Toward Computational Argumentation with Reasoning and Knowledge

Yohan Jo

CMU-LTI-21-004

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
Eduard Hovy, Carnegie Mellon University (Chair)
Alan W. Black, Carnegie Mellon University
Graham Neubig, Carnegie Mellon University
Chris Reed, University of Dundee

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies

© 2021 Yohan Jo

Abstract

Our society today is overloaded with information and opinions. While they are important resources for decision-making for the general public and policy makers in organizations, the staggering amount of them is making people more passive and dependent on information delivered by technologies. This issues an urgent call for technologies that support human decision-making in a truthful way. Truthful language technologies need the ability to reason and use knowledge, beyond memorizing patterns in data and relying on irrelevant statistics and biases. To achieve this goal, our field needs a better understanding of how humans reason and how to incorporate human-like reasoning and knowledge into computational models.

In response to this need, this thesis studies one of the most common communication modes that is full of reasoning: argumentation. The first goal is to provide computational models for analyzing argumentation quantitatively and shedding light on human reasoning reflected in language. The second goal is to incorporate the findings from our study and argumentation theory into computational models via proper knowledge to improve their predictive power and fidelity. By doing so, this thesis argues that integrating reasoning and knowledge, along with argumentation theory, into computational models improves their explanatory and predictive power for argumentative phenomena.

This thesis begins with a study of individual statements in argumentation, in terms of asserted propositions, propositional types, and their effects. We build a model that identifies argumentatively meaningful text spans in text and recovers asserted propositions. Next, we present a methodology for identifying various surface types of propositions (e.g., statistics and comparison) that underlie dialogues and analyzing their associations with different argumentation outcomes (e.g., persuasion). Applying the model on four argumentative corpora, we find 24 generic surface types of propositions in argumentation and their associations with successful editing in Wikipedia, moderation in political debates, persuasion, and formation of pro- and counter-arguments.

We take a step further and study argumentative relations between statements (support, attack, and neutral) by drawing upon argumentation schemes. We first address the challenging problem of annotation in application of argumentation schemes to computational linguistics. We develop a human-machine hybrid annotation protocol to improve the speed and robustness of annotation. By applying it to annotating four main types of statements in argumentation schemes, we demonstrate the natural

affinity between the statement types to form arguments and argumentation schemes. Next, we hypothesize four logical mechanisms in argumentative relations informed by argumentation theory: factual consistency, sentiment coherence, causal relation, and normative relation. Not only do they explain argumentative relations effectively, but incorporating them into a supervised classifier through representation learning further improves the predictive power by exploiting intuitive correlations between argumentative relations and logical relations.

Lastly, we take a closer look at counter-argumentation and study counterargument generation. We first present two computational models to detect attackable sentences in arguments via persuasion outcomes as guidance. Modeling sentence attackability improves prediction of persuasion outcomes. Further, they reveal interesting and counterintuitive characteristics of attackable sentences. Next, given statements to attack, we build a system to retrieve counterevidence from various sources on the Web. At the core of this system is a natural language inference (NLI) model that classifies whether a candidate sentence is valid counterevidence to the given statement. To overcome the lack of reasoning abilities in most NLI models, we present a knowledge-enhanced NLI model that targets causality- and example-based inference. This NLI model improves performance in NLI tasks, especially for instances that require the targeted inference, as well as the overall retrieval system. We conclude by making a connection of this system with the argumentative relation classifier and attackability detection.

The contributions of the thesis include the following:

- This thesis contributes computational tools and findings to the growing literature of argumentation theory on quantitative understanding of argumentation.
- This thesis provides insights into human reasoning and incorporates them into computational models. For instance, logical mechanisms are incorporated into an argumentative relation classifier, and two types of inference are incorporated into counterevidence retrieval through relevant knowledge graphs.
- This thesis draws largely on and borrows frameworks from argumentation theory, thereby bridging argumentation theory, language technologies, and computational linguistics.

Acknowledgments

First, I would like to thank my advisors. Eduard Hovy helped me shape and finish this thesis by showing his passion and appreciation, sharing his vision, and encouraging me throughout. I was lucky to work with Chris Reed, who graciously invited me to his lab in Scotland and offered insights and perspectives from argumentation theory. Carolyn Rosé was very supportive and taught me appropriate skills and perspectives for interdisciplinary research.

I would also like to express my gratitude to the thesis committee for their valuable advice on this thesis. I appreciate having served as a TA for Alan Black's NLP course as well, which was one of the most memorable times at CMU. I received great help from Graham Neubig in his course, where I learned up-to-date modeling techniques and published a paper.

I would like to express my deep appreciation to Jamie Rossi, who graciously spent considerable time chatting with me and listening to me during the darkest time. I thank Bob Frederking and Suzanne Laurich-McIntyre for their help with financial hardships. The Kwanjeong Educational Foundation also sponsored me for six years. I also appreciate the timely service of LTI staff, Stacey Young, Alison Chiocchi, Hannah Muczynski, and more, who allowed a smooth journey.

I want to thank all my collaborators and friends. My deepest thanks go to my family, who always stood by me. Rain or shine.

Contents

Chapter 1

Introduction

Our society today is overloaded with information and opinions. They spread through online social media and news platforms quickly, and people are constantly exposed to them. While they are important resources for decision-making for the general public and policy makers in organizations, the staggering amount of information is making people more passive and dependent on information delivered by technologies. For instance, when one watches a video on YouTube, the technology suggests what to watch next (Figure 1.1). When one searches for information on Google, the technology displays up front the document that it believes is the most relevant. When one enters the news portal of Microsoft Bing, the technology delivers news articles that it believes may be of interest to the user. In the circumstance where information consumers are gradually losing control in the flood of information, how can they be protected from delivered information that is potentially ungrounded or partial (Noble, 2018)?

A key is reasoning—justifying a belief using logic and existing information (Kompridis, 2000). The practice of reasoning has been emphasized throughout human history, from ancient Greek philosophers to the Critical Thinking Movement in education (Schwarz and Baker, 2016). However, the importance of reasoning applies not only to humans but also to language technologies. The overabundance of information has made it a norm that machine learning models become more complex and trained on larger data. Despite the great achievements of such models in

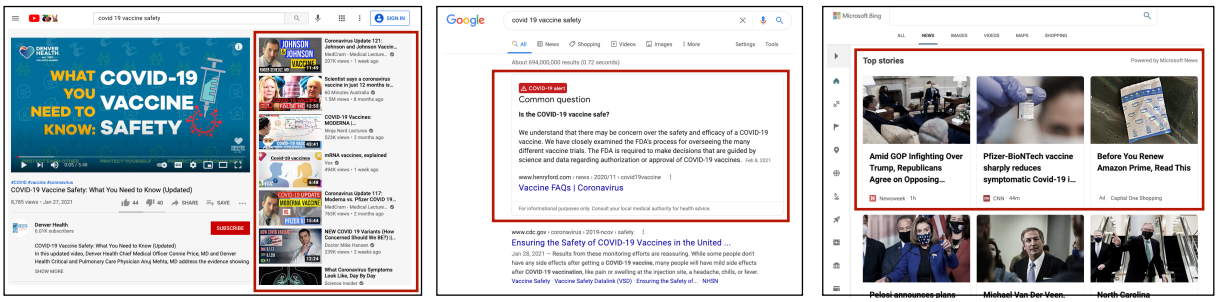


Figure 1.1: Examples of information delivery by technologies. (Left: YouTube, middle: Google, right: Microsoft Bing)

many NLP fields, naive models have shown problematic behaviors, such as relying on spurious statistics and biases in data when making important decisions (Vig et al., 2020; Tan and Celis, 2019; Utama et al., 2020). Due to the impact technologies have on what information people consume and base their decisions on, there is an urgent call for technologies that support human decision-making in truthful ways. For this, language technologies need the ability to reason and use knowledge, beyond memorizing patterns in data. To achieve this goal, our field needs a better understanding of, first, how humans reason and, second, how to incorporate human-like reasoning and knowledge into computational models to process information.

In response to the need, this thesis studies one of the most common communication modes in daily life that is full of reasoning: argumentation. Simply put, argumentation is the process of drawing a conclusion through reasoning and information. Argumentation is pervasive. People share and discuss opinions over diverse issues with many goals, such as persuading other people, learning different perspectives, and accomplishing a collaborative task. The computational study of argumentation has recently garnered a lot of popularity in the subfield of NLP *computational argumentation* or *argument mining*. For instance, the annual Workshop on Argument Mining has been held for eight years, and major NLP conferences have dedicated sessions for argument mining (usually coupled with sentiment analysis) for the past three years. Research in this field has addressed various problems, such as text segmentation for elementary argument units, classifying argumentative relations between statements, and generating pro- and counter-arguments. More practically, IBM has developed an AI system called Project Debater that debates with people on different topics (Slonim et al., 2021; Reed, 2021). Despite drastic advances in this field, most approaches still focus on data annotation and model training without a deep consideration of how humans reason and how to incorporate reasoning and knowledge into models. Such models have shown to rely overly on superficial cues (Niven and Kao, 2019; Allaway and McKeown, 2020; Opitz and Frank, 2019) and have the potential risk of segregating unpopular opinions in training data.

The goal of this thesis is to overcome this limitation by enlarging our understanding of human reasoning reflected in argumentative text and developing effective methods for integrating this insight into computational models. In doing so, this thesis actively draws upon argumentation theory to borrow useful frameworks from the long-standing field. Ultimately, this thesis aims to argue that integrating reasoning and knowledge, along with argumentation theory, into computational models improves their explanatory and predictive power for argumentative phenomena.

To motivate the problems addressed in this thesis more concretely, we use a snapshot of argumentation between two online users on the recent impeachment of Donald Trump, as shown in Figure 1.2. Arguer1 (top) argues that Republicans should convict Trump and ban him from running for office in the future, and Arguer2 (bottom) responds to this argument with the goal of changing Arguer1's viewpoint. For now, we keep our description rather informal for illustration purposes. More formal and technical terms and concepts will be introduced in the next section.

To understand this argumentation and how the arguers reason, we first need a proper understanding of individual sentences. The sentences in this argumentation seem to contribute to the development of the argumentation by putting forward assertions. But are all these sentences equally relevant and meaningful argumentatively? How about a formal debate where a lot of

CMV: The best strategy forward for Republicans would have been to convict Trump and ban him from running for office in the future.

[1]Right now Trump holds the balls of the Republican Party. [2]He's got 30% of the base locked up and he can push them in any direction he wants, including a 2024 Presidential Bid. [3]That would absolutely be the worst case scenario for Republicans. [4]And even if he doesn't end up running in 2024, you can guarantee he will dangle that 2024 run out there just to get leverage on Republican leadership.

[5]That all goes away if just ten Republican senators voted to convict and then ban him from running for future office. [6]All of the headaches, the hand-wringing, the groveling, the bowing down to Trump, all of it goes away if he has zero political leverage. [7]If he has no chance at ever becoming President again his supporters will dry up and Trump will focus on other ventures, just like he does every time a business venture of his fails.

[8]No one in Republican leadership wants Trump to run in 2024. [9]Everyone is deathly afraid that he is going to fracture the party. [10]So why didn't they convict him? [11]If he was convicted, what leverage would he have to fracture the party? [12]I posit that his political position becomes extremely tenuous if that were to have happened. [13]Whereas since he was acquitted it will only embolden him more to hold on to as much of the party's base as possible.

[14]One counterpoint that I foresee coming up is that if Trump was convicted and barred from future office that his core base (about 30% of the party) would become disenfranchised with Republicans and stay home. [15]My argument against that is voter's memories are short so they should not presume to lose those votes if Trump is no longer on the table.

> He's got 30% of the base locked up and he can push them in any direction he wants, including a 2024 Presidential Bid.

[16]where did you get that number? [17]Trump's popularity in the Republican party continues to be around 80%. [18]That's massive. [19]It's a supermajority, not a minority, which 30% would be. [20]81% of Republican respondents give him positive marks. [21]Trump was at 77% approval among Republicans on Jan. 7 and 74% on Jan. 25.

[22]Even after the insurrection 59% of Republican voters said they want Trump to play a major role in their party going forward. [23]Him not being able to run will not change that.

> One counterpoint that I foresee coming up is that if Trump was convicted and barred from future office that his core base (about 30% of the party) would become disenfranchised with Republicans and stay home.

[24]Again, where is this 30% number from? [25]The number of Republicans who back Trump is far more than 30%.

[26]And yes, Republicans would absolutely lose votes if they turned on Trump. [27]He would make sure of it. [28]He's already having his people start to run for office. [29]Imagine how many of his stooges would run if the Republican party was seen as "betraying" Trump. [30]Don't forget that he's one of the most popular figures within the Republican party ever.

> My argument against that is voter's memories are short so they should not presume to lose those votes if Trump is no longer on the table.

[31]You underestimate the anger and passion that Trump can inspire. [32]The primaries are in less than 2 years. [33]People won't forget.

Figure 1.2: Example argumentation from ChangeMyView. Each sentence is prepended with a sentence number.

utterances are used for moderation and structure rather than for contributing to the content of argument itself? Further, sentence 9 uses a rhetorical question and sentences 28–29 use imperatives, both of which do not assert any content in their grammatical form. What are the pragmatic, hidden assertions they are making? And more generally, how can we decompose an argument into argumentatively relevant building blocks and recover the meaning of each? These are fundamental questions, since most computational models assume and are built on some kind of basic units of arguments, but usually in a simplistic way of text segmentation. And the adequacy of these

units and their meaning is crucial for proper assessment of the argument and the transparency of computational models. We will cover these topics in Part I Chapter 2.

We also observe that the two arguments use various types of sentences that may have different rhetorical functions and effects. For instance, sentences 2, 20, and 21 use specific numbers and statistics, and sentences 5–7 make predictions about hypothetical circumstances. Sentences 9 and 31 express emotion, and sentences 16 and 31 directly address the other arguer. A naturally occurring question is: how do these different types of sentences affect the outcome of argumentation? What types are positively correlated with successful decision making or persuasion? And even before that, what types are commonly used in argumentation at all and how can we automatically identify those types in an empirical and data-driven way? Quantitative investigation of these questions will shed light on rhetorical devices in human reasoning. We will address these questions in Part I Chapters 3–4.

Next, we shift our focus from individual sentences to the interaction between sentences. Sentences in the example argumentation interact with one another to build the support or attack relations. For example, sentence 10 is supported by sentences 8–9, and sentence 29 is by sentence 30. On the other hand, sentence 14 is attacked by sentence 15, and sentences 17 and 31 attack Arguer1's points quoted from the original argument. Identifying the argumentative relations among sentences is key to understanding and assessing the argumentation. It is also a core part in many NLP tasks, such as Q&A and fact verification. And especially in the era of information overload, this technology is important to verify whether a piece of information is supported or attacked by evidence. How do humans usually reason about the argumentative relations? And how can we incorporate the same reasoning into a computational model? We address these questions in Part II Chapters 5–6, by deeply drawing upon the mechanisms of logical reasoning and argumentation theory.

Lastly, we take a closer look at how one refutes an argument, which may inform how to build a debating machine or feedback generator. In the example argumentation, Arguer2 refutes Arguer1's argument by first choosing specific points to attack and then by presenting counterevidence to each point. Assuming we want to build an automated feedback system, e.g., for legal reasoning or essay scoring, can we automatically detect "attackble" points in an argument? Once we identify specific points to attack, how do we find counterevidence to each point? Although counterevidence could be simple factual contradiction as in sentence 17, counterevidence often requires complex reasoning as in sentences 31–32. In Part III Chapter 7–8, we will discuss how to find attackable points in an argument and counterevidence to each point from different sources of documents, by focusing on example- and causality-based inference and incorporating relevant knowledge graphs.

In the remainder of this chapter, we introduce theoretical background and formal terminology this thesis draws upon. And then we present the structure of this thesis and emphasize main contributions.

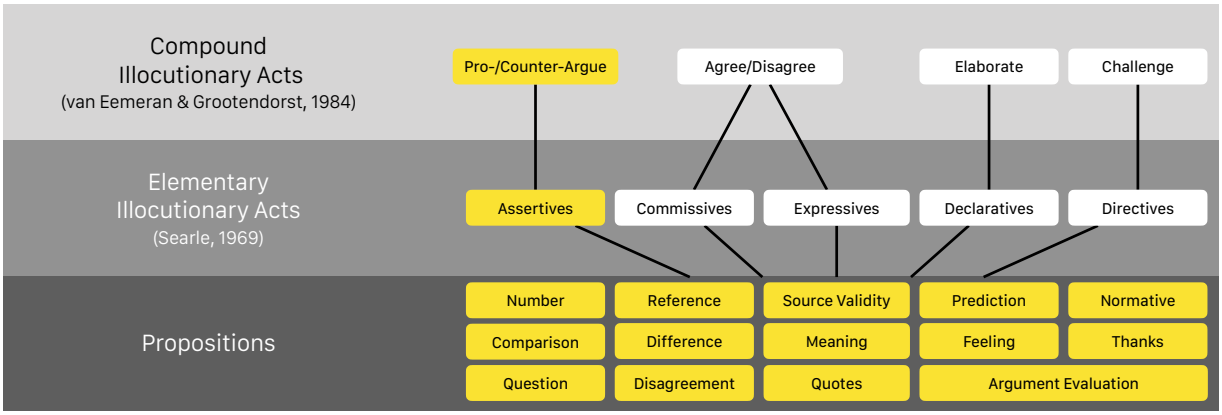


Figure 1.3: Concept hierarchy of speech acts. An upper-level concept consists of lower-level concepts. Small boxes are example elements of the concepts that are relevant to argumentation. The elements covered in this thesis are highlighted in yellow.

1.1 Theoretical Background

This thesis draws largely upon the terminology and view of argumentation in informal logic. This section begins with the definition and structure of argument from the pragma-dialectics perspective (van Eemeren and Grootendorst, 1984). Next, it discusses the modern view of how argumentative relations (the relations between statements within an argument) are defined and assessed on the basis of argumentation schemes (Walton et al., 2008).

1.1.1 Definition and Structure of Argument

An **argument** is defined as a claim-reason complex (Hitchcock, 2007), consisting of (i) an act of concluding, (ii) one or more acts of premising that assert propositions in favor of the conclusion, (iii) and a stated or implicit inference word that indicates that the conclusion follows from the premises. A simple example is as follows[1]:

> **Conclusion:** "All humans should be vegan."
> **Inference word:** ("Because")
> **Premise 1:** "The meat production industry is unethical."
> **Premise 2:** "A vegan diet can avoid meat production."

Hence, an argument consists of multiple statements. A statement is called **conclusion** or **claim** if it is supported by other statements, and the supporting statements are called **premises**. The claim-hood and premise-hood are not intrinsic features of a statement; they are rather determined by the relationships between statements. We also distinguish arguments from argumentation. **Argumentation** is the act of making arguments, whereas an argument is a set of actual statements.

In the pragma-dialectics theory (van Eemeren and Grootendorst, 1984), argumentation is viewed as an illocutionary act. In the above example, putting forward premise 1 and premise 2

[1] Adapted from https://www.kialo.com/all-humans-should-be-vegan-2762

is the illocutionary act of supporting the conclusion. The theory further distinguishes between pro-argumentation and counter-argumentation. **Pro-argumentation** is to put forward premises to support the conclusion, whereas **counter-argumentation** is to utter premises to defeat the conclusion. For instance, putting forward

> **Premise 3:** "A vegan diet lacks specific essential nutrients."

can be seen as the illocutionary of counter-argumentation toward the conclusion above.

Viewing argumentation as an illocutionary act, however, gives rise to a need for resolving some conflicts with the illocutionary acts as defined in the speech act theory (Searle, 1969). One of the main differences is that an instance of argumentation may involve multiple sentences (e.g., premises 1–2) and thus multiple instances of more basic illocutionary acts from the speech act theory (e.g., multiple assertives). Hence, the pragma-dialectics theory introduces the important concept of **compound illocutionary act**, which may consist of multiple instances of **elementary illocutionary acts** defined in the speech act theory. Pro- and counter-argumentation are compound illocutionary acts and placed at a higher level than elementary illocutionary acts. This distinction is illustrated as the top two levels of the concept hierarchy in Figure 1.3.

We next focus on the elementary illocutionary act. According to the speech act theory, performing an illocutionary act is often accompanied by a propositional act. A **proposition** is the actual meaning of the content in the utterance and, in the speech act theory, consists of an object to be described and a description of it that has a truth value. Propositions are distinguished from illocutionary acts, and the same proposition may be used for different illocutionary acts. For example, the proposition "X means Y", which is common in argumentation, can be used with different elementary illocutionary acts, as in:

> **Assertives:** "X means Y" (as a factual statement)
> **Declaritives:** "We define X as Y from now on"
> **Commissives:** "What do you mean by X?"

This distinction between elementary illocutionary acts and propositions is illustrated as the bottom two levels of our concept hierarchy (Figure 1.3). Pragma-dialectics maintains that argumentation is an act of putting forward asserted propositions.

While different types of compound illocutionary acts and elementary illocutionary acts have been identified and studied relatively well by pragma-dialectics and the speech act theory, less has been established about the types of propositions that play main roles in argumentation. We take a close look at a methodology for identifying various surface types of propositions and understanding their roles in Part I.

1.1.2 Definition and Assessment of Argumentative Relations

Going back to Hitchcock's definition and the pragma-dialectic view, an argument consists of asserted propositions. Asserted propositions interact with one another to form pro- or counter-arguments. To define and assess these argumentative relations, we take the view of informal logic. Unlike formal logic, which uses deductive reasoning based on formal language (e.g., logical expressions), informal logic has been developed with a focus on analysis and assessment of

everyday argument. Arguments in our daily lives are often defeasible rather than deductive. A classical example is as follows:

> **Conclusion:** "Tweety can fly."
> **Premise 1:** "Tweety is a bird."
> **Premise 2:** "Birds generally fly."

This argument can be considered reasonable and acceptable. However, if we have the additional information that Tweety is the name of a penguin, the premises no longer make the argument sound. That is, the argument is defeasible and subject to defaults and exceptions. This is the nature of arguments we handle in this thesis. We do not assume that argumentative relations (support, attack, and neutral) are logically deductive. Instead, we rely on potentially subjective human intuitions and assume that the argumentative relation between asserted propositions depends on how they would generally be accepted by people.

To describe and assess defeasible arguments more systematically, argumentation theory has developed argumentation schemes (Walton et al., 2008). **Argumentation schemes** specify reasoning patterns that are commonly used in daily arguments and that are generally accepted as reasonable but subject to defaults depending on exceptions and additional information. Each scheme is a template that represents a specific type of inferential link between a conclusion and premise(s). For example, the scheme *argument from consequences* has the following form:

> **Conclusion:** "We should do X."
> **Premise:** "X may lead to a positive consequence Y."

Another scheme *argument from cause-to-effect* has the following form:

> **Conclusion:** "Y may occur."
> **Premise 1:** "X has occurred."
> **Premise 2:** "Y generally occurs if X occurs."

More than 80 argumentation schemes have been identified in the literature.

Although it has not been firmly established in argumentation theory, each argumentation scheme often can accommodate both support and attack relations. For *argument from consequences*, if we slightly modify the premise to "X may lead to a negative consequence Y", then the new statement and the conclusion still have a very similar form to *argument from consequences* except that this statement counters the conclusion. Similarly, if we slightly modify premise 2 in *argument from cause-to-effect* to "Y generally does not occur if X occurs", then this statement and the conclusion have a similar form to *argument from cause-to-effect*, but the statement counters the conclusion. This thesis draws largely upon argumentation schemes and this property in modeling human reasoning about argumentative relations, argumentative relation classification, and couterevidence retrieval (Part III Chapter 6 and Part III Chapter 8).

1.2 Thesis Overview

The goal of this thesis is to develop a theory about how humans reason and how to incorporate human reasoning and knowledge into computational models in the context of argumentation. We begin with the basic unit of argument—proposition—and investigate its meaning, types, and effects in argumentation. We next investigate argumentative relations among the building blocks of argument in terms of the way humans reason about their relations and methods for incorporating the reasoning into computational models. Lastly, we zoom in on counter-argumentation and examine what characteristics make sentences attackable and how to find counterevidence effectively by integrating certain types of reasoning and related knowledge. For each of the subjects, we draw insights into human reasoning reflected in argumentative language and build a suite of computational models informed by these insights.

In Part I, we study individual propositions in terms of their meaning, types, and effects. In Chapter 2, we present a cascade model that takes an utterance as input and returns asserted propositions in the utterance. Most argument systems obtain the basic building blocks of an argument through simple text segmentation, which produces fragmented texts and non-assertions whose meaning and argumentative contribution are unclear. As a result, it is obscure how their meaning is interpreted in downstream components, reducing the transparency of the system. Our cascade model identifies argumentatively meaningful text spans from the input utterance, reconstructs fragmented text spans, and recover implicitly asserted propositions from reported speech, questions, and imperatives.

In Chapters 3–4, we present a methodology for identifying different types of propositions that underlie argumentative dialogue and analyzing the association between these types and argumentation outcomes. Unlike most prior work in rhetoric, marketing, and communication sciences, we derive various types of propositions from large argumentation corpora in an empirical and data-driven way and quantify them for further analyses. To that end, in Chapter 3, we present and evaluate a model that aims to identify latent types of propositions in a given set of dialogues. In Chapter 4, we apply this model to four corpora of argumentative dialogue and identify 24 main surface-level types of propositions that are generic in argumentation, such as references, definitions, and comparisons. We further examine how these types are correlated with various argumentation outcomes, such as successful decision-making on editing in Wikipedia, moderation bias in political debates, and effective persuasion in deliberative dialogue.

In Part II, we dive into the argumentative relations between statements, i.e., their support, attack, and neutral relations. We draw upon argumentation schemes to analyze argumentative relations. To better understand argumentation schemes, in Chapter 5, we annotate four major types of statements used in argumentation schemes, namely, normative, prediction, desire, and reported speech. Annotating argumentation schemes is challenging due to their fuzzy nature, subjective interpretations, and logical reasoning. Regardless, not many methodological protocols have been proposed and explored for robust and efficient annotation. We present a human-machine hybrid protocol, where a machine is trained on a subset of human annotations and serves as an additional annotator to process easy instances and validate human annotations. We present a desirable property appearing in argumentation data for this protocol and demonstrate that this protocol

improves the speed and robustness of the annotation. We also analyze the tendency of certain statement types to form more natural arguments and argumentation schemes in debates.

In Chapter 6, we examine how humans reason about argumentative relations between statements and how to incorporate this reasoning into a computational model. We hypothesize four logical mechanisms that may be used by humans, namely, factual consistency, sentiment coherence, causal relation, and normative relation. Our operationalization of these mechanisms explains argumentative relations well without supervised learning, signifying their effectiveness in determining argumentative relations. We incorporate these mechanisms to a supervised classifier using representation learning, which further improves prediction performance and shows an intuitive connection the model makes between argumentative relations and logical relations.

In Part III, we take a closer to at counterarguments and examine counterargument generation. We see counterargument generation as a three-step process: detect "attackble" points in the given argument, find counterevidence to each point, and combine this evidence to make a fluent and coherent argument. This thesis addresses the first two steps. In Chapter 7, we define the attackability of sentences in arguments in terms of how addressing them affects persuasion outcomes. We explore two computational methods to detect attackable sentences. One is based on neural representations of sentences and the other on interpretable hand-crafted features informed by argumentation theory. This work is the first large-scale analysis of this problem in NLP.

In Chapter 8, we present a system that finds counterevidence to a given statement. It retrieves relevant documents from different sources, ranks them, selects a set of candidate sentences of counterevidence, and classifies each sentence as valid or invalid counterevidence. The last component, the core of this system, uses natural language inference (NLI). However, many NLI models show a lack of reasoning abilities and fail to capture instances that require complex inference. We enhance NLI models by focusing on example- and causality-based inference and incorporating relevant knowledge graphs into NLI models. The knowledge-enhanced NLI models achieve higher performance in NLI tasks, inference tasks, and the counterevidence retrieval task.

1.3 Contributions

Overall, this thesis advocates for the study of human reasoning and methods for incorporating reasoning and knowledge into computational models. It showcases analyses of human reasoning in the particular context of argumentation and tackles important problems in argumentation using computational approaches informed by reasoning mechanisms and argumentation theory. Specific contributions made in the thesis include:

- A cascade model for recovering asserted propositions (either explicitly or implicitly) in argumentative discourse. This recovery is a missing link in our field between segmenting text into meaningful units of argument and using the units for downstream tasks. The recovered assertions of these units make an argument system more transparent and accountable for its decisions.
- A methodology for identifying various types of propositions in argumentative discourse and analyzing their associations with argumentation outcomes. Applying this methodology

to four argumentation corpora reveals 24 generic surface-level types of propositions in argumentation. Four case studies demonstrate these types are highly associated with various argumentation outcomes, including the success of decision-making for Wikipedia editing, the success of persuasion, moderation in debates, and the support or attack relations between propositions.

- A human-machine hybrid annotation protocol for annotating statement types in argumentation schemes. By training and utilizing a machine annotator, this hybrid protocol expedites the annotation process and makes the annotation more robust than human-only annotation. The corpus study based on the resulting annotations reveals the affinity between statement types to form arguments and argumentation schemes. The study also demonstrates different argument styles of U.S. presidential candidates in 2016.
- A method for examining important logical mechanisms in argumentative relations and a representation learning method to incorporate the mechanisms into classifiers. The study reveals that the factual consistency, sentiment coherence, causal relation, and normative relation between two statements are effective mechanisms that determine the argumentative relations between the statements. The mechanisms integrated into a classifier through the proposed representation learning method further improve the classifier's prediction accuracy.
- Methods for detecting attackable sentences in arguments. The study proposes a computational way of measuring attackability based on persuasion outcomes. The first model based on neural representations of sentences shows that modeling the attackability of individual sentences improves the accuracy of predicting persuasion outcomes. The second model based on hand-crafted features demonstrates that different characteristics of sentences are correlated with their attackability. A large-scale analysis reveals some interesting characteristics of attackable sentences.
- A counterevidence retrieval system enhanced by a knowledge-integrated NLI model. An effective method is proposed for incorporating causality- and example-based inference and relevant knowledge graphs into NLI. It improves performance in general NLI tasks, especially for instances that require the targeted inference, and counterevidence retrieval.

Part I

Propositions: Meaning, Types, and Effects

As a starting point of studying argumentation, Part I examines individual propositions in argument in terms of their meaning, types, and effects. In Chapter 2, we present a cascade model that recovers either explicitly or implicitly asserted propositions in argumentative text. This model identifies text spans that serve as basic argumentative units, reconstruct fragmented text spans, and recover implicitly asserted propositions in reported speech, questions, and imperatives. In Chapters 3–4, we present a methodology for identifying different types of propositions that underlie argumentative dialogues and analyzing the associations between these types and argumentation outcomes. Specifically, we first present and evaluate a model that aims to learn latent, surface-level types of propositions in a given set of dialogues in Chapter 3. And in Chapter 4, we apply this model to four corpora of argumentative dialogues to identify underlying types of propositions, and examine how these types are associated with various argumentation outcomes, such as Wikipedia edits, moderation, persuasion, and formation of pro-/counter-arguments.

Chapter 2

Extracting Asserted Propositions

According to the pragma-dialectics theory (van Eemeren and Grootendorst, 1984), argumentation is the compound illocutionary act of putting forward premises to support or attack an expressed opinion. Furthermore, this act consists of multiple instances of assertives (also called representatives)—a main type of elementary illocutionary acts stating that something is the case (Figure 1.3). In other words, argumentation is the process of asserting propositions to support or attack another asserted proposition. Therefore, the first and foundational step for identifying pro-arguments and counter-arguments is to extract asserted propositions from dialogue. According to our data of 2016 U.S. presidential debates and online commentary, roughly 90% of text comprises propositions that are asserted. Among them, 89% are explicitly asserted and the other 11% are implicitly asserted (e.g., questions and reported speech).

In most work in NLP, asserted propositions are usually substituted by *argumentative discourse units (ADUs)* obtained via segmenting text into smaller grammatical pieces (usually clauses). This approach may yield text segments that lack semantic information necessary for downstream tasks. It may also fail to capture propositions that are asserted implicitly. For instance, reported speech and rhetorical questions play important roles in dialogical argumentation, contributing propositional contents that are not apparent in their surface forms. However, prior argument mining research has paid little attention to extracting these implicit propositions, resulting in missing information necessary for identifying relations between propositions. Text segments without their intended meaning being recovered also make an argument system less transparent and accountable for its downstream decisions.

In this chapter, we present a model to tackle this fundamental but understudied problem in computational argumentation: extracting asserted propositions. Our cascade model aims to extract complete, asserted propositions by handling anaphora resolution, text segmentation, reported speech, questions, imperatives, missing subject reconstruction, and revision. We formulate each task as a computational problem and test various models using a corpus of the 2016 U.S. presidential debates. We show promising performance for some tasks and discuss main challenges in extracting asserted propositions.

2.1 Introduction

Most argument mining models for identifying the argumentative structure (pro- and counter-arguments) of argumentative text build upon elementary text spans that serve argumentative functions, such as premise and conclusion. In the pragma-dialectics theory (van Eemeren and Grootendorst, 2004) and argumentation theory in general (Blackburn, 2016), it is commonly accepted that these building blocks are *asserted* propositions, i.e., assertions that are either true or false. Despite their foundational role, however, extracting asserted propositions from text has been little studied in computational argumentation. Instead, most models rely on argumentative discourse units (ADUs)—text spans obtained by surface-level text segmentation, usually at clause levels (Stede et al., 2016; Al-Khatib et al., 2016). In what follows, we discuss limitations of ADUs that potentially impinge upon subsequent argument mining processes, and then describe our approach.

One limitation of ADUs is that they may lack important semantic information, such as the **referents of anaphors** and the **subject of an incomplete clauses**, necessary for subsequent argument mining steps. For example, for two consecutive text segments "Alice complained to Bob" and "He is upset", if we do not know "he" refers to Bob, it would be confusing whether the first segment supports the second or vice versa. In another example, suppose "Alice was faithful to Bob, keeping the secret" is split into two ADUs, each associated with the main clause and the adverbial participle, respectively. While mere text segmentation leaves the subject of the participle (Alice) missing, tracing and reconstructing the subject makes it clear that the participle supports the main clause. As illustrated in these examples, anaphora resolution and subject reconstruction recover semantic information that has potential benefits for argument mining systems.

Moreover, ADUs include locutions that are seemingly not assertives, such as questions and imperatives used as rhetorical devices, which may seem inconsistent with the theory. In fact, questions, imperatives, and reported speech in argumentation often assert propositions implicitly. Therefore, in order to understand certain argumentation and identify pro-/counter-arguments properly, locutions in argumentation should not be taken literally in their surface forms; instead, we need to go further and understand what propositions are implicitly asserted and argumentatively relevant in those locutions. The following example dialogue illustrates how questions, reported speech, and imperatives assert propositions implicitly in argumentation.

A : "All human should be vegan." (2.1)

"Look at how unethical the meat production industry is." (2.2)

"Environmental scientists proved that vegan diets reduce meat production by 73%." (2.3)

B : "Well, don't vegan diets lack essential nutrients, though?" (2.4)

In this dialogue, speaker *A* is supporting conclusion 2.1 using sentences 2.2 and 2.3, whereas speaker *B* is attacking the conclusion using sentence 2.4. Sentence 2.2 is an imperative, but in this argumentation, it is *asserting* that the meat production industry *is* unethical. In sentence 2.3, the primary proposition asserted in support of the conclusion is the content of this reported speech—"vegan diets reduce meat production by 73%"; the "environmental scientists" is presented as the source of this content in order to strengthen the main proposition in this sentence. Lastly, sentence

2.4 is in question form, but it is in fact *asserting* that vegan diets *lack* essential nutrients. These examples suggest that properly understanding arguments requires comprehension of what is meant by questions, reported speech, and imperatives, that is, what they assert implicitly.

To solve this problem, we present a cascade model that aims to extract propositions from argumentative dialogue, recovering important semantic information and implicitly asserted propositions. Our model consists of seven modules, namely, anaphora resolution, locution extraction, reported speech, question, imperative, subject reconstruction, and revision (Figure 2.2). For each module, we formulate the task as a computational problem and test various models to solve it. Our analyses and evaluation are based on the transcripts of the 2016 U.S. presidential debates and online commentary (Visser et al., 2019).

2.2 Related Work

In computational argumentation, the basic unit of an argument is often called an argumentative discourse unit (ADU). In this section, we first review how existing studies define and obtain ADUs from text, and then some theoretical framework to obtain asserted propositions from ADUs.

2.2.1 From Text to ADUs

In most studies, ADUs are obtained via text segmentation. While some studies leave the choice of the boundary of an ADU to the annotator's judgment (Stab and Gurevych, 2014), many studies employ a set of syntactic rules as a basis. For instance, an ADU can be as fine-grained as a phrase that plays a discrete argumentative function (Stede et al., 2016). In other cases, an ADU may be a clause (Peldszus and Stede, 2015) or a series of clauses that must include a subject, a verb, and an object if necessary (Al-Khatib et al., 2016).

Based on annotated ADUs, some studies have proposed methods for automatically segmenting ADUs using machine learning. This task is commonly formulated as tagging each word in the text as either the beginning, inside, or outside of an ADU (BIO tagging). The tagging has been incorporated into an end-to-end argument mining (Eger et al., 2017) or conducted separately on various domains (Ajjour et al., 2017). Instead of tagging, a retrieval approach has also been used, where candidate ADUs are generated and the best is retrieved (Persing and Ng, 2016a).

All these approaches to ADU segmentation share most of the concerns mentioned in Section 2.1. For better-informed argument mining, we need to go further to obtain asserted propositions from ADUs, and thus a relevant framework will be discussed in the following section.

2.2.2 From ADUs to Asserted Propositions

Following the speech act theory (Austin, 1962; Searle, 1969), the connection between text segments and propositions can be modeled as illocutionary acts—the application of particular communicative intentions to propositional contents, e.g., *asserting* that a proposition is true, or *questioning* whether it is true. Focusing on argumentatively relevant speech acts (van Eemeren and Grootendorst, 1984), Inference Anchoring Theory (IAT) (Reed and Budzynska, 2011) explains

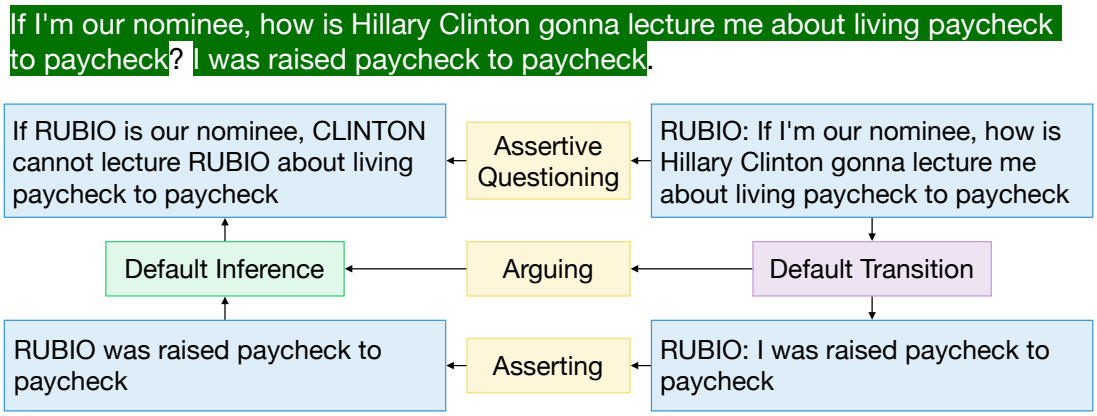


Figure 2.1: A snippet of the US2016 corpus. The top text is the original utterance. The blue boxes on the right are locutions, which are also highlighted with green on the utterance. The blue boxes on the left are propositions anchored in the locutions, via illocutionary acts (yellow boxes).

how propositional contents and the argumentative relations between them are anchored in the expressed locutions by means of illocutionary connections.

IAT has been applied to annotate argumentative dialogues of various kinds, including the corpus of 2016 U.S. presidential debates and online commentary (Section 4.2.4). IAT annotation comprises, amongst other things, segmenting the original text into locutions[1], identifying the illocutionary force instantiated by the locution, and reconstructing its propositional content asserted; an example snippet is shown in Figure 2.1. Each locution generally conveys one proposition. Conjuncts conjoined by a conjunction and conditional clauses may be separated if they each fulfill a discrete argumentative function. In addition, punctuation, discourse indicators, and epistemic modalities (e.g., "I think") should be excluded. Anaphoric references are typically reconstructed, resulting in full grammatical sentences understandable without context.

2.3 Data

We use the US2016 corpus (Visser et al., 2019), which contains transcripts of televized debates for the 2016 U.S. presidential election and reaction to the debates on Reddit. Specifically, the corpus includes the first Republican candidates debate for the primaries, the first Democratic candidates debate for the primaries, and the first general election debate. It also include Reddit discussions on these debates.

All dialogues have been manually segmented and annotated with locutions, illocutionary connections, and asserted propositions based on IAT (Reed et al., 2016) (Figure 2.1). The corpus was annotated by 4 annotators, yielding an overall Cohen's κ of 0.610 (considered substantial agreement). The corpus is further annotated with support (inference) or attack (conflict) relations between propositions (the green box "Default Inference" in Figure 2.1). We downloaded the annotations from the corpus webpage and separately scraped the original dialogues.

[1] Analogous to ADUs. We use the terms interchangeably.

For data preparation, we aligned each locution with the original dialogue; e.g., in Figure 2.1, the locutions (in the right blue boxes) are aligned with the original utterance (at the top) using string matching. This allows us to build a model to extract locutions from utterances, and asserted propositions from locutions. The processed corpus includes 2,672 utterances and 8,008 locutions.

This corpus is ideal for our analysis, since the debates cover a wide range of political issues and are interactive among the debaters. The debates also accommodate diverse rhetorical devices, such as questions, reported speech, and imperatives, in both formal (main debates) and informal (Reddit) debate settings. While most parts of our system are based on this corpus as explained so far, some parts need additional processing or additional data. They will be described in the respective sections.

2.4 Propositions in Argumentation

Before we move on to the cascade model, in this section, we consider what kinds of processing are necessary in order to extract asserted propositions. Specifically, we take a close look into the following 9 aspects, based mostly on the IAT annotation guidelines[2]:

- Anaphora resolution
- Extraction of proposition content and its source from reported speech
- Extraction of propositional content from questions
- Extraction of propositional content from imperatives
- Removal of non-propositional expressions, discourse markers, and epistemic modalities
- Qualification/hedges
- Time/location
- When/if-statements
- Segmentation granularity and reconstruction

Anaphora resolution: Most NLP tools (e.g., Stanford CoreNLP, AllenNLP, SpaCy) has the functionality of anaphora resolution. In order to see if it is enough to apply these tools or something more should be considered, we ran Stanford CoreNLP on locutions in the US2016 and compared the results with the annotated propositions. Besides accuracy problems inherent in the NLP tool, we find five main sources of errors.

First, the NLP tool cannot resolve speakers and hearers:

> **Original:** "In Florida, they called me Jeb"
> **Anaphora resolved:** "In Florida, they called Chris Jeb"
> **Annotation:** "In Florida, they called BUSH Jeb"
>
> **Original:** "He provided a good middle-class life for us"
> **Anaphora resolved:** "my late father provided a good middle-class life for us"
> **Annotation:** "CLINTON's late father provided a good middle-class life for his family"

[2] https://typo.uni-konstanz.de/add-up/wp-content/uploads/2018/04/IAT-CI-Guidelines.pdf

Original: “But I can’t vote for someone ...”
Anaphora resolved: “But Bernie can’t vote for someone ...”
Annotation: “I can’t vote for Bernie Sanders ...”

In the first example, the speaker is Bush but “me” is wrongly replaced with “Chris”. In the second example, “He” is replaced with “my late father”, where “my” should be further replaced with “Clinton”. In the third example, the speaker is a Reddit user, but “I” is wrongly replaced with “Bernie”. These problems may be rectified by rule-based resolution of speakers and hearers, e.g., first and second singular pronouns (“I”, “me”, “you”, “your”, etc.) are replaced with speaker or hearer names.

Second, case/POS tag mismatches happen:

Original: “You are a successful neurosurgeon”
Anaphora resolved: “Dr. Carson are a successful neurosurgeon”
Annotation: “CARSON is a successful neurosurgeon”

Original: “I don’t think he cares too much about war crimes.”
Anaphora resolved: “I don’t think Trump ’s cares too much about war crimes.”
Annotation: “I don’t think TRUMP cares too much about war crimes”

In the first example, “You” is replaced with “Dr. Carson” but “are” remains the same. In the second example, “he” is replaced with “Trump ’s”, because the NLP tool recognized “Trump ’s” as the representative mention. These problems may be rectified by the postprocessing of case/POS tag matching. We could also replace pronouns with only the head word of the reference.

Third, the NLP tool tends to resolve non-pronouns as well:

Original: “We left the state better off”
Anaphora resolved: “We left Florida better off”
Annotation: “We left the state better off”

In this example, “the state” is replaced with “Florida”. This is correct, but the corpus usually does not resolve non-pronouns, making it difficult to evaluate our model. This conflict may be rectified by skipping resolution of non-pronouns.

Fourth, the corpus does not resolve pronouns following an antecedent within the same sentence:

Original: “Bernie can’t easily answer whether he is a capitalist”
Anaphora resolved: “Bernie can’t easily answer whether Bernie is a capitalist”
Annotation: “Bernie Sanders can’t easily answer whether he is a capitalist”

In the first example, “he” has not been resolved in the annotation. Similarly, in the second example, “they” has not. Resolving all pronouns might be better for argumentation structure analysis. But for evaluation purposes, a quick workaround of this issue would be to apply the same rule to anaphora resolution, that is, only the first occurring pronoun is resolved in each sentence.

Lastly, generic “you” and “they” are left unresolved in the corpus:

Original: “Rather than trying to fix the broken system, they would rather break it entirely”

Anaphora resolved: "Rather than trying to fix the broken system, the Bernie supporters would rather break it entirely"
Annotation: "Rather than trying to fix the broken system, they would rather break it entirely"

There seems to be no easy solution to this issue for now.

Reported speech: In argumentation, it is common to quote someone's speech or beliefs (e.g., authority claims). Reported speech consists of **speech content** that is borrowed from a **speech source** external to the speaker. Speech content can be a direct quote of the original utterance or an indirect, possibly paraphrased utterance. Reported speech is a common rhetorical device in argumentation and performs various functions, including:

- Appeals to authority by referencing experts or rules (Walton et al., 2008) (e.g., "Environmental scientists proved that vegan diets reduce meat production by 73%.")
- Sets a stage for dis/agreeing with the position (Janier and Reed, 2017) (e.g., "You say that you want attention, but, at the same time, you don't want me to bring attention to you.")
- Commits straw man fallacies by distorting the original representation or selecting part of the original utterance (Talisse and Aikin, 2006)

While reported speech as a whole is an assertion, its primary contribution to the argumentation usually comes from the speech content, as in:

Original text: "Environmental scientists suggest that vegan diets reduce meat production by 73%."
Proposition(s): "Vegan diets reduce meat production by 73%." (Source: environmental scientists)

where the speech source "environmental scientists" is used to support the speech content.

Questions: Grammatically, questions are not asserted propositions, because they cannot be judged true or false. However, in argumentation, questions play important argumentative roles, e.g., by challenging the listener (Copi et al., 2016) or asking critical questions (Walton et al., 2008). Questions in argumentation may be categorized into rhetorical questions and pure questions. Rhetorical questions are not intended to require an answer; instead, they often make an implicit assertive (as in sentence 2.4). Zhang et al. (2017) identified finer-grained types of rhetorical questions, such as sharing concerns, agreeing, and conceding. Our system is not aiming to classify these types, but instead focuses on extracting implicit assertives in rhetorical questions.

Pure questions, on the other hand, are intended to seek information. According to the speech act theory, non-binary questions have incomplete propositions (Searle, 1969). For instance, the question "How many people were arrested?" has the proposition "X people were arrested", with the questioned part underspecified and denoted by "X". Although the proposition is semantically underspecified, subsequent arguments may build on this, making this proposition an important argumentative component. Hence, our system covers extracting semantically underspecified propositions from pure questions as well. (See Bhattasali et al. (2015) for computational methods to distinguish between rhetorical questions and pure questions.)

Question	Possible interpretations
Why would X do Y? (e.g., "Why would you buy it?")	X would not do Y. Y is not necessary.
How many/much of X are Y? (e.g., "How many of them are military experts?")	No/few/little X are Y.
How ADJECTIVE is X? (e.g., "How big is the debt really?")	X is not ADJECTIVE.
What is X? (e.g., "What's the problem with that?")	There is no X.
Did/do/does X do Y? (e.g., "Did they steal the money?")	X did/do/does not do Y.
What/how can X do Y? (e.g., "How can he solve issue?")	X cannot do Y.
Why not do X? (e.g., "Why not buy it?")	You should/had better do X.

Table 2.1: Possible interpretations of challenge questions.

Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011) categorizes questions into four big types, depending on the function: challenge question, assertive question, pure question, and directive question. Challenge questions and assertive questions, despite their question form, have the main function of asserting propositional content. First, the challenge question is probably the most common type of question in debates. Its function is to make assertions to challenge the argumentation partner. For example:

> **Original text:** "Isn't it a little bit hard to call this just a partisan issue?"
> **Proposition(s):** "It's a little bit hard to call this just a partisan issue"
>
> **Original text:** "What has he not answered?"
> **Proposition(s):** "He has answered questions"
>
> **Original text:** "What the fuck is he supposed to say to that?"
> **Proposition(s):** "There is nothing he is supposed to say to that"
>
> **Original text:** "Would you want the future president to remember you as the guy who cut your mic off while you were talking?"
> **Proposition(s):** "You would not want the future president to remember you as the guy who cut your mic off while you were talking"

Given a challenge question, one way to think of what the question asserts is to consider what would be implied when the partner does not answer the question. Here we see some mechanical transformation happening for challenge questions (Table 2.1).

An assertive question similarly asserts that its propositional content is true. For example:

> **Original text:** "Do you want to know what she'll do? It's all on her website."
> **Proposition(s):** "You want to know what she'll do" + "What she'll do is all on her website"

In contrast, the main function of pure questions is information seeking. IAT posits that even a pure question implies or presupposes some propositional content, and this content is a crucial building block for the subsequent argumentation. In the US2016 corpus, the underspecified semantic information in a pure question is replaced with a placeholder variable "xxx". For example:

Original text: "When was the last time Chafee got to speak?"
Proposition(s): "The last time Chafee got to speak was xxx"

Original text: "Who is Chafee?"
Proposition(s): "Chafee is xxx"

Original text: "Do all lives matter?"
Proposition(s): "All lives do / do not matter"

Even these seemingly incomplete propositions, such as the second example, open a way to further elaboration in the dialogue. Similarly, in the third example, the proposition does not have much meaning on its own, but serves as a building block for further argumentation.

Lastly, directive questions have imperative forces as in:

Original text: "Any specific examples?"
Proposition(s): "Provide any specific examples"

Imperatives: Like questions, imperatives are not propositions grammatically, but they are often important and common argumentative components. Imperatives are common in argumentation as in "Stop raising the sales tax" and "Look how bad the system is". However, to our knowledge, there is little theoretical work on what propositional content is asserted by imperatives in argumentation. There have been theories about the semantics of imperatives in general context; for example, the *you-should* theory suggests that an imperative of the form "Do X" may imply "X should be done" (Hamblin, 1987; Schwager, 2005), as in:

Original text: "Yes, of course, raise the minimum wage."
Proposition(s): "The minimum wage should of course be raised"

While applicable in many general cases, this mechanism is not satisfactory in argumentation. For instance, while this transformation preserves the literal meaning of both the first and second examples above, it does not capture the main proposition asserted in the following example.

Original text: "Look how bad the system is."
Proposition(s): "The system is bad"

This example is unlikely arguing for "looking" per se; it rather asserts that the system is bad, which is the main content that contributes to the argumentation. Some other examples include:

Original text: "Let me address college affordability"
Proposition(s): "CLINTON would like to address college affordability"
Original text: "Look at the mess that we're in."
Proposition(s): "We're in a mess"

No simple transformation rules apply here, and such irregularities call for more case studies. Our work aims to make an initial contribution in that direction.

Removal of non-propositional expressions, discourse markers, and epistemic modalities: Non-propositional expressions, discourse markers, and epistemic modalities are not included in asserted propositions.

Non-propositional expressions include:

- Non-propositional answers ("Yes.", "Very true.", "Funny.")
- Filling words ("Well...", "Look.", "I mean")
- Incomplete sentences ("What do you ...")
- Utterances for moderating a dialogue ("Senator...", "It is time to start the debate.")

Discourse markers include "But", "And", etc.

The general function of an epistemically qualified statement is to assert the statement. Hence, epistemic modalities may be removed[3]. For example:

> **Original text:** "I think Sanders is winning gun owners over."
> **Proposition(s):** "Sanders is winning gun owners over"
>
> **Original text:** "I'm fairly convinced he threw that first debate."
> **Proposition(s):** "He threw that first debate"

Qualification (hedges): Qualifiers (or hedges) are a main component in Toulmin's argumentation structure, related to *Qualifier* and *Rebuttal*. A qualifier can be as specific as a conditional as in "under the condition that ..." or as simple as a modality, such as "probably". We think conditionals should be included in a proposition, because they are essential in determining the meaning of a claim. For example, in the following dialogue, the critic is attacking the initiator's claim using qualification.

> **Initiator:** "A practicing vegan is less likely to harm animals and the environment."
> **Critic:** "When properly conducted, the practice of eating meat is far better for the environment than eating purely vegetables."

Without the conditional, the critic's claim is another overgeneralization and is not what the critic means. More thoughts on when- and if-statements are discussed later.

For simple modalities, such as "probably" and "certainly", there exist different views as to whether they are part of a proposition. Toulmin includes those modalities as part of a claim, whereas Freeman (1991) argues that those modalities may modify the inferential link between premises and conclusion instead of the conclusion itself. For example, given premises, a claim being *certainly* true indicates that the claim is strongly supported by the given premises, rather than the claim is by nature certainly true. Freeman suggests that such cases should be distinguished from the cases where modalities directly modify a claim, as in "2 + 2 is certainly 4". This distinction is theoretically interesting but would not matter much in practice. Leaving simple modalities as part of a proposition is consistent with conditionals, and either choice would have little effect on argumentation structure analysis.

[3]This decision can be controversial, as an epistemic modality can be a target of attack by a critic (e.g., "Do you really believe that?"). However, This may be an edge case and we may follow IAT. On the other hand, treating epistemically modified statements in the same way as reported speech (i.e., duplicate the *that*-clause and extract two propositions) may have the benefit of system-wide consistency.

Time and location: Time and location information is essential part of a proposition, and the meaning of a proposition can be changed significantly without them.

When/if-statements: As discussed above, conditionals and time information that are expressed through "when" or "if" may need to be kept in an asserted proposition in most cases. However, we acknowledge that "when" and "if" can be used not strictly to indicate a condition. Consider the following sentence:

> "When you look at what ISIS is doing with the Internet, they're beating us at our own game."

The when-clause does not qualify the main clause. In the corpus, this sentence is indeed annotated with only one asserted proposition ("ISIS are beating USA at their own game"), by judging that the when-clause adds no information to the argumentation. This judgment can be controversial, however.

Similarly, in the following dialogue:

> **Initiator:** "Having fewer debates is going to be a massive advantage for Hillary."
> **Critic:** "If you're basing your political opinion based on which speaker is smoother, you shouldn't be voting."

the critic's if-clause might be interpreted as an accusation rather than a true conditional. In the corpus, this sentence is annotated with two asserted propositions: "You're basing your political opinion based on which speaker is smoother" and "You shouldn't be voting".

The distinction between true conditional and assertive conditional is highly subject to interpretation and sensitive to the context. It is not clear if the corpus has been annotated with clear instructions for this distinction, and considering how much effort we would need to re-annotate the corpus for this distinction with good inter-rater agreement and how much benefit that would allow us at subsequent stages of argumentation analysis, we do not make the distinction between true conditionals and assertive conditionals.

Segmentation granularity and reconstruction: There is no theoretical consensus on the granularity of argumentative propositions, and depending on the granularity, we can think of a sentence as having varying numbers of propositions. The most common practice in computational argument systems is to treat individual clauses as propositions. However, theoretically we can think of other choices of granularity, such as event levels (e.g. "destruction of the building" has the meaning that "the building was destructed"), potential-negation levels (i.e., separate apart all components that can be negated), and many more. However, there are at least two important practical considerations. First, unless a proposition contains too many predicates, the subtle choice of granularity may not have a huge impact on argumentation structure analysis, because the analysis will likely be performed on predicate-level information anyway. Second, choosing a granularity different than the available corpus means that the corpus has to be re-annotated not only at the proposition level but also at the propositional relation level.

We think that the US2016 corpus has been annotated with a reasonable granularity, which seems to follow the work by Stede et al. (2016). A sentence is basically segmented into clauses and further segmented into (1) rhetorical participle phrases as in:

> **Original text:** "I may need to call out of work tomorrow due to drinking words"
> **Proposition(s):** "I may need to call out of work tomorrow" + "This debate has so many drinking words"

(2) conjoined verb phrases as in:

> **Original text:** "Obama lost because he seemed disinterested and relied to heavily on trying to actually explain the issue."
> **Proposition(s):** "Obama lost" + "Obama seemed disinterested" + "Obama relied to heavily on trying to actually explain the issue"

(3) objects with different polarity as in:

> **Original text:** "I want to see a factual Hillary going in, not this one that is trying to be the High School president ..."
> **Proposition(s):** "I want to see a factual Hillary Clinton going" + "I do not want to see Hillary Clinton that is trying to be the High School president ..."

(4) interpolated texts as in:

> **Original text:** "The liquid, because it is so dangerous, is not allowed in the building."
> **Proposition(s):** "The liquid is not allowed in the building" + "The liquid is so dangerous"

(5) nonrestrictive relative clauses as in:

> **Original text:** "She's staying above the noise, which is all she has to do with a 30% lead in the polls."
> **Proposition(s):** "Hillary Clinton is staying above the noise" + "Staying above the noise is all Hillary Clinton has to do with a 30% lead in the polls"

2.5 Cascade Model

Based on the theoretical concerns and practical examples described in the previous section, in this section, we present a cascade model that takes an utterance as input and extracts asserted propositions as output. The model consists of seven modules as shown in Figure 2.2. The functions of individual modules can be summarized as follows:

1. **Anaphora resolution:** Replace pronoun anaphors with their referents.
2. **Locution extraction:** Extract locutions (ADUs) from the utterance.
3. **Reported speech:** Determine if the locution is reported speech; if so, identify the the speech content and speech source.

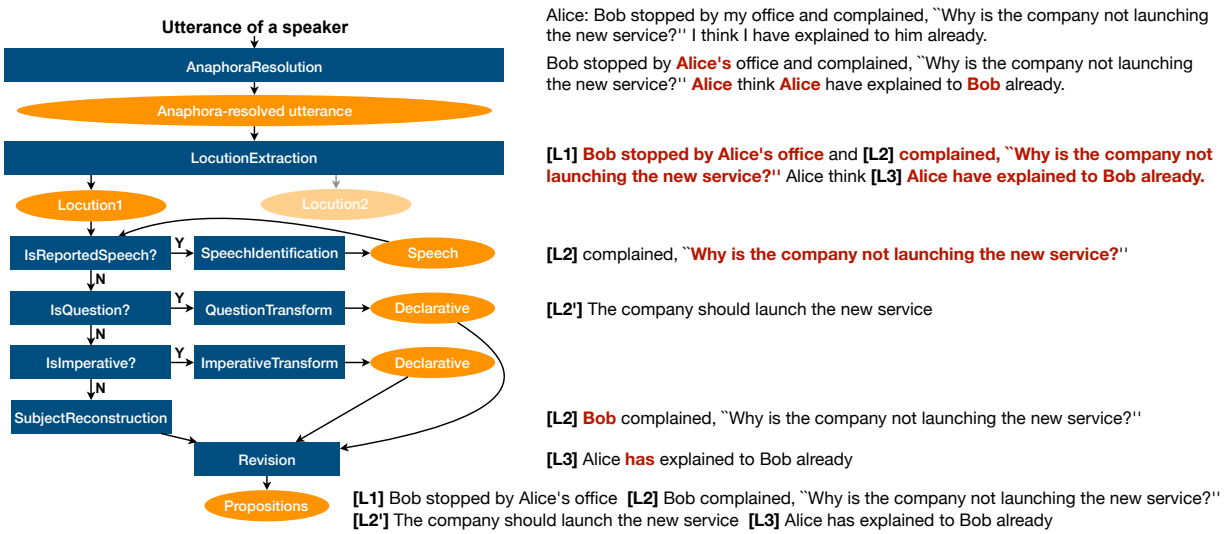


Figure 2.2: Cascade model of proposition extraction. The input is each utterance, blue boxes are individual (sub)modules and orange circles are the outputs of the modules. We made up the utterance used in the figure in order to cover the functions of most modules.

4. **Question:** Determine if the locution or speech content is a question; if so, extract its propositional content.
5. **Imperative:** Determine if the locution or speech content is an imperative; if so, extract its propositional content.
6. **Subject reconstruction:** Reconstruct the missing subject, if any, of the locution or speech content.
7. **Revision:** Make additional adjustments necessary for final propositions.

In the remainder of this section, we describe how to formulate the task of each module as a computational problem, and present various approaches with their performance. Each module is evaluated separately, instead of using the result of the previous module; this setting prevents error propagation and helps evaluate the performance of each module more accurately. Some methods we use are based on machine learning and thus requires a split of training and test sets. Hence, we randomly split the entire corpus into five folds and conduct cross validation with the same folds throughout the section.

2.5.1 Anaphora Resolution

Anaphora resolution is based on Stanford CoreNLP 3.8.0. Yet, blindly applying it induces several challenges as shown in the previous section, such as incorrect resolution of speakers and hearers (as this information is often missing in the text), resolution of non-pronouns, and errors inherent in the tool. To rectify these challenges, we decompose the task into the following subtasks.

- **1st-person singular:** Replace “I”, “my”, “me”, “mine” with the speaker’s name.
- **2nd-person singular:** Replace “you”, “your”, “yours” with the previous turn’s speaker name.

	BLEU	Dep	Dep-SO	Noun
Locution (no resolution)	69.3	65.1	55.8	71.4
CoreNLP	62.8	61.7	53.8	70.4
1S	**70.1**	**65.7**	58.9	74.8
1S+2S	69.7	65.5	58.3	74.6
1S+3SG	69.3	65.4	**60.1**	**75.7**
1S+3SG+3SN	68.5	64.9	59.2	75.6

Table 2.2: Performance of anaphora resolution. (**1S:** 1st-person singular, **2S:** 2nd-person singular, **3SG:** 3rd-person singular gender, **3SN:** 3rd-person singular gender-neutral, **Dep:** Dependency, **Dep-SO:** Dependency for subjects and objects.)

- **3rd-person singular gender:** Resolve "he", "his", "him", "she", "her", "hers" using CoreNLP.
- **3rd-person singular gender-neutral:** Resolve "it", "that" using CoreNLP.
- **3rd-person plural:** Resolve "they", "their", "them", "theirs" using CoreNLP.

Inaccurate anaphora resolution can rather distort the original meaning of text. Hence, the goal here is to find the best combination of the subtasks. The first two subtasks are applied only to TV debates, as Reddit user names have not been resolved in the corpus. All possessive pronouns are replaced with references suffixed with "'s" (e.g., "his" $\rightarrow$ "Trump's").

For evaluation, we assume that effective anaphora resolution would make a locution more "similar" to the annotated proposition. Hence, we compare the similarities between a locution and the annotated proposition before and after anaphora resolution, using the following metrics:

- **BLEU:** Generic string similarity based on n-grams ($n = 1, 2, 3, 4$).
- **F1-score of dependency tuples:** String similarity based on dependencies. Less sensitive than BLEU to the exact locations of words.
- **F1-score of nsubj/dobj dependency tuples:** Rough semantic information pieces representing who did what to whom/what.
- **F1-score of nouns:** How accurately anaphora resolution retrieves nouns (as our anaphora resolution replaces only nouns).

Results

As shown in Table 2.2, blindly applying CoreNLP (row 2) significantly hurts all similarity measures (compared to row 1). In contrast, speaker resolution (row 3) plays a key role in improving all measures over original locutions, especially semantic information (subject/object) and nouns. Additional resolution of hearers (row 4) does not help, as "you" is used in a more general way than referring specifically to the hearer.

Resolving 3rd-person gender pronouns (row 5) further improves performance for semantic information and noun retrieval over speaker resolution, at the expense of slightly lower BLEU and dependency similarites. Additional resolution of "it", "its", and "that" turns out to rather hurt

performance.

For argument mining, it may be desired to resolve as many anaphors as possible unless the original meaning is significantly hurt, because pronouns provide little information for identifying propositional relations. Hence, we conclude that resolution of speakers and 3rd-person gender pronouns is ideal for this module, and the subsequent modules use the result of this configuration. However, we find that resolution of 3rd-person gender-neutral pronouns is critical, as will be discussed in Section 2.5.7, and eventually they should be resolved depending on the availability of proper anaphora resolution tools.

2.5.2 Locution Extraction

For each utterance with anaphors resolved, the LocutionExtraction module identifies locutions (ADUs), from which asserted proposition(s) will be extracted. This task is almost identical to conventional ADU segmentation, and many methods have already been proposed (Section 2.2.1). Beating prior models for this task is beyond the scope of this thesis; rather, we focus on understanding what causes confusion for locution boundaries. Following the convention for this task (Eger et al., 2017; Ajjour et al., 2017), the task is formulated as tagging each word with B/I/O (beginning/inside/outside of a locution).

Models

We explore the state-of-the-art BiLSTM model (Ajjour) (Ajjour et al., 2017), as well as a regular CRF (R-CRF) and BiLSTM-CRF (Huang et al., 2015). A CRF showed strong performance for cross-domain segmentation, and BiLSTM-CRF is an extension of CRFs, where emission scores are calculated through BiLSTM. For all models, we use the following features, adopted from or informed by the prior work (Ajjour et al., 2017):

- **word:** Current word (i.e., word index for R-CRF and pre-trained GloVe.840B.300d word embeddings for BiLSTM-CRF and Ajjour).
- **pos:** Part-of-speech tag of the current word.
- **ne:** Named entity type of the current word.
- **prev_1gram:** Previous word of the current word, as conjunctions and discourse markers are good indicators of locution boundaries. (R-CRF only, as BiLSTM considers context.)
- **bos/eos:** Indicator of whether the current word marks the beginning/end of a sentence, as locution boundaries are often restricted by sentence boundaries.
- **boc/eoc:** Indicator of whether the current word marks the beginning/end of a clause, as locution boundaries are closely related to clause boundaries. We obtain clauses from the constituency parse of the sentence, taking phrases tagged with S. For nested clauses, we take the deepest clauses to avoid overlap.

We use the following model settings. For R-CRF, we use sklearn-crfsuite 0.3.6. We conducted grid search, exploring all combinations of the bias feature ($\{1, 0\}$) and the following optimization parameters:

- Gradient descent using the L-BFGS method

Model	F1
R-CRF	78.8
BiLSTM-CRF	78.9
Ajjour	**79.4**

Table 2.3: F1-score of locution extraction.

  - L1 regularization: 0, 0.05, 0.1
  - L2 regularization: 0, 0.05, 0.1
- Passive Aggressive (PA)
  - Aggressiveness parameter: 0.5, 1, 2

For BiLSTM-CRF, we used the following parameter values:

- BiLSTM hidden dim: 128
- Optimizer: Adam
- Learning rate: 0.001

For Ajjour, we used the following parameter values:

- Encoder BiLSTMs hidden dim: 128
- Output BiLSTM hidden dim: 5, 10, 20
- Optimizer: Adam
- Learning rate: 0.001

We evaluate the models using the macro F1-score across the BIO tags with 5-fold CV.

Results

Ajjour et al. (2017)'s model outperforms the CRF-based models (Table 2.3). The model tends to underproduce locutions (7,767 compared to 8,008 annotated), i.e., produce coarser and longer locutions than ground truth locutions, missing signals for splitting them further into smaller locutions. To examine those signals, we gathered extracted locutions that overlap with two consecutive annotated locutions, and counted the words between the two locutions.

As shown in Table 2.4, frequently, the model failed to make a split at a comma (31%) or between locutions that are back-to-back without any separator in between (10%). In the majority of these cases, the locutions are two independent clauses, indicating that the model needs a more robust mechanism to make use of clause boundaries. Although not very common, a locution also serves as a subordinate clause, adverb phrase, particle phrase, yes/no answer, or relative clause (Table 2.5). Deciding whether to separate a subordinate clause from the main clause is not trivial. For instance, if- and when-clauses, the most common subordinate clauses in the analysis, are separated off or attached to the main clause depending on the strength of their dependency, which is often vague. If we are to build a system to make this decision automatically, we may consider the truth value of the subordinate clause and whether it is idiomatic.

Other frequent separators include conjunctions "and" (21%) and "but" (6%). As in the case

Top 1-8	Top 9-16	Top 17-24
, (31%)	– (2%)	or (1%)
and (12%)	, because (1%)	? (1%)
NONE (10%)	-lrb- (1%)	. and (1%)
, and (9%)	, which (1%)	to (1%)
, but (4%)	; (1%)	as (1%)
. (3%)	... (1%)	, so (1%)
because (2%)	- (1%)	that (1%)
but (2%)	when (1%)	if (0%)

Table 2.4: Words that separate two annotated locutions that overlap with one predicted locution. NONE indicates that the locutions are back-to-back without any separator.

	1st locution1	2nd locution
Subordinate clauses	7%	6%
Adverb phrases	4%	8%
Particle phrases	1%	4%
Yes/no	2%	-
Relative clauses	-	5%

Table 2.5: Breakdown of locution types that are separated by a comma or that are back-to-back (total 293 pairs).

above, the model sometimes has difficulty deciding whether to split conjoined phrases and clauses. According to the data, phrases conjoined by these words are sometimes separated and sometimes not. Again, the decision becomes harder when clauses are conjoined. The module sometimes makes split errors at punctuation marks, although such errors are not frequent. Punctuation marks, such as ".", "–", "?", and "!", often conclude a sentence or clause, but sometimes a locution is annotated across these punctuation marks. We looked at the cases where annotated locutions are separated by these marks but not by the module. They seem to be a simple mistake made by the module, and we did not find any specific patterns except that locutions tend to be short.

Lastly, we examined what prevents the module from making precise locution boundaries and if there are any patterns. Specifically, for each utterance, we count words that are included only in annotated locutions or in extracted locutions, but not both. These counts may reveal words or phrases that the module tends to include or ignore incorrectly. As shown in Table 2.6a and Table 2.6b, the module does not show any strong tendency to include or ignore certain words. However, annotated locutions tend to include more periods and questions marks than extracted locutions. On the other hand, annotated locutions tend to include more commas and conjunction "and", as we have already discussed before.

Top 1-10	Top 11-20
, (23%)	! (1%)
. (15%)	: (1%)
? (10%)	well (1%)
and (7%)	secretary clinton (1%)
but (4%)	so (1%)
– (2%)	mr. trump (1%)
” (2%)	i think (1%)
... (2%)	yes (1%)
because (2%)	yeah (1%)
“ (1%)	lol (0%)

(a) Frequency of words contained only in annotated locutions ($N = 1736$).

Top 1-10	Top 11-20
, (31%)	: (1%)
and (7%)	“ (1%)
– (4%)	i think (1%)
. (3%)	no (0%)
but (2%)	! (0%)
” (2%)	then (0%)
? (1%)	you know (0%)
because (1%)	though (0%)
so (1%)	-rrb- (0%)
... (1%)	that (0%)

(b) Frequency of words contained only in predicted locutions ($N = 3397$).

Table 2.6: Frequency of words misaligned between annotated locutions and predicted locutions.

2.5.3 Reported Speech

A locution extracted above is examined by the IsReportedSpeech submodule to decide if it is reported speech. If so, we extract two main pieces of information: the source and the content of speech, by the SourceExtraction and ContentExtraction submodules, respectively.

Classifying whether a locution is reported speech or not is a typical classification problem. We trained a BERT model for sequence classification (Devlin et al., 2018) on the annotations, using the implementation from HuggingFace (Wolf et al., 2020) with the pretrained, uncased base model. The trained model achieved an AUC of 97.0 and an F1 of 85.1 for 5-fold cross validation. We did not conduct further experiments with other classifiers, because the BERT accuracy is reasonably high.

Due to the important roles of speech content and source, computational models have been proposed to identify them, based on rules (Krestel et al., 2008), conditional random fields (Pareti et al., 2013), and a semi-Markov model (Scheible et al., 2016). Our work is different from these studies in two ways. First, they are based on news articles, whereas our work is on argumentative dialogue. Second, they use rules or features that reflect typical words and structures used in reported speech, whereas our work explores a neural method that does not require feature engineering. We aim to show how well a state-of-the-art neural technique performs on extraction of speech content and source. A slightly different but related strain of work is to identify authority claims in Wikipedia discussions (Bender et al., 2011), but this work does not identify speech content and source.

The tasks of identifying the source and content of speech are both formulated as BIO sequence tagging (we conduct separate experiments for sources and content). For the text span of a source or content, the first word is tagged with B and the other words with I; all other words are tagged with O.

Models

We explore three models: a conditional random field (CRF) with hand-crafted features, the BERT token classifier with a pretrained language model, and a semi-Markov model as the baseline. For all models, the input is a sequence of words and the output is a BIO tag for each word. We conduct separate experiments for content and source, because we do not assume that they are mutually exclusive (although they are in most cases).

Conditional Random Field (CRF): Our CRF uses the following features:

- Current word.
- Named entity type of the word.
- POS tag of the word.
- Unigram and bigram preceding the word.
- Unigram and bigram following the word.
- Indicator of if the word is a subject ("nsubj*" on the dependency parse tree).
- Indicator of if the current word is the beginning/end of a clause ("S" on the parse tree).

The features were extracted using Stanford CoreNLP 0.9.2 (Manning et al., 2014).

For model parameters, we explore two optimization functions: (i) L-BFGS with the combinations of L1/L2 regularization coefficients $\{0, .05, .1, .2\}$; (ii) Passive Aggressive with aggressiveness parameter values $\{.5, 1, 2, 4\}$. The model was implemented using sklearn_crfsuite 0.3.6.

BERT: The second model is the BERT token classifier (Devlin et al., 2018), which classifies the tag of each word. BERT has shown significant performance boosts in many NLP tasks and does not require hand-crafted features. We use the pretrained, uncased base model with the implementation provided by Hugging Face (Wolf et al., 2020). The model is fine-tuned during training.

Baseline: The baseline is the state-of-the-art semi-Markov model for speech content identification (Scheible et al., 2016). This model first identifies cue words (e.g., reporting verbs) and iteratively identifies the boundaries of speech content using a set of hand-crafted features. This model does not identify speech sources and thus is compared with other models only for content identification.

For a methodological note, the original source code was hard-coded to work for the PARC3.0 dataset, and we could not replicate the model to train on other data. Therefore, all accuracies of this model in the next section result from training it on the training set of the PARC3.0 dataset. We will show its performance on both PARC3.0 and US2016.

Data

PARC3.0: The first dataset is 18,201 instances of reported speech in news data (Pareti, 2016). The original dataset was built upon the Wall Street Journal articles in the Penn Discourse TreeBank

(PDTB) (Prasad et al., 2008), where each instance of reported speech has been annotated with the content, source, and cue word (e.g., reporting verbs). The reliability of the annotations were measured by the overlap of annotated text spans between annotators. The overlap for speech content is 94% and that for speech source is 91%, suggesting the high reliability of the annotations.

This dataset consists of 24 sections corresponding to the PDTB sections. The original paper suggests using sections 00-22 for training (16,370 instances), section 23 for testing (667 instances), and section 24 for validation (1,164 instances).

US2016: The second dataset is the instances of reported speech in the US2016 corpus. Reported speech is not properly annotated in the original US2016 corpus. Hence, we conducted an additional layer of annotation on top of the original corpus (the details are available in Chapter 5). Briefly, the annotations include 242 instances of reported speech annotated with speech content and source. The reliability of the annotations was measured by the number non-overlapping words between annotators. The average number of words that are outside of the overlapping text span was 0.2 for speech content and 0.5 for speech sources, suggesting the high reliability of the annotations.

Experiment Settings

The CRF and BERT models are trained and tested on both PARC3.0 and US2016, separately. For PARC3.0, we use the split of train, validation, and test as suggested by the original paper. For US2016, we use 5-fold cross validation; for each iteration, three folds are used for training, one for testing, and the other for choosing the optimal hyperparameters (CRF) or the optimal number of epochs (BERT).

The baseline model is trained and tested on PARC3.0 using the same training, validation, and test split. US2016 is used only for testing after it is trained on the training set of PARC3.0 (as mentioned in 2.5.3).

We use various evaluation metrics. For speech content, the **F1-score** is calculated based on the true and predicted BIO tags of individual words, as well as the **BLEU** score of the predicted text span against the true text span. For speech sources, the F1-score is calculated based on the match between the true source's text and the predicted text. Two texts are considered matched if they are identical (**Strict**) or if their words overlap (**Relaxed**). We do not measure the F1-score based on BIO tags for speech sources, because the source may be mentioned multiple times in reported speech and we do not want to penalize the model when the mention identified by the model is the true source but different from the annotated mention.

Results

Content Identification: The accuracies of all models are summarized in Table 2.7a. The baseline model (Scheible) has two rows: row 1 is its accuracy on all test instances, and row 2 is on test instances where the model was able to identify cue words. We find that the BERT model (row 4) outperforms the feature-based CRF and the baseline model for both corpora, achieving a macro F1-score of 82.6% at tag levels and a BLEU score of 82.0% for PARC3.0 and an F1-score of 87.1% and a BLEU score of 89.3% for US2016. These scores show the high reliability of the

	PARC3.0		US2016	
	F1	BLEU	F1	BLEU
Scheible (All)	64.4	57.1	*37.9*	*23.4*
Scheible (Matched)	75.8	72.7	*79.3*	*76.5*
CRF	71.3	66.3	72.5	68.7
BERT	82.6	82.0	87.1	89.3

(a) Accuracy of identifying speech content. The accuracies of Scheible for US2016 (italic) result from training it on the training data of PARC3.0.

	PARC3.0		US2016	
	Strict F1	Relaxed F1	Strict F1	Relaxed F1
CRF	52.4	59.8	62.4	71.6
BERT	71.0	78.6	70.3	84.8

(b) Accuracy of identifying speech source.

Table 2.7: Accuracy of identifying speech content and source.

BERT model for extracting main propositions asserted in reported speech. In addition, the high accuracy on US2016 despite its small size suggests that the pretrained language model effectively encodes important semantic information, such as reporting verbs and dependencies among subject, verb, and object.

The baseline model, which was trained on PARC3.0, performs poorly on US2016 (row 1). The main obstacle is that it fails to detect cue words (e.g., reporting verbs) in 168 out of 242 instances (69%). This shows one weakness of the baseline model: since this model works at two steps—detect cue words and find content boundaries—identifying speech content is strongly subject to cue word detection. When the baseline is evaluated only on the instances where a cue word was detected, its accuracy boosts significantly (row 2), outperforming the CRF but still worse than BERT.

A qualitative analysis of the BERT model reveals that most instances are tagged accurately, and errors are concentrated on a few instances. One of the main issues is whether a reporting verb should be included or not as speech content. In the annotation process for US2016, a reporting verb was included as speech content only if the verb has meaning other than merely "to report" (e.g., "**blamed** his idea", "**declared** their candidacy"). As a result, the model often has difficulty judging a reporting verb to be part of the speech content or not.

In some cases, the exact boundary of speech content is ambiguous. For instance, in the sentence

> "Bush has promised **four percent economic growth and 19 million new jobs** if Bush is fortunate enough to serve two terms as president."

the annotated speech content is in bold, while the model included the if-clause as the content

(underlined). However, it may seem more appropriate to include the if-clause as part of the promise.

Source Identification: The accuracies of all models are summarized in Table 2.7b. The BERT model (row 2) again significantly outperforms the CRF (row 1), achieving F1-scores of 75.7% for strict evaluation (exact match) and 85.1% for relaxed evaluation (overlap allowed). It is usually when a source is a long noun phrase that a predicted source and the true source overlap without exact match (e.g., "President Obama" vs. "Obama").

Our qualitative analysis of the BERT model reveals two common error cases. First, the model tends to capture subjects and person names as a speech source, which is not correct in some cases:

> "We have been told through investigative reporting that he owes about $650 million to Wall Street and foreign banks"

where the model identifies "we" as the speech source, while the true source is the "investigative reporting". The model also sometimes fails to detect any source candidate if reported speech has an uncommon structure, such as "The record shows that ..." and "No one is arguing ... except for racists", where the speech sources are underlined. These problems may be rectified with larger training data that include more diverse forms of reported speech.

2.5.4 Question

A locution or the speech content of reported speech is examined by the IsQuestion submodule to decide if it is a question. If so, it is transformed to its asserted proposition by the QuestionTransformation submodule. Let's begin with question detection.

Models

We explore three approaches: parse, regex, and neural classifier.

Parse: For the parse approach, we rely on the constituency parse tree result of CoreNLP. Specifically, a locution is classified as a question if any part of the locution is tagged with SBARQ (direct question introduced by wh-element) or SQ (yes/no questions and subconstituent of SBARQ excluding wh-element).

Regex: For the regex approach, we compile regex patterns, capturing if the locution as a question mark or if the text begins with words that often initiate a question (e.g., "how", "do") (Table 2.9).

BiLSTM For the BiLSTM classifier, words are encoded through a BiLSTM layer, weighted and combined using an attention mechanism, and fed to a single-layer neural network. The final output is the probability of the input being a question. The locution is classified as a question if the probability is greater than or equal to 0.5.

	Prec	Recl	F1
Parse	78.4	54.5	64.3
Question Mark	75.1	93.8	83.4
Regex-All	58.8	**97.2**	73.3
BiLSTM	78.1	92.2	84.6
BERT	**81.2**	92.1	**86.2**

Table 2.8: Accuracy of question detection. The accuracy of the first three approaches is calculated on the entire data at once, whereas that of the neural classifiers is the average across the folds.

BERT: Lastly, we try the BERT sequence classifier, using the pretrained, uncased base model provided by Hugging Face.

Data

From the US2016 corpus, we filtered 565 pairs of a locution and its asserted proposition that are annotated with the following question types:

- **Pure:** e.g., "Who is Chafee?" → "Chafee is xxx"; "Do lives matter?" → "Lives do / do not matter" (Semantically underspecified parts are denoted by "xxx" and the slash "/".)
- **Assertive:** e.g., "What does that say about your ability to handle challenging crises as president?" → "Clinton does not have the ability to handle challenging crises as president"
- **Challenge:** e.g., "What has he not answered?" → "He has answered questions"
- **Directive:** e.g., "Any specific examples?" → "Provide any specific examples"

Note that only pure questions are semantically underspecified (indicated by "xxx" and "/"); the other types contain concrete propositions to be asserted. Our models are trained on all question types.

Results

Table 2.8 shows the accuracy of the models. Overall, the BERT model achieves the highest F1-score. Unsurprisingly, the parse approach has fairly high precision but suffers from low recall. The parse approach fails to capture ill-formed questions that are in declarative form grammatically, short words, and incomplete sentence, as in the following examples:

- "semen is a human as well?"
- "one douchebag?"
- "attack obama then?"

Often the NLP tool misses even well-formed questions due to its inherent inaccuracy.

Regex achieves a higher F1-score, especially due to high recall (Table 2.9). Interestingly, a question mark by itself is strongly indicative of a question and has a high coverage (first row, first column in the table). It can successfully capture the example questions above, but fails to

Regex	Prec	Recl	F1	Regex	Prec	Recl	F1
\?	75.1	93.8	83.4	^should	80.0	01.4	02.8
^do	48.5	08.7	14.7	^would	53.8	01.2	02.4
^how	75.9	07.8	14.1	^will	100.0	01.1	02.1
^what	46.2	06.4	11.2	^was	66.7	01.1	02.1
^is	77.5	05.5	10.2	^where	71.4	00.9	01.7
^why	42.3	03.9	07.1	^when	07.1	00.9	01.6
^did	84.2	02.8	05.5	^which	28.6	00.7	01.4
^are	80.0	02.1	04.1	^have	50.0	00.5	01.1
^who	70.6	02.1	04.1	^were	100.0	00.4	00.7
^can	61.1	01.9	03.8	^could	18.2	00.4	00.7
^does	58.8	01.8	03.4	^has	33.3	00.2	00.4
All	58.8	97.2	73.3				

Table 2.9: Accuracy of question detection for regex.

capture questions that do not have a question mark, as in “how can you buy a man worth 10b+...”. Although it has relatively high precision, some false-positive cases include:

- A text uses a question merely for emphasis. (e.g., “it also could be somebody sitting on their bed that weighs 400 pounds , ok ?”)
- A text reports, not asks, a question. (e.g., “so you say to yourself , why did n’t they make the right deal ?”)
- A text uses a question form to expresses confusion. (e.g., “bernie? ... come again?”)

Including question-initiating words into regex patterns increases recall but significantly hurts precision. Some of these words are used for other purposes than a question; for example, “when” may initiate a subordinate clause, and “which” is used as a relative pronoun. The low precision of some words is due to incomplete sentences with subject “I” missing, as in “Could barely understand”.

As can be seen in the error cases above, detecting a question sometimes requires considering a combination of several factors. And the neural classifiers seem more effective in doing that than simpler models. According to a qualitative analysis, the neural classifiers rely heavily on a question mark, probably due to the fact that most questions (94%) have a question mark. However, the neural classifiers are better than simple regex in detecting non-questions that have a question mark.

In the remainder of this section, we move on to extracting implicitly asserted propositions from questions in argumentation. The task is formulated as transforming a question into its asserted proposition.

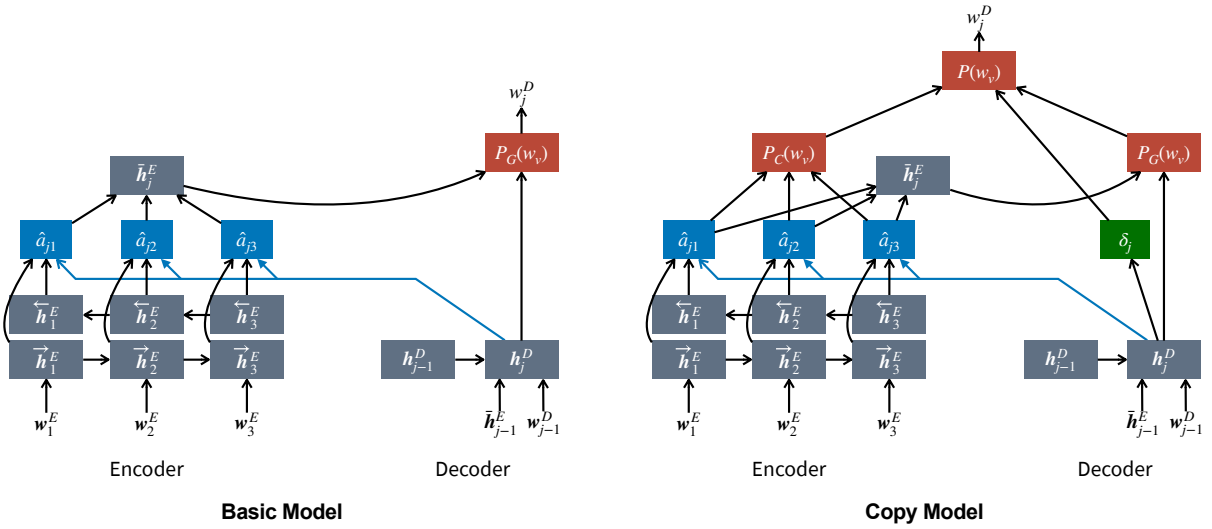


Figure 2.3: Basic model and copy model for question transformation. The snapshots for the jth output word.

Neural Models

We test two RNN-based seq2seq models. First, the **basic** model encodes a question using BiLSTM and decodes a proposition using LSTM and the standard attention mechanism (Luong et al., 2015). Figure 2.3 illustrates the snapshot of the model for the jth output word.

Formally, the input is a sequence of words $w_1^E, \cdots, w_N^E$, and the embedding of w_i^E is denoted by w_i^E. BiLSTM encodes each word w_i^E and outputs forward/backward hidden states $\overrightarrow{h}_i^E$ and $\overleftarrow{h}_i^E$:

$$\overrightarrow{h}_i^E, \overleftarrow{h}_i^E = \text{BiLSTM}(w_i^E, \overrightarrow{h}_{i-1}^E, \overleftarrow{h}_{i+1}^E),$$
$$\overrightarrow{h}_0^E = \overleftarrow{h}_{N+1}^E = 0.$$

For the jth word to be generated, an LSTM decoder encodes the concatenation of the previously generated word w_{j-1}^D and context vector $\bar{h}_{j-1}^E$ (explained below), and the previous hidden state:

$$h_j^D = \text{LSTM}([w_{j-1}^D; \bar{h}_{j-1}^E], h_{j-1}^D),$$
$$h_0^D = [\overleftarrow{h}_1^E; \overrightarrow{h}_N^E].$$

Next, the decoder attends to the encoder's hidden states using an attention mechanism. The attention weight of the ith hidden state is the dot product of the hidden states from the encoder and the decoder:

$$a_{ji} = h_j^D \cdot [\overleftarrow{h}_i^E; \overrightarrow{h}_i^E], \quad \hat{a}_{ji} = \frac{\exp(a_{ji})}{\sum_{i'} \exp(a_{ji'})},$$
$$\bar{h}_j^E = \sum_i \hat{a}_{ji} [\overrightarrow{h}_i^E; \overleftarrow{h}_i^E].$$

The probability of the vth word in the vocabulary being generated is calculated as in the standard attention decoder mechanism:

$$P_G(w_v) = \text{softmax}(W_G[h_j^D; \bar{h}_j^E] + b_G)_v,$$

where W_G and b_G are trainable weight matrix and bias vector.

The basic seq2seq model requires a lot of training data, whereas according to our observation, question transformation is often formulaic, consisting largely of word reordering. Hence, our **copy** model uses a copying mechanism to learn to re-use input words. A prior model (Gu et al., 2016) does not perform well in our task, so we modified it as follows (Figure 2.3).

Our copy model is based on the basic model and has the same process for the generating part. When an output word is copied from the input text, instead of being generated, the probability of the ith input word being copied is proportional to the attention weight of the ith hidden state. That is, the probability of the vth word in the vocabulary being copied is:

$$P_C(w_v) = \sum_{i=1}^{N} \hat{a}_{ji} I(w_i^E = w_v).$$

The final probability of w_v being output is a weighted sum of $P_C(w_v)$ and $P_G(w_v)$, where the weight δ is calculated as

$$\delta_j = \sigma(W_\delta h_j^D + b_\delta),$$
$$P(w_v) = \delta P_C(w_v) + (1-\delta) P_G(w_v),$$

where W_δ and b_δ are trainable weight matrix and bias vector. The main difference of our model from existing ones is that we compute the mixture weight δ_j for P_C and P_G using a separate neural network. In contrast, existing models do not explicitly compute this weight (Gu et al., 2016) or do not use attentional hidden states (Allamanis et al., 2016).

We try the following hyperparameter values:

- Encoder/decoder hidden dim: 96, 128, 160, 192 (basic model) / 128, 192 (copy model)
- Beam size: 4
- Optimizer: Adam
- Learning rate: 0.001
- Gradient clipping: 1
- Word embedding: GloVe 840B

Rule-Based Model

As question transformation is often formulaic, a rule-based method may be effective for small data. For each question, the most relevant parts for transformation are the first word (wh-adverb or auxiliary verb), subject, auxiliary verb, negation, and main verb (i.e., "be"+adjective, "be"+gerund, or else). For instance, the question "Why would you not pay the tax?" might be rearranged to "You would pay the tax", where "why" and "not" are removed. We compile rules that match combinations of these components, starting with a rule that has a high coverage and breaking it down to more specific ones if the rule makes many errors. An example rule is "Why [MODAL] [SUBJECT] "not"" → "[SUBJECT] [MODAL]", which applies to the above example. As a result, we compiled total 94 rules for 21 first words (4.5 rules per first word on average) based on the US2016 dataset (see Table 2.10 for a summary of these rules).

Data

US2016: Our main data is the US2016 data described above for question detection.

From	To
why [MD]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [MD]$_1$ not [*]$_3$.
why [MD]$_1$ not [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [MD]$_1$ [*]$_3$.
why do [SBJ]$_1$ [*]$_2$?	[SBJ]$_1$ [*]$_2$.
why [does|did]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [does|did]$_1$ [*]$_3$.
why is [SBJ]$_1$ [*]$_2$?	[SBJ]$_1$ is [*]$_2$ because xxx.
why [are|were|was]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [are|were|was]$_1$ [*]$_3$.
why [is|are|am]$_1$ not [SBJ]$_2$ [ADJ]$_3$?	[SBJ]$_2$ [is|are|am]$_1$ [ADJ]$_3$.
why [is|are|am]$_1$ not [SBJ]$_2$ [VP]$_3$?	[SBJ]$_2$ should be [VP]$_3$.
why not [VP]$_1$?	should [VP]$_1$.
where [do|did|does|MD]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [do|did|does|MD]$_1$ [*]$_3$ at xxx.
when [did|has]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [did|has]$_1$ not [*]$_3$.
how can [SBJ]$_1$ [*]$_2$?	[SBJ]$_1$ cannot [*]$_2$.
how [MD\can]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [MD\can]$_1$ [*]$_3$ by xxx.
how [do|does]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [*]$_3$ by xxx.
how [MD|do|does|did]$_1$ [SBJ]$_2$ not [*]$_3$?	[SBJ]$_2$ should [*]$_3$.
how are [SBJ]$_1$ going to [*]$_2$?	[SBJ]$_1$ need to [*]$_2$.
how are [SBJ]$_1$ supposed to [*]$_2$?	[SBJ]$_1$ cannot [*]$_2$.
how [am|are|is]$_1$ [SBJ]$_2$ not [*]$_3$?	[SBJ]$_2$ should be [*]$_3$.
how much [*]$_1$?	xxx [*]$_1$.
how [ADJ|ADV]$_1$ [VB|MD]$_2$ [SBJ]$_3$ [VP]$_4$?	[SBJ]$_3$ [VB|MD]$_2$ [VP]$_4$.
what [MD|did]$_1$ [SBJ]$_2$ [VB]$_3$ [*]$_4$?	[SBJ]$_2$ [MD|did]$_1$ [VB]$_3$ xxx [*]$_4$.
what [does|do]$_1$ [SBJ]$_2$ [VB]$_3$ [*]$_4$?	[SBJ]$_2$ [VB]$_3$ xxx [*]$_4$.
what am [SBJ]$_1$ [VB]$_2$ [*]$_3$?	[SBJ]$_1$ am [VB]$_2$ xxx [*]$_3$.
what [is|was|are]$_1$ [SBJ]$_2$?	[SBJ]$_2$ [is|was|are]$_1$ xxx.
what [VB|did|does|do|am|was|is|are]$_1$ [*]$_2$?	xxx [VB|did|does|do|am|was|is|are]$_1$ [*]$_2$.
which [*\VB]$_1$ [*]$_2$?	[*\VB]$_1$ xxx.
which [*\VB]$_1$ [VB]$_2$ [SBJ]$_3$ [*]$_4$?	[SBJ]$_3$ [VB]$_2$ [*]$_4$ [*\VB]$_1$ xxx.
who [VB]$_1$ [SBJ]$_2$ [VP]$_3$?	[SBJ]$_2$ [VB]$_1$ [VP]$_3$ xxx.
who is [SBJ]$_1$?	[SBJ]$_1$ is xxx.
who is [VP]$_1$?	xxx is [VP]$_1$.
who [*\is]$_1$ [*]$_2$?	xxx [*\is]$_1$ [*]$_2$.
have you not [*]$_1$?	you have not [*]$_1$.
[have|has]$_1$ [SBJ\you]$_2$ [*]$_3$?	[SBJ\you]$_2$ [have|has]$_1$ [*]$_3$.
is [SBJ]$_1$ [NP]$_2$?	[SBJ]$_1$ is [NP]$_2$.
is [SBJ]$_1$ [*\NP]$_2$?	[SBJ]$_1$ is / is not [*\NP]$_2$.
are [SBJ]$_1$ [*]$_2$?	[SBJ]$_1$ are not [*]$_2$.
[was|were]$_1$ [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [was|were]$_1$ [*]$_3$.
[is|are|was|were]$_1$ not [SBJ]$_2$ [*]$_3$?	[SBJ]$_2$ [is|are|was|were]$_1$ [*]$_3$.
can [SBJ]$_1$ [VP]$_2$?	[SBJ]$_1$ can [VP]$_2$.
[MD\can]$_1$ [SBJ]$_2$ [VP]$_3$?	[SBJ]$_2$ [MD\can]$_1$ / [MD\can]$_1$ not [VP]$_3$.
[MD]$_1$ not [SBJ]$_2$ [VP]$_3$?	[SBJ]$_2$ [MD]$_1$ [VP]$_3$.
does [SBJ]$_1$ [VP]$_2$?	[SBJ]$_1$ does not [VP]$_2$.
[does|do]$_1$ not [SBJ]$_2$ [VP]$_3$?	[SBJ]$_2$ [VP]$_3$.
[does|do]$_1$ [SBJ]$_2$ not [VP]$_3$?	[SBJ]$_2$ [VP]$_3$.
do [SBJ]$_1$ [VP]$_2$?	[SBJ]$_1$ do / do not [VP]$_2$.
did [SBJ]$_1$ [*]$_2$?	[SBJ]$_1$ did not [*]$_2$.
did not [SBJ]$_1$ [*]$_2$?	[SBJ]$_1$ did not [*]$_2$.

Table 2.10: A summary of question transformation rules. Some rules have been combined into one rule expression for clarity. **(Notations)** SBJ: subject, MD: modal verb, VB: verb, VP: verb phrase, ADJ: adjective, ADV: adverb, NP: noun phrase, backslash (\): exclusion. "xxx" and a forward slash indicate being semantically underspecified.

	US2016		MoralMaze	
	BLEU	%M	BLEU	%M
Original Questions	47.5	–	50.7	–
Basic Model	5.3	–	6.5	–
Copy Model	41.5	–	44.1	–
Rules	54.5	64%	51.9	48%
Rules (well-formed)	56.7	85%	54.5	69%

Table 2.11: Accuracy of extracting implicitly asserted propositions from questions. "%M" is the percentage of questions matched with any hand-crafted rules.

MoralMaze: This dataset consists of 8 episodes of the BBC Moral Maze Radio 4 program from the 2012 summer season[4] (Lawrence et al., 2015). The episodes deal with various issues, such as the banking system, welfare state, and British empire. In each episode, the BBC Radio presenter moderates argumentation among four regular panelists and three guest participants. This dataset has been annotated in the same way as US2016, and we filtered 314 pairs of a question and its asserted proposition. This dataset is not used for training or compiling rules; instead, it is only used as a test set to examine the domain-generality of the models.

Experiment Settings

For the neural models, we conduct two sets of experiments. First, we train and test the models on US2016 using 5-fold cross validation. Second, to examine domain generality, we train the models on the entire US2016 dataset and test on MoralMaze.

For the rule-based model, we compile the rules based on US2016 and test them on US2016 (previously seen) and MoralMaze (unseen).

The accuracy of the models is measured in terms of the BLEU score, where the references are asserted propositions annotated in the dataset.

Results

As shown in Table 2.11, the basic seq2seq model (row 2) performs poorly, because of the small size of the training data. On the other hand, the copy model (row 3) significantly improves the BLEU scores by 36.2–37.6 points, by learning to re-use words in input texts[5]. However, it still suffers the small data size, and its outputs are worse than the original questions without any transformation (row 1).

In contrast, the hand-crafted rules (rows 4–5) significantly improve performance and outperform the original questions. The effectiveness of the rule-based method on MoralMaze, which was not

[4] http://corpora.aifdb.org/mm2012

[5] Our model also outperforms a prior copy model (Gu et al., 2016) by more than 20 BLEU scores.

First word	% Matched Questions	BLEU			Exact Match		
		Before	After	Δ	Before	After	Δ
what, which	42 / 57 (74%)	43.9	51.9	7.9	5.3	12.3	7.0
who	19 / 19 (100%)	37.3	50.0	12.7	10.5	26.3	15.8
how	45 / 53 (85%)	44.3	61.0	16.7	7.5	34.0	26.4
why	26 / 31 (84%)	35.4	47.3	11.9	3.2	12.9	9.7
where, when	5 / 8 (62%)	45.3	49.8	4.5	25.0	25.0	0.0
do, does, did	68 / 71 (96%)	52.9	61.5	8.6	4.2	8.5	4.2
have, has	8 / 8 (100%)	52.4	69.7	17.2	0.0	37.5	37.5
is, are, was, were	61 / 65 (94%)	48.3	57.5	9.2	4.6	15.4	10.8
can, will, should, would, could	32 / 41 (78%)	49.8	61.6	11.8	4.9	12.2	7.3
All above	262 / 309 (85%)	45.7	56.7	11.0	5.8	18.4	12.6
All questions	359 / 565 (64%)	47.5	54.5	7.0	11.2	17.3	6.2

Table 2.12: BLEU and exact match (%) scores of questions before and after applying the hand-crafted rules (from US2016). "% Matched Questions" is the percentage of questions that match any of the hand-crafted rules.

used for compiling the rules, indicates that these rules generalize across argumentative dialogue[6]. The effectiveness of the rule-based method also suggests that there exist a high degree of syntactic regularities in how propositions are asserted implicitly in question form, and the hand-crafted rules provide interpretable insights into these regularities (Table 2.10).

Taking a closer look at the rule-based method, we find that many questions are subordinated or ill-formed, and thus the rules match only 64% of questions for US2016 and 48% of questions for MoralMaze. When we focus only on well-formed questions (that begin with a wh-adverb or auxiliary verb), the rules match 85% and 69% of questions for the respective dataset, and the BLEU scores improve by 2.2–2.6 points (row 4 vs. row 5). When analyzed by the first word of a question (Table 2.12), questions beginning with "have", "do", and modal verbs achieve the highest BLEU scores. Why-questions achieve the lowest, probably due to many variants possible; for example, "why isn't [SUBJECT] [ADJECTIVE]?" is most likely to be transformed to "[SUBJECT] is [ADJECTIVE]", whereas "why isn't [SUBJECT] [VERB]?" is to "[SUBJECT] should be [VERB]".

One limitation of the rule-based method, however, is that it cannot distinguish between questions that have the same syntactic structure but assert opposite propositions. For example, "Would you ...?" can mean both "You would ..." and "You would not ..." depending on the context. In order to separate these cases properly, we may need to take into account more nuanced features and context, and machine learning with large data would be the most promising direction eventually.

[6]Yet, we do not believe these rules would be effective beyond argumentation if the distribution of rhetorical questions and pure questions is significantly different from argumentative dialogue.

Top 1-8	Top 9-16	Top 17-24	Top 25-32
let (39)	fuck (5)	say (3)	bring (2)
look (7)	stop (5)	ask (2)	love (2)
have (7)	do (4)	vote (2)	drink (2)
wait (6)	check (3)	help (2)	pay (2)
thank (6)	give (3)	keep (2)	are (2)
please (6)	make (3)	find (2)	believe (2)
go (5)	get (3)	think (2)	talk (2)
take (5)	use (3)	forget (2)	screw (2)

Table 2.13: Root verbs and counts in imperatives.

2.5.5 Imperative

In this section, we collect imperatives in argumentative dialogue and examine a simple method for extracting propositions asserted in them. We do not build automated models for transformation (as in questions), because US2016 had no clear guidelines on how to annotate asserted propositions in imperatives when the dataset was built.

Model

No automated model is used in this section, but instead, we examine the applicability of the *you-should* theory in argumentation. Specifically, we analyze whether each imperative preserves the original intent when it is transformed to an assertive by adding "should", along with appropriate changes in the verb form, (implicit) subject, and object. We additionally analyze the argumentative relevancy of the transformed verb, that is, whether the imperative is mainly asserting that it should happen.

Data

We use imperatives in US2016 (Jo et al., 2019). We assume that a sentence is an imperative if its root is a verb in the bare infinitive form and has no explicit subject. Using Stanford CoreNLP, we chose locutions that are not questions and whose root is a verb with base form or second-person present case (VB/VBP), neither marked (e.g., "to go") nor modified by an auxiliary modal verb (e.g., "would go"). We found total 191 imperatives, and the most common root verbs are listed in Table 2.13.

Results

We found that 74% of the imperatives can be transformed to an assertion by adding "should" while preserving their original meaning[7]. And 80% of the transformed assertions were found to be argumentatively relevant content. For example, the imperative "Take away some of the pressure

[7]Many of the other cases are attributed to subject drop (e.g., "Thank you", "Doesn't work") and CoreNLP errors (e.g., "Please nothing on abortion", "So do police jobs").

placed on it" can be transformed to (and at the same time asserts that) "some of the pressure placed on it should be taken away". This result suggests that we can apply the *you-should* theory to many imperatives and extract implicitly asserted propositions in consistent ways.

Some imperatives were found to be rather rhetorical, and the propositions they assert cannot be obtained simply by adding "should". Those imperatives commonly include such verbs as "let", "fuck", "look", "wait", and "have". The verb "let" can assert different things. For instance, "Let's talk about the real issues facing america" asserts that "there are real issues facing america", while "Let's solve this problem in an international way" asserts that "we should solve this problem in an international way". The words "fuck" and "screw" are used to show strong disagreement and often assert that something should go away or be ignored.

We cannot apply the same transformation rule to the same verb blindly, as a verb can be argumentatively relevant sometimes and only rhetorical at other times depending on the context. For instance, the verb "take" in the above example is argumentatively relevant, but it can also be used only rhetorically as in "Take clean energy (as an example)".

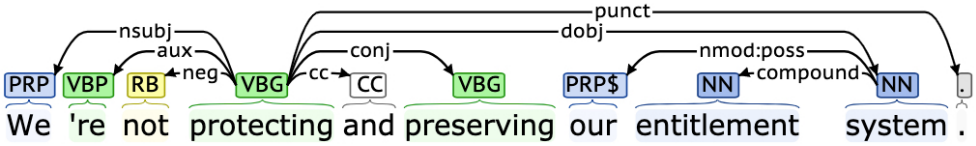
Based on our analyses, we propose rough two-step guidelines for annotating propositions that are implicitly asserted in imperatives. First, we may group imperatives by their semantics based on theories, such as *you-should* and *you-will* (Schwager, 2005). Second, for these imperatives, we may annotate whether the root verb is argumentatively relevant. For instance, if the *you-should* theory is applicable to an imperative, we may annotate whether its verb is at the core of the main argumentative content that the speaker asserts should happen; the assertive form of this imperative is likely to be a statement that proposes a policy or action (Park and Cardie, 2018). Argumentatively relevant imperatives may be annotated with asserted propositions using predefined transformation templates appropriate for their semantics. On the other hand, argumentatively irrelevant verbs may simply be rhetorical and need to be replaced properly. Annotation of these imperatives should handle many irregular cases, relying on the domain of the argumentation and the annotator's expertise.

2.5.6 Subject Reconstruction

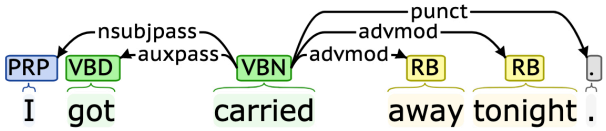
A locution or the speech content of reported speech may miss its subject due to segmentation. Hence, the SubjectReconstruction module aims to reconstruct the subject if it exists within the same sentence. We first trace the subject of each verb in every sentence, and then reconstruct the subject (along with auxiliary verbs) of a segmented text that begins with a verb whose subject is outside the text.

We trace the subject of a verb using basic dependency relations (from CoreNLP) as follows. When a verb has no subject relation with any words, we move to the word that is connected with the current verb through a dependency relation of the types: conjunct (conj), auxiliary (aux/auxpass), copula (cop), and open clausal complement (xcomp). The intuition is that this new word and the current word are likely to have the same subject. We repeat this process until we find a subject or no more move is available. In what follows, we illustrate the intuition behind using these dependency relations

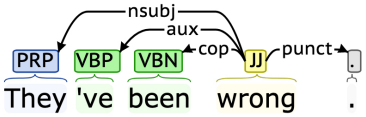
Conjunct (conj): Two verbs that are conjoined by a conjunction are likely to have the same subject. In the following example, "preserving" has the same subject as "protecting" does.



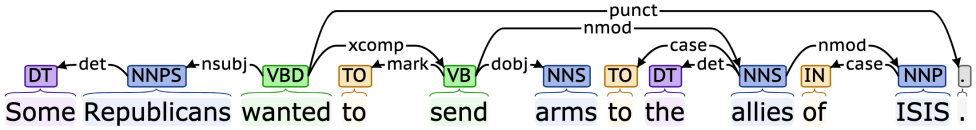
Auxiliary, passive auxiliary (aux, auxpass): An auxiliary verb that modifies a (passive) verb is likely to have the same subject as the modified verb does. In the following example, "got" has the same subject as "carried" does.



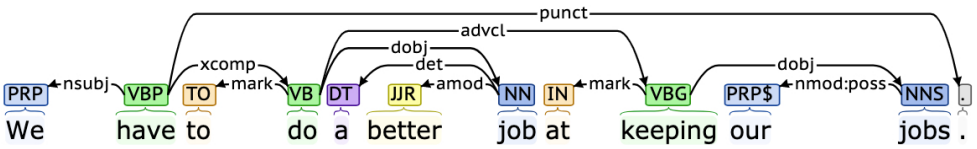
Copula (cop): A copula that joins a verb with its subject is likely to have the same subject as the verb. In the following example, "'ve" has the same subject as "wrong" does.



Open clausal complement (xcomp): An open clausal complement of a verb is likely to have the same subject as the verb does. In the following example, "send" has the same subject as "wanted" does.



Adverbial clause modifier (advcl): An adverbial clause modifier of a verb may or may not have the same subject as the verb does. In the following examples, the two sentences have the same structure of verb + object + marked adverbial clause modifier. However, in the first sentence, "keeping" has the same subject as "do" does, whereas in the second sentence, "leaving" has a different subject than "stop" does. For reliability, we do not include adverbial clause modifiers for tracing a subject.



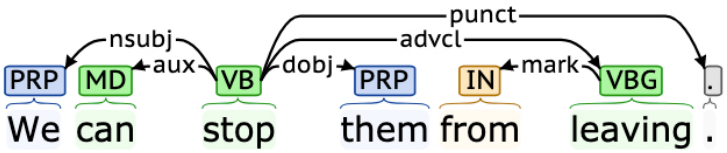
Prec	BLEU-Reconst	BLEU-Locution
71.4	62.6	59.1

(a) Performance of subject reconstruction.

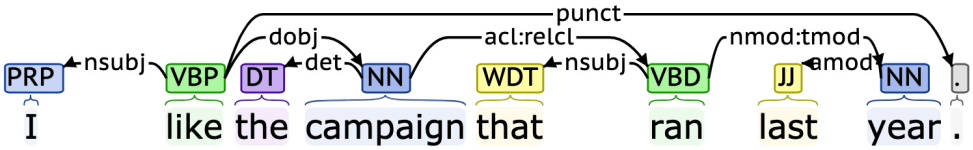
Reason	%
Ill-formed sentence	25%
No subject in the sentence	25%
Trace mistake	20%
Complex sentence	10%
Phrasal/clausal subject	10%
Wrong antecedents of relative pronouns	10%

(b) Reasons for subject identification errors.

Table 2.14: Results of subject identification.



Relative clause modifier (acl:relcl): Sometimes a verb's direct subject is a relative pronoun, in which case we move to the word modified by the verb via the acl:relcl relation. In the following example, "ran" modifies "campaign", which is the proper subject.



However, "which" may often refer to a phrase or a clause, and this method may not be able to capture that.

Results

We identified 96 locutions (1.2% of locutions) beginning with a verb whose subject is identified to be in the sentence yet outside the locution. We focus on 73% of them whose subjects are recovered in annotated propositions. Note that annotated subjects can be lexically different from the ones that are correctly identified by our method, due to imperfect anaphora resolution. Hence, our evaluation is based on manual comparison, checking if identified subjects and annotated subjects refer to the same thing.

	BLEU	Exact
Locution	75.5	47.3
Attention	47.2	12.4
Copy	76.2	49.3
Copy (short)	**76.6**	**50.1**

Table 2.15: Accuracy of revision. **Copy (short)** revises only short input texts.

As shown in Table 2.14a, the method identified subjects correctly for 71% of the locutions. Accordingly, the BLEU score improved by 3.5, compared to before subject reconstruction. Table 2.14b breaks down the reasons for errors. Sometimes the tracing method made a mistake (20%) or failed to capture a phrasal/clausal subject (10%). However, more commonly, CoreNLP could not properly handle sentences that are ill-formed (25%), missing a subject (25%), or too long/complex (10%). In some cases, it incorrectly identified the antecedents of relative pronouns (10%).

There exists other work that addresses recovering elided materials in sentences using dependencies (Schuster et al., 2018). Following some of the work, it would be an interesting direction to explore a richer set of dependency relations, such as the enhanced dependencies (Schuster and Manning, 2016).

2.5.7 Revision

While the previous modules handle major tasks, a processed locution may still need additional adjustments, including grammar correction. Hence, the Revision module makes adjustments to a processed locution and outputs final, asserted proposition(s). This task is formulated as a seq2seq problem, i.e., a model automatically learns and decides how to change the input, based on the data.

Models

We explore two models: standard attention (Luong et al., 2015) and copy mechanism. Both encode an input text using BiLSTM and decode proposition(s) using LSTM. The attention model computes the probability of a word being generated, using attention over the encoder's hidden states. It requires a lot of training data, whereas we already know that most input words remain unchanged. The copy model, on the other hand, decides internally whether to copy an input word or generate a new word. We use the same copy model as in Section 2.5.4.

We use two evaluation metrics: BLEU and exact match (percentage of outputs identical to the annotated propositions). We exclude locutions of reported speech and questions, to better focus on this module's performance. The baseline is to treat each locution as a proposition without modification. Accuracy is based on 5-fold CV.

Results

As shown in Table 2.15, the baseline (row 1) already achieves high performance, because locutions are often very similar to the propositions extracted from them unless they are reported speech or questions. For this reason, the attention model (row 2) performs poorly, as it tends to make many unnecessary adjustments to input locutions. The copy model (row 3) performs significantly better than the attention model, but sometimes it cannot handle long input texts and generated irrelevant content toward the end of an output. Leaving long input texts (25+ words) unmodified (row 4) slightly improved performance. Overall, the improvement over the baseline is rather modest.

The most notable and useful role of the copy model is correcting a verb case that was left incorrect due to anaphora resolution (e.g., "cooper want to" → "cooper wants to", "webb have had" → "webb has had"). This behavior is quite desirable. The model also sometimes removed non-propositional content and changed a person's first name to the full name as reflected in annotations. In general, the roles of the model remain lexical conversion rather than semantic conversion.

We found that the differences between generated and annotated propositions are derived mainly from unresolved non-personal anaphors (e.g., "it", "this", "that"). Furthermore, annotators sometimes insert omitted verb phrases (e.g., "You should." → "You should clinge to capitalism."; "not hard to do" → "not hard to dominate"). Such semantic information is not recovered by the current copy model.

2.5.8 End-to-end Extraction

So far, we have modularized the system and tested individual modules separately in order to find the optimal model and configuration for each module. In this section, we conduct a small experiment to see how well the cascade model extracts asserted propositions in an end-to-end way, i.e., the module takes an input utterance, goes through all modules, and outputs asserted propositions in the utterance. For this, we fix the optimal setting for each module learned in the previous sections. We use the same 5-fold cross validation to measure the extraction performance across the folds.

For evaluation, Figure 2.4 shows a possible metric for calculating precision, recall, and F1-score. For precision, each extracted proposition is matched with the most similar annotated proposition in terms of BLEU, and the precision score is the average BLEU score of the pairs. For recall, similarly, each annotated proposition is matched with the most similar extracted proposition, and the recall score is the average BLEU score of all pairs. The F1-score is the average of these two scores.

Prec	Recl	F1
60	57	59

Table 2.16: BLEU scores of the final end-to-end system in terms of precision, recall, and F1-score.

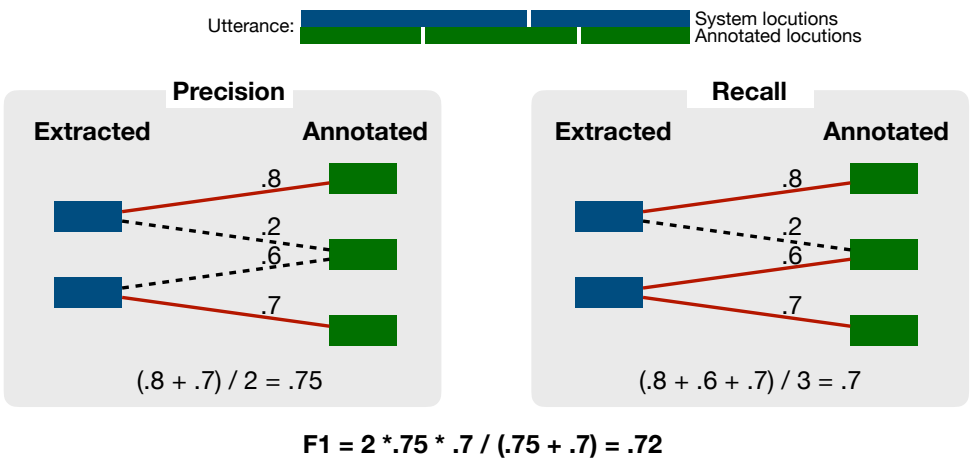


Figure 2.4: Evaluation metric for the cascade system. For precision, each extracted proposition is matched with the most similar annotated proposition. For recall, each annotated proposition is matched with the most similar extracted proposition. These two scores are averaged for the F1-score.

As shown in Table 2.16, the precision is 60. In other words, an extracted proposition has an average BLEU score of 60 in reference to its most similar annotated proposition. A similar trend holds for annotated propositions as well, i.e., an average BLEU score of 57 in reference to their most similar extracted propositions. Combining these two cases results in an average BLEU score of 59, which we believe is pretty high.

2.6 Conclusion

Pro- and counter-argumentation are constituted by asserted propositions, and thus the first step for identifying pro- and counter-arguments in argumentative dialogue is to extract asserted propositions. In this chapter, we presented a cascade model to extract asserted propositions from argumentative dialogue. This cascade model has seven subtasks: anaphora resolution, locution extraction, reported speech processing, question processing, imperative processing, subject reconstruction, and revision.

Anaphora resolution is crucial for recovering the semantic information of propositions, and the main bottleneck is to resolve 2nd-person singular and 3rd-person gender-neutral pronouns (e.g., “it” and “that”). However, we believe this problem will be solved soon as anaphora resolution is an active research problem in NLP.

Identifying locutions or ADUs is quite robust now. One fundamental question is whether we really need this as an explicit step. A positive effect is that it increases the explainability of the cascade model by locating where each proposition comes from. A potential downside is that it requires additional effort for data annotation, which can be more costly than necessary due to the fuzzy nature of exact locution boundaries. Furthermore, locution segmentation requires reconstructing missing subjects as a component, as our model currently has, whereas the cascade

model might be able to identify and decode missing subjects better by abstracting out locution segmentation.

For identifying asserted propositions in rhetorical questions, translating a rhetorical question to an asserted proposition seems to be relatively easy, as this process, as we found, is often mechanical. A bigger challenge is rather to classify whether a certain question is a pure question or a rhetorical question. While there is prior work on this problem (Bhattasali et al., 2015), we might additionally benefit from sarcasm detection (Joshi et al., 2017).

Identifying asserted propositions in imperatives should be accompanied by theoretical studies. For instance, to our knowledge, there is currently no theoretical background to draw upon in order to identify what exactly is asserted in an imperative. Does "Look how bad the system is" assert that we should look or that the system is bad? The former interpretation is in accordance with the you-should theory, which is clearly limited in this context, whereas the later interpretation is more suitable but is less clear as to how we come to this decision. The good thing is that there is rather a limited set of verbs that commonly accommodate multiple interpretations (e.g., "look", "let"), so research may begin with those verbs and gradually extend to infrequent ones.

Besides these modules, reported speech can be detected with fairly high accuracy. In addition, the source and the content of speech can also be extracted reliably. For subject reconstruction, our tracing method is fairly effective, and the accuracy is bounded mainly by the robustness of dependency parsing to ill-formed and complex sentences. The final revision with a seq2seq model remains mostly grammar error correction, and substantial semantic revision may require significantly different models.

Additional challenges outside of the current components of the cascade model are discussed in §2.4. For instance, the main challenge in identifying asserted propositions in conditionals comes from deciding whether a conditional clause is purely hypothetical or is an assertion. Many challenges could be resolved by collecting large data with careful annotation guidelines.

Some of the aforementioned challenges, such as anaphora resolution, do not need to be specific to argumentation. But others may greatly benefit if addressed in the context of argumentation. For instance, the proportions of rhetorical questions and pure questions in argumentation may be substantially different than in other genres like medical consultation. Similarly, interpretations of imperatives in argumentation may be different than in other genres. Hence, such problems seem to require argumentation- and perhaps domain-specific data and approaches.

Informal argumentative dialogue often accommodates locutions that are only rhetorical and do not contribute to argumentative structure, such as meta-argumentation. Such locutions could be identified in the locution segmentation component, which currently aims to filter out locutions irrelevant to the content of the ongoing argumentation based on human-annotated data. In addition, the model we will introduce in Chapter 3 distinguished meta-argumentation as a notable type of propositions (see §4.3.2 and Table 4.1), which indicates that it has some characteristic lexical features. We also see that this problem could benefit from metaphor detection and topical contrast, that is, how likely a certain locution should be interpreted rhetorically versus literally.

Chapter 3

Identifying Surface Types of Propositions

In this chapter, we present a novel unsupervised model for identifying different surface types of propositions that underlie a given set of dialogues (general, not necessarily argumentative, dialogues). This model is built on the basic assumption that a compound illocutionary act is a mixture of different surface types of propositions used. We strengthen the model by encoding several characteristics of linguistic phenomena. For instance, an illocutionary act to be performed at a specific time point depends on both the illocutionary act of the preceding utterance as well as the speaker's own preferences for certain acts. In addition, we observe that the surface types of propositions are characterized mainly by non-topical words and function words. Hence, our model de-emphasizes topical words in order to focus on words that signal distinctive types of propositions, by modeling topical themes as transitioning more slowly than illocutionary acts in dialogue.

We evaluate the model on two dissimilar corpora, CNET forum and NPS Chat corpus. The effectiveness of each modeling assumption is found to vary depending on the characteristics of data; de-emphasizing topical words yields improvement on the CNET corpus, while utilizing speaker preferences is advantageous on the NPS corpus. The components of our model complement one another to achieve robust performance on both corpora and outperform state-of-the-art baseline models.

3.1 Introduction

The main assumption of our model is that each utterance (or a turn) in dialogues performs a compound illocutionary act and can consist of more than one sentence. Each sentence in an utterance is then assumed to take one proposition-level type. As an example, let's think about dialogues in a tech forum (Figure 3.1). One possible compound illocutionary act in a tech forum is *to ask a question about a system problem*. An utterance performing this illocutionary act may include various types of propositions at surface levels, such as a system environment, an error message encountered, and a specific question. Similarly, another hypothetical illocutionary act is *to provide a solution*, and this act may include propositions of such types as general explanation,

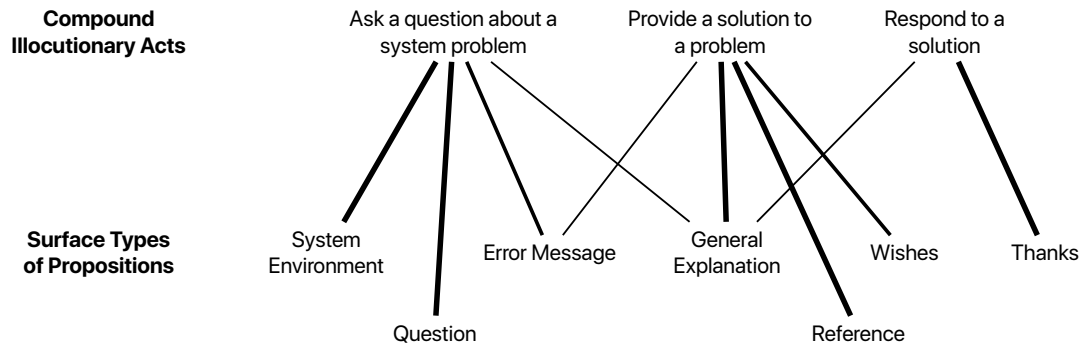


Figure 3.1: Examples of compound illocutionary acts and composing types of propositions in a tech forum. The thickness of an edges indicates the strength of association between an illocutionary act and a type of proposition.

references, and wishes. For modeling purposes, we represent each compound illocutionary act as a probability distribution over different surface types of propositions.

A challenge here is that in most cases, we do not know a complete set of compound illocutionary acts in a given corpus. Hence, our model assumes that not only the surface types of propositions are latent, but compound illocutionary acts are also latent. Now then the question is how can we better identify the latent compound illocutionary acts so that we can obtain proposition-level types that compose those illocutionary acts. To achieve this goal, we make three important modeling assumptions.

The first assumption is the conditional relevance between two compound illocutionary acts (Levinson, 1983; Martin and Rose, 2003), e.g., asking a question is likely to be followed by answering the question, a greeting by a greeting, and inviting by accepting the invitation (Sidnell, 2011). To encode conditional relevance, our model assumes a transition probability between every pair of compound illocutionary acts.

While there are general tendencies about what compound illocutionary act would likely follow what illocutionary act, an illocutionary act to be performed depends also on the speaker's personal preferences (Appling et al., 2013). For instance, suppose there is a system expert in a tech forum who answers questions most of the time rather than asking a question. By taking into account speaker preferences, the model may be able to better identify the compound illocutionary acts performed by this person, especially when an utterance has mixed characteristics of different illocutionary acts (e.g., a response to a question includes some clarification question). Therefore, the second assumption of our model is that each speaker has preferences for certain illocutionary acts, which are represented as a probability distribution over compound illocutionary acts.

The third assumption of the model is that different types of propositions are mainly characterized by different function words and those words that are less specific to discussion topics. For instance, questions are characterized by wh-adverbs ("why", "how") and the question mark, rather than topical words ("Windows", "iPhone"). Similarly, error messages are characterized by some template words ("Traceback", ":", "Error"). Hence, the model learns the language model of

each surface-level type of proposition while trying to separate out topical words. This goal is accomplished by encoding that the types of propositions transition at utterance levels whereas the background topic is consistent throughout the dialogue. As a result two kinds of language models are learned: one for different surface-level types of propositions and one for background topics. While some existing models assume a background or domain-specific language model to filter out common words (Lee et al., 2013; Paul, 2012; Ritter et al., 2010), they either require domain labels or do not learn background topics underlying dialogues.

To illustrate the effectiveness of our model, we evaluate it on two dialogue corpora with very different characteristics in terms of utterance length, the number of speakers per dialogue, and domain: CNET and NPS Chat Corpus. Since these corpora are annotated with utterance-level compound illocutionary acts, we directly evaluate the model's ability to identify latent compound illocutionary acts and demonstrate that our model is more effective than baseline models. Furthermore, we qualitatively analyze the latent types of propositions learned by the model and demonstrate that they are reasonable components of the identified illocutionary acts. Lastly, by exploring different settings of the model parameters for each corpus, we use our model as a lens to understand the nature of the corpus dialogues, which may inform future model design.

3.2 Related Work

Speech act theory (Austin, 1975) makes a distinction between the illocutionary, social intention of an utterance (as seen in the indirect sentence "Can you pass the salt?") and the locutionary act of an utterance, which includes the ostensible surface-level meaning of the words. Although the theory focuses mainly on basic sentence-level illocutionary acts (e.g., assertives, imperatives), more complex illocutionary acts can be thought of in real-life dialogue by considering compound illocutionary acts that consist of multiple sentences and sentence-level illocutionary acts (van Eemeren and Grootendorst, 1984). Example (compound) illocutionary acts used in computational systems include *yes-no question*, *statement*, *backchannel*, and *opinion* (Jurafsky et al., 1998). In this work, illocutionary acts refer to compound illocutionary acts.

Winograd and Flores (1986) were some of the first to conceptualize illocutionary acts[1] with state transitions as a model for conversation. Similarly, contemporary unsupervised models often use a hidden Markov model (HMM) to structure a generative process of utterance sequences (Ritter et al., 2010). It is commonly assumed that each hidden state corresponds to an illocutionary act, but different approaches use different representations for states.

One common representation of a state is a multinomial distribution over words, from which words related to an illocutionary act are generated. Often, this generative process includes domain- or topic-related language models that are independent of states and used to filter out words unrelated to illocutionary acts (Lee et al., 2013; Ritter et al., 2010). However, these language models have some limitations. For instance, Lee et al. (2013) rely on domain labels for learning domain-specific language models, which may require human annotation, whereas our model

[1] In the NLP literature, illocutionary acts are more commonly referred to as dialogue acts. For consistency, we use the term "illocutionary acts".

learns them without labels. Ritter et al. (2010) learn conversation-specific language models to filter out topical words. We take a different approach, simultaneously learning background topics underlying the entire corpus and filtering out these topical words. Although most models incorporate a general language model to separate out common words (Lee et al., 2013; Paul, 2012; Ritter et al., 2010), we do not learn it because we assume that many common words (e.g., function words) are relevant to illocutionary acts.

Word embedding vector representations have also been researched as the outputs of latent states. For example, Brychcín and Král (2017) represent an utterance as a weighted sum of word vectors from GloVe[2]. Each utterance vector is generated from a Gaussian distribution that parameterizes a latent state. This model has been shown to capture illocutionary acts effectively for short utterances.

Illocutionary acts are not completely determined by preceding acts (Levinson, 1983), and this difficulty can be overcome partly by modeling speaker preferences, as there is evidence that each speaker has preferences for certain illocutionary acts (Appling et al., 2013). Joty et al. (2011) model speakers as outputs generated by an HMM, but this structure makes it hard to adjust the contribution of speaker preferences and may overestimate the influence of speakers. We model speaker preferences more directly such that the preceding illocutionary act and the speaker preferences together determine an utterance's probability distribution over illocutionary acts.

One reason for the nondeterministic nature of illocutionary acts is that one utterance can involve more than one act (Levinson, 1983); this is a similar concept to compound illocutionary acts, suggesting that one language model per illocutionary may not be enough. Paul (2012) represents latent states as mixtures of topics, but there is no one-to-one relationship between states and illocutionary acts. Joty et al. (2011) assume that words are drawn individually from a fixed number of language models specific to each illocutionary act. However, the speech act theory (Searle, 1969) suggests that usually one sentence performs one elementary illocutionary act and a propositional act of a certain type. So, we constrain each sentence in an utterance to one language model, which represents a surface-level type of proposition. Thus, utterances, which may consist of multiple sentences, are represented as a mixture of those types of propositions.

Word order in an utterance may play an important role in determining an illocutionary act, as in the difference between "I am correct" and "am I correct". Ezen-Can and Boyer (2015) compute the similarity between utterances based on word order using a Markov random field and cluster similar utterances to identify illocutionary acts. This model, however, does not consider transitions between clusters.

Online conversations often have asynchronous, deeper than two-level tree structure (e.g., nested replies). In Joty et al. (2011)'s model, individual reply paths from the first utterance to terminal utterances are teased apart into separate sequential conversations by duplicating utterances. However, this method counts the same utterance multiple times and requires an aggregation method for making a final decision of the illocutionary act for each utterance. We address multi-level structure without duplicating utterances.

[2] https://nlp.stanford.edu/projects/glove/

	Speaker Preference	Transitions between Illocutionary Acts	Language Models Unrelated to Illocutionary Acts	Multi-Level Structure Support	Mixture of Language Models for Illocutionary Acts
Brychcín and Král (2017)	N	Y	-	N	N
Ezen-Can and Boyer (2015)	N	N	-	N	N
Lee et al. (2013)	N	Y	GD	N	N
Paul (2012)	N	Y	G	N	Y
Joty et al. (2011)	Y	Y	U	Y	Y
Ritter et al. (2010)	N	Y	GD	N	N
Our model	Y	Y	D	Y	Y

Table 3.1: Properties of baseline models. (G: general background, D: domain-specific, U: unspecified)

The properties of the models explained so far are summarized in Table 3.1.

The relative importance of each structural component in a model may not be identical across all corpora. Differences, especially as they are attributed to meaningful contextual variables, can be interesting both practically and theoretically. One contribution of our work is to consider how differences in these kinds of contextual variables lead to meaningful differences in the utility of our different modeling assumptions. More typical work in the field has emphasized methodological concerns such as minimization of parameter tuning, for example, by using a hierarchical Dirichlet process to determine the number of latent illocutionary acts automatically (Lee et al., 2013; Ritter et al., 2010) or by simply assuming that a word is equally likely to be related or unrelated to an illocutionary act (Paul, 2012). While these efforts are useful, especially when maximizing the likelihood of the data, searching for the optimal values of parameters for illocutionary act recognition may allow us to better understand the contribution of each model component depending on the characteristics of the dialogue, which in turn can inform future model design.

3.3 Model Design

Our model, CSM (content word filtering and speaker preferences model), is based on an HMM combined with components for topical word filtering and speaker preferences. In the model, each latent state represents an utterance-level compound illocutionary act as a mixture of language models, each of which represents a sentence-level surface type of proposition; each sentence in an utterance is assigned one such type. To filter topical words, there is a set of *background topics* shared across dialogues, and each dialogue is assigned a background topic that underlies it.

A transition between states is defined on every parent-child (or, two consecutive) utterance pair, supporting multi-level tree structure. The state of an utterance is dependent on both the its preceding utterance's state and its speaker. Speakers are specific to each conversation, i.e., a speaker participating in multiple dialogues is treated as different speakers for different dialogues.

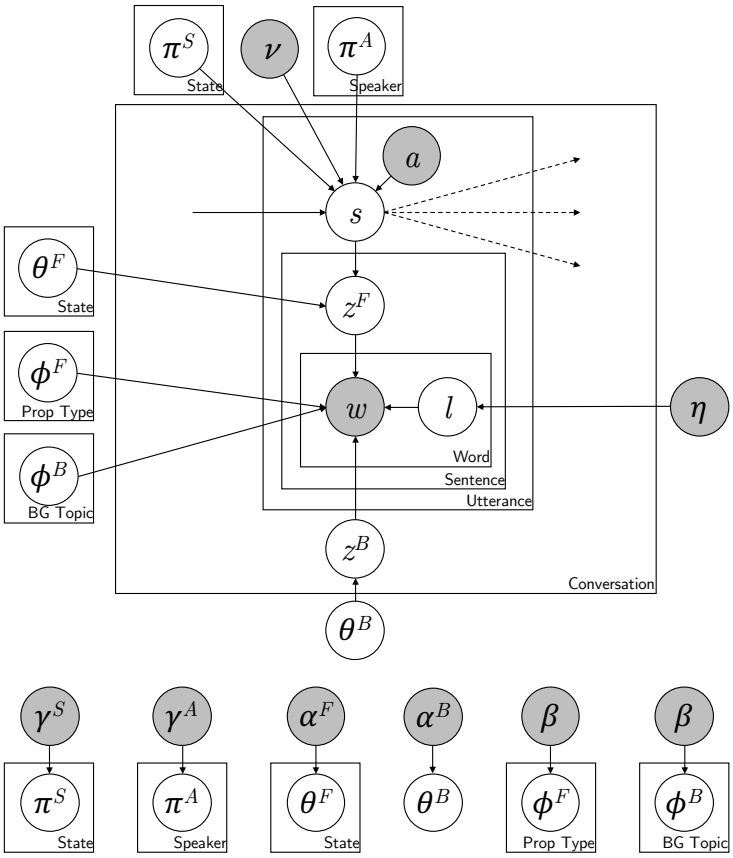


Figure 3.2: Graphical representation. Shaded nodes represent observable variables.

The graphical representation of CSM is in Figure 3.2.

The formal generative process of dialogues is as follows:

- For each speaker a, draw a preference distribution over states $\pi_a^A \sim \text{Dir}(\gamma^A)$.
- For each state s
  - Draw a transition probability distribution over states $\pi_s^S \sim \text{Dir}(\gamma^S)$.
  - Draw a probability distribution over the types of propositions $\theta_s^F \sim \text{Dir}(\alpha^F)$.
- For each type of proposition t, draw a probability distribution over words $\phi_t^F \sim \text{Dir}(\beta)$.
- For each background topic t, draw a probability distribution over words $\phi_t^B \sim \text{Dir}(\beta)$.
- For the corpus, draw a distribution over background topics $\theta^B \sim \text{Dir}(\alpha^B)$.
- For each dialogue
  - Draw a background topic $z^B \sim \text{Cat}(\theta^B)$.
  - For each utterance u, with its speaker a_u, its preceding utterance p, and the preceding utterance's state s_p,
    - Draw a state $s_u \sim \text{Cat}(\nu\pi_{s_p}^S + (1-\nu)\pi_{a_u}^A)$.
    - For each sentence
      - Draw a type of proposition $z^F \sim \text{Cat}(\theta_{s_u}^F)$.
      - For each word
        - Draw an indicator of "proposition-level type" or "background topic" $l \sim$

Notation	Meaning
N_{ij}^{SS}	Transition from state i to state j
N_{ij}^{AS}	Assignment of speaker i to state j
N_{ij}^{SF}	Assignment of state i to proposition-level type j
N_{j}^{B}	Assignment to background topic j
N_{ij}^{FW}	Assignment of proposition-level type i to word j
N_{ij}^{BW}	Assignment of background topic i to word j

Table 3.2: Descriptions of counter matrices.

$\text{Cat}((\eta, 1-\eta))$.
- If l is "proposition-level type", draw a word $w \sim \text{Cat}(\phi^F_{z^F})$.
- If l is "background topic", draw a word $w \sim \text{Cat}(\phi^B_{z^B})$.

According to this model, topical words are separated out into background topics in several ways. A background topic does not transition as frequently as the types of propositions do within a dialogue. Accordingly, words that are consistently used across utterances in a dialogue are likely to be clustered into the background topic z^B, whereas words whose use is sensitive to the previous state and the speaker are likely to be clustered into one type z^F. However, this design may cause common function words, such as pronouns, prepositions, and punctuations to be separated out as well. Hence, η, the probability of a word being from a proposition-level type, adjusts the degree of filtering. The higher the η value, the more likely words are to be generated for a proposition-level type, and thus the more function words are included in proposition-level types, leaving background topics with topical words. Hence, we may set η high if we believe common words play an important role in determining the types of propositions in a corpus and low otherwise. Background topics capture topical words underlying the entire corpus, as they are shared across dialogues.

Speaker preferences are captured as a probability distribution over illocutionary acts (π^A), which, along with the preceding state, affects the probability of the current state. ν adjusts the contribution of general transition tendencies between illocutionary acts (as opposed to the speaker preferences); hence, the higher ν, the weaker the contribution of speaker preferences. So, we may set ν low if each speaker is believed to have invariant preferences for certain illocutionary acts. If there is not enough such evidence and the dialogue is driven without specific preferences of the speakers, then we may set ν high. We find that different corpora have different optimal values of ν depending on their nature.

We use collapsed Gibbs sampling for inference to integrate out π^S, π^A, θ^F, θ^B, ϕ^F, and ϕ^B. Given dialogue text with speakers for each utterance, along with the hyperparameters, ν, and η, the Gibbs sampler estimates the following variables using count matrices explained in Table 3.2:

$$\pi^S_{ij} = \frac{N^{SS}_{ij} + \gamma^S}{\sum_{j'}(N^{SS}_{ij'} + \gamma^S)}, \pi^A_{ij} = \frac{N^{AS}_{ij} + \gamma^A}{\sum_{j'}(N^{AS}_{ij'} + \gamma^A)}$$

$$\theta_{ij}^{F} = \frac{N_{ij}^{SF} + \alpha^{F}}{\sum_{j'}(N_{ij'}^{SF} + \alpha^{F})}, \theta_{j}^{B} = \frac{N_{j}^{B} + \alpha^{B}}{\sum_{j'}(N_{j'}^{B} + \alpha^{B})}$$

$$\phi_{ij}^{F} = \frac{N_{ij}^{FW} + \beta}{\sum_{j'}(N_{ij'}^{FW} + \beta)}, \phi_{ij}^{B} = \frac{N_{ij}^{BW} + \beta}{\sum_{j'}(N_{ij'}^{BW} + \beta)}.$$

We may use slice sampling (Neal, 2003) to estimate ν and η too, but the estimated values of ν and η may not be optimal for illocutionary act recognition. We can also obtain state assignments for utterances by taking a sample from the Gibbs sampler. Detailed derivation for Gibbs sampling and the code are available online[3].

3.4 Experiment Settings

This section describes our evaluation method and settings.

3.4.1 Task and Metrics

We evaluate our model in terms of accuracy in utterance-level compound illocutionary act recognition. Since the output of the model is assignments to latent states (not pre-determined illocutionary act labels) for utterances, we use a clustering evaluation method, as adopted by previous work on unsupervised modeling of illocutionary acts. Specifically, we use homogeneity, completeness, and v-measure as metrics (Rosenberg and Hirschberg, 2007). Borrowing the original notations, suppose there are N utterances, a set of true illocutionary acts $C = \{c_i | i = 1, \cdots, n\}$, and a set of learned clusters $K = \{k_j | j = 1, \cdots, m\}$. Let a_{ij} denote the number of utterances whose true illocutionary act is c_i and assigned cluster is k_j. Homogeneity represents the degree to which utterances assigned to the same cluster by the model share the same illocutionary act in the labeled corpus. This measure is reflected in the conditional entropy of the illocutionary act distribution given the proposed clustering $H(C|K)$, which is 0 in the perfectly homogeneous case. Since the range of this value depends on the size of each illocutionary act, it is normalized by the entropy of the true illocutionary act distribution $H(C)$. Following the convention of 1 being desirable and 0 undesirable, homogeneity is defined as:

$$h = \begin{cases} 1 & \text{if } |C| = 1 \text{ or } |K| = 1 \\ 1 - \frac{H(C|K)}{H(C)} & \text{otherwise,} \end{cases},$$

where

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N}\log\frac{a_{ck}}{\sum_{c=1}^{|C|}a_{ck}},$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{\sum_{k=1}^{|K|}a_{ck}}{N}\log\frac{\sum_{k=1}^{|K|}a_{ck}}{N}.$$

Completeness represents the degree to which utterances that have the same illocutionary act according to the gold standard are assigned to the same cluster. This measure is identical to

[3] https://github.com/yohanjo/Dialogue-Acts

	CNET	NPS
# dialogues	310	15
# utterances	1,332	10,567
# compound illocutionary acts	12	15
# domains	24	-
Median # utterances/dialogue	3	706
Median # words/utterance	51	2
Median # speakers/dialogue	2	94

Table 3.3: Corpora statistics.

homogeniety except that we measure entropies for K instead of C. Completeness is defined as:

$$c = \begin{cases} 1 & \text{if } |C| = 1 \text{ or } |K| = 1 \\ 1 - \frac{H(K|C)}{H(K)} & \text{otherwise,} \end{cases},$$

where

$$H(K|C) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{a_{ck}}{N}\log\frac{a_{ck}}{\sum_{k=1}^{|K|}a_{ck}},$$

$$H(K) = -\sum_{k=1}^{|K|}\frac{\sum_{c=1}^{|C|}a_{ck}}{N}\log\frac{\sum_{c=1}^{|C|}a_{ck}}{N}.$$

V-measure is the harmonic mean of homogeneity and completeness:

$$v = \frac{2 * h * c}{h + c}.$$

These metrics are easy to interpret and have been demonstrated to be invariant to dataset size and number of clusters. This enables a meaningful comparison of accuracy across different corpora.

3.4.2 Corpora and Preprocessing

We evaluate on two corpora: CNET and NPS Chat (see Table 3.3 for statistics).

CNET (Kim et al., 2010) is a set of post threads from the Operating System, Software, Hardware, and Web Development sub-forums of CNET. This corpus is tagged with 12 compound illocutionary acts at post levels, including *Question-Question*, *Question-Confirmation*, *Answer-Add*, *Resolution*, and *Other* (Table 3.4). Note that question- and answer-related acts are two-level. Most posts are tagged with one act; in case a post is tagged with multiple acts, we choose the first act in the meta-data[4]. Each post was annotated by two annotators, achieving Cohen's κ of 0.59.

Each post is considered an utterance and each thread as a dialogue. Each thread has only a few posts (median 3) and involves a few speakers (median 2). There are total 310 dialogues and 1,332 utterances, covering such domains as hardware (e.g., drive, RAM, CPU, motherboard), networks, operating systems, registry, and email. Since there are many URLs, email addresses, and numbers

[4]Some tagging systems, such as the DAMSL-style, break down an utterance that has multiple illocutionary acts.

CNET	NPS
Question-Question	Accept
Question-Add	Bye
Question-Confirmation	Clarify
Question-Correction	Continuer
Answer-Answer	Emotion
Answer-Add	Emphasis
Answer-Confirmation	Greet
Answer-Correction	Reject
Answer-Objection	Statement
Resolution	System
Reproduction	yAnswer
Other	nAnswer
	whQuestion
	ynQuestion
	Other

Table 3.4: Compound illocutionary acts tagged in the corpora.

in text, we normalize them with special tokens using regular expressions, and conduct tokenization with the Stanford PTBTokenizer included in Stanford Parser 3.7.0[5].

NPS Chat (Forsythand and Martell, 2007) is a set of dialogues from various online chat services. This corpus is tagged with 15 compound illocutionary acts at turn levels, including *Emotion*, *System*, and *whQuestion* (Table 3.4). The details of the annotation process is not available from the original paper. But a naive Bayes classifier with 22 hand-crafted features achieved F1-scores between 0 and 98.7 across the acts, and without the acts that scored 0 (due to too low frequency), the average F1-score was 51.1.

Each turn is considered an utterance. There are total 15 dialogues and 10,567 utterances, covering a wide range of casual topics in daily life. Each dialogue is long (median 706 utterances) and involves many speakers (median 94). This corpus has already been tokenized, so we only replace usernames with a special token. Dialogues in NPS have no reply structure, but we build in multi-level tree structure, simply treating an utterance that mentions another user as a child of the nearest utterance of the mentioned user. We compare the accuracy of the multi-level structure and the original linear structure in Section 3.5.

3.4.3 Models and Parameters

We set the numbers of states and background topics to the numbers of illocutionary acts and domains, respectively, if these numbers are available. For NPS, we search for the optimal number of background topics between 1 and 2, because there are only a few dialogues. The optimal

[5] https://nlp.stanford.edu/software/lex-parser.html

number of proposition-level types is chosen among multiples of five between the number of states and four times the number of states, and the weights for state transition (ν) and foreground topics (η) are chosen among multiples of 0.1. For Dirichlet hyperparameters, we use $\alpha^F = 0.1, \gamma^A = 0.1, \beta = 0.001$ to induce sparsity, and $\gamma^S = 1, \alpha^B = 1$ for the uniform distribution over all configurations.

We randomly split each corpus into five groups and use three groups for training, one for parameter tuning, and one for testing. We run 5-fold cross-validation and report the average optimal parameter values and accuracy across the folds. The number of sampling iterations was chosen such that the log-likelihood of the data has converged. For each fold, we take 10 samples during inference on the test data with interval of 10 iterations and compute the mean and standard deviation of the 50 samples from all folds.

We compare our model with the three most recent unsupervised models we surveyed. The baseline models and settings are as follows.

Gaussian mixture HMM (Brychcín and Král, 2017), based on an HMM, has a characteristic output representation: utterance vectors. These vectors are generated from Gaussian distributions instead of using language models as in most existing models. After following their preprocessing steps, we trained the model on the training data, chose the optimal word vector dimensionality on the validation data (among 50, 100, 200, and 300, as used in the original model), and performed inference on the test data. We used the original source code from the authors for training and modified the code for inference.

MRF-based clustering (Ezen-Can and Boyer, 2015) considers word order within an utterance to calculate similarity between utterances using an MRF. Then k-medoids clustering is conducted based on the similarity scores, resulting in clusters that represent illocutionary acts. The similarity score between two utterances is asymmetric, so we took the average value of each direction and inversed it to obtain the distance between two utterances. We trained the model on the training data, chose the optimal parameter values ($\lambda_i, \lambda_t, \alpha_d$ in the original paper) on the validation data, and assigned clusters to the test data. We implemented the algorithm since the original code was not available.

HDP-HMM (Lee et al., 2013) is based on an HMM, and each word comes from either the state-specific, general background, or domain-specific language model. HDP-HMM automatically decides the number of states using a hierarchical Dirichlet process, but we manually set the number of illocutionary acts in our experiment, assuming that we know the number of the acts of interest. We trained the model on the training data and performed inference on the test data; the validation data was not used since there are no parameters to tune. We used the original source code from the authors for training and modified the code for inference.

3.5 Results

The accuracy of illocutionary act recognition in terms of homogeneity, completeness, and v-measure on both corpora is summarized in Table 3.5. We also tested the following configurations:

Model	CNET			NPS		
	H	C	V	H	C	V
Brychcín and Král (2017)	.13$_{\pm .00}$	.09$_{\pm .00}$	.10$_{\pm .00}$	.24$_{\pm .10}$	**.33**$_{\pm .06}$	.28$_{\pm .08}$
Ezen-Can and Boyer (2015)	.03$_{\pm .00}$	.37$_{\pm .00}$	.05$_{\pm .00}$	.26$_{\pm .00}$	.33$_{\pm .00}$	.28$_{\pm .00}$
Lee et al. (2013)	.09$_{\pm .03}$	.16$_{\pm .03}$	.11$_{\pm .03}$	**.36**$_{\pm .02}$	.28$_{\pm .02}$	.31$_{\pm .02}$
CSM	.24$_{\pm .03}$	**.38**$_{\pm .04}$	**.29**$_{\pm .03}$	.35$_{\pm .04}$	.31$_{\pm .04}$	**.33**$_{\pm .04}$
CSM + Domain	**.27**$_{\pm .02}$	.33$_{\pm .11}$	.29$_{\pm .05}$		N/A	
CSM – Speaker	.24$_{\pm .03}$	**.38**$_{\pm .04}$	**.29**$_{\pm .03}$	.21$_{\pm .03}$	.19$_{\pm .05}$	.20$_{\pm .04}$
CSM – Multi-level	.23$_{\pm .04}$	.33$_{\pm .06}$	.27$_{\pm .04}$	.35$_{\pm .02}$	.30$_{\pm .04}$	.32$_{\pm .03}$
CSM – Background Topics	.15$_{\pm .03}$	.11$_{\pm .02}$	.12$_{\pm .02}$	.35$_{\pm .04}$	.31$_{\pm .04}$	**.33**$_{\pm .04}$

Table 3.5: Accuracy of illocutionary act recognition (the higher the better). Smaller numbers are population standard deviations (H: homogeneity, C: completeness, V: v-measure). Optimal parameter values for CSM: # proposition-level types=34, $\eta = .86$, $\nu = 1.00$ for CNET and # proposition-level types=35, $\eta = 1.00$, $\nu = 0.58$ for NPS.

- **CSM + Domain** uses true domain labels when learning background topics by forcefully assigning a dialogue the background topic corresponding to the true label.
- **CSM – Speaker** does not use speaker preferences, by setting $\nu = 1$.
- **CSM – Multi-level** ignores multi-level structure; that is, utterances in each dialogue are linearly ordered by time.
- **CSM – Background Topics** uses only one background topic.

Overall, our model performs significantly better than the baselines for CNET and marginally better for NPS. The baseline models show a large variance in performance depending on the characteristics of the corpus. In contrast, our model has a low variance between the corpora, because the topical word filtering, distinction between utterance-level illocutionary acts and sentence-level types of propositions, and speaker preferences complement one another to adapt to different corpora. For example, topical word filtering and the distinction between illocutionary acts and proposition-level types play more significant roles than speaker preferences on CNET, whereas their effects are reversed on NPS. The details will be described later with qualitative analyses.

There may be several reasons for the poor performance of the baseline models on CNET. First, in our model, each illocutionary act (latent state) is a probability distribution over different types of propositions, which better characterizes compound illocutionary acts, especially for long utterances in CNET. The utterances in CNET may be too complex for the baseline models, which use a simpler representation for compound illocutionary acts. Another reason for the low performance could be that the baseline models do not filter out topical words as our model does.

In the remainder of this section, we describe our qualitative analysis on the results. All examples

Topic	Top 5 words
BT0	drive partition drives partitions c
BT1	router wireless network connected connection
BT2	vista camera canon windows scanner
BT3	drive ipod touch data recovery
BT4	speakers firewall sound still no
BT5	/ \blaster dos drive
BT6	windows cd i xp boot
BT7	page xp sp3 ! content
BT8	ram mhz 1gb 512mb screen
BT9	his rupesh to company he
BT10	xp drive drivers new hard
BT11	tv port cpu motherboard grounded
BT12	file files copy external mac
BT13	“ password flash ##NUMBER## ?
BT14	fan fans cpu case air
BT15	ram card 2.4 graphics nvidia
BT16	registry file shutdown machines screen
BT17	div site % ie6 firefox
BT18	printer sound would card contact
BT19	hosting web hostgator they host
BT20	ubuntu linux memory boot reader
BT21	mac compression archive format trash
BT22	bluetooth router wireless laptop 802.11
BT23	email address account mail bounce

Table 3.6: Background topics learned from CNET.

shown in the analysis are from the result with the optimal parameter values for the first fold.

Filtering topical words: Our model effectively separates topical words from words that are related to proposition-level types, even without using the domain label of each dialogue. As an example, the background topics learned by our model from CNET are shown in Table 3.6. These topics are clearly related to the subjects of the forum, rather than reflecting specific types of propositions, and the topics are distinctive from one another and cohesive in themselves.

The main purpose of learning background topics is to better identify illocutionary acts by filtering out topical words in characterizing proposition-level types. The learned background topics serve this purpose well, as including these topics in the model increases v-measure by 0.17 (CSM vs. CSM – Background Topics). It is also promising that the background topics learned without domain labels perform as well as when they are learned with domain labels (CSM vs. CSM + Domain), because domain labels may not always be available.

proposition-level Type	Top Words
Environments (FT20)	. i a ##NUMBER## and have -rrb- xp -lrb- : windows my is the dell vista
Error msgs (FT12)	. the # * messages / : it log
Asking (FT19)	any help you ? ! . appreciated i suggestions
Thanking (FT17)	thanks . for the ! in advance help your all response
Problem (FT8)	: \file is the c corrupted following missing or error
Wishes (FT14)	. bob good luck
Reference (FT5)	##URL##
Praise (FT1)	. thank you ˜ sovereign , and are excellent recommendations
Explanation (FT10)	the . to , i and a it you is that of

(a) Proposition-level types learned from CNET.

Proposition-level Type	Top Words
Wh question (FT7)	##USERNAME## ? how you are u good is round where who . ??
Wh question (FT27)	##USERNAME## ? you i u what how , ok 'm for up do have
YN question (FT1)	chat any wanna / me pm to ? anyone f guys m want here
Greeting (FT5)	##USERNAME## hi hey :) hello wb ! ... hiya ty
Laughing (FT0)	##USERNAME## lol lmao yes ! hey up !!!! ?
Laughing (FT12)	lol ##USERNAME## haha ! brb omg nite hiyas hb :p !!! . ha lmfao
Emotion (FT30)	ok ! im lol my its in " ... oh always
System logs (FT25)	part join

(b) Proposition-level types learned from NPS.

Table 3.7: Proposition-level types learned from the corpora.

Common words play an important role in distinguishing different surface types of propositions in CNET as indicated by the high optimal value of $\eta = 0.86$ (the probability of a word being drawn from a proposition-level type). The higher η means more common words are included in proposition-level types, leaving background topics with highly topical words (Section 3.3). The high η is evidence contrary to the common practice of designating a general background topic to filter out common words and assuming that a word is equally likely to be related to an illocutionary act or a background topic (Lee et al., 2013; Paul, 2012).

The effectiveness of our method of separating background topics turns out to diminish when there are no consistent conversational topics within and across dialogues as in NPS. Our model learns not to use background topics ($\eta = 1$) for NPS, because background topics may filter out common words that occur more consistently throughout a dialogue than topical words do.

Mixture of proposition-level types: As a consequence of filtering out topical words, the learned surface types of propositions reflect various types of propositions that characterize compound illocutionary acts in each corpus. Some of the learned types from CNET are shown in Table 3.7a. They capture important types that constitute compound illocutionary acts that are assigned to each post in CNET. For example, *Question-Question* is a compound illocutionary act that often starts a dialogue, and conducting this act typically includes types, such as explaining the system

environment and situation, asking a question, and thanking, as shown in the following post:

> "I am currently running Windows XP Media Edition on a 500G hard drive." (FT20) / "I want to move my XP to it's own partition, move all of my files(music, games, work) to another, and then install the Windows 7 beta on another partition." (FT10) / "I don't know if this is possible or not, but I have access to Partition Magic 8, and am wondering if I can do it with that or not." (FT10) / "I am not worried about installing 7 on another partition, but am not sure if I can move my files onto a separate one while keeping XP intact." (FT10) / "Any help is great, thank you." (FT17)

Likewise, the compound illocutionary act *Answer-Answer* includes such types as wishes or URLs, as in the posts:

> "Simple - Download and install the Vista Rebel XT drivers from canon usa.com." (FT10) / "Once installed.....go to camera menu and switch the communication to Print/PTP." (FT10) / "Don't forget to switch it back if you're connecting to an XP machine." (FT10) / "Good Luck" (FT14)
>
> http://forums.microsoft.com/MSDN/ShowPost.aspx?PostID=1996406&SiteID=1 (FT5)

When a problem is resolved, the illocutionary act of *Resolution* may be performed with thanking and praising:

> "Excellent summary Thank you." (FT1) / "Sounds like at some point it's worth us making the transition to a CMS..." (FT10)

FT10 seems like general explanations and statements that do not belong to any other types specifically. Modeling each compound illocutionary act as a mixture of different surface types of propositions is effective for CNET, as our model beats the baselines significantly.

The types learned from NPS also reflect those that characterize compound illocutionary acts in the corpus (Table 3.7b). Distinguishing compound illocutionary acts and proposition-level types is not beneficial for NPS, probably because each utterance is short and usually contains only one type of proposition. As a consequence, the model has difficulty grouping different surface types of propositions into compound illocutionary acts; for CNET, on the other hand, proposition-level types that co-occur in the same utterance tend to cluster to the same state.

It is worth noting that some proposition-level types learned represent rather topical clusters. However, they do not have undue influence in our model.

Speaker preferences: Speaker preferences substantially increase the v-measure by 0.13 for NPS (CSM vs. CSM – Speaker). Notably, speaker preferences complement the mixture of proposition-level types, which is not good at clustering related proposition-level types into the same compound illocutionary act for short utterances. More specifically, when each speaker is modeled to have sparse preferences for illocutionary acts (i.e., states), proposition-level types

used by the same speaker, often representing the same illocutionary act, tend to cluster to the same state.

Speaker preferences also capture the characteristic styles of some speakers. Among speakers who are found to have sparse preferences by our model, some actively express reactions and often mark laughter (FT12). Others frequently agree (FT0), greet everyone (FT5), or have many questions (FT7, FT27). Accordingly, the model finds a relatively high optimal weight for speaker preferences in NPS ($\nu = 0.58$).

In contrast, CNET benefits little from speaker preferences ($\nu = 1$), partly because there is not enough information about each speaker in such short dialogues. Speakers also show little preference for illocutionary acts as their granularity is too fine in the corpus. For instance, while a thread initiator tends to ask questions in successive posts, these questions are annotated as different illocutionary acts (e.g., *Question-Question*, *Question-Add*, *Question-Confirmation*, etc.) depending on the position of the post within the thread.

Multi-level structure: Our model's ability to account for multi-level structure improves the accuracy of illocutionary act recognition for both corpora (CSM vs. CSM – Multi-level). For NPS, where multi-level structure is not explicit, this improvement comes from simple heuristics for inferring multi-level structure based on user mentions.

Limitations: A main limitation of the model is that a set of short and similar sentences can have too high influence, forming a proposition-level type that simply reflects these sentences. The reason is that, in our model, all words in the same sentence are assigned to the same proposition-level type. This assignment is based on the similarity of words in a sentence to other sentences in the same proposition-level type, and short sentences often find similar sentences more easily than long sentences do. Therefore, learned types tend to be characteristic of short sentences that are similar enough to form separate types. As a result, long sentences may be lumped to a small number of "garbage" types without reflecting their distinctive roles.

Another notable limitation is our assumption that one dialogue has one background domain. While this assumption holds quite well for CNET, it does not hold for NPS and perhaps other dialogues that do not discuss a cohesive topic. To rectify this issue, we could consider the new modeling assumption that even background domains change over time within a dialogue but more slowly than illocutionary acts. We leave this direction to future work.

3.6 Conclusion

We have presented an unsupervised model that learns the language models of different surface types of propositions underlying given dialogues. The assumption that a compound illocutionary act is a mixture of proposition-level types helped identify latent compound illocutionary acts better. The model separates out topical words to better characterize main types of propositions and also incorporates speaker preferences. We find that different characteristics and nature of dialogue require different modeling assumptions. Whereas the baseline models show a large

variance in performance across the evaluation corpora, our model is robust for both CNET and NPS corpora due to the model components complementing one another. Specifically, topical word filtering is found to be effective when each dialogue has a consistent conversational topic, and modeling a compound illocutionary act as a mixture of proposition-level types is beneficial for long utterances. Speaker preferences are found to be helpful when speakers have characteristic preferences for illocutionary acts. These findings, in addition to the fact that many common words are not filtered out as background, may help inform future model design.

Chapter 4

Analyzing Surface Types and Effects

Using the CSM model from Chapter 3, which learns various surface types of propositions (henceforth, surface types) underlying dialogue, in this chapter we apply it to four corpora of argumentative dialogue with two main goals:

- Identifying what surface types are common and consistently occurring across argumentative dialogue with different domains and goals.
- Analyzing how these surface types are associated with argumentation outcomes.

We first apply CSM to discussions among Wikipedia editors, political debates on Ravelry, persuasion dialogues on ChangeMyView, and the 2016 U.S. presidential debates among candidates and online commentary on Reddit. Based on the surface types learned from the four corpora, we identify 24 generic surface types in argumentation that occur consistently across the corpora. Next, we conduct four case studies using these corpora to examine how different surface types are associated with argumentation outcomes. We reveal that use of certain surface types has strong correlations with different argumentation outcomes.

4.1 Introduction

While various rhetorical strategies in argumentation have been studied intensively in rhetoric, marketing, and communication sciences, less has been studied about a comprehensive list of such strategies in a bottom-up fashion. We assume that surface types represent such strategies (e.g., using numbers and statistics, making comparisons) and identify surface types occurring across argumentative dialogue in an empirical and data-driven way. The results contribute to the literature of generic strategies in argumentation.

Our approach also allows for quantitative analysis of these strategies. We conduct four case studies that examine how these surface types are associated with argumentation outcomes. In the first study, we analyze five different roles of Wikipedia editors reflected in surface types they use often. The association between these roles and the success of editing is examined, revealing how the use of certain surface types correlates with successful editing in Wikipedia.

In the second study, we investigate surface types that are often perceived as "inappropriate" in political debates on Ravelry. We identify high-risk surface types (e.g., argument evaluation, expression of feelings) that are likely to lead the containing post to be moderated. Using these surface types as a lens, we also reveal that moderators in the forum have biases against minority opinions.

In the third study, we examine the association of surface types and the success of persuasion. We identify the effectiveness of surface types when used by persuadees and when used by persuaders (e.g., expression of confusion by the persuadee and using definitions by the persuader are positively correlated with successful persuasion). In addition, we further analyze the effectiveness of the interactions of surface types between the persuadee and persuader (e.g., the persuadee presenting statistics, followed by the persuader presenting a reference is positively correlated with successful persuasion).

In the fourth study, we analyze the association between surface types and the formation of pro- and counter-arguments. We look at the surface types of premises and show that some surface types have a strong tendency to be used for either pro-argument or counter-argument.

Before we go into details, we present the full list of surface types learned from the corpora in Table 4.1.

4.2 Data

We use four corpora of argumentative dialogue: Wikipedia discussions, online political discussions on Ravelry, ChangeMyView discussions, and the 2016 U.S. presidential debates and online commentary. These corpora cover different domains and have different goals (e.g., persuasion, winning votes, accomplishing a collaborative task, and sharing opinions). A brief summary of these corpora is as follows:

- **Wikipedia**: Discussions among Wikipedia editors on Wikipedia talk pages. The discussions focus on how to edit an article, where the goal is to make optimal edits on Wikipedia articles in a collaborative way.
- **Ravelry**: Argumentative discussions on the Ravelry Big Issues Debate forum. The discussions are mainly around political issues, where the goal is to discuss political opinions mostly for fun.
- **ChangeMyView**: Argumentative dialogues from the *ChangeMyView* subreddit. The dialogues cover a wide range of issues, where the goal is to change other users' viewpoints.
- **US2016**: 2016 U.S. presidential debates and online commentary on Reddit. The dialogues cover mostly political issues, where the goal is to sway votes.

4.2.1 Wikipedia Discussions

Wikipedia talk pages are explicitly designed to support coordination in editing of their associated article pages; they are not stand-alone discussion forums. We extracted all versions (*revisions*) of English Wikipedia articles from 2004 to 2014 and removed much of the Mediawiki markup

Surface Type	Description	Representative Forms	CMV	US2016	Ravelry	Wikipedia
Question	WH- or binary questions	"what/why/how is X ...?" / "Does X ...?"	✓	✓	✓	✓
Answer Elicitation	Eliciting an answer	"please answer X"		✓		
Agenda	Structure of the dialogue	"we discuss X and Y"		✓		✓
Meta-Argumentation	Reflection on the argumentation	"I answered your post"			✓	
Feeling	Feelings	Interjectives		✓	✓	
Thanks	Thanks	"thank you"	✓	✓	✓	
Number	Numbers, %, $, time	Number / "X%" / "$X"	✓	✓	✓	
Source	URLs and sources	"based on X" / "http://..."	✓	✓	✓	✓
Policy Reference	Referencing a policy	"according to WP:.."				✓
Source Validity	Validity of a source	"source is invalid/reliable"				✓
Comparison	Making a comparison	"X is ... than Y"	✓		✓	
Difference	Pointing out differences	"X is different from Y"	✓		✓	
Choice	Presenting choices	"whether X or Y"	✓			
Prediction	Predicting a future event	"X will do Y"		✓		
History	Past event	"X was/did Y"	✓		✓	
Normative	Normative statement	"X should/can Y"	✓	✓		
Disagreement	Expressing disagreement	"no" / "doesn't make sense" / "I don't think"	✓	✓	✓	
Confusion	Expressing confusion	"I'm not sure X"	✓			
Negated Expression	Negated expressions	"X is/does not Y"		✓		
Argument Evaluation	Evaluation on the listener's argument	"Your argument is X"	✓		✓	
Meaning	Definition or meaning	"X means Y"	✓		✓	✓
Quotes	Using quotation marks	"X is/does "Y"" / ""X" is/does Y"	✓		✓	
I	Statement about the speaker	"I am/do X" / "my X"	✓	✓	✓	
You	Directly addressing the listener	"you are/do X" / "your X"	✓		✓	

Table 4.1: Surface types learned by the models from the four corpora.

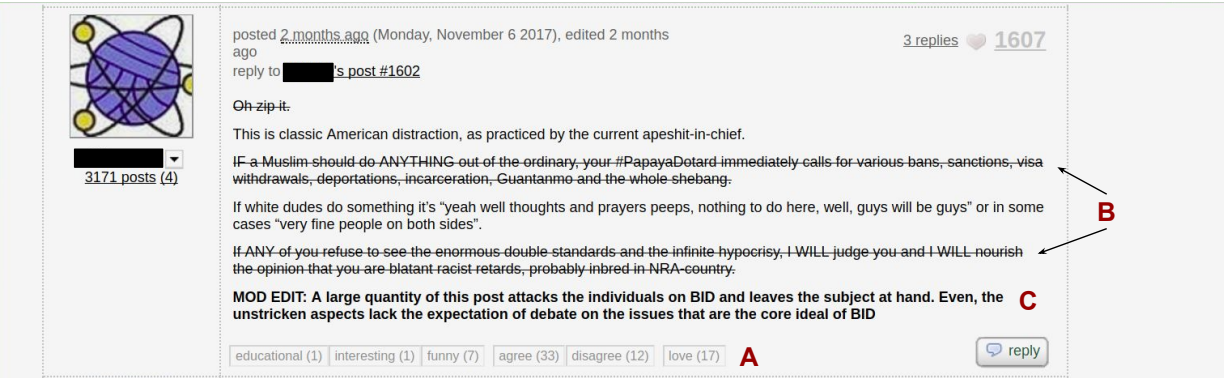


Figure 4.1: Example of a BID post that was also moderated. (A) shows the *tags* associated with the post. The text of the post that was crossed out (B) was not crossed out by the original poster but by the moderators after judging the text as a violation of the rules of BID. (C) gives the moderators' reasoning for how the post violates the rules of BID. Note that although the post was moderated, more users in the group *agree* with the post than *disagree*.

using the Java Wikipedia Library (JWPL) (Ferschke et al., 2011). The most recent revisions of talk pages corresponding to the articles were split into turns using paragraph boundaries and edit history. We grouped discussion posts under the same section headings as *discussion threads*. We sampled 100,000 articles with talk page discussions and filtered to include discussion threads with 2 or more participants who made edits to the article page from 24 hours before the discussion began to 24 hours after the discussion ended. Discussion thread beginnings and endings are defined as the time of the first post and last post, respectively. Statistics on our discussion dataset can be seen in Table 4.2.

Number of articles	7,211
Number of utterances	161,525
Number of discussion threads	21,108
Average #editors/discussion	2.52

Table 4.2: Statistics of the Wikipedia corpus.

4.2.2 Ravelry Big Issues Debate

Ravelry is a free social networking site for people interested in the fiber arts, such as knitting, crocheting, weaving, and spinning. With over 7.5 million users in December 2017[1], Ravelry is one of the largest active online communities that has been relatively understudied. While the broader Ravelry community is primarily focused on the fiber arts, social participation on Ravelry centers around tens of thousands of user-created and -moderated subcommunities, called *groups*. Groups act as discussion boards centered around a certain theme. Any user on Ravelry can create a group covering any variety of topics, special interests, or identities, which may or may not be related to

[1] https://www.ravelry.com/statistics/users

the fiber arts. For example, *Men Who Knit* provides a space for men, an underrepresented group in the fiber arts, while *Remrants* allows users to post rants about nearly any aspect of their lives.

Our study focuses on the Big Issues Debate group on Ravelry. Big Issues Debate, commonly referred to as BID, is described as a space "for everyone who likes to talk about big issues: religion, politics, gender, or anything that is bound to start a debate". Receiving over 3,500 posts a month, BID is the largest group dedicated to political and social issues and one of the most active groups overall on Ravelry[2].

Debates on BID begin with a user creating a thread and posting their view on an issue. Other users post responses to the original user's post or to other posts in the thread. An example BID post is given in Figure 4.1. Every post in the thread, including the original post, has a set of six associated *tags* (Figure 4.1, A) that users can interact with: *educational*, *interesting*, *funny*, *agree*, *disagree*, and *love*. Clicking on one of the tags allows a user to anonymously increase the value of a particular tag once per post, though these values do not affect the order in which posts are displayed.

There are three officially recognized and regulated formats of debate on BID: *Order* (default debate format), *Rigor* (stronger standards for sourcing/citations), and *BID* (discussion about policies and practices on BID). Thread creators can choose which format they want their debate to be in by tagging it in the thread title (e.g. "ORDER - Media Responsibility in Politics", "RIGOR: Bigotry and the 2016 US presidential race"). If not tagged, the thread is assumed to be in the Order format. In all of the recognized formats on BID, users are expected to follow these rules:

1. Abide by Ravelry's Community Guidelines and Terms of Service.
2. No personal attacks.
3. Behave civilly.
4. Debate the topic, not the person.
5. Do not bring in other groups, users not participating in the debate or baggage from one thread to another thread.
6. Don't derail the thread.

Within a discussion thread, users can flag another user as being in violation of one of the 6 main rules. Whether or not a post is flagged is only public to the moderation team, the user who made the flag, and the user who received the flag. Moderators then judge whether flagged posts are in violation of the BID rules. If the post is judged to be in violation of the rules, it is hereinafter referred to as *moderated*. In almost all cases, moderated posts are kept visible, but the offending part of the post is crossed out with a strikethrough (Figure 4.1, B). Moderators are also expected to give reasons for why a post was moderated (Figure 4.1, C), though they do not post their username. Users who repeatedly make offensive posts may have posting privileges suspended for a period of 24 hours or banned from the group for a longer period of time based on severity of the offense. Moderators may also delete posts, but this is only practiced in the Ask the Mods thread (where only specific types of posts are allowed) or in cases of "extreme spam"[3]. One key limitation on moderator privileges is that moderators cannot participate in debate threads they moderate, which

[2] https://www.ravelry.com/groups/search#sort=active

[3] https://www.ravelry.com/groups/big-issues-debate/pages/Information-on-Moderation-for-Members

prevents moderators from making explicit decisions against users they are debating.

Post data was scraped from the Big Issues Debate group on Ravelry from the beginning of the group in October 16, 2007 until June 6, 2017, including posts from threads that were publicly archived by the moderators and ignoring posts that were deleted. For each post, we collect its thread number, title, post number, author, date of creation, and the value of its tags on June 6, 2017. We also determined whether the post was moderated. We consider a post to be moderated if it contains the phrase "mod post", "mod edit", or "this post was moderated for", which all signal that a moderator has edited the post for inappropriate behavior. Moderators are expected to cross out the portions of text that were judged to have violated the BID rules, so in almost all cases we can recover the original text of the post that was moderated. We remove the very few "moderated" posts that do not have any portions that have been crossed out from our dataset, as we cannot ensure that these posts still contain the original behavior that they were moderated for. Some statistics of our final dataset from BID are shown in Table 4.3.

Dialogues	4,213
Utterances	350,376
# Utterances/dialogue	83.2 (average)
# Users	3,320

Table 4.3: Statistics of the Ravelry corpus.

4.2.3 ChangeMyView Discussions

Our next corpus is online discussions from the ChangeMyView (CMV) subreddit[4]. In this forum, users post their views on various issues and invite other users to challenge their views. If a comment changes the original poster (OP)'s view, the OP acknowledges it by replying to the comment with a Δ symbol. An example post and a comment that received a Δ is shown in Figure 4.2.

The goal of this subreddit is to facilitate civil discourse. The high quality of the discussions in this forum is maintained based on several moderation rules, such as the minimum length of an original post and the maximum response time of OPs. OPs who show a hostile attutude to new perspectives are also moderated. As a result, CMV discussions have been used in many NLP studies (Chakrabarty et al., 2019; Morio et al., 2019; Jo et al., 2018; Musi, 2017; Wei et al., 2016; Tan et al., 2016).

We scraped CMV posts and comments written between January 1, 2014 and September 30, 2019, using the Pushshift API. We split them into a dev set (Jan 2014–Jan 2018 for training and Feb 2018–Nov 2018 for validation) and a test set (Dec 2018–Sep 2019), based on the ratio of 6:2:2.

In order to see what issues are covered in this forum, we categorized the posts into domains using LDA. For each post, we chose as its domain the topic that has the highest proportion after

[4]https://www.reddit.com/r/changemyview

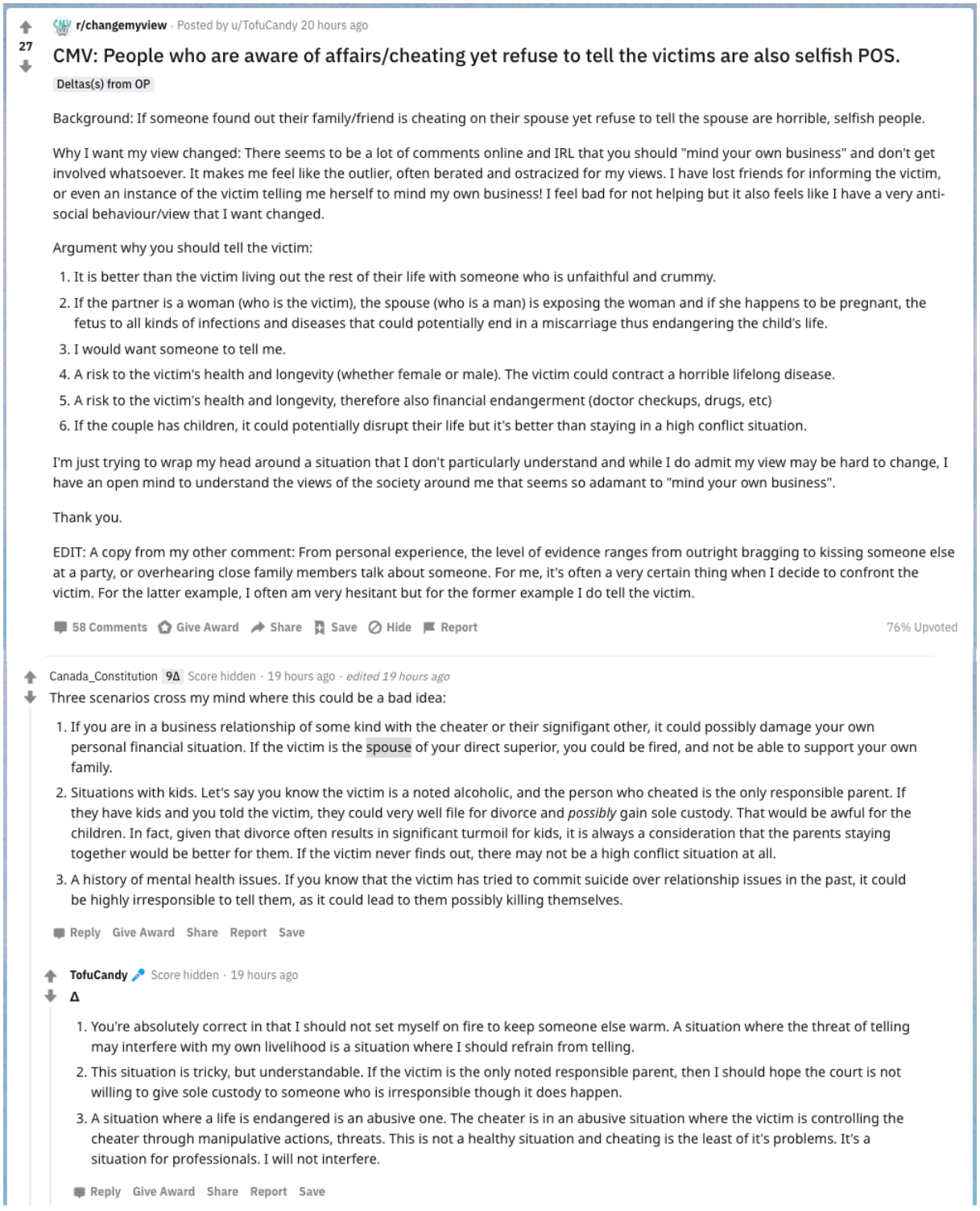


Figure 4.2: An example post and comments from ChangeMyView. User Canada_Constitution made a comment to the OP's post, and the OP (TofuCandy) gave this user a Δ, indicating the comment changed the OP's view.

Domain	%	Domain	%	Domain	%	Domain	%
media	5	race	4	tax	3	food	3
abortion	4	family	4	law	3	power	3
sex	4	life	4	money	3	school	3
election	4	crime	4	drug	3	college	3
reddit	4	relationship	3	war	3	music	2
human	4	movie	3	religion	3	gun	2
economy	4	world	3	job	3	israel	2
gender	4	game	3				

Table 4.4: Domain distribution in the CMV corpus. 10 LDA topics were excluded as they are irrelevant to discussion issues.

Threads	3,207 (Δ-awarded: 1,361)
Utterances (posts or comments)	204,679 (Δ-awarded: 2,748)
# Utterances/thread	63.8 (average) / 37.0 (median)
# Users	28,062

Table 4.5: Statistics of the CMV corpus.

standardization; topics comprising common words were excluded. We tried different numbers of topics (25, 30, 35, 40) and finalized on 40, as it achieves the lowest perplexity. This process resulted in 30 domains (after excluding 10 topics comprising common words): media, abortion, sex, election, Reddit, human economy, gender, race, family, life, crime, relationship, movie, world, game, tax, law, money, drug, war, religion, job, food, power, school, college, music, gun, and Jewish (Table 4.4).

Some statistics of the final corpus are shown in Table 4.5.

4.2.4 2016 U.S. Presidential Debates

Our next corpus is the transcripts of the 2016 U.S. presidential debates and online commentary on Reddit (Visser et al., 2019). The corpus contains debates for the primaries of the Republican (on August 6, 2015) and Democratic parties (on October 13, 2015), and for the general elections (September 26, 2016); the transcripts are from the American Presidency Project. Each debate is split into mini "dialogues" based on the topic. The corpus also contains online commentary on Reddit. This commentary is a compilation of sub-threads on Reddit that were (i) created during the debates, (ii) longer than four turns, (iii) directly relevant to the debates, and (iv) argumentative (rather than phatic) based on manual inspection. The dialogues in this corpus are annotated with asserted propositions and their relations in terms of *support* and *attack* (Figure 4.3). Each instance has been annotated by two annotators, achieving Cohen's κ of 61.0 across the corpora. We cleaned up the corpus, and some statistics of the resulting data are shown in Table 4.6.

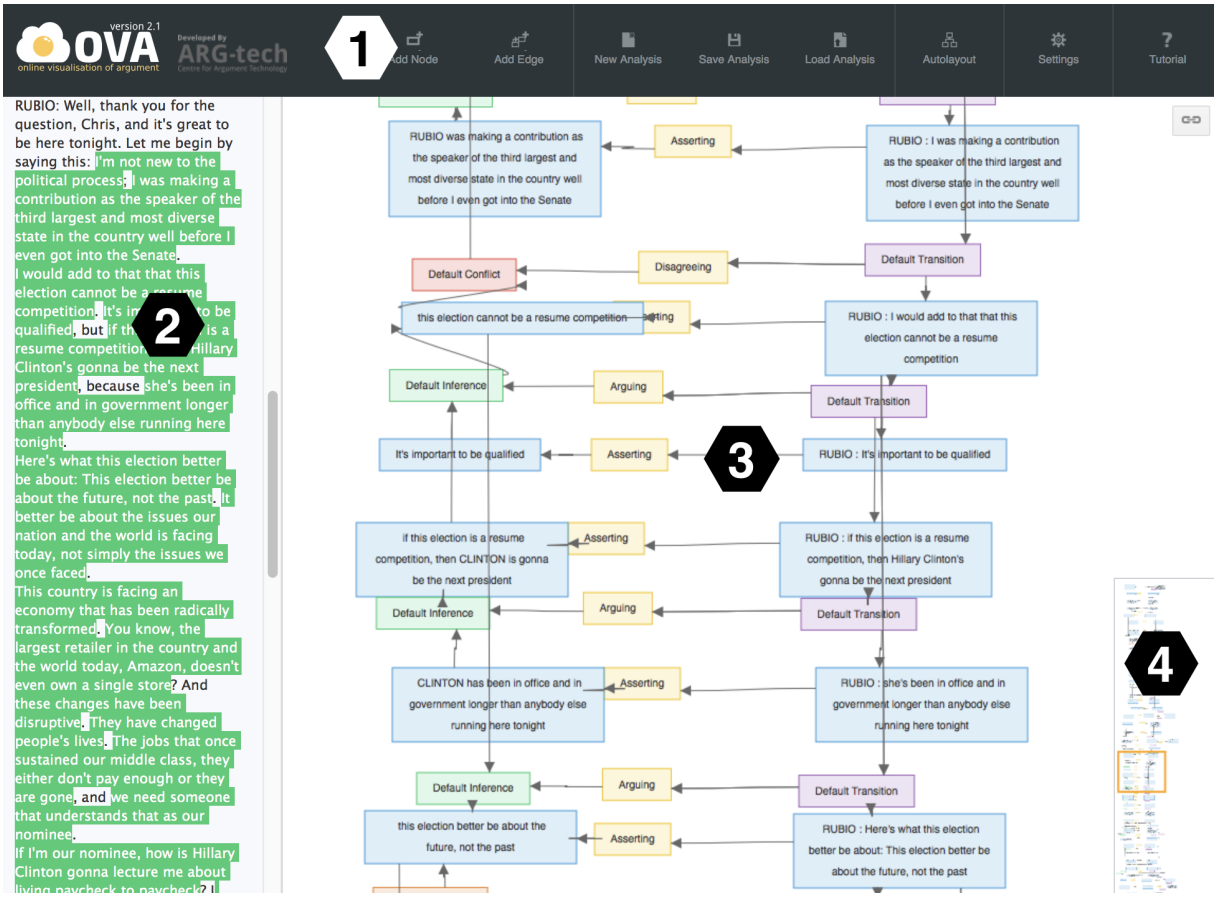


Figure 4.3: An example of the US2016 corpus (Fig. 3 in the original paper (Visser et al., 2019)). (1) is the annotation menu bar, (2) is the original dialogue text, (3) is annotations (on the left is asserted propositions and their relations—Default inference (support) and Default Conflict (attack); on the right is locutions in the dialogue), and (4) is a navigation tool.

Dialogues	369
Utterances	3,021
# Utterances/dialogue	8.2 (average) / 6.0 (median)
# Speakers	32 for debates (excluding Reddit users)

Table 4.6: Statistics of the US2016 corpus used in this chapter.

4.3 Surface Types in Argumentative Dialogue

In this section, we apply CSM on the four corpora to identify surface types in these corpora. We first describe the model settings and then present the results.

4.3.1 Model Settings

Model Variants: For the CMV and US2016 corpora, we use CSM as it was introduced in Chapter 3. For the Wikipedia and Ravelry corpora, we slightly modify CSM to better incorporate the properties of these corpora.

For the Wikipedia corpus, the discussions on talk pages are centered around the content of the actual article pages. In other words, the content of the article pages can be thought to serve as background topics on top of which the discussions are conducted. Hence, we run a separate LDA model (Blei et al., 2003)[5] to learn 100 topics from article sections. Then, for each discussion on talk pages, we assume that the probability distribution over background topics is the probability distribution over LDA topics of the corresponding article section. The modified model is illustrated in Figure 4.4a, and the generation process is modified as follows:

- For each word
  - Draw an indicator of "surface type" or "background topic" $l \sim \text{Cat}((\eta, 1-\eta))$.
  - Draw a background topic $z^B \sim \text{Cat}(\theta^B)$, where θ^B is a probability distribution over LDA topics for the article section.
  - If l is "surface type", draw a word $w \sim \text{Cat}(\phi^F_{z^F})$.
  - If l is "background topic", draw a word $w \sim \text{Cat}(\phi^B_{z^B})$.

For the Ravelry corpus, each post (or comment) has a label of whether it has been moderated or not. We assume that some surface types may contribute to moderation, and we hope that the learned surface types reflect this phenomenon. Hence, the model encodes that the probability of an utterance being moderated is computed by logistic regression where explanatory variables are the probabilities of surface types in the utterance. The modified model is illustrated in Figure 4.4b, and the following step is added to the generation process:

- For each utterance u with its surface types $\mathbf{z}^F$,
  - Draw an outcome measure $y \sim P(y=1|\mathbf{w}, \mathbf{z}^F) = \sigma(\mathbf{w}^\top \bar{\mathbf{z}}^F)$,

[5] We used the MALLET library from http://mallet.cs.umass.edu.

where $\bar{\mathbf{z}}^F$ is normalized frequencies of surface types in $\mathbf{z}^F$, $\mathbf{w}$ is a weight vector over surface types, and $\sigma(\cdot)$ is the sigmoid function. $\mathbf{w}$ is optimized in every iteration of the inference phase.

Speakers: For CMV, one thing we are interested in is how surface types are associated with persuasion. We assume that OPs, successful challengers, and unsuccessful challengers may have characteristic preferences for certain surface types. Hence, instead of treating individual users separately, we define three aggregate "speakers" for these three categories. A user is classified as the *OP* only for the threads where this user is the OP; for the other threads, the user is classified as the *Successful Challenger* if the user receives a Δ or as the *Unsuccessful Challenger* otherwise. Assuming the three hypothetical speakers helps to learn surface types that reflect potential differences across these three categories of users.

For the other corpora, individual speakers are treated as separate speakers. For the US2016 corpus, however, we introduce an aggregate user *Commentator* for all Reddit users, assuming that online commentators have characteristic preferences for certain surface types.

Input Unigrams: For CMV, US2016, and Ravelry, we obtain model input unigrams using the following preprocessing:

- Each token is converted to its lemma.
- Based on dependency parses, negated words are prefixed with "NOT_" and negating words (e.g., "not", "never") are removed.
- URLs are normalized to a special token.
- Numbers are normalized to a special token.

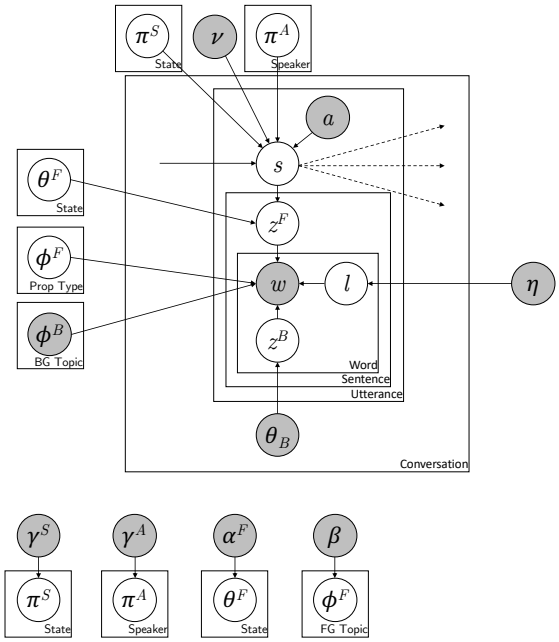
For the Wikipedia corpus, we skip this preprocessing and just use raw tokens from tokenization, because background topics are already given by LDA.

Other Parameters: Other model parameters are summarized in Table 4.7. We explored various values for the number of surface types, the number of background topics, the number of states, η, and ν. In order to choose the optimal parameter values, we first filtered top two configurations based on the likelihood of the data, and then chose the final one based on the interpretability of surface types. We did not conduct extensive parameter exploration for the Wikipedia corpus.

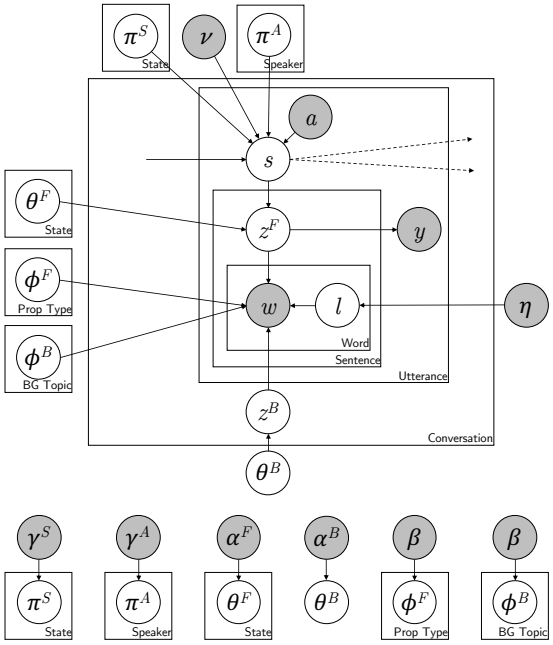
4.3.2 Result

Some surface types learned by the model are rather topical and domain-specific (e.g., money, jobs). Although these types may be meaningful for specific corpora and for specific tasks, our main interest is in surface types that are less topical and more generic to argumentation. Hence, we filtered such surface types through manual inspection and categorized them into 24 surface types shown in Table 4.1.

The first group of surface types is related to questioning and moderation. *Question* propositions generally include wh-adverbs and the question mark, and are used to express questions. This surface type roughly corresponds to underspecified propositions in the speech act theory (Searle,



(a) Variant of CSM for the Wikipedia corpus. Unlike the original CSM, the background topics and the probability distribution over background topics for each discussion are assumed to be observable.



(b) Variant of CSM for the Ravelry corpus. Unlike the original CSM, each utterance has an observable variable representing whether the utterance is moderated or not; its value depends on the distribution of surface types in the utterance.

Figure 4.4: Variants of CSM for the Wikipedia and Ravelry corpora.

	CMV	US2016	Wikipedia	Ravelry
Speakers	OP, Successful Challenger, Unsuccessful Challenger	Individual debaters, Commentator	Individual editors	Individual users
Model	CSM	CSM	CSM + background topics	CSM + moderation label
# Surface Types	20, 30, **40**	20, 30, **40**	**20**	20, 30, **40**
# Background Topics	40, 50, **60**	10, 20, **30**	**100**	40, 50, **60**
# States	**10**, 20	10, **20**	**5**	10, **20**
η	**0.25**, 0.5, 0.75	**0.25**, 0.5, 0.75	**0.75**	**0.25**, 0.5, 0.75
ν	0.25, **0.5**, 0.75	**0.25**, 0.5, 0.75	**0.75**	**0.25**, 0.5, 0.75
# Iterations	500	3,000	10,000	500

Table 4.7: Model settings for identifying surface types. Bolds are the values finally chosen.

1969); for example, the question "What is your name?" is viewed as having the underspecified proposition "Your name is (underspecified)", in which the underspecified part may be replaced with an wh-adverb in our case. *Answer Elicitation* propositions elicit questions from other people, and *Agenda* propositions describe the structure of the dialogue. *Meta-Argumentation* propositions usually reflect on the argumentation and are common only in the Ravelry corpus.

The second group of surface types is related to feelings. *Feeling* propositions express feelings and often consist of short interjectives (e.g., "lol"). While such interjectives may not be seen as propositional content technically, we interpret them as describing the speaker's feelings and thus include them as a surface type. *Thanks* propositions express gratefulness. They can be seen as a subset of *Feeling*, but we distinguish between the two types because the model usually does so, probably due to the characteristic and frequent uses of thanking.

The third group of surface types is related to data. *Number* propositions include specific numbers and measurements. *Source* propositions contain URLs and other references. Sometimes, they are composed of a simple URL; a URL may not be seen as a proposition technically, but we interpret it as indicating that the specified reference has necessary information. These two surface types are highly common across the corpora. *Policy Reference* propositions appear mostly in Wikipedia discussions, where editors refer to specific Wikipedia policies to determine if certain text in an article violates Wikipedia norms. *Source Validity* propositions mention the validity of a source.

The fourth group of surface types is related to juxtaposition. *Comparison* propositions compare multiple objects, and *Difference* propositions point out that there is a difference or a distinction should be made between multiple objects. These types may not be dominant in argumentative dialogue, but they occur in the two largest corpora. *Choice* propositions enumerate multiple choices.

The fifth group of surface types is related to tense and imperative mood. *Prediction* and *History* propositions each describe future and past events, respectively. *Normative* propositions express

imperative mood, indicating something is needed or should be carried out.

The sixth group of surface types is related to disagreement. *Disagreement* propositions directly express disagreement on a statement and are common across the corpora. *Confusion* propositions do not directly disagree, but indirectly express confusion. *Negated Expression* propositions use negation; they may not explicitly express disagreement or confusion. *Argument Evaluation* propositions are evaluation on an argument (usually the hearer's argument).

The seventh group of surface types is related to meaning. *Meaning* propositions point out the definition or meaning of a term or someone's speech. This type is common across many corpora. *Quotes* propositions use quotation marks. They are used for many purposes, such as emphasizing some words or expressions, and indicating commonly named terms or improperly named terms.

The eighth group of surface types is related to addressees. *I* propositions mention first-person singular nouns, often expressing the speaker's own stories or thoughts. On the other hand, *You* propositions mention second-person singular nouns, directly addressing the hearer.

In the following four sections, we study how these surface types are associated with various outcomes of argumentative dialogue.

4.4 Study 1. Surface Types and Edits in Wikipedia

In this study, we explore the argumentative strategies and configurations of conversational roles that allow Wikipedia editors to influence the content of articles. The goal is to examine how the surface types of propositions form different conversational roles of editors and how these types and roles are associated with the success of Wikipedia editors measured by an operationalization of the lasting impact of their edits in the article. In so doing, we propose a probabilistic graphical model that advances earlier work inducing latent conversational roles. This model allows the interpretation of configurations of roles that are conducive or detrimental to the success of individual editors; for instance, one of our findings is that the greatest success is achieved by detail-oriented editors working in cooperation with editors who play more abstract organizational roles.

Online production communities like Wikipedia, an online encyclopedia which anyone can edit, have the potential to bring disparate perspectives together in producing a valuable public resource. Individual Wikipedia editors unavoidably carry their own perspectives; these voices can explicitly or subtly influence the jointly produced article content even when editors strive for neutrality[6]. Wikipedia editors discuss article improvements, coordinate work and resolve disagreements on talk pages associated with each article (Ferschke, 2014). Pairing talk page discussions with simultaneous edits in shared content, we introduce a task predicting the success of a particular editor's article edits based on the corresponding discussion.

This study fits with research on editor behavior on Wikipedia, which is relatively well-studied on article pages and somewhat less studied on talk pages. Wikipedia has been a popular source of data

[6] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

for modeling social interaction and other issues of language behavior from multiple perspectives including collaboration (Ferschke et al., 2012), authority (Bender et al., 2011), influence (Bracewell et al., 2012; Swayamdipta and Rambow, 2012), and collegiality and adversity (Bracewell et al., 2012).

Much work analyzing behavior in Wikipedia has focused on types of edit behavior. Yang et al. (2016) use an LDA-based model to derive editor roles from edit behaviors. They then find correlations between certain editor roles and article quality improvements. Their approach differs from ours in that our model is supervised with an outcome measure and that we define editor roles based on talk page behavior. Viégas et al. (2007) categorize talk page contributions into 11 classes, and find that the most common function of talk page behavior is to discuss edits to the corresponding article, but that requests for information, references to Wikipedia policies, and off-topic remarks are also commonly found. Bender et al. (2011) annotate authority claims and agreement in Wikipedia talk pages.

Above the level of individual contributions to discussion, the notion of a conversational role is relevant both for characterizing the rights and responsibilities an individual has within an interaction as well as the configuration of conversational behaviors the person is likely to engage in. Therefore, it is not surprising that prior work has revealed that the process of becoming a Wikipedia moderator is associated both with changes in language use and in the roles editors play on the talk pages (Danescu-Niculescu-Mizil et al., 2012). In order to understand roles Wikipedia editors play, Arazy et al. (2017) find self-organizing roles based on the edit behavior of thousands of editors. Editors frequently move in and out of those roles, but on the aggregate the proportions of these roles are relatively stable.

Our work is similar to that of Ferschke et al. (2015), who apply the role identification model of Yang et al. (2015) to Wikipedia talk page contributions. This model learns a predefined number of user roles, each of which is represented as weights on a set of user behaviors, and assigns the roles to the participants in each discussion. The roles are induced by rewarding latent role representations with high utility in selecting users whose behavior was highly predictive of the task outcome of article quality. We extend this work by incorporating proposition-level surface types in defining editors' roles. We also predict an outcome that is specific to one discussion participant, i.e., the editing success of a particular editor within an interaction. Our model relaxes the strong constraint that every role must be assigned to a single participant and that each participant can take at most one role, making our model more flexible in capturing more nuanced configurations of roles.

Editor Success Scores

Since our goal is to see the relationship between conversational roles and an editor's success, we first quantify editor success scores in terms of how long edits would last as a result of argumentative discussions on talk pages. Our approach is similar to prior work (Priedhorsky et al., 2007). We define a success score y for each editor in a specific discussion. Intuitively, this measure is computed as the change in word frequency distribution associated with an editor's edits between the article revision prior to discussion and the article revision when the discussion

ends. In particular, this score is the proportion of an editor's edits—words deleted and words added—that remain 1 day after the discussion ends. Note that this score only reflects changes in word frequencies and does not take word re-ordering into account.

Formally, we consider each edit $\mathbf{e}$ as a vector of word frequency changes, both positive (additions) and negative (deletions) for each word type, stopwords removed. For an example in English, an edit that changed one instance of *suggested* to *insinuated*, as well as adding *old* might be represented as {'suggested': -1, 'insinuated': +1, 'old': +1'}. For each edit $\mathbf{e}_i$, let vector $\mathbf{c}_i$ be the changes in word frequencies from that edit to the final revision after the discussion. This change vector represents how many tokens that an editor deleted were put back and how many tokens the editor added were afterward deleted. Let $|\mathbf{e}|$ be the number of tokens changed in that edit and $|\mathbf{c}|$ be the total word frequency changes (deletions if tokens of the word were added in the edit, or vice versa) in those specific word types from the edit to the final revision. The score y of a particular Wikipedia editor u in thread t across edits $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ made by u in t is:

$$y(u,t) = 1 - \frac{\sum_{i=1}^{n} |\mathbf{c}_i|}{\sum_{i=1}^{n} |\mathbf{e}_i|}$$

Each editor's score is the proportion of tokens they changed that remain changed, so $s \in [0,1]$.

The goal of this editor score is to capture the "ground truth" of an editor's influence on the article page. To validate this editor success measure, we sampled 20 conversations, read through the corresponding article edits by those editors, and made sure our automated editor success scores were reasonable compared with the success that editors seemed to achieve.

In our experiments, we aim to predict this editor success measure calculated from article revisions with behaviors and interactions simultaneously occurring on the talk page. This assumes that talk page discussions in our data are related to the simultaneous article edits that those same editors are doing. To validate that editors who were editing the article while having a discussion on the talk page simultaneously were talking about those simultaneous article edits, and not something else, we manually went through 20 conversations and simultaneous edits. Nineteen out of the 20 conversations directly related to simultaneous edits, and the only one not specifically about simultaneous edits related to similar content on the article page.

4.4.1 Probabilistic Role Profiling Model

We propose a lightly supervised probabilistic graphical model of conversational roles that offers advances over the prior role modeling work of Yang et al. (2015), which employs a more restricted conceptualization of role taking. While the earlier model only allowed each role to be played by one editor, our extended model learns a distribution over roles for each editor. Furthermore, it can assign roles to an arbitrary number of editors rather than being restricted to a specific number. This model allows the interpretation of configurations of roles that are conducive or detrimental to the success of individual editors.

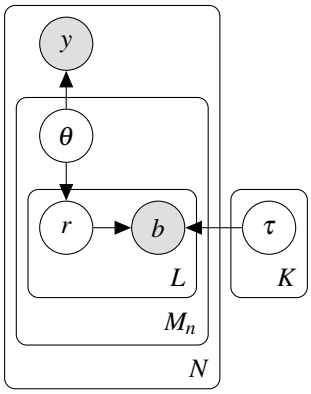


Figure 4.5: PRPM plate diagram relating for each conversation N the outcome measure y and each user M's L behaviors b.

Model Design

Our model attempts to learn both discussion behaviors of the target editor (editor we are predicting the success of) and roles of other discussion participants that influence the success of a particular editor. The task of role modeling as described is to identify latent patterns of behavior in discourse which explain some conversational outcome measure. The learned roles can then be intuitively interpreted to better understand the nature of the discourse and the interactions between the participants with respect to the chosen outcome measure.

For modeling roles in discourse, we propose a generative model shown in Figure 4.5, whose generative process is as follows:

- For each role $k \in \{1,\ldots,K\}$,
  - Draw behavior distribution $\tau_k \sim \text{Dir}(\alpha)$.
- For each conversation $n \in \{1,\ldots,N\}$,
  - For each user $m \in \{1,\ldots,M\}$,
    - Observe user participation z_{nm}.
  - For each user $m \in M_n$, where $M_n = \{m|z_{nm} = 1\}$,
    - Draw role distribution $\theta_{nm} \sim \text{Dir}(\gamma)$.
    - For each behavior $l \in \{1,\ldots,L\}$,
      - Draw role $r_{nml} \sim \text{Multi}(\theta_{nm})$.
      - Draw behavior $b_{nml} \sim \text{Multi}(\tau_{r_{nml}})$.
  - Draw outcome $y_n \sim \mathcal{N}(\mu_n, \sigma)$, where $\mu_n = \sum_m z_{nm}\theta_{nm} \cdot \beta$.

The values of the parameters η, β, and τ are inferred from data, and represent the settings with which the data is best explained (i.e., has the highest likelihood) under the generative process. We implement the model sampler using the JAGS framework (Plummer, 2003), which uses Gibbs sampling to generate dependent samples from the posterior distribution. These samples are used to obtain posterior mean estimates of the model parameters.

Surface Type	Description
Question	Questions. (“I wonder how difficult it would be to try to track down a replacement.”)
Moderation	Providing structure to discussion. (“The following discussion is an archived discussion of a requested move.”)
Sectioning	Sectioning, merging, and archiving of articles. (“Please consider providing additional references to such information and/or moving it to a new section within the article’s subject.”)
Image	Fair use of images. (“That there is a non-free use rationale on the image’s description page for the use in this article.”)
Policy	Mentions of Wikipedia policies. (“If there was ever a good reason to invoke WP:UNDUE this is it.”)
Terms	Spelling and use of terms. (“... it should probably be reworded to “compatible with North American NES cartridges” or something similar ...”)
Source	Mentions of content sources. (“And the source for Rock and Roll, could not be more solid and can not be excluded.”)

Table 4.8: Surface types learned from the Wikipedia corpus.

Features

Surface Type Features: We are interested in argumentative moves in terms of surface types that characterize editors in the argumentative dialogue on Wikipedia talk pages. Some of the surface types we learned are listed in Table 4.8. The surface types were found to yield better performance with our model than unigrams with tf-idf selection.

Position of the editor in a discussion:

- Number of editor turns
- Number of other editors’ turns
- Whether the editor takes the first turn
- Whether the editor takes the last turn

Style characteristics: Style characteristics may reflect the style and state of editors.

- Number of definite/indefinite articles
- Number of singular/plural personal pronouns
- Examples: number of occurrences of “for example”, “for instance”, and “e.g.”
- URLs: number of URLs that end with “.com”, “.net”, “.org”, or “.edu”
- Questions: number of question marks that follow an alphabetic character

Authority claims: Bender et al. (2011) define these authority claim categories annotate them in Wikipedia talk pages. For each word type in their annotated data, we calculated the pointwise mutual information for each category. In our data, we scored each sentence with the log sum of the word scores for each category. The categories used are:

- Credentials: education or occupation

- Experiential: personal involvement
- Forum: policy or community norms
- External: outside authority, such as a book
- Social expectations: expected behavior of groups

Emotion expressed by editors: For a simple measure of emotion, we use LIWC (Tausczik and Pennebaker, 2010).

- Counts of positive/negative emotion words

4.4.2 Experiment Settings

We frame our task as a regression problem, predicting editor scores based on discussion behaviors of the target editor and the other editors. Our outcome measure is the editor success score of a single editor. Since there are multiple editors in a discussion, we have multiple instances per discussion.

We use root mean squared error (RMSE) between the true scores and the predicted scores as an evaluation metric. We hypothesize that in specifying our model with latent roles as mediators between the raw discussion data and the predictive task we can achieve a lower RMSE than from a baseline that takes only the behaviors into account, especially for conversations with a greater number of participants, for which there can be more interaction. Furthermore, to the extent to which the proposed graphical model better captures a valid conceptualization of roles, we hypothesize that we can achieve a lower RMSE than the model of Yang et al. (2015). In this section we first specify the baselines used for comparison in our experiments, and then explain the testing process with our own model and experimental design.

Data processing

We pair discussions with the record of concurrent edits to the associated article page. Once a discussion has been paired with a sequence of edits, an assessment can be made for each editor who participated both in the discussion and in article edits of how successful that editor was in making changes to the article page. It is this assessment that forms the class value of our predictive task. In this study we explore negotiation strategies and role configurations that affect article editing; each data point in our task provides both discussion and an article edit success value for each editor involved.

The dataset comprises 53,175 editor-discussion pairs in which a "target" editor interacts with one or more other editors in a talk page discussion and achieves a measured influence on the associated article page[7].

Comparison Models

We want to test two hypotheses. The first hypothesis is that introducing a model with latent roles improves over simply using discussion features, and the second is that PRPM better captures

[7]This dataset is available at http://github.com/michaelmilleryoder/wikipedia-talk-scores

interaction than the prior RIM model. This goal leads to the following baseline models.

Linear Regression: The full set of features in this model are included twice, once from the target editor in the discussion, and once from an aggregation across all non-target editors in the discussion.

Role Identification Model (RIM): A similar task was explored by (Ferschke et al., 2015) and (Yang et al., 2015), who represented role modeling as a bipartite matching problem between participants and roles. More specifically, RIM learns conversational roles from discussion behaviors, supervised by discussion outcome. A role is defined as a weight vector over discussion behaviors, where the weights represent the positive or negative contribution of the behaviors toward outcome measures.

However, this approach suffers from several simplifying assumptions which reduce its applicability to realistic conversation settings:

1. All roles are present in every conversation.
2. Each role is played by exactly one editor.
3. Each editor plays exactly zero or one roles.
4. All behaviors from editors with a role contribute to the outcome metric under that role.
5. No behaviors from editors without a role contribute to the outcome metric.

Our model addresses these limitations by using a probabilistic graphical model that encodes a more appropriate hierarchical structure for the task.

We evaluate our model against RIM, introduced by Yang et al. (2015). RIM was originally applied to Wikipedia talk page discussions in Ferschke et al. (2015), who assigned a single success score to each page. In our work, for each discussion, we evaluate the success of each editor in each discussion thread separately. Since there is differential success between editors in the same interaction, the same interaction is associated with multiple different success measures. We handle this by slightly tweaking the original RIM model such that the first role is reserved exclusively for target editors, i.e., editors whose success measure is being evaluated. The other roles represent the roles of other editors in terms of their influence on the success of the target editor. Additionally, for conversations having fewer editors than the number of roles, we leave some of the roles unassigned by adding dummy editors whose behavior values are zero.

To predict the success measure of an editor for a test instance, RIM first assigns the learned roles to the editors. This process is identical to the training process, except that there is only the role assignment step without the weight adjustment step. Specifically, the first role is assigned to the target editor as in training, and the other roles are assigned according to the original model. Once the roles are assigned, the predicted score is simply the sum over roles of the inner product of a role's weight vector and the behavior vector of the editor who is assigned the role.

PRPM: For our model, to infer role distributions for each editor in a test instance conversation, we first fix the model parameters to the estimates learned during the training phase. Gibbs sampling is then used to infer the non-target users' role distributions θ_m and the conversation

Model	Setting	2	3	4	5+	All
LinReg	tgt editor	**0.286**	**0.302**	**0.287**	0.302	0.292
LinReg	all	0.287	**0.302**	0.289	0.301	0.292
RIM	K=2	0.316	0.317	0.308	0.342	0.318
RIM	K=3	0.307	0.320	0.310	0.337	0.314
RIM	K=4	0.307	0.314	0.311	0.327	0.311
RIM	K=5	0.309	0.315	0.308	0.321	0.312
PRPM	K=2	**0.286**	**0.302**	0.288	0.297	0.292
PRPM	K=3	**0.286**	**0.302**	0.288	**0.295**	**0.291**
PRPM	K=4	**0.286**	**0.302**	0.289	**0.295**	**0.291**
PRPM	K=5	**0.286**	**0.302**	0.288	**0.295**	**0.291**

Table 4.9: RMSE for baselines and models. Rows are model settings. Scores are reported for different numbers of participants, which are the columns headings. (LinReg: editor uses only the target editor's features, and all uses all participants' features. RIM and PRPM: K is the number of roles.)

outcome measure y over the unseen data. The role distributions for each non-target editor are then averaged together and concatenated with the target editor role distribution. Finally, a linear regressor is used analogously to the above baseline to evaluate the predictive power of the PRPM roles in aggregating the information from editor behavior features.

Parameters

In order to evaluate our approach and model, we split our data into a training set of 60%, a development set of 20% to train regression weights on the roles learned from the training set, and a test set of 20%. For the original and proposed role identification models, we manipulated the number of latent roles the learned models were allowed to include.

4.4.3 Results

Results from baselines and PRPM are presented in Table 4.9. We do not include scores with unigram tf-idf counts as features, as this decreases the performance of all models. The pattern of results is consistent with the hypotheses, i.e., role information and our model's configuration improves performance over both baselines.

First, the relatively high RMSE values indicate the challenging nature of this task. Talk page discussion is only one factor in editor success, and undoubtedly much interaction between editors comes from edit behavior, past interactions between editors, and even the short edit comments that editors leave about their edits. We were not able to find a comprehensive study of the effect of Wikipedia talk pages on article pages, but links from discussion features to outcomes in collaborative editing are often tenuous (Wen et al., 2016).

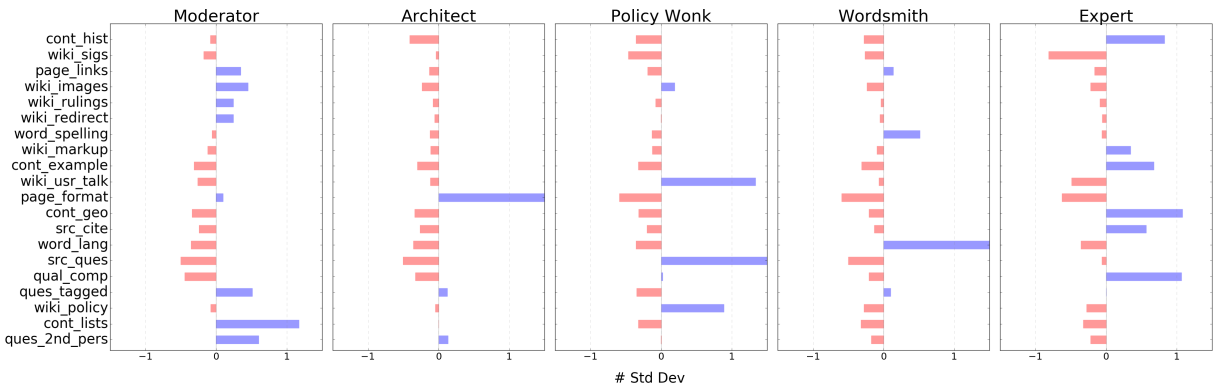


Figure 4.6: Behavior distributions for each role, expressed for each behavior as the number of standard deviations above the mean.

Our model performs slightly better than the linear regression baseline, though it performs substantially better than the previously proposed RIM model. One advantage of our role-based model above the linear regression baseline is clear when looking at conversations with more editors (columns in Table 4.9 denote the number of discussion participants in analyzed conversations). This points to the utility of using role information with larger groups, when roles are likely more relevant.

Another advantage of PRPM over the linear regression baseline is that it allows interpretation of both target editor strategies and group dynamics that characterize the success or failure of a target editor. Where linear regression allows only the characterization of behaviors that make individual editors successful, PRPM captures roles in interaction with other roles in group conversation. In this way, PRPM allows a more full interpretation of group interaction.

PRPM Role Analysis

Our best-performing model classified editors into 5 different roles. We identified the combinations of roles that are predictive of editor success (or failure). To assess roles, we examined the text and discussion features of editors who scored highly, as well as considered the weights assigned to each feature for each role. The relative frequencies of each behavior for each role are shown in Figure 4.6. A characteristic example discussion post for each role is given in Table 4.10. Each role is named and described qualitatively below.

Moderator. This role primarily helps discussion flow without getting too involved, performing and summarizing the results of administrative tasks. High probability surface types for this role include asking questions of other editors and discussing itemized content. The moderator role is less likely than other roles to have success as a target editor and has the lowest target editor success when paired with other editors playing the moderator role.

Architect: This role is predominantly focused on page hierarchy, with the bulk of its probability focused on surface types related to formatting, which are relevant to discussions of adding new page sections, merging, archiving, and creating new pages. The architect role is moderately likely

Role	Example post
Moderator	It was requested that this article be renamed but there was no consensus for it be moved.
Architect	I think a section in the article should be added about this.
Policy Wonk	The article needs more WP:RELIABLE sources.
Wordsmith	The name of the article should be ““Province of Toronto"" because that is the topic of the article.
Expert	There actually was no serious Entnazifizierung in East Germany.

Table 4.10: Examples of discussion posts from users in certain learned roles

to have success as a target editor.

Policy Wonk: This role is an knowledgeable Wikipedia user, frequently mentioning source accountability, fair use or copyright policy for images. Surface types that have high probability for the policy wonk include appealing to Wikipedia policy and discussing engagement with other users on user talk pages. The policy wonk role is moderately unlikely to have success as a target editor.

Wordsmith: This role is predominantly concerned with the naming, creation, and wording of pages. Surface types that have high probability for the wordsmith include discussing the spelling, pronunciation, or translation of words and phrases, as well as discussing the (re-)naming of new or existing pages or sections. The wordsmith role is strongly correlated with target editor success, especially when combined with the moderator or architect.

Expert: This role is the most content-oriented role learned by our model. Surface types that have high probability for the expert include making comparisons, discussing historical and geopolitical content, giving examples, and citing sources. The expert role is most strongly correlated with target editor success when combined with other users playing the expert role.

We find that the roles that lend themselves most strongly to target editor success (the Wordsmith and Expert) are more concrete edit-focused roles, while the roles associated with lower target editor success (the Moderator, Architect, and Policy Wonk) are more conceptual organizational roles. Note that it is not necessarily the case that editors that edit more frequently have higher scores. We find frequent editors across all roles.

Additionally, we find that configurations with multiple conceptual organizational roles lead to diminished outcomes for individual editors, suggesting that individual conceptual editors are unlikely to have their edits universally accepted. This could mean that talk page conversations that have multiple conceptual voices (which could be a measure of interesting discussion) are more likely to result in compromises or failure for a target editor. It is important to recognize that we are focusing on strategies and configurations of roles always in relation to the success of one editor; this editor score does not necessarily refer to a good, well-rounded discussion.

Conclusion

The nature of collaboration on Wikipedia is still not fully understood, and we present a computational approach that models roles of talk page users with relation to success on article pages. The proposed probabilistic graphical role model is unique in its structure of roles in relation to the outcome of one particular participant instead of group performance, and allows flexible mappings between roles and participants, assigning each participant a distribution over roles. The model we present retains one limitation of the RIM model, the assumption that editors in one conversation exist independently from those same editors in other conversations. Future work should address this.

Our model lends interpretability to combinations of talk page discussion roles. We find that detail-oriented roles are associated with success in combination with organizational roles, but that multiple participants taking organizational roles can lessen individual editing success.

Acknowledgement: *This study was led mainly by Keith Maki. Yohan Jo's contributions include: discussing the design of the role profiling model and extracting surface types and features (§4.4.1), running baseline models (§4.4.2), and interpreting results (Section §4.4.3).*

4.5 Study 2. Surface Types and Censorship in Debates

Moderators are believed to play a crucial role in ensuring the quality of discussion in online political debate forums, but the line between moderation and illegitimate censorship is often contentious and causes some participants to feel being treated unfairly. Hence, the main aim of this study is to examine if the perception of moderation bias is grounded. Specifically, we model users' actual posting behavior using surface types, and analyze if some users are indeed moderated unfairly based on irrelevant factors, such as their minority viewpoints and moderation history.

Online discussion forums create space for communities with similar interests to share thoughts and debate issues. However, the technological facilitation of conversation on these forums does not ensure that high-quality deliberation takes place. Discussion forums are vulnerable to problems such as trolling, flaming, and other types of nonconstructive content (Pfaffenberger, 2003). Furthermore, when the topic is controversial, such as religion or politics, discussions can become toxic or inflammatory. Perceived anonymity in many online forums often exacerbates this problem by weakening self-censorship, as people are less likely to regulate their own behavior if they believe that it is difficult to trace back what they say (Chadwick, 2006; Davis, 1999).

To address these issues, online political discussion forums often rely on moderators to enforce rules and boundaries for how users behave and what they can say. However, the line between legitimate forms of regulation, which are used to discourage behavior defined as inappropriate, and *illegitimate censorship*, where particular individuals, opinions, or forms of communication are unfairly suppressed, is often difficult to define (Wright, 2006). Censorship is usually defined subjectively, and in cases where there is room for interpretation, the unconscious biases of regulators may affect their judgments. On the other hand, a user's own bias may lead them to perceive unfair treatment where there is none.

In this paper, we contribute new insight into the differences between perceived and actual bias in an online community's attempt to facilitate productive exchange on controversial issues. Fair moderation without illegitimate censorship is fundamental for creating safe, engaging online spaces for deliberation on controversial topics (Carter, 1998). Research in this area not only can improve the quality of discussion in online political forums but also can allow insight into the process of developing norms of behavior and effective moderation in online communities. Regardless of whether censorship actually takes place, the perception of illegitimate censorship itself can create an atmosphere where users feel unfairly treated and trust in the forum is undermined (Wright, 2006). Thus, it is important to understand the sources of perceived censorship and recognize when and how perceived censorship is actually manifested.

Guided by these issues, we explore the following research questions:

(1) Do moderators unfairly target users with specific viewpoints? If so, to what degree?
(2) What are possible sources of bias that could lead moderators to censor unfairly?
(3) What are possible causes for users' perceptions of moderator bias?

To address these questions, we examined the perception of moderation bias against users with unpopular viewpoints and moderation history in the Big Issues Debate forum on Ravelry. Based on the surface types learned from these discussions, we identified high-risk behaviors associated with rule-breaking, then examined the effect of viewpoint and moderation history on the likelihood of moderation, controlling for high-risk behavior. This allows us to investigate whether users with minority viewpoints and moderation history are being unfairly moderated, given the behaviors they exhibit. We find evidence to suggest that certain surface types used by users highly likely induce moderation. But independently of these propositions, moderators still make decisions biased against individuals with unpopular viewpoints and moderation history. We argue that the perception of bias within the group is an issue by itself, as the perception of illegitimate censorship can lead to tension between the moderators and users within a community.

Moderation Issues in Political Discussion

Moderators play an important role in many online forums by helping to maintain order and facilitate discussion within their community (Kittur et al., 2009; Lindsay et al., 2009). While conventional wisdom suggests that moderators positively influence the quality of discussion in forums (Hron and Friedrich, 2003), the role of a moderator is often diverse (Maloney-Krichmar and Preece, 2005), unclear (Wright, 2006), or emergent (Huh, 2015) across different communities. Thus, it is important to consider how moderators operate within the context of the community that they are trying to maintain. In online political forums, moderators are considered critical in ensuring quality discussions by creating and enforcing regulations for proper behavior (Edwards, 2002), as useful debates require that participants maintain order, respect, and civility towards each other (Carter, 1998; Wilhelm, 2000).

However, when these political discussions are facilitated by interested groups, moderation can quickly be labeled as censorship. These claims are common on online political forums administered by national governments, a focus of research on the potential for new forms of deliberative democracy (Wright and Street, 2007; Khatib et al., 2012). Wright (2006) reviews the

process for moderation in two of the UK government's online political discussion forums. They find that moderation must be done carefully to avoid the "shadow of control", the perception that some entity of power can control what is said (Edwards, 2002). Ideally, rules for censorship must be detailed, openly available, and enforced by an independent party (Wright, 2006). Moderation should also be done in a way that explicitly facilitates the goals of the forum.

In non-governmental political discussion forums, the concept of a "shadow of control" is less obvious, as these forums are not explicitly run by a centralized entity with particular goals. Nevertheless, unconscious cognitive biases may arise from the structural organization of political discussion forums and from cognitive tendencies. Bazerman et al. (2002), in their investigation into why accountants make biased decisions, noted that ambiguity in interpreting information gave accountants the room to make self-serving decisions. In the context of political discussions, ambiguity in the rules for how to engage appropriately in a debate may allow moderators to make unfair decisions against particularly troublesome users or viewpoints they disagree with. Another (more surprising) condition that often promotes unconscious cognitive biases is the belief in one's personal impartiality (Kaatz et al., 2014). While moderators are expected to act impartially, as they are often removed from debate, they may unconsciously make more biased decisions because they are primed to believe that they are genuinely impartial, instead of recognizing these biases.

Issues with Moderation

Ravelry's Big Issues Debate (BID) group provides an interesting setting for studying perceptions of censorship in political discussions not only because it is an active debate group with formal moderation but also because of its controversial reputation. BID's formal moderation is crucial in creating a space where users with different viewpoints can discuss political and social issues, compared to other Ravelry political discussion groups with less formal moderation, which tend to be more homogeneous. However, BID is infamous in the broader Ravelry community for tension between users and its moderation team, providing an ideal setting for studying frustrations about moderation from perceived bias. Meta-discussion threads also provide insight into user opinions and perceptions about the organization of the group. As an example of frustration with the perceived censorship on BID, one conservative-leaning user comments

> "Never have I seen bold faced disregard for opinion. Am I surprised? Not with the group we have as mods ... A sorrier bunch of biased, preachy people with unlimited authority seldom seen ... we don't have a freaking chance of having any of our issues addressed. When we're outnumbered 50 to 1 (at the very least)- seriously????"

expressing their perception that moderators are biased against conservative users, who are in the minority on BID. A liberal-leaning user, on the other hand, commented

> "The one thing we can say with some certainty is that a lot of conservative voices have come forward saying they're not being treated fairly. I don't think that's true, but then I wouldn't, would I?"

questioning whether the perception that conservative users in BID are actually unfairly treated.

Some users argue another view on how moderation in BID is biased, where moderators may be biased against certain individuals based on their past behavior:

> "I think there are people who draw a moderation when others wouldn't. I don't think it has anything to do with political leanings. It's embarrassingly apparent at times."

> "It's not unusual for people in BID who have been modded to double down, rationalize their actions, cast blame on someone else, or toss a word salad to "explain" why they shouldn't have been modded. The mods' reaction to their being modded is just par for the course for BID."

Users who have been moderated in the past or users who have complained about moderation in the past, for example, may be given less leeway for offenses than someone who has never been moderated, as it is in the moderators' interests to quickly shut down dissent from high-risk individuals.

The widespread idea that the moderators are biased against certain viewpoints or individuals raises the question of what forms these perceived biases take. We find that users on BID primarily consider "censorship" to be a problem of false negatives in moderation. Most users that have been moderated accept that their behavior is inappropriate under the rules of BID. However, users also argue that if their behavior is considered inappropriate, then many similar posts that have escaped moderation should be moderated as well:

> "However none of those were struck through / given a "mod edit". This was only done to XXXX. Yep. Modding isn't biased at all"

> "If my posts were deleted why not XXX's?."

> "I also see certain liberals constantly get away with rule breaking. I don't quite understand why. But they do."

> "I was also modded for not furthering the discussion. I wonder how many other posts don't further the discussion?"

Thus, the primary issue of perceived bias appears to be derived not from direct suppression of a user or viewpoint but from uneven standards in how the rules are applied.

Contrasting Views of Bias

Based on our examination of the organizational structure of BID, we hypothesize that there is opportunity for moderator bias in deciding whether to moderate a post. The guidelines of BID are as follows:

1. Abide by Ravelry's Community Guidelines and Terms of Service.
2. No personal attacks.
3. Behave civilly.
4. Debate the topic, not the person.

5. Do not bring in other groups, users not participating in the debate or baggage from one thread to another thread.
6. Don't derail the thread.

These guidelines are ambiguous, using vague statements such as "Behave civilly" and "Debate the topic", which leaves room for interpretation at the discretion of the moderators. This ambiguity may allow moderators to make self-serving judgments in favor of users who they agree with. Thus, one hypothesis is that moderators could be biased against certain viewpoints. On the other hand, this same ambiguity in the rules could allow users to make the self-serving interpretation that moderators are unfair against them or their viewpoints. This supports the hypothesis that there is little to no actual moderator bias, only a user's strong perception of bias. The goal of our analysis is to test these hypotheses through a series of statistical modeling experiments.

4.5.1 Experiment Settings

To assess whether the moderation team is actually making biased decisions based on the viewpoints of users or moderation history, we present an approach for evaluating moderator decisions alongside users' actual behavior in posts considered for moderation. In order to determine whether or not user viewpoint plays a role in moderation decisions, we need to characterize viewpoints on BID. We also need to identify the behaviors that may put a user at risk of being moderated, as certain types of users may contribute offensive content more often. If users of a certain group more often behave inappropriately, they may be deserving of more moderation. After operationalizing these relevant variables of viewpoint and behavior, we include them in a binary logistic regression model with odds ratios (OR) to predict whether a given post is moderated. This model allows interpretation of the factors that may increase the likelihood that a post would be moderated; odds ratios allows us to estimate the effect of a variable on the probability that the post is moderated.

Model Specification

Our model is designed to measure the effect of user viewpoint (*minority*), moderation history (*mod_prev*), and actual posting behavior operationalized with surface types (*high_risk*) on the likelihood of being moderated (we explain how to measure these variables below). We also define pairwise interaction terms among the three main effect variables as an input to the regression to tease apart the relationships between the main effect variables in conjunction with each other. The final set of variables that we use as input to the regression are:

Dependent variable:

- *moderated*: A binary variable indicating whether the given post was moderated or not.

Independent variables:

- *mod_prev*: The number of times the user has been moderated in the previous 30 days. We normalize this variable to have a mean of 0 and standard deviation of 1 across all posts in our dataset for rescaling purposes.

- *minority*: A binary variable indicating whether the user who made the post is a minority-view user in BID (see "Assigning Viewpoint" section).
- *high_risk*: A continuous variable indicating whether a post has an unusually large amount of high-risk behaviors (see "Characterizing Behavior in BID Posts" section).
- *high_risk* $\times$ *mod_prev*
- *high_risk* $\times$ *minority*
- *mod_prev* $\times$ *minority*

Assigning Viewpoint

Assigning viewpoints to posts: In order to determine whether users who hold unpopular views are moderated more, we need to label users with whether or not they tend to hold the same view as the majority of the group. To determine whether a user holds majority or minority views, we use the *agree* and *disagree* tags on the posts they have made. The agree and disagree tags on a user's post provide an indication of how closely the post aligns with the views of the general user-base on BID.

The general perception on BID is that right-leaning, conservative users and viewpoints are in the minority while left-leaning, liberal users and viewpoints make up the majority. To verify that the agree and disagree tags align with this liberal-conservative conception of majority-minority on BID, we sampled 20 posts with higher agree than disagree tag values and 20 posts with higher disagree than agree tag values. Posts were sampled across threads to determine the general trend of views on BID on a variety of issues. We then presented the posts, along with the title of the relevant thread and the preceding post in the reply structure as context, to two native English speakers with moderate political knowledge and asked them to separately determine whether the opinion expressed in a post leaned more towards a liberal viewpoint or a conservative viewpoint. We define *liberal* viewpoints as those that favor social progressivism and government action for equal opportunity and *conservative* viewpoints as those that favor limited government, personal responsibility, and traditional values.

We then treat the agree/disagree tags on the sampled posts as another annotator who rates a post as liberal if the post has a higher agree than disagree tag value and conservative otherwise. Comparing this "agree/disagree" annotator with our human judges, we obtain a Fleiss' kappa of 0.916. This indicates high agreement among the human annotators' judgment of liberal and conservative and the agree/disagree tags associated with the post. Thus, we can aggregate the values of the agree and disagree tags of a particular user across BID to get an overview of their political viewpoint.

Assigning viewpoints to users: To label the viewpoint of a particular user, we first find every thread they have participated in on BID. For each thread, we sum the agree tag values for each post the user made in that thread. We repeat the same process for the disagree tag values in the same thread. As threads on BID are intended to be centered around a particular issue of debate (e.g. gun control, immigration, tax reform), the summed agree and disagree tag values should indicate how much the other users on BID agree or disagree with the user on that particular issue. If the total disagree tag value is greater than the total agree tag value for a user on a particular

thread, we label that user as having the minority viewpoint on the issue discussed in the thread. This thread-level notion of viewpoint is analogous to the *issue-oriented viewpoint* described in the literature (Kelly et al., 2005).

However, simply holding a minority view on one thread does not indicate that a user holds the minority viewpoint across BID; users may have particular issues where their viewpoints do not align with the ideological group closest to their general beliefs (e.g. a primarily liberal user who is pro-life). Thus, in order to get a general viewpoint for each user, we compare the number of threads where they hold the majority viewpoint with the number of threads where they hold the minority viewpoint. If the number of threads where they hold the minority viewpoint is greater, we label that user as a *minority-view user*. This notion of viewpoint is analogous to the *ideological viewpoints* described in the literature (Kelly et al., 2005), which are coherent systems of positions across issues. We focus on ideological viewpoints in our analyses because users participate across threads and recognizably carry their ideological positions with them. This is apparent in BID meta-discussion threads where users will refer to each other with ideological labels (e.g., "conservative", "liberal"). Thus, we predict that moderator impressions of users are based on their activity beyond the level of single-issue threads.

Identifying High-Risk Behaviors In the section "Issues with Moderation", we presented evidence that the primary sources of the perception of bias in BID are false negative judgments. Thus, in our analyses, we want to control for the case where users make high-risk, potentially offensive acts in their posts.

In order to identify the types of behavior that are associated with getting moderated, we choose to focus on surface types within posts. While previous work has characterized offensive behavior using lists of curated terms associated with hate speech or profanity (Chandrasekharan et al., 2017; Hine et al., 2017), we found that this method is unsuited for identifying the types of behavior associated with moderation. First, lists of unacceptable words or phrases will not fully capture more subtle, implicit ways of attacking or offending other users, such as sarcasm or passive aggressive statements. Second, the use of offensive terms is acceptable behavior on BID in certain contexts. Profanity is generally accepted (e.g., "We do not mod for profanity, no matter what people have tried to flag for.", "I have no issues whatsoever with profanity and often sprinkle my posts with it just for my own amusement."), while hateful terms are often quoted or referenced in debates about language use (e.g., "I nearly blew a gasket when my stepmother referred to Obama as 'that nigger in the White House"', "Do you think homosexual people are bullying others when they speak up about people using 'gay' and 'faggot' as insults?").

The learned surface types are summarized in Table 4.11. The model we used internally runs a logistic regression and calculates the weights of the surface types to predict the probability of the post being moderated. As shown in Figure 4.7, surface types that are positively associated with moderation include meta-argumentation (*MetaArg*), argument evaluation (*ArgEval*), asking questions related to the hearer (*YouQuest*), directly addressing the hearer (*You*), using racial terms (*Race*), feelings (*Feeling*), edits (*Edit*), talking about debates and topics (*Debate*), and references (*Reference*).

Surface Type	Description
Disagree	Expressing disagreement ("I don't think there's any harm in voting a certain way because you have an opinion.")
I	Personal stories. ("I can't get past that initial burst of "Oh HELL no" to get to the second half of the question.")
Quotes	Quotation marks. ("Would it make you feel better if I used the word "peaceful" instead of "correct"?")
Number	Numbers. ("I said further up that I pay 19.6% of my income to some type of federal tax.")
YouQuest (10)	Questions, especially directly addressing the hearer. ("If it is not your own, how can you expect to ever have any governance over who touches it and how?")
Feeling (11)	Feelings. ("Just wow", "What a stupid question.")
Reference	References and validity. ("Your sources don't address problems in states like Washington and Oregon.")
Difference	Pointing out differences. ("There needs to be a distinction between car maintenance finances needed and car payments needing to be made.")
Comparison	Making a comparison. ("Beans are a much less expensive source of protein than meat is.")
Edit	Marking edits. ("ETA: Actually by that definition, what happened in USSR was not "feminism" by any means.")
Race	Race-related terms. ("White is considered default.")
ArgEval	Evaluation on an argument. ("Can't understand it because everything you've said is your opinion.")
Money/Time	Numbers, especially about money and time. ("If 3 weeks in a class and one shift in a clinical setting is all it takes to be qualified, I wouldn't be surprised if these folks 'acted to their best ability'!")
You	Directly addressing the hearer. ("If you'd been made to wear school uniform, you would have thought of that.")
History	Past events. ("That's over decades ago now.")
Debate	Debates and topics. ("Are these topics truly debatable?", "When you walk into a debate and use personal experiences, it makes the debate almost impossible to continue.")
MetaArg	Reflection on the argumentation. ("The reason I had responded was the quote and reply.", "It isn't discussion or debating people are after when they badger someone who has answered.")

Table 4.11: Surface types learned from the Ravelry corpus.

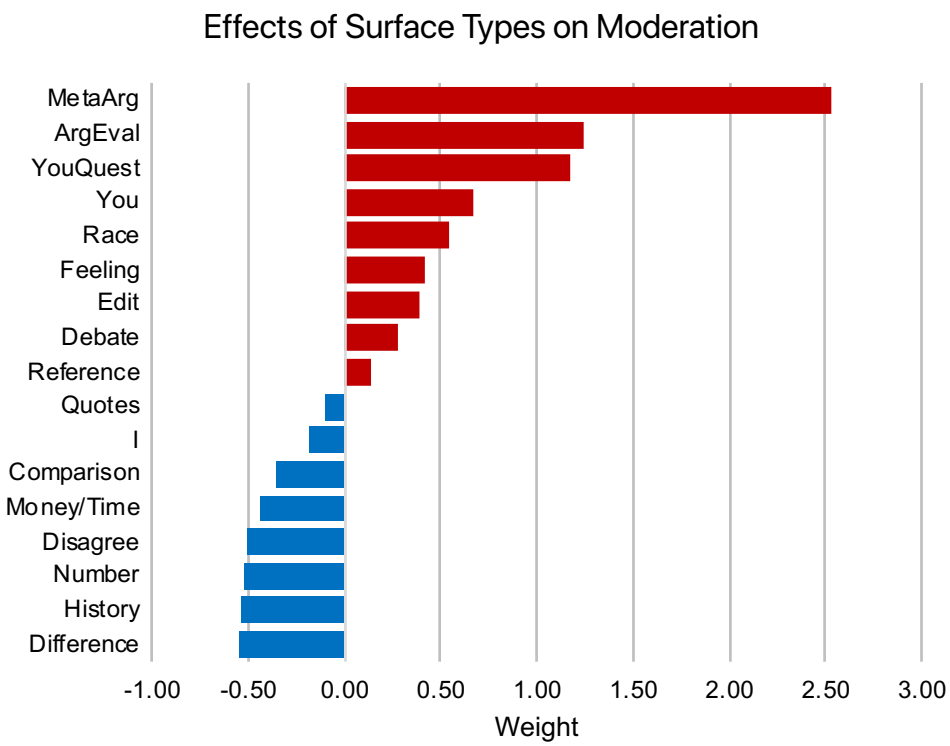


Figure 4.7: Effects of surface types to moderation, learned by the logistic regression component in CSM.

Among these surface types, we filtered those that fit with violations of BID's moderation guidelines, and call them *high-risk behavior*. These surface types include MetaArg, ArgEval, YouQuest, You, Race, Feeling, and Debate. Propositions of the types You and YouQuest directly address another user and may contain harsh personal judgments that were moderated for being uncivil or attacking. The rules of BID espouse argumentation around the topic and not the users participating in the debate. Propositions of the type Feeling, which are largely made up of exclamations and short comments, contain many snippy statements that could come off as being uncivil and dismissive to another user. They sometimes sarcastically dismiss a previous comment as being beneath the author's attention. Propositions of the types ArgEval, MetaArg, and Debate probe and evaluate other perspectives and can be inherently threatening to other users. Lastly, propositions of the type Race contain racial terms and content, which can be offensive to some users.

After identifying this set of high-risk surface types, we combine their weights to create the control variable *high_risk*, which characterizes to what extent a given post has some form of high-risk behavior. Before we combine them, we standardize the proportions of each of the high-risk surface types across all posts to account for differences in scale between surface types. This also allows us to measure the intensity of a surface type in terms of standard deviations from its mean. For a given post, we then take its maximum weight over the high-risk surface types as the value of the *high_risk* variable. Taking the maximum weight allows us to indicate if at least one of the high-risk types has a high intensity in a post. Thus, the *high_risk* gives us a measure of

Features	Model 1	Model 2	Model 3	Model 4
high_risk			2.95***	2.65***
minority	4.27***			3.01***
mod_prev		2.06***		1.62***
high_risk x minority				1.03
high_risk x mod_prev				0.99
minority x mod_prev				1.00
Cross-Validation F1	50.34	57.71	67.89	73.34

Table 4.12: Odds ratios of different features as explanatory variables for predicting whether a post is moderated or not (***$p < 0.001$).

whether a post has an unusually large amount of the identified high-risk surface types.

4.5.2 Results

Table 4.12 summarizes the findings from our regression on which factors contribute to the likelihood of a post being moderated. Models 1-3 test the effect of each feature separately, and Model 4 includes all the features and their interactions. All the features have significant positive effects on being moderated. But the high-risk behavior has stronger predictive power than minority viewpoints and moderation history (Cross-validation F1 for Model 3 vs. Models 1-2). In addition, the interaction variables have no effects, indicating that the high-risk behavior captures a different cause for moderation than minority viewpoints and moderation history. These results suggest that the surface types are a robust indicator of moderation independent of a user's minority/majority viewpoint and moderation history.

A user's perception of moderation bias is perhaps best reflected in Model 1 and Model 2. Users who consistently express minority viewpoints are indeed more likely to be moderated than users who consistently express majority viewpoints (OR=4.272 in Model 1). Similarly, Users who were moderated in the near past are more likely to be moderated again (OR=2.060 in Model 2). Although their effects slightly decrease when the high-risk behavior is introduced (Model 4), they still show significant positive effects on being moderated (OR=3.005 for minority viewpoints and 1.622 for moderation history).

From our regression analysis, we find evidence that some users' perception of unfair censorship is not absurd; the moderators of BID are more likely to moderate the posts of users with minority viewpoints and moderation history, even after accounting for high-risk surface types that appear in the post. In the remainder of this section, we discuss explanations for the actual bias we see in BID, the issues surrounding the perception of bias in political discussions, and future work to address the dual problems of actual and perceived bias on political discussion forums.

Sources of Actual Bias in BID

In the case of BID, moderators can be susceptible to bias against certain viewpoints for a number of reasons. One of the most notable systemic reasons for bias (Bazerman et al., 2002) is ambiguity in how rules and guidelines can be interpreted. Users of BID explicitly raise this issue of rule ambiguity:

> "It's been said so many times I've lost count but the answer is: decide on clear, unambiguous rules; state them clearly; moderate for breaking those rules. Instead we keep going for nonsense like "be excellent" "be civil" "civil discourse"."

This type of ambiguity can make moderation susceptible to the cognitive biases of individual moderators (Bazerman et al., 2002) and mask subjectivity in determining who is acting in a "civil" way. When moderators are not aware of these biases and instead believe they are acting objectively, this can make moderation even more biased (Kaatz et al., 2014).

Specific cognitive biases that could influence moderators to moderate unfairly include the *ecological fallacy*, making assumptions about individuals based on judgments about a group (Kaatz et al., 2014). In the context of BID, moderators likely recognize users who express conservative viewpoints and make judgments based on that group membership instead of individual behavior. *In-group/out-group bias* (Kaatz et al., 2014) may also be a factor in moderator bias. Moderators may more easily make negative judgments about users expressing positions that differ from their own group's. Unfortunately, we cannot easily compare the ideological positions of the moderators in BID with the users they judge. Moderators do not give their names with mod edits and the current Ravelry API does not include logs of post edits, so pinpointing the specific moderator who handed down judgment is impossible. Additionally, it is difficult to determine the viewpoints of the moderation team on BID with our current approach for assigning ideology. Though moderators can in theory participate in debate threads they are not moderating, moderators in practice almost never post outside of their moderating duties. This is likely due to the high workload of the moderator role and a previous prohibition against all moderator participation in debate, which some moderators still follow.

Even without biased behavior from the moderation team, users with minority viewpoints in BID could still be more likely to be moderated if more of their posts are flagged. The moderation process in BID begins with users anonymously flagging posts as potentially violating the rules of discussion, which moderators then judge. Posts from majority-view users may be less likely to be flagged as there are, by definition, fewer users who have the incentive to flag offensive posts from majority-view users. In this case, even if moderators make fair judgments given what they see, due to imbalance in flagging they may miss posts that should be moderated from majority-view users.

Sources of Perceived Bias

Ambiguity in the moderator guidelines may also play a role in why users perceive bias against them when they are moderated. Vague rules, such as "Behave civilly" in BID, allow users to

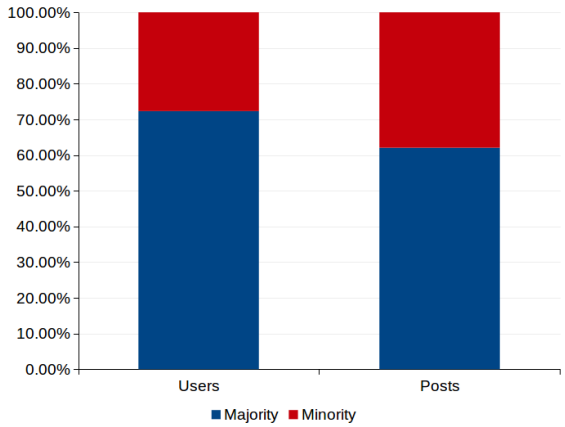


Figure 4.8: Comparison of viewpoint distributions over users vs. posts. The proportion of majority vs. minority are different between users and posts with statistical significance $p < 0.001$ by Pearson's chi-square test. Note that the distribution of viewpoints over posts is more balanced than the distribution of viewpoints over users.

make judgments about their behavior in their own self-interest (Bazerman et al., 2002). As it is in their interest not to get moderated, a user may be prone to *blind-spot bias* (Kaatz et al., 2014) and perceive themselves as being more civil than they actually are. If these users are then moderated, they may be inclined to believe that moderators made an unfair judgment by moderating them for their "civil" behavior. While we saw that most users viewed the main issue of censorship in BID to be false negative judgments, some users do argue that they have been moderated without cause:

> "Excuse me Pop but who did I personally attack ... Could you please clarify why my post was modded?"
> "Again, can you explain how this post is off topic/about myself?"

Another possible explanation for the perception of biased moderation from minority-view users in general is that minority users may experience a *halo effect* where their perception of the moderators are shaped by their experiences with other users within the group. Kelly et al. (2005) found that in political Usenet groups, minority-view posts are overrepresented compared to the population of minority-view authors, meaning minority-view users generate more posts per person than majority-view users. We see this same pattern in BID (Figure 4.8). This pattern suggests that individual minority users must spend more effort on defending their views, as there are fewer people on their side who can help support their arguments. As a result, these minority-view users may feel like they are outnumbered and targeted by majority-view users, who can afford to spend less effort individually. These feelings of unfairness could be transferred to the moderation team, as the moderators are responsible for regulating conversations and maintaining order within the group.

Potential Interventions

One way of addressing the image of moderators as biased dictators is to shift both the power and burden of moderation in the group. Studying the political branch of the technology news

aggregator Slashdot, Lampe et al. (2014) argue for the success of a distributed moderation system in which users with highly rated comments can become moderators, who in turn are allowed to rate others' comments higher or lower. Along with a "meta-moderation" system that broadly crowdsources the review of moderator actions, they argue that this model can filter out unproductive behaviors as well as develop and pass on community norms. Such a meta-moderation system could not only counter moderator bias, but improve feelings of ownership in the moderation system for users who are not moderators. A danger of these meta-moderation systems that rely on the user base, however, is that minority-view users have fewer protections against the majority. An independent panel of judges may be helpful in protecting minority-view users from the tyranny of the majority, yet these judges should be made aware of their own biases to avoid introducing blind-spot biases (Kaatz et al., 2014).

Moderators accused of censorship are often criticized for providing little evidence for why a particular post is moderated while others are not. One possible intervention in these cases is an automated system that does not directly classify posts as needing moderation, but instead provides better grounding for the discussions between moderators and those being moderated (Gweon et al., 2005). An example of such a grounding is an automated metric of inflammatory language that also provides comparisons to similar past posts that have been moderated. Making this visible to both the moderators and users could lend greater transparency and objectivity to how moderators operate, though this method would have to be safeguarded against the possibility of reproducing the bias of previous moderation.

Finally, it may be possible to address some of the sources of perceived and actual bias by working towards reducing ambiguity in how rules of proper debate are written. Most moderated discussion forums, like BID, frame their rules primarily in terms of what NOT to do (e.g. No personal attacks, don't derail the thread, etc.) Even the positively worded statement "Behave civilly" in BID is framed in terms of what not to do, as it is unclear what it means to behave in a civil manner. It instead implicitly tells users not to be uncivil. These negatively framed rules, however, are unlikely to capture the full range of offensive or inappropriate behavior, as users will try to find ways to circumvent the rules. One possible way of reducing the number of users skirting around ambiguous, negatively-framed rules is reframing rules in terms of positive discussion behaviors that users should include before they post. Encouraging political moderators to enforce rules in terms of what users should do may reduce both inappropriate behaviors and rule ambiguity by clearly defining what is expected of users.

Conclusion

Moderation in political discussion forums can be controversial, especially when claims of illegitimate censorship of specific views and individuals arise. In this study, we examined what surface types are likely to induce moderation and whether perceived unfairness against minority-view conservative users is grounded when these surface types are accounted for in Ravelry's Big Issues Debate forum. We found that users holding minority views and moderation history are more likely to be moderated, even after accounting for levels of potentially offensive posting behaviors. The perception that there is bias against certain subgroups remains an issue in political forums, as it may lead to tension and conflict over how moderation should be handled. We argue that

ambiguity in how guidelines are laid out exacerbates cognitive biases, explaining how both actual bias from the moderators and the perception of bias from users arise. We make recommendations for interventions that mitigate these biases by reducing ambiguity and increasing transparency in moderation decisions. While our study focuses primarily on Big Issues Debate, the techniques presented can easily be applied to other political debate forums and it is likely that our findings about the issue of perception of bias are not exclusive to this context.

Acknowledgement: *This study was led mainly by Qinlan Shen. Yohan Jo's contributions include: extracting and interpreting surface types (Section 4.5.1) and interpreting the logistic regression results (Section 4.5.2).*

4.6 Study 3. Surface Types and Persuasion

In this study, we focus on deliberative dialogue, the kind of argumentation whose goal is to broaden the understanding of an issue. The ChangeMyView (CMV) forum provides a platform for such dialogue, where users (OPs) post their viewpoints on diverse issues and other users (challengers) try to persuade OPs to change their viewpoints. While the social sciences and communication studies have researched the change of attitudes (Petty and Cacioppo, 1986) and rhetorical tools for literature and advertisements, we still lack quantified analyses of effective argumentative moves in deliberative dialogue. As a step for addressing this problem, we examine surface types and their influence on persuasion.

4.6.1 Experiment Settings

We use the CMV corpus for our analysis. We collect pairs of an OP's argument and a direct response by a challenger; hence, each pair is one exchange of utterances between an OP and a challenger. Many factors would affect whether this exchange changes the OP's viewpoint or not. The main factor of our interest is the surface types used by the challenger. However, the success of persuasion also depends on the domain (people are very stubborn for certain topics like religion and politics) and the OP's argumentative moves (surface types). Hence, in our first analysis, we conduct a logistic regression where the response variable is the result of persuasion (1 if successful and 0 otherwise) and the explanatory variables are the surface types used by the OP and the challenger (binary), plus the domain (categorical):

$$\begin{aligned}\text{Persuasion result} \sim &\text{OP's surface types} + \text{Challenger's surface types} \\ &+ \text{Domain}.\end{aligned}$$

Moreover, the success of persuasion may also depend on the challenger's argumentative moves *in relation to* the OP's. Hence, we estimate the effects of the interactions of surface types between the OP and the challenger. We use a logistic regression similar to the one above, but add interaction variables:

$$\begin{aligned}\text{Persuasion result} \sim &\text{OP's surface types} + \text{Challenger's surface types} \\ &+ \text{Interactions of surface types between OP and Challenger} \\ &+ \text{Domain}.\end{aligned}$$

4.6.2 Results

Table 4.13 summarizes the surface types learned from the CMV corpus.

Table 4.14 shows the odds ratios of surface types used by OPs and challengers, and domains. Some surface types used by OPs have significant impacts on persuasion outcomes. For instance, when the OP expresses confusion (*Confusion*), their viewpoint is more likely to be changed. In contrast, using numbers (especially percentages *Percent*) and emphasizing specific terms (*Term*) signal a less likelihood of their viewpoint being changed, probably because using these surface types implies that the OP has concrete evidence or a specific intention for using certain terms. Directly addressing the hearer (*You*) might reflect the OP's aggressiveness and is indicative of the failure of persuasion.

For challengers' uses of surface types, providing concrete numbers (especially percentages *Percent*) and references (*URL*), and clarifying definitions (*Definition*) significantly contribute to the success of persuasion. Presenting different choices (*Choice*) and making a comparison (*Comparison*) also help to change the OP's viewpoint. Directly expressing confusion (*Confusion*) and disagreement (*NoSense*), and even asking questions (*Question*) generally have positive effects on persuasion.

As expected, OPs have the tendency to not change their views for some domains, such as sex, gender, race, media, religion, drug, abortion, life, gun, job, and human.

Table 4.15 shows the effects of challengers' argumentative moves *in relation to* the OP's. The OP's directly addressing the hearer (*You*) was found to be negatively correlated with successful persuasion in general, but when the challenger responds with history (*History*) or emphasis specific terms (*Term*), the chance of success increases. Similarly, the OP's using numbers (*Percent*) reduces the chance of success in general, responding with meaning (*Meaning*), specific numbers (*Number*), or references (*URL*) alleviates this tendency. The challenger's expressing disagreement (*NoSense*) generally has a positive effect on persuasion, but this effect decreases if the OP uses a normative statement (*Normative*), and explains meaning (*Meaning*) or definitions (*Definition*).

A main take-away here is that the effect of a surface type can vary depending on the context. That is, some surface types have positive effects on persuasion in general but the effects decrease if certain surface types are used by the discussion partner, and vice versa. Furthermore, some surface types do not have a consistent effect in general, but show a significant effect in certain contexts. Our analysis here does not provide much detail about the mechanisms of interactions between surface types. Nevertheless, it suggests what kinds of interactions are meaningful and would be interesting topics for further nuanced analyses.

4.7 Study 4. Surface Types and Pro-/Counter-Argumentation

Before we delve into relations between propositions in terms of pro-/counter-argumentation in the next chapter, we dedicate this short section as a bridge and conduct some preliminary analysis on how surface types are associated with pro- and counter-argumentation. Specifically, we aim to see if the types of two propositions signal whether they form a pro-argument or a counter-argument.

Surface Type	Description
You	Directly addressing the hearer. (“You are a bad person if you do this.”)
Percent	Numbers, especially including percentage. (“Most of the wealth that the 1% has comes from investments or from ownership of different companies/projects/properties.”, “The owner still pays 100’s of millions of dollars.”)
Normative	Normative statements. (“We need to do something about climate change.”, “We can’t avoid it.”)
Meaning	Meaning of a statement. (“Did you perhaps mean “part genetic part nature”?”, “He clearly was not talking about “woman”.”)
Quotes	Quotations. (“If they respond by saying “saying ‘sorry’ doesn’t mean anything,” it’s a little harsh, but understandable given the context.”)
NotThink	Expressing disagreement. (“I don’t think he came nearly as close to making a threat like these.”)
Number	Numbers. (“Healthy does not mean organic fancy 20$/kg rice.”)
Saying	Pointing out a saying. (“Saying ""evidence of a non-trivial nature"" isn’t very illuminating.”)
ArgEval	Evaluation on an argument. (“That is a perfectly accurate statement.”, “It’s absurd to say the government has the power to make some random Joe take you skydiving.”)
Question	Questions. (“Then by what basis do you make a case for the existence of white privilege?”)
Difference	Pointing out that there is a difference. (“Again, that’s the difference between prescriptive and descriptive.”)
Choice	Presenting choices. (“Men should therefore be more feminine or have a new set of masculine values”)
History	Past event. (“Tea Party-Trumpism is the direction the party has been heading in for years now.”)
I	Personal belief or stories. (“I could care less that the drug is illegal in terms of its moral for you to use it.”)
NoSense	Expressing disagreement. (“It doesn’t make sense to argue that God shouldn’t take responsibility for the destruction and pain that they indiscriminately cause to people.”)
Confusion	Expressing confusion. (“I’m not at all sure why.”)
Term	Emphasizing terms. (“It does not “change” an existing definition.”)
URL	URLs. (“Being able to transition [**vastly improves trans youth’s mental health**](https://thinkprogress.org/allowing-transgender-youth-to-transition-improves-their-mental-health-study-finds-dd6096523375#.pqspdcee0)”)
Definition	Definitions. (“That’s what “black” denotes for me – my membership in a specific diaspora.”)
Comparison	Making a comparison. (“The top-down control of the economy in a communist country has a greater tendency towards famine than the competition of capitalism.”)

Table 4.13: Surface types learned from the CMV corpus.

OP			Challenger		
You	0.89	(*)	You	1.03	( )
Percent	0.87	(*)	Percent	1.35	(***)
Normative	1.03	( )	Normative	1.13	(*)
Meaning	0.96	( )	Meaning	1.05	( )
Quotes	0.86	( )	Quotes	1.19	( )
NotThink	1.09	( )	NotThink	1.01	( )
Number	0.98	( )	Number	1.14	( )
Saying	1.08	( )	Saying	1.19	(**)
ArgEval	0.91	( )	ArgEval	1.01	( )
Question	1.09	( )	Question	1.13	(*)
Difference	1.01	( )	Difference	1.10	( )
Choice	0.97	( )	Choice	1.16	(*)
History	1.01	( )	History	1.12	( )
I	1.04	( )	I	1.11	( )
NoSense	0.94	( )	NoSense	1.14	(*)
Confusion	1.11	(*)	Confusion	1.18	(**)
Term	0.71	(***)	Term	1.20	( )
URL	1.08	( )	URL	1.29	(***)
Definition	0.99	( )	Definition	1.27	(***)
Comparison	1.02	( )	Comparison	1.25	(***)
Domain			**Domain**		
food	0.97	( )	gun	0.91	(***)
music	0.98	( )	job	0.94	(*)
college	0.97	( )	world	0.97	( )
israel	0.97	( )	gender	0.92	(***)
family	1.04	( )	tax	0.95	( )
money	0.98	( )	power	0.99	( )
religion	0.94	(**)	relationship	1.01	( )
law	0.98	( )	race	0.90	(***)
drug	0.93	(**)	economy	0.98	( )
abortion	0.95	(*)	game	0.99	( )
war	0.98	( )	human	0.92	(***)
crime	1.00	( )	media	0.89	(***)
school	1.02	( )	election	0.98	( )
life	0.93	(**)	movie	1.00	( )
sex	0.87	(***)	reddit	0.97	( )

Table 4.14: Odds ratio (OR) and statistical significance of features. An effect is positive if OR $>$ 1 (blue) and negative if OR $<$ 1 (red). (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

OP→Challenger		OP→Challenger	
You→History	1.39 (*)	You→Term	1.71 (*)
You→URL	0.65 (*)	Percent→You	0.74 (*)
Percent→Meaning	1.49 (*)	Percent→Number	1.56 (*)
Percent→Difference	0.63 (**)	Percent→URL	1.50 (**)
Normative→NoSense	0.71 (*)	Normative→Confusion	1.39 (*)
Meaning→You	1.50 (*)	Meaning→Choice	0.67 (*)
Meaning→NoSense	0.67 (*)	Quotes→You	1.53 (*)
Quotes→Saying	0.53 (*)	Number→Normative	1.66 (**)
Number→Meaning	0.56 (*)	Quotes→Normative.1	1.35 (*)
ArgEval→Normative	0.69 (*)	ArgEval→History	1.51 (**)
Difference→Quotes	1.62 (*)	Choice→History	0.72 (*)
History→NotThink	1.37 (*)	Definition→Quotes	1.61 (*)
Definition→NoSense	0.71 (*)	Comparison→NotThink	1.47 (**)
Comparison→ArgEval	0.77 (*)		

Table 4.15: Odds ratio (OR) and statistical significance of features. An effect is positive if OR $>$ 1 (blue) and negative if OR $<$ 1 (red). (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

4.7.1 Experiment Settings

The US2016 corpus is annotated with pro or counter relations between propositions; a proposition may be supported or attacked by a set of propositions. Here, the supported/attacked proposition is called a *claim* and the supporting/attacking propositions are called *premises*. Given the surface types learned from this corpus, we first examine the association between the surface type of a premise and the pro/counter relation it forms with the claim. For this analysis, we use a logistic regression where the explanatory variable is the surface types of a premise (categorical) and the response variable is whether the premise supports (1) or attacks (0) the claim:

$$\text{Pro-argument} \sim \text{Premise's surface type.}$$

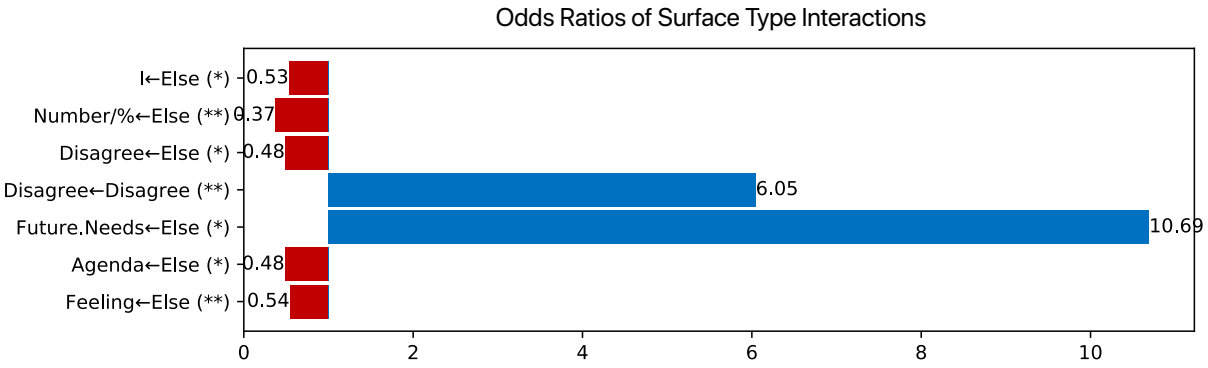
In the second analysis, we look at the interaction of surface types between a premise and the claim. In other words, we want to see if the pair of surface types between a premise and the claim signals whether they form a pro-argument or a counter-argument. The logistic regression model includes interaction variables in additional to the premise's surface type.

$$\begin{aligned}\text{Pro-argument} \sim \text{Premise's surface type} + \\ \text{Premise's surface type} * \text{Claim's surface type.}\end{aligned}$$

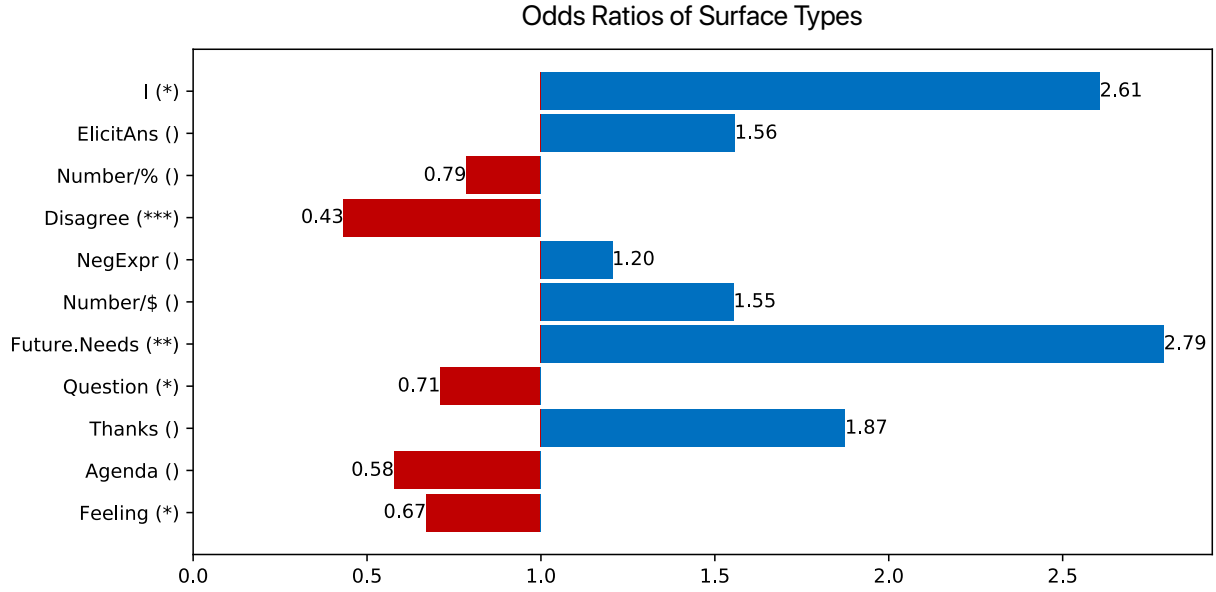
4.7.2 Results

We summarize the surface types learned from this corpus in Table 4.16.

Figure 4.9a shows the odds ratios (OR) of surface types that contribute to pro-argumentation (as opposed to counter-argumentation). Here is a summary of the trend. Talking about the speaker's own story (*I*) is highly and significantly associated with supporting premises, indicating



(a) Odds ratios of surface types. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)



(b) Odds ratios of surface type interactions. Only statistically significant interactions are shown. $P1 \leftarrow P2$ means that $P1$ is the surface type of the claim and $P2$ is that of the premise. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

Surface Type	Description
I	Talking about the speaker's story. ("I'm not a gun advocate by any stretch.", "As a young person, I'm very concerned about climate change and how it will affect my future.")
ElicitAns	Eliciting answers from debaters. ("Sanders please answer the questions.")
Number	Using numbers. ("He's 10 years old.")
Number/%	Using numbers, especially percentage. ("Global warming increases by 22.4%", "she has had more air time then over half the candidates have.")
Number/$	Using numbers, especially about money. ("I don't know how they do on $11,000 , $12,000 , $13,000 a year")
Disagree	Expressing disagreement. ("Sorry you are flat out wrong.", "No its tremendous.")
NegExpr	Using negated expressions. ("It was not too late for Wisconsin.", "I do not have a lot of time to research further.")
Future/Needs	Expressing future events, including needs. ("We need more debates with people who just say shit with no filter.", ""if you are learning, you're gonna change your position."")
Question	Questions. ("What would it take to destroy ISIS in 90 days?")
Thanks	Thanks. ("Thank you")
Agenda	Structure of debates. ("You know, tonight we hear about what people want to do", "Well, I do not expect us to cover all the issue of this campaign tonight, but I remind everyone there are two more presidential debates scheduled.")
Feeling	Feeling. ("that's what scares me.", "lol")

Table 4.16: Surface types learned from the US2016 corpus.

that debaters likely use their own stories and experiences to form a pro-argument. Similarly, predicting a future event or presenting a need (*Future/Needs*) is associated pro-argumentation; what will happen and what is needed are common content of supporting premises. Not surprisingly, expressing disagreement (*Disagree*) and throwing questions (*Question*) are strongly correlated with counter-argumentation. Expressing feelings (*Feeling*) is also significantly correlated with counter-argumentation, probably because premises of this type often laugh at a claim or express a grief. We have not found significant correlations with pro-argumentation for such surface types as eliciting answers (*ElicitAns*), using numbers (*Number/%*, *Number/$*), thanking (*Thanks*), and presenting an agenda (*Agenda*).

Figure 4.9b shows the odds ratios (OR) of the interactions of surface types between a premise and the claim toward pro-argumentation. We find that when the speaker expresses disagreement, it is likely to emphasize and strengthen a claim that expresses disagreement (*Disagree*←*Disagree*). However, overall, not many interactions strongly and significantly signal a particular relation type between propositions, implying that similar combinations of surface types are used to form pro-argumentation and counter-argumentation, and thus it is difficult to predict pro- and counter-argumentation based solely on surface types. In other words, we really have to look into the meaning of propositional content in order to identify the relations between propositions, which will be discussed in depth in the next part of the thesis.

4.8 Conclusion

In this chapter, we applied CSM to four corpora of argumentative dialogue, and identified underlying surface types. We found that despite the different domains and goals of these corpora, there are surface types that are common across argumentative dialogue, such as questions, thanks, numbers, references, disagreement, meaning, and personal stories. Based on the identified surface types, we demonstrated that certain surface types are strongly associated with various outcomes of argumentative dialogue. These analyses are by no means exhaustive, but they suggest interesting directions for future work; based on the findings from these analysis, more nuanced analyses may be conducted to understand the sophisticated mechanisms of surface types in achieving the goal of argumentative dialogue.

Part II

Argumentative Relations

In Part I, we focused on individual propositions and their types. In Part II, we focus on relations between propositions and how these relations constitute pro- and counter-argumentation. To clarify our terminology and scope, we see an argument as consisting of a claim and a premise, and each claim or premise in turn consists of one or more asserted propositions. Henceforth, we use the general notion of **statement** to refer to a claim or premise that comprises proposition(s). This work focuses on the interaction between statements rather than individual propositions within them.

In informal logic, argumentation schemes play an important role in categorizing and assessing reasoning used in an argument. In NLP, researchers have struggled to apply argumentation schemes to computational work due to the difficulty of data annotation. In Chapter 5, we propose an effective human-machine hybrid annotation protocol and apply it to annotate four main types of statements in argumentation schemes. We further show the affinity between these types in formation of natural arguments and argumentation schemes. In Chapter 6, we investigate four logical and theory-informed mechanisms that constitute argumentative relations between statements (support, attack, and neutral): factual consistency, sentiment coherence, causal relation, and normative relation. They explain argumentative relations effectively and can further improve supervised classifiers through representation learning.

Chapter 5

Annotating Proposition Types in Argumentation Schemes

Modern machine learning pipelines for analyzing argument have difficulty distinguishing between types of statements based on their factuality, rhetorical positioning, and speaker commitment. Inability to properly account for these facets leaves such systems inaccurate in understanding of fine-grained proposition types. In this chapter, we demonstrate an approach to annotating for four proposition types common in the statements of argumentation schemes, namely *normative propositions*, *desires*, *future possibility*, and *reported speech*. We develop a hybrid machine learning and human workflow for annotation that allows for efficient and reliable annotation of complex linguistic phenomena, and demonstrate with preliminary analysis of structure and rhetorical strategies in presidential debates. We develop a corpus of the 2016 U.S. presidential debates and commentary, containing 4,648 argumentative statements annotated with the four proposition types. This new dataset and method can support technical researchers seeking more nuanced representations of argument, as well as argumentation theorists developing new quantitative analyses.

5.1 Introduction

Argument mining is a broad field of computational linguistics that seeks to identify the structure of written and spoken argument and extract meaningful content based on that understanding. But as the domains that we can tackle with NLP grow more diverse and expand from newswire text to social media and real-world dialogue, we are reaching an inflection point. These domains are not characterized solely by objective statements with clean reporting of facts and details; opinion, hedging, and reported speech are commonplace. In recent years, researchers have found that argument mining pipelines struggle to identify factual content and disambiguate it from fiction, lies, or mere hypotheticals in real-world data (Feng et al., 2012; Thorne et al., 2018). In today's politically charged atmosphere, this poses a challenge for developers of systems like fake news detectors and recommender systems: when algorithmic systems cannot even reliably detect the presence or assertion of facts in statements, how can they address the ethical challenges of

deployed machine learning systems at scale (Leidner and Plachouras, 2017; Gonen and Goldberg, 2019)?

In this chapter, we introduce new resources for understanding statements that appear in speech and text, based on the 2016 U.S. presidential debates. We define a fine-grained, four-dimensional annotation schema for how propositions are introduced rhetorically in debates: namely, **normative**, **desire**, **future possibility**, and **reported speech** propositions. These proposition types are tied closely to practical reasoning, causal reasoning, and authority claims in argumentation schemes (Walton et al., 2008) and represent varying levels of speaker commitment to individual statements (Lasersohn, 2009).

While these definitions are tractable for reliable human annotators, we find that occurrences in running text are rare and annotation is both difficult and inefficient. In response, we develop a machine learning model with high recall for finding likely candidates for positive labels, and describe a hybrid annotation workflow that boosts the efficiency of human annotators by 39-85% while further improving reliability. Using this process we produce a corpus of annotated statements. We conclude with a preliminary analysis of how these proposition types are used in political debate and commentary. Our contributions in this chapter are as follows:

- A multi-dimensional annotation schema for fine-grained proposition types that are tied to argumentation schemes and speaker commitment. In our work this schema has been proven to be tractable and robust for both human and automated annotation.
- An effective, efficient, and novel methodology for hybrid machine-aided annotation. To address logistic challenges with annotating sparse labels in our task, we introduce additional best practices for hybrid human-machine systems for building datasets. This method produces efficient machine filtering, especially of likely negative instances, which covers a large percentage of our corpus. Human annotator time is prioritized on potential positive instances, which are harder to recognize automatically with high precision.
- A public sample annotated corpus of statements using that schema, along with full annotation manuals and baseline classification code. This dataset contains annotated instances of novel proposition types, such as reported speech, and is more than three times larger than comparable recent corpora. All these materials may enable further progress in the community.

5.2 Related Work

5.2.1 Argument Mining and Statement Types

Argument mining is an expansive field with many applications. Datasets include the Internet Argument Corpus for online debate on political topics (Walker et al., 2012; Swanson et al., 2015), student argument in course essays (Stab and Gurevych, 2017), and parliamentary debate (Duthie et al., 2016). State-of-the-art results have been produced using a range of methods including random forests (Aker et al., 2017), integer linear programming for constraint-based inference (Persing and Ng, 2016a), graph-based methods that focus on relations between claims (Niculae et al., 2017; Nguyen and Litman, 2018), and more recently, end-to-end neural methods (Cocarascu

and Toni, 2018; Frau et al., 2019). But these systems struggle to distinguish between distinctions in argumentative strategy that look intuitively obvious to casual observers, instead relying on coarse notions of claims and premises.

Today, automated systems fail to understand the nuanced factuality of these statements when they appear in argumentation. Perceived factuality of statements, it turns out, are heavily tied to an author's intent (Wentzel et al., 2010); this concept of authors or speakers making claims with only partial certainty or factuality have been collectively studied under the umbrella term of "commitment" to a truth value for claims (Lasersohn, 2009). Naderi and Hirst (2015) give examples of statements that are not straightforwardly factual, but instead contain statements deeply embedded in hypotheticals and shifts in tense, beyond the current bounds of today's NLP:

> "Who among us would dare consider returning to a debate on the rights of women in our society or the rights of visible minorities?"
> "How can we criticize China for imprisoning those who practise their religion when we cannot offer protection of religious beliefs in Canada?"

Later, Haddadan et al. (2018) describe the context-dependent annotation task of identifying premises and claims in political discourse, providing the following statement from the 1960 Nixon-Kennedy presidential debate:

> "Communism is the enemy of all religions; and we who do believe in God must join together. We must not be divided on this issue."

It turns out ideas are not only factual or fictitious, but lie on a many-dimensional gradient. They can be positioned carefully when making arguments, negotiating, or manipulating a discourse (Potter, 1996), and authors take care to distinguish between claims they know to be true, desires they have for the future, amid other epistemological states of reported knowledge (Walton et al., 2008).

In argumentation theory and communication sciences, statements are typically divided into three types: fact, value, and policy (Hollihan and Baaske, 2015; Wagemans, 2016). Statements of *fact* have contents whose truth value is verifiable with empirical evidence, whereas statements of *value* are subjective judgments. Statements of *policy* propose that an action be carried out. These types have been extended by prior studies. For instance, Park and Cardie (2018) extended *fact* into *non-experiential fact* and *testimony*, and added *reference*—a text of information source (but not reported speech in itself). Egawa et al. (2019) further added *rhetorical statement*, judgments of *value* using figurative language and discourse structure.

While most prior work extended statement types based on the needs of the task at hand, our taxonomy has been motivated mainly by argumentation theory. In particular, the argumentation schemes of Walton et al. (2008) are a set of reasoning types commonly used in daily life. Each scheme defines the form of a conclusion and the form(s) of one or more premises. As an example, the scheme of *argument from consequences* is as follows:

> **Premise:** "If *A* is brought about, good consequences will plausibly occur."
> **Conclusion:** "*A* should be brought about."

These schemes have been adopted by many studies as a framework for analyzing reasoning pat-

terns (Song et al., 2017; Nussbaum, 2011). Researchers in computational linguistics have tried to code the schemes, but this task turned out to be very challenging; as a result, annotations have low agreement between annotators (Lindahl et al., 2019) or are available only from experts (Lawrence et al., 2019). But different schemes are associated with different proposition types, and therefore, we speculate that reliably annotating proposition types may ease the annotation of argumentation schemes. The proposition types in this chapter are closely related to common argumentation schemes, including practical reasoning, argument from consequence, argument from cause to effect, and argument from expert opinion.

5.2.2 Efficient Linguistic Annotation

In their overview of argument mining today, Lippi and Torroni (2016) identify three key challenges that limit the field:

1. The subtlety of the task requires more time-consuming and expensive training to achieve high inter-rater reliability, compared to tasks like object detection in computer vision, limiting the size and breadth of corpora available to researchers.
2. Because of the lack of existing data, there are few automation tools available to expedite the annotation of future datasets, leaving the field with too much unsupervised data and not enough labels.
3. The structured nature of claims and premises limits the utility of widely-used classification algorithms.

More recent reviews of the field have made similar observations (Lawrence and Reed, 2019; Janier and Saint-Dizier, 2019). Researchers have suspected that part of the challenge in these problems is data collection and reliable annotation. Collecting span- and sentence-level annotations is a frequently used tool for machine learning researchers seeking to improve their systems. Accurate annotation is time-consuming and expensive, though, and even when funding is available, annotation tasks often require subject matter expertise that comes from either lived experience or extensive training. This problem is exacerbated by rare phenomena, which results in imbalanced datasets in many domains, like emotional crisis or suicidal ideation detection online and in medical records (Pestian et al., 2012; Imran et al., 2016; Losada and Crestani, 2016), rare occurrence of high- and low-end scores in student data in education domains (Woods et al., 2017; Lugini and Litman, 2018), and rare social behaviors in healthcare settings (Mayfield et al., 2013; Carrell et al., 2016). Our annotation also handles rare phenomena, and using a conventional annotation methodology allows only moderate inter-annotator agreement even after intensive annotator training, reflecting the difficulty of our task.

Many previous papers on text annotation have relied on crowdsourcing, relying on inexperienced editors on services such as Crowdflower and Amazon Mechanical Turk (Snow et al., 2008; Swanson et al., 2015). While this approach works for many common-sense tasks, prior work has shown that achieving high inter-rater reliability with these services is arduous and relies on many strict methodological choices and narrowing of task type (Alonso et al., 2015; Hoffman et al., 2017). When converting real-world phenomena into categorical judgments that can achieve high reliability, nuance is often lost in the name of inter-annotator agreement. This requires researchers

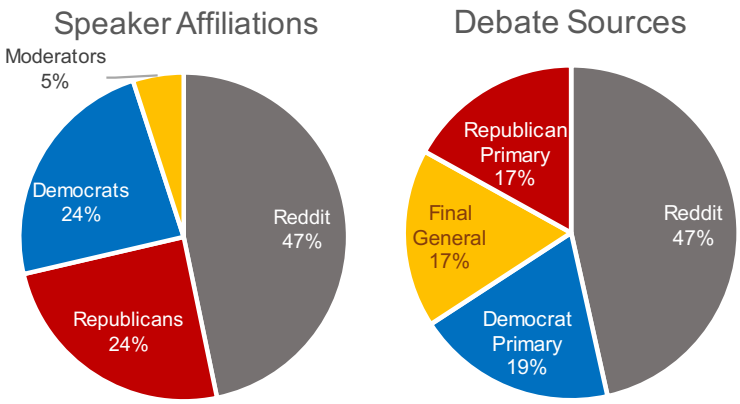


Figure 5.1: Speaker affiliations and debate sources.

to make a trade-off between, on one hand, the expressiveness and fidelity of the linguistic construct they are attempting to capture, and on the other the potential for operationalization and quantification in coding manuals and fully automated systems. Particularly in imbalanced tasks, these choices can have the effect of producing an inaccurate picture of the minority class and producing datasets that are no longer a valid representation of the original construct (Corbett-Davies and Goel, 2018).

To expedite annotation without sacrificing validity, researchers have developed annotation tools that incorporate machine learning (Pianta et al., 2008; Yimam et al., 2014; Klie et al., 2018). These tools train a machine learning algorithm on a subset of annotations and suggest predicted annotations for new data, producing a hybrid "human-in-the-loop" model (da Silva et al., 2019). Our work here follows in this tradition, seeking effective and efficient methods for collecting reliable new data.

5.3 Domain Description

For all annotation and experiments in this work, we use transcripts of the 2016 U.S. presidential debates and reaction to the debates on Reddit (Visser et al., 2019) (Section 4.2.4). This corpus is appropriate for our task as it includes various rhetorical moves by both politicians and observers in social media. In addition, human annotators have extracted propositions from all dialogues and posts, and identified claim-premise pairs with support and attack relations. Our work focuses on 4,648 propositions that are part of claim-premise pairs with support relations. Approximately half of our data comes directly from debate transcripts, with the remainder coming from social media response. From the transcripts of the debates themselves, approximately 10% of statements come from moderators while the remainder comes from candidates themselves. The full distributions of speaker affiliations and debate sources are shown in Figure 5.1.

We are not the first researchers to study this domain. Haddadan et al. (2018) annotated similar presidential debates dating back to 1960, while numerous researchers have studied argumentation on Reddit and similar social media sites (Jo et al., 2018). Datasets have also been developed for

similar annotation schemes, like the more syntactically and lexically constrained CommitmentBank (Jiang and de Marneffe, 2019), and for the 2016 U.S. presidential election in particular (Savoy, 2018). Our work, however, is the first to date to examine argumentation frames in this context, at this level of depth, in primarily computational work.

5.4 Defining Proposition Types

This work does not attempt to cover all of argumentation theory; instead, we focus on four important proposition types: normative, desire, future possibility, and reported speech. Using the language from prior work, in our taxonomy *future possibility*, *desire*, and *reported speech* are subtypes of *fact*, while *normative* is close to *policy*. We do not assume that these proposition types are mutually exclusive, choosing to adopt binary annotation for each proposition type. More details and examples are available in the full annotation manuals.

5.4.1 Normative

A normative proposition is defined as a proposition where the speaker or someone else proposes that a certain situation should be achieved or that an action should be carried out. A normative proposition, under our definition, carries the explicit force of community norms and policies[1], as opposed to a mere desire or valuation, and includes commands, suggestions, expression of needs, and prohibitive "can't". An example proposition is:

> "The major media outlets **should not be the ones** dictating who wins the primaries."

Normative propositions are tightly related to several argumentation schemes. For instance, the argument from consequences scheme proposes that a certain action should (or shouldn't) be carried out because of a potential consequence. Practical reasoning also asserts a normative conclusion in order to achieve a certain goal (Walton et al., 2008). Prior studies have referred to similar normative statements as "policy" annotations (Park et al., 2015; Egawa et al., 2019).

5.4.2 Desire

A desire proposition is defined as a proposition that explicitly claims that the speaker or someone else desires to own something, do something, or desires for a certain situation to be achieved. A desire is usually *weaker* than normative propositions and carries no explicit force of proposal or norm. Actively desiring something is also different than merely valuing that thing or asserting a future possibility. An example proposition is:

> "At the very least for the first debate **I'd like to see everyone get a fair shot** at expressing themselves."

In practical reasoning, a normative conclusion is supported by a certain goal to achieve, and this goal is often expressed as another normative proposition or a desire as in:

[1] Albeit through the implicit lens of the speaker or writer's interpretation and understanding of those norms.

Claim: "Let's have paid family leave."
Premise: "**I want US to do more** to support people who are struggling to balance family and work."

Prior work has paid little attention to annotating desire statements. In NLP, the closest work is in subjectivity annotation and the more narrow task of annotating subjectively beneficial events (Somasundaran and Wiebe, 2010; Deng et al., 2013), but these approaches have typically been applied in the context of sentiment analysis; our approach focusing on argument is, to our knowledge, a new contribution in computational linguistics.

5.4.3 Future Possibility

A future possibility proposition claims a possibility or prediction that something may be the case in the future. These future possibilities are independent of whether the speaker desires the forecast to be true, or believes they *should* be true; the claimed future possibility is just the speaker's own, or someone else's, belief about what the future may hold:

"US shooting down a Russian jet **could easily turn ugly**."

Speakers describing their own future plans are also counted as a future possibility. Propositions with future possibilities are often used to support conclusions in the argument from consequences scheme, as in the following example:

Claim: "Bring us to a 350 ship Navy again, and bring our Air Force back to 2,600 aircraft."
Premise: "Those are the kind of things **that are going to send a clear message around the world**."

An additional scheme, *argument from cause to effect*, also makes use of future possibility as a conclusion, supported by factors that may cause the future event.

5.4.4 Reported Speech

Our last proposition type is reported speech. A reported speech proposition must convey an explicit or implicit predicate borrowed from a source external to the speaker. We extend the scope of "speech" to belief, thoughts, and questions, in order to capture a wider range of contents borrowed from external sources:

"**Many in the Black Lives Matter movement, and beyond, believe that** overly-aggressive police officers targeting young African Americans is the civil rights issue of our time."

For each proposition of reported speech, we also annotate text spans that represent the source and the content, and mark the credibility of the source as high, low, or unsure.

Reported speech plays a critical role in discourse; the alignment of a statement with a third-party source allows for both distancing an author from the claim, and for simultaneously strengthening that claim by appealing to the authority of the original source (Walton et al., 2008). In practice,

this is used as a sophisticated rhetorical tool in argument, as a trigger to agree or disagree with the position (Janier and Reed, 2017), to make authority claims (Walton et al., 2008), or even to commit straw man fallacies (Talisse and Aikin, 2006). In the NLP community, a prior study identified authority claims in Wikipedia talk pages (Bender et al., 2011), but the ways of referring to task-oriented norms in these pages are different from general reported speech in argumentation. Park et al. (2015) annotated references (e.g., URLs) in policy-related argumentation, but reported speech was not included as references.

As a methodological note, in the original corpus the pronoun "I" has been resolved to the speaker's name in the process of annotating propositions from locutions (e.g., for the sentence "I believe Americans do have the ability to give their kids a better future", "I believe" has been replaced with "O'MALLEY believes") (Jo et al., 2019). As a result, it is difficult to tell whether the source of a reported speech proposition is indeed the speaker or not. For annotation, we are faithful to the text of each proposition as it is, resulting in many instances of reported speech that can be used for machine learning. Since some of these instances are not reported speech in the original debates, however, our post hoc analyses (Section 5.6) exclude instances whose speaker and report source are identical.

5.5 Annotation Workflow

The workflow of our annotation process is designed to manage three concurrent problems. First, our annotations require detailed reading of an annotation manual and are difficult to acquire from the minimally trained workers typically used in contexts like crowdsourcing. Second, positive instances are rare (less than 15% of the total dataset for each statement type), in which case capturing positive instances is challenging but crucial for high inter-annotator agreement and the high quality of annotations. And third, because of the high engagement needed by individual annotators and the lack of positive examples in freely occurring text, the collection and labeling of a dataset sufficiently large to perform quantitative studies and train downstream argument mining classifiers is expensive and logistically challenging.

We solve these problems by leveraging a machine annotator trained on a set of annotations. After we train two human annotators on a subset of data, the remaining corpus is split between them. To expedite annotation, the machine annotator annotates the data first and separates it into a large percentage of instances that are covered by highly reliable machine annotation and only need quick review by humans (mostly negative instances), and a remaining small portion that needs to be annotated as usual. To maintain high quality of the final annotated dataset, as a final step all human annotations are compared with the machine annotations, and discrepancies are resolved by an adjudicator (the author of this thesis).

An overview of this annotation process is shown in full in Figure 5.2. For each proposition type, our annotation follows a three-stage process. Stage 1 is to train two human annotators. In Stage 2, we train a machine annotator and calibrate it to optimize dataset coverage and accuracy. In Stage 3, the remaining data is annotated by the human and machine annotators in collaboration; final discrepancies are resolved by the adjudicator.

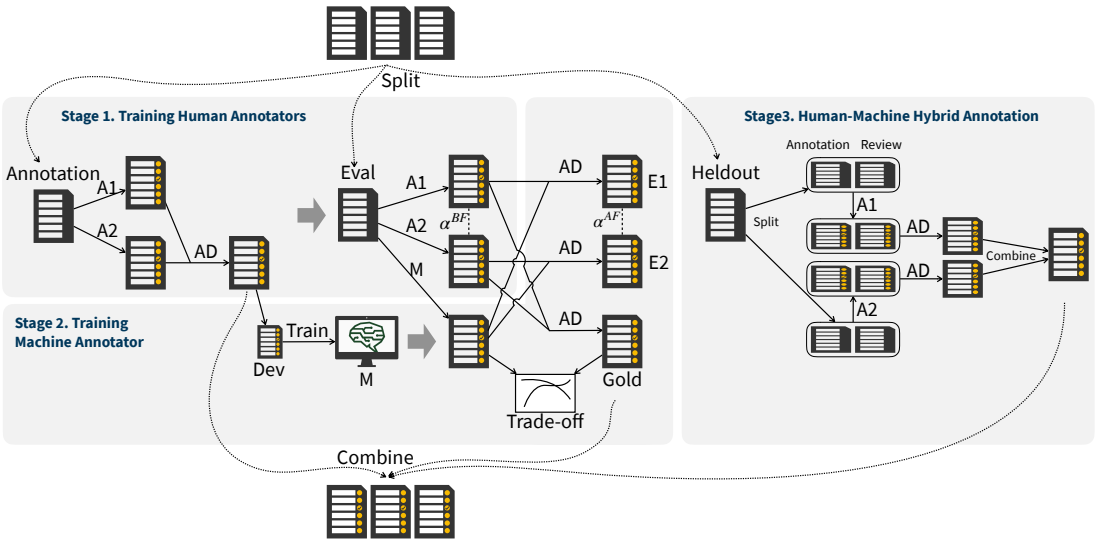


Figure 5.2: Workflow of annotation process. A1, A2, and AD are two human annotators and the adjudicator, respectively. M is the machine annotator.

Category	Annotation (Dev)	Eval	Heldout Annotation	Heldout Review
Normative	1,497 (924)	400	461	2,290
Future	1,497 (424)	400	433	2,318
Desire	1,497 (424)	400	340	2,411
Rep. Speech	997 (997)	400	541	2,710

Table 5.1: Statistics of data splits.

5.5.1 Initial Training for Human Annotators

In this stage, we train two human annotators and evaluate their inter-annotator agreement. We recruited two undergraduate students as annotators; they have no particular experience in argumentation or rhetoric. Approximately 30% of the data (**Annotation**) is used for developing annotation guidelines and training human annotators iteratively over multiple rounds. We then evaluate the final annotation guidelines for reliability on the **Eval** set, approximately 10% of the entire data (Table 5.1).

Inter-annotator agreement (IAA) was measured using Krippendorff's alpha. We achieve results of $\alpha = 0.67$ for the normative type, $\alpha = 0.59$ for the desire type, $\alpha = 0.66$ for the future possibility type, and $\alpha = 0.71$ for the reported speech type. Despite quite intensive training of human annotators, the main challenge for achieving substantially high IAA is the small number of positive instances; missing a few positive instances greatly affects the IAA score. This motivates our use of the machine annotator as the third annotator.

For reported speech, we also annotated the text spans of sources and contents, and the credibility of the sources. To evaluate the annotators' agreement on sources or contents, we first filtered

Category	α^{BF}	α^{AF}	DA (A1)	DA (A2)
Normative	0.67	0.97	6.8%	11.0%
Desire	0.59	0.86	2.5%	4.5%
Future	0.66	0.96	4.8%	4.5%
Reported Speech	0.71	0.83	13.0%	13.0%

Table 5.2: IAA on the Eval set. α^{BF} and α^{AF} are the IAA before and after machine involvement, respectively. "DA (A1)" and "DA (A2)" are the instance-level disagreement rates between the machine and the two human annotators.

Category	Prec	Recl	F1	AUC
Normative	84.1	88.7	86.1	98.1
Desire	100.0	70.0	80.0	95.1
Future	60.0	81.4	62.8	98.2
Reported Speech	44.6	92.9	59.4	96.4

Table 5.3: Machine performance using 5-fold cross validation.

statements that both annotators marked as reported speech, and for each statement, we obtained the longest common sequence of words between two text spans from the annotators. The average number of words that are outside of the common span is 0.5 for sources and 0.2 for contents. Most mismatch comes from articles ("the experts" vs. "experts") or modifiers ("President Clinton" vs. "Clinton"). For credibility annotations, the annotators agreed on 85% of the annotations. These results show that the annotations of sources, contents, and credibility are reliable.

5.5.2 Training Machine Annotator

In this stage, we train a machine annotator and calibrate it to optimize the amount of dataset it covers and annotation accuracy. A subset of the Annotation set is annotated on the final, independently reliable annotation guidelines (**Dev**) and used for training the machine annotator for each proposition type (Table 5.1). For machine learning feature representation and labeling, we use the single sentence classification model in BERT[2] (Devlin et al., 2018). The input is the full text of a proposition, and the output is the probability that the input proposition is an instance of the corresponding proposition type. Representation is fully automated in a deep neural model that makes extensive use of attention weights and intermediate representations. We used the pretrained uncased, base model with the implementation provided by Hugging Face (Wolf et al., 2020). The machine annotator's accuracy on the Dev set using 5-fold cross validation is shown in Table 5.3.

To evaluate how the machine annotator can improve the reliability of annotations, the **Eval** set was also annotated by the machine, and discrepancies between the machine predictions and

[2]We tried logistic regression with extensive hand-crafted features as well, but BERT performed significantly better.

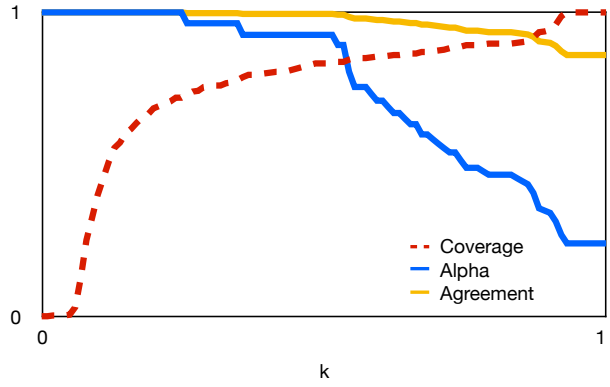


Figure 5.3: Trade-off between data coverage and annotation accuracy as the threshold of machine-predicted probability k varies. This graph is for Reported Speech, but other proposition types have similar tendencies.

original human annotations from both annotators were resolved by the adjudicator (E1 and E2 in Figure 5.2). As expected, the IAA improved significantly from before adjudication (α^{BF}) to after adjudication (α^{AF} in Table 5.2); the final adjudicated agreement between annotators is between 0.83 and 0.97. The disagreement rate between a human annotator and the machine annotator—annotations that need to be adjudicated—ranges between 2.5% and 13.0%.

We next move to questions for developing a hybrid human-machine annotation pipeline. We take advantage of the distribution of classifier output probabilities, finding that the machine annotator has very high AUC scores (Table 5.3), and that the shape of the probability distributions is well-suited to filtering out instances that are unlikely to contain positive examples of our proposition types. We define a probability threshold k and say that instances with probability of a positive label less than k are *covered* by the model.

We analyzed how k affects the coverage and annotation accuracy on the Eval set. For this analysis, we first created gold standard annotations for the Eval set by the adjudicator resolving disagreements between the annotations of the two human annotators (**Gold** in Figure 5.2). Then, for each value of k, we replaced the labels of instances whose predicted probability is lower than k with the machine annotator's decisions and measured the IAA and agreement rate between these partially replaced annotations and the Gold set. Figure 5.3 shows visually the trade-off between this threshold, quantity of data covered, and annotation accuracy as k increases:

- **Dataset coverage (red line):** A large percentage of instances, over half, are clumped together and assigned probabilities of positive labels of approximately $k = 0.2$. After this large group of negative instances comes a steadier growth in coverage between k =0.2–0.9.
- **Agreement (Krippendorf's α, blue line; accuracy, yellow line):** This estimates the lower bound of accuracy from human-machine hybrid annotation without final adjudication. Initially, for low values of k, accuracy remains at or approximately 100%, because the machine filters out likely negative instances well. As k grows, overall model accuracy decreases.

Category	Metric	k	Coverage	α	Agree
Normative	mean	.19	78%	.98	99.0%
Desire	max	.34	81%	.95	99.6%
Future	mean	.39	88%	.95	99.1%
Reported Speech	mean	.35	78%	.96	99.7%

Table 5.4: Final configurations of the machine annotator.

This resulting model is a good fit for a hybrid human-machine annotation workflow. The models efficiently filter out negative samples with high coverage at a relatively low value of k, producing a much smaller and more balanced set of candidate statements for human annotation. Below this threshold, instances are assigned negative labels automatically and are only subject to very efficient human review; above this threshold, humans are required for a more time-consuming full annotation process. Table 5.4 shows the hyperparameter selection of mean or max probabilities of the 5-fold classifiers; the tuned threshold k for each proposition type; and the resulting data coverage, α, and agreement rate (accuracy).

5.5.3 Human-Machine Hybrid Annotation

In the last stage of our workflow, the remaining data (**Heldout**) is split between the two human annotators. Each split is further split into an annotation set and a review set (Table 5.1); the annotation set is annotated by the human annotator as usual, and the review set is pre-annotated by the machine, and reviewed and corrected by the human annotator. Since human annotators may make mistakes, the annotations of a human annotator for both the annotation and review sets are compared with the machine annotations, and disagreements are resolved by the adjudicator.

Detailed statistics of annotation speed and disagreement rates are listed in Table 5.5. On average, the review session is three times faster than the annotation session, expediting annotation significantly for a large portion of the data. Both annotators see efficiency boosts of between 39.0% and 85.3%, depending on proposition type, when moving from the full annotation process to review of machine annotations. We observe that the two human annotators have different annotation paces for each proposition type. This situation is common in many annotation tasks where data is split among annotators; although it could potentially result in inconsistent annotations, many annotation studies do not take a further step of quality control. In our task, when all human annotations were compared with the machine annotations, on average 6% of instances had disagreement, which was resolved by the adjudicator (Table 5.5). This emphasizes the value of our approach, using a machine annotator to double check human annotations and resolve potentially incorrect annotations with a small effort of adjudication. The prevalence of each proposition type for the entire annotated dataset is shown in Table 5.6. Labels are not exclusive or conditioned on each other; in total, 30% of the final dataset contains at least one positive annotation, and most other positions describe judgments and facts.

Category	A1				A2			
	Annotation	Review	Gain	Agreement	Annotation	Review	Gain	Agreement
Normative	17.3	3.0	82.7%	93.2%	6.4	1.9	70.3%	96.5%
Desire	8.3	3.4	59.0%	96.5%	4.9	1.8	63.3%	95.0%
Future	10.4	5.3	49.0%	93.0%	10.9	1.6	85.3%	99.0%
Reported Speech	10.6	6.5	39.0%	86.6%	22.4	7.1	67.4%	91.2%

Table 5.5: Annotation speed (sec/statement) and efficiency gain moving from full annotation to review of machine labels, and instance-level agreement rates between single human and machine annotation on the Heldout set.

	Num of statements
Normative	602 (13%)
Desire	147 (3%)
Future	453 (10%)
Reported Speech	242 (5%)
Total	4,648 (100%)

Table 5.6: The number of positive instances and their proportion for each proposition type for the entire data.

5.6 Analysis of U.S. Presidential Debates

Our annotations readily allow us to conduct some interesting analyses of the 2016 U.S. presidential debates. First, different speakers in the debates use different rhetorical strategies, and our proposition types shed light on how the strategies differ in terms of the kinds of statements made by the speakers. Next, we analyze varying types of claims made in the debates and what types of premises are commonly used to support those claims.

5.6.1 Use of Proposition Types by Main Speakers

Across individual speakers: As representative examples of how our annotations can be used to evaluate language in use, we first chose five main speakers to examine how they differ in their use of proposition types: Donald Trump, Hillary Clinton, Bernie Sanders, Anderson Cooper, and Reddit users (as an aggregated group). Trump and Clinton were the nominees of the Republican and Democratic Parties, while Sanders was a competitive rival of Clinton. Cooper was a main moderator of the debates. For each of these speakers, we calculated the proportion of each proposition type and then normalized these proportions to z-scores.

As shown in Figure 5.4, these five exemplar speakers use proposition types differently (their distributions of the types are significantly different with $p <$ 1e-5 for a χ^2 test). When compared to Trump, the Democratic candidates make much greater use of normative language. In particular,

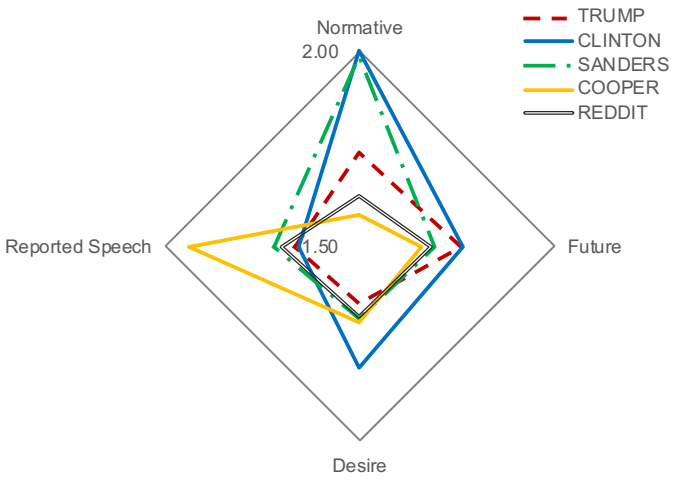


Figure 5.4: Use of proposition types by five main speakers, normalized to z-scores.

language from the two Democratic candidates uses normative statements and expresses desires a lot more than Trump, often to make the case for specific policies based on normative values. Clinton makes the most use of normative language, while Sanders mentions future possibilities less and uses reported speech slightly more. But the major differentiator is in normative language, where he mirrors Clinton.

> **Clinton:** "We also, though, need to have a tax system that rewards work and not just financial transactions"
> **Sanders:** "War should be the last resort that we have got to exercise diplomacy"

These normative judgments are not absent entirely from Trump's language, but they are less prevalent. Among moderators, Cooper uses significantly more reported speech and less normative statements and future possibilities than candidates or online commenters, which matches his role as a moderator in contemporary politics. While early debates in the television era leaned on questions from moderators that were "unreflective of the issues salient to the public" (Jackson-Beeck and Meadow, 1979), moderators today view themselves as serving a "gatekeeping" function that is nevertheless representative of a curated version of engaging questions that are relevant to the public interest, expressed through reported speech (Turcotte, 2015). Lastly, Reddit users make use of less rhetorical structure than candidates of either party or moderators, instead focusing more on past/current events, facts, and more straightforward rhetoric. This is reflected in their lower use of normative statements, future possibilities, and desire compared to the candidates.

Across affiliations: Next, we examined whether there is a significant difference in use of proposition types among Republican candidates, Democratic candidates, and Reddit users. We split statements into the three groups (excluding moderators) and tested for differences in proportion of each proposition type across groups, using χ^2 tests.

As shown in Table 5.7, Democratic candidates as a whole continue the trend we observe in individual speakers. They use more normative statements and desire expressions than Republican candidates, and this result across groups is highly significant. However, they had no significant

	Normative	Desire	Reported Speech
Dem vs. Rep	+++	+	
Reddit vs. Dem	−−	−	++
Reddit vs. Rep	−−		+

Table 5.7: Comparison of proposition types used by Republicans, Democrats, and Reddit users. +/− represents whether the group on the left uses a higher/lower proportion of the proposition type than the group on the right, and the numbers of +/− indicate significance levels (one: $p < .05$, two: $p < .01$, three: $p < .001$). There was no significant difference in use of future possibilities among the groups.

difference in use of reported speech and future possibilities. Reddit users make less use of argumentation proposition types in general: they use less normative language than the candidates and express less desire than Republican candidates. However, they use reported speech often, partly because much of their discussions occurs after the debates have occurred. As a result, these texts often refer back to speech from the debates themselves and the reported speech of the candidates.

5.6.2 Proposition Types in Claim-Premise Pairs

The statements in our work are drawn from claim-premise pairs, as annotated in the original corpus. As such, we are able to merge our annotations with this pre-existing structure for deeper analysis. We do so first by examining the correlations of the proposition types between claims made in the debates and their supporting premises. We computed the correlations between proposition types in claim-premise pairs as follows. First, since a few propositions have more than one proposition type, we chose the main type in the importance order of normative, desire, reported speech, and future possibility. Propositions that do not belong to any of these types are classified as other. For each type of claims, we calculated the distribution over proposition types for their premises and normalized again to z-scores (Figure 5.5).

Each proposition type has different degrees of correlations with other proposition types. Naturally, proposition types often match between claims and premises—the appearance of a particular proposition type in a premise conditioned on that type appearing in a claim is high (the diagonal of the table). We see many instances of normative statements supported by a normative premise, constituting practical reasoning:

> **Claim:** "We need to control our border."
> **Premise:** "It's our responsibility to pick and choose who comes in."

Similarly, many claims of future possibility are supported by the same type of premise, constituting an argument from cause to effect:

> **Claim:** "Families' hearts are going to be broken."
> **Premise:** "Their kids won't be able to get a job in the 21st Century."

On the other hand, certain pairings are deeply unnatural and rarely-occurring in natural text.

	Normative	Future	Desire	Reported Speech	Other
Normative	1.59	-0.43	-0.70	-0.37	-0.67
Future	-0.24	1.70	-0.31	-1.35	-0.78
Desire	0.26	0.07	1.74	0.33	-0.74
Reported Speech	-0.96	-0.70	-0.59	1.39	1.11
Other	-0.65	-0.64	-0.14	0.00	1.08

Figure 5.5: Normalized z-scores of correlations between proposition types in claim-premise pairs. Rows are claim types and columns are premise types.

Pairs comprised of future-looking claims based on premises from reported speech, for instance, are the least likely pairing in our dataset. The correlation analysis supports our belief that proposition types can be useful information for studying argument.

This analysis has application to tasks like argument generation, where correlation information may inform systems of what kind of premise should likely follow a certain type of claim in natural speech, allowing parameterization beyond mere topic and into natural language generation that controls for style and structure, a goal of recent work (Prabhumoye et al., 2018). For argumentation scheme annotation, high correlations between proposition types imply that the proposition types may reflect different argumentation schemes, and may provide a structured crosswalk between argumentation theory, which are often nuanced and resist quantification at scale, and NLP advances that are often limited to labeling tasks.

5.7 Conclusion

Through the introduction of this new corpus of the U.S. 2016 presidential debates and commentary, annotated with four proposition types that capture nuanced building blocks of argumentation schemes, we hope to advance the state of the art in argument mining. For effective annotation, we presented a human-machine hybrid annotation protocol that allows for efficient and reliable annotation for difficult annotation tasks involving complex reasoning and rare occurrences of positive instances. We believe this methodology is replicable in the identification, annotation, and study of sociolinguistic or argument features more broadly that appear rarely. Today's machine learning systems struggle with such skewed distributions in a fully automated context, but we demonstrated that both the speed and inter-annotator reliability of these annotations can be enhanced with a hybrid approach that makes targeted, selective use of machine learning methods. Future research should test whether the distributional properties that make this approach effective in our domain, like high recall and near-100% precision in low-probability negative instances, are part of a more general pattern in annotation of rare linguistic phenomena in text and speech.

Chapter 6

Classifying Argumentative Relations

While argument mining has achieved significant success in classifying argumentative relations between statements (support, attack, and neutral), we have a limited computational understanding of logical mechanisms that constitute those relations. Most recent studies rely on black-box models, which are not as linguistically insightful as desired. On the other hand, earlier studies use rather simple lexical features, missing logical relations between statements. To overcome these limitations, our work classifies argumentative relations based on four logical and theory-informed mechanisms between two statements, namely (i) factual consistency, (ii) sentiment coherence, (iii) causal relation, and (iv) normative relation. We demonstrate that our operationalization of these logical mechanisms classifies argumentative relations without directly training on data labeled with the relations, significantly better than several unsupervised baselines. We further demonstrate that these mechanisms also improve supervised classifiers through representation learning.

6.1 Introduction

There have been great advances in argument mining—classifying the argumentative relation between statements as support, attack, or neutral. Recent research has focused on training complex neural networks on large labeled data. However, the behavior of such models remains obscure, and recent studies found evidence that those models may rely on spurious statistics of training data (Niven and Kao, 2019) and superficial cues irrelevant to the meaning of statements, such as discourse markers (Opitz and Frank, 2019). Hence, in this work, we turn to an *interpretable* method to investigate *logical relations* between statements, such as causal relations and factual contradiction. Such relations have been underemphasized in earlier studies (Feng and Hirst, 2011; Lawrence and Reed, 2016), possibly because their operationalization was unreliable then. Now that computational semantics is fast developing, our work takes a first step to computationally investigate how logical mechanisms contribute to building argumentative relations between statements and to classification accuracy with and without training on labeled data.

To investigate what logical mechanisms govern argumentative relations, we hypothesize that governing mechanisms should be able to classify the relations without directly training on

relation-labeled data. Thus, we first compile a set of rules specifying logical and theory-informed mechanisms that signal the support and attack relations (§6.3). The rules are grouped into four mechanisms: factual consistency, sentiment coherence, causal relation, and normative relation. These rules are combined via probabilistic soft logic (PSL) (Bach et al., 2017) to estimate the optimal argumentative relations between statements. We operationalize each mechanism by training semantic modules on public datasets so that the modules reflect real-world knowledge necessary for reasoning (§6.4). For normative relation, we build a necessary dataset via rich annotation of the normative argumentation schemes *argument from consequences* and *practical reasoning* (Walton et al., 2008), by developing a novel and reliable annotation protocol (§6.5).

Our evaluation is based on arguments from kialo.com and debatepedia.org. We first demonstrate that the four logical mechanisms explain the argumentative relations between statements effectively. PSL with our operationalization of the mechanisms can classify the relations without direct training on relation-labeled data, outperforming several unsupervised baselines (§6.7). We analyze the contribution and pitfalls of individual mechanisms in detail. Next, to examine whether the mechanisms can further inform supervised models, we present a method to learn vector representations of arguments that are "cognizant of" the logical mechanisms (§6.8). This method outperforms several supervised models trained without concerning the mechanisms, as well as models that incorporate the mechanisms in different ways. We illustrate how it makes a connection between logical mechanisms and argumentative relations. Our contributions are:

- An interpretable method based on PSL to investigate logical and theory-informed mechanisms in argumentation computationally.
- A representation learning method that incorporates the logical mechanisms to improve the predictive power of supervised models.
- A novel and reliable annotation protocol, along with a rich schema, for the argumentation schemes *argument from consequences* and *practical reasoning*. We release our annotation manuals and annotated data.[1]

6.2 Related Work

There has been active research in NLP to understand different mechanisms of argumentation computationally. Argumentative relations have been found to be associated with various statistics, such as discourse markers (Opitz and Frank, 2019), sentiment (Allaway and McKeown, 2020), and use of negating words (Niven and Kao, 2019). Further, as framing plays an important role in debates (Ajjour et al., 2019), different stances for a topic emphasize different points, resulting in strong thematic correlations (Lawrence and Reed, 2017).

Such thematic associations have been exploited in stance detection and dis/agreement classification. Stance detection (Allaway and McKeown, 2020; Stab et al., 2018; Xu et al., 2018) aims to classify a statement as pro or con with respect to a topic, while dis/agreement classification (Chen et al., 2018a; Hou and Jochim, 2017; Rosenthal and McKeown, 2015) aims to decide whether two statements are from the same or opposite stance(s) for a given topic. Topics are usually discrete,

[1]The annotations, data, and source code are available at: https://github.com/yohanjo/tacl_arg_rel

and models often learn thematic correlations between a topic and a stance (Xu et al., 2019). Our work is slightly different as we classify the *direct* support or attack relation between two *natural* statements.

The aforementioned correlations, however, are rather byproducts than core mechanisms of argumentative relations. In order to decide whether a statement supports or attacks another, we cannot ignore the *logical* relation between them. Textual entailment was found to inform argumentative relations (Choi and Lee, 2018) and used to detect arguments (Cabrio and Villata, 2012). Similarly, there is evidence that the opinions of two statements toward the same concept constitute their argumentative relations (Gemechu and Reed, 2019; Kobbe et al., 2020). Causality between events also received attention, and causality graph construction was proposed for argument analysis (Al-Khatib et al., 2020). Additionally, in argumentation theory, Walton's argumentation schemes (Walton et al., 2008) specify common reasoning patterns people use to form an argument. This motivates our work to investigate logical mechanisms in four categories: factual consistency, sentiment coherence, causal relation, and normative relation.

Logical mechanisms have not been actively studied in argumentative relation classification. Models based on hand-crafted features have used relatively simple lexical features, such as *n*-grams, discourse markers, and sentiment agreement and word overlap between two statements (Stab and Gurevych, 2017; Habernal and Gurevych, 2017; Persing and Ng, 2016b; Rinott et al., 2015). Recently, neural models have become dominant approaches (Chakrabarty et al., 2019; Durmus et al., 2019; Eger et al., 2017). Despite their high accuracy and finding of some word-level interactions between statements (Xu et al., 2019; Chen et al., 2018a), they provide quite limited insight into governing mechanisms in argumentative relations. Indeed, more and more evidence suggests that supervised models learn to overly rely on superficial cues, such as discourse markers (Opitz and Frank, 2019), negating words (Niven and Kao, 2019), and sentiment (Allaway and McKeown, 2020) behind the scene. We instead use an interpretable method based on PSL to examine logical mechanisms (§6.7) and then show evidence that these mechanisms can inform supervised models in intuitive ways (§6.8).

Some research adopted argumentation schemes as a framework, making comparisons with discourse relations (Cabrio et al., 2013) and collecting and leveraging data at varying degrees of granularity. At a coarse level, prior studies annotated the presence of particular argumentation schemes in text (Visser et al., 2020; Lawrence et al., 2019; Lindahl et al., 2019; Reed et al., 2008) and developed models to classify different schemes (Feng and Hirst, 2011). However, each scheme often accommodates both support and attack relations between statements, so classifying those relations requires semantically richer information within the scheme than just its presence. To that end, Reisert et al. (2018) annotated individual components within schemes, particularly emphasizing *argument from consequences*. Based on the logic behind this scheme, Kobbe et al. (2020) developed an unsupervised method to classify the support and attack relations using syntactic rules and lexicons. Our work extends these studies by including other normative schemes (*practical reasoning* and property-based reasoning) and annotating richer information.

6.3 Rules

We first compile rules that specify evidence for the support and attack relations between **claim** C and **statement** S (Table 6.1)[2]. These rules are combined via probabilistic soft logic (PSL) (Bach et al., 2017) to estimate the optimal relation between C and S[3].

We will describe individual rules in four categories: factual consistency, sentiment coherence, causal relation, and normative relation, followed by additional chain rules.

6.3.1 Factual Consistency

A statement that supports the claim may present a fact that naturally entails the claim, while an attacking statement often presents a fact contradictory or contrary to the claim. For example:

> **Claim:** "Homeschooling deprives children and families from interacting with people with different religions, ideologies or values."
>
> **Support Statement:** "Home school students have few opportunities to meet diverse peers they could otherwise see at normal schools."
>
> **Attack Statement:** "Homeschool students can interact regularly with other children from a greater diversity of physical locations, allowing them more exposure outside of their socio-economic group."

This logic leads to two rules:

$$\textbf{R1:} \quad \text{FactEntail}(S,C) \rightarrow \text{Support}(S,C),$$
$$\textbf{R2:} \quad \text{FactContradict}(S,C) \rightarrow \text{Attack}(S,C)$$
$$\textit{s.t.} \ \text{FactEntail}(S,C) = P(S \text{ entails } C),$$
$$\text{FactContradict}(S,C) = P(S \text{ contradicts } C).$$

In our work, these probabilities are computed by a textual entailment module (§6.4.1).

In argumentation, it is often the case that an attacking statement and the claim are not strictly contradictory nor contrary, but the statement contradicts only a specific part of the claim, as in:

> **Claim:** "Vegan diets are healthy."
>
> **Attack Statement:** "Meat is healthy."

Formally, let $(A^S_{i,0}, A^S_{i,1}, \cdots)$ denote the ith relation tuple in S, and $(A^C_{j,0}, A^C_{j,1}, \cdots)$ the jth relation tuple in C. We formulate the conflict rule:

$$\textbf{R3:} \quad \text{FactConflict}(S,C) \rightarrow \text{Attack}(S,C)$$

[2]We do not assume that claim-hood and statement-hood are intrinsic features of text spans; we follow prevailing argumentation theory in viewing claims and statements as roles determined by virtue of relationships between text spans.

[3]Predicates in the rules are probability scores, and PSL aims to estimate the scores of Support(S,C), Attack(S,C), and Neutral(S,C) for all (S,C). The degree of satisfaction of the rules are converted to a loss, which is minimized via maximum likelihood estimation.

		Rules
Factual Consist.	R1	$\text{FactEntail}(S,C) \rightarrow \text{Support}(S,C)$
	R2	$\text{FactContradict}(S,C) \rightarrow \text{Attack}(S,C)$
	R3	$\text{FactConflict}(S,C) \rightarrow \text{Attack}(S,C)$
Senti Cohe.	R4	$\text{SentiConflict}(S,C) \rightarrow \text{Attack}(S,C)$
	R5	$\text{SentiCoherent}(S,C) \rightarrow \text{Support}(S,C)$
Causal Relation		***CAUSE-TO-EFFECT REASONING***
	R6	$\text{Cause}(S,C) \rightarrow \text{Support}(S,C)$
	R7	$\text{Obstruct}(S,C) \rightarrow \text{Attack}(S,C)$
		EFFECT-TO-CAUSE REASONING
	R8	$\text{Cause}(C,S) \rightarrow \text{Support}(S,C)$
	R9	$\text{Obstruct}(C,S) \rightarrow \text{Attack}(S,C)$
Normative Relation		***ARGUMENT FROM CONSEQUENCES***
	R10	$\text{BackingConseq}(S,C) \rightarrow \text{Support}(S,C)$
	R11	$\text{RefutingConseq}(S,C) \rightarrow \text{Attack}(S,C)$
		PRACTICAL REASONING
	R12	$\text{BackingNorm}(S,C) \rightarrow \text{Support}(S,C)$
	R13	$\text{RefutingNorm}(S,C) \rightarrow \text{Attack}(S,C)$
Relation Chain	R14	$\text{Support}(S,I) \wedge \text{Support}(I,C) \rightarrow \text{Support}(S,C)$
	R15	$\text{Attack}(S,I) \wedge \text{Attack}(I,C) \rightarrow \text{Support}(S,C)$
	R16	$\text{Support}(S,I) \wedge \text{Attack}(I,C) \rightarrow \text{Attack}(S,C)$
	R17	$\text{Attack}(S,I) \wedge \text{Support}(I,C) \rightarrow \text{Attack}(S,C)$
Const-raints	C1	$\text{Neutral}(S,C) = 1$
	C2	$\text{Support}(S,C)+\text{Attack}(S,C)+\text{Neutral}(S,C) = 1$

Table 6.1: PSL rules. (S: statement, C: claim)

$$s.t.\ \text{FactConflict}(S,C) = \max_{i,j,k} P(A^S_{i,k} \text{ contradicts } A^C_{j,k}) \prod_{k' \neq k} P(A^S_{i,k'} \text{ entails } A^C_{j,k'}).$$

We use Open IE 5.1[4] to extract relation tuples, and the probability terms are computed by a textual entailment module (§6.4.1).

6.3.2 Sentiment Coherence

When S attacks C, they may express opposite sentiments toward the same target, whereas they may express the same sentiment if S supports C (Gemechu and Reed, 2019). For example:

Claim: "Pet keeping is morally justified."

[4] https://git.io/JTr3Y

Attack Statement: "Keeping pets is hazardous and offensive to other people."

Support Statement: "Pet owners can provide safe places and foods to pets."

Let (t_i^S, s_i^S) be the ith expression of sentiment $s_i^S \in \{\text{pos, neg, neu}\}$ toward target t_i^S in S, and (t_j^C, s_j^C) the jth expression in C. We formulate two rules:

$$\textbf{R4:} \quad \text{SentiConflict}(S,C) \rightarrow \text{Attack}(S,C),$$
$$\textbf{R5:} \quad \text{SentiCoherent}(S,C) \rightarrow \text{Support}(S,C)$$
$$\textit{s.t.}\ \text{SentiConflict}(S,C) =$$
$$\max_{i,j} P(t_i^S = t_j^C) \Big\{ P(s_i^S = \text{pos}) P(s_j^C = \text{neg}) + P(s_i^S = \text{neg}) P(s_j^C = \text{pos}) \Big\},$$
$$\text{SentiCoherent}(S,C) =$$
$$\max_{i,j} P(t_i^S = t_j^C) \Big\{ P(s_i^S = \text{pos}) P(s_j^C = \text{pos}) + P(s_i^S = \text{neg}) P(s_j^C = \text{neg}) \Big\}.$$

In this work, targets are all noun phrases and verb phrases in C and S. $P(t_i^S = t_j^C)$ is computed by a textual entailment module (§6.4.1), and $P(s_i^S)$ and $P(s_j^C)$ by a target-based sentiment classifier (§6.4.2).

6.3.3 Causal Relation

Reasoning based on causal relation between events is used in two types of argumentation: *argument from cause to effect* and *argument from effect to cause* (Walton et al., 2008). In cause-to-effect (C2E) reasoning, C is derived from S because the event in S may cause that in C. If S causes (obstructs) C then S is likely to support (attack) C. For example:

Claim: "Walmart's stock price will rise."

Support Statement: "Walmart generated record revenue."

Attack Statement: "Walmart had low net incomes."

This logic leads to two rules:

$$\textbf{R6:} \quad \text{Cause}(S,C) \rightarrow \text{Support}(S,C),$$
$$\textbf{R7:} \quad \text{Obstruct}(S,C) \rightarrow \text{Attack}(S,C),$$
$$\textit{s.t.}\ \text{Cause}(S,C) = P(S \text{ causes } C),$$
$$\text{Obstruct}(S,C) = P(S \text{ obstructs } C).$$

Effect-to-cause (E2C) reasoning has the reversed direction; S describes an observation and C is a reasonable explanation that may have caused it. If C causes (obstructs) S, then S is likely to support (attack) C, as in:

Claim: "St. Andrew Art Gallery is closing soon."

Support Statement: "The number of paintings in the gallery has reduced by

half for the past month."

Attack Statement: "The gallery recently bought 20 photographs."

$$\textbf{R8:} \quad \text{Cause}(C,S) \rightarrow \text{Support}(S,C),$$
$$\textbf{R9:} \quad \text{Obstruct}(C,S) \rightarrow \text{Attack}(S,C).$$

The probabilities are computed by a causality module (§6.4.3).

6.3.4 Normative Relation

In argumentation theory, Walton's argumentation schemes specify common reasoning patterns used in arguments (Walton et al., 2008). We focus on two schemes related to normative arguments, whose claims suggest that an action or situation be brought about. Normative claims are one of the most common proposition types in argumentation (Chapter 5) and have received much attention in the literature (Park and Cardie, 2018).

Argument from Consequences: In this scheme, the claim is supported or attacked by a positive or negative consequence, as in:

Claim: "Humans should stop eating animal meat."

Support Statement: "The normalizing of killing animals for food leads to a cruel mankind. (S1)"

Attack Statement: "Culinary arts developed over centuries may be lost. (S2)"

In general, an argument from consequences may be decomposed into two parts: (i) whether S is a positive consequence or a negative one; and (ii) whether the source of this consequence is consistent with or facilitated by C's stance (S2), or is contrary to or obstructed by it (S1)[5].

Logically, S is likely to support C by presenting a positive (negative) consequence of a source that is consistent with (contrary to) C's stance. In contrast, S may attack C by presenting a negative (positive) consequence of a source that is consistent with (contrary to) C's stance. Given that S describes consequence Q of source R, this logic leads to:

$$\textbf{R10:} \quad \text{BackingConseq}(S,C) \rightarrow \text{Support}(S,C),$$
$$\textbf{R11:} \quad \text{RefutingConseq}(S,C) \rightarrow \text{Attack}(S,C)$$
$$\textit{s.t.}\ \text{BackingConseq}(S,C) =$$
$$P(S \text{ is a consequence}) \times$$
$$\{P(Q \text{ is positive}) \cdot P(R \text{ consistent with } C)$$
$$+ P(Q \text{ is negative}) \cdot P(R \text{ contrary to } C)\},$$

[5] "Losing culinary arts" is a consequence of "stopping eating animal meat", which is the claim's stance itself and hence "consistent". In contrast, "a population with no empathy for other species" is a consequence of "the normalizing of killing animals for food", which is contrary to the claim's stance.

$$
\begin{aligned}
&\text{RefutingConseq}(S,C) = \\
&P(S \text{ is a consequence}) \times \\
&\quad \{P(Q \text{ is negative}) \cdot P(R \text{ consistent with } C) \\
&\quad + P(Q \text{ is positive}) \cdot P(R \text{ contrary to } C)\}\,.
\end{aligned}
$$

Practical Reasoning: In this scheme, the statement supports or attacks the claim by presenting a goal to achieve, as in:

> **Claim:** "Pregnant people should have the right to choose abortion."
> **Support Statement:** "Women should be able to make choices about their bodies. (S3)"
> **Attack Statement:** "Our rights do not allow us to harm the innocent lives of others. (S4)"

The statements use a normative statement as a goal to justify their stances. We call their target of advocacy or opposition (underlined above) a **norm target**. Generally, an argument of this scheme may be decomposed into: (i) whether S advocates for its norm target (S3) or opposes it (S4), as if expressing positive or negative sentiment toward the norm target; and (ii) whether the norm target is a situation or action that is consistent with or facilitated by C's stance, or that is contrary to or obstructed by it[6].

Logically, S is likely to support C by advocating for (opposing) a norm target that is consistent with (contrary to) C's stance. In contrast, S may attack C by opposing (advocating for) a norm target that is consistent with (contrary to) C's stance. Given that S has norm target R, this logic leads to:

$$
\begin{aligned}
&\textbf{R12:} \quad \text{BackingNorm}(S,C) \rightarrow \text{Support}(S,C), \\
&\textbf{R13:} \quad \text{RefutingNorm}(S,C) \rightarrow \text{Attack}(S,C) \\
&\textit{s.t.}\ \text{BackingNorm}(S,C) = \\
&\quad P(S \text{ is normative}) \times \\
&\quad \{P(S \text{ advocates for } R) \cdot P(R \text{ consistent with } C) \\
&\quad\ + P(S \text{ opposes } R) \cdot P(R \text{ contrary to } C)\}\,, \\
&\quad \text{RefutingNorm}(S,C) = \\
&\quad P(S \text{ is normative}) \times \\
&\quad \{P(S \text{ opposes } R) \cdot P(R \text{ consistent with } C) \\
&\quad\ + P(S \text{ advocates for } R) \cdot P(R \text{ contrary to } C)\}\,.
\end{aligned}
$$

The probabilities are computed by modules trained on our annotation data (§6.5).

[6]Harming innocent lives is facilitated by the right to choose abortion ('consistent'), whereas making choices about their bodies is obstructed by the right ('contrary').

6.3.5 Relation Chain

A chain of argumentative relations across arguments may provide information about the plausible relation within each argument. Given three statements S, I, and C, we have four chain rules:

R14: $\text{Support}(S,I) \wedge \text{Support}(I,C) \rightarrow \text{Support}(S,C)$,

R15: $\text{Attack}(S,I) \wedge \text{Attack}(I,C) \rightarrow \text{Support}(S,C)$,

R16: $\text{Support}(S,I) \wedge \text{Attack}(I,C) \rightarrow \text{Attack}(S,C)$,

R17: $\text{Attack}(S,I) \wedge \text{Support}(I,C) \rightarrow \text{Attack}(S,C)$.

For each data split, we combine two neighboring arguments where the claim of one is the statement of the other, whenever possible. The logical rules R1–R13 are applied to these "indirect" arguments.

6.3.6 Constraints

C and S are assumed to have the neutral relation (or the attack relation for binary classification) if they do not have strong evidence from the rules mentioned so far (Table 6.1 **C1**). In addition, the probabilities of all relations should sum to 1 (**C2**).

6.4 Modules

In this section, we discuss individual modules for operationalizing the PSL rules. Each module takes a text or a pair of texts as input and computes the probabilities of classes relevant to the module. For each module, we fine-tune the pretrained uncased BERT-base (Devlin et al., 2018), which has shown great performance in many NLP tasks. We use the transformers library v3.3.0 (Wolf et al., 2020) for high reproducibility and low development costs. But any other models could be used instead.

Each dataset used is randomly split with a ratio of 9:1 for training and test. Cross-entropy and Adam are used for optimization. To address the imbalance of classes and datasets, the loss for each training instance is scaled by a weight inversely proportional to the number of its class and dataset.

6.4.1 Textual Entailment

A textual entailment module is used for rules about factual consistency and sentiment coherence (R1–R5). Given a pair of texts, it computes the probabilities of entailment, contradiction, and neutral.

Our training data include two public datasets: MNLI (Williams et al., 2018) and AntSyn (Nguyen et al., 2017) for handling antonyms and synonyms. An NLI module combined with the word-level entailment handles short phrases better without hurting accuracy for sentence-level entailment. Since AntSyn does not have the neutral class, we add 50K neutral word pairs by randomly pairing two words among the 20K most frequent words in MNLI; without them, a trained model can

		Dataset (Classes, *N*)	Accuracy
Textual Entailment (R1–R5)	1	MNLI (ent/con/neu, 412,349)	F1=82.3
	2	AntSyn (ent/con, 15,632)	F1=90.2
	3	Neu50K (neu, 50,000)	R=97.5
	4	**MicroAvg (ent/con/neu, 477,981)**	**F1=84.7**
Sentiment Classification (R4–R5)	5	SemEval17 (pos/neg/neu, 20,632)	F1=64.5
	6	Dong (pos/neg/neu, 6,940)	F1=71.4
	7	Mitchell (pos/neg/neu, 3,288)	F1=62.5
	8	Bakliwal (pos/neg/neu, 2,624)	F1=69.7
	9	Norm (pos/neg, 632)	F1=100.0
	10	**MicroAvg (pos/neg/neu, 34,116)**	**F1=69.2**
Causality (R6–R9)	11	PDTB (cause/else, 14,224)	F1=68.1
	12	PDTB-R (cause/else 1,791)	F1=75.7
	13	BECauSE (cause/obstruct, 1,542)	F1=46.1
	14	BECauSE-R (else, 1,542)	R=86.5
	15	CoNet (cause, 50,420)	R=88.6
	16	CoNet-R (else, 50,420)	R=91.7
	17	WIQA (cause/obstruct, 31,630)	F1=88.2
	18	WIQA-P (else, 31,630)	R=90.2
	19	**MicroAvg (cause/obstr/else, 183,119)**	**F1=87.7**
Normative Relation (R10–R13)	20	JustType (conseq/norm, 1,580)	F1=90.2
	21	ConseqSenti (pos/neg, 824)	F1=71.8
	22	NormType (adv/opp, 758)	F1=91.1
	23	*RC*-Rel (consist/contra/else, 1,924)	F1=70.1

Table 6.2: F1-scores and recall of modules.

hardly predict the neutral relation between words. The accuracy for each dataset is in Table 6.2 rows 1–4.

6.4.2 Target-Based Sentiment Classification

A sentiment classifier is for rules about sentiment coherence (R4–R5). Given a pair of texts T_1 and T_2, it computes the probability of whether T_1 has positive, negative, or neutral sentiment toward T_2.

Our training data include five datasets for target-based sentiment classification: SemEval17 (Rosenthal et al., 2017), entities (Dong et al., 2014), open domain (Mitchell et al., 2013), Irish politics (Bakliwal et al., 2013), and our annotations of positive/negative norms toward norm targets (§6.5.1). These annotations highly improve classification of sentiments expressed through advocacy and opposition in normative statements. Pretraining on general sentiment resources—subjectivity lexicon (Wilson et al., 2005) and sentiment140 (Go et al., 2009)—also helps (Table

Corpus	Corpus-Specific Labels	Our Label (N)
PDTB	Temporal.Asynchronous	Cause (1,255)
	Temporal.Synchrnonous	Cause (536)
	Comparison, Expansion	Else (12,433)
PDTB-R†	Temporal.Asynchronous	Else (536)
	Temporal.Synchronous	Cause (1,255)
BECauSE	Promote	Cause (1,417)
	Inhibit	Obstruct (142)
BECauSE-R†	Promote, Inhibit	Else (1,613)
WIQA	RESULTS_IN	Cause (12,652)
	NOT_RESULTS_IN	Obstruct (18,978)
WIQA-P‡	RESULSTS_IN, NOT_RESULTS_IN	Else (31,630)
ConceptNet	Causes, CausesDesire, HasFirstSubevent, HasLastSubevent, HasPrerequisite	Cause (50,420)
ConceptNet-R†	Causes, CausesDesire, HasFirstSubevent, HasLastSubevent, HasPrerequisite	Else (50,420)

Table 6.3: Mapping between corpus-specific labels and our labels for the causality module. †The order of two input texts are reversed. ‡The second input text is replaced with a random text in the corpus.

6.2 rows 5–10).

6.4.3 Causality

A causality module is used for rules regarding causal relations (R6–R9). Given an input pair of texts T_1 and T_2, it computes the probability of whether T_1 causes T_2, obstructs T_2, or neither.

Our training data include four datasets about causal and temporal relations between event texts. PDTB 3.0 (Webber et al., 2006) is WSJ articles annotated with four high-level discourse relations, and we map the sub-relations of 'Temporal' to our classes[7]. BECauSE 2.0 (Dunietz et al., 2017) is news articles annotated with linguistically marked causality. WIQA (Tandon et al., 2019) is scientific event texts annotated with causality between events. ConceptNet (Speer et al., 2017) is a knowledge graph between phrases, and relations about causality are mapped to our classes. To prevent overfitting to corpus-specific characteristics (e.g., genre, text length), we add adversarial data by swapping two input texts (PDTB-R, BECauSE-R, ConceptNet-R) or pairing random texts (WIQA-P). The mapping between corpus-specific labels and ours is in Table 6.3, and the module

[7]We use explicit relations only for pretraining, since they often capture linguistically marked, rather than true, relations between events. We also exclude the Contingency relations as causal and non-causal relations (e.g., justification) are mixed.

accuracy in Table 6.2 rows 11–19.

6.4.4 Normative Relation

All the modules here are trained on our annotations of normative argumentation schemes (§6.5).

$P(S$ **is a consequence / norm**$)$ **(R10–R13):** Given a statement, one module computes the probability that it is a consequence, and another module the probability of a norm. Both modules are trained on all claims and statements in our annotations, where all claims are naturally norms, and each statement is annotated as either norm or consequence (Table 6.2 row 20).

$P(Q$ **is positive / negative**$)$ **(R10–R11):** Given a statement assumed to be a consequence, this module computes the probability of whether it is positive or negative. It is trained on all statements annotated as consequence (Table 6.2 row 21).

$P(S$ **advocates / opposes**$)$ **(R12–R13):** Given a statement assumed to be a norm, this module computes the probability of whether it is advocacy or opposition. It is trained on all claims, plus statements annotated as norm (Table 6.2 row 22).

$P(R$ **consistent / contrary to** $C)$ **(R10–R13):** For a pair of S and C, the module computes the probability of whether R (the norm target or the source of consequence in S) and C's stance are consistent, contrary, or else. In our annotations, R and C are 'consistent' if both (1a and 3a in Figure 6.1) are advocacy or opposition, and 'contrary' otherwise. To avoid overpredicting the two classes, we add negative data by pairing C with a random statement in the annotations. The module is pretrained on MNLI and AntSyn (Table 6.2 row 23).

6.5 Annotation of Normative Argumentation Schemes

In this section, we discuss our annotation of the argumentation schemes *argument from consequences* and *practical reasoning* (Figure 6.1). The resulting annotations are used to train the modules in §6.4.4 which compute the probability terms in R10–R13.

For each pair of normative claim C and statement S, we annotate the following information: (1a) Whether C advocates for or opposes its norm target, and (1b) the norm target T (Figure 6.1 TASK 1); (2a) Whether S uses a norm, consequence, or property for justification, and (2b) the justification J (Figure 6.1 TASK 2); (3a) Whether J's focus is on advocating for T or opposing T, and (3b) whether J is positive or negative (Figure 6.1 TASK 3).[8]

Our annotation schema is richer than existing ones (Lawrence and Reed, 2016; Reisert et al., 2018). Due to the increased complexity, however, our annotation is split into three pipelined

[8]This annotation schema provides enough information for the classifiers in §6.4.4. $P(S$ is a consequence / norm) is from (2a), and both $P(Q$ is positive / negative) and $P(S$ advocates / opposes) are from (3b). $P(R$ consistent / contrary to $C)$ can be obtained by combining (1a) and (3a): 'consistent' if both advocate or both oppose, and 'contrary' otherwise.

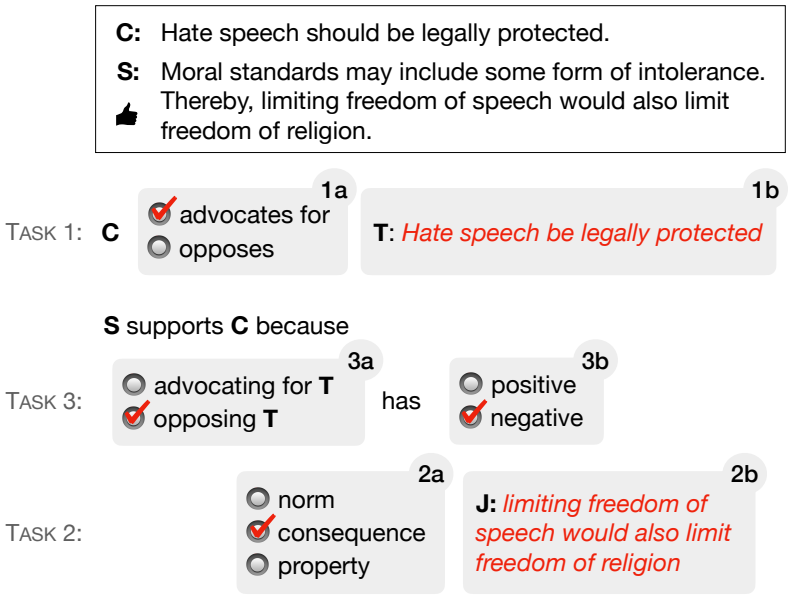


Figure 6.1: Example annotations (checks and italic) of the normative argumentation schemes. It depends on the argument whether *S* supports or attacks *C*.

tasks. For this annotation, we randomly sampled 1,000 arguments from Kialo whose claims are normative (see §6.6 and Table 6.4 for details).

6.5.1 Task 1. Norm Type/Target of Claim

For each *C*, we annotate: (1a) the norm type—advocate, oppose, or neither—toward its norm target; and (1b) the norm target *T*. Advocacy is often expressed as "should/need T", whereas opposition as "should not T", "T should be banned"; 'neither' is noise (2.8%) to be discarded. *T* is annotated by rearranging words in *C* (Figure 6.1 TASK 1).

There are 671 unique claims in the annotation set. The first author of this paper wrote an initial manual and trained two undergraduate students majoring in economics, while resolving disagreements through discussion and revising the manual. In order to verify that the annotation can be conducted systematically, we measured inter-annotator agreement (IAA) on 200 held-out claims. The annotation of norm types achieved Krippendorff's α of 0.81 with 95% CI=(0.74, 0.88) (with the bootstrap). To measure IAA for annotation of *T*, we first aligned words between each annotation and the claim[9], obtaining a binary label for each word in the claim (1 if included in the annotation). As a result, we obtained two sequences of binary labels of the same length from the two annotators and compared them, achieving an F1-score of 0.89 with 95% CI=(0.86, 0.91). The high α and F1-score show the validity of the annotations and annotation manual. All disagreements were resolved through discussion afterward.[10]

[9]We iteratively matched and excluded the longest common substring.

[10]These annotations are used for the sentiment classifiers in §6.4.2 too. For example, "the lottery should be banned" is taken to express negative sentiment toward the lottery. Such examples are underrepresented in sentiment datasets,

6.5.2 Task 2. Justification Type of Premise

For each pair of C and S, we annotate: (2a) the justification type of S—norm, consequence, property, or else; and (2b) the justification J. The justification types are defined as follows:

- **Norm:** J states that some situation or action should be achieved (practical reasoning).
- **Consequence:** J states a potential or past outcome (argument from consequences).
- **Property:** J states a property that (dis)qualifies C's stance (argument from consequence).

The difference between consequence and property is whether the focus is on extrinsic outcomes or intrinsic properties, such as feasibility, moral values, and character (e.g., "Alex shouldn't be the team leader because he is dishonest"). We consider both as argument from consequences because property-based justification has almost the same logic as consequence-based justification. The 'else' type is rare (3.4%) and discarded after the annotation.

The process of annotation and IAA measurement is the same as Task 1, except that IAA was measured on 100 held-out arguments due to a need for more training. For justification types, Krippendorff's α is 0.53 with 95% CI=(0.41, 0.65)—moderate agreement. For justification J, the F1-score is 0.85 with 95% CI=(0.80, 0.90). The relatively low IAA for justification types comes from two main sources. First, a distinction between consequence and property is fuzzy by nature, as in "an asset tax is the most fair system of taxing citizens". This difficulty has little impact on our system, however, as both are treated as argument from consequences. If we combine these two categories, Krippendorff's α increases to 0.58 with 95% CI=(0.37, 0.77). Second, some statements contain multiple justifications of different types. If so, we asked the annotators to choose one that they judge to be most important (for training purposes). They sometimes chose different justifications, although they usually annotated the type correctly for the chosen one. Lastly, since the 'else' type is rare, disagreements on it hurt IAA significantly.

6.5.3 Task 3. Justification Logic of Statement

Given C with its norm target T, and S with its justification J, we annotate: (3a) whether the consequence, property, or norm target of J is regarding advocating for T or opposing T; and (3b) whether J is positive or negative. J is positive (negative) if it's a positive (negative) consequence/property or expresses advocacy (opposition).

For instance, in the following argument

> **Claim:** "People should eat whatever they feel like eating."
>
> **Statement:** "There is no reason to deny oneself a pleasure. (norm targets underlined)"

S is a *negative* norm because it is opposing its norm target, by saying "there is no reason to". This norm target is consistent with *opposing* the claim's norm target T, because denying oneself a pleasure is contrary to eating whatever they feel like eating. Some ambiguous cases include:

> **Claim:** "The roles an actor can play should be limited by that actor's race,

resulting in inaccurate sentiment classification for normative statements.

gender, or sexuality."

Attack Statement: "An actor's talent may be more important than getting an exact visual match in some cases."

Here, the statement may be considered positive or negative depending on the perspective of talent or race. In this case, we annotate the overall sentiment of the statement, which in this case is positive reflected in "more important".

This task was easy, so only one annotator worked with the first author. Their agreement measured on 400 heldout arguments is Krippendorff's α of 0.82 with 95% CI=(0.77, 0.86) for positive/negative and 0.78 with 95% CI=(0.72, 0.83) for advocate/oppose.

6.5.4 Analysis of Annotations

We obtained 962 annotated arguments with claims of advocacy (70%) and opposition (30%), and statements of consequence (54%), property (32%), and norm (14%). Supporting statements are more likely to use a positive justification (62%), while attacking statements a negative one (68%), with significant correlations ($\chi^2 = 87, p < .00001$). But 32–38% of the time, they use the opposite sentiment, indicating that sentiment alone cannot determine argumentative relations.

Supporting statements tend to emphasize the positivity of what the claim advocates for (74%) or the negativity of what the claim opposes (66%). While attacking statements often emphasize the negativity of what the claim advocates for (76%), positivity and negativity are equally emphasized (50%) for claims that show opposition.

Statements tend to present a direct indication (consequence or norm) of the claim's stance rather than an indication of the opposite of the claim's stance, while attacking statements are more likely so (68%) than supporting statements (60%) ($\chi^2 = 5.9, p < .05$). Especially when attacking claims that advocate for something, statements tend bring up direct negativity of it (76%).

6.6 Data

Kialo: Our first dataset is from kialo.com, a collaborative argumentation platform covering contentious topics. Users contribute to the discussion of a topic by creating a statement that either supports or attacks an existing statement, resulting in an argumentation tree for each topic. We define an **argument** as a pair of parent and child statements, where the parent is the **claim** and the child is the **support or attack statement**. Each argument is labeled with support or attack by users and is usually self-contained, not relying on external context, anaphora resolution, or discourse markers.

We scraped arguments for 1,417 topics written until Oct 2019, and split into two subsets. **Normative arguments** have normative claims suggesting that a situation or action be brought about, while **non-normative arguments** have non-normative claims. This distinction helps us understand the two types of arguments better. We separated normative and non-normative claims using a BERT classifier trained on the dataset of statement types from Chapter 5 (AUC=98.8%),

		Kialo				Debatepedia		
		Annot-ation	Fit	Val	Test	Fit	Val	Test
Normative	Sup	480	4,621	1,893	6,623	6,598	229	356
	Att	520	5,383	2,124	7,623	4,502	243	351
	Neu	–	9,984	4,000	14,228	–	–	–
Non-normative	Sup	–	4,953	10,135	21,138	3,302	243	178
	Att	–	5,043	9,848	20,197	3,278	253	152
	Neu	–	10,016	20,000	40,947	–	–	–

Table 6.4: Numbers of arguments in datasets.

as binary classification of normative statement or not. A claim is considered normative (non-normative) if the predicted probability is higher than 0.97 (lower than 0.4); claims with probability scores between these thresholds (total 10%) are discarded to reduce noise.

In practice, an argument mining system may also need to identify statements that seem related but do not form any argument. Hence, we add the same number of "neutral arguments" by pairing random statements within the same topic. To avoid paired statements forming a reasonable argument accidentally, we constrain that they be at least 9 statements apart in the argumentation tree, making them unlikely to have any support or attack relation but still topically related to each other.

Among the resulting arguments, 10K are reserved for fitting; 20% or 30% of the rest (depending on the data size) are used for validation and the others for test (Table 6.4). We increase the validity of the test set by manually discarding non-neutral arguments from the neutral set. We also manually inspect the normativity of claims, and if they occur in the fitting or validation sets too, the corresponding arguments are assigned to the correct sets according to the manual judgments. For normative arguments, we set aside 1,000 arguments for annotating argumentation schemes (§6.5).

The data cover the domains economy (13%), family (11%), gender (10%), crime (10%), rights (10%), God (10%), culture (10%), entertainment (7%), and law (7%), as computed by LDA. The average number of words per argument is 49 (45) for normative (non-normative) arguments.

Debatepedia: The second dataset is Debatepedia arguments (Hou and Jochim, 2017). 508 topics are paired with 15K pro and con responses, and we treat each pair as an **argument** and each topic and response as **claim** and **statement**, respectively.

One important issue is that most topics are in question form, either asking if you agree with a stance ("yes" is pro and "no" is con) or asking to choose between two options (the first is pro and the second is con). Since our logical mechanisms do not handle such questions naturally, we convert them to declarative claims as follows. The first type of questions are converted to a claim that proposes the stance (e.g., "Should Marijuana be legalized?" to "Marijuana should be legalized"), and the second type of questions to a claim that prefers the first option (e.g., "Mission

to the Moon or Mars?" to "Mission to the Moon is preferred to Mars"). The first author and an annotator converted all topics independently and then resolved differences.

We split the arguments into **normative** and **non-normative** sets as we do for Kialo, manually verifying all claims. There is no neutral relation. We use the original train, validation, and test splits (Table 6.4). Debatepedia claims are shorter and less diverse than Kialo claims. They focus mostly on valuation, while Kialo includes a lot of factual claims.

6.7 Experiment 1. Probabilistic Soft Logic

The goal here is to see how well the logical mechanisms alone can explain argumentative relations.

6.7.1 PSL Settings

We use the PSL toolkit v2.2.1[11]. The initial weights of the logical rules R1–R13 are set to 1. The importance of the chain rules R14–R17 may be different, so we explore {1, 0.5, 0.1}. The weight of C1 serves as a threshold for the default relation (i.e., neutral for Kialo and attack for Debatepedia), and we explore {0.2, 0.3}; initial weights beyond this range either ignore or overpredict the default relation. C2 is a hard constraint. The optimal weights are selected by the objective value on the validation set (this does not use true relation labels).

6.7.2 Baselines

We consider three baselines. **Random** assigns a relation to each argument randomly. **Sentiment** assigns a relation based on the claim and statement's agreement on sentiment: support if both are positive or negative, attack if they have opposite sentiments, and neutral otherwise. This generally outperforms labeling based on the statement's sentiment only. We compute a sentiment distribution by averaging all target-specific sentiments from our sentiment classifier (§6.4.2). **Textual entailment** assigns support (attack) if the statement entails (contradicts) the claim, and neutral otherwise (Cabrio and Villata, 2012). We use our textual entailment module (§6.4.1). For Debatepedia, we choose between support and attack whichever has a higher probability.

6.7.3 Results

Tables 6.5a and 6.5b summarize the accuracy of all models for Kialo and Debatepedia, respectively. Among the baselines, sentiment (row 2) generally outperforms textual entailment (row 3), both significantly better than random (row 1). Sentiment tends to predict the support and attack relations aggressively, missing many neutral arguments, whereas textual entailment is conservative and misses many support and attack arguments. PSL with all logical rules R1–R13 (row 4) significantly outperforms all the baselines with high margins, and its F1-scores are more balanced across the relations.

[11] https://psl.linqs.org/wiki/2.2.1/

		Normative Arguments						Non-normative Arguments					
		ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$
1	Random	33.5	50.2	32.6	27.8	30.1	39.9	33.4	49.9	32.5	28.7	28.8	40.0
2	Sentiment	40.8	64.1	40.7	40.6	39.1	42.4	43.7	61.1	42.2	40.0	35.2	51.5
3	Text Entail	51.8	61.8	36.7	12.8	30.4	67.0	52.1	62.8	38.6	18.4	31.0	66.4
4	PSL (R1–R13)	54.0‡	73.8‡	52.1‡	47.0‡	43.6‡	65.7‡	57.0‡	76.0‡	54.0‡	50.1‡	42.6‡	69.3‡
5	\ Fact	55.1‡	74.3‡	52.4‡	47.1‡	41.6‡	68.4‡	58.6‡	77.1‡	55.1‡	50.5‡	42.2‡	72.7‡
6	\ Sentiment	**62.1**‡	77.6‡	57.5‡	49.1‡	45.8‡	**77.7**‡	61.3‡	77.8‡	56.7‡	50.3‡	44.1‡	75.7‡
7	\ Causal	54.4‡	73.1‡	52.3‡	45.4‡	45.4‡	66.0‡	57.6‡	76.1‡	54.3‡	48.7‡	43.4‡	70.7‡
8	\ Normative	51.8‡	68.6‡	49.4‡	44.3‡	40.4†	63.4‡	54.7‡	70.3‡	51.4‡	47.0‡	40.3‡	66.8‡
9	\ Sentiment + Chain	61.9‡	**77.7**‡	**57.7**‡	**49.3**‡	**46.2**‡	77.6‡	**61.5**‡	**78.0**‡	**57.2**‡	**50.8**‡	**44.7**‡	**76.1**‡

(a) Kialo

		Normative Arguments					Non-normative Arguments				
		ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$
1	Random	47.7	49.4	50.2	49.0	51.4	53.0	54.6	52.4	53.7	51.1
2	Sentiment	59.3	63.9	59.2	61.0	57.4	69.1	73.4	68.5	72.7	64.3
3	Text Entail	52.2	55.8	49.4	37.6	61.2	70.6	74.2	70.5	69.0	72.0
4	PSL (R1–R13)	**63.9**⋆	**68.3**⋆	**63.9**⋆	63.8	64.0†	73.0	76.1	73.0	74.2	71.7
5	\ Fact	63.4⋆	67.1	63.4⋆	**64.0**	62.7⋆	71.8	75.6	71.7	73.2	70.3
6	\ Sentiment	63.1⋆	67.2	63.1⋆	62.7	63.5⋆	70.9	74.0	70.9	71.6	70.2
7	\ Causal	62.4⋆	66.3	62.1⋆	58.6	**65.5**⋆	**74.5**	**78.7**	**74.5**	**75.4**	**73.6**
8	\ Normative	61.0	64.7	61.0	60.3	61.6⋆	68.2	72.4	68.2	68.3	68.1

(b) Debatepedia

Table 6.5: PSL accuracy. $p < \{0.05^{\star}, 0.01^{\dagger}, 0.001^{\ddagger}\}$ with paired bootstrap compared to the best baseline.

To examine the contribution of each logical mechanism, we conducted ablation tests (rows 5–8). The most contributing mechanism is clearly normative relation across all settings, without which F1-scores drop by 2.6–4.8 points (row 8). This indicates that our operationalization of *argument from consequences* and *practical reasoning* can effectively explain a prevailing mechanism of argumentative relations.

Quite surprisingly, normative relation is highly informative for non-normative arguments as well for both datasets. To understand how this mechanism works for non-normative arguments, we analyzed arguments for which it predicted the correct relations with high probabilities. It turns out that even for non-normative claims, the module often interprets negative sentiment toward a target as an opposition to the target. For the following example,

> **Claim:** "Schooling halts individual development."
>
> **Attack Statement:** "Schooling, if done right, can lead to the development of personal rigor ..."

the module implicitly judges the "schooling" in the claim to be opposed and thus judges the

"schooling" in the statement (the source of consequence) to be contrary to the claim's stance while having positive sentiment (i.e., R11 applies). This behavior is reasonable, considering how advocacy and opposition are naturally mapped to positive and negative norms in our annotation schema (§6.5.3).

The utility of normative relation for non-normative arguments is pronounced for Debatepedia. Excluding this mechanism leads to a significant drop of F1-scores by 4.8 points (Table 6.5b row 8). One possible reason is that most claims in the non-normative set of Debatepedia are valuation; that is, they focus on whether something is good or bad, or preferences between options. As discussed above, valuation is naturally handled by this mechanism. And in such arguments, causal relation may provide only little and noisy signal (row 7).

Sentiment coherence is the second most contributing mechanism. For Kialo, including it in the presence of normative relation is rather disruptive (Table 6.5a row 6). This may be because the two mechanisms capture similar (rather than complementary) information, but sentiment coherence provides inaccurate information conflicting with that captured by normative relation. Without normative relation, however, sentiment coherence contributes substantially more than factual consistency and causal relation by 4.4–5.9 F1-score points (not in the table). For Debatepedia, the contribution of sentiment coherence is clear even in the presence of normative relation (Table 6.5b row 6).

Factual consistency and causal relation have high precision and low recall for the support and attack relations. This explains why their contribution is rather small overall and even obscure for Kialo in the presence of normative relation (Table 6.5a rows 5 & 7). However, without normative relation they contribute 0.7–1.1 F1-score points for Kialo (not in the table). For Debatepedia, factual consistency contributes 0.5–1.3 points (Table 6.5b row 5), and causal relation 1.8 points to normative arguments (row 7). Their contributions show different patterns in a supervised setting, however, as discussed in the next section.

To apply the chain rules (R14–R17) for Kialo, we built 16,328 and 58,851 indirect arguments for the normative and non-normative sets, respectively. Applying them further improves the best performing PSL model (Table 6.5a row 12). It suggests that there is a relational structure among arguments, and structured prediction can reduce noise in independent predictions for individual arguments.

There is a notable difference in the performance of models between the three-class setting (Kialo) and the binary setting (Debate). The binary setting makes the problem easier for the baselines, reducing the performance gap with the logical mechanisms. When three relations are considered, the sentiment baseline and the textual entailment baseline suffer from low recall for the neutral and support/attack relations, respectively. But if an argument is guaranteed to belong to either support or attack, these weaknesses seem to disappear.

6.7.4 Error Analysis

We conduct an error analysis on Kialo.

Normative Relation

For the mechanism of normative relation, we examine misclassifications in normative arguments by focusing on the 50 support arguments and 50 attack arguments with the highest probabilities of the opposite relation. Errors are grouped into four types: *R-C* consistency/contrary (60%), consequence sentiment (16%), ground-truth relation (8%), and else (16%).

Claim-statement consistency: The most common error type is the model's misjudgment of whether the source of consequence or the norm target in the statement is consistent with the claim's stance. Frequently, the model fails to capture simple lexical cues that signal inconsistency, such as "without", "instead of", "the absence of", "deprived of", "failing to", "be replaced", and "exception". The following are some examples, where the underlined words are important signals missed by the model:

> **Claim:** "Governments should provide postal services."
>
> **Support Statement:** "Without universal, affordable postal services, people in rural communities would be unfairly overcharged and/or under-served."

> **Claim:** "Genetic engineering should not be used on humans or animals."
>
> **Attack Statement:** "Failing to use genetic modification on humans will lead to the preventable deaths of countless victims of genetic disease."

We could improve the model by collecting such words and giving more weights or attention to them during training.

Another common error is the failure to capture antonymy relations. The model sometimes misses rather straightforward antonyms, such as "reduction ↔ increased", "multi-use plastic items ↔ single-use plastic items", and "collective presidency ↔ unitary presidency", as in:

> **Claim:** "The USA should adopt an elected collective presidency, like Switzerland's."
>
> **Support Statement:** "The unitary Presidency invests too much power into just one person."

Some antonymy relations require advanced knowledge and are context-dependent, such as "lifetime appointments ↔ fixed 18-year terms", "faith ↔ compatibility and affection", "marketplace of ideas ↔ deliver the best ideas", and "a character witness ↔ government lists", as in:

> **Claim:** "Explicit internet content should be banned."
>
> **Attack Statement:** "Personal freedom must not be denied."

While the classifier was pretrained on textual entailment, that seems not enough; we may need knowledge bases or corpora with a broader coverage for antonymy detection.

There are some occasions where "without it" is implicit in the statement, so the model fails to capture inconsistency, as in:

> **Claim:** "The European Union should become a United States of Europe."

Support Statement: "[Without it,] Too many resources are used by countries in Europe to support their own defense."

Recognizing such cases is very challenging and requires substantial real-world knowledge.

Consequence sentiment: This error type refers to the failure of classifying positivity and negativity of the consequence in the statement. This classification can be challenging without knowing the arguer's true intent, as in:

Claim: "It should be possible to buy and sell nationalities on an open market."

Attack Statement: "The wealthy would create their own super nationality. It would function just like a private club except on a national scale."

The underlined expression may sound positive on its own but is negative in this context.

The majority of failure cases, however, are simpler than the above example. Confusion usually arises when a statement includes both positive and negative words, as in:

Claim: "Student unions should be allowed in schools and universities."

Support Statement: "Student unions could prevent professors from intentionally failing students due to personal factors."

Sentiment classification could be improved by adding more challenging examples in training data.

Ground-truth relation: Sometimes the ground-truth relations from Kialo are questionable, as in:

Claim: "The West should build working autonomous killing machines (AKMs) as quickly as possible."

Support Statement: "More missions and wars could be carried out if the military had AKMs at its disposal."

This statement actually raised a question in Kialo as shown in Figure 6.2. It is common in Kialo that users debate author-marked relations and switch original relations. This indicates the inherent ambiguity of argumentative relations.

Else: Some errors have other sources. The most common case is that the statement includes multiple justifications with different sentiments and different consistency relations with the claim's stance, as in:

Claim: "Single sex schools should be banned."

Support Statement: "Mixed-sex schools can have certain single-sex oriented activities while single-sex schools cannot have mixed-sex activities."

The first part of the statement is a positive consequence and its source is consistent with the claim's stance, whereas the second part is a negative consequence and its source is contrary to the claim's stance. When different modules attend to different parts of the statement, the combination of their judgments can be incorrect even if each judgment is correct. To rectify this problem, we

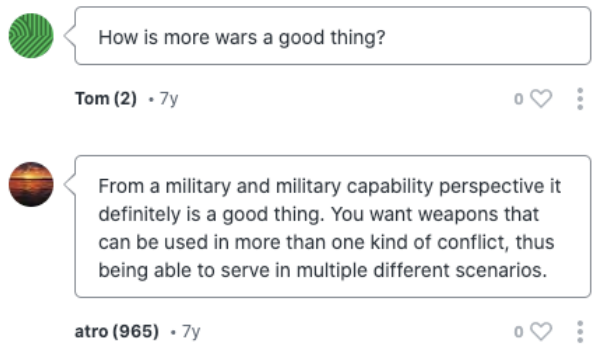


Figure 6.2: Example discussion on ground-truth relations.

could specify a part of the statement available to the modules so their decisions are based on the same ground. We may instead develop a unified module that makes decisions for all components simultaneously.

Other Mechanisms

For the other mechanisms, we examine non-normative arguments that each mechanism judged to have strong signal for a false relation. To that end, for each predicate in R1–R9, we choose the top 20 arguments that have the highest probabilities but were misclassified. Many errors were simply due to the misclassification of the classification modules, which may be rectified by improving the modules' accuracy. But we also found some blind spots of each predicate.

FactEntail: FactEntail often fails to handle concession, elaboration, and scoping. Among the 20 arguments that have the highest probabilities of FactEntail but have the attack relation, 12 have some form of concession, elaboration, and scoping, as in:

> **Claim:** "Fourth wave feminists espouse belief in equality."
>
> **Attack Statement:** "It is belief in equality of outcome not opportunity that fourth wave feminists are espousing with quotas and beneficial bias. (Scoping)"

> **Claim:** "Social media can be used to promote body positivity."
>
> **Attack Statement:** "Successful social media influencers often appropriate the language of body positivity, despite being entirely conventionally attractive. Thus a lot of content which claims to be body-positive in fact reinforces traditional beauty standards. (Elaboration)"

> **Claim:** "Having gender equality in a film does not necessarily make that film good."
>
> **Attack Statement:** "While gender equality in itself may not make for a good film, its absence or the perception that gender-relations have not been taken into account can clearly harm a film. (Concession)"

SentiConsist: SentiConsist shows a similar tendency. a statement can have the same ground of value as the claim without supporting it:

> **Claim:** "Religion as a required course allows free practice of religion on school grounds."
>
> **Attack Statement:** "Free practice only for those believing in the religion being taught."

> **Claim:** "The education of women is the most important objective to improve the overall quality of living"
>
> **Attack Statement:** "Education of both men and women will have greater effects than that of women alone. Both must play a role in improving the quality of life of all of society's members."

In these arguments, the statements attack the claims while expressing the same sentiment toward the same target (underlined).

FactContradict: FactContradict identified many arguments whose ground-truth relations are questionable. Among 20 arguments that have the highest probabilities of FactContradict but whose ground-truth relation is support, 8 indeed seem to have the attack relation. We verified some of them by confirming that their relation had changed to attack in Kialo; the others were deleted from Kialo.

FactConflict: Among 20 arguments that have the highest probabilities of FactConflict but whose ground-truth relation is support, 7 were due to errors from the textual entailment module, and 4 were the failure of Open IE to process negation properly. For another 6 arguments, factual conflicts were in hypotheticals, as in:

> **Claim:** "Skips eslint hook rules if anonymous function is used."
>
> **Support Statement:** "Works properly with eslint hook rules and devtools if named function is used."

The frequency of such cases, along with errors for FactEntail due to concession, indicates that subordinate clauses may need to be handled more carefully for these mechanisms.

Cause and Obstruct: Cause and Obstruct produced many errors due to the mistakes of the causality module. However, there was some difference between cause-to-effect (C2E) reasoning and effect-to-cause (E2C) reasoning. For C2E reasoning, 39 out of 40 errors were purely misclassifications by the causality module. However, such errors are fewer for E2C reasoning (30 of 40). Instead, the Cause predicate captured reasonable temporal relationships between claim and statement that did not lead to the support relation, as in:

> **Claim:** "Society is now less racist and homophobic than it was."
>
> **Attack Statement:** "We should continue the trend of becoming less discriminatory by stopping discrimination against the obese."

On the other hand, the Obstruct predicate identified 4 arguments whose ground-truth relation is questionably support, as in:

> **Claim:** "Ecotourism is not sustainable."
>
> **?Support Statement:** "Ecotourism can help promote the economic growth of a country."

6.8 Experiment 2. Representation Learning

We have established that factual consistency, sentiment coherence, causal relation, and normative relation are working mechanisms in argumentation and can predict argumentative relations to a certain degree without training on relation-labeled data. However, argumentative relations have correlations with other statistics as well, such as thematic associations between a topic and a stance (framing), use of negating words, and sentiment, and supervised models are good at leveraging them, even excessively sometimes (Niven and Kao, 2019; Allaway and McKeown, 2020). Here, we examine if our logical mechanisms can further inform them. We describe a simple but effective representation learning method, followed by baselines and experiment results.

6.8.1 Method

We present a simple method that aims to learn a vector representation of the input argument that is "cognizant of" our logical mechanisms. Our logical mechanisms are based on textual entailment, sentiment classification, causality classification, and four classification tasks for normative relation (§6.4). We call them **logic tasks**. We combine all minibatches across the logic tasks using the same datasets from §6.4 except the heuristically-made negative datasets. Given uncased BERT-base, we add a single classification layer for each logic task and train the model on the minibatches for five epochs in random order. After that, we fine-tune it on our fitting data (Table 6.4), where the input is the concatenation of statement and claim. Training stops if AUC does not increase for 5 epochs on the validation data. We call our model **LogBERT** (the base model does not have to be BERT). To avoid catastrophic forgetting, we tried blending the logic and main tasks (Shnarch et al., 2018) and using regularized fine-tuning (Aghajanyan et al., 2020), but they did not help.

6.8.2 Baselines

The first goal of this experiment is to see if the logical mechanisms improve the predictive power of a model trained without concerning them. Thus, our first baseline is **BERT** fine-tuned on the main task only. This method recently yielded the (near-) best accuracy in argumentative relation classification (Durmus et al., 2019; Reimers et al., 2019).

In order to see the effectiveness of the representation learning method, the next two baselines incorporate logical mechanisms in different ways. **BERT+LX** uses latent cross (Beutel et al., 2018) to directly incorporate predicate values in R1–R13 as features; we use an MLP to encode the predicate values, exploring (i) one hidden layer with D=768 and (ii) no hidden layers. BERT+LX

		Normative Arguments						Non-normative Arguments					
		ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$
1	TGA Net	71.5	88.3	62.2	43.5	54.3	88.7	76.6	90.8	69.8	62.9	53.9	92.5
2	Hybrid Net	66.8	78.2	56.2	42.9	42.4	83.4	71.8	82.2	65.7	55.6	51.4	90.2
3	BERT	79.5	92.4	73.3	60.5	65.2	**94.2**	83.8	94.6	79.2	72.3	68.8	96.6
4	BERT+LX	79.2	92.1	72.7	58.7	65.6*	93.8	83.7	94.6	79.2	70.8	69.9‡	**96.9**‡
5	BERT+MT	79.3	92.6*	73.4	**63.8**‡	63.6	92.7	83.6	94.7	79.2	71.8	69.7‡	96.1
6	LogBERT	**80.0**‡	**92.8**‡	**74.3**‡	63.6‡	**66.2**‡	93.2	**84.3**‡	**95.0**‡	**80.2**‡	**73.1**‡	**71.4**‡	96.1

(a) Kialo

		Normative Arguments					Non-normative Arguments				
		ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$
1	TGA Net	66.1	75.0	65.4	69.8	60.9	66.5	74.3	65.9	70.1	61.7
2	Hybrid Net	67.2	70.1	67.2	68.1	66.3	59.7	62.6	58.8	64.5	53.2
3	BERT	79.1	88.3	79.4	79.8	79.0	80.7	87.6	80.7	81.4	79.9
4	BERT+LX	78.4	88.1	78.4	79.2	77.5	**81.6**	**88.8**	**81.5**	**82.3**	**80.8**
5	BERT+MT	79.6	88.2	79.6	80.0	79.1	77.6	86.3	77.5	78.9	76.0
6	LogBERT	**81.0***	**88.8**	**80.7***	**81.1***	**80.4***	81.2	88.3	80.8	81.7	80.0

(b) Debatepedia

Table 6.6: Accuracy of supervised models. $p < \{0.05^{\star}, 0.001^{\ddagger}\}$ with paired bootstrap compared to BERT.

consistently outperforms a simple MLP without latent cross. **BERT+MT** uses multitask learning to train the main and logic tasks simultaneously.

Lastly, we test two recent models from stance detection and dis/agreement classification. **TGA Net** (Allaway and McKeown, 2020) takes a statement-topic pair and predicts the statement's stance. It encodes the input using BERT and weighs topic tokens based on similarity to other topics. In our task, claims serve as "topics". We use the published implementation, exploring {50, 100, 150, 200} for the number of clusters and increasing the max input size to the BERT input size. **Hybrid Net** (Chen et al., 2018a) takes a quote-response pair and predicts whether the response agrees or disagrees with the quote. It encodes the input using BiLSTM and uses self- and cross-attention between tokens. In our task, claims and statements serve as "quotes" and "responses", respectively.

6.8.3 Results

Tables 6.6a (Kialo) and 6.6b (Debatepedia) summarize the accuracy of each model averaged over 5 runs with random initialization. For non-normative arguments, the causality task is excluded from all models as it consistently hurts them for both datasets.

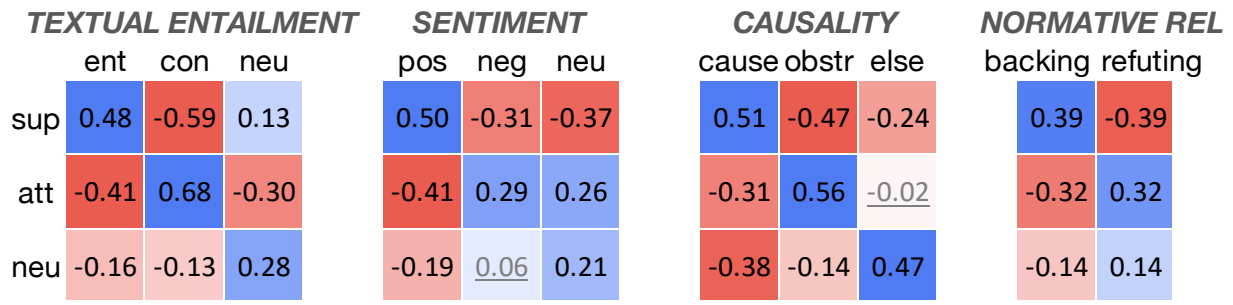
Regarding the baselines, TGA Net (row 1) and Hybrid Net (row 2) underperform BERT (row 3). TGA Net, in the original paper, handles topics that are usually short noun phrases. It weighs input topic tokens based on other similar topics, but this method seems not as effective when topics are replaced with longer and more natural claims. Hybrid Net encodes input text using BiLSTM, whose performance is generally inferior to BERT.

BERT trained only on the main task is competitive (row 3). BERT+LX (row 4), which incorporates predicate values directly as features, is comparable to or slightly underperforms BERT in most cases. We speculate that predicate values are not always accurate, so using their values directly can be noisy. LogBERT (row 6) consistently outperforms all models except for non-normative arguments in Debatepedia (but it still outperforms BERT). While both BERT+MT and LogBERT are trained on the same logic tasks, BERT+MT (row 5) performs consistently worse than LogBERT. The reason is likely that logic tasks have much larger training data than the main task, so the model is not optimized enough for the main task. On the other hand, LogBERT is optimized solely for the main task after learning useful representations from the logic tasks, which seem to lay a good foundation for the main task.

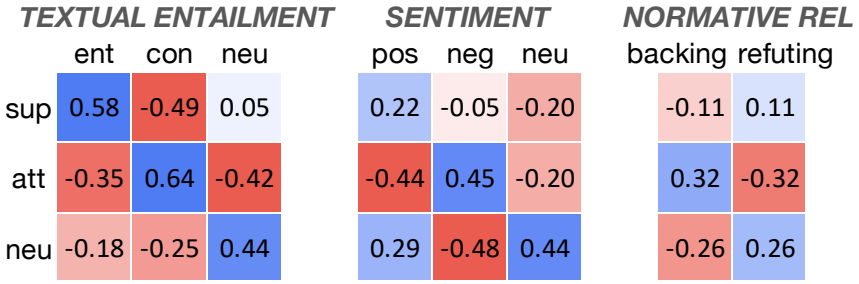
We examined the contribution of each logic task using ablation tests (not shown in the tables). Textual entailment has the strongest contribution across settings, followed by sentiment classification. This contrasts the relatively small contribution of factual consistency in Experiment 1. Moreover, the tasks of normative relation have the smallest contribution for normative arguments and the causality task for non-normative arguments in both datasets. Three of the normative relation tasks take only a statement as input, which is inconsistent with the main task. This inconsistency might cause these tasks to have only small contributions in representation learning. The small contribution of the causality task in both Experiments 1 and 2 suggests large room for improvement in how to effectively operationalize causal relation in argumentation.

To understand how LogBERT makes a connection between the logical relations and argumentative relations, we analyze "difficult" arguments in Kialo that BERT misclassified but LogBERT classified correctly. If the correct decisions by LogBERT were truly informed by its logic-awareness, the decisions may have correlations with (internal) decisions for the logic tasks as well, e.g., between attack and textual contradiction. Figure 6.3 shows the correlation coefficients between the probabilities of argumentative relations and those of the individual classes of the logic tasks, computed simultaneously by LogBERT (using the pretrained classification layers for the logic tasks). For sentiment, the second text of an input pair is the sentiment target, so we can interpret each class roughly as the statement's sentiment toward the claim. For normative relation, we computed the probabilities of backing (R10+R12) and refuting (R11+R13).

The correlations are intuitive. The support relation is positively correlated with textual entailment, positive sentiment, 'cause' of causality, and 'backing' of normative relation, whereas the attack relation is positively correlated with textual contradiction, negative sentiment, 'obstruct' of causality, and 'refuting' of normative relation. The neutral relation is positively correlated with the neutral classes of the logic tasks. The only exception is the normative relation for non-normative arguments. A possible reason is that most claims in non-normative arguments do not follow the typical form of normative claims, and that might affect how the tasks of normative relation contribute for these arguments. We leave a more thorough analysis to future work.



(a) Normative arguments.



(b) Non-normative arguments.

Figure 6.3: Pearson correlation coefficients between argumentative relations and logic tasks from LogBERT. All but underlined values have $p < 0.0001$.

LogBERT's predictive power comes from its representation of arguments that makes strong correlations between the logical relations and argumentative relations. Though LogBERT uses these correlations, it does not necessarily *derive* argumentative relations *from* the logic rules. It is still a black-box model with some insightful explainability.

6.9 Conclusion

We examined four types of logical and theory-informed mechanisms in argumentative relations: factual consistency, sentiment coherence, causal relation, and normative relation. We operationalized these mechanisms through machine-learned modules and probabilistic soft logic, to find the optimal argumentative relations between statements. To operationalize normative relation, we also built rich annotation schema and data for the argumentation schemes *argument from consequences* and *practical reasoning*.

Evaluation on arguments from Kialo and Debatepedia revealed the importance of these mechanisms in argumentation, especially normative relation and sentiment coherence. Their utility was further verified in a supervised setting via our representation learning method. Our model learns argument representations that make strong correlations between logical relations and argumentative relations in intuitive ways. Textual entailment was found to be particularly helpful in the supervised setting.

Some promising future directions are to probe fine-tuned BERT to see if it naturally learns

logical mechanisms and to improve PSL with more rules.

It may be worth discussing our view on the difference between argumentative relations and textual entailment. Argumentative relations (support/attack/neutral) and the relations in textual entailment, or often called natural language inference, (entail/contradict/neutral) are very similar as they stand in current NLP. We see argumentative relations as a broader and looser concept than textual entailment, at least in terms of how data are typically annotated. When annotating textual entailment between premise and hypothesis, we typically ask whether the hypothesis is *definitely* true or *definitely* false given the premise; otherwise, they are considered to have the neutral relation. This definition is stricter than argumentative relations, where we typically consider whether the hypothesis is *likely* to be true or false given the premise. This looser definition has to do with the view in informal logic that daily arguments are defeasible; a hypothesis that is currently supported by a premise can be defeated given additional premises that suggest otherwise. That is, the goal of argumentative relations is not to decide whether the hypothesis is deductively derived from the premise as in textual entailment, but instead those relations are determined by the degree to which they are accepted by general listeners based on available information.

Causality and normative relation, in this sense, are closer to argumentative relations than textual entailment, because they do not deductively validate or invalidate the hypothesis. For instance, although the fact that a company had high net incomes can support the prediction that its stock price will rise, the former is not generally seen as entailing the latter because there could be other factors that obstruct the rise of the stock price, such as unethical affairs of the company.

The four mechanisms we considered—textual consistency, sentiment coherence, causal relation, and normative relation—cover more than half the instances of argumentation schemes in the 2016 U.S. presidential debates (Visser et al., 2020). Adding additional mechanisms may capture long-tail schemes, such as argumentation from expert opinion or argument from analogy. Some of those schemes have characteristic lexical features and are relatively easy to capture; for example, argument from expert opinion is closely related to reported speech, and identifying the content and source in reported speech can be automatically conducted robustly §2.5.3. Some schemes like argument from analogy are not straightforward to capture and thus require semantically richer models. Such schemes may be more common in certain domains, for instance, argument from expert opinion in legal argumentation and argument from analogy in mental health-related argumentation. Hence, capturing long-tail schemes would be a good direction in order to handle diverse domains of argumentation.

Part III

Counter-Argumentation

In Part II, we focused on argumentative relations between statements. We assumed that all statement pairs were given and we only classified their relations. In Part III, we take a step further and investigate generating counterarguments in ongoing argumentation. Counterargument generation may help people make better decisions informed by counterevidence to their arguments and may be used for practical applications, such as feedback generation and fact verification.
We see counterargument generation as a three-step process: given an argument, (i) detect attackable sentences, (ii) find valid counterevidence to them, and (iii) combine the counterevidence as a coherent and fluent argument. This thesis covers the first two steps. In Chapter 7, we present two computational models for detecting attackable sentences in arguments, one based on neural representations of sentences and the other based on hand-crafted features that represent various characteristics of sentences. In Chapter 8, we build a system that retrieves counterevidence to a given statement, from various sources. We enhance the system to handle nontrivial cases that require causality- and example-based reasoning, by incorporating relevant knowledge graphs.

Chapter 7

Detecting Attackable Sentences

Finding attackable sentences in an argument is the first step toward successful counter-argumentation. When attacking an argument in deliberative dialogue, for example, it is crucial to identify important parts in the reasoning of the argument that are key to impacting the arguer's viewpoint. In the field of computational argumentation, it has been understudied what makes certain sentences attackable, how to identify them, and how addressing them affects the persuasiveness of the counterargument. In this chapter, we present large-scale studies to tackle these problems. Specifically, we present two approaches to modeling attackable sentences in arguments in persuasive dialogue. The first approach uses a neural method for jointly modeling sentence attackability and persuasion success. The second approach uses a semantic method for quantifying different characteristics of sentences and analyzing differences in the characteristics between attackable sentences and non-attackable sentences. Computationally identified attackability information would help people make persuasive refutations and strengthen an argument by solidifying potentially attackable points.

7.1 Introduction

Effectively refuting an argument is an important skill in persuasion dialogue, and the first step is to find appropriate points to attack in the argument. Prior work in NLP has studied various aspects of argument quality (Wachsmuth et al., 2017a; Habernal and Gurevych, 2016a), such as clarity and topical relevance. But these studies mainly concern an argument's *overall* quality, instead of providing guidance of which parts of the argument can be effective targets for attacks. There are also studies on counterargument generation (Hua et al., 2019; Wachsmuth et al., 2018b), but most of them focus on making counterarguments toward a *main claim*, instead of refuting the reasoning of another argument. Accordingly, we have a limited understanding of how to detect attackable points in an argument, what characteristics they have, and how attacking them affects persuasion.

To motivate these problems, example arguments are shown in Figure 7.1. The attacked argument at the top presents a negative view on DNA tests, along with reasoning and experiences that justify the view. Challenger 1 attacks this argument by addressing the argument's general statement and

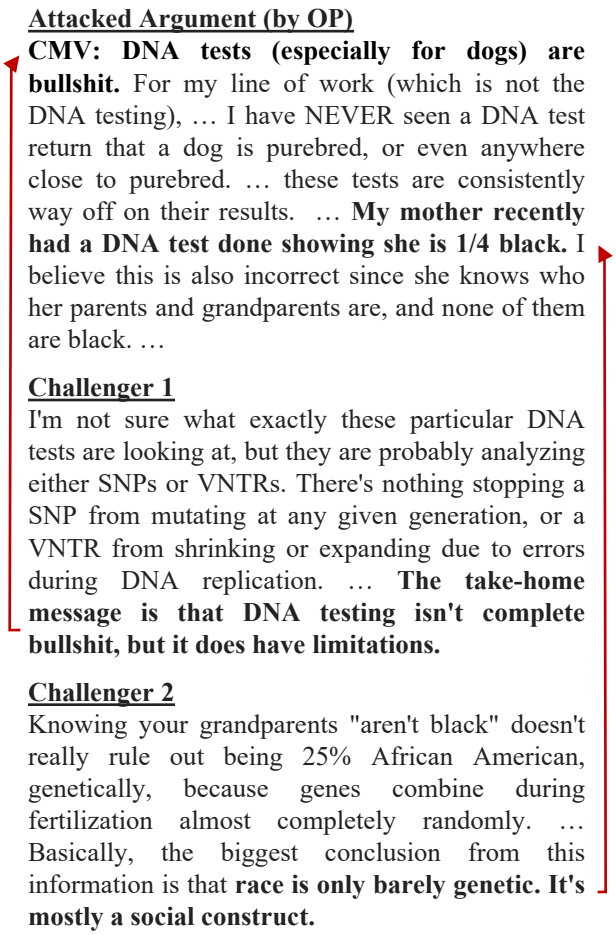


Figure 7.1: An example discussion from the ChangeMyView subreddit. The first argument is attacked by the following two challengers. The arrows show which sentences of the attacked arguments are attacked.

providing a new fact. This challenger successfully changed the attacked arguer's view. On the other hand, Challenger 2 attacks the race issues and failed to persuade the attacked arguer. This example suggests that some points in an argument are more attackable than others, and effectively attacking those points could increase the chance of successful persuasion.

To tackle these problems, we present two models in this chapter. In Section 7.3, we present a neural approach that jointly models sentence attackability and persuasion success. Given two arguments, one attacked and one attacking, this model is built on the assumption that the model attends less to non-attackable sentences than to attackable sentences when predicting if the attacking argument would successfully refute the attacked argument. Specifically, the model first encodes the attackability score of each sentence in the attacked argument via an attention mechanism, and then leverages these scores to predict the success or failure of persuasion. For instance, in the attacked argument in Figure 7.1, attackable sentences (e.g., "CMV: DNA tests

(especially for dogs) ...") are assigned high attention weights, and less attackable sentences (e.g., "My mother recently had a DNA test done ...") are assigned low attention weights ideally. As a result, when a challenger attacks this argument, a high degree of interaction around the less attackable sentences does not contribute much to the persuasion outcome. This neural model is trained end-to-end, taking a pair of arguments as input and predicting a persuasion success label as output. Each sentence's attackability score is reflected in the attention weights computed by the intermediate attention layer.

Our second approach in Section 7.4 is more interpretable and explicitly models various semantic properties of sentences. We assume that if the attacking argument attacks certain sentences in the attacked argument and successfully refutes the argument, the attacked sentences can be considered attackable, i.e., worth attacking. In contrast, if certain sentences are attacked but that does not lead to successful refutation, these sentences may be considered less attackable, i.e., less worth attacking. Hence, in this approach, we take sentences in attacked arguments that are directly quoted by attacking arguments, quantify various semantic characteristics of these sentences that are relevant to attackability, and analyze how these characteristics are different between successfully attacked sentences and unsuccessfully attacked sentences. For instance, as reflected in Figure 7.1 and confirmed in our experiment that comes later, sentences that describe a personal story (e.g., "My mother recently had a DNA test done...") are not much attackable.

The studies are conducted on online discussion from the ChangeMyView (CMV) subreddit (Section 4.2.3). The discussions consist of many pairs of attacked arguments and attacking arguments, along with the labels of whether each attacking argument successfully changed the attacked arguer's viewpoint. Although there is no direct information about sentence attackability, we use the persuasion success labels and quote information to analyze sentence attackability indirectly. One benefit of this approach is that we can avoid relying on annotators' intuitions on sentence attackability, which can be subjective; instead, we use the self-report of argumentation participants to reflect on which sentences are actually worth attacking. While both studies are based on CMV discussions, they use different subsets of discussions, which will be explained more in detail in each section.

7.2 Related Work

The strength of an argument is a long-studied topic, dating back to Aristotle and Kennedy (2007), who suggested three aspects of argument persuasiveness: ethos (the arguer's credibility), logos (logic), and pathos (appeal to the hearer's emotion). More recently, Wachsmuth et al. (2017b) summarized various aspects of argument quality studied in argumentation theory and NLP, such as clarity, relevance, arrangement. Some researchers took empirical approaches and collected argument evaluation criteria from human evaluators (Habernal and Gurevych, 2016a; Wachsmuth et al., 2017a). By adopting some of these aspects, computational models have been proposed to automatically evaluate argument quality in various settings, such as essays (Ke et al., 2019), online comments (Gu et al., 2018), and pairwise ranking (Habernal and Gurevych, 2016b). While these taxonomies help us understand and evaluate the quality of an argument as a whole, little empirical analysis has been conducted in terms of what to attack in an argument to persuade the

arguer.

What can be attacked in an argument has been studied more in argumentation theory. Particularly, Walton et al. (2008) present argumentation schemes and critical questions (CQs). Argument schemes are reasoning types commonly used in daily argumentation. For instance, the scheme of *argument from cause to effect* has the following structure:

> **Premises:** Generally, if *A* occurs, *B* will occur. In this case, A occurs.
> **Conclusion:** *B* will occur.

Each scheme is associated with a set of CQs for judging the argument to be good or fallacious. CQs for the above scheme include "How strong is the causal generalization?" and "Are there other factors that interfere with the causal effect?" Unlike the general argument quality described in the previous paragraph, CQs serve as an evaluation tool that specify local attackable points in an argument. They have been adopted as a framework for grading essays (Song et al., 2017) and teaching argumentation skills (Nussbaum et al., 2018). In our paper, some of the sentence characteristics we consider are informed by argumentation schemes and CQs.

NLP researchers have widely studied the effectiveness of counterarguments on persuasion (Tan et al., 2016; Cano-Basave and He, 2016; Wei et al., 2016; Wang et al., 2017; Morio et al., 2019) and how to generate counterarguments (Hua et al., 2019; Wachsmuth et al., 2018b). Most of the work focuses on the characteristics of counterarguments with respect to topics and styles, without consideration of what points to attack.

7.3 Neural Modeling of Attackability and Persuasion

In this section, we present a neural model for modeling argumentative dialogue that explicitly models the interplay between an attacked argument and an attacking argument. The model encodes sentence attackability in the process of predicting if the attacking argument successfully changes the attacked arguer's viewpoint. The model has two main components: (1) *sentence attackability scoring*, an attention layer that computes attackability score of each sentence in the attacked argument, and (2) *interaction encoding*, which encodes the interaction between the sentences in the attacked argument and those in the attacking argument.

The first component, sentence attackability scoring, aims to identify important sentences in the attacked argument that are key to impacting their viewpoint. The intuition behind our model is that addressing certain points of an argument often has little impact in changing the arguer's view, even if the arguer realizes the reasoning is flawed. On the other hand, there are certain points in the argument that are more open to debate, and thus, it is reasonable for the model to learn to attend to those attackable points, attacking which leads to a better chance to change the viewpoint of the attacked arguer. As a result, attention weights learned by the model may reflect the degrees of attackability of individual sentences in the attacked argument.

The second component, interaction encoding, aims to identify the connection between the sentences in the attacked argument and those in the attacking argument. Meaningful interaction in argumentation may include agreement/disagreement, topic relevance, or logical implication.

Our model encodes the interaction between every sentence pair of the attacked argument and attacking argument, and computes interaction embeddings. These embeddings are then aggregated for predicting the success or failure of persuasion. Intuitively, not all interactions are equally important; rather, interactions with attackable sentences are more critical. Thus, in our complete model, the interaction embeddings are weighted by the attackability scores of sentences computed in the first component.

Using this model, we hope to better understand if computational models can identify attackable sentences and what properties constitute attackability, if the joint modeling of sentence attackability and interaction encoding helps to predict persuasion results better, and what kinds of interactions between arguments are captured by the model.

This study is also situated in modeling knowledge co-construction through persuasive argumentation. Through engagement in argumentative dialogue, interlocutors present arguments with the goals of contributing to the joint construction of knowledge. Modeling this process requires understanding of both the substance of viewpoints and how the substance of an argument connects with what it is arguing against. Prior work on argumentation in the NLP community, however, has focused mainly on the first goal and has often reduced the concept of a viewpoint as a discrete side (e.g., pro vs against, or liberal vs conservative), missing more nuanced and complex details of viewpoints. In addition, while the strength of the argument and the side it represents have been addressed relatively often, the dialogical aspects of argumentation have received less attention.

Argumentation theories have identified important dialogical aspects of persuasive argumentation, which motivate our attempt to model the interaction between arguments. Persuasive arguments build on the hearer's accepted premises (Walton, 2008) and appeal to emotion effectively (Aristotle and Kennedy, 2007). From a challenger's perspective, effective strategies for these factors could be derived from the hearer's background and reasoning. On the other hand, non-persuasive arguments may commit fallacies, such as contradicting the hearer's accepted premises, diverting the discussion from the relevant and salient points in the original argument, failing to address the issues in question, misrepresenting the hearer's reasoning, and shifting the burden of proof to the hearer by asking a question (Walton, 2008). These fallacies can be identified only when we can effectively model how the attacking argument argues *in relation to* the attacked argument.

While prior work in the NLP community has studied argumentation, such as predicting debate winners (Potash and Rumshisky, 2017; Zhang et al., 2016; Wang et al., 2017; Prabhakaran et al., 2013) and winning negotiation games (Keizer et al., 2017), our approach addresses a different angle: predicting whether an attacking argument will successfully impact the attacked arguer's view. Some prior work investigates factors that underlie viewpoint changes (Tan et al., 2016; Lukin et al., 2017; Hidey et al., 2017; Wei et al., 2016), but none target our task of identifying the specific arguments that impact an arguer's view.

Changing someone's view depends highly on argumentation quality, which has been the focus of much prior work. Wachsmuth et al. (2017b) reviewed theories of argumentation quality assessment and suggested a unified framework. Prior research has focused mainly on the presentation of an argument and some aspects in this framework without considering the attacked arguer's reasoning. Specific examples include politeness, sentiment (Tan et al., 2016; Wei et al., 2016),

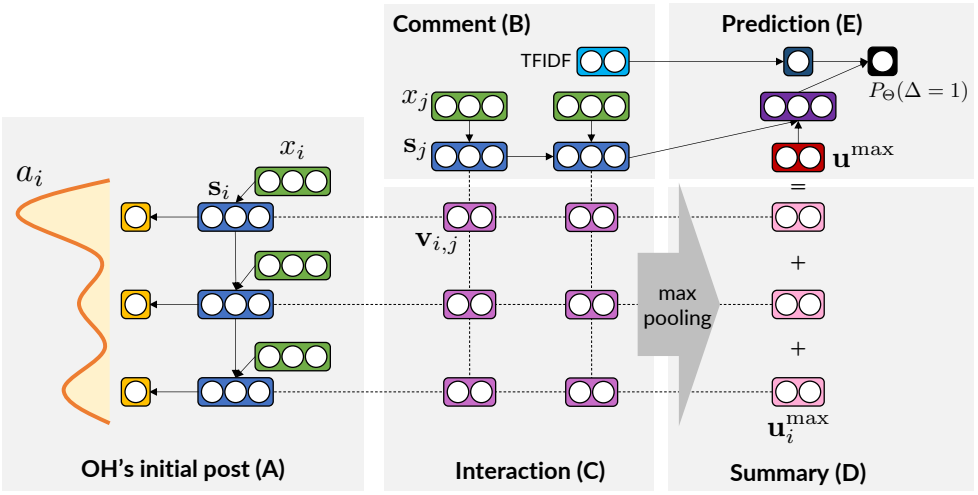


Figure 7.2: Architecture of Attentive Interaction Model.

grammaticality, factuality, topic-relatedness (Habernal and Gurevych, 2016b), argument structure (Niculae et al., 2017), topics (Wang et al., 2017), and argumentative strategies (e.g., anecdote, testimony, statistics) (Al Khatib et al., 2017). Some of these aspects have been used as features to predict debate winners (Wang et al., 2017) and view changes (Tan et al., 2016).

The persuasiveness of an attacking argument, however, is highly related to the attacked reasoning and how the argument connects with it. Nonetheless, research on this relationship is quite limited in the NLP community. Existing work uses word overlap between attacked and attacking arguments as a feature in predicting the success or failure of persuasion (Tan et al., 2016). Some studies examined the relationship between the attacked arguer's personality traits and receptivity to arguments with different topics (Ding and Pan, 2016) or degrees of sentiment (Lukin et al., 2017).

Our approach in this section is highly relevant to the task by Tan et al. (2016) that predicts the success or failure of persuasion in CMV discussions. They examined various stylistic features (sentiment, hedging, question marks, etc.) and word overlap features to identify discussions that impacted the attacked arguer's view. However, our task is different from theirs in that they made predictions on the challenger's initial argument (comment) only, while we did so for the challenger's all arguments (comments). Our task is more challenging because challengers' arguments that come later in a discussion have a less direct connection to the attacked arguer's original argument. Another challenge is the extreme skew in class distribution in our setting; Tan et al. (2016) ensured a balance between the positive and negative classes (i.e., persuasion success and failure).

7.3.1 Attentive Interaction Model

Suppose there is a pair of an attacked argument and an attacking argument. We call the arguer of the attacked argument the **OP** (original poster), and the arguer of the attacking argument the **challenger**. Each pair has an outcome label Δ set to 1 if the attacking argument successfully changes the OP's viewpoint, and 0 otherwise.

Our **Attentive Interaction Model** predicts the probability of the attacking argument changing the OP's original view, $P(\Delta = 1)$. The architecture of the model (Figure 7.2) consists of (i) computing the attackability scores of the sentences in the OP's argument, (ii) embedding the interactions between every sentence in the attacked argument and the attacking argument, (iii) summarizing the interactions weighted by the attackability scores of OP sentences, and (iv) predicting $P(\Delta = 1)$.

The main idea of this model is the architecture for capturing interactions around attackable sentences, rather than methods for measuring specific argumentation-related features (e.g., agreement/disagreement, contraction, attackability, etc.). To better measure these features, we need much richer information than the dataset provides (discussion text and Δs). Therefore, our architecture is not to replace prior work on argumentation features, but rather to complement it at a higher, architectural level that can potentially integrate various features. Moreover, our architecture serves as a lens for analyzing attackable sentences in attacked arguments and their interactions with attacking arguments.

Formal notations (Figure 7.2 (A) and (B)): Denote the attacked argument by $d^O = (x_1^O, ..., x_{M^O}^O)$, where x_i is the ith sentence, and M^O is the number of sentences. The sentences are encoded via an RNN, yielding a hidden state for the ith sentence $\mathbf{s}_i^O \in \mathbb{R}^{D^S}$, where D^S is the dimensionality of the hidden states. Similarly, for the attacking argument $d^C = (x_1^C, ..., x_{M^C}^C)$, hidden states of the sentences $\mathbf{s}_j^C, j = 1, \cdots, M^C$, are computed.

Sentence attackability scoring (Figure 7.2 (A)): Given the sentences of the attacked argument, the model computes the attackability measure of the ith sentence $g(\mathbf{s}_i^O) \in \mathbb{R}^1$ (e.g., using a feedforward neural network). From this measure, the attention weight of the sentence is calculated as

$$a_i = \frac{\exp g(\mathbf{s}_i^O)}{\sum_{i'=1}^{M^O} \exp g(\mathbf{s}_{i'}^O)}.$$

Interaction encoding (Figure 7.2 (C)): The model computes the interaction embedding of every pair of the attacked argument's ith sentence and the attacking argument's jth sentence,

$$\mathbf{v}_{i,j} = \mathbf{h}(\mathbf{s}_i^O, \mathbf{s}_j^C) \in \mathbb{R}^{D^I},$$

where D^I is the dimensionality of interaction embeddings, and $\mathbf{h}$ is an interaction function between two sentence embeddings. $\mathbf{h}$ can be a simple inner product (in which case $D^I = 1$), a feedforward neural network, or a more complex network. Ideally, each dimension of $\mathbf{v}_{i,j}$ indicates a particular type of interaction between the pair of sentences.

Interaction summary (Figure 7.2 (D)): Next, for each sentence in the attacked argument, the model summarizes what types of meaningful interaction occur with the challenger's sentences. That is, given all interaction embeddings for the OP's ith sentence, $\mathbf{v}_{i,1}, \cdots, \mathbf{v}_{i,M^C}$, the model conducts max pooling for each dimension,

$$\mathbf{u}_i^{\max} = \left(\max_j(\mathbf{v}_{i,j,1}), \cdots, \max_j(\mathbf{v}_{i,j,D^I}) \right),$$

where $\mathbf{v}_{i,j,k}$ is the kth dimension of $\mathbf{v}_{i,j}$ and $\mathbf{u}_i^{\max} \in \mathbb{R}^{D^I}$. Intuitively, max pooling is to capture the existence of an interaction and its highest intensity for each of the OP's sentences—the interaction does not have to occur in all sentences of the attacking argument. Since we have different degrees of interest in the interactions in different parts of the OP's post, we take the attention-weighted sum of $\mathbf{u}_i^{\max}$ to obtain the final summary vector

$$\mathbf{u}^{\max} = \sum_{i=1}^{M^O} a_i \mathbf{u}_i^{\max}.$$

Prediction (Figure 7.2 (E)): The prediction component consists of at least one feedforward neural network, which takes as input the summary vector $\mathbf{u}^{\max}$ and optionally the hidden state of the last sentence in the comment $\mathbf{s}_{M^C}$. More networks may be used to integrate other features as input, such as TFIDF-weighted n-grams of the comment. The outputs of the networks are concatenated and fed to the final prediction layer to compute $P(\Delta = 1)$. Using a single network that takes different kinds of features as input does not perform well, because the features are in different spaces, and linear operations between them are probably not meaningful.

Loss: The loss function is composed of binary cross-entropy loss and margin ranking loss. Assume there are total N^D attacked arguments, and the lth argument has N_l attacking arguments. The binary cross-entropy of the lth attacked argument and its tth attacking argument measures the similarity between the predicted $P(\Delta = 1)$ and the true Δ as:

$$\begin{aligned} BCE_{l,t} = &- \Delta_{l,t} \log P_\Theta(\Delta_{l,t} = 1) \\ &- (1 - \Delta_{l,t}) \log(1 - P_\Theta(\Delta_{l,t} = 1)), \end{aligned}$$

where $\Delta_{l,t}$ is the true $\Delta \in \{0, 1\}$ and P_Θ is the probability predicted by our model with parameters Θ. Since our data is skewed to negatives (persuasion failure), the model may overpredict $\Delta = 0$. To adjust this bias, we use margin ranking loss to drive the predicted probability of positives to be greater than the predicted probability of negatives to a certain margin. The margin ranking loss is defined on a pair of some attacking arguments C_1 and C_2 with $\Delta_{C_1} > \Delta_{C_2}$ as:

$$\begin{aligned} &MRL_{C_1,C_2} = \\ &\quad \max\{0, P_\Theta(\Delta_{C_2} = 1) - P_\Theta(\Delta_{C_1} = 1) + \varepsilon\}, \end{aligned}$$

where ε is a margin. Combining the two losses, our final loss is

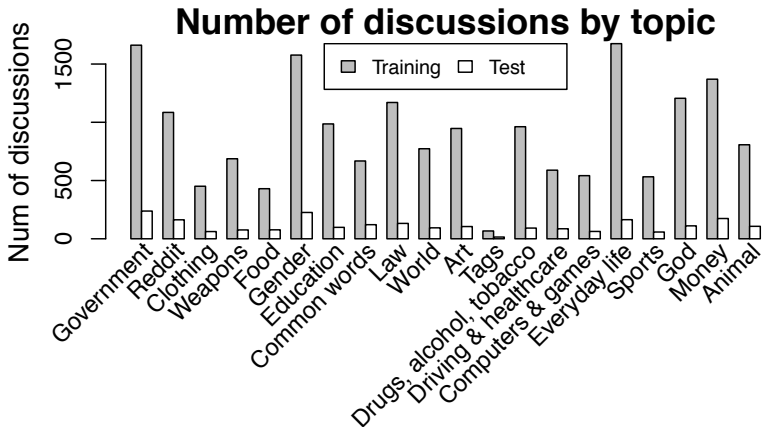
$$\frac{1}{N^D} \sum_{l=1}^{N^D} \frac{1}{N_l} \sum_{t=1}^{N_l} BCE_{l,t} + \mathbb{E}_{C_1,C_2}\left[MRL_{C_1,C_2}\right].$$

For the expectation in the ranking loss, we consider all pairs of attacking arguments in each minibatch and take the mean of their ranking losses.

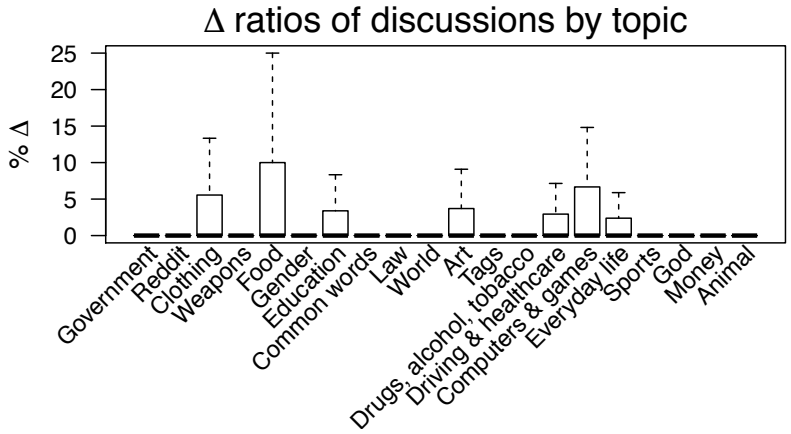
We implemented the model using PyTorch 0.3.0.

7.3.2 Experimental Settings

Our task is to predict whether a challenger's comment would receive a Δ, given the OP's initial post and the comment. We formulate this task as binary prediction of $\Delta \in \{0, 1\}$. Since our data is highly skewed, we use as our evaluation metric the AUC score, which measures the probability



(a) Number of discussions by topic.



(b) Delta ratios in discussions by topic. (e.g., a discussion has a 10% ratio if 10% of the OP's replies have a Δ.)

Figure 7.3: Discussion characteristics by topic.

of a positive instance receiving a higher probability of $\Delta = 1$ than a negative instance.

Data

We use the CMV dataset compiled by Tan et al. (2016)[1]. The dataset is composed of 18,363 discussions from January 1, 2013–May 7, 2015 for training data and 2,263 discussions from May 8–September 1, 2015 for test data. Note that this dataset is different from the CMV dataset we introduced in §4.2.3, although they are from the same subreddit.

We conducted qualitative analysis to better understand the data. First, to see if there are topical effects on changes in view, we examined the frequency of view changes across different topics. We ran Latent Dirichlet Allocation[2] (Blei et al., 2003) with 20 topics, taking each discussion as

[1] https://chenhaot.com/pages/changemyview.html

[2] Toolkit: LatentDirichletAllocation in scikit-learn v0.19.1 / `n_components`: 20 / `max_iter`: 200 / `learning_method`: online / `learning_offset`: 50

one document. We assigned each discussion the topic that has the highest standardized probability. The most discussed topics are government, gender, and everyday life (Figure 7.3a). As expected, the frequency of changes in view differs across topics (Figure 7.3b). The most malleable topics are food, computers & games, clothing, art, education, and everyday life. But even in the food domain, OPs give out a Δ in less than 10% of their replies in most discussions.

In order to see some behavior of users, we sampled discussions not in the test set and compared comments that did and did not receive a Δ. When a comment addresses the OP's points, its success relies on various interactions, including the newness of information, topical relatedness, and politeness. For example, in the discussion in Figure 7.1, Challenger 1 provides new information that is topically dissimilar to the OP's original reasoning. In contrast, Challenger 2's argument is relatively similar to the OP's reasoning, as it attempts to directly correct the OP's reasoning. These observations motivate the design of our Attentive Interaction Model, described in the next section.

Data Preprocessing

We exclude (1) DeltaBot's comments with no content, (2) comments replaced with "[deleted]", (3) system messages that are included in OP posts and DeltaBot's comments, (4) OP posts that are shorter than 100 characters, and (5) discussions where the OP post is excluded. We treat the title of an OP post as its first sentence. After this, every comment to which the OP replies is paired up with the OP's initial post. A comment is labeled as $\Delta = 1$ if it received a Δ and $\Delta = 0$ otherwise.

More specifically, in the CMV forum, DeltaBot replies to an OP's comment with the confirmation of a Δ, along with the user name to which the OP replied. For most OP replies, the (non-)existence of a Δ indicates whether a comment to which the OP replied changed the OP's view. However, an OP's view is continually influenced as they participate in argumentation, and thus a Δ given to a comment may not necessarily be attributed to the comment itself. One example is when a comment does not receive a Δ when the OP reads it for the first time, but the OP comes back and gives it a Δ after they interact with other comments. In such cases, we may want to give a credit to the comment that actually led the OP to reconsider a previous comment and change the view.

Hence, we use the following labeling method that considers the order in which OPs read comments. We treat the (non-)existence of a Δ in an OP comment as a label for the last comment *that the OP read*. We reconstruct the order in which the OP reads comments as follows. We assume that when the OP writes a comment, they have read all prior comments in the path to that comment.

Based on this assumption, we flatten the original tree structure of the initial post and all subsequent comments into a linear sequence S. Starting with empty S, for each of the OP's comments in chronological order, its ancestor comments that are yet to be in S and the comment itself are appended to S. And for each of the OP's comments, its preceding comment in S is labeled with $\Delta = 1$ if the OP's comment has a Δ and 0 otherwise.

This ensures that the label of a comment to which the OP replied is the (non-)existence of a Δ

	Train	Val	Test	CD
# discussions	4,357	474	638	1,548
# pairs	42,710	5,153	7,356	18,909
# positives	1,890	232	509	1,097

Table 7.1: Data statistics. (CD: cross-domain test)

in the OP's first reply. If an OP reply is not the first reply to a certain comment (as in the scenario mentioned above), or a comment to which the OP replied is missing, the (non-)existence of a Δ in that reply is assigned to the comment that we assume the OP read last, which is located right before the OP's comment in the restructured sequence.

The original dataset comes with training and test splits. After tokenization and POS tagging with Stanford CoreNLP (Manning et al., 2014), our vocabulary is restricted to the most frequent 40,000 words from the training data. For a validation split, we randomly choose 10% of training discussions for each topic.

We train our model on the seven topics that have the highest Δ ratios (Figure 7.3b). We test on the same set of topics for in-domain evaluation and on the other 13 topics for cross-domain evaluation. The main reason for choosing the most malleable topics is that these topics provide more information about people learning new perspectives, which is the focus of our paper. Some statistics of the resulting data are in Table 7.1.

Model Input

We use two basic types of input: sentence embeddings and TFIDF vectors. We acknowledge that these basic input types are not enough for our complex task, and most prior work utilizes higher-level features (politeness, sentiment, etc.) and task-specific information. Nevertheless, in this thesis, our experiment is limited to the basic input types to minimize feature engineering and increase replicability, but our model is general enough to incorporate other features as well.

Sentence embeddings: Our input sentences are sentence embeddings obtained by a pretrained sentence encoder (Conneau et al., 2017) (this is different from the sentence encoder layer in our model). The pretrained sentence encoder is a BiLSTM with max pooling trained on the Stanford Natural Language Inference corpus (Bowman et al., 2015) for textual entailment. Sentence embeddings from this encoder, combined with logistic regression on top, showed good performance in various transfer tasks, such as entailment and caption-image retrieval (Conneau et al., 2017).

TFIDF: A whole post or comment is represented as a TFIDF-weighted bag-of-words, where IDF is based on the training data[3]. We consider the top 40,000 n-grams ($n = 1, 2, 3$) by term frequency.

[3] TfidfVectorizer in scikit-learn v0.19.1, with the default setting

Word Overlap: Although integration of hand-crafted features is behind the scope of this paper, we test the word overlap features between a comment and the OP's post, introduced by Tan et al. (2016), as simple proxy for the interaction. For each comment, given the set of its words C and that of the OP's post O, these features are defined as $\left[|C \cap O|, \frac{|C \cap O|}{|C|}, \frac{|C \cap O|}{|O|}, \frac{|C \cap O|}{|C \cup O|}\right]$.

Model Settings

Network configurations: For sentence encoding, Gated Recurrent Units (Cho et al., 2014) with hidden state sizes 128 or 192 are explored. For attention, a single-layer feedforward neural network (FF) with one output node is used. For interaction encoding, we explore two interaction functions: (1) the inner product of the sentence embeddings and (2) a two-layer FF with 60 hidden nodes and three output nodes with a concatenation of the sentence embeddings as input. For prediction, we explore (1) a single-layer FF with either one output node if the summary vector $\mathbf{u}^{\max}$ is the only input or 32 or 64 output nodes with ReLU activation if the hidden state of the comment's last sentence is used as input, and optionally (2) a single-layer FF with 1 or 3 output nodes with ReLU activation for the TFIDF-weighted n-grams of the comment. The final prediction layer is a single-layer FF with one output node with sigmoid activation that takes the outputs of the two networks above and optionally the word overlap vector. The margin ε for the ranking margin loss is 0.5. Optimization is performed using AdaMax with the initial learning rate 0.002, decayed by 5% every epoch. Training stops after 10 epochs if the average validation AUC score of the last 5 epochs is lower than that of the first 5 epochs; otherwise, training runs 5 more epochs. The minibatch size is 10.

Input to prediction layer: The prediction component of the model takes combinations of the inputs: MAX ($\mathbf{u}^{\max}$), HSENT (the last hidden state of the sentence encoder $\mathbf{s}^{C}_{M^C}$), TFIDF (TFIDF-weighted n-grams of the comment), and WDO (word overlap).

Baselines

The most similar prior work to ours (Tan et al., 2016) predicted whether an OP would ever give a Δ in a discussion. The work used logistic regression with bag-of-words features. Hence, we also use logistic regression as our baseline to predict $P(\Delta = 1)$[4]. Simple logistic regression using TFIDF is a relatively strong baseline, as it beat more complex features in the aforementioned task.

Model configurations: Different regularization methods (L1, L2), regularization strengths ({.5, 0, 2, 4}), and class weights for positives (1, 2, 5) are explored. Class weights penalize false-negatives differently from false-positives, which is appropriate for the skewed data.

Input configurations: The model takes combinations of the inputs: TFIDF (TFIDF-weighted n-grams of the comment), TFIDF (+OH) (concatenation of the TFIDF-weighted n-grams of the comment and the OP's post), WDO (word overlap), and SENT (the sum of the input sentence embeddings of the comment).

[4] LogisticRegression in scikit-learn v0.19.1, with the default settings

Model	Inputs	In-domain	Cross-domain
LR	SENT	62.8	62.5
LR	TFIDF (+OH)	69.5	69.1
LR	TFIDF	70.9	**69.6**
LR	SENT+TFIDF	64.0	63.1
LR	TFIDF+WDO	71.1	69.5
AIM	MAX	70.5	67.5
AIM	MAX+TFIDF	**72.0***	69.4
AIM	MAX+TFIDF+WDO	70.9	68.4
(A)IM	HSENT	69.6	67.6
(A)IM	HSENT+TFIDF	69.0	67.6
(A)IM	MAX+TFIDF	69.5	68.1

Table 7.2: AUC scores. (LR: logistic regression, AIM: Attention Interaction Model, (A)IM: AIM without attention.) *: $p < 0.05$ using the DeLong test compared to LR with TFIDF.

7.3.3 Results

Table 7.2 shows the test AUC scores for the baseline and our model in different input configurations. For each configuration, we chose the optimal parameters based on validation AUC scores. Both interaction information learned by our model and surface-level *n*-grams in TFIDF have strong predictive power, and attending to attackable sentences helps. The highest score is achieved by our model (AIM) with both MAX and TFIDF as input (72.0%). The performance drops if the model does not use interaction information—(A)IM with HSENT (69.6%)—or attackability information—(A)IM with MAX+TFIDF (69.5%).

TFIDF by itself is also a strong predictor, as logistic regression with TFIDF performs well (70.9%). There is a performance drop if TFIDF is not used in most settings. This is unsurprising because TFIDF captures some topical or stylistic information that was shown to play important roles in argumentation in prior work (Tan et al., 2016; Wei et al., 2016). Simply concatenating both comment's and OP's TFIDF features does not help (69.5%), most likely due to the fact that a simple logistic regression does not capture interactions between features.

When the hand-crafted word overlap features are integrated to LR, the accuracy is increased slightly, but the difference is not statistically significant compared to LR without these features nor to the best AIM configuration. These features do not help AIM (70.9%), possibly because the information is redundant, or AIM requires a more deliberate way of integrating hand-crafted features.

For cross-domain performance, logistic regression with TFIDF performs best (69.6%). Our interaction information does not transfer to unseen topics as well as TFIDF. This weakness is alleviated when our model uses TFIDF in addition to MAX, increasing the cross-domain score (from 67.5% to 69.4%). We expect that information about attackability would have more impact within domain than across domains because it may learn domain-specific information about which

kinds of reasoning are attackable.

The rest of the section reports our qualitative analysis based on the best model configuration.

Can the model identify attackable sentences? If so, what properties constitute attackability? Our rationale behind sentence attackability scoring is that the model is able to learn to pay more attention to sentences that are more likely to change the OP's view when addressed. If the model successfully does this, then we expect more alignment between the attention mechanism and sentences that are actually addressed by successful comments that changed the OP's view.

To verify if our model works as designed, we randomly sampled 30 OP posts from the test set, and for each post, the first successful and unsuccessful comments. We asked a native English speaker to annotate each comment with the two most relevant sentences that it addresses in the OP post, without knowledge of how the model computes attackability scores and whether the comment is successful or not.

After this annotation, we computed the average attention weight of the two selected sentences for each comment. We ran a paired sample t-test and confirmed that the average attention weight of sentences addressed by successful comments was significantly greater than that of sentences addressed by unsuccessful comments ($p < 0.05$). Thus, as expected in the case where the attention works as designed, the model more often picks out the sentences that successful challengers address.

As to what the model learns as attackability, in most cases, the model attends to sentences that are not punctuation marks, bullet points, or irrelevant to the topic (e.g., "can you cmv?"). A successful example is illustrated in Figure 7.4 and Figure 7.5. Figure 7.6 shows some unsuccessful examples. All examples are from the test set.

What kinds of interactions between arguments are captured by the model? We first use argumentation theory as a lens for interpreting interaction embeddings. For this, we sampled 100 OP posts with all their comments and examined the 150 sentence pairs that have the highest value for each dimension of the interaction embedding (the dimensionality of interaction embeddings is 3 for the best performing configuration). 22% of the pairs in a dimension capture the comment asking the OP a question, which could be related to shifting the burden of proof. In addition, 23% of the top pairs in one dimension capture the comment pointing out that the OP may have missed something (e.g., "you don't know the struggles ..."). This might represent the challengers' attempt to provide premises that are missing in the OP's reasoning.

As providing missing information plays an important role in our data, we further examine if this attempt by challengers is captured in interaction embeddings even when it is not overtly signaled (e.g., "You don't know ..."). We first approximate the novelty of a challenger's information with the topic similarity between the challenger's sentence and the OP's sentence, and then see if there is a correlation between topic similarity and each dimension of interaction embeddings. The topic similarity between a pair of sentences is computed as the consine similarity between

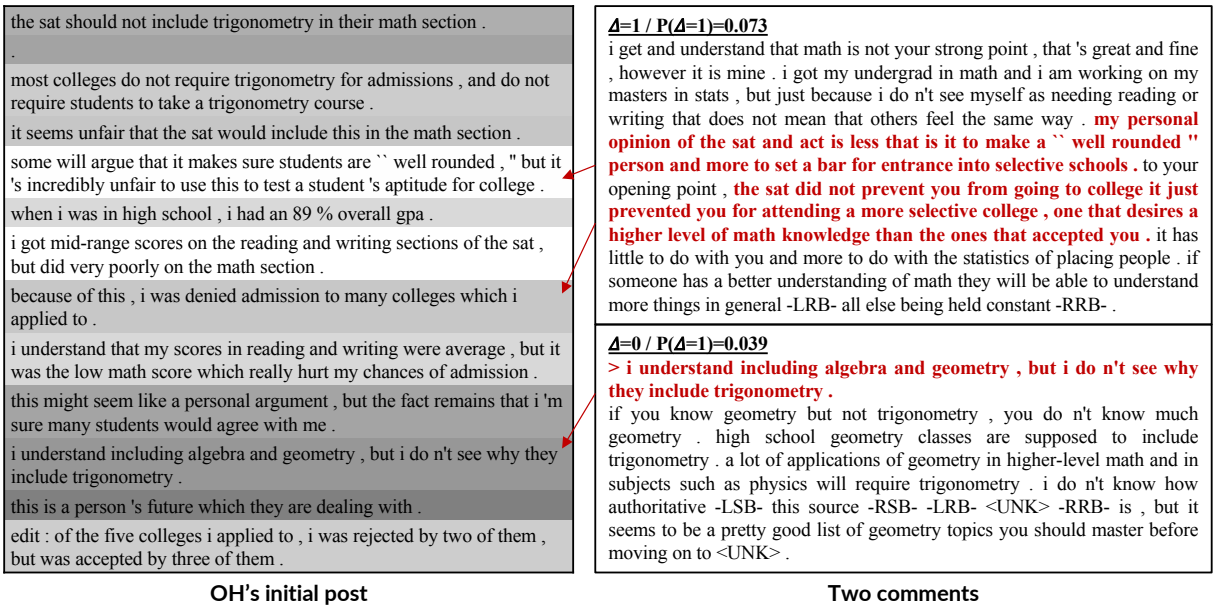


Figure 7.4: Example discussion with the OP's initial post (left), a successful comment (top right), and an unsuccessful comment (bottom right). The OP's post is colored based on attention weights (the higher attention the brighter). Sentences with college and SAT sections (*reading*, *writing*, *math*) get more attention than sentences with other subjects (*algebra*, *geometry*). The successful comment addresses parts with high attention, whereas the unsuccessful comment addresses parts with low attention.

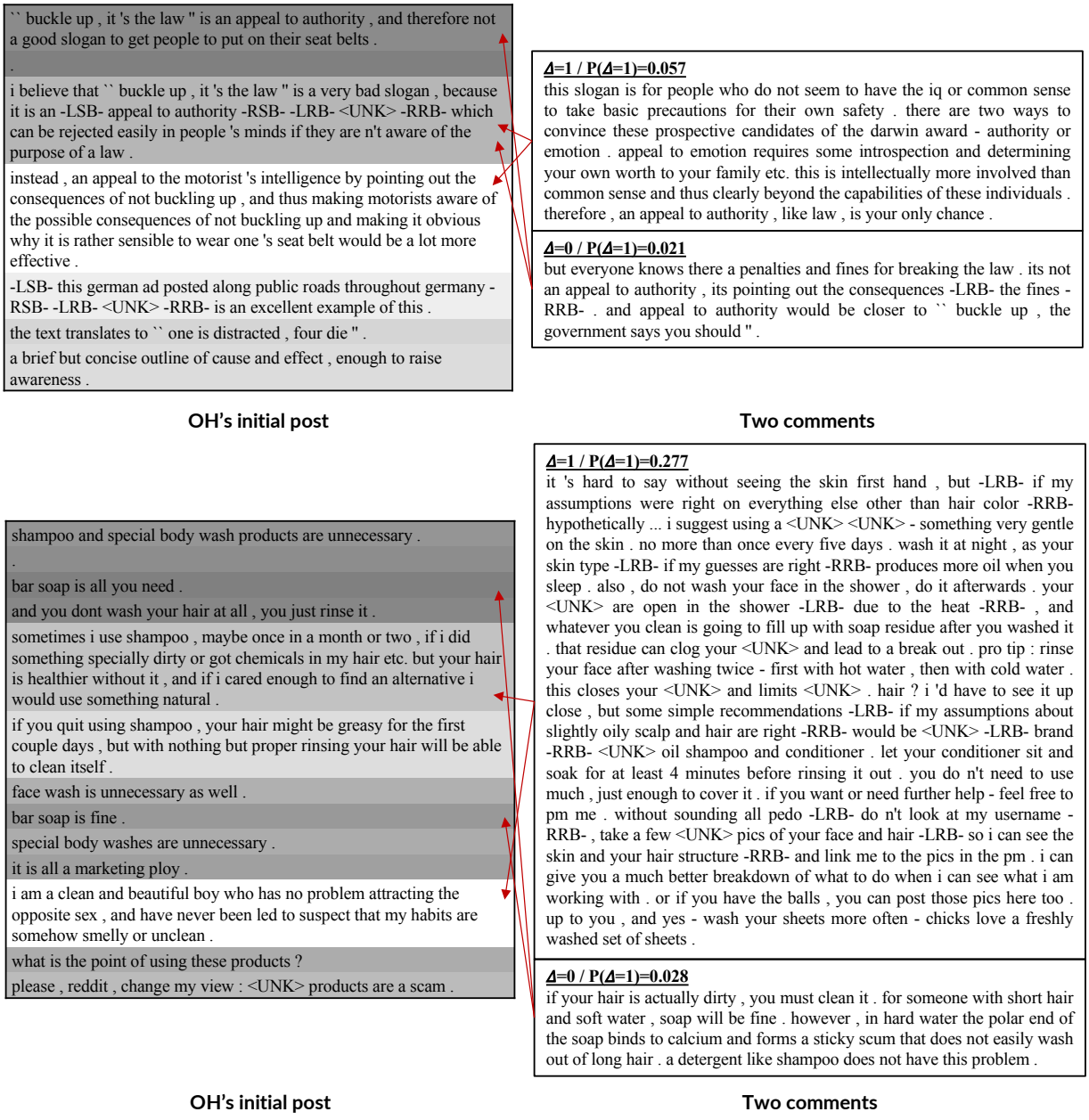


Figure 7.5: Successful examples of attackable sentence detection.

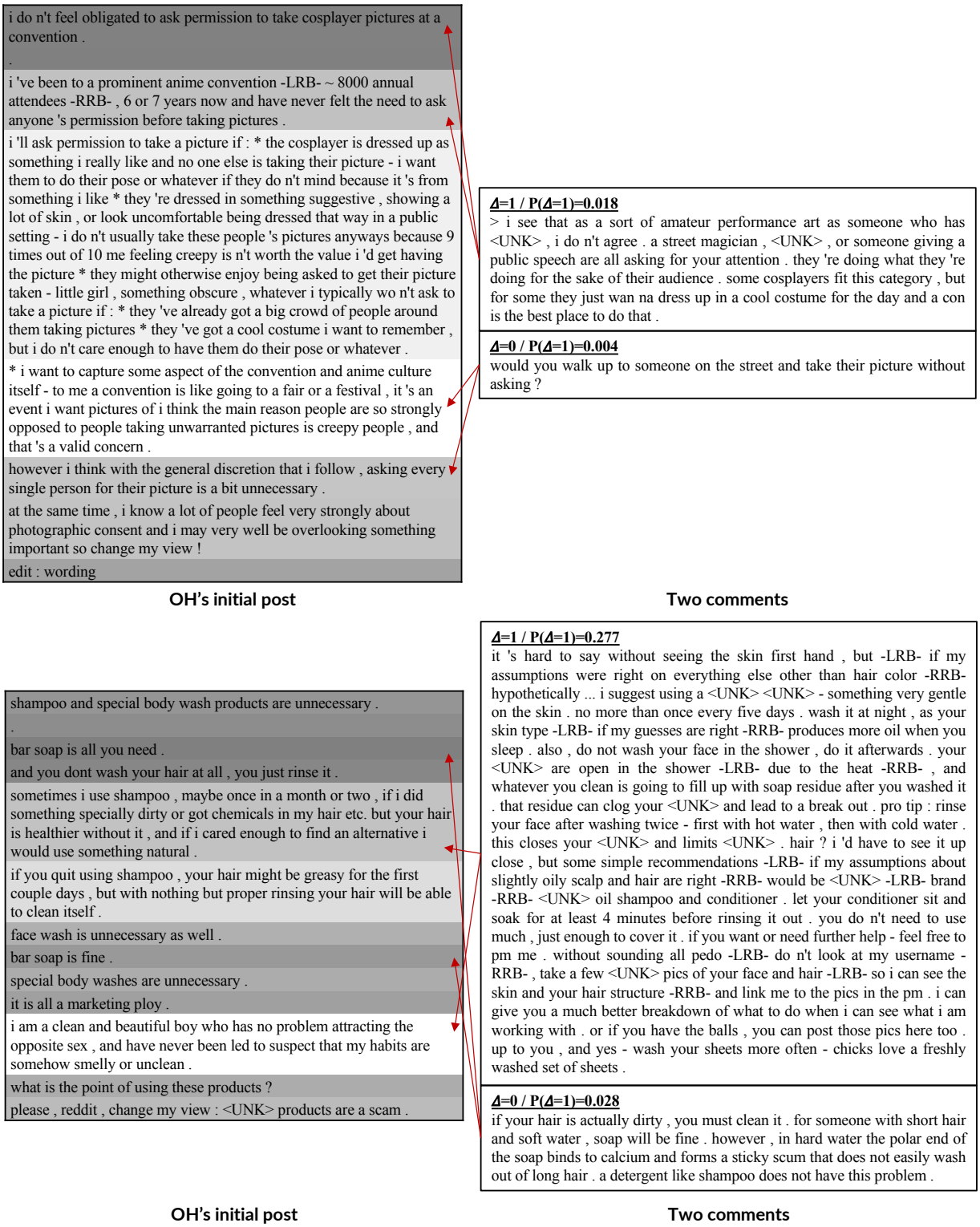


Figure 7.6: Unsuccessful examples of attackable sentence detection.

n-grams for $\Delta = 1$	n-grams for $\Delta = 0$
and, in, for, use, it, on, thanks, often, delta, time, depression, -RRB-, lot, -LRB-, or, i, can, &, with, more, as, band, *, #, me, - LRB-_-RRB-, can_be, has, deltas, when	?, >, sex, why, do_you, wear, relationship, child, are_you, op, mother, should, wearing, teacher, then, it_is, same, no, circumcision, you_are, then_you, baby, story

Table 7.3: Top n-grams with the most positive/negative weights for logistic regression.

the topic distributions of the sentences. The first step is to extract topics. We ran LDA[5] on the entire data with 100 topics, taking each post and comment as a document. We treat the top 100 words for each topic as topic words. The second step is to compute the topic distribution of each sentence. We simply counted the frequency of occurrences of topic words for each topic, and normalized the frequencies across topics. Lastly, we computed the cosine similarity between the topic distributions of a pair of sentences.

We found only a small but significant correlation (Pearson's $r = -0.04$) between topic similarity with one of the three dimensions. Admittedly, it is not trivial to interpret interaction embeddings and find alignment between embedding dimensions and argumentation theory. The neural network apparently learns complex interactions that are difficult to interpret in a human sense. It is also worth noting that the top pairs contain many duplicate sentences, possibly because the interaction embeddings may capture sentence-specific information, or because some types of interaction are determined mainly by one side of a pair (e.g., disagreement is manifested mostly on the challenger's side).

Lastly, we examine n-grams that are associated with the success and the failure of persuasion, reflected in TFIDF-weighted n-grams, based on their weights learned by logistic regression. The top n-grams with the highest and lowest weights are shown in table 7.3. First, challengers are more likely to change the OP's view when talking about themselves than mentioning the OP in their arguments. For instance, first-person pronouns (e.g., "i" and "me") get high weights, whereas second-person pronouns (e.g., "you_are" and "then_you") get low weights. Second, different kinds of politeness seem to play roles. For example, markers of negative politeness ("can" and "can_be", as opposed to "should" and "no") and negative face-threatening markers ("thanks"), are associated with receiving a Δ. Third, asking a question to the OP (e.g., "why", "do_you", and "are_you") is negatively associated with changing the OP's view.

Conclusion

To summarize this section, we presented the Attentive Interaction Model, which predicts an arguer's change in view through argumentation by sentence attackability scoring in the attacked argument and modeling the interaction between this argument and the attacking argument. According to the evaluation on discussions from the ChangeMyView forum, sentences identified by our model to be attackable were addressed more by successful challengers than by unsuccessful

[5] LatentDirichletAllocation in scikit-learn v0.19.1

ones. The model also effectively captured interaction information so that both attackability and interaction information increased accuracy in predicting the attacked arguer's change in view.

One limitation of our model is that making a prediction based only on one attacking argument (without considering the entire discussion) is not ideal because we miss context information that connects successive arguments. As a discussion proceeds, the topic may digress from the initially attacked argument. In this case, detecting attackable sentences and encoding interactions for this argument may become irrelevant. We leave the question of how to transfer contextual information from the overall discussion as future work.

Another key limitation is that it is not clear what makes certain sentences attackable. Although our neural representations have been found to be correlated with sentence attackability and to help to improve predicting persuasion outcomes, we still need a better understanding of the properties of attackable sentences. Therefore, we will discuss a more interpretable and semantic-oriented approach in the next section.

7.4 Semantic Modeling of Attackability

In the previous section, we presented a neural approach that computes the attackability scores of sentences in the attacked argument. However, this approach does not provide interpretable insights into characteristics that make certain sentences attackable. In this section, we present a more interpretable and semantic-oriented approach to examine the characteristics of sentences that distinguish attackable and non-attackable sentences.

Our key observation is that attacking specific points of an argument is common and effective; in our data of online discussions, challengers who successfully change the original poster (OP)'s view are 1.5 times more likely to quote specific sentences of the argument for attacks than unsuccessful challengers (Figure 7.7).

To examine the characteristics of attackable sentences in an argument, we first conduct a qualitative analysis of reasons for attacks in online arguments. Our data comes from discussions in the ChangeMyView (CMV) forum on Reddit. In CMV, users challenge the viewpoints of OPs, and those who succeed receive a Δ from the OPs. In this setting, sentences that are attacked and lead to the OP's view change are considered "attackable", i.e., targets that are worth attacking.

This analysis of reasons for attacks, along with argumentation theory and discourse studies, provide insights into what characteristics of sentences are relevant to attackability. Informed by these insights, we extract features that represent relevant sentence characteristics, clustered into four categories: content, external knowledge, proposition types, and tone. We demonstrate the effects of individual features on sentence attackability, in regard to whether a sentence would be *attacked* and whether a sentence would be attacked *successfully*.

Building on these findings, we examine the efficacy of machine learning models in detecting attackable sentences in arguments. We demonstrate that their decisions match the gold standard significantly better than several baselines and comparably well to laypeople.

>A society where everyone is equal seems great to me

That's one of the big problems with communism - what is equality? Is everyone equal? [...]

>it removes some of the basic faults in society, such as poverty, homelessness, joblessness, as well as touching on moral values such as greed, and envy

Yes there are problems within society but this doesn't mean there is a fault with society. [...]

>I believe a proper Communist society (I.E. one that is not a dictatorship like Joseph Stalin or Fidel Castro)

furthermore, it is unlikely we could ever get a true communist society due to human nature. [...]

Figure 7.7: A comment to a post entitled "I believe that Communism is not as bad as everyone says". It quotes and attacks some sentences in the post (red with ">")

7.4.1 Data and Labeling

We use the CMV corpus described in Section 4.2.3. As a reminder, we scraped CMV posts and comments written between January 1, 2014 and September 30, 2019, using the Pushshift API. We split them into a dev set (Jan 2014–Jan 2018 for training and Feb 2018–Nov 2018 for validation) and a test set (Dec 2018–Sep 2019), with the ratio of 6:2:2. We split the data by time to measure our models' generality to unseen subjects.

As the characteristics of arguments vary across different issues, we categorized the posts into domains using LDA. For each post, we chose as its domain the topic that has the highest standard score; topics comprising common words were excluded. We tried different numbers of topics (25, 30, 35, 40) and finalized on 40, as it achieves the lowest perplexity. This process resulted in 30 domains (excluding common-word topics): media, abortion, sex, election, Reddit, human economy, gender, race, family, life, crime, relationship, movie, world, game, tax, law, money, drug, war, religion, job, food, power, school, college, music, gun, and Jewish (from most frequent to least, ranging 5%–2%).

Since we are interested in which parts of a post are attacked by comments and whether the attacks lead to successful view changes, our goal here is to label each sentence in a post as *successfully attacked*, *unsuccessfully attacked*, or *unattacked*. We only consider comments directly replying to each post (top-level comments), as lower-level comments usually address the same points as their parent comments (as will be validated at the end of the section).

Attacked vs. Unattacked: Some comments use direct quotes with the $>$ symbol to address specific sentences of the post (Figure 7.7). Each quote is matched with the longest sequence of sentences in the post using the Levenshtein edit distance (allowing a distance of 2 characters for typos). A matched text span should contain at least one word and four characters, and cover at least 80% of the quote to exclude cases where the $>$ symbol is used to quote external content. As a result, 98% of the matched spans cover the corresponding quotes entirely. Additionally, a sentence

in the post is considered to be quoted if at least four non-stopwords appear in a comment's sentence. For example:

> **Post:** "... If you do something, you should be prepared to accept the consequences. ..."
> **Comment:** "... I guess my point is, even if you do believe that "**If you do something, you should be prepared to accept the consequences**," you can still feel bad for the victims. ..."

We considered manually annotating attacked sentences too, but it turned out to be extremely time-consuming and subjective. We tried to automate it using heuristics (word overlap and vector embeddings), but precision severely deteriorated. As we value the precision of labels over recall, we only use the method described in the previous paragraph. Chakrabarty et al. (2019) used the same method to collect attack relations in CMV.

Here is the specific steps we took to capture sentences in posts that are addressed by comments but not directly quoted: To see its feasibility, we randomly sampled 100 post-comment pairs that do not contain direct quotes and then asked an undergraduate native speaker of English (who has no knowledge about this work) to mark attacked sentences in each post, if any. This revealed two challenges. First, human annotation is subjective when compared to a co-author's result and very time-consuming (2.5 min/comment). Second, we tried several methods to automatically identify attacked sentences. We compared the similarity between each post sentence with the comment (first sentence of the comment, first sentence of each paragraph, or all comment text) based on word overlap with/without synonym expansion and the GloVe embeddings. But it turned out to be difficult to get similar results to human annotations. Therefore, we decided to use only those sentences that are direct quoted or have at least 4 common words with a comment's sentence as the most reliable labels.

Successfully vs. Unsuccessfully Attacked: After each sentence in a post is labeled as attacked or not, each attacked sentence is further labeled as *successfully attacked* if any of the comments that attack it, or their lower-level comments win a Δ.

We post-process the resulting labels to increase their validity. First, as a challenger and the OP have discussion down the comment thread, the challenger might attack different sentences than the originally attacked ones and change the OP's view. In this case, it is ambiguous which sentences contribute to the view change. Hence, we extract quotes from all lower-level comments of Δ-winning challengers, and if any of the quotes attack new sentences, this challenger's attacks are excluded from the labeling of *successfully attacked*. This case is not common, however (0.2%).

Second, if a comment attacks many sentences in the post and change the OP's view, some of them may not contribute to the view change but are still labeled as *successfully attacked*. To reduce this noise, comments that have more than three quotes are excluded from the labeling of *successfully attacked*[6]. This amounts to 12% of top-level comments (63% of comments have only one quote, 17% two quotes, and 8% three quotes).

[6]This allows our subsequent analyses to capture stronger signals for successful attacks than without this process.

Dataset		Train	Val	Test
Attacked	#posts	25,839	8,763	8,558
	#sentences	420,545	133,090	134,375
	#attacked	119,254	40,163	40,354
Successful	#posts	3,785	1,235	1,064
	#sentences	66,628	20,240	17,129
	#successful	8,746	2,718	2,288

Table 7.4: Data statistics. "Attacked" contains posts with at least one attacked sentence. "Successful" contains posts with at least one successfully attacked sentence.

Rationale	%
S is true but does not support the main claim	19%
S misses a case suggesting the opposite judgment	18%
S has exceptions	17%
S is false	12%
S misses nuanced distinctions of a concept	8%
S is unlikely to happen	6%
S has no evidence	6%
S uses an invalid assumption or hypothetical	4%
S contradicts statements in the argument	4%
Other	4%

(a) Rationales for attacking a sentence (*S*).

Factor	%
Personal opinion	28%
Invalid hypothetical	26%
Invalid generalization	13%
No evidence	11%
Absolute statement	7%
Concession	5%
Restrictive qualifier	5%
Other	5%

(b) Motivating factors for attacks.

Table 7.5: Rationales and motivating factors for attacks.

Lastly, we verified if quoted sentences are actually attacked. We randomly selected 500 comments and checked if each quoted sentence is purely agreed with without any opposition, challenge, or question. This case was rare (0.4%)[7], so we do not further process this case. Table 7.4 shows some statistics of the final data.

7.4.2 Quantifying Sentence Characteristics

As the first step toward analyzing the characteristics of attackable sentences, we examine driving reasons for attacks and quantify relevant sentence characteristics.

Rationales and Motivation for Attacks

To analyze rationales for attacks, two authors examined quotes and rebuttals in the training data (one successful and one unsuccessful comment for each post). From 156 attacks, we identified 10 main rationales (Table 7.5a), which are finer-grained than the refutation reasons in prior work (Wei

[7]Further, this case happened in only one out of the 500 comments (0.2%), where the author agreed with 4 quoted sentences. In CMV, challengers do use concessions but hardly quote the OP's sentences just to agree.

et al., 2016). The most common rationale is that the sentence is factually correct but is irrelevant to the main claim (19%). Counterexample-related rationales are also common: the sentence misses an example suggesting the opposite judgment to the sentence's own (18%) and the sentence has exceptions (17%).

This analysis is based on polished rebuttals, which mostly emphasize logical aspects, and cannot fully capture other factors that motivate attacks. Hence, we conducted a complementary analysis, where an undergraduate student chose three sentences to attack for each of 50 posts and specified the reasons in their own terms (Table 7.5b). The most common factor is that the sentence is only a personal opinion (28%). Invalid hypotheticals are also a common factor (26%). The tone of a sentence motivates attacks as well, such as generalization (13%), absoluteness (7%), and concession (5%).

Feature Extraction

Based on these analyses, we cluster various sentence characteristics into four categories—content, external knowledge, proposition types, and tone.[8]

Content: Content and logic play the most important role in CMV discussions. We extract the content of each sentence at two levels: TFIDF-weighted ***n*-grams** ($n = 1, 2, 3$) and sentence-level **topics**. Each sentence is assigned one topic using Sentence LDA (Jo and Oh, 2011). We train a model on posts in the training set and apply it to all posts, exploring the number of topics $\in \{10, 50, 100\}$.[9]

External Knowledge: External knowledge sources may provide information as to how truthful or convincing a sentence is (e.g., Table 7.5a-R2, R3, R4, R7 and Table 7.5b-F4). As our knowledge source, we use kialo.com—a collaborative argument platform over more than 1.4K issues (§6.6). Each issue has a main statement, and users can respond to any existing statement with pro/con statements (1-2 sentences), building an argumentation tree. Kialo has advantages over structured knowledge bases and Wikipedia in that it includes many debatable statements; many attacked sentences are subjective judgments (§7.4.2), so fact-based knowledge sources may have limited utility. In addition, each statement in Kialo has pro/con counts, which may reflect the convincingness of the statement. We scraped 1,417 argumentation trees and 130K statements (written until Oct 2019).

For each sentence in CMV, we retrieve similar statements in Kialo that have at least 5 common words[10] and compute the following three features. **Frequency** is the number of retrieved statements; sentences that are not suitable for argumentation are unlikely to appear in Kialo. This feature is computed as $\log_2(N+1)$, where N is the number of retrieved statements. **Attractiveness** is the average number of responses for the matched statements, reflecting how debatable

[8] Some rationales in Table 7.5a (e.g., R1 and R9) are difficult to operationalize reliably using the current NLP technology and thus are not included in our features.

[9] We also tried features based on semantic frames using SLING (Ringgaard et al., 2017), but they were not helpful.

[10] Similarity measures based on word embeddings and knowledge representation did not help. All these methods are described in Section 7.4.5 for interested readers.

the sentence is. It is computed as $\log_2(M+1)$, where $M = \frac{1}{N}\sum_{i=1}^{N} R_i$ and R_i is the number of responses for the ith retrieved statement. Lastly, **extremeness** is $\frac{1}{N}\sum_{i=1}^{N} |P_i - N_i|$, where P_i and N_i are the proportions (between 0 and 1) of pro responses and con responses for the ith retrieved statement. A sentence that most people would see flawed would have a high extremeness value.

Proposition Types: Sentences convey different types of propositions, such as predictions and hypotheticals. No proposition types are fallacious by nature, but some of them may make it harder to generate a sound argument. They also communicate different moods, causing the hearer to react differently. We extract 13 binary features for proposition types. They are all based on lexicons and regular expressions, which are available in Appendix 7.6).

Questions express the intent of information seeking. Depending on the form, we define three features: **confusion** (e.g., "I don't understand"), **why/how** (e.g., "why ...?"), and **other**.

Normative sentences suggest that an action be carried out. Due to their imperative mood, they can sound face-threatening and thus attract attacks.

Prediction sentences predict a future event. They can be attacked with reasons why the prediction is unlikely (Table 7.5a-R6), as in critical questions for *argument from cause to effect* (Walton et al., 2008).

Hypothetical sentences may make implausible assumptions (Table 7.5a-R8 and Table 7.5b-F2) or restrict the applicability of the argument too much (Table 7.5b-F7).

Citation often strengthens a claim using authority, but the credibility of the source could be attacked (Walton et al., 2008).

Comparison may reflect personal preferences that are vulnerable to attacks (Table 7.5b-F1).

Examples in a sentence may be attacked for their invalidity (Walton et al., 2008) or counterexamples (Table 7.5a-R3).

Definitions form a ground for arguments, and challengers could undermine an argument by attacking this basis (e.g., Table 7.5a-R5).

Personal stories are the arguer's experiences, whose validity is difficult to refute. A sentence with a personal story has subject "I" and a non-epistemic verb; or it has "my" modifying non-epistemic nouns.

Inclusive sentences that mention **"you"** and **"we"** engage the hearer into the discourse (Hyland, 2005), making the argument more vulnerable to attacks.

Tone: Challengers are influenced by the tone of an argument, e.g., subjectiveness, absoluteness, or confidence (Table 7.5b). We extract 8 features for the tone of sentences. The lexicons and regular expressions used for feature extraction are listed in Table 7.7

Subjectivity comprises judgments, which are often attacked due to counterexamples (Table 7.5a-R2) or their arbitrariness (Table 7.5b-F1, Walton et al. (2008)). The subjectivity of a sentence

Feature	Pattern
Question - Confusion	`r"(^| )i (\S + ){,2}(not|n't|never) (understand|know)"`, `r"(not|n't) make sense"`, `r"(^| )i (\S + ){,2}(curious|confused)"`, `r"(^| )i (\S + ){,2}wonder"`, `r"(me|myself) wonder"`
Question - Why/How	`r"(^| )(why|how).*\?"`
Question - Other	?
Normative	should, must, "(have|has) to", "have got to", "'ve got to", gotta, need, needs
Prediction	`r"(am$\|$'m$\|$are$\|$'re$\|$is$\|$'s) (not )?(going to$\|$gonna)"`, will, won't, would, shall
Hypothetical	`r"(^|, )if|unless"`
Citation	`r" {PATTERN} that [^.,!?]"` (`PATTERN`: said, reported, mentioned, declared, claimed, admitted, explained, insisted, promised, suggested, recommended, denied, blamed, apologized, agreed, answered, argued, complained, confirmed, proposed, replied, stated, told, warned, revealed), according to, `r"https?:"`
Comparison	than, compared to
Examples	`r"(^| )(for example|for instance|such as|e\.g\.)( |$)"`
Personal Story	**Epistemic verbs:** think, believe, see, know, feel, say, understand, mean, sure, agree, argue, consider, guess, realize, hope, support, aware, disagree, post, mention, admit, accept, assume, convince, wish, appreciate, speak, suppose, doubt, explain, wonder, discuss, view, suggest, recognize, respond, acknowledge, clarify, state, sorry, advocate, propose, define, apologize, curious, figure, claim, concede, debate, list, oppose, describe, suspect, reply, bet, realise, defend, convinced, offend, concern, intend, certain, conclude, reject, challenge, thank, condone, value, skeptical, contend, anticipate, maintain, justify, recommend, confident, promise, guarantee, comment, unsure, elaborate, posit, swear, dispute, imply, misunderstand. **Epistemic nouns:** view, opinion, mind, point, argument, belief, post, head, position, reasoning, understanding, thought, reason, question, knowledge, perspective, idea, way, stance, vote, best, cmv, response, definition, viewpoint, example, claim, logic, conclusion, thinking, comment, statement, theory, bias, assumption, answer, perception, intention, contention, word, proposal, thesis, interpretation, reply, guess, evidence, explanation, hypothesis, assertion, objection, criticism, worldview, impression, apology, philosophy
Use of "You"	you, your, yours
Use of "We"	`r"(^| )we |(?<!the) (us|our|ours)( |$)"`

Table 7.6: Lexicons and regular expressions used for extracting proposition type features.

Feature	Pattern
Subjectivity	Wilson et al. (2005)
Concreteness	Brysbaert et al. (2014)
Hedges	**Downtoners (score=1):** allegedly, apparently, appear to, conceivably, could be, doubtful, fairly, hopefully, i assume, i believe, i do not believe, i doubt, i feel, i do not feel, i guess, i speculate, i think, i do not think, if anything, imo, imply, in my mind, in my opinion, in my understanding, in my view, it be possible, it look like, it do not look like, kind of, mainly, may, maybe, might, my impression be, my thinking be, my understanding be, perhaps, possibly, potentially, presumably, probably, quite, rather, relatively, seem, somehow, somewhat, sort of, supposedly, to my knowledge, virtually, would. **Boosters (score=-1):** be definite, definitely, directly, enormously, entirely, evidently, exactly, explicitly, extremely, fundamentally, greatly, highly, in fact, incredibly, indeed, inevitably, intrinsically, invariably, literally, necessarily, no way, be obvious, obviously, perfectly, precisely, really, be self-evident, be sure, surely, totally, truly, be unambiguous, unambiguously, be undeniable, undeniably, undoubtedly, be unquestionable, unquestionably, very, wholly (Hyland, 2005 ; URL1 ; URL2)
Qualification	**Qualifiers (score=1):** a bit, a few, a large amount of, a little, a lot of, a number of, almost, approximately, except, generally, if, in general, largely, likely, lots of, majority of, many, more or less, most, mostly, much, nearly, normally, occasionally, often, overall, partly, plenty of, rarely, roughly, several, some, sometimes, tend, ton of, tons of, typically, unless, unlikely, usually. **Generality words (score=-1):** all, always, every, everybody, everyone, everything, never, no, no one, nobody, none, neither, not any, ever, forever (Hyland, 2005 ; URL2 ; URL3)
Arousal	Warriner et al. (2013)
Dominance	Warriner et al. (2013)

Table 7.7: Lexicons and regular expressions used for extracting tone features.

is the average subjectivity score of words based on the Subjectivity Lexicon (Wilson et al., 2005) (non-neutral words of "weaksubj" = 0.5 and "strongsubj" = 1).

Concreteness is the inverse of abstract diction, whose meaning depends on subjective perceptions and experiences. The concreteness of a sentence is the sum of the standardized word scores based on Brysbaert et al. (2014)'s concreteness lexicon.

Qualification expresses the level of generality of a claim, where absolute statements can motivate attacks (Table 7.5b-R3). The qualification score of a sentence is the average word score based on our lexicon of qualifiers and generality words.

Hedging can sound unconvincing (Durik et al., 2008) and motivate attacks. A sentence's hedging score is the sum of word scores based on our lexicon of downtoners and boosters.

Sentiment represents the valence of a sentence. Polar judgments may attract more attacks than neutral statements. We calculate the sentiment of each sentence with BERT (Devlin et al., 2018) trained on the data of SemEval 2017 Task 4 (Rosenthal et al., 2017). **Sentiment score** is a continuous value ranging between -1 (negative) and +1 (positive), and **sentiment categories** are nominal (positive, neutral, and negative)[11]. In addition, we compute the scores of **arousal** (intensity) and **dominance** (control) as the sum of the standardized word scores based on Warriner et al. (2013)'s lexicon.

7.4.3 Attackability Characteristics

One of our goals in this paper is to analyze what characteristics of sentences influence a sentence's attackability.

Hence, in this section, we measure the effect size and statistical significance of each feature toward two labels: (i) whether a sentence is attacked or not, using the dev set of the "Attacked" dataset (N=553,635), (ii) whether a sentence is attacked successfully or unsuccessfully, using all attacked sentences (N=159,417).[12] Since the effects of characteristics may depend on the issue being discussed, the effect of each feature is estimated conditioned on the domain of each post using a logistic regression, and the statistical significance of the effect is assessed using the Wald test (Agresti and Kateri, 2011).

For each feature, we use the following logistic regression model:

$$\log \frac{\mathbb{P}(\mathbf{Y}=1)}{1-\mathbb{P}(\mathbf{Y}=1)} = \beta_0 + \beta_X \mathbf{X} + \alpha_1 \mathbf{D}_1 + \cdots + \alpha_{|D|} \mathbf{D}_{|D|},$$

where $\mathbf{X}$ is a continuous or binary explanatory variable that takes the value of a characteristic that we are interested in. $\mathbf{D}_d$ ($d = 1, \cdots, |D|$) is a binary variable that takes 1 if the sentence belongs to the d-th domain. $\mathbf{Y}$ is a binary response variable that takes 1 if the sentence is attacked or if the sentence is attacked successfully. β_X is the regression coefficient of the characteristic $\mathbf{X}$, which is the main value of our interest for examining the association between the characteristic and the

[11] We achieved an average recall of 0.705, which is higher than the winner team's performance of 0.681.

[12] Simply measuring the predictive power of features in a prediction setting provides an incomplete picture of the roles of the characteristics. Some features may not have drastic contribution to prediction due to their infrequency, although they may have significant effects on attackability.

response; $\exp(\beta_X)$ is the *odds ratio* (OR) that is interpreted as the change of odds (i.e., the ratio of the probability that a sentence is (successfully) attacked to the probability that a sentence is not (successfully) attacked) when the value of the characteristic increases by one unit. If β_X is significant, we can infer that **X** has an effect on **Y**. If β_X is positive (and significant), we can infer that the characteristic and the response have positive association, and vice versa.

For interpretation purposes, we use *odds ratio* (OR)—the exponent of the effect size. Odds are the ratio of the probability of a sentence being (successfully) attacked to the probability of being not (successfully) attacked; OR is the ratio of odds when the value of the characteristic increases by one unit.

Content

Attacked sentences tend to mention big issues like gender, race, and health as revealed in topics 47, 8, and 6 (Table 7.8) and *n*-grams "life", "weapons", "women", "society", and "men" (Table 7.9). These issues are also positively correlated with successful attacks. On the other hand, mentioning relatively personal issues ("tv", "friends", topic 38) seems negatively correlated with successful attacks. So do forum-specific messages ("cmv", "thank", topic 4).

Attacking seemingly evidenced sentences appears to be effective for persuasion when properly done. Successfully attacked sentences are likely to mention specific data ("data", "%") and be the OP's specific reasons under bullet points ("2." and "3.").

n-grams capture various characteristics that are vulnerable to attacks, such as uncertainty and absoluteness ("i believe", "never"), hypotheticals ("if i"), questions ("?", "why"), and norms ("should").

External Knowledge

The Kialo-based knowledge features provide significant information about whether a sentence would be attacked successfully (Table 7.8). As the frequency of matched statements in Kialo increases twice, the odds for successful attack increase by 7%. As an example, the following attacked sentence has 18 matched statements in Kialo.

> "I feel like it is a parents right and responsibility to make important decisions for their child."

The attractiveness feature has a stronger effect; as matched statements have twice more responses, the odds for successful attack increase by 18%, probably due to higher debatability. A sentence being completely extreme (i.e., the matched sentences have only pro or con responses) increases the odds for successful attack by 19%.

As expected, the argumentative nature of Kialo allows its statements to match many subjective sentences in CMV and serves as an effective information source for a sentence's attackability.

	Feature	Attacked	Successful
Content	Topic47: Gender†	1.37 (***)	1.34 (***)
	Topic8: Race†	1.19 (***)	1.21 (**)
	Topic6: Food†	1.00 ( )	1.39 (***)
	Topic38: Movie & Show†	1.03 ( )	0.78 (***)
	Topic4: CMV-Specific†	0.16 (***)	0.36 (**)
Knowledge	Kialo Frequency (log2)	1.18 (***)	1.07 (***)
	Kialo Attractiveness (log2)	1.30 (***)	1.18 (***)
	Kialo Extremeness	1.51 (***)	1.19 (***)
Proposition Types	Question - Confusion†	0.97 ( )	1.29 (*)
	Question - Why/How†	1.77 (***)	1.27 (***)
	Question - Other†	1.16 (***)	1.11 (*)
	Citation†	0.53 (***)	1.17 (*)
	Definition†	1.04 ( )	1.32 (**)
	Normative†	1.26 (***)	1.10 (**)
	Prediction†	1.22 (***)	1.02 ( )
	Hypothetical†	1.29 (***)	1.07 ( )
	Comparison†	1.25 (***)	1.02 ( )
	Example†	1.20 (***)	1.17 (*)
	Personal Story†	0.70 (***)	1.09 (**)
	Use of *You*†	1.18 (***)	1.04 ( )
	Use of *We*†	1.24 (***)	0.98 ( )
Tone	Subjectivity‡	1.03 (***)	0.97 (***)
	Concreteness‡	0.87 (***)	0.92 (***)
	Hedges‡	1.04 (***)	1.06 (***)
	Quantification‡	0.97 (***)	1.02 ( )
	Sentiment Score‡	0.87 (***)	1.00 ( )
	Sentiment: Positive†	0.76 (***)	0.99 ( )
	Sentiment: Neutral†	0.82 (***)	1.00 ( )
	Sentiment: Negative†	1.34 (***)	1.00 ( )
	Arousal‡	1.02 (***)	0.95 (***)
	Dominance‡	1.07 (***)	1.08 (***)

Table 7.8: Odds ratio (OR) and statistical significance of features. An effect is positive (blue) if OR > 1 and negative (red) if OR < 1. (†: binary, ‡: standardized / *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

	Attacked (vs. Unattacked)	Attacked Successfully (vs. Unsuccessfully)
High	is are no - ? life women why should to society men a nothing 1_) would money they if_i n't people if * someone 2_. human believe never 3_. 2_) your and i_believe 5_. americans , tax 4 being :_- :_* feel because than *_the could do republicans be government sex) 3_) why_should nobody "_i the_government religion their seems ca ca_n't less 4_. war world pay an)_the 6_. without ,_why science reason animals 4_) humans racism of military selfish racist 3 when social makes have gun you climate get speech kids can white ,_is should_i *_** proven how_can	without -) data the public * ever someone are_a war way as_it % weapons ,_if how_can which ,_they were , since gender_is waste ? way_of land we_do travel effectively you_like 1_: problem_. transportation really important job the_us with up 3_) where c set dog countries rational use pretty_much that_can n't_the result_of is the_news song market rates for_people single have_. need happen guess porn years problem issues made so what_is 7 less organizations a_significant changes destruction cultural for_the so_you that_some actions second driving why_does weaker that_i problem_with an_individual coffee process investment pc boring does how_is high most_people ..._. and_if and
Low	edit cmv i /_? / thanks (edit_: [! post]]_(this thank thank_you comments please | view > discussion here topic sorry changed my_view some cmv_. posts ._" my delta comment i_will points responses :_1_. of_you /_) title article i_'ll = 'll thanks_for now 'm got & i_'m was **_edit above recently reddit view_. lot i_was below change_my hi 's a_few edit_2 on_this again " my_view_.)_. this_post discuss arguments you_all deltas few /_)_| i_'ve there_are 1_. i_have currently edit_2_: comments_. let_me let a_lot hello i_still)_| here_. course background context you_guys appreciate perspective respond thread posted	i_was suffering free edit linux tv moral kill 's_no everyone ,_for remember so_much go bitcoin their_own above there_'s_no developed alive why_should i some n't_. simple take_the that_this this_, of_" wealth / people_, video do_this ,_" hypocritical approach while_the mean to_me_. too_. poetry asking game whole articles you_do_n't meat possible poverty vegan a_great results more_to die would_be here_'s day words ' trump president security smoking cmv what_they i_have difficult you_do working due_to_the n't_feel dont warming rhetoric couple with_that the_human chose he basic_income skill for_everyone honest spend for_my basis ads to_see **_1 cop attracted saying lack_of machines along ad is_not nobody exclude

Table 7.9: n-grams ($n = 1, 2, 3$) with the highest/lowest weights. Different n-grams are split by a space, and words within an n-gram are split by "_".

Proposition Types

Questions, especially why/how, are effective targets for successful attack (Table 7.8). Although challengers do not pay special attention to expressions of confusion (see column "Attacked"), they are positively correlated with successful attack (OR=1.29).

Citations are often used to back up an argument and have a low chance of being attacked, reducing the odds by half. However, properly attacking citations significantly increases the odds for successful attack by 17%. Similarly, personal stories have a low chance of being attacked and definitions do not attract challengers' attacks, but attacking them is found to be effective for successful persuasion.

All other features for proposition types have significantly positive effects on being attacked (OR=1.18–1.29), but only normative and example sentences are correlated with successful attack.

Tone

Successfully attacked sentences tend to have lower subjectivity and arousal (Table 7.8), in line with the previous observation that they are more data- and reference-based than unsuccessfully attacked sentences. In contrast, sentences about concrete concepts are found to be less attackable.

Uncertainty (high hedging) and absoluteness (low qualification) both increase the chance of attacks, which aligns with the motivating factors for attacks (Table 7.5b), while only hedges are positively correlated with successful attacks, implying the importance of addressing the arguer's uncertainty.

Negative sentences with high arousal and dominance have a high chance of being attacked, but most of these characteristics have either no or negative effects on successful attacks.

Discussion

We have found some evidence that, somewhat counter-intuitively, seemingly evidenced sentences are more effective to attack. Such sentences use specific data ("data", "%"), citations, and definitions. Although attacking these sentences may require even stronger evidence and deeper knowledge, arguers seem to change their viewpoints when a fact they believe with evidence is undermined. In addition, it seems very important and effective to identify and address what the arguer is confused (confusion) or uncertain (hedges) about.

Our analysis also reveals some discrepancies between the characteristics of sentences that challengers commonly think are attackable and those that are indeed attackable. Challengers are often attracted to subjective and negative sentences with high arousal, but successfully attacked sentences have rather lower subjectivity and arousal, and have no difference in negativity compared to unsuccessfully attacked sentences. Furthermore, challengers pay less attention to personal stories, while successful attacks address personal stories more often.

7.4.4 Attackability Prediction

Now we examine how well computational models can detect attackable sentences in arguments.

Problem Formulation

This task is cast as ranking sentences in each post by their attackability scores predicted by a regression model. We consider two types of attackability: (i) whether a sentence will be attacked or not, (ii) whether a sentence will be successfully attacked or not (attacked unsuccessfully + unattacked). For both settings, we consider posts that have at least one sentence with the positive label (Table 7.4).

We use four evaluation metrics. **P@1** is the precision of the first ranked sentence, measuring the model's accuracy when choosing one sentence to attack for each post. Less strictly, **A@3** gives a score of 1 if any of the top 3 sentences is a positive instance and 0 otherwise. **MAP** is mean average precision, which measures the overall quality of ranking. **AUC** measures individual sentence-level accuracy—how likely positive sentences are assigned higher probabilities.

Comparison Models

For machine learning models, we explore two logistic regression models to compute the probability of the positive label for each sentence, which becomes the sentence's *attackability score*. **LR** is a basic logistic regression with our features[13] and binary variables for domains. We explored feature selection using L1-norm and regularization using L2-norm.[14] **BERT** is logistic regression where our features are replaced with the BERT embedding of the input sentence (Devlin et al., 2018). Contextualized BERT embeddings have achieved state-of-the-art performance in many NLP tasks. We use the pretrained, uncased base model from Hugging Face (Wolf et al., 2020) and fine-tune it during training.

We explore two baseline models. **Random** is to rank sentences randomly. **Length** is to rank sentences from longest to shortest, with the intuition that longer sentences may contain more information and thus more content to attack as well.

Lastly, we estimate laypeople's performance on this task. Three undergraduate students each read 100 posts and rank three sentences to attack for each post. Posts that have at least one positive instance are randomly selected from the test set.[15]

Results

All computational models were run 10 times, and their average accuracy is reported in Table 7.10. Both the LR and BERT models significantly outperform the baselines, while the BERT model performs best. For predicting attacked sentences, the BERT model's top 1 decisions match the

[13]We tried the number of topics $\in \{10, 50, 100\}$, and 50 has the best AUC on the val set for both prediction settings.

[14]We also tried a multilayer perceptron to model feature interactions, but it consistently performed worse than LR.

[15]We were interested in the performance of young adults who are academically active and have a moderate level of life experience. Their performance may not represent the general population, though.

	Attacked				Successfully Attacked			
	P@1	Any@3	MAP	AUC	P@1	Any@3	MAP	AUC
Random	35.9	66.0	48.0	50.1	18.9	45.0	34.0	50.1
Length	42.9	73.7	53.7	54.5	22.3	52.1	38.8	55.7
Logistic Regression	47.1	76.2	56.5	61.7	24.2	54.5	41.0	59.3
(×) Content	45.2	74.4	54.7	58.1	24.0	52.6	39.9	57.0
(×) Knowledge	47.0	76.0	56.4	61.7	24.1	54.3	40.5	59.0
(×) Prop Types	46.7	75.9	56.2	61.5	24.4	53.6	40.7	59.0
(×) Tone	47.0	76.0	56.4	61.9	25.2	56.2	41.4	59.4
BERT	49.6	77.8	57.9	64.4	28.3	57.2	43.1	62.0
Human	51.7	80.1	–	–	27.8	54.2	–	–

Table 7.10: Prediction accuracy.

gold standard 50% of the time; its decisions match 78% of the time when three sentences are chosen. Predicting successfully attacked sentences is harder, but the performance gap between our models and the baselines gets larger. The BERT model's top 1 decisions match the gold standard 28% of the time—a 27% and 10% boost from random and length-based performance, respectively.

To examine the contribution of each feature category, we did ablation tests based on the best performing LR model (Table 7.10 rows 4–7). The two prediction settings show similar tendencies. Regarding P@1 for successful attack, content has the highest contribution, followed by knowledge, proposition types, and tone. This result reaffirms the importance of content for a sentence's attackability. But the other features still have significant contribution, yielding higher P@1 and AUC (Table 7.10 row 4) than the baselines.

BERT can capture complex interactions among input features (i.e., input words) by taking advantage of a large number of parameters, complex attention mechanisms, and pretrained representations. As a result, BERT seems to capture various statistics related to sentence attackability better than the smaller set of hand-crafted features in LR does. One way to test this is to use adversarial training with reverse gradients (Ganin et al., 2016). That is, BERT is trained to predict attackability, as well as the values of our hand-crafted features (e.g., sentiment score or whether the sentence is a question or not). During training, we use reverse gradients for the prediction of hand-crafted features, where the goal is to make the model unable to predict the values of hand-crafted features and thus able to predict attackability from an input representation that lacks information about the hand-crafted features. If the model's performance of attackability prediction is still better than random, that means the model captures and uses additional information of input sentences other than the hand-crafted features.

It is worth noting that our features, despite the lower accuracy than the BERT model, are clearly informative of attackability prediction as Table 7.10 row 3 shows. Moreover, since they directly operationalize the sentence characteristics we compiled, it is pretty transparent that they capture relevant information that contributes to sentence attackability and help us better understand what

characteristics have positive and negative signals for sentence attackability. We speculate that transformer models like BERT are capable of encoding these characteristics more sophisticatedly and may include some additional information, e.g., lexical patterns, leading to higher accuracy. But at the same time, it is less clear exactly what they capture and whether they capture relevant information or irrelevant statistics, as is often the case in computational argumentation (Niven and Kao, 2019).

Figure 7.8 illustrates how LR allows us to interpret the contribution of different features to attackability, by visualizing a post with important features highlighted. For instance, external knowledge plays a crucial role in this post; all successfully attacked sentences match substantially more Kialo statements than other sentences. The attackability scores of these sentences are also increased by the use of hypotheticals and certain *n*-grams like "could". These features align well with the actual attacks by successful challengers. For instance, they pointed out that the expulsion of Russian diplomats (sentence 2) is not an aggressive reaction because the diplomats can be simply replaced with new ones. Kialo has a discussion on the relationship between the U.S. and Russia, and one statement puts forward exactly the same point that the expulsion was a forceful-looking but indeed a nice gesture. Similarly, a successful challenger pointed out the consistent attitude of the U.S. toward regime change in North Korea (sentence 3), and the North Korean regime is a controversial topic in Kialo. Lastly, one successful challenger attacked the hypothetical outcomes in sentences 4 and 5, pointing out that those outcomes are not plausible, and the LR model also captures the use of hypothetical and the word "could" as highly indicative of attackability.

For the erroneous example in Figure 7.9a, the successfully attacked sentence does not have many characteristics that are associated with attackability; rather, it has some characteristics that are negatively associated with attackability, such as neutrality and mentioning the hearer. The model misses most of its content and gives it a low attackability score. In contrast, other sentences have characteristics that are indicative of attackability, such as comparison and Kialo matches, or do not have characteristics negatively correlated with attackability.

For the erroneous example in Figure 7.9b, the successfully attacked sentence does not match many statements in Kialo, whereas sentences 3-5 match many statements in Kialo. As can be seen in this case, sentences that have few matched sentences in Kialo can be severely penalized in a ranking setting if other sentences have many matched statements.

Laypeople perform significantly better than the BERT model for predicting attacked sentences, but only comparably well for successfully attacked sentences (Table 7.10 row 9). Persuasive argumentation in CMV requires substantial domain knowledge, but laypeople do not have such expertise for many domains. The BERT model, however, seems to take advantage of the large data and encodes useful linguistic patterns that are predictive of attackability. A similar tendency has been observed in predicting persuasive refutation (Guo et al., 2020), where a machine-learned model outperformed laypeople. Nevertheless, in our task, the humans and the BERT model seem to make similar decisions; the association between their choices of sentences is high, with odds ratios ranging between 3.43 (top 1) and 3.33 (top 3). Interestingly, the LR model has a low association with the human decisions for top 1 (OR=2.65), but the association exceeds the BERT model for top 3 (OR=3.69). It would be interesting to further examine the similarities and

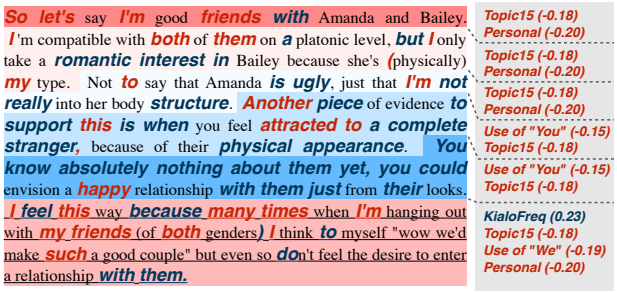
Figure 7.8: Prediction visualization. Background color indicates predicted attackability (blue: high, red: low). Successfully attacked sentences are underlined. Features with high/low weights are indicated with blue/red.

differences in how humans and machines choose sentences to attack.

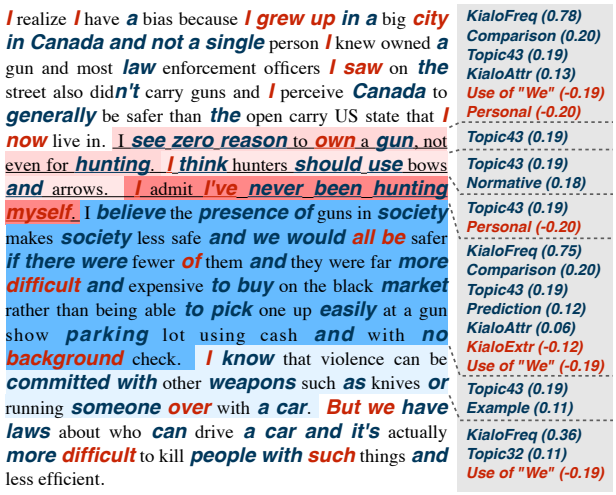
Conclusion

In this section, we studied how to detect attackable sentences in arguments for successful persuasion. Using CMV arguments, we demonstrated that a sentence's attackability is associated with many of its characteristics regarding its content, proposition types, and tone, and that Kialo provides useful information about attackability. Based on these findings we demonstrated that machine learning models can automatically detect attackable sentences, comparably well to laypeople.

Our work contributes a new application to the growing literature on causal inference from text (Egami et al., 2018), in the setting of "text as a treatment". Specifically, our findings in Section 7.4.3 pave the way towards answering the causal question: would attacking a certain type of



(a) Erroneous example 1.



(b) Erroneous example 2.

Figure 7.9: Examples of unsuccessful prediction. Background color indicates predicted attackability (blue: high, red: low). Successfully attacked sentences are underlined. Features with high/low weights are indicated with blue/red, respectively. For each sentence, features and their weights are sorted by absolute weight in the side bar.

sentence (e.g., questions or expressions of confusion) in an argument increase the probability of persuading the opinion holder? While our findings suggest initial hypotheses about the characteristics of sentences that can be successfully attacked, establishing causality in a credible manner would require addressing confounders, such as the challenger's reputation (Manzoor et al., 2020) and persuasive skill reflected in their attack (Tan et al., 2014). We leave this analysis to future work.

Our work could be improved also by including discourse properties (coherence, cohesiveness). Further, argumentation structure (support relations between sentences or lack thereof) might provide useful information about each sentence's attackability.

7.4.5 Appendix: Methods for Using External Knowledge

In this subsection, we describe the methods that we tried to use Kialo as a knowledge base but that were not successful.

Knowledge Feature	Attacked	Successful
Word Overlap Frequency (log2)	1.18 (***)	1.07 (***)
Word Overlap Attractiveness (log2)	1.30 (***)	1.18 (***)
Word Overlap Extremeness	1.51 (***)	1.19 (***)
UKP Avg Distance 10‡	0.93 (***)	0.98 (*)
UKP 0.1 Frequency†	1.08 (*)	0.99 ( )
UKP 0.1 Attractiveness†	1.11 (*)	1.08 ( )
UKP 0.1 Extremeness	3.49 (*)	6.77 ( )
UKP 0.2 Frequency†	1.02 (**)	1.01 ( )
UKP 0.2 Attractiveness†	1.05 (***)	1.06 ( )
UKP 0.2 Extremeness	1.69 (***)	1.76 ( )
UKP 0.3 Frequency†	1.04 (***)	1.01 ( )
UKP 0.3 Attractiveness†	1.09 (***)	1.02 ( )
UKP 0.3 Extremeness	2.44 (***)	1.40 ( )
UKP 0.4 Frequency†	1.04 (***)	1.01 (**)
UKP 0.4 Attractiveness†	1.12 (***)	1.01 ( )
UKP 0.4 Extremeness	2.35 (***)	1.02 ( )
Frame Knowledge Consistent	1.28 (***)	1.01 ( )
Frame Knowledge Conflict	1.37 (***)	1.08 ( )
Word Sequence Knowledge Consistent	1.05 ( )	0.98 ( )
Word Sequence Knowledge Conflict	1.18 ( )	1.49 ( )

Table 7.11: Odds ratio (OR) and statistical significance of features. An effect is positive (blue) if OR > 1 and negative (red) if OR < 1. (†: log2, ‡: standardized / *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

UKP Sentence Embedding-Based Retrieval

We measured the similarity between CMV sentences and Kialo statements using the UKP sentence embedding—BERT embeddings fine-tuned to measure argument similarity (Reimers et al., 2019). Specifically, the authors provide pretrained embeddings constructed by appending a final softmax layer to BERT to predict a numerical *dissimilarity* score between 0 and 1 for each sentence pair in the UKP ASPECT corpus. The 3,595 sentence pairs in this corpus were drawn from 28 controversial topics and annotated via crowd workers to be "unrelated" or of "no", "some" or "high" similarity. They report a mean F1-score of 65.39% on a held-out subset of this corpus, which was closest to human performance (F1=78.34%) among all competing methods that were not provided with additional information about the argument topic.

We used this fine-tuned model to measure the dissimilarity between each CMV sentence and Kialo statements. Based on this information, we extracted the feature **UKP Avg Distance 10**, which is the average dissimilarity score of the 10 Kialo statements that are closest to the sentence. This score is expected to be low if a sentence has many similar statements in Kialo. In addition, we extracted the same **frequency**, **attractiveness**, and **extremeness** features as in §7.4.2. Here, we determine whether a CMV sentence and a Kialo statement are "matched" based on several

dissimilarity thresholds (0.1, 0.2, 0.3, 0.4); A Kialo statement is considered matched with a CMV sentence if the dissimilarity is below the selected threshold.

Semantic Frame-Based Knowledge

We extracted semantic frames from CMV sentences and Kialo statements, using Google SLING (Ringgaard et al., 2017). For each frame in a sentence or statement, a "knowledge piece" is defined as the concatenation of the predicate and arguments (except negation); the predicate is lemmatized and the arguments are stemmed to remove differences in verb/noun forms. We also mark each knowledge piece as negated if the frame contains negation. Example knowledge pieces include:

- ARG0:peopl-ARG1:right-ARGM-MOD:should-PRED:have (Negation: true)
- ARG1:person-ARG2:abl-ARGM-MOD:should-PRED:be (Negation: false)

For each CMV sentence, we extracted two features: the count of knowledge pieces in Kialo that are **consistent** with those in the sentence, and the count of knowledge pieces in Kialo that are **conflicting** with those in the sentence. Two knowledge pieces are considered consistent if they are identical, and conflicting if they are identical but negated. Attackable sentences are expected to have many consistent and conflicting knowledge pieces in Kialo. If we assume that most statements in Kialo are truthful, attackable sentences may have more conflicting knowledge pieces than consistent knowledge pieces.

Word Sequence-Based Knowledge

Treating each frame as a separate knowledge piece does not capture the dependencies between multiple predicates within a sentence. Hence, we tried a simple method to capture this information, where a knowledge pieces is defined as the concatenation of verbs, nouns, adjectives, modal, prepositions, subordinating conjunctions, numbers, and existential "there" within a sentence; but independent clauses (e.g., a "because" clause) were separated off. All words were lemmatized. Each knowledge piece is negated if the source text has negation words. Example knowledge pieces include:

- gender-be-social-construct (Negation: true)
- congress-shall-make-law-respect-establishment-of-religion-prohibit-free-exercise (Negation: false)

For each CMV sentence, we extracted the same two features as in semantic frame-based knowledge pieces: the count of knowledge pieces in Kialo that are **consistent** with those in the sentence, and the count of knowledge pieces in Kialo that are **conflicting** with those in the sentence.

Effects and Statistical Significance

The effects and statistical significance of the above features were estimated in the same way as §7.4.3 and are shown in Table 7.11. Word sequence-based knowledge has no effect, probably because not many knowledge pieces are matched. Most of the other features have significant effects only for "Attacked". We speculate that a difficulty comes from the fact that both vector

embedding-based matching and frame-based matching are inaccurate in many cases. UKP sentence embeddings often retrieve Kialo statements that are only topically related to a CMV sentence. Similarly, frame-based knowledge pieces often cannot capture complex information conveyed in a CMV sentence. In contrast, word overlap-based matching seems to be more reliable and better retrieve Kialo statements that have similar content to a CMV sentence.

7.5 Conclusion

In order to refute an argument, the first step is to detect attackable sentences in the argument. In this chapter, we presented two approaches to detecting attackable sentences in arguments. The first approach jointly models sentence attackability and persuasion outcomes using an end-to-end neural network. We demonstrated that this joint modeling identifies attackable sentences that align with human judgments, even without explicit labels of attackability scores. In addition, modeling sentence attackability helps to better predict whether an attacking argument would successfully change the attacked arguer's view or not. The second approach complements the first approach by explicitly modeling various characteristics of sentences. We demonstrated that several characteristics of sentences have significant associations with sentence attackability, such as certain topics, expression of uncertainly, use of data and references. These findings and methods may be able to help people strengthen their argument by solidifying potentially attackable points, as well as help the machine to build convincing counterarguments.

Our problem setting can capture a wide range of attacks. Table 7.5a lists various types of attacks, including undercutters and meta-argumentation. The most common case is that the main claim of the attacked argument is not strongly supported by the presented premises in the argument. In this case, we can consider the main claim to be attacked. Similarly, undercutters, by definition, attack an inferential link between premise and claim (e.g., by saying that "this claim is not reasonably derived from your premise"), so we can still locate claims that they attack, which fit into the current problem setting of the attackability detection models.

However, automatically detecting a lack of strong evidence or weak inferential links requires understanding the argumentative structure of the attacked argument and assessing the inferential links. The current attackability detection models do not take argumentative structure into account nor assess the quality of inferential links. The former could be handled by using an argumentative relation classifier, such as the one in Chapter 6, whereas the latter is much more difficult.

Some attacks do not fit into our problem setting. For example, personal and situational attacks (Krabbe and van Laar, 2011) may not target specific sentences. Personal attacks (ad hominem) are generally considered to be unsound, so computational models may not want to use this strategy. Situational attacks are often concerned with how well the attacked argument fits in the context. Our models do not handle such attacks. They require an understanding of high-level dialogue structures, which would be a useful future direction that can benefit argumentation-enabled dialogue systems.

The quality of arguments can be assessed from various angles. There is a large body of literature in automated essay scoring that develops and uses various rubrics, such as grammaticality and

coherence. One aspect that is extremely important and currently missing in the NLP field is to measure the soundness of arguments. For this we may use critical questions associated with argumentation schemes (Walton et al., 2008). For instance, given an argument from expert opinion, critical questions include how credible the referenced expert is, whether the expert's opinion is consistent with other experts, and to what extent the conclusion is implied by the actual assertions made by the expert. The answer to each critical question may be scored using extensive real-world knowledge, and all the scores may be aggregated to quantify the quality of the argument.

Chapter 8

Retrieving Counterevidence

We build a system that, given a statement, retrieves counterevidence from diverse sources on the Web. At the core of this system is a natural language inference (NLI) model that determines whether a candidate sentence is valid counterevidence or not. Most NLI models to date, however, lack proper reasoning abilities necessary to find counterevidence that involves complex inference. Thus, we present a knowledge-enhanced NLI model that aims to handle causality- and example-based inference by incorporating knowledge graphs. Our NLI model outperforms baselines for NLI tasks, especially for instances that require the targeted inference. In addition, This NLI model further improves the counterevidence retrieval system, notably finding complex counterevidence better.

8.1 Introduction

Generating counterarguments is key to many applications, such as debating systems (Slonim, 2018), essay feedback generation (Woods et al., 2017), and legal decision making (Feteris et al., 2017). In NLP, many prior studies have focused on generating counterarguments to the main conclusions of long arguments, usually motions. Although such counterarguments are useful, argumentative dialogue is usually interactive and synchronous, and one often needs to address specific statements in developing argument. For instance, in the ChangeMyView (CMV) subreddit, challengers often quote and counter specific statements in the refuted argument, where 41% of these attacks are about factual falsehood, such as exceptions, feasibility, and lack of evidence (Jo et al., 2020). Hence, the scope of our work is narrower than most prior work. Instead of generating a counterargument to a *complete argument*, we aim to find counterevidence to *specific statements* in an argument. This counterevidence may serve as essential building blocks for developing a larger counterargument and also allow for more interactive development of argumentation.

We adopt a popular fact-verification framework (Thorne et al., 2018): given a statement to refute, we retrieve relevant documents from the Web and select counterevidence (Figure 8.1). At the core of this framework is a module that determines whether a candidate sentence is valid counterevidence to the given statement. A natural choice for this module is a natural language inference (NLI) model. But NLI models to date have shown a lack of reasoning abilities (Williams

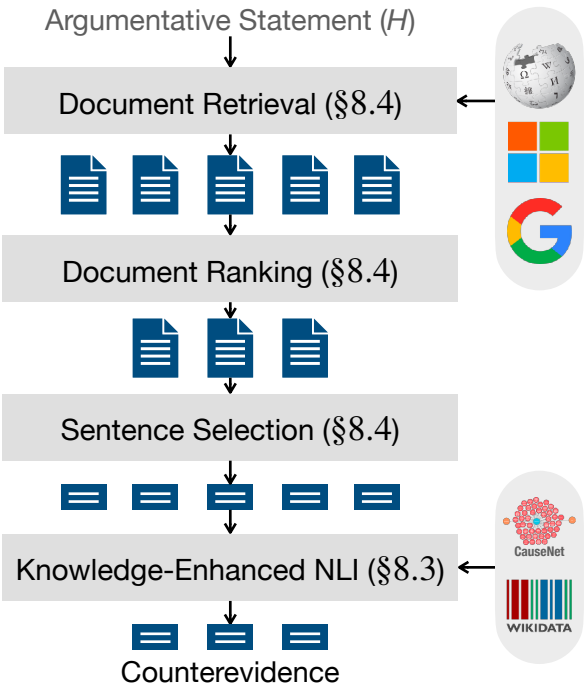


Figure 8.1: Architecture overview.

et al., 2020), which is problematic because counterarguments often involve complex inference. To overcome this limitation, we enhance NLI by focusing on two types of inference informed by argumentation theory (Walton et al., 2008). The first is *argument from examples*, as in:

> **Claim:** "Vegan food reduces the risk of diseases."
>
> **Counterevidence Statement:** "Legume protein sources can result in phytohemagglutinin poisoning."

The inference is based on the fact that "legume protein sources" and "phytohemagglutinin poisoning" are **examples** of "vegan food" and "diseases", respectively. The second type of inference is *argument from cause-to-effect*, as in:

> **Claim:** "Veganism reduces the risk of diabetes."
>
> **Counterevidence Statement:** "Vegan diets suffer from poor nutrition."

The inference is based on the fact that poor nutrition can **cause** diabetes.

In order to handle causality- and example-based reasoning, we develop a knowledge-enhanced NLI model (§8.3). By incorporating two knowledge graphs—CauseNet and Wikidata—into the model while training on public NLI datasets, the accuracy of the NLI model improves across NLI datasets, especially for challenging instances that require the targeted inference.

We integrate this NLI model into the entire retrieval system to find counterevidence to argumentative statements from two online argument platforms, ChangeMyView (CMV) and Kialo (§8.4). We demonstrate that our knowledge-enhanced NLI model improves the system, finding

more complex counterevidence. We also conduct in-depth analyses of the utility of different types of source documents and document search methods (Wikipedia, Bing, and Google).

8.2 Related Work

8.2.1 Counterargument Generation

Retrieval approach: Given an argument, Wachsmuth et al. (2018b) retrieves an argument in the debate pool idebate.org that is similar to the conclusion and dissimilar to the premise. Orbach et al. (2020) created a dataset of arguments and their counter arguments, collected from existing debates. Reisert et al. (2015) generates a counterargument to a claim that expresses advocacy for or opposition to a topic. Using lexical patterns, it retrieves sentences that specify consequences of the topic from the web corpus ClueWeb12. Sentences about a negative consequence are presented as a counterargument if the claim is advocacy, and positive sentences are presented otherwise. Sato et al. (2015) uses a similar approach except that for each claim, it first decides what value (e.g., health, poverty) to emphasize and then selects sentences that contain consequences of the topic in relation to the chosen value. Le et al. (2018) retrieves sentences in the debate forum ConvinceMe that are similar to the claim.

Neural generation: Hua and Wang (2018)'s model consists of two steps: evidence retrieval and text generation. For the evidence retrieval part, it retrieves relevant Wikipedia articles using sentences in the original argument and reranks the sentences using TF-IDF similarity to the argument. For the text generation part, the sentences are concatenated with the input argument. The model decodes a counterargument in two steps, first by decoding keyphrases and then by decoding a counterargument with attention to the keyphrases. Unlike retrieval-only approaches, the model does not directly evaluate whether the sentences are counterevidence to the argument. However, that is indirectly handled as the model learns to decode text that is similar to human-made counterarguments. Hua et al. (2019)'s model improves the previous method in several ways. First, it extends evidence passage selection to four news media. Second, it extracts (instead of generating) keyphrases from the input statement and candidate evidence passages using heuristics. Third, it ranks evidence passages by their keyphrase overlap with the input statement and their sentiment toward the input statement (to encourage counterevidence). For both methods, human evaluation shows that the quality of fully-generated counterarguments is yet lower than that of a simple concatenation of evidence passages in terms of topical relevance and counterness. The quality of evidence passages was not examined in detail. Hence, our work is complementary to these studies, as high-quality counterevidence is essential to generating counterarguments in natural language.

Argument generation: Wachsmuth et al. (2018a) prepares ADUs from existing arguments and analyze how humans synthesize arguments with these ADUs and rhetorical strategies. Baff et al. (2019) formulates ADU selection as language modeling and trains it on actual human-written arguments.

Most of these studies aim to build a counterargument against an entire argument. In extreme cases, generated counterarguments do not need to address specific points in the argument and still counter the main conclusion of the argument. On the other hand, our focus is slightly different and narrower. Our work aims to find counterevidence that directly addresses specific statements in the attacked argument. While some of the prior studies have the same aim, they focus on motions by retrieving their consequences.

8.2.2 Fact Verification

Due to this goal, our work is closely related to fact verification (Li and Zhou, 2020). Recently, this research area has garnered much attention, especially with the emergence of the FEVER (Fact Extraction and VERification) task (Thorne et al., 2018). The FEVER task aims to predict the veracity of statements. In the provided dataset, each statement has been labeled as supported, refuted, or not-enough-info, based on Wikipedia articles as a source; for supported or refuted statements, evidence sentences have also been annotated. Most approaches to this task follow three steps: document retrieval, sentence selection, and claim verification. A lot of initial studies focused on evidence representations (Ma et al., 2019; Tokala et al., 2019), while later studies began to examine homogeneous model architectures across different steps (Tokala et al., 2019; Nie et al., 2020a, 2019). Recently, BERT has been shown to be effective in both retrieval and verification (Soleimani et al., 2020), and a joint model of BERT and pointer net achieved state-of-the-art performance in this task (Hidey et al., 2020). Our work builds on this model, which will be discussed more in detail in §8.4. While multi-hop reasoning among evidence sentences has lately been studied (Liu et al., 2020b; Zhou et al., 2019), we do not consider this in our work as our main goal is to find individual counterevidence statements rather than verification of a claim. There are also studies on fact verification over tabular information (Yang et al., 2020; Zhong et al., 2020), which is beyond the scope of our work.

8.2.3 Knowledge-Enhanced Language Models

The last step of fact verification, i.e., claim verification, relies heavily on natural language inference (NLI) between an evidence text and a statement to verify. NLI is a long-studied topic in NLP, and recent transformer-based language models (LMs) have renewed state-of-the-art performance over the years (Nie et al., 2020b). However, evidence suggests that these models still lack reasoning abilities (Williams et al., 2020). Hence, here we discuss some research on integrating knowledge into LMs. Depending on the mechanism, there are three main types of approaches. The first is to exploit knowledge mainly to learn better embeddings of tokens and entities. Once learning is done, the triple information of entities in the input text is not activated during inference. KEPLER (Wang et al., 2019) has two components. One is a transformer-based text encoder, and the other is a KG encoder that computes each entity's embedding using the text encoder. The TransE objective function is optimized for the KG encoder jointly with the masked language model (MLM) task for the text encoder. For each downstream task, the pretrained text encoder is fine-tuned and used just like a transformer model. KnowBert (Peters et al., 2019) exploits entity embeddings from TransE. For a downstream task, token embeddings from BERT are updated based on self-attention scores among entities in the input text and similarity scores between each entity and similar entities.

ERNIE (Zhang et al., 2019) updates entity embeddings from TransE by self-attention among entities in the input text. The updated entity embeddings are aggregated with token embeddings from BERT through information fusion.

Unlike these models, the second type of models use the triple information of entities in the input text by encoding a subgraph during inference. BERT-MK (He et al., 2020a) encodes entities in the input text, along with their neighboring entities and relations, using a transformer. The updated entity embeddings are aggregated with token embeddings from BERT via information fusion. K-BERT (Liu et al., 2020a) converts entities in the input text and their neighboring entities and relations into text tokens, and combines them with the input text. BERT encodes the aggregate input text via manipulation of position embeddings and attention masks. KIM (Chen et al., 2018b) is designed for NLI. The two texts in the input pair are separately encoded by BiLSTM. Token embeddings are updated via co-attention between the two texts, where relations between tokens (synonymy, antonymy, hypernymy, hyponymy) are aggregated with the co-attention scores.

The third type of models incorporate knowledge using adapters (Houlsby et al., 2019) to avoid pretraining the entire language model, as well as the catastrophic forgetting of the pretrained representations. For instance, AdaptBERT (Lauscher et al., 2020) injects bottleneck adapters into BERT layers and trains them using a MLM task over a verbalized knowledge graph (each triple is converted to a sentence). For a downstream task, the adapters are fine-tuned.

8.3 Knowledge-Enhanced NLI

A natural language inference (NLI) model is the core of our entire system (Figure 8.1). Given a statement to refute, the system retrieves and ranks relevant documents, and then obtains a set of candidate sentences for counterevidence. For each candidate, the NLI model decides whether it entails, contradicts, or neither the statement. In this section, we first motivate the model design and explain our Knowledge-Enhanced NLI model (KENLI), followed by evaluation settings and results.

8.3.1 Motivation

Many NLI models have difficulty in capturing the relation between statements when their words are semantically far apart. For instance, if a statement refutes another based on example- or causality-based inference using technical terms (e.g., legume protein sources as an example of vegan food), the semantic gap between words can make it hard to capture the relation between the two statements without explicit knowledge.

To reduce semantic gaps between words, our method aims to bridge entities in the two statements using a knowledge graph (KG) so that the information of an entity in one statement flows to an entity in the other statement, along with the information of the intermediate entities and relations on the KG. This information updates the embeddings of the tokens linked to the entities.

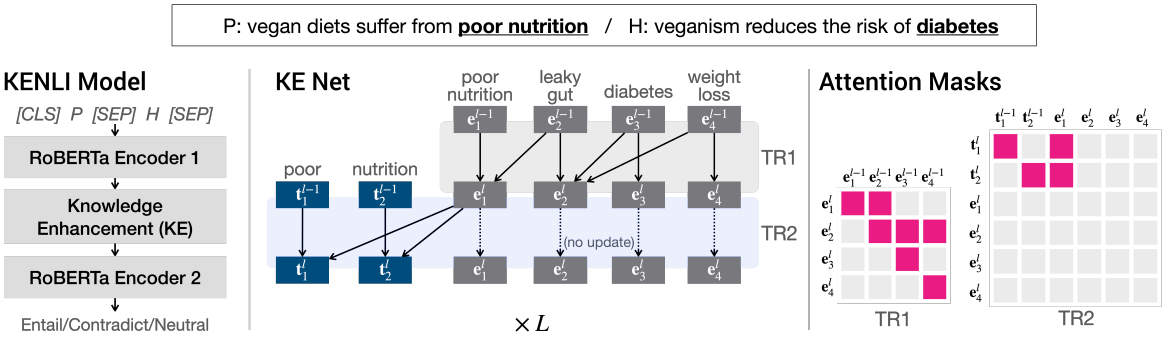


Figure 8.2: KENLI Model. This example illustrates using two KG paths: "poor nutrition $\xrightarrow{\text{cause}}$ leaky gut $\xrightarrow{\text{cause}}$ diabetes" and "poor nutrition $\xrightarrow{\text{cause}}$ leaky gut $\xrightarrow{\text{cause}}$ weight loss".

8.3.2 Model

KENLI (Figure 8.2 left) is based on RoBERTa-base (Liu et al., 2019), which takes a pair of premise P and hypothesis H as input and computes the probability of whether their relation is entailment, contradiction, or neutral. To bridge entities between P and H, the **Knowledge Enhancement (KE) Net** is inserted between two layers (e.g., 10th and 11th layers), splitting RoBERTa into **Encoder1** and **Encoder2**. It updates intermediate token embeddings from Encoder1 and feeds them to Encoder2. The final prediction is made through a fully-connected layer on top of the CLS embedding.

The KE Net (Figure 8.2 middle) exploits a knowledge graph (KG) where nodes are entities and edges are directed relations between entities (e.g., 'instance_of', 'cause'). Its main goal is to let information flow between entities in P and H through the KG. Suppose the KG has a set of relations $R = \{r_i\}_{i=1}^{|R|}$. For each input text pair, $T = \{t_i\}_{i=1}^{|T|}$ is the tokens in P that are linked to entities. Their initial embeddings $\{\mathbf{t}_i^0\}_{i=1}^{|T|}$ are the intermediate token embeddings from Encoder1. $E = \{e_i\}_{i=1}^{|E|}$ denotes entities under consideration, with initial embeddings $\{\mathbf{e}_i^0\}_{i=1}^{|E|}$. Considering all entities in the KG for every input pair is computationally too expensive. Recall that our motivation is to bridge entities between P and H. Hence, for each input pair, we first include entity paths whose source is in P and destination is in H. We add more destinations with the constraint that the total number of considered entities is no greater than λ and the length of each path is no greater than ν (λ and ν are hyperparameters). To obtain $\mathbf{e}_i^0$, we simply encode the name of each entity with RoBERTa Encoder1 and sum all the token embeddings.

The KE Net is a stack of KE cells. Each KE cell handles one-hop inference on the KG using two transformers TR1 and TR2. TR1 updates each entity embedding based on its neighboring entities, and TR2 updates token embeddings based on the embeddings of linked entities. More specifically, in the l-th KE cell, TR1 takes $\{\mathbf{e}_i^{l-1}\}_{i=1}^{|E|}$ as input and updates their embeddings using self-attention. Each attention head corresponds to each relation, and the attention mask for the k-th head $M^k \in \mathbb{R}^{|E|\times|E|}$ allows the information flow between entities that have the k-th relation:

$$M_{ij}^k = \begin{cases} 1 & \text{if } i = j \text{ or } (e_i, r_k, e_j) \in \text{KG} \\ 0 & \text{otherwise.} \end{cases}$$

TR2 takes the concatenation of $\{\mathbf{t}_i^{l-1}\}_{i=1}^{|T|}$ and $\{\mathbf{e}_i^l\}_{i=1}^{|E|}$ as input and updates the token embeddings using one attention head with attention mask $M \in \mathbb{R}^{|T+E|\times|T+E|}$:

$$M_{ij} = \begin{cases} 1 & \text{if } i \leq |T| \text{ and} \\ & \quad (i = j \text{ or } t_i \text{ is linked to } e_{j-|T|}) \\ 0 & \text{otherwise.} \end{cases}$$

Entity embeddings are not updated in TR2.

After token embeddings are updated by L KE cells (i.e., L-hop inference), the token embedding of t_i is updated as $\mathbf{t}_i \leftarrow \mathbf{t}_i^0 + \mathbf{t}_i^L$ and fed to Encoder2 along with the other token embeddings in the input.

8.3.3 Knowledge Graphs

Our work uses two knowledge graphs: CauseNet and Wikidata. **CauseNet** (Heindorf et al., 2020) specifies *claimed* causal relations between entities, extracted from Wikipedia and ClueWeb12 based on linguistic markers of causality (e.g., "cause", "lead") and infoboxes. We discard entity pairs that were identified by less than 5 unique patterns, since many of them are unreliable. This results in total 10,710 triples, all having the 'cause' relation.

Wikidata (Vrandečić and Krötzsch, 2014) is a database that specifies a wide range of relations between entities. We use the October 2020 dump and retain triples that have 8 example-related relations: instance_of, subclass_of, part_of, has_part, part_of_the_series, located_in_the_administrative_territorial_entity, contains_administrative_territorial_entity, and location. The importance of information about physical and temporal containment in NLI was discussed recently (Williams et al., 2020). This filtering results in 95M triples, which we call **WikidataEx**.

8.3.4 Data

Our data mainly come from public NLI datasets: MNLI (Williams et al., 2018), ANLI (Nie et al., 2020b), SNLI (Bowman et al., 2015), and FEVER-NLI (Nie et al., 2019). We split the data into train, validation, and test sets as originally or conventionally set up for each dataset (Table 8.1). Due to limited computational resources, our training set includes only MNLI and ANLI.

The public NLI datasets alone may not include enough instances that require example- and causality-based inference. As a result, the NLI model may not learn to exploit the KGs well. To alleviate this issue, we generate synthetic NLI pairs that are built on example-based inference as follows (details are in §8.5.2). Given a pair of P and H in the public datasets, we modify P to P' by replacing an entity that occurs in both P and H with an incoming entity on WikidataEx (e.g., "England" with "Yorkshire"). This achieves two goals. First, P' includes an entity that is an example of another entity in H so that the (P', H) pair requires example-based inference, with the same expected relation as the (P, H) pair. Second, this example relation comes from our KG so that the NLI model learns how to use the KG. Generating similar NLI pairs for causality-based inference is more challenging, and we leave it to future work.

Dataset	Train	Val	Test
MNLI	392,702	–	9,815
MNLI-MM	–	–	9,832
ANLI	162,865	3,200	3,200
SNLI	–	9,842	9,824
SNLI-Hard	–	–	3,261
FEVER-NLI	–	9,999	9,999
Example-NLI	30,133	2,867	3,468
ANLI-Contain	–	–	277
ANLI-Cause	–	–	1,078
BECauSE	–	–	2,814

Table 8.1: Number of NLI pairs by dataset.

Inference Evaluation: We use additional datasets to evaluate NLI models' inference abilities. For example-based inference, we first use a diagnostic subset of ANLI that has been annotated with various categories of required inference, such as counting, negation, and coreference (Williams et al., 2020). We choose the instances of the 'Containment' category, which requires inference on part-whole and temporal containment between entities (**ANLI-Contain**). In addition, we use the test set of our **Example-NLI** data after manually inspecting their labels.

For causality-based inference, we use the instances in the diagnostic ANLI set that belong to the 'CauseEffect' and 'Plausibility' categories (**ANLI-Cause**). They require inference on logical conclusions and the plausibility of events. In addition, we use **BECauSE 2.0** (Dunietz et al., 2017), which specifies the 'Cause' and 'Obstruct' relations between text spans based on linguistic markers of causality. Since it has only two classes, we randomly pair up text spans to generate 'neutral' pairs. For reliability, we discard pairs where at least one text comprises only one word. Although this data is not for NLI, we expect that the better NLI models handle the causality between events, the better they may distinguish between the cause, obstruct, and neutral relations. See Table 8.1 for statistics.

8.3.5 Experiment Settings

For KENLI, the KE Net is inserted between the 10th and 11th layers of RoBERTa, although the location of insertion has little effect on NLI performance. The KE Net has a stack of two KE cells, allowing for 2-hop inference on a KG. We test KENLI with CauseNet (**KENLI+C**) and with WikiedataEx (**KENLI+E**); we do not combine them so we can understand the utility of each KG more clearly. The maximum number of entities for each input (λ) and the maximum length of each KG path (ν) are set to 20 and 2, respectively. To see the benefit of pretraining the KE Net (as opposed to random initialization) prior to downstream tasks, we also explore pretraining it with masked language modeling on the training pairs while the original RoBERTa weights are fixed (**KENLI+E+Pt** and **KENLI+C+Pt**). The Adam optimizer is used with a learning rate of 1e-5.

We compare KENLI with three baselines. The first two are state-of-the-art language models

	NLI Evaluation							Inference Evaluation			
	MNLI	MNLI-MM	ANLI	SNLI	SNLI-Hard	FEVER-NLI	Micro Avg	Example-NLI	ANLI-Contain	ANLI-Cause	BECauSE
AdaptBERT+C	83.0	83.6	44.7	78.8	68.2	68.2	75.4	58.4	42.3	35.0	27.6
AdaptBERT+E	83.2	83.5	44.7	78.7	68.4	67.8	75.4	58.8	43.4	34.7	27.5
K-BERT+C	83.7	83.9	45.2	80.0	70.3	68.9	76.2	58.9	42.4	34.7	27.2
K-BERT+E	83.4	83.7	46.0	79.3	69.5	69.2	76.0	59.0	44.2	**35.8**	26.9
RoBERTa	87.3	87.0	48.6	84.2	74.6	71.9	79.7	61.8	47.7	35.0	27.6
KENLI+C	87.5	87.1	48.2	84.3	74.8	71.4	79.7	62.0	48.2	35.1	27.9
KENLI+C+Pt	87.3	86.9	**48.8**	84.2	74.2	71.9	79.7	61.7	48.4	35.2	27.8
KENLI+E	87.3	**87.2**	48.5	84.2	**75.1**	**72.5***	79.9*	61.9	**49.2**	35.5	**28.0**
KENLI+E+Pt	**87.6**	87.1	48.4	**84.6**	75.1	72.5*	**80.0**†	**62.0**	46.9	35.2	27.6

Table 8.2: F1-scores of NLI models by dataset. Statistical significance was measured by the paired bootstrap against the best baseline ($p < 0.05^{\star}, 0.01^{\dagger}$). Bold and underline each indicate top1 and top2 results, respectively.

enhanced with knowledge graphs. **K-BERT** (Liu et al., 2020a) exploits a KG during both training and inference, by verbalizing subgraphs around the entities linked to the input and combining the verbalized text into the input. **AdaptBERT** (Lauscher et al., 2020) uses a KG to enhance BERT using bottleneck adapters (Houlsby et al., 2019); after that, it is fine-tuned for downstream tasks like normal BERT. We pretrain AdaptBERT for masked language modeling on sentences that verbalize CauseNet (10K) and a subset of WikidataEx (10M) for four epochs. We use the hyperparameter values as suggested in the papers. The last baseline is **RoBERTa**-base fine-tuned on the NLI datasets. RoBERTa trained with the ANLI dataset recently achieved a state-of-the-art performance for NLI (Nie et al., 2020b).

Input texts are linked to WikidataEx entities by the Spacy Entity Linker[1]. CauseNet has no public entity linker, so we first stem all entities and input words using Porter Stemmer and then use exact stem matching for entity linking. The stemming allows verbs in input texts to be linked to entities (e.g., "infected–infection", "smokes–smoking").

8.3.6 Results

Table 8.2 shows the F1-scores of each model averaged over 5 runs with random initialization.

In the NLI evaluation, KENLI (rows 6–9) generally outperforms the baseline models (rows 1–5) across datasets. Especially KENLI with WikidataEx (rows 8–9) performs best overall and notably well for difficult datasets (SNLI-Hard, FEVER-NLI, and ANLI). This suggests that KENLI effectively incorporates example-related knowledge, which benefits prediction of nontrivial relations between statements. KENLI with CauseNet (rows 6–7) slightly underperforms KENLI+E, and its average F1-score across datasets is comparable to RoBERTa (row 5). Without

[1] pypi.org/project/spacy-entity-linker/

pretraining (row 6), it performs slightly better than RoBERTa overall except for two difficult datasets ANLI and SNLI-Hard. With pretraining (row 7), its performance is best for the most difficult dataset ANLI, but slightly lower than or comparable to RoBERTa for the other datasets. This variance in performance across datasets makes it hard to conclude the benefit of CauseNet in general cases.

However, according to the inference evaluation, KENLI's strength is clearer compared to other models. For example-based inference, KENLI+E (row 8) significantly outperforms the other models (ANLI-Contain) or performs comparably well (Example-NLI). Its performance is best or second-best for causality-based inference as well (ANLI-Cause and BECauSE). This suggests that the benefit of example-related knowledge is not limited to example-based inference only. Although KENLI+C (rows 6–7) shows comparable performance to RoBERTa for the general NLI tasks, it consistently outperforms RoBERTa when example- and causality-based inference is required.

Pretraining KENLI (rows 7 & 9) does not show a conclusive benefit compared to no pretraining (rows 6 & 8). Particularly for difficult datasets and inference evaluation, KENLI+E without pretraining (row 8) performs better than pretraining (row 9). The benefit of pretraining for KENLI+C varies depending on the dataset and inference task, making no substantial difference overall.

8.4 Evidence Retrieval

Our system for counterevidence retrieval builds on DeSePtion (Hidey et al., 2020), a state-of-the-art system for the fact extraction and verification (FEVER) task (Thorne et al., 2018). As Figure 8.1 shows, given a statement to verify, it retrieves and ranks relevant documents, ranks candidate evidence sentences, and predicts whether the statement is supported, refuted, or neither. Our main contribution is to strengthen the last stage via our knowledge-enhanced NLI model. In this section, we first explain individual stages and then describe evaluation settings and results.

8.4.1 Stages

Document Retrieval: In this stage, documents that may contain counterevidence are retrieved. Given a statement to verify, DeSePtion retrieves candidate documents from Wikipedia in four ways: (1) using named entities in the statement as queries for the wikipedia library[2], (2) using the statement as a query for Google, (3) TF-IDF search using DrQA (Chen et al., 2017), and (4) some heuristics. Note that all documents are from Wikipedia, in accordance with the FEVER task.

We make several adaptations that better suit our task. First, in addition to Wikipedia articles, we also retrieve web documents using Microsoft Bing and Google (wikipedia pages are excluded from their search results). The three sources provide documents with somewhat different characteristics, and we will compare their utility in §8.4.4. Second, we use the Spacy Entity Linker to retrieve the articles of Wikidata entities linked to the statement. And for each linked entity, we additionally

[2] https://pypi.org/project/wikipedia/

sample at most five of their instance entities and the corresponding articles. These expanded articles potentially include counterexamples to the statement[3]. We will analyze the utility of these documents. Lastly, we do not use the heuristics.

Document Ranking: Given a set of candidate documents, DeSePtion ranks them using a pointer net combined with fine-tuned BERT. First, BERT is trained to predict whether each document is relevant or not, using the FEVER dataset; it takes the concatenation of the page title and the statement to verify as input. The output is used as the embedding of the document. Next, a pointer net takes these embeddings of all documents and sequentially outputs pointers to relevant documents.

In our adaptation, we use RoBERTa in place of BERT. More importantly, we use search snippets in place of page titles to take advantage of the relevant content in each document provided by search engines. The ranker is still trained on the FEVER dataset, but since it does not include search snippets, we heuristically generate snippets by concatenating the title of each Wikipedia page with its sentence that is most similar to the statement[4]. This technique substantially improves document relevance prediction on the FEVER dataset by 7.4% F1-score points. For web documents, we use search snippets provided by Bing and Google.

The number of retrieved documents varies a lot depending on the search method; the Google API retrieves much fewer documents than Wikipedia and Bing in general. Since this imbalance makes it difficult to compare the utility of the different search methods, we make the number of candidate documents the same across the methods, by ranking documents from different search methods separately and then pruning low-ranked documents of Wikipedia and Bing. This process lets the three methods have the same average number of candidate documents per statement ($\sim$8).

Sentence Selection: DeSePtion considers all sentences of the ranked documents. However, web documents are substantially longer than Wikipedia articles in general, so it is computationally too expensive and introduces a lot of noise to process all sentences. Therefore, for each statement to verify, we reduce the number of candidate sentences by selecting the top 200 sentences (among all ranked documents) whose RoBERTa embeddings have the highest cosine similarity to the statement.

Relation Prediction: In this stage, we classify whether each candidate sentence is valid counterevidence to the statement to verify. Here, instead of DeSePtion, we simply use an NLI model as-is[5]. We compute the probability score that each sentence contradicts the statement and rank

[3]We considered retrieving web documents in a similar way, using query expansion, but ended up not doing it. One reason is that search engines already include example-related documents to some extent. For instance, for the query "Vegan diets can cause cancer", Bing returns a document with the title "Can the Keto and Paleo Diets Cause Breast Cancer?". Another practical reason is that query expansion requires arbitrarily many search transactions that are beyond the capacity of our resources.

[4]We combine all token embeddings in the last layer of RoBERTa and measure the cosine similarity between these vectors.

[5]The main goal of DeSePtion is to predict the veracity of the statement, rather than whether each sentence supports or refutes the statement. Thus, it assumes that once the statement is found to be supported or refuted, considering

sentences by these scores. This simple approach has been found to be effective in information retrieval (Dai and Callan, 2019).

8.4.2 Data

Input statements to our system come from the two argument datasets in §7.4 collected from the ChangeMyView (CMV) subreddit and Kialo. On CMV, the user posts an argument, and other users attempt to refute it often by attacking specific sentences. Each sentence in an argument becomes our target of counterevidence. On Kialo, the user participates in a discussion for a specific topic and makes a statement (1–3 sentences) that either supports or attacks an existing statement in the discussion. We find counterevidence to each statement. To use our resources more efficiently, we discard CMV sentences or Kialo statements that have no named entities or Wikidata entities, since they often do not have much content to refute. We also run coreference resolution for third-person singular personal pronouns using the neuralcoref 4.0 library[6]. We randomly select 94 posts (1,599 sentences) for CMV and 1,161 statements for Kialo for evaluation.

8.4.3 Evaluation

We evaluate four NLI models. The first three models are directly from §8.3. That is, **RoBERTa** is fine-tuned on the NLI data. **KENLI+C** and **KENLI+E** are trained with CauseNet and WikidataEx, respectively, without pretraining. The last baseline is **LogBERT**, a state-of-the-art model for argumentative relation classification from Chapter 6. Given a pair of statements, it predicts whether the first statement supports, attacks, or neither the second statement based on four logical relations between them, namely, textual entailment, sentiment, causal relation, and normative relation[7]. Since LogBERT captures the support and attack relations beyond textual entailment, this baseline would show whether NLI is sufficient for finding counterevidence.

We collect a ground-truth set of labeled data using MTurk. First, for each statement to refute, we include in the ground-truth set the top candidate sentence from each model if the probability of contradiction is ≥ 0.5 (i.e., max four sentences). As a result, the ground-truth set consists of 4,783 (CMV) and 3,479 (Kialo) candidate sentences; they are challenging candidates because at least one model believes they are valid counterevidence.

Each candidate sentence is scored by two Turkers with regard to how strongly it refutes the statement (very week=0, week=1, strong=2, and very strong=3). Each candidate sentence is displayed with the surrounding sentences in the original source document as context, as well as a link to the source document. If a candidate is scored as both very weak and very strong, these scores are considered unreliable, and thus is further scored by a third Turker. For each candidate,

more sentences results in the same prediction. This assumption is justified for the FEVER task, where a statement cannot be both supported and refuted. In real-world arguments, a statement can be both supported and refuted, and our goal is to find refuting sentences.

[6] github.com/huggingface/neuralcoref

[7] For implementation, BERT-base is fine-tuned for the four classification tasks and then for argumentative relation classification on the Kialo arguments (both normative and non-normative).

	CMV				Kialo			
	Prec	Recl	F1	τ	Prec	Recl	F1	τ
RoBERTa	48.3	63.6	54.9	0.2	58.0	57.0	57.5	2.2
KENLI+C	48.8	65.0	55.8	1.4	59.0	$62.2^{\ddagger}$	$60.6^{\ddagger}$	$3.8^{\dagger}$
KENLI+E	48.9	**71.3**‡	**58.0**‡	1.5	58.0	$65.2^{\ddagger}$	$61.4^{\ddagger}$	$3.8^{\dagger}$
LogBERT	**51.4**†	61.8	56.1	**3.1***	**60.0**	**66.2**‡	**62.9**‡	**4.5**†

Table 8.3: Accuracy of evidence retrieval. For precision, recall, and F1-score, statistical significance was calculated using the paired bootstrap against RoBERTa; for Kendall's τ, the statistical significance of each correlation value was calculated ($p < 0.05^{\star}, 0.01^{\dagger}, 0.001^{\ddagger}$).

the mean score s is taken as the ground-truth validity as counterevidence: 'valid' if $s \geq 1.5$ and 'invalid' otherwise. More details are described in §8.5.3.

According to the additional question of whether reading the source document is necessary to make a decision for each candidate, about 40% of answers and 65% of candidates required reading source documents. This might indicate that three sentences are insufficient for making robust decisions about counterevidence, but it could also be the case that, since our system checks all documents and filter them by relevance in earlier stages, it would not benefit much from more than three sentences.

We use four evaluation metrics on the ground-truth set. **Precision**, **recall**, and **F1-score** are computed based on whether the model-predicted probability of contradiction for each candidate is ≥ 0.5. These metrics, however, make the problem binary classification, missing the nuanced degree of validity for each candidate. Thus, we measure **Kendall's** τ between mean validity scores from human judgments and each model's probability scores. High τ indicates a good alignment between the human judgment and the model judgment about the strength of validity of each candidate.

8.4.4 Results

Table 8.3 summarizes the accuracy of evidence retrieval. Both KENLI+C (row 2) and KENLI+E (row 3) outperform RoBERTa (row 1) for both CMV and Kialo. The motivation behind KENLI was to capture statement pairs that require complex inference, by bridging entities with KGs. As expected, KENLI identifies more instances of contradiction that are missed by RoBERTa, as indicated by its high recall. The recall of KENLI+E is substantially higher than RoBERTa's by 7.7 and 8.3 points for CMV and Kialo, respectively, while its improvement of precision is relatively moderate. KENLI+C has a similar pattern but with a smaller performance gap with RoBERTa.

To see if KENLI-E indeed effectively captures counterevidence that requires example-based inference, we broke down its F1-score into one measured on candidate sentences for which KG paths exist between their tokens and the statement's tokens and one measured on the other candidate sentences with no connecting KG paths (Figure 8.3). The F1-score gap between KENLI-E and RoBERTa is substantially higher for the candidate sentences where KG paths exist. The gap

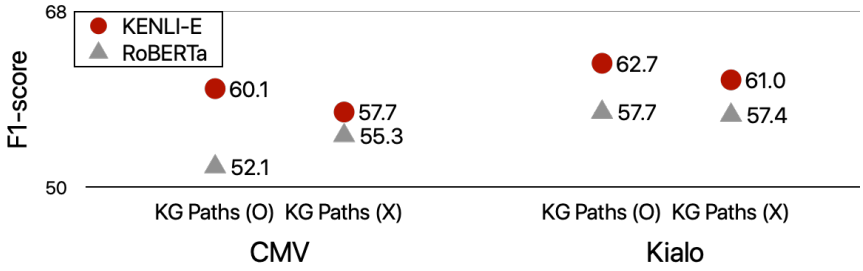


Figure 8.3: F1-scores of KENLI-E and RoBERTa by the existence of KG paths in candidate sentences.

of recall is even higher, indicating that KENLI-E indeed captures complex counterevidence more effectively than RoBERTa. KG paths that benefit KENLI-E the most include "player PART_OF game", "Tel Aviv District LOCATED_IN Israel", and "neurovascular system HAS_PART brain". That is, KENLI-E and RoBERTa show the largest accuracy gap for candidate sentences that include these KG paths.

LogBERT slightly underperforms KENLI+E for CMV, but it outperforms KENLI+E for Kialo, possibly because LogBERT is trained on arguments from Kialo and may learn argumentative patterns in Kialo. In contrast to KENLI, LogBERT is notable for relatively high precision compared to the other models. This is somewhat counterintuitive, because LogBERT uses four logical relations between two statements (textual entailment, sentiment, causal relation, and normative relation), which might improve recall substantially by capturing a broad range of mechanisms for contradiction. Although that may be the case for Kialo as reflected in the high recall, the four relations rather seem to correct each other, i.e., wrong decisions based on one relation are adjusted by other relations. All the results so far suggest the promising direction of combining KENLI and LogBERT: we may be able to capture more instances of contradiction by incorporating different KGs and, at the same time, improve precision by incorporating different types of signals (e.g., sentiment). We leave this direction to future work.

According to Kendall's τ, LogBERT shows the best alignment with human judgments on the validity scores of candidate sentences among the four models. KENLI shows better correlations with human judgments than RoBERTa. However, overall correlation values are rather small, ranging between 0.2% and 4.5%.

The reason that the models have higher accuracy for Kialo than for CMV is not clear. We analyzed accuracy by the length of refuted statements and by whether the refuted statement is normative or not, but we did not find conclusive evidence that they are important factors for accuracy.

Utility of Search Methods: One difference between our system and prior work is that we retrieved web documents using Bing and Google, whereas no prior work did that to our knowledge. Hence, comparing candidate sentences from Wikipedia, Bing, and Google will shed light on the usefulness of the search engines and inform future system designs. Table 8.4 shows candidate

	CMV			Kialo		
	P	R	F	P	R	F
Wikipedia	42.4	64.5	51.1	55.1	60.8	57.8
Bing	53.1	66.2	58.9	59.5	63.5	61.4
Google	47.0	64.8	54.5	59.9	62.6	61.2

Table 8.4: Accuracy of counterevidence retrieval by search methods.

sentences retrieved from Bing and Google generally achieve higher F1-scores than those from Wikipedia. While Wikipedia provides comparably good recall, its precision is substantially lower than the other methods. This suggests that Wikipedia is a great source of a vast amount of relevant information, but the other search methods are worth resorting to if one needs more precise and nuanced counterevidence.

Utility of Document Types: One question we want to answer is: what types of documents are useful sources of counterevidence to argumentative statements? Prior work focuses mostly on Wikipedia articles (Thorne et al., 2018; Hua and Wang, 2018), debates (Orbach et al., 2020; Wachsmuth et al., 2018b), and occasionally news articles (Hua et al., 2019). In contrast, our candidate sentences come from more diverse types of documents, such as academic papers and government reports. To analyze the utility of different document types, we first annotated each candidate sentence with 13 different types using MTurk (Table 8.5). Each network location[8] was tagged by Turkers until the same label was annotated twice; if no such label occurred for five annotations, we chose the label of the Turker who had the highest acceptance rate. More details are described in §8.5.4.

First of all, Figure 8.4 shows the distribution of document types for valid counterevidence. A lot of counterevidence exists in knowledge archives (27–37%), followed by mainstream news (8–13%), magazines about social issues (7–12%), personal blogs (5–10%), and research journals (6–8%). This suggests the benefit of using broader types of documents in counterevidence and fact verification than conventionally used Wikipedia and debates.

Table 8.5 summarizes the F1-score of counterevidence retrieval by document types (averaged across all models). For both CMV and Kialo, financial magazines and Q&A platforms are useful document types providing high F1-scores. For CMV, magazines about culture and research journals are beneficial, while in Kialo, general-domain magazines and forums are useful types. As we also observed in the earlier analysis of search methods, Wikipedia, which is conventionally used in fact verification, and mainstream news are relatively less reliable. So are reports that contain a lot of detailed information.

Attackability: Our system was originally designed in consideration of the scenario where we counter an argument by first detecting attackable sentences and then finding proper counterevi-

[8]A network location is generally the part in a URL that directly follows the URL scheme (e.g., "http") until the first slash; for instance, "www.cnn.com", "docs.python.org".

Type	Description	Examples
Mainstream News	Mainstream news about daily issues and general topics.	www.cnn.com, www.bbc.com
Research Journal	Peer-reviewed papers or dissertations.	link.springer.com, www.nature.com
Report	Surveys, statistics, and reports. Should be a source of substantial data rather than a summary of reports.	www.whitehouse.gov, www.irs.gov, www.cdc.gov
Personal Blog	Personal blogs.	medium.com, jamesclear.com
Magazine–Psychology	Magazines about psychology, mental health, relationships, family.	www.psychologytoday.com
Magazine–Society	Magazines about social and political issues.	www.hrw.org, www.pewresearch.org
Magazine–Finance	Magazines about finance, business, management.	www.hbr.org
Magazine–Culture	Magazines about culture, education, entertainment, fashion, art.	www.vulture.com
Magazine–Scitech	Magazines about science, medicine, technology.	www.techdirt.com, www.webmd.com
Magazine–General	Magazines about multiple domains.	thedickinsonian.com
Knowledge Archives	Information archives for knowledge transfer, such as encyclopedias, books, dictionaries, lectures.	plato.stanford.edu, quizlet.com, www.wikihow.com
Q&A	Question and answering platforms.	stackoverflow.com, www.quora.com
Forum	Forums for opinion sharing and reviews.	www.reddit.com, www.debate.org
Broken	URLs are not accessible.	

Table 8.5: Evidence document types.

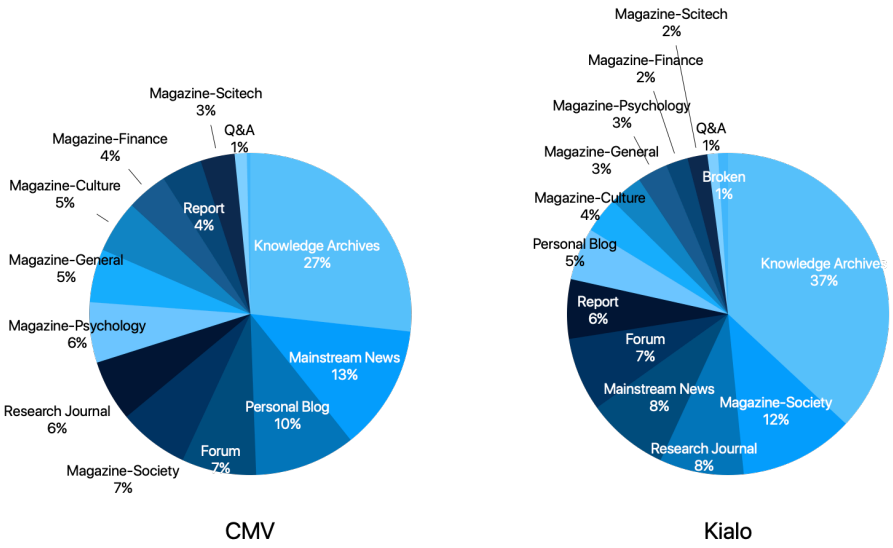


Figure 8.4: Distribution of document types for valid counterevidence.

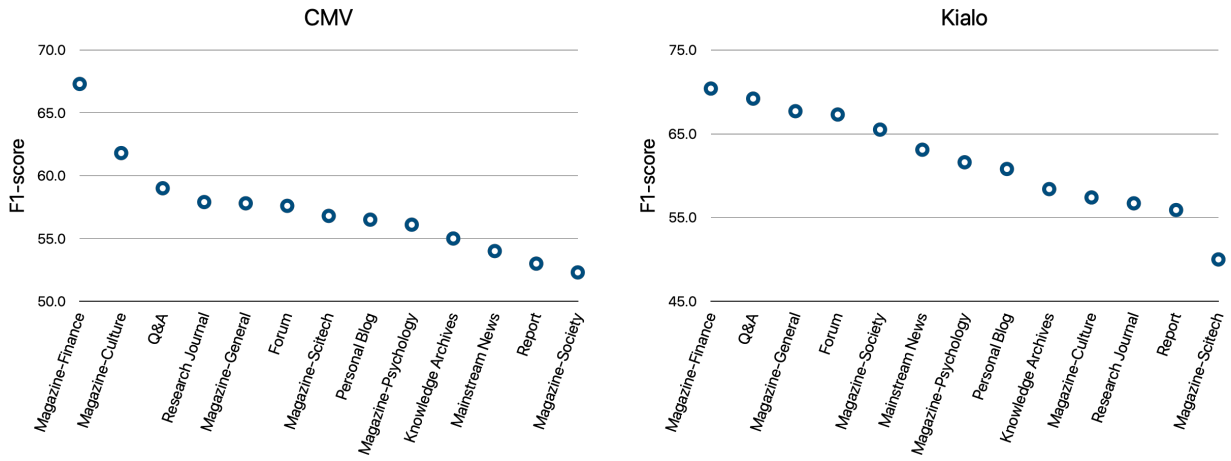


Figure 8.5: F1-scores of counterevidence retrieval by evidence document types.

dence to them. Detecting attackable sentences in arguments has been studied for CMV based on persuasion outcomes in §7.4. Here, we test if this method can help us find statements for which counterevidence exists.

We assume that statements in our dataset are attackable if they have at least one candidate sentence that is valid counterevidence. Figure 8.6 shows the distribution of the attackability scores of statements for which counterevidence was found (Found) and statements for which no counterevidence was found (Not Found). As expected, the attackability scores of statements that have counterevidence are higher than the other statements for both CMV ($p = 0.001$) and Kialo ($p = 0.003$ by the Wilcoxon rank-sum test).

One reason that the attackability scores have high variance is that the attackability detection model tends to predict scores that are close to either end (0 or 1). This has to do with the binary

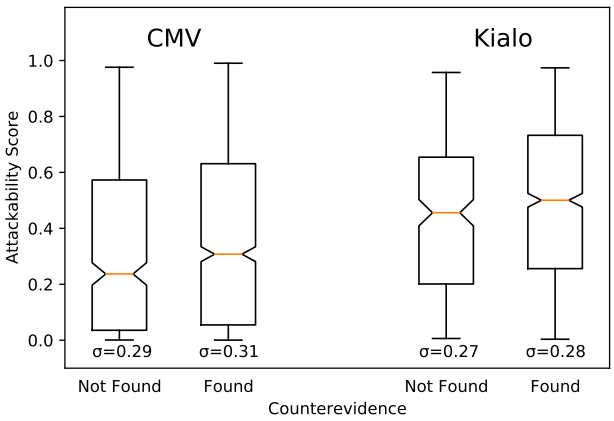


Figure 8.6: Attackability.

classification setting of the problem. One way to rectify this issue is to relax the binary setting to regression, that is, the model predicts an attackability score on the continuum between 0 and 1. But this introduces the additional challenge of data annotation, since each sentence must be annotated with the degree of attackability, which requires very careful guidelines to avoid annotator biases and subjectivity.

The attackability score of each statement has a positive correlation with the number of candidate sentences that are valid counterevidence, resulting in Kendall's τ of 5.7% (CMV, $p = 0.002$) and 8.5% (Kialo, $p = 0.006$). These results suggest that synergies can be made by integrating our system with attackability detection to build a complete counterargument generation system. We leave this direction to future work.

8.5 Appendix: Annotation Tasks

8.5.1 Annotation Principle

For all annotation tasks, we recruited annotators on Amazon Mechanical Turk (MTurk). Participants should meet the following qualifications: (1) residents of the U.S., (2) at least 500 HITs approved, and (3) HIT approval rate greater than 97%. Each HIT includes several questions and one attention question. The attention question asks the annotator to select a specific option, and we rejected and discarded a HIT if the annotator failed the attention question.

8.5.2 Annotation of Example-Based NLI data

This section describes our method for synthetic building of example-based NLI data that was augmented with the public NLI datasets in our experiments. The entire process consists of two steps. First, we generate synthetic example-based pairs using a pretrained language model (§8.5.2). Next, we annotate their labels using MTurk (§8.5.2).

ID	Statement	Perplexity
P	a breakdancer man is performing for the kids at school	3.08
P'_1	a breakdancer man is performing for the kids at licensed victuallers' school	2.67
P'_2	a breakdancer man is performing for the kids at preschool	2.95
P'_3	a breakdancer man is performing for the kids at boys republic	3.09
P'_4	a breakdancer man is performing for the kids at language teaching	3.10

Table 8.6: Examples of generated example-based statement and its perplexity measured by GPT2.

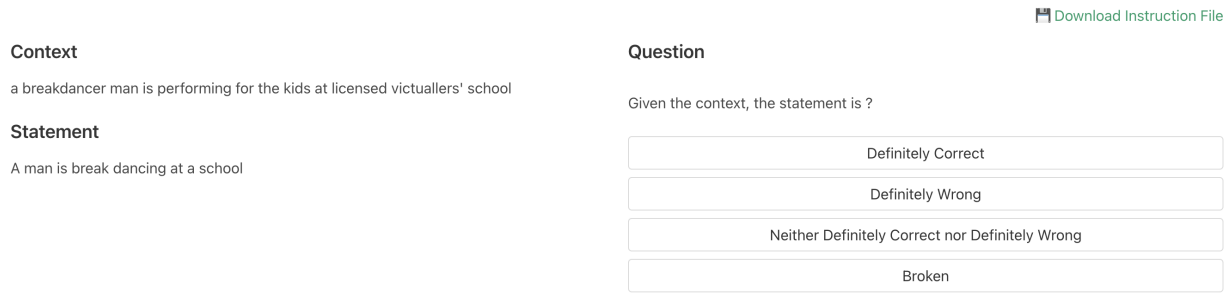


Figure 8.7: Example-based NLI labeling page.

Generating Synthetic NLI Pairs

We synthetically generate example-based NLI pairs as follows. Given a pair of P and H in the public datasets in Table 8.1, we modify P to P' by replacing an entity that occurs in both P and H with an incoming entity on WikidataEx. For example, for the following pair

> ***P*:** "a breakdancer man is performing for the kids at school"
>
> ***H*:** "a man is break dancing at a school"

"school" occurs in both P and H, so we may generate P' by replacing "school" with an instance of the school (e.g., "preschool") based on WikidataEx. To avoid broken or implausible sentences, we retain P' only if its perplexity is lower than or equal to that of P based on GPT2. Table 8.6 shows examples of synthetically generated P' and their perplexity. P is the original statement from the SNLI dataset, and P'_1–P'_4 are generated statements after the entity "school" is replaced. The perplexity of P'_1 and P'_2 is lower than that of the original statement P, so we pair each of them with H and add the pairs to our synthetic NLI data. However, P'_3 and P'_4 are discarded because their perplexity is higher than that of P.

Label Annotation

For each of the generated NLI pairs, we ask annotators whether H is correct or wrong given the context P'. They can choose from the four options: definitely correct (entail), definitely wrong (contradict), neither definitely correct nor definitely wrong (neutral), and broken English (Figure 8.7). Each HIT consists of 10 pairs and one attention question. Each pair is labeled by three annotators and is discarded if the three annotators all choose different labels.

		Original Label		
		Entail	Neutral	Contradict
New Label	Entail	1,698	548	373
	Neutral	228	543	139
	Contradict	151	302	1,030
	Broken	20	15	24
	No Majority	336	333	291

Table 8.7: Confusion matrix of example-based NLI data labels.

Analysis

To see how the labels of the generated pairs (P', H) differ from the labels of their original pairs (P, H), we manually analyzed 6,031 pairs (Table 8.7). Only 59 sentences were labeled as broken, meaning that our GPT2-based generation method effectively generates sensible statements P'. Most original pairs of entailment and contradiction keep their labels, but many of originally neutral pairs turn to either entailment or contradiction after entity replacement.

8.5.3 Annotation of Evidence Validity

In this task, annotators were asked to mark how strongly each counterevidence candidate sentence refutes the statement it attempts to refute (i.e., statements from CMV or Kialo). The four options of strength are very weak, weak, strong, and very strong, with corresponding scores 0, 1, 2, and 3 (Figures 8.8 and 8.9). For each statement from CMV, the entire post is displayed with the target statement highlighted so the annotator can consider the context of the statement when making a decision. For each candidate sentence, the annotators should also answer whether reading the source document is necessary to make a judgment.

Each HIT includes four statements to refute, along with at most four candidate counterevidence sentences for each statement, and one attention question. Each candidate sentence was labeled by two annotators. If a candidate sentence was labeled as both *very weak* and *very strong*, we treated the labels as unreliable (146 candidates in 131 sentences from CMV, 71 candidates in 65 statements from Kialo) and allocated a third annotator. We average their scores, which becomes the candidate sentence's final strength. The average variance of scores for each candidate sentence is 0.48, meaning that annotators on average have a score difference less than 1 point.

8.5.4 Annotation of Document Types

In this task, we annotate the type of source document for each candidate sentence. Each annotator was shown the network location identifier of a URL (e.g., "www.cnn.com", "docs.python.org") and asked to choose the type of the site from 14 categories (Table 8.5 and Figure 8.10). Total 1,987 unique location identifiers were annotated. Each HIT consists of 10 identifiers and one attention question. Each identifier was annotated until two annotators chose the same category. If

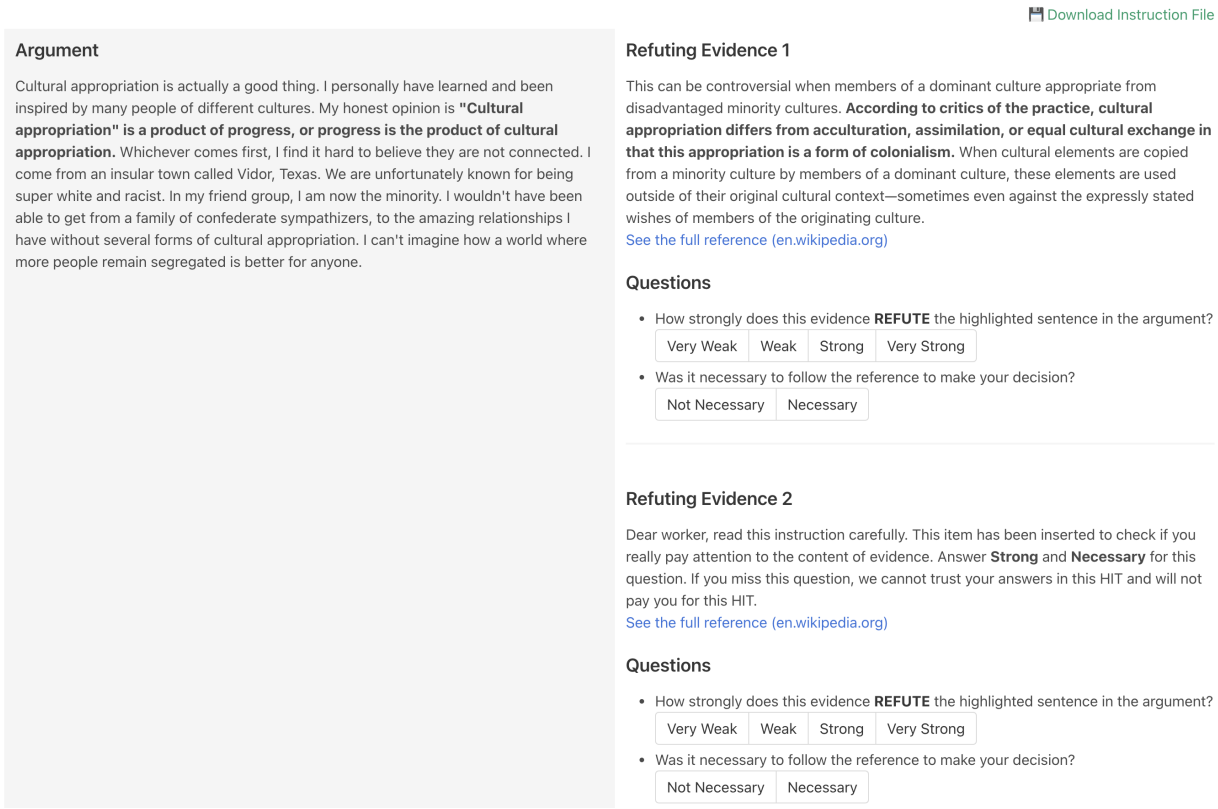


Figure 8.8: CMV evaluation page. Refuting Evidence 2 is an attention question.

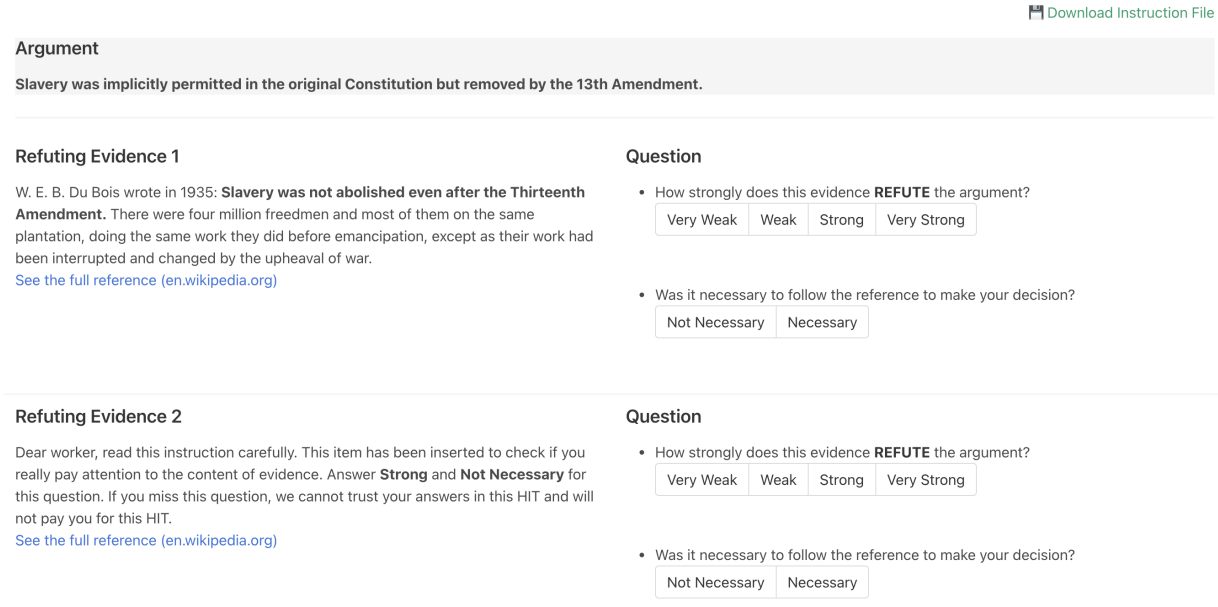


Figure 8.9: Kialo evaluation page. The second question is an attention question.

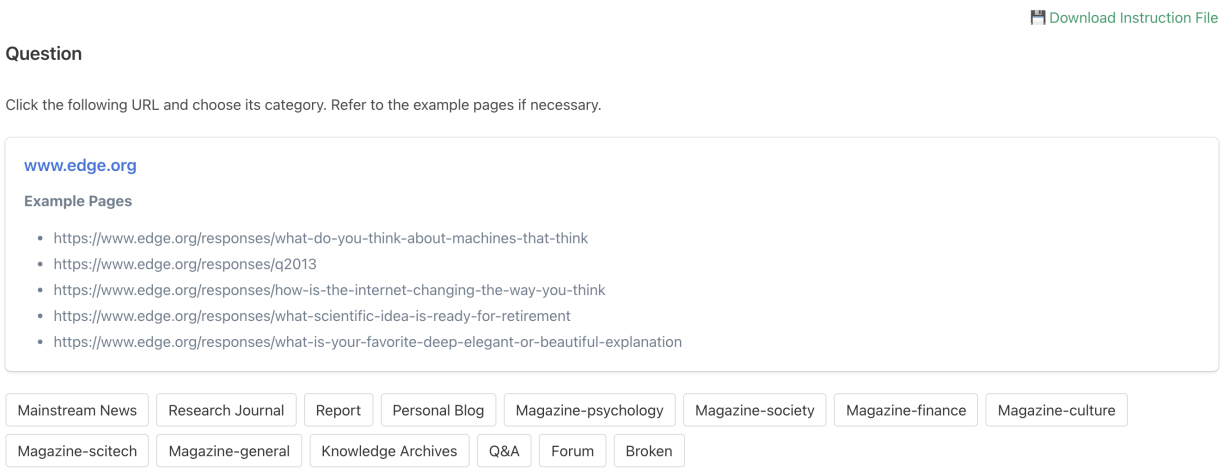


Figure 8.10: Document type annotation page.

there was no such category for five annotators, we selected the decision of the most "trustworthy" annotator, who had the highest rate of decisions selected for the other identifiers.

8.5.5 Ethical Considerations on Human Annotation

We consider ethical issues on our annotation tasks. The first consideration is fair wages. We compute the average time per HIT based on a small pilot study, and set the wage per HIT to be above the federal minimum wage in the U.S. ($7.25[9]). Table 8.8 shows that the expected hourly wage is higher than the federal minimum wage for all the annotation tasks.

We also preserve the privacy of crowdworkers. We do not ask for their personal information, such as names and gender. We collect Worker IDs to map each HIT result with the annotator and to accept or reject their work on MTurk. But the Worker IDs are discarded afterward to preserve their privacy.

Our annotation tasks are upfront and transparent with annotators. We provide the instruction manual of each task at the starting page, which informs the annotators of various task information, such as an estimated time needed for the task. Some annotators complained when their work was rejected. We generally responded within a business day with evidence of our decision (i.e., their failure of the attention question).

8.6 Conclusion

In this chapter, we built a counterevidence retrieval system. To allow the system to retrieve counterevidence that involves complex inference, we presented a knowledge-enhanced NLI model with specific focus on causality- and example-based inference. The NLI model demonstrates improved performance for NLI tasks, especially for instances that require the targeted inference.

[9] https://www.dol.gov/general/topic/wages/minimumwage

Task	# Questions/HIT	Time/HIT (secs)	Wage/HIT	Expected Hourly Wage
Example-based NLI	10	324	$0.7	$7.78
Evidence Validity – CMV	4	247	$0.5	$7.28
Evidence Validity – Kialo	4	240	$0.5	$7.50
Document Type	10	351	$0.5	$7.79

Table 8.8: Expected hourly wage of each annotation task. All wages are over the federal minimum wage in the U.S. ($7.25). The number of questions per HIT does not include attention questions.

Integrating the NLI model into the retrieval system further improves counterevidence retrieval performance, especially recall, showing the effectiveness and utility of our method of incorporating knowledge graphs in NLI.

Part IV

Conclusion

Chapter 9

Conclusion

This thesis began with an urgent call for a better understanding of human reasoning reflected in language and of methods for incorporating such reasoning and knowledge into computational models. In response, we studied one of the most common communication modes in human life that is full of reasoning: argumentation. Overall, the thesis covers three aspects of argumentation, from individual propositions, to argumentative relations between statements, to counter-argumentation. Throughout the study, we focused on and tried to shed light on human reasoning reflected in argumentative text. In addition, in later chapters, we incorporated human reasoning and knowledge into computational models to improve their performance and fidelity. A lot of parts in this study are informed by argumentation theory, making a nice connection among argumentation theory, language technologies, and computational linguistics.

9.1 Summary of Findings

Propositions: Meaning, Types, and Effects In Part I, we examined the building blocks of argument—asserted propositions—in terms of their meaning, types, and effects. In Chapter 2, we presented a cascade model to recover asserted propositions in argumentative discourse, by resolving anaphors, identifying meaningful locutions, recovering implicitly asserted propositions in reported speech, questions, and imperatives, reconstructing missing subjects, and revising the output. Our main findings are as follows. Anaphora resolution is crucial for recovering the semantic information of propositions, and the main bottleneck is to resolve 2nd-person singular and 3rd-person gender-neutral pronouns. Locution boundaries are often confused around clause boundaries, but they are generally well identified. Identifying speech source and content from reported speech is highly reliable, whereas recovering asserted propositions from questions and imperatives still have a long way to go. They both suffer from a scarcity of training data, and especially imperatives first need a strong theoretical foundation for data collection. For subject reconstruction, the tracing method is fairly effective, and the accuracy is bounded mainly by the robustness of dependency parsing to ill-formed and complex sentences. The final revision remains mostly grammar error correction, and substantial semantic revision (if needed) may require significantly different approaches. The recovered asserted propositions allow for a clear

picture of argumentation structure and provide transparency to downstream tasks.

In Chapters 3–4, we presented a methodology for identifying surface types of propositions and analyzing their associations with argumentation outcomes. In Chapter 3, we developed an unsupervised model for learning surface types that underlie a set of dialogues. The main assumption is that different compound illocutionary acts have different compositions of surface types at proposition levels, and we can identify such types well via good modeling of compound illocutionary acts. We additionally assumed that illocutionary acts transition faster than the background topic of discussion, and each speaker has their own preferences for certain illocutionary acts. This model demonstrated better performance in identifying illocutionary acts than previous models and identified reasonable surface types in data.

In Chapter 4, we applied the model on four argumentative corpora: Wikipedia discussions, political debates on Ravelry, persuasion dialogues on Reddit, and U.S. presidential debates. As a result, we found 24 generic surface types in argumentation, such as mega-argumentation, feelings, numbers, and comparison (Table 4.1). Using these surface types, we further conducted case studies to better understand their associations with argumentation outcomes. Our main findings are as follows:

- In the case study of Wikipedia discussion, we found that surface types contribute to defining different roles of Wikipedia editors, and we identified five main roles of Wikipedia editors: moderator, architect, policy wonk, wordsmith, and expert. Combining these roles yielded better performance in predicting the success of editing than prior models.
- In the case study of political debates in Ravelry, nine surface types were found to be perceived inappropriate often in political debates, such as meta-argumentation, argument-evaluation, direct question, direct mention of the conversation partner, using racial terms, and expression of feelings. Using these surface types, we found evidence that moderators in these debates have biases against minority opinions.
- In the case study of persuasion dialogues on Reddit, some surface types were found to be associated with successful and unsuccessful persuasion. From the persuader's side, presenting concrete numbers, references, definitions, different choices, and comparisons are associated with successful persuasion. From the persuadee's side, expression of confusion is related to successful persuasion, but using numbers and emphasizing specific terms are related to unsuccessful persuasion.
- In the case study of U.S. presidential debates, we found relatively weak associations between surface types and formation of pro-/counter-arguments. Presenting personal stories and future needs is more likely to be used to support a claim, whereas expression of disagreement is used mainly to attack a claim. But other than that, formation of pro- and counter-arguments is not explained well by surface types; it rather seems to require understanding the content of propositions.

These case studies provide insights into how people reason and present their ideas in argumentation. They also show the potential of surface types as an analytic tool for understanding and diagnosing argumentation, which could be incorporated into decision support systems.

Argumentative Relations In Part II, we took a step further from individual propositions and examined argumentative relations (support, attack, and neutral) between statements via the lens of argumentation schemes and logical mechanisms. In Chapter 5, we addressed a big challenge in application of argumentation schemes to computational linguistics: annotation. We developed a human-machine hybrid annotation protocol and applied it to annotation of four main types of statements in argumentation schemes: normative statement, desire, future possibility, and reported speech. By training a machine annotator, this protocol allowed for efficient and reliable annotation for difficult annotation tasks involving complex reasoning and rare occurrences of positive instances. The machine annotator effectively identified negative instances with high accuracy, which allowed human annotators to prioritize their resources for the more challenging task of identifying positive instances. Furthermore, machine annotator can be used to validate final human annotations, enhancing both the speed and inter-annotator reliability of output annotations. A case study of these annotations from U.S. presidential debates demonstrates natural affinities between these statement types to form arguments and argumentation schemes. For example, people tend to use the same statement type for both claim and premise. In addition, different parings of statement types result in different argumentation schemes (e.g., normative claim and premise form *practical reasoning* and normative claim and future possibility premise form *argument from consequences*). The statement types were also used to analyze the rhetorical strategies of presidential candidates, revealing highly normative tone of democratic candidates.

In Chapter 6, we investigated various mechanisms in argumentative relations between statements. We hypothesized four logical mechanisms informed by computational linguistics and argumentation theory: factual consistency, sentiment coherence, causal relation, and normative relation. They were operationalized through semantic modules (classifiers) trained on separate datasets, which can be seen as exploiting "soft" knowledge embedded in natural-language datasets and necessary for commonsense reasoning. These mechanisms were found to effectively explain argumentative relations, especially normative relation and sentiment coherence. Next, they were incorporated into a supervised classifier through representation learning, showing higher predictive power than models that do not incorporate these mechanisms or that incorporate the mechanisms using different methods. The resulting model learns good representations of input arguments that make intuitive correlations between logical relations and argumentative relations. Alongside the model itself, we also developed a rich annotation protocol for the argumentation schemes *argument consequences* and *practical reasoning*, which was needed to operationalize normative relation. The annotation protocol provides reliable inter-annotator agreement and contributes to the literature of argumentation scheme annotation.

Counter-Argumentation In Part III, we took a closer look at counter-argumentation and studied counterargument generation as a three-step process: detecting attackable sentences in a target argument, finding counterevidence to the sentences, and combining the counterevidence to a fluent and coherent argument. This thesis focused on the first two steps. In Chapter 7, we presented two computational models to detect attackable sentences in arguments using persuasion outcomes as guidance. The first model uses neural representations of sentences and jointly models the attackability of each sentence and the interaction of sentences between attacking and attacked arguments. The model detects attackable sentences effectively, and modeling the attackability

improves prediction of persuasion outcomes. The second model turns to a more interpretable representation of sentences, featurizing various characteristics relevant to sentence attackability. We found interesting associations between sentence characteristics and attackability. For instance, seemingly evidenced sentences (e.g., using data, references, and definitions) are more effective to attack. Although attacking these sentences may require even stronger evidence and deeper knowledge, arguers seem to change their viewpoints when a fact they believe with evidence is undermined. It is also very important to identify and address the arguer's confusion and uncertainty. Challengers are often attracted to subjective and negative sentences with high arousal, but successfully attacked sentences have rather lower subjectivity and arousal, and have no difference in negativity compared to unsuccessfully attacked sentences. Furthermore, challengers tend to pay less attention to personal stories, while successful attacks address personal stories more often.

After finding attackable points, in Chapter 8, we built a system for retrieving counterevidence from various sources on the Web, using Wikipedia, Microsoft Bing, and Google. At the core of this system is a natural language inference (NLI) system that classifies whether each candidate sentence is valid counterevidence or not. To overcome the limitation of most NLI systems—a lack of reasoning abilities—we presented a knowledge-enhanced NLI model that focuses on causality- and example-based reasoning and incorporates relevant knowledge graphs. The main idea is to reduce semantic gaps between words used in two statements by bridging them via a knowledge graph. This NLI model demonstrated improved performance in NLI tasks, especially for instances that require the targeted reasoning. Integrating this NLI model into the retrieval system also improved the retrieval performance, especially recall. We also explored the utility of our argumentative relation classifier from Chapter 6 in this retrieval system, showing its comparable performance to the knowledge-enhanced NLI model, especially for precision. The result suggests that different knowledge graphs can help the model enlarge its coverage to capture nontrivial cases of relations and that different kinds of signals (e.g., sentiment, normative relation) can help the model make more precise predictions by adjusting its decision based on different angles. Lastly, we tested our attackability detection model from Chapter 7 in this task and found that statements with higher attackability scores tend to have more instances of counterevidence. It suggests the promising direction of combining the attackability detection model and the counterevidence retrieval system into a complete counterargument generation system.

9.2 Sensitivity to Argumentation Types and Domains

Argumentative dialogue can be classified into various types, such as persuasion, negotiation, information-seeking, deliberation, and eristic debates (Walton and Krabbe, 1995). Different types have different goals and thus different representations of rhetorical strategies. For instance, information-seeking and pedagogical argumentation may include more pure questions, whereas eristic debates may face more rhetorical questions. Negotiation and deliberative argumentation may include more normative and action-oriented statements than information-seeking argumentation. Overall, this thesis aims to cover topics and computational models that are general in

argumentation. For instance, the methodology for identifying surface types (Chapter 3) and their associations with argumentation outcomes (Chapter 4) does not make any assumptions about the type of argumentation and was applied to four corpora across different types. Similarly, argumentative relation classification (Chapter 6) aims to focus on the meaning of statements and identify their relations using logical mechanisms rather than relying on the type of argumentation.

On the other hand, some parts of the thesis may be more sensitive to argumentation types. For instance, the proposition extraction model (Chapter 2) that is trained on a certain type of argumentation may not generalize to other types, although the model itself is type-agnostic. One example is extracting propositions asserted in questions, for which the model was trained and tested on political debates that include a lot of rhetorical questions than pure questions. In this case, the model performance is expected to be sensitive to the type of argumentation it is trained on. In addition, the attackability detection models (Chapter 7) mainly target persuasion dialogue and use the definition of attackable sentences being sentences that are likely to change the arguer's viewpoints when attacked. The thesis, however, does not deeply examine what sentences should be considered attackable in eristic or deliberative argumentation. Hence, more theoretical and empirical considerations are necessary to develop fully type-general argumentation technologies.

Regarding domains or topics of argumentation, argumentation generally requires domain-specific knowledge. Due to different vocabularies and knowledge used in different domains, most models in this thesis may require domain-specific training data in order to achieve high performance. An encouraging observation, however, is that a large portion of argumentative relations is explained by sentiment coherence and normative relation (§6.7.3), which are less specific to individual domains than factual consistency and causal relation; of course, even sentiment analysis requires a certain degree of domain knowledge. At the same time, this observation implies that less normative domains (e.g., scientific discussions) may require more domain-specific training data. Developing methods that are domain-independent or robust across domains would be an important future direction.

9.3 Theoretical and Practical Implications

9.3.1 Theoretical Implications for Argumentation Theory and Social Sciences

Argumentation theorists categorize and study argumentative patterns (e.g., argumentation schemes), but their choice of patterns and study subjects is often based heavily on their instincts. This thesis, on the other hand, can provide more data-driven and systematic ways of identifying and categorizing such patterns. For example, the surface type identification model in Chapter 3 aims to categorize surface types generally used in argumentative dialogue in a data-driven manner, by imposing certain assumptions like speaker preferences and dependency between consecutive compound illocutionary acts. As a result, the model identifies 24 generic surface types that have distinctive lexical and dependency patterns, which have not been presented as such in argumentation theory to our knowledge. Some of these surface types (e.g., presenting statistics and references) have already been examined by argumentation theorists (Hoeken and Hustinx,

2009-10; Janier and Reed, 2017), but many of them have received little attention and remain understudied despite their prevalence in argumentative discourse (e.g., presenting different choices, making comparisons). Hence, along with an analysis of associations between surface types and argumentation outcomes as in Chapter 4, our computational methodology could offer interesting and data-driven hypotheses about argumentative strategies reflected in surface types and open a way for theorists to conduct in-depth and qualitative analyses on them.

Argumentation theory has suggested many frameworks to explain argumentative phenomena, one of which is argumentation schemes. While argumentation schemes are intuitive and plausible, how much they can explain and constitute actual argumentation in practice has little been verified empirically and at scale. Our study of logical mechanisms (Chapter 6) reveals that normative argumentation schemes indeed contribute the most to argumentative relations among several other mechanisms in a large debate forum. Another framework proposed by argumentation theory is a taxonomy of argument quality (Wachsmuth et al., 2017a). While such a taxonomy presents general aspects of argument quality, such as credibility and emotional appeal, more practical and empirical evidence is still missing as to what kinds of sentences are important to address to change the arguer's viewpoint. Our large-scale analysis in Chapter 7 reveals interesting characteristics of sentences that attract attacks and lead to view changes. For instance, quite counterintuitively, sentences that are seemingly evidenced by data and references and that have objective tone are more attackable targets for a view change. On the other hand, challengers are often attracted to attack negative sentences, but sentiment is not a meaningful factor for view changes.

Parts of the thesis also ask important questions for argumentation theory. For instance, our proposition extraction model (Chapter 2) draws upon Inference Anchoring Theory (Budzynska and Reed, 2011) and examines the feasibility and practicality of translating the theory into computational methods. Our study spots the weakness of theoretical foundation for interpreting imperatives and provides an initial categorization of verbs depending on the applicability of the "you-should" theory. This study calls for a need for further theoretical and empirical studies while suggesting some annotation guidelines.

How to manage successful dialogues is an important topic in communication science. A lot of research focuses on what factors lead to successful and unsuccessful dialogue in various settings. Our study of surface types and censorship in debates (§4.5) algorithmically reveals that certain surface types, such as meta-argumentation and direct questions, are often perceived as inappropriate in political debates and provides evidence for moderation biases. The authors of this study are working on a book chapter titled "The Politics of Moderation: Perceived Censorship in Volunteer-Run Online Communities" for the book "Gendered Digital Labor" published by Routledge. Similarly, collaboration success in Wikipedia has been a popular subject among social scientists and information scientists (Kraut et al., 2011). Our study of Wikipedia (§4.4) identifies five roles of Wikipedia editors—moderator, architect, policy wonk, wordsmith, and expert-based on surface types and examines combinations of these roles for successful decision-making in Wikipedia discussions. Lastly, persuasion is a long-standing subject in rhetoric, business, and economics, and many books suggest different strategies for successful persuasion (Cialdini, 2009). Our large-scale analysis of persuasion (§4.2.3) examines how surface types used by persuaders and persuadees have associations with successful persuasion based on a large sample of discussions

in a real-world online forum. It offers the new insight that successful persuasion is associated with the persuader's presentation of numbers, references, and different choices, as well as the persuadee's expression of confusion, whereas the persuadee's use of confusion and emphasis on specific terms are associated with unsuccessful persuasion. All these studies provide systematic and data-driven explanations of argumentative phenomena at scale.

In addition, in Chapter 5, we applied our human-machine hybrid annotation protocol and identified various argumentative styles of U.S. presidential candidates in 2016. The study provides the new insight that Democratic candidates (Clinton and Sanders) have substantially high normative tone, whereas Republican candidate Trump has more balanced tone across normativity, desire, and reported speech. The study also provides a reliable and fast annotation protocol that communication researchers can adopt to investigate various communication styles in a similar manner to our study.

9.3.2 Practical Implications

Visualization of argumentative structure is of interest to many. Argument diagramming is a popular pedagogical tool that more and more educators adopt to teach argumentation skills (Harrell, 2011). Accordingly, HCI researchers also have built educational systems centered around argumentation structures (Wambsganss et al., 2020; Scheuer et al., 2010; Pinkwart et al., 2009). The users of such systems learn from the structures of example arguments, which can greatly benefit from automated methods for argumentation structure construction like ours in Chapter 6. Argumentation structure is of interest to the general public as well. For instance, kialo.com is an online collaborative argumentation platform to "engage in thoughtful discussion, understand different points of view, and help with collaborative decision-making" in its own words. Each discussion is represented as an argumentation tree where each node is an asserted proposition(s) and nodes are connected by the support or attack relation. Using proposition extraction (Chapter 4) and argumentative relation classification (Chapter 6) to automatically visualize such a tree from a naturally occurring discussion or debate would be a great helper to many of those users who like a well-structured representation of argumentation. Kialo also offers educational resources based on their argumentation trees (kialo-edu.com). News media, such as BBC, have been interested in analyzing the structure of editorials[1] (Figure 9.1) and educating readers to build sound arguments using support and attack relations[2].

Conversation analysis is a field of interest to social scientists and practitioners. Our analysis of surface types (Chapter 3) and their associations with moderation biases (§4.5), for example, can inform moderators in debates of their (subconscious) biases and provide an opportunity to inspect them. The association between surface types and successful persuasion (§4.6) could be extended to a practical support system that intervenes in ongoing argumentation and guides the argumentative moves of the arguers for successful persuasion, or could be used for marketing purposes. Surface type identification may become a useful application in psychotherapy as well, where cognitive behavior therapy is commonly used to identify certain types of reasoning, such as

[1] https://evidence-toolkit-moral-maze.pilots.bbcconnectedstudio.co.uk/
[2] https://www.bbc.co.uk/taster/pilots/moral-maze

catastrophizing. Automated identification of surface types may help therapists monitor clients' thought processes at scale (e.g., in daily lives) and make appropriate interruptions.

Detecting attackable sentences (Chapter 7) have potential use cases in editing tools (e.g., Grammarly) and essay scoring tools (Burstein, 2003). Such tools aim to spot and improve weaknesses of writing, and our attackability models may be able to serve that purpose at scale. Another area of practical applications is combating fake news. Fake news is an important societal issue today, and many efforts have been made to fight against it. The counterevidence retrieval system in Chapter 8 can clearly contribute to this effort, by retrieving counterevidence to statements we want to verify. This system may be of interest to security sectors in governments (Roudik et al., 2019), news media like BBC[3], and social media like Facebook[4].

Overall, argumentation is a common communication mode in human life. Hence, complete conversational AI (e.g., Amazon Alexa) and decision support systems would eventually need the ability to argue in decision-making. IBM recently has developed Project Debater, an AI system that debates with people on different topics (Slonim et al., 2021). A lot of information in the world is more nuanced than either right or wrong, and assessing such information requires diverse perspectives. In my view, AI technology will advance from delivering the right information to delivering a means to navigate and assess information. Therefore, the research on computational argumentation in this thesis hopefully provides important building blocks and techniques toward such practical AI systems.

9.4 Future of Argumentation Technology with Advanced Language Models

Today neural network models are fast developing, and traditional models have been outperformed by pretrained language models, such as BERT (Devlin et al., 2018). I believe that such advanced language models would increase the capacity of argumentation models, since they contain linguistic information and knowledge useful in argumentation. Especially syntax-oriented tasks like proposition extraction in Chapter 2 may benefit a lot from such language models. For example, we ran a post hoc analysis for question transformation (§2.5.4) using an advanced pretrained transformer model T5 (Raffel et al., 2019). In contrast to our original experiment, where a rule-based method outperformed a BiLSTM-based model, T5 effectively learned to translate questions to asserted propositions from a small size of data and outperformed the rule-based method. This evidence supports that pretrained language models are indeed quite data-efficient for syntax-oriented tasks. Similarly, we found that neural networks perform well for many components in the cascade model, such as locutions (§2.5.2) and reported speech (§2.5.3). Therefore, I see it a promising direction to replace the entire cascade model with an end-to-end neural network.

More semantic-oriented tasks like argumentative relation classification (Chapter 6) can also benefit from pretrained language models. In many experiments in this thesis, transformer models outperform non-transformer baselines that were previously state-of-the-art models. Furthermore,

[3] https://www.bbc.co.uk/taster/pilots/evidence-toolkit-moral-maze
[4] https://www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news

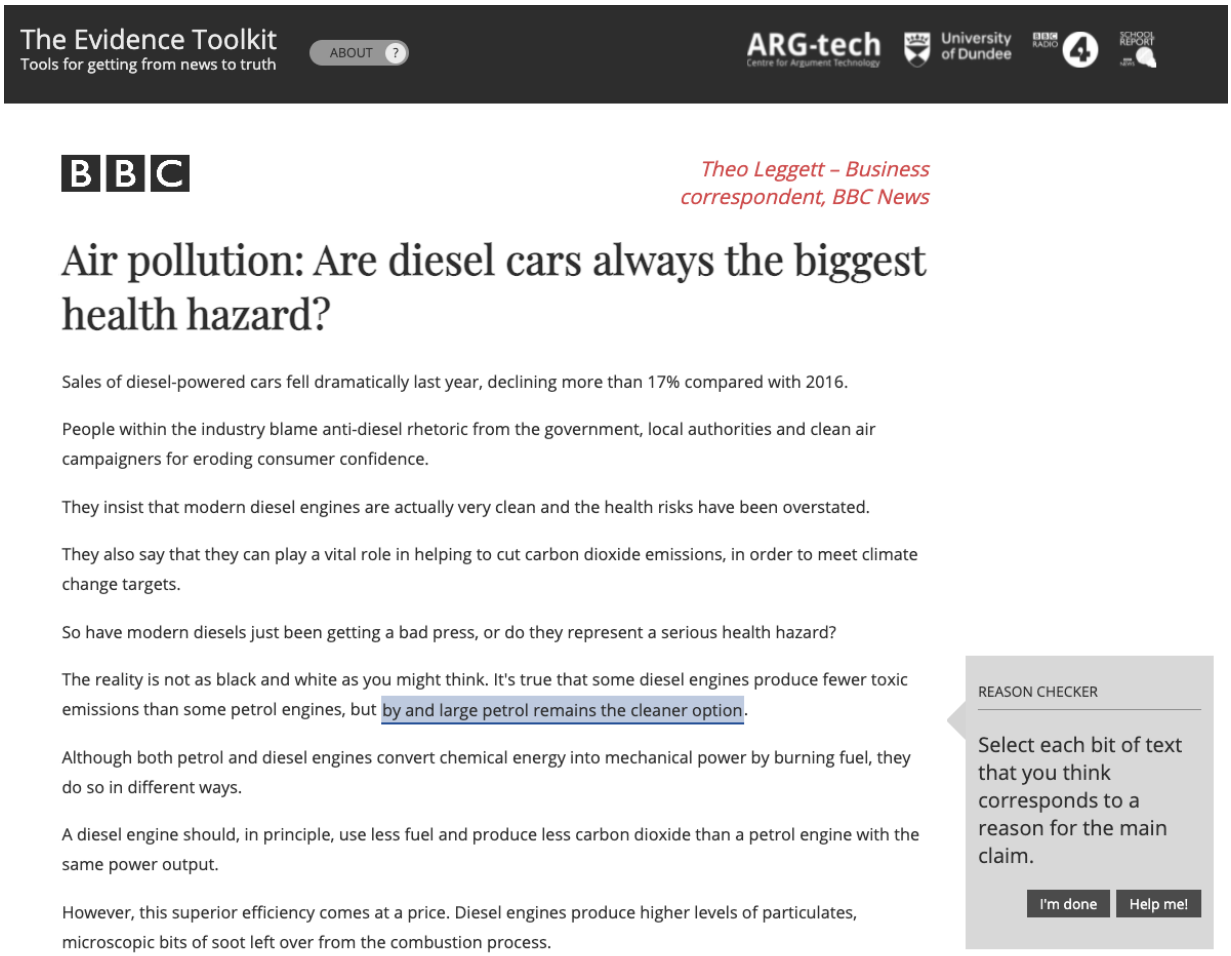


Figure 9.1: The Evidence Toolkit by BBC.

some parts of these tasks can be done in a very data-efficient way by levering those models. For instance, we collected a relatively few annotations (1,000 instances) for normative argumentation schemes in Chapter 6, and BERT trained on them operationalized normative relation quite effectively.

Although advanced language models show some promising data efficiency, data collection itself still remains very challenging in many argumentative tasks, and this could be a significant bottleneck. For example, we currently lack a strong theoretical foundation for asserted propositions in imperatives (§2.5.5), which makes it hard to even collect small data for this rhetorical device. Hence, I think it is still important to zoom in on individual components in an argumentative phenomenon as much as we can, understand linguistic phenomena there, and identify what's missing and what should be done further. This takes a long time; the annotation of statement types in Chapter 5 took about six months and the annotation of argumentation schemes in §6.5 took almost seven months, despite their relatively small sizes. Tasks that do not have solid theoretical foundations may take even longer. It is hopeful, however, that once we collect well-representative and solid data, we may be able to take advantage of the power of pretrained language models in efficient ways.

Despite the positive side of advanced pretrained language models, we do not have strong evidence that such language models would suffice without argumentation-specific representations in all argumentative tasks. Rather, there has been evidence that explicit integration of reasoning improves the argumentative capability of the language models. For example, GPT trained on deductive arguments better completes conclusions of arguments (Betz, 2020), and incorporating a knowledge graph about commonsense, such as cause-effect relations, improves predicting unseen events (Sap et al., 2019).

In addition, the good performance of neural models often comes at the cost of potentially spurious behavior behind the scenes. Especially for complex and semantic-oriented argumentative tasks, large pretrained language models have been found to overly rely on spurious cues. For example, a BERT-based model selects valid warrants for claims by picking up on annotation biases, such as use of negation (Niven and Kao, 2019), and stance detection models rely heavily on the sentiment of statements (Allaway and McKeown, 2020). All this evidence suggests that language models alone may not be sufficient for truthful argumentation technology. Some nice effects of imposing inductive biases and manipulating internal representations are to alleviate the model's dependency on such cues, as we trained our model on several logic tasks (Chapter 6), and to enable the model to use useful resources, as we infused external knowledge graphs into our model (Chapter 8). Another aspect to consider is that the success of argumentation depends on the participants and their value systems. Hence, psychology and mind games play an important role in certain argumentation, and inductive biases may still be useful to handle the complexity.

9.5 Future Directions

Causal Inference Our findings and model outcomes can potentially be integrated into a practical decision support system. For example, in Chapter 4, we identified various surface types and

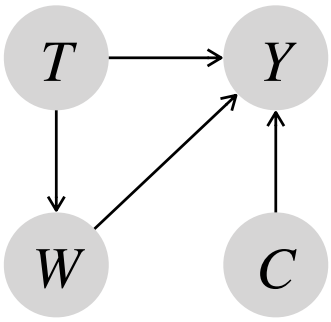


Figure 9.2: Representation of causal relationships.

how they are associated with different argumentation outcomes, such as successful persuasion. Based on the findings, a decision support system can intervene in ongoing argumentation and guide the argumentative moves of the speakers toward desirable outcomes. In order to translate these findings to a practical system, however, it is imperative to establish their *causal* impact on outcomes, beyond mere *correlations*. There is a growing literature on causal inference from observational text data (Egami et al., 2018; Zhang et al., 2020; Pryzant et al., 2020).

For instance, our work on detecting attackable sentences in arguments (Chapter 7) could be extended to find the causal effects of sentence characteristics on the arguer's view change. One possible approach is as follows. Let T be a continuous or binary variable that represents a characteristic of attacked sentences (e.g., sentiment score or being a question) and Y be the outcome of whether the attack is successful or not (e.g., whether the attacked arguer's view is changed or not). Our goal is to establish the causal effect of T on Y. However, the refutation text of the challenger W is a backdoor path from T to Y, since the characteristics of attacked sentences affect the content of the refutation, and the refutation affects the outcome of the attack. There are also external factors C that influence Y, such as the displayed reputation of the challenger (Manzoor et al., 2020). The relations among these variables can be represented graphically as in Figure 9.2.

Now, the average treatment effect (ATE) for a unit value of the treatment T (Δt) is

$$\begin{aligned}\psi &= \mathbb{E}[Y;T=t]-\mathbb{E}[Y;T=t-\Delta t]\\ &= \mathbb{E}_{W,C}[\mathbb{E}[Y;T=t,W]-\mathbb{E}[Y;T=t-\Delta t,W]]\\ &= \frac{1}{N}\sum_{i=1}^{N}\{P(Y=1|T=t_i,w_i)-P(Y=1|T=t_i-\Delta t,w_i)\},\end{aligned}$$

where $i=1,\cdots,N$ is the ith instance in the data, t_i is the treatment value, and w_i is the challenger's text. Note that a binary treatment can be handled naturally by setting $\Delta t=1$. For simplicity, we assume that T is accurately measurable. We estimate $P(Y=1|T,W)$ by training a neural classifier $f_\theta : T,W \rightarrow Y$ with trainable parameters θ (a model like BERT). Classifier errors may underestimate the effect size of the treatment but does not change the sign as long as the classifier is better than chance (Pryzant et al., 2020). We may improve the model by taking into account propensity scores and training the classifier with the additional objective to predict W (Shi et al.,

2019).

Argument Generation In Chapter 8, we built a counterevidence retrieval system. This system uses an NLI model to classify whether a candidate sentence entails, contradicts, or neither the input statement. Hence, the retrieval system can easily be extended to retrieve supporting evidence as well. This thesis did not cover combining retrieved evidence to make a fluent and coherent argument. There are at least two relevant fields in that direction. The first is natural language generation via retrieve-and-edit methods (Hossain et al., 2020; He et al., 2020b). Instead of generating text from a vector, these methods use existing text as a backbone and generate new text based on it. In our task, retrieved evidence statements may be used as a quality backbone. Another relevant field is to incorporate rhetorical strategies into argument generation (Wachsmuth et al., 2018a; Hua et al., 2019). Generating persuasive arguments requires prioritizing different pieces of evidence to suit the characteristics of the listener (Longpre et al., 2019; Durmus and Cardie, 2018, 2019). Eventually, we will need an evaluation method that is scalable and effectively measures the quality or impact of generated arguments. Recent studies on argument generation rely heavily on human evaluation or similarly-based metrics (Hua and Wang, 2018; Durmus and Cardie, 2018). A good evaluation metric may combine various aspects of argument, such as factuality (Goodrich et al., 2019) and personality traits of listeners (Shmueli-Scheuer et al., 2019).

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better Fine-Tuning by Reducing Representational Collapse. *arXiv* .

Alan Agresti and Maria Kateri. 2011. *Categorical Data Analysis*, Springer Berlin Heidelberg, pages 206–208. https://doi.org/10.1007/978-3-642-04898-2_161.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling Frames in Argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. pages 2922–2932. https://doi.org/10.18653/v1/d19-1290.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit Segmentation of Argumentative Texts. In *Proceedings of the 4th Workshop on Argument Mining*. Association for Computational Linguistics, Copenhagen, Denmark, pages 118–128.

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*. pages 91–96.

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-End Argumentation Knowledge Graph Construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 34, pages 7367–7374. https://doi.org/10.1609/aaai.v34i05.6231.

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of Argumentation Strategies across Topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1362–1368.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A News Editorial Corpus for Mining Argumentation Strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, pages 3433–3443.

Miltiadis Allamanis, Hao Peng, and Charles Sutton. 2016. A Convolutional Attention Network for Extreme Summarization of Source Code. In *International Conference on Machine Learning (ICML)*.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Omar Alonso, Catherine C Marshall, and Marc Najork. 2015. Debugging a crowdsourced task with low inter-rater agreement. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, pages 101–110.

D Scott Appling, Erica J Briscoe, Heather Hayes, and Rudolph L Mappus. 2013. Towards automated personality identification using speech acts. In *AAAI Workshop - Technical Report*.

Ofer Arazy, Johannes Daxenberger, Hila Lifshitz-Assaf, Oded Nov, and Iryna Gurevych. 2017. Turbulent Stability of Emergent Roles: The Dualistic Nature of Self-Organizing Knowledge Co-Production. *Information Systems Research* page Forthcoming.

Aristotle and George Alexander Kennedy. 2007. *On Rhetoric*. A Theory of Civic Discourse. Oxford University Press, USA.

John L. Austin. 1962. *How to Do Things with Words*. Clarendon Press.

John L Austin. 1975. *How to Do Things with Words*. Harvard University Press.

Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. *Journal of Machine Learning Research* 18(109):1–67.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, Manfred Stede, and Benno Stein. 2019. Computational Argumentation Synthesis as a Language Modeling Task. In *Proceedings of the 12th International Conference on Natural Language Generation*. pages 54–64. https://doi.org/10.18653/v1/w19-8607.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment Analysis of Political Tweets: Towards an Accurate Classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*. pages 49–58.

Max H Bazerman, George Loewenstein, and Don A Moore. 2002. Why good accountants do bad audits. *Harvard Business Review* 80(11):96–103.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Association for Computational Linguistics, Portland, Oregon, pages 48–57. https://www.aclweb.org/anthology/W11-0707.

Gregor Betz. 2020. Critical Thinking for Language Models. In *arXiv*.

Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H. Chi. 2018. Latent Cross: Making Use of Context in Recurrent Recommender Systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, page 46–54. https://doi.org/10.1145/3159652.3159727.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic Identification of Rhetorical Questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. pages 743–749. https://doi.org/10.3115/v1/p15-2122.

Simon Blackburn. 2016. *The Oxford Dictionary of Philosophy*. Oxford University Press.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR '03: The Journal of Machine Learning Research* 3.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 632–642. https://doi.org/10.18653/v1/d15-1075.

David B Bracewell, Marc T Tomlinson, Mary Brunson, Jesse Plymale, Jiajun Bracewell, and Daniel Boerger. 2012. Annotation of Adversarial and Collegial Social Actions in Discourse. *6th Linguistic Annotation Workshop* pages 184–192.

Tomáš Brychcín and Pavel Král. 2017. Unsupervised dialogue act induction using gaussian mixtures. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 485–490. http://www.aclweb.org/anthology/E17-2078.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3):904–911.

Katarzyna Budzynska and Chris Reed. 2011. Whence inference. *University of Dundee Technical Report* .

Jill Burstein. 2003. The E-rater® scoring engine: Automated essay scoring with natural language processing. *Automated essay scoring: A cross-disciplinary perspective* .

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From Discourse Analysis to Argumentation Schemes and Back: Relations and Differences. In *Computational Logic in Multi-Agent Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 1–17.

Elena Cabrio and Serena Villata. 2012. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 208–212.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A Study of the Impact of Persuasive Argumentation in Political Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1405–1413.

David S Carrell, David J Cronkite, Bradley A Malin, John S Aberdeen, and Lynette Hirschman. 2016. Is the juice worth the squeeze? costs and benefits of multiple human annotators for clinical text de-identification. *Methods of information in medicine* 55(04):356–364.

Stephen L Carter. 1998. *Civility: Manners, morals, and the etiquette of democracy*. Basic Books (AZ).

A Chadwick. 2006. *Internet Politics: States, Citizens, and New Communication Technologies*. Oxford UP.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 2926–2936.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *CSCW 2017*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, ACL, pages 1870–1879. https://doi.org/10.18653/v1/p17-1171.

Di Chen, Jiachen Du, Lidong Bing, and Ruifeng Xu. 2018a. Hybrid Neural Attention for Agreement/Disagreement Inference in Online Debates. In *Proceedings of the 2018 Conference of Empirical Methods in Natural Language Processing*. pages 665–670.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018b. Neural Natural Language Inference Models Enhanced with External Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. NAACL, pages 2406–2417. https://doi.org/10.18653/v1/p18-1224.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, pages 103–111.

HongSeok Choi and Hyunju Lee. 2018. GIST at SemEval-2018 Task 12: A network transferring inference knowledge to Argument Reasoning Comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, New Orleans, Louisiana, pages 773–777.

R.B. Cialdini. 2009. *Influence: The Psychology of Persuasion*. Collins Business Essentials. HarperCollins e-books. https://books.google.co.kr/books?id=5dfv0HJ1TEoC.

Oana Cocarascu and Francesca Toni. 2018. Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics* 44(4):833–858.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Lo i c Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 681–691.

Irving M Copi, Carl Cohen, and Kenneth McMahon. 2016. *Introduction to Logic*. Routledge, 14 edition.

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* .

Ticiana L Coelho da Silva, Regis Pires Magalhães, José AF de Macêdo, David Araújo, Natanael Araújo, Vinicius de Melo, Pedro Olímpio, Paulo Rego, and Aloisio Vieira Lira Neto. 2019. Improving named entity recognition using deep learning with human in the loop. In *EDBT*. pages 594–597.

Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, page 985–988. https://doi.org/10.1145/3331184.3331303.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA, page 699–708. https://doi.org/10.1145/2187836.2187931.

Richard Davis. 1999. *The web of politics: The Internet's impact on the American political system*. Oxford UP.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the Association for Computational Linguistics*. pages 120–125.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* .

Tao Ding and Shimei Pan. 2016. Personalized Emphasis Framing for Persuasive Message Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1432–1441.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 49–54.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations. In *Proceedings of the 11th Linguistic Annotation Workshop*. pages 95–104. https://doi.org/10.18653/v1/w17-0812.

Amanda M Durik, M Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The Effects of Hedges in Persuasive Arguments: A Nuanced Analysis of Language. *Journal of Language and Social Psychology* 27(3):217–234.

Esin Durmus and Claire Cardie. 2018. Exploring the Role of Prior Beliefs for Argument Persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 1035–1045.

Esin Durmus and Claire Cardie. 2019. A Corpus for Modeling User and Language Effects in Argumentation on Online Debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 602–607. https://doi.org/10.18653/v1/p19-1057.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining Relative Argument Specificity and Stance for Complex Argumentative Structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 4630–4641.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *COMMA*. pages 299–310.

Arthur R Edwards. 2002. The moderator as an emerging democratic intermediary: The role of the moderator in Internet discussions about public issues. *Information Polity* 7(1):3–20.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163* .

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and Analyzing Semantic Role of Elementary Units and Relations in Online Persuasive Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 422–428.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 11–22. https://doi.org/10.18653/v1/p17-1002.

Aysu Ezen-Can and Kristy Elizabeth Boyer. 2015. Understanding Student Language: An Unsupervised Dialogue Act Classification Approach. *JEDM - Journal of Educational Data Mining* 7(1):51–78.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 171–175.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 987–996.

Oliver Ferschke. 2014. *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. Ph.D. thesis, Technische Universität, Darmstadt.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics EACL 2012* .

Oliver Ferschke, Diyi Yang, and Carolyn P. Rosé. 2015. A Lightly Supervised Approach to Role Identification in Wikipedia Talk Page Discussions. *Ninth International AAAI Conference on Web and Social Media* pages 43–47.

Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. 2011. Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. In *Proceedings of the ACL-HLT 2011 System Demonstrations*. Association for Computational Linguistics, Portland, Oregon, pages 97–102. http://www.aclweb.org/anthology/P11-4017.

Eveline T Feteris, Feteris, and Olivier. 2017. *Fundamentals of legal argumentation*, volume 1. Springer.

Eric N Forsythand and Craig H Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. In *International Conference on Semantic Computing (ICSC 2007)*. pages 19–26.

Johanna Frau, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2019. Different flavors of attention networks for argument mining. In *Proceedings of FLAIRS*.

James B Freeman. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. A Theory of Argument Structure. Walter de Gruyter, Berlin, Boston.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Fran c ois Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* 17(59):1–35.

Debela Gemechu and Chris Reed. 2019. Decompositional Argument Mining: A General Purpose Approach for Argument Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 516–526.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report* .

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pages 609–614.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing The Factual Accuracy of Generated Text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, USA, KDD, page 166–175. https://doi.org/10.1145/3292500.3330955.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O K Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1631–1640.

Yunfan Gu, Zhongyu Wei, Maoran Xu, Hao Fu, Yang Liu, and Xuan-Jing Huang. 2018. Incorporating Topic Aspects for Online Comment Convincingness Evaluation. *Proceedings of the 5th Workshop on Argument Mining* pages 97–104.

Zhen Guo, Zhe Zhang, and Munindar Singh. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In *Proceedings of The Web Conference 2020*. Association for Computing Machinery, New York, NY, USA, WWW, page 2388–2399. https://doi.org/10.1145/3366423.3380302.

Gahgene Gweon, Carolyn Penstein Rosé, Joerg Wittwer, and Matthias Nueckles. 2005. Supporting Efficient and Reliable Content Analysis Using Automatic Text Processing Technology. In *Human-Computer Interaction–INTERACT 2005*.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1214–1223.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1589–1599.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics* 43(1):125–179. https://doi.org/10.1162/coli_a_00276.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2018. Annotation of argument components in political debates data. In *Proceedings of the Workshop on Annotation in Digital Humanities*.

Charles Leonard Hamblin. 1987. *Imperatives*. Basil Blackwell.

Maralee Harrell. 2011. Argument diagramming and critical thinking in introductory philosophy. *Higher Education Research & Development* 30(3):371–385. https://doi.org/10.1080/07294360.2010.502559.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020a. BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, EMNLP Findings, pages 2281–2290. https://www.aclweb.org/anthology/2020.findings-emnlp.207.

Junxian He, Taylor Berg-Kirkpatrick, and Graham Neubig. 2020b. Learning Sparse Prototypes for Text Generation. In *Proceedings of the 34th Conference on Neural Information Processing Systems*. NIPS.

Stefan Heindorf, Yan Scholten, Hening Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. CauseNet: Towards a Causality Graph Extracted from the Web. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual Sequence Prediction and Adversarial Examples for Improved Fact-Checking. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics. Association for Computational Linguistics, Online, ACL, pages 8593–8606. https://doi.org/10.18653/v1/2020.acl-main.761.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*. Association for Computational Linguistics, Copenhagen, Denmark, pages 11–21.

Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *ICWSM 2017*.

David Hitchcock. 2007. Informal Logic and the Concept of Argument. In Dale Jacquette, editor, *Handbook of the Philosophy of Science*, North-Holland, Amsterdam, pages 101–129.

Hans Hoeken and Lettica Hustinx. 2009-10. When is Statistical Evidence Superior to Anecdotal Evidence in Supporting Probability Claims? The Role of Argument Type. *Human Communication Research* 35(4):491–510. https://doi.org/10.1111/j.1468-2958.2009.01360.x.

Erin R Hoffman, David W McDonald, and Mark Zachry. 2017. Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW):52.

Thomas A. Hollihan and Kevin T. Baaske. 2015. *Arguments and Arguing: The Products and Process of Human Decision Making, Third Edition*. Waveland Press.

Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer. 2020. Simple and Effective Retrieve-Edit-Rerank Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, ACL, pages 2532–2538. https://doi.org/10.18653/v1/2020.acl-main.228.

Yufang Hou and Charles Jochim. 2017. Argument Relation Classification Using a Joint Inference Model. In *Proceedings of the 4th Workshop on Argument Mining*. pages 60–66. https://doi.org/10.18653/v1/w17-5107.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, Long Beach, California, USA, volume 97 of *ICML*, pages 2790–2799. http://proceedings.mlr.press/v97/houlsby19a.html.

Aemilian Hron and Helmut F Friedrich. 2003. A review of web-based collaborative learning: factors beyond technology. *Journal of Computer Assisted Learning* 19(1):70–79.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument Generation with Retrieval, Planning, and Realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, ACL, pages 2661–2672. https://doi.org/10.18653/v1/p19-1255.

Xinyu Hua and Lu Wang. 2018. Neural Argument Generation Augmented with Externally Retrieved Evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 219–230.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* .

Jina Huh. 2015. Clinical questions in online health communities: the case of see your doctor threads. In *CSCW 2015*.

Ken Hyland. 2005. *Metadiscourse : Exploring Interaction in Writing.*. Continuum Discourse Series. Continuum.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *10th International Conference on Language Resources and Evaluation, LREC 2016*. European Language Resources Association (ELRA), pages 1638–1643.

Marilyn Jackson-Beeck and Robert G Meadow. 1979. The triple agenda of presidential debates. *Public Opinion Quarterly* 43(2):173–180.

Mathilde Janier and Chris Reed. 2017. I didn't say that! Uses of SAY in mediation discourse:. *Discourse Studies* 19(6):619–647.

Mathilde Janier and Patrick Saint-Dizier. 2019. *Argument Mining*. Linguistic Foundations. Wiley, 1 edition.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Evaluating bert for natural language inference: A case study on the commitmentbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 6088–6093.

Yohan Jo, Seojin Bang, Eduard Hovy, and Chris Reed. 2020. Detecting Attackable Sentences in Arguments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, EMNLP, pages 1–23. https://www.aclweb.org/anthology/2020.emnlp-main.1.

Yohan Jo and Alice H Oh. 2011. Aspect and Sentiment Unification Model for Online Review Analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*. pages 815–824.

Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Penstein Rosé, and Graham Neubig. 2018. Attentive Interaction Model: Modeling Changes in View in Argumentation. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1–10.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 2019. A Cascade Model for Proposition Extraction in Argumentation. In *Proceedings of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, Florence, Italy, pages 11–24.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic Sarcasm Detection. *ACM Computing Surveys (CSUR)* 50(5):1–22. https://doi.org/10.1145/3124420.

Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*. AAAI Press, pages 1807–1813.

Daniel Jurafsky, Rebecca Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, Paul Taylor, and C. Van Ess-Dykema. 1998. Automatic detection of discourse structure for speech recognition and understanding. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* pages 88–95. https://doi.org/10.1109/ASRU.1997.658992.

Anna Kaatz, Belinda Gutierrez, and Molly Carnes. 2014. Threats to objectivity in peer review: the case of gender. *Trends in pharmacological sciences* 35(8):371–373.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 3994–4004.

Simon Keizer, Markus Guhe, Heriberto Cuayahuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, and Oliver Lemon. 2017. Evaluating Persuasion Strategies and Deep Reinforcement Learning methods for Negotiation Dialogue agents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 480–484.

John Kelly, Danyel Fisher, and Marc Smith. 2005. Debate, division, and diversity: Political discourse networks in USENET newsgroups. In *Online Deliberation Conference 2005*.

Lina Khatib, William Dutton, and Michael Thelwall. 2012. Public diplomacy 2.0: A case study of the US digital outreach team. *The Middle East Journal* 66(3):453–472.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and Linking Web Forum Posts. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Uppsala, Sweden, pages 192–202.

Aniket Kittur, Bryan Pendleton, and Robert E Kraut. 2009. Herding the cats: the influence of groups in coordinating peer production. In *WikiSym 2009*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Santa Fe, New Mexico, pages 5–9.

Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pages 50–60. https://doi.org/10.18653/v1/2020.emnlp-main.4.

Nikolas Kompridis. 2000. So We Need Something Else for Reason to Mean. *International Journal of Philosophical Studies* 8(3):271–295. https://doi.org/10.1080/096725500750039282.

Erik C W Krabbe and Jan Albert van Laar. 2011. The Ways of Criticism. *Argumentation* 25(2):199–227.

Robert E. Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. 2011. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31(2):317–326.

Peter Lasersohn. 2009. Relative truth, speaker commitment, and control of implicit arguments. *Synthese* 166(2):359–374.

Anne Lauscher, Olga Majewska, Leonardo F R Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common Sense or World Knowledge? Investigating Adapter-Based Knowledge Injection into Pretrained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. EMNLP.

John Lawrence, Mathilde Janier, and Chris Reed. 2015. Working with open argument corpora. In *Proceedings of the 1st European Conference on Argumentation (ECA)*.

John Lawrence and Chris Reed. 2016. Argument Mining Using Argumentation Scheme Structures. In *Proceedings of the Sixth International Conference on Computational Models of Argument*. pages 379–390.

John Lawrence and Chris Reed. 2017. Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models. In *Proceedings of the 4th Workshop on Argument Mining*. pages 39–48. https://doi.org/10.18653/v1/w17-5105.

John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Computational Linguistics* 0(0):1–54.

John Lawrence, Jacky Visser, and Chris Reed. 2019. An Online Annotation Assistant for Argument Schemes. In *The 13th Linguistic Annotation Workshop*. pages 1–8.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*. AMW, pages 121–130.

Donghyeon Lee, Minwoo Jeong, Kyungduk Kim, Seonghan Ryu, and Gary Geunbae Lee. 2013. Unsupervised Spoken Language Understanding for a Multi-Domain Dialog System. *IEEE Transactions on Audio, Speech, and Language Processing* 21(11):2451–2464.

Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by design: ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. pages 30–40.

Stephen C Levinson. 1983. Conversational structure. In *Pragmatics*, Cambridge University Press, chapter 6, pages 284–333.

Qifei Li and Wangchunshu Zhou. 2020. Connecting the Dots Between Fact Verification and Fake News Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), pages 1820–1825. https://doi.org/10.18653/v1/2020.coling-main.165.

Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. Towards Assessing Argumentation Annotation - A First Step. In *Proceedings of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, Florence, Italy, pages 177–186.

Sally Lindsay, Simon Smith, Paul Bellaby, and Rose Baker. 2009. The health impact of an online heart disease support group: a comparison of moderated versus unmoderated support. *Health education research* 24(4):646–654.

Marco Lippi and Paolo Torroni. 2016. Argumentation Mining: State of the Art and Emerging Trends. *ACM Trans. Internet Technol.* 16(2):10:1–10:25.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020a. K-BERT: Enabling Language Representation with Knowledge Graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 34 of *AAAI*, pages 2901–2908. https://doi.org/10.1609/aaai.v34i03.5681.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. https://arxiv.org/abs/1907.11692.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020b. Fine-grained Fact Verification with Kernel Graph Attention Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 7342–7351. https://www.aclweb.org/anthology/2020.acl-main.655.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the Undecided: Language vs. the Listener. In *Proceedings of the 6th Workshop on Argument Mining*. pages 167–176.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pages 28–39.

Luca Lugini and Diane Litman. 2018. Argument component classification for classroom discussions. *Proceedings of Empirical Methods in Natural Language Processing* page 57.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 742–753.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 2561–2571. https://doi.org/10.18653/v1/p19-1244.

Diane Maloney-Krichmar and Jenny Preece. 2005. A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12(2):201–232.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Emaad Manzoor, George H. Chen, Dokyun Lee, and Michael D. Smith. 2020. Influence via ethos: On the persuasive power of reputation in deliberation online. *arXiv preprint arXiv:2006.00707* .

J R Martin and David Rose. 2003. *Negotiation: interacting in dialogue*. New Century Series. Bloomsbury Academic. https://books.google.com/books?id=Hoby40oSnUIC.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. 2013. Recognizing rare social phenomena in conversation: Empowerment detection in support group chatrooms. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pages 104–113.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open Domain Targeted Sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1643–1654.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and Predicting Online Persuasion Strategy with Elementary Units. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 6273–6278.

Elena Musi. 2017. How did you change my view? A corpus-based study of concessions' argumentative role. *Discourse Studies* 20(2):270–288.

Nona Naderi and Graeme Hirst. 2015. Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems*, Springer, pages 16–25.

Radford M Neal. 2003. Slice sampling. *The Annals of Statistics* 31(3):705–767.

Huy V Nguyen and Diane J Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *AAAI Conference on Artificial Intelligence*.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Distinguishing Antonyms and Synonyms in a Pattern-based Neural Network. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 76–85.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the Association for Computational Linguistics*. pages 985–995.

Yixin Nie, Lisa Bauer, and Mohit Bansal. 2020a. Simple Compounded-Label Training for Fact Extraction and Verification. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*. pages 1–7. https://doi.org/10.18653/v1/2020.fever-1.1.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 33, pages 6859–6866. https://doi.org/10.1609/aaai.v33i01.33016859.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020b. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, pages 4885–4901. https://doi.org/10.18653/v1/2020.acl-main.441.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 4658–4664.

Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

E Michael Nussbaum. 2011. Argumentation, Dialogue Theory, and Probability Modeling: Alternative Frameworks for Argumentation Research in Education. *Educational Psychologist* 46(2):84–106.

E Michael Nussbaum, Ian J Dove, Nathan Slife, CarolAnne M Kardash, Refika Turgut, and David Vallett. 2018. Using critical questions to evaluate written and oral arguments in an undergraduate general education seminar: a quasi-experimental study. *Reading and Writing* 19(2):1–22.

Juri Opitz and Anette Frank. 2019. Dissecting Content and Context in Argumentative Relation Analysis. In *Proceedings of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, Florence, Italy, pages 25–34.

Matan Orbach, Yonatan Bilu, Assaf Toledo, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2020. Out of the Echo Chamber: Detecting Countering Debate Speeches. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, pages 7073–7086. https://doi.org/10.18653/v1/2020.acl-main.633.

Silvia Pareti. 2016. PARC 3.0: A corpus of attribution relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, pages 3914–3920. https://www.aclweb.org/anthology/L16-1619.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 989–999. https://www.aclweb.org/anthology/D13-1101.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward Machine-assisted Participation in eRulemaking: An Argumentation Model of Evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*. ACM, New York, NY, USA, pages 206–210.

Joonsuk Park and Claire Cardie. 2018. A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Languages Resources Association (ELRA), Miyazaki, Japan.

Michael J. Paul. 2012. Mixed membership markov models for unsupervised conversation modeling. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pages 94–104. http://www.aclweb.org/anthology/D12-1009.

Andreas Peldszus and Manfred Stede. 2015. Towards Detecting Counter-considerations in Text. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 104–109.

Isaac Persing and Vincent Ng. 2016a. End-to-End Argumentation Mining in Student Essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 1384–1394.

Isaac Persing and Vincent Ng. 2016b. End-to-End Argumentation Mining in Student Essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1384–1394. https://doi.org/10.18653/v1/n16-1164.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights* 5:BII–S9042.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). EMNLP-IJCNLP, pages 43–54. https://doi.org/10.18653/v1/d19-1005.

Richard E Petty and John T Cacioppo. 1986. *Communication and Persuasion*. Central and Peripheral Routes to Attitude Change. Springer New York, New York, NY.

Bryan Pfaffenberger. 2003. A Standing Wave in the Web of Our Communications: Usenet and the Socio-Technical Construction of Cyberspace Values. In *From Usenet to Cowebs*, Springer, pages 20–43.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.

Niels Pinkwart, Kevin Ashley, Collin Lynch, and Vincent Aleven. 2009. Evaluating an Intelligent Tutoring System for Making Legal Arguments with Hypotheticals. *Int. J. Artif. Intell. Ed.* 19(4):401–424.

Martyn Plummer. 2003. Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Peter Potash and Anna Rumshisky. 2017. Towards Debate Automation: a Recurrent Model for Predicting Debate Winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2455–2465.

Jonathan Potter. 1996. *Representing reality: Discourse, rhetoric and social construction*. Sage.

Vinodkumar Prabhakaran, Ajita John, and Dor e e D Seligmann. 2013. Who Had the Upper Hand? Ranking Participants of Interactions Based on Their Relative Power. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, pages 365–373.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the Association for Computational Linguistics*. pages 866–876.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco.

Reid Priedhorsky, Jilin Chen, Shyong Tony K Lam, Katherine Panciera, Loren Terveen, and John Riedl. 2007. Creating, destroying, and restoring value in wikipedia. *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP '07* page 259. https://doi.org/10.1145/1316624.1316663.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2020. Causal Effects of Linguistic Properties. *arXiv* .

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* .

Chris Reed. 2021. Argument technology for debating with humans. *Nature* 591(7850):373–374. https://doi.org/10.1038/d41586-021-00539-5.

Chris Reed and Katarzyna Budzynska. 2011. How dialogues create arguments. In F. H. van Eemeren, B. Garssen, D. Godden, and G. Mitchell, editors, *Proceedings of the 7th Conference of the International Society for the Study of Argumentation (ISSA)*. SicSat.

Chris Reed, Katarzyna Budzynska, and Jacky Visser. 2016. IAT annotation guidelines for US2016. `http://arg.tech/US2016-guidelines`.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language Resources for Studying Argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/pdf/648_paper.pdf.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 567–578.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates. In *Proceedings of the 5th Workshop on Argument Mining*. Association for Computational Linguistics, Brussels, Belgium, pages 79–89. https://doi.org/10.18653/v1/w18-5210.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A Computational Approach for Generating Toulmin Model Argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*. pages 45–55. https://doi.org/10.3115/v1/w15-0507.

Michael Ringgaard, Rahul Gupta, and Fernando C. N. Pereira. 2017. SLING: A framework for frame semantic parsing. *CoRR* abs/1710.07032. http://arxiv.org/abs/1710.07032.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 440–450. https://doi.org/10.18653/v1/d15-1050.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, pages 172–180.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based

External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, pages 410–420.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 502–518.

Sara Rosenthal and Kathy McKeown. 2015. I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 168–177.

Peter Roudik, Graciela Rodriguez-Ferrand, Edouardo Soares, Tariq Ahmad, Laney Zhang, George Sadek, Nicolas Boring, Jenny Gesley, Ruth Levush, Sayuri Umeda, Hanibal Goitom, Kelly Buchanan, Norma C Gutierrez, Astghik Grigoryan, Elin Hofverberg, and Clare Feikert-Ahalt. 2019. Initiatives to Counter Fake News in Selected Countries. Technical report, The Law Library of Congress, Global Legal Research Directorate. https://www.loc.gov/law/help/fake-news/counter-fake-news.pdf.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):3027–3035. https://ojs.aaai.org/index.php/AAAI/article/view/4160.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end Argument Generation System in Debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. pages 109–114. https://doi.org/10.3115/v1/p15-4019.

Jacques Savoy. 2018. Trump's and clinton's style and rhetoric during the 2016 presidential election. *Journal of Quantitative Linguistics* 25(2):168–189.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model Architectures for Quotation Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 1736–1745. https://doi.org/10.18653/v1/p16-1164.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M Mclaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning* 5(1):43–102. https://doi.org/10.1007/s11412-009-9080-x.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia, pages 2371–2378. https://www.aclweb.org/anthology/L16-1376.

Sebastian Schuster, Joakim Nivre, and Christopher D Manning. 2018. Sentences with Gapping: Parsing and Reconstructing Elided Predicates. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 1156–1168.

Johanna Magdalena Schwager. 2005. *Interpreting Imperatives*. Ph.D. thesis, Johann-Wolfgang-Goethe Universitat.

Baruch B Schwarz and Michael J Baker. 2016. *Dialogue, Argumentation and Education*. History, Theory and Practice. Cambridge University Press.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Claudia Shi, David M Blei, and Victor Veitch. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*.

Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, and Tommy Sandbank. 2019. Detecting Persuasive Arguments Based on Author-Reader Personality Traits and Their Interaction. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP, pages 211–215. https://doi.org/10.1145/3320435.3320467.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 599–605.

Jack Sidnell. 2011. *Conversation Analysis: An Introduction*. Language in Society. Wiley. https://books.google.com/books?id=uS-eHxYck3EC.

Noam Slonim. 2018. Project Debater. In *Computational Models of Argument*. volume 305, pages 4–4. https://doi.org/10.3233/978-1-61499-906-5-4.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature* 591(7850):379–384. https://doi.org/10.1038/s41586-021-03215-w.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the*

conference on empirical methods in natural language processing. Association for Computational Linguistics, pages 254–263.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for Evidence Retrieval and Claim Verification. In *Proceedings of the 42nd European Conference on IR Research*. pages 359–366. https://doi.org/10.1007/978-3-030-45442-5_45.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, pages 116–124.

Yi Song, Paul Deane, and Beata Beigman Klebanov. 2017. Toward the Automated Scoring of Written Arguments: Developing an Innovative Approach for Annotation. *ETS Research Report* ... 152(3):157.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 1(1):1–62. https://doi.org/10.1162/coli_a_00295.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. pages 3664–3674.

Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. Parallel Discourse Annotations on a Corpus of Short Texts. In Nicoletta Calzolari Conference chair, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*. pages 217–226.

Swabha Swayamdipta and Owen Rambow. 2012. The pursuit of power and its manifestation in written dialog. *Proceedings - IEEE 6th International Conference on Semantic Computing, ICSC 2012* pages 22–29. https://doi.org/10.1109/ICSC.2012.49.

Robert Talisse and Scott F Aikin. 2006. Two Forms of the Straw Man. *Argumentation* 20(3):345–352.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 175–185.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 613–624.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., volume 32. https://proceedings.neurips.cc/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf.

Niket Tandon, Bhavana Dalvi Mishra, Keisuke Sakaguchi, Antoine Bosselut, and Peter Clark. 2019. WIQA: A dataset for "What if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. pages 6076–6085. https://doi.org/10.18653/v1/d19-1629.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1):24–54. https://doi.org/10.1177/0261927X09351676.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 809–819. https://doi.org/10.18653/v1/N18-1074.

Santosh Tokala, Vishal G, Avirup Saha, and Niloy Ganguly. 2019. AttentiveChecker: A Bi-Directional Attention Flow Mechanism for Fact Verification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 2218–2222. https://doi.org/10.18653/v1/n19-1230.

Jason Turcotte. 2015. The news norms and values of presidential debate agendas: An analysis of format and moderator influence on question content. *Mass Communication and Society* 18(3):239–258.

URL1. . `http://www-di.inf.puc-rio.br/~endler/students/Hedging_Handout.pdf`. http://www-di.inf.puc-rio.br/ endler/students/Hedging_Handout.pdf.

URL2. . `https://github.com/words/hedges/blob/master/data.txt`. https://github.com/words/hedges/blob/master/data.txt.

URL3. . https://www.janefriedman.com/hedge-word-inflation-words-prune/. https://www.janefriedman.com/hedge-word-inflation-words-prune/.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards Debiasing NLU Models from Unknown Biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, EMNLP, pages 7597–7610. https://doi.org/10.18653/v1/2020.emnlp-main.613.

Frans H van Eemeren and Rob Grootendorst. 1984. *Speech Acts in Argumentative Discussions*. A Theoretical Model for the Analysis of Discussions Directed towards Solving Conflicts of Opinion. Walter de Gruyter, Berlin, New York.

Frans H van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge University Press.

Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. 2007. Talk before you type: coordination in Wikipedia. *40th Hawaii International Conference on System Sciences* 1:1–10. https://doi.org/10.1017/CBO9781107415324.004.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Proceedings of the 34th Conference on Neural Information Processing Systems*.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2019. Argumentation in the 2016 US Presidential Elections . *Language Resources and Evaluation* pages 1–35.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2020. Annotating Argument Schemes. *Argumentation* pages 1–39. https://doi.org/10.1007/s10503-020-09519-x.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57(10):78–85. https://doi.org/10.1145/2629489.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 250–255.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 176–187.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018a. Argumentation Synthesis following Rhetorical Strategies. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pages 3753–3765.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018b. Retrieval of the Best Counterargument without Prior Topic Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 241–251.

Jean H M Wagemans. 2016. Constructing a Periodic Table of Arguments. In *OSSA Conference Archive*. pages 1–13.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*. Istanbul, pages 812–817.

Douglas Walton. 2008. *Informal Logic*. A Pragmatic Approach. Cambridge University Press.

Douglas Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, CHI '20, page 1–14. https://doi.org/10.1145/3313831.3376732.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. *Transactions of Association of Computational Linguistics* 5:219–232.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2019. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *arXiv* .

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4):1191–1207.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2006. The Penn Discourse Treebank 3.0 Annotation Manual.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is This Post Persuasive? Ranking Argumentative Comments in Online Forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 195–200.

Miaomiao Wen, Keith Maki, Xu Wang, Steven P Dow, James Herbsleb, and Carolyn Rose. 2016. Transactivity as a predictor of future collaborative knowledge integration in team-based learning in online courses. *Proceedings of the 9th International Conference on Educational Data Mining* .

Daniel Wentzel, Torsten Tomczak, and Andreas Herrmann. 2010. The moderating effect of manipulative intent and cognitive resources on the evaluation of narrative ads. *Psychology & Marketing* 27(5):510–530.

Anthony G Wilhelm. 2000. *Democracy in the digital age: Challenges to political life in cyberspace*. Psychology Press.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pages 1112–1122. https://doi.org/10.18653/v1/n18-1101.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. ANLIzing the Adversarial Natural Language Inference Dataset. *arXiv* .

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Vancouver, British Columbia, Canada, pages 347–354.

Terry Winograd and Fernando Flores. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Language and being. Ablex Publishing Corporation. https://books.google.com/books?id=2sRC8vcDYNEC.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Bronwyn Woods, David Adamson, Shayne Miel, and Elijah Mayfield. 2017. Formative Essay Feedback Using Predictive Scoring Models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, New York, USA, pages 2071–2080.

Scott Wright. 2006. Government-run online discussion fora: Moderation, censorship and the shadow of control. *The British Journal of Politics and International Relations* 8(4):550–568.

Scott Wright and John Street. 2007. Democracy, deliberation and design: the case of online discussion forums. *New media & society* 9(5):849–869.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-Target Stance Classification with Self-Attention Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, pages 778–783. https://doi.org/10.18653/v1/p18-2123.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2019. Recognising Agreement and Disagreement between Stances with Reason Comparing Networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 4665–4671. https://doi.org/10.18653/v1/p19-1460.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia. *Proc. ICWSM* pages 446–455.

Diyi Yang, Miaomiao Wen, and Carolyn Rosé. 2015. Weakly Supervised Role Identification in Teamwork Interactions. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* pages 1671–1680. http://www.aclweb.org/anthology/P15-1161.

Xiaoyu Yang, Feng Nie, Yufei Feng, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020. Program Enhanced Fact Verification with Verbalization and Graph Attention Network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 7810–7825. https://doi.org/10.18653/v1/2020.emnlp-main.628.

Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. 2014. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 91–96.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 136–141.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the Causal Effects of Conversational Tendencies. *Proceedings of ACM on Human Computer Interaction* 4(CSCW2).

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. Asking too much? The rhetorical role of questions in political discourse. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1558–1572.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, pages 1441–1451. https://doi.org/10.18653/v1/p19-1139.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Compu-

tational Linguistics, Online, pages 6170–6180. https://www.aclweb.org/anthology/2020.acl-main.549.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 892–901. https://doi.org/10.18653/v1/p19-1085.