Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios

Vamshi Ambati

CMU-SCS-11-020
January 11, 2012

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

Thesis Committee:
Jaime Carbonell (Chair)
Stephan Vogel (co-chair)
Raj Reddy
Aniket Kittur
Kevin Knight, University of Southern California, ISI

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

This research was partially supported by DARPA under grant NC 10-1326.

Keywords: Machine Translation, Active Learning, Crowdsourcing, Applied Machine Learning, Low-resource Languages, Natural Language Processing, Computer Science

Abstract

Corpus based approaches to automatic translation such as Example Based and Statistical Machine Translation systems use large amounts of parallel data created by humans to train mathematical models for automatic language translation. Large scale parallel data generation for new language pairs requires intensive human effort and availability of fluent bilinguals or expert translators. Therefore it becomes immensely difficult and expensive to provide state-of-the-art Machine Translation (MT) systems for rare languages.

In this thesis, we explore active learning to reduce costs and make best use of human resources for building low-resource MT systems. Active learning approaches help us identify sentences, which if translated have the potential to provide maximal improvement to an existing system. We then apply active learning to other relevant tasks in MT such as word alignment, classifying monolingual text by topic, extracting comparable corpora from the web. In all these tasks we reduce annotated data required by the underlying supervised learning models. We also extend the traditional active learning approach of optimizing selection for a single annotation to handle cases of multiple-type annotations and show further reduction of costs in building low-resource MT systems.

Finally, as part of this thesis, we have implemented a new framework - Active Crowd Translation (ACT), a cost sensitive active learning setup for building MT systems for low-resource language pairs. Our framework will provide a suitable platform for involving disparately spread out human translators around the world, in a timely and sparingly fashion for rapid building of translation systems. We first explore the ACT paradigm with expert translators and then generalize to full-scale crowdsourcing with non-expert bilingual speakers. In case of Machine Translation, although crowdsourcing services like Amazon's Mechanical Turk have opened doors to tap human potential, they do not guarantee translation expertise nor extended availability of translators. We address several challenges in eliciting quality translations from an unvetted crowd of bilingual speakers.

Acknowledgments

Carnegie Mellon is always brimming with potential, and sometimes the best solutions and collaborations are made at an informal social or in a hallway conversation. I am grateful to everyone that I have had a chance to interact with during my graduate years.

Firstly, I would like to thank my advisors Jaime Carbonell and Stephan Vogel and my fabulous committee. Jaime has always been an unparalleled judge of what was not possible, while still pushing the limits on what was possible. He has taught me how to focus on the big picture and systematically approach the pieces of the puzzle. I will always be thankful for all the freedom and support he has given me to pursue what I was interested in, and being always available to offer advise and mentorship when I was lost. No problem seems too big with an advisor like Jaime. Thanks to Stephan for all the long discussions that helped me build the analytical skills and critical thinking required to face the reality of research. Always cheerful and enthusiastic, I am yet to find a more approachable advisor than Stephan. I am going to dearly miss my Thursday evening advisor meetings, my knowledge enrichment supplements.

My heartfelt thanks to Prof. Raj Reddy for introducing me to CMU and persisting with me while I picked the necessary skills to embark upon the PhD path. Dr. Reddy is a true visionary and working closely with him has been a dream come true for me. Every trip to his office filled me with new ideas and renewed my vigor for research. He leads by example and his life will always be my source of inspiration. Thanks to Niki Kittur for discussions on crowdsourcing and also getting me interested in the area of collaborative computing. I am fortunate to have Kevin Knight, a person I look up to, as my external advisor. I would consider my life a success if I can come close to imitating any one of the members of my committee.

Thanks to Alon Lavie for getting me started at CMU on the exciting research of Machine Translation. Alon's passion for the field is contagious. I am indebted to Prof. Rajeev Sangal, my undergraduate thesis advisor, for introducing me to the fields of NLP and AI. Thanks to friends and colleagues at CMU - Kenneth Heafield, Greg Hanneman, Jonathan H. Clark, Abhaya Agarwal, Alok Parlikar, Erik Petersen, Michael Denkowski, Kevin Gimpel, Qin Gao, Sanjika Hewavitharana and other past and present members of the Avenue, InteractLab

and other research groups at LTI. Your advice, criticisms and constructive feedback were critical in shaping me as a researcher. Thanks also to all the people with lit offices working on weekends. You guys were my inspiration, and someday you will see the light at the end of the tunnel as well.

I have worked with some great collaborators and made good friends all along my graduate years, who have in some way or the other impacted my life as a student and otherwise. Thanks to Kishore Prahallad, Rohini Uppuluri, Hemant Gogineni, Madhavi Ganapathiraju, Sriram Venkatapathy, Gopala Anumanchipalli, Pradeep Varakantham, Rohit Kumar, Ponnurangam Kumaraguru, Vasudeva Varma and other visitors and students that I have met by the fortune of being part of Dr. Reddy's group and at LTI. Special thanks to Vivian Lee, for the sanity and the Bethel Bakery cakes.

My deepest gratitude to my parents back home, for supporting me and believing in my goals and abilities. Dad's commitment and dedication to educating his kids has shaped my life and character. I am blessed to have two brothers, both of whom treat me as their elder kid rather than as a sibling. At times, when I had my own share of lows and tough times, I needed to go no farther than calling my loving mom, to whom I will always remain a hero and a bit of a mystery.

Last but not least, this thesis would not have been possible without the love and strength of my dearest friend and now my wife, Pooja. This thesis journey is as much hers as mine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We live in a world driven by information, and a generation that is constantly making a leap towards globalization. While information technology has made abundant progress to bridge the social, political and economic gap across culturally diversified population, a true sense of communication and information exchange is still impeded by the language barrier. The current world can be seen as a batch of communicating islands, at best. We want people to maintain their cultures, preserve their languages and so for the foreseeable future data will be produced in local languages. We want intelligent systems that interpret this content and deliver information where needed.

Machine translation has remained an interesting problem and a challenging sub-field of natural language processing, aiming to bridge this exact language barrier, by building machines that can automatically translate between human languages. The first known application of Machine Translation (MT) was its usage for military purposes in decoding Russian sentences into English after the second World War. Although, its usage is still predominant in government domains trying to understand intelligence across borders, it is now increasingly becoming a technology that is used in translating legal documents, patents, studying history and cultural heritages of foreign countries etc. With the dropping hardware and computational costs, the translation technology that was once only in the hands of a privileged few with access to supercomputing resources in the government is now actually available to the common man, running on a commodity personal laptops and in some cases, on mobile and hand-held devices as well, e.g in the hands of a tourist, translating foreign language on sign-boards in real-time.

The World Wide Web (WWW) which was dominated by the English language content in its nascent years, is now closely tied with Chinese, followed by languages like Spanish, Japanese, Portuguese etc [1]. With substantially decreasing costs of bandwidth in developing

[1] http://www.internetworldstats.com/

countries, the Internet population which was also mostly from the North Americas and Europe is now dominated by countries like China and India. Seamless communication, dissemination of knowledge and uninterrupted equal access to information to this diverse population puts a heavy emphasis on the need for Machine Translation.

In the past decade we have seen great progress in Machine Translation (MT) with a drift from traditional rule based approach to Statistical Machine Translation (SMT). SMT has now become the dominant paradigm in MT as evidenced in government evaluations like NIST[2] and industrial online MT services, such as Google Translation service [3]. Both the dominance of the statistical approach in MT and the progress made in recent years are clearly demonstrated in MT evaluations as organized by NIST, Workshop for Machine Translation (WMT[4]), International Workshop for Spoken Language Translation (IWSLT[5]) etc. These initiatives have not only nurtured the overall interest in MT research but have also led to the development of successful statistical MT algorithms - as can be affirmed by the increasing number of SMT related papers at major conferences and in leading journals.

The state-of-the-art approaches to Machine Translation (MT) are data-driven requiring voluminous translation corpora. Given a large set of sentences in a source language and the equivalent translations in a target language, referred to as parallel corpora, the underlying algorithms learn word level dictionaries and other phrase patterns to support the translation of an unseen sentence. Therefore, progress in MT has been observed only in a small number of language-pairs where substantial amounts of parallel data exist - such as Spanish-English, French-English, Arabic-English, Chinese-English, etc. While there are clearly some improvements achieved by better modeling the translation process - from word-based to phrase-based to hierarchical to syntax-based systems - continued improvements in MT are to a large extent due to just throwing more and more data at the problem. Given that linear improvement in translation quality metrics requires exponential growth in training data, one can only extrapolate how large a parallel corpus is needed (e.g., for Thai, Swahili or Finish) to achieve the same quality as we already see in Arabic or Spanish to English translations.

Despite efforts to compile or even build from scratch relevant resources for such low-resource or less commonly taught languages, the situation is still very unsatisfactory [Simpson et al., 2008]. Only a few languages in the world enjoy sustained research interest and continuous financial support for the development of automatic translation systems. For most remaining languages there is very little interest or funding available and/or a limited or expensive access to experts for data elicitation. Building translation systems for those low resource languages therefore requires better ways to actively build and make do with smaller but perhaps more useful parallel text collections.

[2] http://www.itl.nist.gov/iad/mig/tests/mt/
[3] http://translate.google.com
[4] http://www.statmt.org/wmt10/
[5] http://mastarpj.nict.go.jp/IWSLT2009/

In this thesis, we resort to Active Learning (AL) techniques for building MT systems for minority languages. Active learning is a suite of techniques whose objective is to rank a set of instances in an optimal manner for an external oracle to label them so as to provide maximal benefit to the learner. Active learning is indispensable in supervised machine learning in low-resource scenarios where further annotation of new instances constrained by limited budget. In a complex system like MT, different models combine forces to produce the final translation and therefore, the annotations are multiple, structural and could be of different types. Depending upon the MT system and the paradigm, the resource requirements for translation may also encompass annotations such as morphological analysis, named-entity tagging, part-of-speech tagging, sense disambiguation, syntactic parses, semantic analysis, etc. We will focus on translation data acquisition, word-alignment and topic categorization, but our techniques should be useful and generalize to other types of annotations as well. Traditionally, active learning strategies have been applied in learning a classifier, where the annotation is of a single data-type, the class label. In this thesis, we also extend the traditional setup of active learning which elicits a single kind of annotation to improve the quality of a particular task to multiple annotations for different tasks.

Large scale parallel data generation is an onerous task requiring intensive human effort and availability of bilingual speakers. Consequently, much of the efforts made and progress seen has been confined to majority languages, where 'majority' can be interpreted as languages with large amount of data or funding or political interest. However, a lot of languages in the world do not enjoy this status, which I refer to in this thesis as 'low-resource' languages. However for most of the language-pairs, while there is a lack of access to language experts, there are a large number of speakers available for each of the language, some of whom may also be bilingual. Earlier it was difficult to have rapid access to such bilingual speakers for a language-pair, but with the upcoming platforms and techniques for micro-task markets, also called 'crowdsourcing' [Howe, 2006], it is now possible to tap into users for completing tasks.

In this thesis, we use crowdsourcing market places like Amazon's Mechanical Turk[6] for collecting translation data. A recent study in crowdsourcing shows that there exists a wide range of users on Mechanical Turk (MTurk), speaking multiple languages[7]. One of the advantages of crowdsourcing platforms is the associated reduction of cost due to the increased access to non-expert bilinguals. For instance, to translate 1000 sentences from Telugu to English on MTurk, the total cost involved is 45 USD at the rate of 3 cents per sentence and service charges for using the platform. A typical state-of-the-art SMT system for any language pair today requires 10M sentence pairs to reach reasonable translation quality. In order to create a similar sized corpus to build a Telugu to English MT system the overall expenditure would exceed 50M USD. This is assuming a 100M word corpus and 5 cents per word professional translation rate. We would expect that building an equal

[6]http://www.mturk.com/mturk/

[7]http://www.junglelightspeed.com/amt_language/

sized corpus using crowdsourcing would require far less expenditure (500K USD approx.). However, crowd sources are inherently unreliable requiring us obtain repeated labels. In this thesis we will investigate proactive methods for estimating reliability and for combining results via weighted voting, jointly optimizing sample-selections, annotation-type selections, source-selections and multi-source combination strategies.

Also, unlike language translation in majority languages, where we need more general translation systems, language translation systems for low-resource languages, when built, are driven by an immediate need of a small group of audience and are therefore, focused around a problem. For instance, projects that build MT systems for African languages in order to aid the rehabilitation of refugees in the United States or other foreign countries etc. Another example is where MT systems were required to bridge the communication gap during disaster situations, similar to the recent relief efforts at the earthquake in Haiti and Japan, where doctors and volunteers from all over the world came down to help the victims. Such projects are of importance for humanitarian causes are also very time critical and typically have pre-specified , limited budgets. One cannot afford the translation of a million sentences to train high accuracy systems, neither can we wait for the time taken for data entry. Hence the need of the day is to build algorithms that provide usable translation systems at a very low budget that takes less time for development. We will therefore, combine both active learning and crowdsourcing under a single framework and show that the overall costs of building low-resource translation systems can be drastically reduced.

1.1 Statistical Machine Translation

In this section, we will briefly introduce the phrase-based statistical machine translation (SMT) paradigm that we will use in the rest of the thesis. Also, active Learning is dependent upon the definition of the model representation and so in this section we will discuss the representation of the SMT model that we use to optimize in the rest of the chapter. We will only discuss models pertaining to finite state methods, as this thesis does not consider syntax-based SMT approaches that use the context-free and other grammar formalisms [Lopez, 2008].

Machine Translation (MT) can be seen as a function, $f : S \rightarrow T$ that transforms a sentence in a source language S to an equivalent translation into a target language T. Learning such a function is a pattern recognition problem that involves modeling, parametrization and parameter estimation. The task of translation can be seen as producing the most likely target sentence t for a source sentence s under a particular model $P_\theta(s/t)$, where θ can be any parameterization of the model.

$$s* = argmax_s P_\theta(s|t)$$

The parameterization of the SMT system is motivated and focused in the generalization power expected of the model without sacrificing quality. A good unit of generalization that encapsulates local context, important for cohesive translation, are $phrases$. Phrase-based approaches rely on the existence of techniques for alignment at word-level, which are typically performed by the IBM models [Brown et al., 1993]. A large corpus with reasonably accurate alignment is a good starting point for applying extraction techniques for producing a translation equivalence table of phrase mappings, forming our model parameters. Estimation of these parameters is a problem that has been at best addressed as relative likelihoods as provided by the data. The IBM models also provide other rich parameters in the form of lexical translation tables, and distortion parameters that are dependent upon the model assumptions. They are an important by-product of the translation equivalence computation. These, in conjunction with the phrase tables form the sub-models of translation which are often linearly combined as shown below, the weights of which are optimized for translation performance on a held-out dataset.

$$P(e|f) = \frac{1}{Z(\lambda)} \prod_{i=1}^{m} h_i(e,f)^{\lambda_i}$$

The space as defined by the model $P_\theta(s'/t)$ of a PB-SMT consists of a number of sub-models or feature functions $h_i(e,f)$ computed either from the source s, or target f or both (e,f). Some of the features observed in phrase based systems are as follows:

- $p_w(s_i|t_j)$: Lexical translation probability of the source word s_i being a translation equivalent , given a target word t_j
- $p_w(t_i|s_j)$: Lexical translation probability of the target given source.
- $phr(s_i|t_j)$: Phrase translation probability of the source phrase s_i being a translation equivalent , given a target word t_j
- $phr(t_i|s_j)$: Phrase translation probability of the target given source.
- $lm(t)$: Language model score for the target segment t.

1.2 Active Learning

Supervised learning has become a widely used and successful paradigm in natural language processing tasks. Most approaches rely on large-scale labeled data availability for better accuracy rates. We address the problem of selecting the most informative examples from unlabeled data in order to reduce the human effort via 'Active Learning' methods. In

active learning, the learner has access to a large pool of unlabeled data and sometimes a small portion of seed labeled data. The objective of the active learner is then to select the most informative instances from the unlabeled data and seek annotations from an external expert/oracle, which it then uses to retrain the underlying supervised model for improving performance. This continues in an iterative fashion for convergence, which typically is a threshold on the achievable performance before exhausting all of the unlabeled data set. Since we consider the most informative unlabeled instances for annotation at every iteration, we always expend effort towards improving the performance of the task.

Machine Translation (MT) requires labeled data in the form of translated sentences pairs where a translation can be seen as an annotation of a monolingual sentence. Similarly, other annotation tasks exist for MT such as word alignment, topic identification etc. Figure 1.1 shows a standard active learning setup for creating parallel corpus for an MT system, which can largely be generalized for other annotation tasks as well.

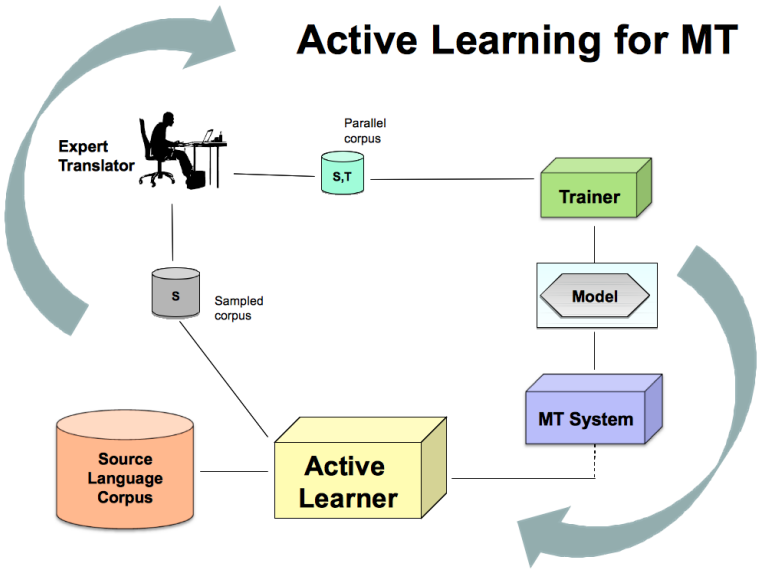

Figure 1.1: Active Learning setup for the Sentence Selection task. A set of sentences are selected by the learner from an unlabeled pool and translated by an expert translator to re-train the MT system

The main components of any active learning setup are therefore, the unlabeled data, a selection criteria and an oracle to annotate the selected data. We always assume availability of a a large pool of unbiased collection of unlabeled data to select from. This is typically called a pool-based active learning. Finally, an important aspect that we do not address in this thesis is the stopping criteria for active learning process. We assume that the process is

continued until a desired accuracy level is accomplished. For a more detailed description of active learning and its applications, we refer the reader to Settles [2009].

1.3 Dimensions of Active Learning

In this section, we try to understand some of the dimensions of active learning that make it an interesting and a challenging problem. We are not aware of work highlighting these dimensions in active learning. We study the following dimensions, both jointly and in isolation. Throughout the thesis, we instantiate these in the context of MT and discuss them in further detail.

1.3.1 Query Selection Frameworks

Query strategies weigh the importance of the unlabeled instances with respect to a given model and prioritize them for annotation. As we will study in the rest of this thesis, different query strategies have varied characteristics that can be exploited for a particular task and data-set. So the choice of the algorithm is crucial to the benefits reaped. The following are two of the query frameworks that we explore in this thesis.

Informativeness

Strategies in this primarily compute the representativeness of the instance among its pool. Selecting such instances provides better coverage over unseen future instances.

$$score(s) = \sum_{s=1}^{|U|} sim(s, s') * \frac{1}{|U|} \tag{1.1}$$

Uncertainty

Ability to score the instances with an uncertainty function, reveals the weaknesses of the model. Sampling instances to fill in for this deficiency is expected to improve the model. In most structural learning problems, the model takes the form of a probability distribution over the possible targets conditioned upon the input.

$$score(s) = \theta(t/s) \tag{1.2}$$

1.3.2 Annotation Variety

Active learning approaches optimize the selection of samples for a particular labeling task involving a single kind of annotation. The choice of annotation constraints the data we select from, the methods for scoring instances, and the expertise area of the annotator. For machine translation, there is a wide variety of possible tasks that can benefit the translation system, some of which include - parallel data creation, alignment of word links in a sentence pair, rich tagging of words, syntactic tree annotation, etc. The choice of annotation also influences the improvement in quality of the supervised algorithm.

1.3.3 Annotator

At the core of the learner is an external oracle that can provide answers/annotations to the selected dataset. One of the assumptions that has lately been questioned is the availability of such an expert and the non-fallible, indefatigable nature of the annotator. For most tasks, it is unlikely that such an oracle might exist, and if it does, it is so expensive that the cost of annotation becomes quickly prohibitive. Now, assume that the highly expensive expert is replaced with a very cheap non-expert who is correct in only 70% of the cases. Can the learning algorithms deal with such noise? A crowd of non-experts may sometimes still be cheaper and more accessible than an expert. Active learning algorithms need to be aware of the budget and accuracy of the annotator in order to decide when and who and how many annotators to consult for annotating an instance.

1.3.4 Annotation Granularity

Another dimension that is often ignored in acquiring labels for classification tasks is the granularity. Labels in classification are often binary and thus cannot further be decomposed. In structural learning though, the annotation granularity of the label is legitimate. For example, a partial sentence translation is still useful when compared to no translation. Seeking a phrase translation may be an easier task, that can be completed by more annotators, and has value in the training of an MT system. A cost-sensitive active learning algorithm can choose to fall back to fine-granular annotations to save on budget or utilize its human resources judiciously.

1.3.5 Operational Ranges

A dimension that has been ignored or not fully explored in most literature is that of an 'operating range'. Availability of initial labeled data (often termed as 'seed data') changes the choice of the other dimensions of an active learning algorithm. For instance, when

dealing with absolutely no labeled data, one might argue that a random selection of data may work well in quickly establishing a decision boundary. A similar argument can be applied to the choice of query sampling strategy when a lot of labeled data is available and only limited unlabeled data is streamed. We term these as operating ranges, the exact definition of which is dependent on the application. We refer to this as 'context-aware active learning'.

1.4 Thesis

Corpus based approaches to automatic translation like Example Based and Statistical Machine Translation are the state-of-the-art MT systems. However, these approaches require significantly large amounts of resources, primarily in the form of parallel data to train mathematical models. Therefore, it becomes difficult and expensive to built translation systems for low-resource languages. Active learning addresses the situation when there is a paucity of labeled data such as bilingual parallel text, but unlabeled data such as monolingual untranslated text is available in abundance, and obtaining labels (e.g. topics, translations, alignments) requires extensive and expensive human effort. For most language-pairs, professional translators are in fact expensive and too scarce and so building MT systems becomes a distant dream. We believe that the contributions from this thesis will provide better understanding into faster, cheaper and effective ways to building MT systems for such low-resource languages.

1.4.1 Statement

This thesis explores active learning for building statistical machine translation systems in low-resource scenarios and also extends the traditional framework of active learning to work with multiple annotation tasks and with non-expert annotators available through crowdsourcing.

1.4.2 Hypotheses

- Active learning will reduce human costs in producing resources for corpus-based translation systems, with significant savings in the case of low resource language pairs
- When working with multiple annotation tasks (e.g. topic classification and translation), joint sample selection through active learning can further reduce the cumulative cost of acquiring labeled data.

- Crowdsourcing translation tasks will significantly reduce the cost of annotation, and algorithms for careful selection of translations in the crowd data will produce data that is comparable to that obtained from expert bilinguals.

1.4.3 Research Summary

The complex nature of machine translation task, poses several challenges to the application of active learning, more so in the context of low-resource languages. We started this thesis with an exploration of active learning as a right fit for building low-resource machine translation systems. We identified a number of dimensions that are left unexplored in active learning, and argue that MT is an appropriate test-bed to experiment and validate some of these dimensions both independently and jointly. From the view of an MT system, we will also categorically study the different query strategies based upon the resource they use in selection. While some strategies only make use of the unlabeled data and parallel data for selection, making them MT paradigm agnostic, others use the underlying model representation or the hypothesized output, making them more closely tied to the model representation and thereby, limiting them to particular MT approaches, ex: SMT or EBMT etc.

We explored a variety of query strategy frameworks for selecting the optimal set of sentences to be translated by humans to aid training of an MT system. Some of the techniques we have looked at are data-driven and focus on the density and diversity statistics for selection. We then explore novel ways of combining these strategies as static and dynamic ensembles. These experiments are carried out for multiple language-pairs to test both feasibility and portability. We also applied active learning to a second task in MT - word alignment, and have shown improved performance at less human effort.

We then propose to explore the concept of 'operating ranges' for active learning. Similar to research in Donmez et al. [2007], we also observe that in most related work, the application of active learning strategies is done in an unified fashion all along the learning curve. However, as the data acquired in the active learning process increases, the MT system moves from being a low-resource to a medium-resource and then eventually becomes a high-resource language. Query strategies need to be introspective of the resources available to the system. We will observe for the existence of these operating ranges in a learning curve and study the extension of introspective query strategies for optimal data selection under varied data scenarios.

A second direction that this thesis explores is that of extending active learning to handle multiple annotations. Active learning traditionally selects instances for a single kind of annotation task. In machine translation several different kinds of data and annotations can be provided for the improvement of the underlying model, each associated with respective cost models. We will address this in the context of building multi-domain translation

systems, where domain information and translation are two different annotations and we explore the benefits from combining them under a single learning algorithm. The learning strategies need to be aware of the annotation variety and optimally select the annotation and the instance together to trade-off between cost and benefit.

The third direction that we explore in greater depth is the assumption of expert oracle availability for providing annotations in an active learning setup. We show that this is not only expensive but also less feasible for low-resource languages where finding speakers of such languages is a difficult proposition. We propose crowdsourcing as an alternative data collection for building MT systems. The challenges with using this paradigm is the effect of quality of the data, inspired by the techniques in proactive learning we will both implement existing and propose new algorithms for dealing with data from non-experts. We also study empirically the feasibility of using such data in training real world MT systems.

Finally, we unify both active learning and crowdsourcing and study them under a cost-sensitive framework which we call 'Active Crowd Translation' (ACT), geared towards building low-resource language translation systems.

1.4.4 Contributions

The major contribution of this thesis is the development and application of a new framework for building machine translation systems for low-resource languages - Active Crowd Translation (ACT). Along the path, we also make the following contributions:

- Improvement of active learning algorithms for the task of parallel data creation with significant improvement on low-resource languages.
- Designing active learning setup for the task of word-alignment and implementing strategies that reduce alignment error rates with significant cost reduction.
- Application of active learning to building a comparable corpora classifier and extending the traditional single-annotation driven active learning to select instances for eliciting multiple types of annotations.
- Extension of active learning setup to also jointly select an annotation type and an instance in the context of building domain-specific translation systems for low-resource languages, where topic classification and translation are two inherent tasks.
- Designing several techniques for quality-effective and cost-effective application of crowdsourcing to the general problem of language translation.
- Implementation of a novel framework called Active Crowd Translation (ACT), that combines active learning and crowdsourcing for building MT systems for different language pairs in low-resource scenarios.

1.4.5 Organization

- Chapter 2 provides a literature survey of the three main broad areas supporting this thesis - active learning, crowdsourcing and machine translation
- Chapter 3 discusses application of active learning applied to the task of building parallel corpora for Machine Translation. We discuss several query selection strategies for the sentence selection task
- Chapter 4 discusses another important task of the translation pipeline - word alignment and applies active learning techniques for improving a semi-supervised word alignment setup
- Chapter 5 provides an extension to the traditional single-annotation focused active learning to a multi annotation setup. We discuss the application of multiple annotation active learning for two tasks in translation - improving a comparable corpora classifier and also building domain specific translation systems
- Chapter 6 introduces crowdsourcing as a viable option for eliciting parallel data for building MT systems. We analyze several issues in this new area of research and provide algorithms for making use of the selective and/or collective wisdom of the crowds
- Chapter 7 discusses the ACT framework to combine active learning and crowdsourcing for building MT systems for a couple of language-pairs
- Chapter 8 presents some concluding words and contributions from this thesis
- Chapter 9 discuses avenues for future work

Chapter 2

Literature Survey

In this chapter we will provide some background and survey literature pertaining to the three major fields that are relevant to this thesis - Statistical Machine Translation, Active Learning and Crowdsourcing.

2.1 Statistical Machine Translation

In recent years, corpus based approaches to machine translation have become predominant, with Phrase Based Statistical Machine Translation (PB-SMT)[Koehn et al., 2003] being the most actively progressing area. While PB-SMT improves traditional word based machine translation approaches by incorporating more contextual information in the form of phrase pairs, it still has limitations in global block level reordering of phrasal units. Such reorderings can be captured by knowledge about the structure of the language. Recent research in syntax based machine translation [Yamada and Knight, 2001] [Marcu et al., 2006] [Chiang, 2005] incorporates syntactic information to ameliorate the reordering problem of phrasal units. Some of the approaches operate within the resources of PB-SMT and induce hierarchical grammars from existing non-syntactic phrasal units, to provide better generality and structure for reordering [Chiang, 2005] [Wu, 1997]. Other approaches use syntactic analysis of sentences on one side of the corpus to induce grammar rules [Galley et al., 2004] [Yamada and Knight, 2001] [Venugopal et al., 2007].

Most approaches that incorporate linguistic syntax start with word level alignments and a parse tree for one side of the language pair, and obtain phrase tables and hierarchical translation rules driven by the syntax. While this has indeed proven successful [Yamada and Knight, 2001] [Marcu et al., 2006], it has been shown that the word alignments which are usually extracted using syntactically uninformed generative models are not optimal for the syntactic phrase extraction problem [DeNeefe et al., 2007, DeNero and Klein, 2007]. Some

approaches [Crego and Habash, 2008, Fossum et al., 2008] have been proposed to modify the word alignments in ways that make them more amenable to building syntactic models.

The hierarchical grammar rules can be understood as context-free re-writing systems with left-hand side (LHS) and a right-hand side (RHS). The LHS of a particular rule can correspond to multiple RHS , resulting in an ambiguity factor that is typically resolved during decoding as a search problem. SCFG independence assumptions reflect upon the ambiguity factor seen in the grammars. Syntax based translation systems usually resort to a number of syntactic features and scoring techniques to resolve this ambiguity. However, SCFG translation models that are often learnt from parallel corpus and syntactic parse trees under certain independence assumptions are often so generalized that it becomes quite difficult to explore a meaningfully small space to obtain the right translation hypothesis.

2.2 Active Learning

2.2.1 Active Learning for Natural Language Processing

Over the past two decades, supervised learning in structured spaces has been quite successful in syntactic analysis problems in natural language processing. With the initial success of such techniques in statistical parsing [Charniak, 2000], sequence labeling tasks [Brill, 1992], more researchers have been focusing on learning from data to solve NLP tasks. These learning techniques exploit large amounts of annotated data to learn models that can perform linguistic analysis on unseen data. The quantities of the annotated data are far from being sufficient for the majority of languages. Acquiring such supervised linguistic annotations for a language is important for natural language processing and it usually involves significant human efforts. Languages like English have been well supported in the linguistics community, and therefore there is a wealth of language analysis tools for them, but even for English it becomes a major challenge to find data while adapting the existing tools to different domains.

Researchers have quickly realized this problem and are moving to ways of reducing the burden of annotation, so as to reduce the time to production of their tools. Active Learning is a field of Machine Learning that studies selective sampling of most informative examples for annotation. In this section, we survey the literature for active learning applications in Natural Language Processing (NLP). For a more general survey of Active Learning we refer to [Settles, 2009]

Active learning has been applied successfully in the area of Computer Vision for collecting data to improve the object identification task and face recognition [Hewitt and Belongie, 2006] among others. In speech technologies, data collection for building automatic speech recognition systems benefited by application of active learning techniques [Hakkani-tr et al.,

2002]. Annotated data is used by most supervised and semi-supervised learners in NLP to solve two kinds of problems, classification tasks such as text classification [McCallum and Nigam, 1998], and structure prediction tasks such as statistical parsing [Charniak, 2000], sequence labeling [Brill, 1992].

[Thompson et al., 1999] applied uncertainty based active learning to the task of training a semantic parser. They consider two systems CHILL [Zelle and Mooney, 1996] and RAPIER that map sentences to their semantic representation. They use an uncertainty criteria which selects sentences that are not parsed by the systems. [Roth and Small, 2006] present a margin-based method for active learning in structured outputs. In particular they apply this to the task of semantic role labeling which is the task of identifying and labeling the semantic arguments of a predicate. In the case of margin-based classifiers, uncertainty translates to the distance from the hyper-plane. They explore the tradeoff between selecting instances based on a global margin or a combination of the margin of local classifiers.

Grammar induction is the task of inferring grammatical structure of a language. Research in statistical parser induction [Charniak, 2000] shows that given enough labeled data with good quality annotations, grammar induction can be done with reasonable accuracy. [Hwa, 2004] use selective sampling to minimize the amount of annotation needed for corpus based grammar induction. They use uncertainty based selective sampling techniques, where the uncertainty is computed as the length of the sentence or a tree entropy metric. They select sentences for annotation that have a uniform parse tree distribution. [Steedman et al., 2003b] experiment uncertainty based metrics as provided by their parser to select sentences and show that improved accuracy can be achieved by annotating fewer sentences. [Baldridge and Osborne, 2003] applies similar approaches to selection of sentences for grammar induction under a HPSG grammar formalism.

2.2.2 Machine translation and active learning

For statistical MT, application of active learning has been focused on the task of selecting the most informative sentences to train the model - in order to reduce cost of data acquisition. Chris Callison-Burch provided a research proposal that lays out a plan of action for active learning for SMT [Callison-burch, 2003]. However it lacked any further experimentation and results. Recent work in this area discussed multiple query selection strategies for a Statistical Phrase Based Translation system [Haffari et al., 2009]. Their framework requires source text to be translated by the system, and the translated data is used in a self-training setting to train MT models. [Haffari and Sarkar, 2009] discuss an active learning task of introducing a new language pair into an existing multilingual set of parallel texts with a high quality MT system for each pair. This novel setup is applicable when working with multi-language parallel corpus, and they exploit it to propose new sentence selection strategies and features.

[Eck et al., 2005] apply active learning as a weighting scheme to select more informative sentences in order to port their MT system to low resource devices like PDAs. The informativeness of a sentence is estimated using unseen n-grams in previously selected sentences. [Eck et al., 2005] use a weighting scheme to select more informative sentences, wherein the importance is estimated using unseen n-grams in previously selected sentences. Although our selection strategy has a density based motivation similar to theirs, we augment this by adding a diminishing effect to discourage the domination of density and favor unseen n-grams. Our approach, therefore, naturally works well in pool-based active learning strategy when compared to [Eck et al., 2005]. In case of instance-based active learning, both approaches work comparably, with our approach working slightly better.

[Gangadharaiah et al., 2009] use a pool-based strategy that maximizes a measure of expected future improvement to sample instances from a large parallel corpus. They assume the existence of target-side translations along with the source-side sentences to compute features useful in estimating utility of a sentence pair. [Gangadharaiah et al., 2009] use a pool-based strategy that maximizes a measure of expected future improvement, to sample instances from a large parallel corpus. Their goal is to select the most informative sentence pairs to build an MT system, and hence they assume the existence of target-side translations along with the source-side sentences. We however are interested in selecting most informative sentences to reduce the effort and cost involved in translation.

In an effort to provide translations for difficult-to-translate phrases, [Mohit and Hwa, 2007] propose a technique to classify phrases as difficult and then call for humans to translate them. This can be seen as an active learning strategy for selection at phrase level. [Kato and Barnard, 2007] implement active learning and semi-supervised approaches to build MT systems for language pairs with scarce resources. They show results with very limited data simulated with languages like Afrikaans, Setswana, etc.

2.2.3 Cost-sensitive Active Learning

We may have to build upon recent work in multi-task active learning that looks at selection of corpus for two relatively different tasks [Reichart et al., 2008b]. They experiment with different ways of combining the ranking of unlabeled data as optimized to each of the tasks. [Krause and Horvitz, 2008] applies cost sensitive learning to the task of asking questions for learning privacy preferences of an individual, while [Melville et al., 2005] look at application in feature annotation vs. instance annotation. In translation optimizing cost for low-resource languages is important and we will be working on optimizing cost in a multi-task scenario: translation vs. word-alignment.

2.3 Crowd Sourcing

Vision, hearing, natural language synthesis and also inference are still done better by human today. Given that it is cheaper to have large amount of data analyzed by computer but more accurate and phenomenally more expensive to have it done by humans, Internet makes it possible to break bigger problems or bigger data into smaller pieces and present them to world-wide community of humans as micro-tasks for micro-payments. Humans are more willing to carry out small tasks at their own convenience to provide the service collectively at a much cheaper cost or even in the interest of the greater good. Crowd-sourcing makes it possible to find people who would be able to combine their passion with pass-time to create a resource for consumption by the society.

2.3.1 Crowdsourcing and Data Annotation

Crowd-sourcing is the process of farming out tasks to a large number of users over the web. These tasks are typically performed by a resident employee or a contractor with a specific area of expertise. With crowd-sourcing such tasks are requested from a crowd, which is a set of non-experts. Recently, crowd-sourcing has become popular as human computing, where tasks that are challenging, difficult or time-consuming for computers are passed to human crowds over the web. These tasks broadly belong to the language or vision community, where for a number of tasks it is still impossible for computers, but only requires a few seconds for a human to complete. For example, identifying a person in a photograph, tagging a video for a particular event, flagging an email for spam, identifying the sentiment of a written text, spotting characters in an image are still some of the challenge research problems to computers.

With the advent of online market places such as Amazon Mechanical Turk, it is now easier to reach annotators on the Web than ever before, even if most of them are not expert translators. Researchers in the Natural Language Processing community are quickly exploiting 'crowd-sourcing' for acquisition of supervised data [Snow et al., 2008] and conducting user studies [Kittur et al., 2008] where annotation tasks are farmed out to a large group of users on the web utilizing micro payments. [Snow et al., 2008] discuss usability of annotations created by using Mechanical Turk for a variety of NLP tasks - primarily supervised learning tasks for classification. These tasks included word sense disambiguation, word similarity, textual entailment, and temporal ordering of events. The interest in Natural Language Processing community is also evident from the recently concluded workshop on crowdsourcing Callison-Burch and Dredze [2010], where researchers have used Amazon's Mechanical Turk to produce annotations relevant to their fields of research like word-alignment, sentiment analysis, dictionary collection, translations etc.

2.3.2 Crowdsourcing and Translation

Large scale parallel data generation for new language pairs requires intensive human effort and availability of bilingual speakers. Only a few languages in the world enjoy sustained research interest and continuous financial support for development of automatic translation systems. For most remaining languages there is very little interest or funding available and limited or expensive access to experts for data elicitation. Crowd-sourcing compensates for the lack of experts with a large pool of expert/non-expert crowd. However, crowd-sourcing has thus far been explored in the context of eliciting annotations for a supervised classification task, typically monolingual in nature Snow et al. [2008]. In this project we test the feasibility of eliciting parallel data for Machine Translation (MT) using Mechanical Turk (MTurk). MT poses an interesting challenge as we require turkers to have understanding/writing skills in both the languages. Our work is similar to some recent work on crowd-sourcing and machine translation Ambati et al. [2010a], Callison-Burch [2009].

Recent efforts in MT include feasibility studies for using crowd sourcing techniques for MT Evaluation; users are provided with translations from multiple systems and asked to select the correct one Callison-Burch [2009], Zaidan and Callison-Burch [2009]. One observation that Callison-Burch [2009] make is about the availability of bilingual speakers for annotation tasks in MT. They observe that it is relatively more difficult to find translators for low-resource languages like Urdu, Thai, etc. than it is to find for Chinese, Arabic, Spanish, etc. With the increasing pervasiveness of the Internet, and more and more people in the developing world gaining computer literacy, the situation should ameliorate.

Recently, Facebook has had great success with turning its user base into a 'hub' of translators to internationalize its portal. For obvious reasons of privacy and sheer size of the volume, such techniques may not be applicable when translating user-generated content. However, the potential of the crowd can be harnessed to build parallel corpora which can then be used to train automatic MT systems.

2.3.3 Learning in the Crowd

In case of Machine Translation, although services like Mechanical Turk have opened doors to tap human potential, they do not guarantee translation expertise nor large-volume availability of translators. Here proactive learning [Donmez and Carbonell, 2008] adds other useful dimensions to active learning in a crowd-sourcing scenario - such as coping with information sources of variable cost and variable reliability of information sources [Donmez et al., 2009, 2010] and jointly optimizing the selection of instance (what to translate or align) with the selection of source (which individual or group to rely upon) modulated by cost (more reliable annotators may be more expensive). These techniques become increasingly beneficial as we move to a crowd-sourcing scenario for eliciting annotations. In

our case, the trade-off may be between an expert translator and multiple cheaper translators of unknown reliability.

Some very recent work also addresses the modeling of turker reliability in an expectation-maximization (EM) framework [Raykar et al., 2010, Ipeirotis et al., 2010]. Raykar et al. [2010] also discuss modeling of turker reliability jointly with task difficulty in a generative framework.

Chapter 3

Active Learning for Parallel Data Creation

Building a large-scale SMT system requires vast amounts of parallel corpora. Such parallel corpora are unavailable for most language-pairs and creating them requires an expert translator, which is expensive. We hypothesize that all sentence-pairs are not made equal and so all sentences-pairs do not contribute equally to the performance of an MT system. Therefore, in this chapter we discuss several approaches to actively prioritize and select sentences from a large pool of source language monolingual corpus, in order to be translated by an expert for creation of parallel corpora useful in building MT systems.

3.1 Introduction

Active learning is appropriate for tasks where unlabeled data is readily available, but obtaining annotations for such data is expensive. While there has been a lot of work done in the application of active learning for classification tasks such as document categorization [McCallum and Nigam, 1998] , active learning for structural learning tasks has received considerably less attention. However, due to the structured nature of labeling tasks, annotating these instances can be rather tedious and time-consuming, making active learning an attractive alternative. Recently successful attempts have been made in applying active learning to sequence labeling problems such as part-of-speech tagging, entity tagging and other tasks like parsing [Steedman et al., 2003b], information extraction [Roth and Small, 2006], etc.

In MT, only a few attempts have been made in applying active learning techniques, but mostly using uncertainty sampling approaches [Haffari et al., 2009, Eck et al., 2005]. The methods of evaluation and applied strategies however make it difficult to draw conclusions

on the comparative effectiveness of various active learning approaches. In this section we make an attempt to establish a setup for active learning in MT, and categorize multiple query strategies based on the resources they use and their effectiveness. We compare them under multiple data settings and evaluation metrics to study their effectiveness on a varied set of language-pairs. We also discuss batch selection mechanisms and their importance in the MT setup. We discuss our novel approach to active sentence selection, which is a combination of the informativeness, uncertainty of a sentence along with the diversity it offers in the batch.

3.2 Active Sentence Selection for Parallel Data Creation

3.2.1 Setup

We first discuss our general framework for active learning in SMT followed by the selection approaches. We start with an unlabeled dataset $U_0 = \{f_j\}$ and a seed labeled dataset $L_0 = \{(f_j, e_j)\}$, where labels are the translations. We then score all the sentences in the U_0 according to our selection strategy and retrieve the best scoring sentence or a small batch of sentences. This sentence is translated and the sentence pair is added to the labeled set L_0. However, re-training and re-tuning an SMT system after translating every single sentence is computationally inefficient and may not have a significant effect on the underlying models. We, therefore continue to select a batch of N sentences before retraining the system on newly created labeled set $L_{k=1}$. Our framework for active learning in SMT is shown in Algorithm 7.

Algorithm 1 ACTIVE LEARNING SETUP FOR SENTENCE SELECTION TASK IN SMT

1: Given Labeled Data Set : L_0
2: Given Unlabeled Data Set: U_0
3: **for** $k = 0$ to T **do**
4: **for** $i = 0$ to N **do**
5: s_i = Query(U_i,L_i)
6: t_i = Human Translation for s_i
7: S_k = $S_k \cup (s_i, t_i)$
8: **end for**
9: U_{k+1} = U_k - S_k
10: L_{k+1} = $L_k \cup S_k$
11: Re-train MT system on L_{k+1}
12: **end for**

3.2.2 Evaluation

The effectiveness of active learning approaches is typically measured in terms of the cost savings that result due to the prioritized selection of data. Most active learning approaches consider a uniform cost for elicitation of instances, which clearly is not true for MT. Estimating the true cost model for translation of sentence is a challenging problem. Sentences, which are the instances here, are of different lengths to start with and a long sentence takes more time to be translated than a short sentence. But, if we only consider time taken to translate a sentence as a metric of cost, we are faced with the issue of accurate measurement of time across browsers and platforms. This situation is exacerbated by the fact that humans work at different paces. We, therefore treat the cost of translation per sentence as the sum of the costs of translating the consisting words. Although we acknowledge that cost of words may not be uniform, with some words being more difficult to translate than others, in this work we assume a uniform cost of word translation. Our success criteria, therefore, is to optimize the number of words translated for the improvement in translation quality. To measure improvement in translation quality, we use automatic translation metrics like BLEU [Papineni et al., 2002] and METEOR [Lavie and Agarwal, 2007].

3.2.3 Batch Mode Active Learning

In most active learning research, queries are selected one at a time. In machine translation and other structure prediction tasks, the training time required to induce a model is slow and, therefore expensive making it difficult to work with labeling single units. While annotation was considered a cumbersome process, as it involves working at much slower speeds than model training, with the advent of web platforms like Amazon's Mechanical Turk (MTurk) it is now possible to do large scale labeling in parallel using large crowds. Batch model learning therefore is an interesting alternative to online active learning.

The challenge in batch mode learning ,however, is to select the batch appropriately so that the selected instances have minimal overlap among each other, in order to not introduce redundancy. Also, the selection of each instance influences the underlying model, and in turn affects the scores of the future instances to be selected. When the retraining of the model is not possible due to computational complexity, the selection of instances is sub-optimal due to the staleness of the model parameters. In our work introduce a decay parameter based on the number of times the instance has been seen in the batch previously. As will be discussed later, this component can be coupled with any of our query strategies in order to combat the batch selection problem.

3.3 Query Selection Strategies

In this section we discuss our various query strategies for sentence selection. We categorize the multiple strategies into three categories based on the resources used in computation of the value of the sentences.

3.3.1 Data-Driven Selection

Data-driven sentence selection strategies only use the monolingual data U and the bilingual parallel corpus L to select sentences for translation. Approaches in this category are independent of the underlying MT system and so are agnostic to the model representation. This makes our approach applicable to any corpus-based MT paradigm and system, even though we test on the statistical phrase-based MT paradigm in this thesis.

Density based approach

An important criteria for building an MT system for a language-pair is to be able to decode an unseen source-language sentence. Therefore, we propose to select and have translated those sentences, that provide a maximal source-side lexical coverage for the monolingual corpus. In other words, we wish to select a very representative sample of sentences that are most similar to the rest of the data, as shown below. This enables an MT system, built using the resulting parallel corpus, to have better lexical coverage for translating future sentences that follow the distribution as the monolingual corpus.

$$s = \arg\max_{s} sim(s, U) \tag{3.1}$$

The basic units of an SMT system are phrases and therefore we measure the similarity across sentences in terms of the consisting phrases . Our scoring strategy is shown in Equation 3.2. We select sentences that have the most representative n-grams of the bilingual corpus. Representativeness or the 'density' of a sentence is computed using $P(x|U)$, the relative likelihood estimates of an n-gram x in the unlabeled monolingual data U. We also introduce a decay on the density of an n-gram, based on its frequency in the labeled data L. We define $Phrases(s)$ as a function that computes the set of all phrases up to size $n = 3$. We use a decay parameter λ to diminish the true density of the n-gram as a count of the number of times it has been seen and already labeled. We choose this decay function instead of completely ignoring already seen phrases for two reasons. Firstly, phrases that are dense are typically polysemic with a higher translation fan-out, which means they are translated into different target sides in different contexts. A gradual decay allows us to sample the phrase

multiple times allowing us to capture the spectrum of variations. Secondly, the decay also works against redundancy within the batch, allowing us to do a batch selection.

$$dden(s) = \frac{\sum_{x \in Phrases(s)} P(x|U) * e^{-\lambda count(x|L)}}{\|Phrases(s)\|} \tag{3.2}$$

$$P(x|U) = \frac{count(x)}{\sum_{x' \in Phrases(U)} count(x')} \tag{3.3}$$

Diversity based approach

One of the desirable properties of a machine translation system is to have a high coverage of the vocabulary of the language pair. Therefore, the active learning selection strategy should favor sentences that provide new vocabulary that is different from the sentences that have already been labeled or translated, thus far. However, given that the data will be used to train a phrase based SMT system, we aim to improve new phrasal coverage instead of single word coverage. Therefore, our strategy, called diversity or novelty, is computed as the number of new phrases that a sentence has to offer on the source side, where phrases can be computed up to a certain length n. Similar to above, we define $Phrases(s)$ as a function that computes the set of all phrases up to size $n = 3$ and therefore the novelty score can be computed from the labeled data as shown in Equation 3.4.

$$div(s) = \frac{\sum_{x}^{Phrases(s)} \alpha}{|Phrases(s)|} \alpha = \begin{cases} 1 & x \notin Phrases(L) \\ 0 & \end{cases} \tag{3.4}$$

Density Weighted Diversity Ensemble

Ensemble techniques tend to work better as they provide different perspectives of the value function [Melville and Mooney, 2004, Freund et al., 1997]. We, therefore, devise a density and diversity ensemble called 'density weighted diversity ensemble' (DWDS) that favors dense and novel phrases, by computing the final score of a sentence as the harmonic mean of the two metrics with a tunable parameter 'β'. This tunable parameter helps us balance the novelty and density factors. We choose $\beta = 1$ and $\lambda = 1$ for our current experiments. Thus far, we have only considered n-grams of size up to 3.

$$dwds(s) = \frac{(1 + \beta^2) d(s) * u(s)}{\beta^2 d(s) + u(s)} \tag{3.5}$$

KL Divergence

In probability theory and information theory, KullbackLeibler divergence (also known as relative entropy, or KL divergence) is a non-symmetric measure of the difference between any two probability distributions P and Q. This measure can also be treated as a distance between the two distributions, although the notion of the distance is not traditional as it is directional as well. Inspired by this notion, we propose a new active learning strategy for active selection of sentences for translation. We, first, define the two distributions as the following. The first is defined by the monolingual unlabeled data U and the second defined by all the source sides of the labeled data L.

We formulate the KL divergence strategy (kl-div) as seen in Equation 3.6, where we select the sentence s that contributes highly to the total KL-divergence between the unlabeled data distribution $P(.|U)$ and the labeled data distribution $P(.|L)$.

$$kldiv(s) = P(s|U)log\frac{P(s|U)}{P(s|L)} \tag{3.6}$$

$$s* = \arg\max_{s} kldiv(s) \tag{3.7}$$

We also tried a variation of the strategy that exploits the directionality of the KL-Divergence formulation and selected a sentence that contributes to the total KL-divergence between the labeled distribution and the unlabeled data distribution. This however did not make a significant difference to the overall performance of the translation system and so we will only report results using the above formulation for KL-Divergence.

3.3.2 Model-Driven Selection

Phrasal Entropy

Unlike the data-driven approaches mentioned above, in the model-driven approaches, we make use of the model, trained using the labeled data. Therefore, the model based approaches are specific to the MT system and the underlying model representation, which in our case is Phrase Based Statistical Machine Translation system (PB-SMT). This strategy can be called an uncertainty sampling technique [Lewis and Catlett, 1994], quite popular in active learning.

The translation model in SMT, typically consists of phrase tables with bidirectional translation scores and bidirectional lexicon tables that can be used to compute source-given-target and target-given-source probabilities at word level. We use $t2s$ to denote computation using 'target-given-source' models and $s2t$ to denote computation using 'source-given-target' models. Given a source-side sentence, we propose an uncertainty metric, inspired by

entropy as shown in Equation 3.9. We select that sentence which has a maximum cumulative entropy under the phrase translation model as computed below, where $Phr(x)$ denotes all the phrases present in a sentence x, $Trans(x)$ denotes all the phrase translations of x and $p(t|s)$ is the conditional translation probability distribution of phrases.

$$cond\_entropy(s) = \frac{\sum_{x}^{Phr(s)} \sum_{y}^{Trans(x)} -p(y|x) * log(p(y|x))}{|Phrases(s)|} \tag{3.8}$$

$$s* = \arg\max_s cond\_entropy(s) \tag{3.9}$$

Since we use the score to prioritize the sentences, we are effectively preferring a sentence containing more number of high uncertain phrases. The intuition is that the translation of such a sentence will lead to maximal reduction of entropy of the current translation model.

Decoding Score

We also propose another uncertainty metric to account for the translation difficulty of a sentence under a given model and translation system. One of the strong indicators of the overall translation quality is the conditional distribution of the target hypothesis over the source $P(t|s)$. Incase of SMT, the decoder computes this target conditional distribution as a log-linear score of the various features, including lexical, phrasal and language model scores and therefore, can also be treated as a cumulative score of the various sub-models. The scores produced by the SMT system are for the given sentence and cannot be compared with scores of a different sentence. We normalize the score by the number of words in order for it to be comparable across different source sentences. We select the sentences with low score for the $decode\_score(s)$, indicative of the weakness of the model in translation of the sentence.

$$decode\_score(s) = \frac{1}{len|s|} \arg\max P_\theta(t|s)$$

$$s* = \arg\min_s decode\_score(s)$$

The decoding score has also been previously used as a confidence metric [Ueffing and Ney, 2007] and is a good indicator of the translation difficulty. It is computationally expensive as it can only be computed by running the SMT system on the entire unlabeled data. Also, this can only be done in a batch mode as it is infeasible to re-build the models and to re-decode in order to compute the scores after every addition of a single parallel sentence-pair.

3.4 Experiments

We test our active learning strategies and report performance on three very different language pairs. We select the languages Spanish (Romance language), Japanese (Altaic language) and Urdu (Indo-European) and build translation systems that translate into the English language.

For all the language-pairs we follow the same approach of experimentation. In order to test the performance of our active learning sentence selection strategies we first start with a seed parallel data of 1000 sentence pairs and train an SMT system using this as the training data and evaluate the performance on a respective test dataset. We then continue to iteratively train the MT system by increasing the training dataset in batches of 1000 sentences at a time. In each iteration, the data is selectively sampled using one of the active learning strategies from the source side of the monolingual unlabeled corpus. We simulate the human translation step in all our experiments, as we already have access to the translations of the entire corpus. We use the resulting parallel corpora to retrain the system and re-tune and test on the held-out data sets to complete the iteration.

We use, Moses, open-source translation system [Koehn et al., 2007] for extraction, training and tuning our system. We build an SRILM language model using English-side of the Europarl corpus, which consists of 1.6M sentences or 300M words. Throughout the experiments in this thesis we use the same language model and do not vary the domain or size of the language model. The weights of the different translation features are tuned using standard MERT [Och, 2003].

In all the experiments, in addition to our different techniques we also compare with two strong baselines. First is a random baseline, where sentence pairs are sampled at random from the unlabeled dataset. Random baselines are strong as they capture the underlying data distribution when sampled in large numbers. The second baseline is where we select data based on the order it appears in the corpus. This is natural as this would be the manner in which we would provide data to an expert for translation. We will then plot graphs to report the performance. The x-axis on each of the graphs is the number of words of parallel data used for training the system on specific language-pair. It is an indicator of the human effort involved in translating the source language sentences. The y-axis of the graphs shows the performance of the final SMT system as measured by BLEU [Papineni et al., 2002] on a held-out dataset.

3.4.1 Spanish-English

For the experiments on the Spanish-English language-pair, we use BTEC parallel corpus [Takezawa et al., 2002] from the IWSLT tasks. This dataset consists of 127K sentence pairs. Our development set consists of 343 sentences. The test set used consists of 500 sentences.

As seen in Figure 3.1, some of our active learning strategies perform better than the random baseline. Our most significant improvements come from the density weighted diversity ensemble (DWDS). The difference between the curves is most prominent in the earlier iterations indicating the importance of a density based sampling strategy for very low resource scenarios. One way to read the results is that for the same amount of parallel sentences used, active learning helps to select more informative sentences and hence achieves better performance. Alternatively, we can understand this as, given an MT system, active learning strategy uses less number of sentences to reach a desired accuracy, thereby reducing the cost of acquiring data. This shows that we can achieve similar performances by spending lot less for translation. For example, upon consumption of 10000 words of parallel data, DWDS selection strategy achieves 2 BLEU points higher than random selection strategy. Another observation from the graph is that DWDS achieves 30.5 BLEU points on a held-out test set by training on about 27% less data than random selection.

3.4.2 Japanese-English

For the Japanese-English language-pair, we use the BTEC travel corpus released under the IWSLT 2004 task. This parallel corpus consists of 162,318 sentence pairs. The tuning set/dev set consists of 500 sentences and the test set consists of 506 sentence pairs. Although we have 16 different references available for the development set as well as the test set, we only use the first one and tune and test with a single reference. The performance of multiple active learning strategies is shown in Figure 3.2. We notice improvement with different query strategies like KL-DIV, DWDS, and the best performing strategy with significant margins is still the DWDS strategy.

3.4.3 Urdu-English

We also tried a true low resource language pair: Urdu-English. Urdu is one of the two official languages of Pakistan, the other being English. It belongs to the Indo-European family of languages, and has its vocabulary developed under Persian, Arabic, Turkic, and Sanskrit. Modern Urdu has a significant influence from Punjabi and English, and use of English words in sentences is not a rare phenomenon even in genres such as newswire.

Language Data Consortium has recently released parallel data for the Urdu-English Machine Translation track at the NIST Evaluation 2008. It consists of web data segments accumulated from multiple news websites and other manually created sentences. The data released by NIST consists of 92K parallel segments which we re-segmented to obtain 62K parallel sentence pairs. We use this dataset to conduct our active learning simulation for Urdu-English. The language model used is again the English side of the Europarl parliamentary dataset used for training the Spanish-English MT system in the previous

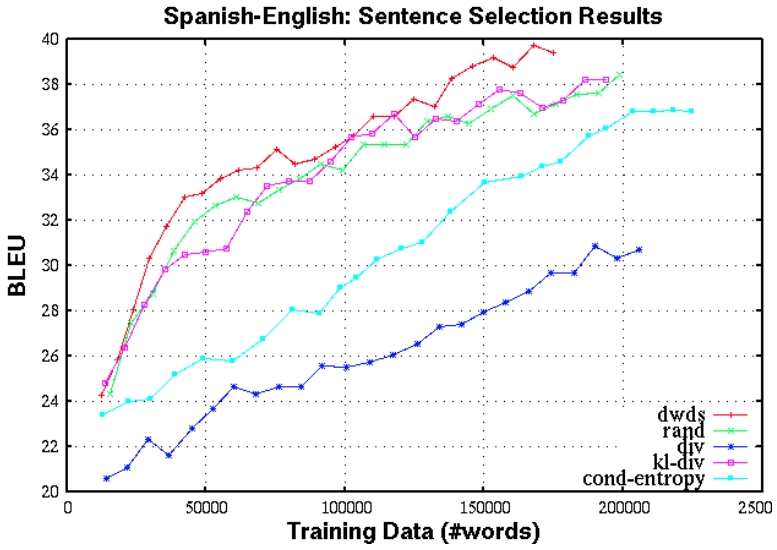

Figure 3.1: Performance of Spanish-English MT systems trained/tuned and tested individually by selecting sentences using various active learning strategies. Cost is equated to # source-language words translated, and performance is plotted using BLEU score on a held-out test set. Density weighted diversity sampling outperforms all baselines in selecting most informative sentences for improvement of MT systems

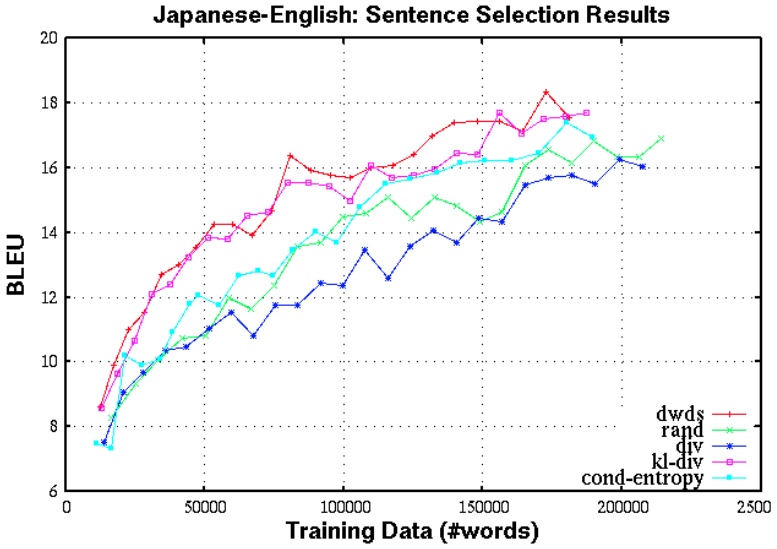

Figure 3.2: Performance of Japanese-English MT systems trained/tuned and tested individually by selecting sentences using various active learning strategies. Cost is equated to # source-language words translated, and performance is plotted using BLEU score on a held-out test set. Density weighted diversity sampling outperforms all baselines in selecting most informative sentences for improvement of MT systems

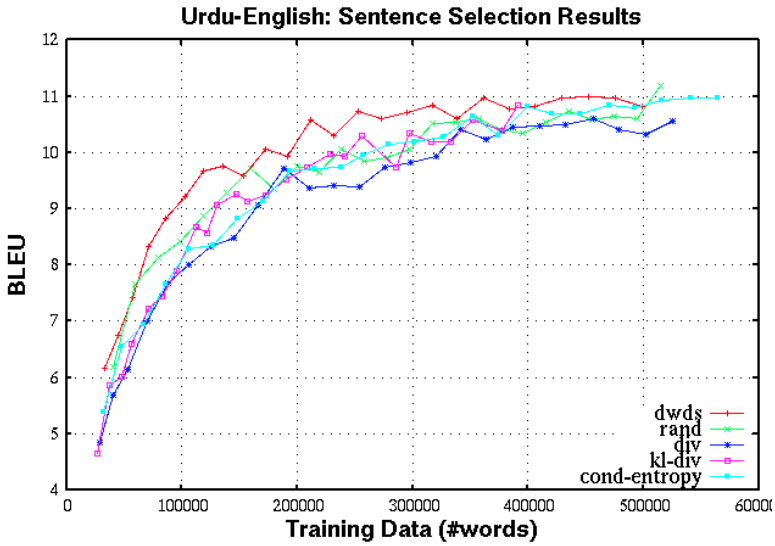

Figure 3.3: Performance of Urdu-English MT systems trained/tuned and tested individually by selecting sentences using various active learning strategies. Cost is equated to # source-language words translated, and performance is plotted using BLEU score on a held-out test set. Density weighted diversity sampling outperforms all baselines in selecting most informative sentences for improvement of MT systems

section. System tuning was done with minimum error-rate training on a subset of 450 sentences, selected from the NIST DEV08 data set with one reference translations available. We use the rest of the 450 sentences from the NIST DEV'08 for the test set. Our post-processing includes a script to reattach as much punctuation as possible to the preceding or following word. Ambiguous marks, such as straight double quotes, are left out as separate tokens.

Even in the Urdu-English case, we see that the density diversity ensemble (DWDS) outperforms other strategies and baselines significantly all across the curve with very noticeable improvement in the initial parts of the curve. For this particular language pair we notice that random selection performs comparably with the rest of our active learning approaches.

3.5 Analysis

3.5.1 Does Domain Affect Selection Strategy?

Given that the success of density based approaches can be attributed heavily to the assumption that the underlying monolingual data follows a certain distribution, we would like to compare the robustness of the active learning approaches in the absence of such in-domain monolingual data. Due to the domain mismatch between training and test conditions, all the active learning strategies are expected to under-perform. However, we would like to see whether the relative distinctions between the strategies still hold.

In order to test this, we conduct an experiment where the Spanish-English monolingual data belongs to the political domain (Europarl corpus) and the MT system we are building is for the travel domain (BTEC corpus). We test our best performing active learning strategy on Spanish-English on data of different genre. We used the Europarl data released for the WMT 2008 experiments and the active learning results with our best performing learning strategy against a standard baseline, which can be seen in Figure 3.4.

We observe that, even while the DWDS approach is affected in general by the domain mismatch, as long as the underlying monolingual data belongs to a single domain, tracking the distribution still proves to be a better strategy. We also notice that the other strategies are also equally affected due to the domain mismatch and the relative differences between the performance of each curves still holds. In later sections of the thesis we will relax this assumption and discuss learning strategies when we are provided with mixed domain data.

3.5.2 Model-based vs. Data-driven

In the previous sections we have discussed two model-driven active learning strategies both of which can broadly be categorized as uncertainty based techniques. Uncertainty selection strategy has been one of the most successful of all active learning methods developed in literature. For most supervised learning tasks involving the training of a classifier, it has been proven that model uncertainty is one of the best performing approaches [Lewis and Catlett, 1994]. In problems related to classification, the intuition behind an uncertainty technique is clearer as we want to redraw the classification decision boundary by sampling in high uncertainty regions of the unlabeled data.

However, in case of MT which is a complex ranking task, it is unclear if uncertainty of the model even makes sense. For instance, a sentence that has not been seen in the training data will always have a much higher level of uncertainty than any other difficult-to-translate sentence which has already been seen in the labeled data. Also, given that a translation system consists of several other sub-models combined at a feature level, it may be the case that uncertainty should be considered for individual models, and a cumulative

understanding of uncertainty is always lacking.

Further, there are two other issues with model-based uncertainty approaches which make them less appealing in case of translation:

- **Computation Cost:** Our active learning framework is a pool-based approach, where a set of sentences are selected from a pool of unlabeled data for labeling. An issue in such an approach is the size of the batch. While small batches are preferable due to a better possibility of error recovery, it becomes infeasible to retrain the SMT system after each batch. Until we come up with incremental approaches for updating SMT models without having to fully re-train the system, model-based approaches may be restricted to constrained data scenarios.
- **Overlap with Diversity Approaches:** We have also observed in our experiments that the model-based strategies closely follow the performance of 'diversity' selection strategy. This also makes intuitive sense, as diversity focuses on reducing the out-of-vocabulary words while selecting sentences. Unseen words also account for the uncertainty of the model during the evolution of a translation system.

Therefore, although we have discussed model based uncertainty approaches for active learning, we have observed that in the low resource scenarios, data-driven approaches outperform the benefits of model-based approaches. In Figure 3.5, we show the learning curves of our diversity sampling strategy and one of the model-based uncertainty approaches (decoding score). We can see that the diversity strategy that is easy to compute and is agnostic to the MT system, and the underlying model training, works similar to a more computationally intensive decoding strategy.

3.5.3 Cost Function

The main aim of active learning is to reduce the overall cost of labeling. However, we do not understand what the cost function is for language translations. In this section, we attempt to list out the various cost functions and discuss merits and demerits of each of them. It is advisable to evaluate using accurate cost accounting so as to precisely understand the effectiveness of the AL algorithms. An inaccurate estimate of cost would provide an incorrect comparison of the performance of the active learning strategies.

We could primarily use the following cost models for sentence selection:

- **# of Sentences**: Each sentence is treated as being equally difficult to translate as any other sentence and therefore, uniform cost is considered.
- **# of Words**: The cost of translation per sentence is treated as the sum of cost of

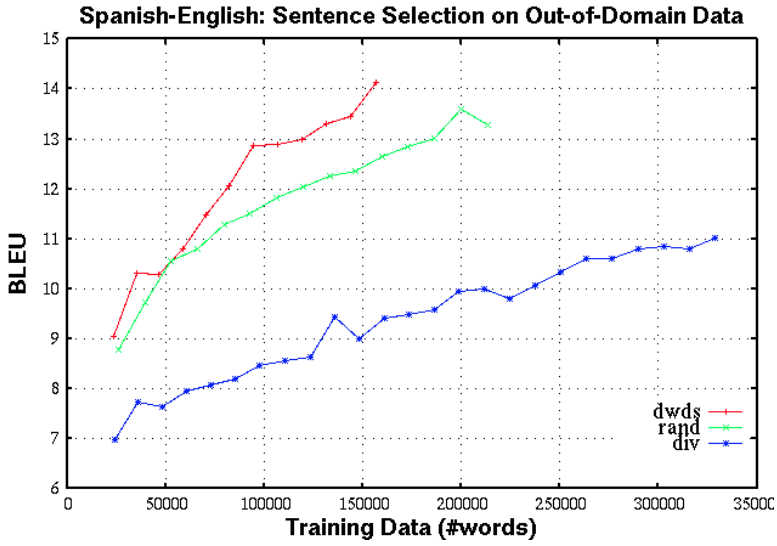

Figure 3.4: Performance of Spanish-English MT systems trained/tuned and tested on Travel domain (BTEC) data, but sentence selection conducted on Politics domain (out-of-domain) monolingual data. Even in this case density weighted diversity sampling outperforms both diversity and random selection baselines.

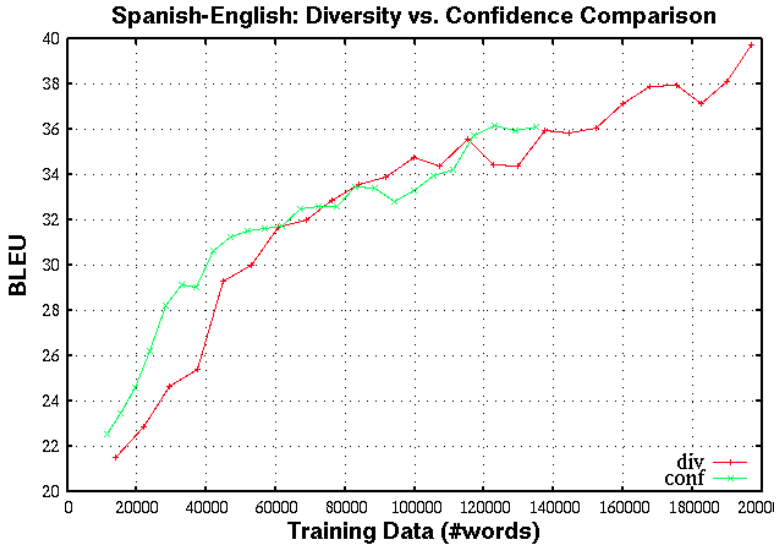

Figure 3.5: Performance of Spanish-English MT systems trained/tuned and tested individually by selecting sentences using diversity strategy (data-driven) and confidence strategy (model-driven). We notice that performance in both cases is similar and comparable, but diversity is easy to compute while confidence strategy requires decoding unlabeled data.

translation of the consisting words. However, cost of translation of each word is uniform.

- **Effort estimate**: A good surrogate for effort can be the time taken to complete the translation. However, such information may not be easily available for existing datasets and so we will not be using it in our experiments.

Haffari et al. [2009] use number of sentences annotated and their affect on the MT system performance as a measure for the effectiveness of AL approaches. We observe that this is not a prudent approach to measure the effectiveness of AL approaches whose goal is typically to reduce the annotation efforts of unlabeled data. While it is still arguable that the number of words in a sentence is an indicator of the difficulty of translation of the sentence, it is safe to assume that a longer sentence takes longer time to translate and therefore stands as an approximate surrogate for difficulty of translation.

3.6 Context- and Resource- Aware Active Learning

Annotations in MT can be of various kinds depending upon the paradigm of translation. In our case, since we work with a Statistical Machine Translation (SMT) system, our task is to seek target-language translation for a source-language sentence. In this thesis, we proposed a novel query strategy, Density Weighted Diversity Sampling (DWDS) which focuses on both diversity and density metrics in selecting a sentence. Our approach works significantly better than other baselines, as reported in our experiments section.

We also explore multiple active learning query strategies for the task of sentence selection. We observe that some methods perform well in initial phases where very few instances have been sampled, while others perform better in later operating ranges upon substantial sampling. For instance, density estimation methods [Nguyen and Smeulders, 2004] perform well with minimal labeled data, since they sample from maximal-density unlabeled regions, and thus build an MT model that is capable of translating majority of the remaining unlabeled data. On the other hand, diversity sampling method focuses more on improving recall by favoring unseen words irrespective of their representativeness in the data. With the awareness of the performance of a query strategy under a particular operating range, we propose multi-strategy query methods that can perform better under a larger operating range by selecting optimal query strategy for different operating ranges. Therefore, an active learning strategy needs to sample sentences in the context of the evolution of the underlying model. It turns out that we can improve the performance of the MT system under a larger operating range by resorting to hybrid approaches that focus on combining different strategies for different operating ranges.

We consider two different strategies for sentence selection in MT, that have varying

returns in different phases of translation. The first method is our density oriented approach (DWDS), which focuses on maximally-dense n-grams in the unlabeled data. The second method is a Diversity sampling (DIV) approach which focuses on n-grams that are different from those already present in the labeled data. Inspired by the work in [Donmez et al., 2007], we propose a multi-strategy approach (DUAL) to switching from a DWDS to a DIV strategy. While Donmez et al. [2007] switch from a density-focused to an uncertainty-focused strategy, we use a diversity-focused approach. Uncertainty of a model has been used as a successful active learning strategy [Lewis and Catlett, 1994]. For the task of translation, we choose diversity as a strategy instead of 'uncertainty', as our experiments show that diversity is much faster to compute and the performance is very similar to the uncertainty sampling approach. Computing the uncertainty of a statistical translation model requires retraining of the model across iterations, which is time consuming. We also extend the DUAL approach and propose a novel ensemble approach called GraDUAL. While DUAL estimates a switch over point to transit to a second querying strategy, GraDUAL chooses an operating range in which it performs a gradual switch over. In the switch over range, we perform a dynamically weighted interpolation for sampling under the two approaches in consideration. This ensures a smooth transition from one strategy to the other and is robust to noise that may false project one query strategy to be better than the other.

For SMT, application of active learning has been focused on the task of selecting the most informative sentences to train the model, in order to reduce the cost of data acquisition. Recent work in this area discussed multiple query selection strategies for a Statistical Phrase Based Translation system [Haffari et al., 2009]. Their framework requires the source text to be translated by the system and the translated data is used in a self-training setting to train MT models. Gangadharaiah et al. [2009] use a pool-based strategy that maximizes a measure of expected future improvement, to sample instances from a large parallel corpus. Their goal is to select the most informative sentence pairs to build an MT system, and hence they assume the existence of target-side translations along with the source-side sentences. We, however, are interested in selecting the most informative sentences to reduce the effort and cost involved in translation.

Ensemble approaches have been proposed in active learning literature and have been successfully applied to classification tasks [Melville and Mooney, 2004, Freund et al., 1997]. Trading off between density and uncertainty has been the focus of several of these active learning strategies [McCallum and Nigam, 1998, Nguyen and Smeulders, 2004]. Baram et al. [2004] propose an online algorithm to select among multiple strategies and decide the strategy to be used for each iteration. Most notably, our approach is inspired from the DUAL approach proposed in Donmez et al. [2007], where the authors differ from earlier ensemble approaches by not focusing on selecting the best strategy for the entire task, but switch between multiple strategies over different ranges. Ensemble methods for active learning in MT have not been explored to our knowledge. Haffari et al. [2009] address an interesting technique of combining multiple query strategies for the task of sentence

selection. Tuning the weights of the combination and optimizing towards translation quality is computationally expensive, and their approach does not perform better than the best performing single strategy approach. This work we discuss in the thesis has also been published [Ambati et al., 2011b].

3.6.1 Active Learning Setup

We use our general framework for active learning in SMT for sentence selection, as discussed above. We start with an unlabeled dataset $U_0 = \{f_j\}$ and a seed labeled dataset $L_0 = \{(f_j, e_j)\}$, where labels are the translations. We then, score all the sentences in the U_0 according to our selection strategy and retrieve the best scoring sentence or a small batch of sentences. This sentence is translated and the sentence pair is added to the labeled set L_0. However, re-training and re-tuning an SMT system after translating every single sentence is computationally inefficient and may not have a significant effect on the underlying models. We, therefore continue to select a batch of N sentences before retraining the system on newly created labeled set $L_{k=1}$. Our framework for active learning in SMT is discussed in Algorithm 7.

3.6.2 DUAL Strategy

Let us consider the DWDS approach in more detail. It has two components for scoring a sentence S; a density component $d(s)$ and a diversity component $u(s)$ as mentioned in the previous section. The DWDS approach favors those sentences that contain dense n-grams and thus, has the largest contribution to the improvement of translation quality. Combining diversity with density of the underlying data is a well known ensemble technique in active learning that improves performance [Nguyen and Smeulders, 2004]. Now consider DIV selection criteria that favors sentences with unseen n-grams. Such a method is prone to selecting uncommon sentences that add very little information to the translation model.

Figure 3.6 displays the translation performance of 'DWDS' and 'DIV' on a held-out dataset as measured in BLEU vs. size of labeled training data in words. One observation is that DWDS, after rapid initial gains exhibits very slow incremental improvements. Diversity sampling shows continuous and consistent improvements over a longer operating range. We computed the overlap of instances selected by the two methods and found that there is a very low overlap, showing that there is significant disagreement in sentence selection by the two approaches.

In the initial phases of evolution of an MT system, there is very little or no labeled data, hence every sentence is highly diverse. DWDS can pick high density sentences which may have been scored lower by the DIV technique. As more data is labeled, explicitly dense sentences may not be found anymore. Therefore, DWDS may score sentences with

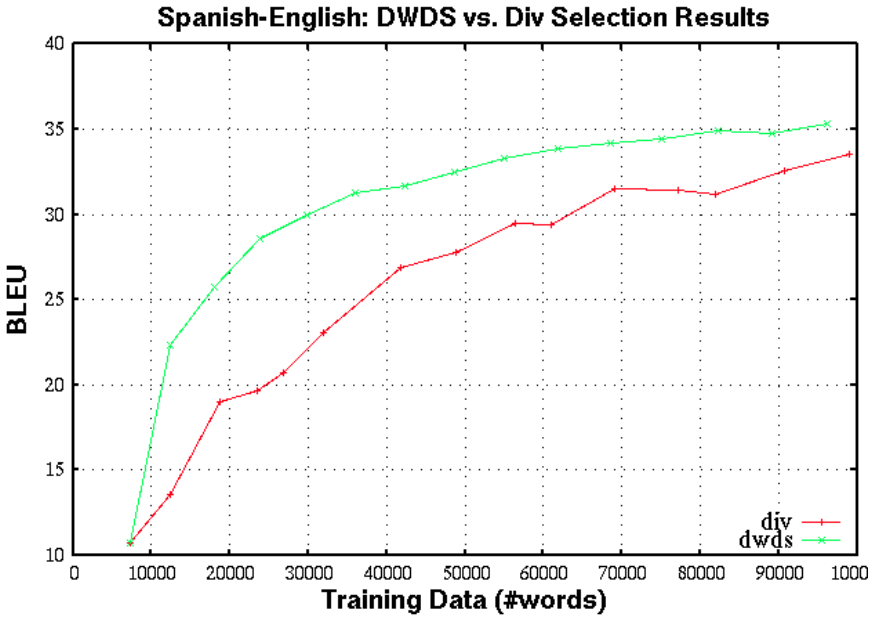

Figure 3.6: Performance curves of Density weighted diversity strategy and pure Diversity strategy. MT system performance measured by BLEU on y-axis and # source-language words translated to create the parallel corpus on x-axis. Notice the accelerated performance of diversity in the later parts of the curve in comparison to the density approach

moderate density higher than the sentences with high diversity, thereby making this criterion suboptimal. It is this weakness that we would like to address using the DUAL approach.

DUAL approach has been applied successfully for text classification problems in Donmez et al. [2007]. We adapt this approach to the task of MT. DUAL approach performs sentence selection using DWDS until a certain switching point is reached. A switching point is that point in the learning process, beyond which DWDS approach tends to provide only slow improvements. In other words, at a switching point we observe the density component of DWDS dominating the diversity component. Beyond the switching point, we use DIV active learning strategy for sentence selection. Algorithm 3 provides details of the DUAL approach.

Algorithm 2 ITERATION

1: Given Unlabeled Data Set: U_k
2: Given Labeled Data Set : L_k
3: **for** $i = 0$ to N **do**
4: s_i = Query(U_i,L_i)
5: t_i = Human Translation for s_i
6: $S_k = S_k \cup (s_i, t_i)$
7: **end for**
8: $U_{k+1} = U_k$ - S_k
9: $L_{k+1} = L_k \cup S_k$
10: Re-train MT system on L_{k+1}

Switching Point

Let us first consider an ideal scenario for switching where we have access to the learning curves from DWDS and DIV, like the ones shown in, Figure 3.6. Looking at the curves , one's natural choice for a switching point is where the slope of DWDS learning curve drops lower than the DIV learning curve. As our experiments later show, this switching point does in fact perform well in terms of translation quality.

The problem with this above approach is that it assumes availability of both the learning curves that have been produced independently. The active learning curve here is over the number of translations on x-axis and the direct improvement in translation quality on y-axis as measured by BLEU metric for MT evaluation. In order to compute such a curve, we need to select a batch of sentences using a querying strategy, translate the batch (or a subset), retrain and retest on a held-out dataset to observe the gradient of improvement across iterations. This is not feasible as we will be spending twice the amount of cost and also retrain the MT system twice. Although computation is not an issue, doubling the cost is unacceptable. Hence, we would like to identify the switching point by an approximation of the translation improvement, which is easy to compute.

Algorithm 3 DUAL APPROACH

Given Unlabeled Data Set: U_0
Given Labeled Data Set : L_0
$k = 0$
$SWITCH = false$
while $SWITCH = false$ **do**
    Query = DWDS
    ITERATION(U_i,L_i)
    β = Compute TTR(U_k,L_k)
    if $\beta > \delta$ **then**
        $k = k + 1$
        SWITCH = $true$
    end if
end while
for $k = k$ to T **do**
    Query = DIVERSITY
    ITERATION(U_i,L_i)
end for

We propose a surrogate metric based on types and token ratios that are computed only using source sentences of the labeled data. Type vs. token curves indicate the growth of vocabulary of the corpus. We use such curves to understand the effects of 'Density' and 'Diversity' in active learning based sentence selection. Density based approaches place an emphasis on the distribution of the data , and therefore provide a larger coverage for tokens. At the same time, the diversity focused component ensures aggregation of new types.

We propose a metric called 'Type-Token Ratio'(TTR) that highlights the balance between the tokens and types of the unlabeled data, and use it in an active learning querying method as formulated below.

$$
\begin{aligned}
Typ_k(L_k, U_0) &= \frac{\sum_x^{Phrases(U_0)} \alpha}{\|Phrases(U_0)\|} \\
\alpha &= \begin{cases} 1 & x \in Phrases(L_k) \\ 0 & \end{cases} \\
Tok_k(L_k, U_0) &= \frac{\|Phrases(L_k) \cap Phrases(U_0)\|}{\|Phrases(U_0)\|} \\
TTR_k(L_k, U_0) &= \frac{2 * Typ_k * Tok_k}{(Typ_k + Tok_k)}
\end{aligned}
$$

It is inexpensive to compute TTR curves for both the DWDS and DIV query methods.

The switching point is chosen where the slope of DWDS curve is lower than the DIV curve by a margin, shown as a constraint below. We set δ to be a very small number, 0.02 in our experiments.

$$\Delta(DWDS_k) > \Delta(DIV_k) + \delta \tag{3.10}$$

3.6.3 GraDUAL Approach

Estimation of the switching point is the key to the success of DUAL approach. Switching too early may take away the benefits of DWDS approach, and switching too late may not yield the benefits of DIV sampling approach. In order to test the effect of the choice of the switching point and on the MT system, in the later part of the section we also conduct an experiment where we switch between strategies at multiple points along the active learning curve with varied benefits. (refer: Figure 3.7).

This may not be robust to cases where noise in the training or data causes a temporary dip in the slope of the TTR curve for DWDS. Noise can cause a false switching from one strategy to another, even when it is not the right sampling strategy to be exploited. Given the multiple factors and parameters in training an MT system, it is natural to expect such unstable behavior in the initial phases of the system. We, therefore, propose a different hybrid strategy called 'GraDUAL', which gradually switches from DWDS to DIV strategies. We do not assume the existence of a 'switching point', but try to estimate a 'switching range' during which the transition between strategies takes place.

GraDUAL approach, as described in Algorithm 4, is motivated from the concept of 'exploration vs. exploitation'. This approach exploits the sampling strategy that is evidently better in a given range. We compute the slope of the TTR curve between two consecutive iterations as Δ. A positive and increasing slope indicates good performance of the approach. When comparing two different TTR curves, we will have operating ranges where the slopes do not project a clear winner. In such cases, GraDUAL approach suggests sampling from both strategies, with a gradual shift towards the second technique. The rate of the shift is controlled by the parameter $f(\beta)$. In our current work, we use a constant $f(\beta) = 0.8$ to sample 80% from the best performing strategy and 20% from the second. We will experiment with other functions for $f(\beta)$.

$$\begin{aligned}
\beta &= Abs(\Delta(DWDS) - \Delta(DIV)) \\
\alpha &= \begin{cases} 1 & \beta > \delta \\ 0 & \beta < \delta \\ f(\beta) & \end{cases} \\
Score(s) &= \alpha DWDS(s) + (1-\alpha)DIV(s)
\end{aligned}$$

Algorithm 4 GRADUAL APPROACH

1: Given Labeled Data Set : L_0
2: Given Unlabeled Data Set: U_0
3: $\beta = 1$
4: **for** $k = 0$ to T **do**
5: Query method = GraDUAL
6: β = Compute Ratio(U_k,L_k)
7: ITERATION(U_i,L_i,β)
8: **end for**

3.6.4 Experiments

Setup

We perform our experiments on the Spanish-English language-pair in order to simulate a resource-poor language pair. We have parallel corpora and evaluation data sets for the Spanish-English language pair allowing us to run multiple experiments efficiently. We use BTEC parallel corpus [Takezawa et al., 2002] from the IWSLT tasks with 127K sentence pairs. We use the standard Moses pipeline [Koehn et al., 2007] for extracting, training and tuning our system. We built an SRILM language model using English side of the Europarl corpus, which consists of 1.6M sentences or 300M words. While experimenting with data sets of varying size, we do not vary the language model. The weights of the different translation features are tuned using standard MERT [Och, 2003]. Our development set consists of 506 sentences and test set consists of 343 sentences. We report results on the test set.

Results: Hybrid AL approaches

We evaluate our multi-strategy approaches and present results. We first compare the robustness of our surrogate metric based switching strategy with 'manual switching'. Figure 3.7 shows results on the development set when switching using BLEU score based learning curves. A human would then visually inspect and select an iteration to switch where the DWDS learning curve's slope is lower than that of the DIV learning curve by a margin. We compare this with results from executing the DUAL approach using our surrogate metric, TTR, to decide the switching point. We observe that switching using feedback from TTR works on par with BLEU, and is also easier to compute. We, therefore, report experiments with multi-strategy approaches using the TTR surrogate.

In Figure 3.8 we compare DUAL and GraDUAL approaches to our best performing active learning strategy DWDS and also DIV. When considered independently, AL approaches

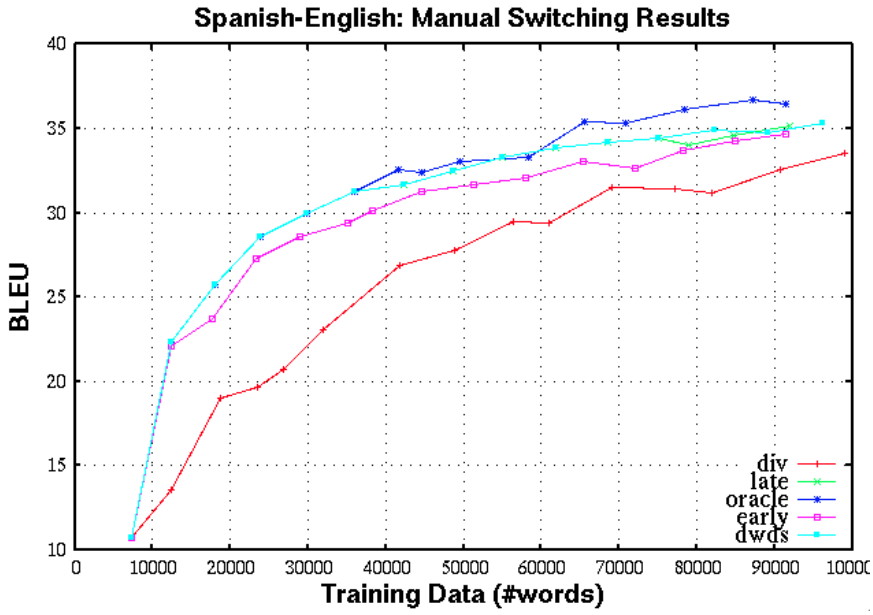

Figure 3.7: Performance of Spanish-English MT system with three different 'manual' switching from DWDS to diversity technique. Switching too early (early) or much later in the curve (late) have adverse affect on the overall performance, but switching by observing performance on dev-set (oracle) has added benefit

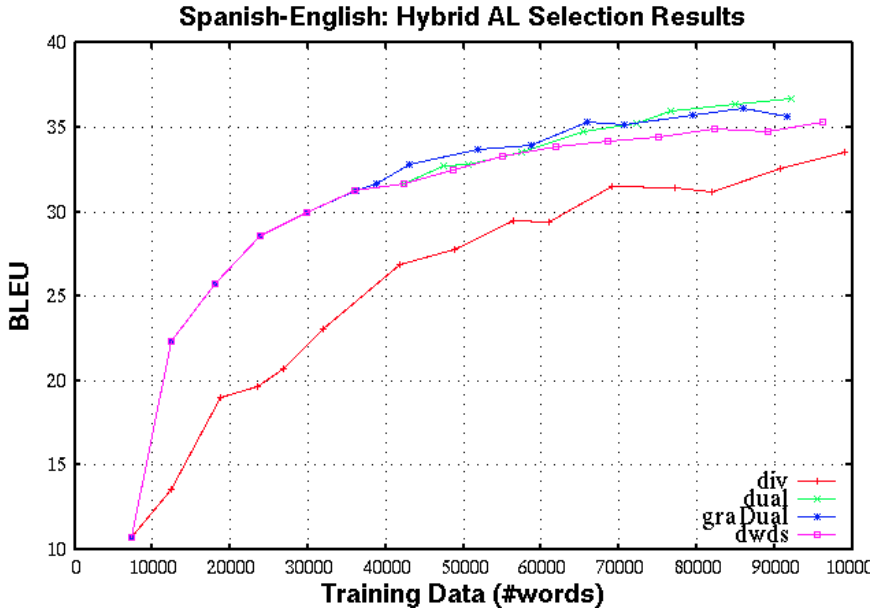

Figure 3.8: Performance of Spanish-English MT system using our ensemble switching techniques: DUAL and graDUAL. Both approaches switch using TTR curves and have additional benefit over either DWDS or DIV only

have a disadvantage that they are hindered by the selection of data made in the earlier iterations. The point of switching strategies is that the second strategy can build on top of better selections made by its predecessor. The results show a similar trend. We observe that both our multi-strategy approaches that include DIV switching strategy perform significantly better than the two baseline approaches even when DIV does not do better than DWDS in isolation. From the results, although GraDUAL and DUAL perform comparably, GraDUAL displays a smoother transition from one strategy to the other. Overall, using multi-strategy ensemble approaches, we have shown that MT systems can reach better performance while requiring much lower amounts of data. At different points on the curves, prior to convergence, we have performed bootstrapped sampling-based significance tests with the baseline and see that the p-value varies between 0.02 and 0.11 (averaging at 0.06). So, the reported results are statistically indicative and with a larger experiment (with more observations) should prove statistically significant.

3.7 Summary

In recent years, corpus based approaches to machine translation have become predominant. Success of these approaches depends on the availability of parallel corpora. In this chapter, we proposed active learning as a cost-effective paradigm for building low-resource language translation systems. To summarize, in this chapter, we have made three major contributions:

- We proposed sentence selection strategies for Statistical Machine Translation that performs significantly better than state-of-the-art baselines and a strong random baseline. Active learning aims at reducing cost of label acquisition by selecting and prioritizing the most informative sentences for translation.
- We have studied the various configurations of the active learning framework for the sentence selection problem and provide empirical observations for the same for three different language-pairs: Spanish-English, Japanese-English and Urdu-English.
- We also proposed two hybrid approaches for sentence selection, one: a modified version of the DUAL [Donmez et al., 2007] approach and two: a novel and robust GraDUAL approach. We experimented our approaches on Spanish-English language pair and have shown significant improvements.

Chapter 4

Active Learning for Word Alignment

The success of statistical approaches to MT can be attributed to the IBM models [Brown et al., 1993] that characterize *word-level* alignments in parallel corpora. Parameters of these alignment models are learnt in an unsupervised manner using the EM algorithm over *sentence-level* aligned parallel corpora.

While the ease of automatically aligning sentences at the word-level with tools like GIZA++ [Och and Ney, 2003] has enabled fast development of SMT systems for various language pairs, the quality of alignment is often quite low, especially for language pairs like Chinese-English, Arabic-English that diverge from the independence assumptions made by the generative models. Increased parallel data enables better estimation of the model parameters, but a large number of language pairs still lack such resources.

4.1 Introduction

Building translation models involves learning in structured spaces. For example the phrase table in SMT is built using the word-alignment models for parallel corpora. Word-Alignment is a particularly challenging problem and has been addressed in a completely unsupervised manner thus far [Brown et al., 1993]. Availability of tools like GIZA++ [Och and Ney, 2003] that implement these generative frameworks, have also made this research quite attractive. While such generative models have been successful, local optimum is a well known problem due to the large output space of word-alignment. Researchers have begun to explore models that use both labeled and unlabeled data to build word-alignment models for MT [Fraser and Marcu, 2006]. They first pose the problem of alignment as a search problem in log-linear space with features coming from the IBM alignment models. The log-linear model is trained on available labeled data to improve performance. They propose a semi-supervised training algorithm which alternates between the discriminative error

training on the labeled data to learn the weighting parameters and the maximum likelihood EM training on unlabeled data to estimate the parameters. [Callison-Burch et al., 2004] also improve alignment by interpolating human alignments with automatic alignments. They observe that while working with such data sets, alignments of higher quality should be given a much higher weight than the lower quality alignments. [Wu et al., 2006] learn separate models from labeled and unlabeled data using the standard EM algorithm. The two models are then interpolated to use as a learner in the semi-supervised AdaBoost algorithm to improve word alignment.

Two directions of research have been pursued for improving generative word alignment. The first is to relax or update the independence assumptions based on more information, usually syntactic, from the language pairs [Cherry and Lin, 2006, Fraser and Marcu, 2007b]. The second is to use extra annotation, typically *word-level* human alignment for some sentence pairs, in conjunction with the parallel data to learn alignment in a semi-supervised manner. Our research is in the direction of the latter, and aims to reduce the effort involved in hand-generation of word alignments by using active learning strategies for careful selection of word pairs to seek alignment.

4.1.1 IBM models

IBM models provide a generative framework for performing word alignment of parallel corpus. Given two strings from source and target languages $s_1^J = s_1, \cdots, s_j, \cdots s_J$ and $t_1^I = t_1, \cdots, t_i, \cdots t_I$, an alignment $\mathcal{A}$ is defined as a subset of the Cartesian product of the word indices as shown in Eq 4.1. In IBM models, since alignment is treated as a function, all the source positions must be covered exactly once Brown et al. [1993].

$$\mathcal{A} \subseteq \{(j, i) : j = 0 \cdots J; i = 0 \cdots I\} \tag{4.1}$$

For the task of translation, we would ideally want to model $P(s_1^I|t_1^J)$, which is the probability of observing source sentence s_1^I given target sentence t_1^J. This requires a lot of parallel corpus for estimation and so it is then factored over the word alignment A for the sentence pair, which is a hidden variable. Word alignment is therefore a by-product in the process of modeling translation. We can also represent the same under some parameterization of θ, which is the model we are interested to estimate.

$$P(s_1^J|t_1^I) = \sum_{a_1^J} Pr(s_1^J, A|t_1^J) \tag{4.2}$$

$$= \sum_A p_\theta(s_1^J, A|t_1^I) \tag{4.3}$$

Given a parallel corpus U of sentence pairs $\{(s_k, t_k) : k = 1, \cdots, K\}$ the parameters can be estimated by maximizing the conditional likelihood over the data. IBM models Brown et al. [1993] from 1 to 5 are different ways of factoring the probability model to estimate

the parameter set θ. For example in the simplest of the models, IBM model 1, only the lexical translation probability is considered treating each word being translated independent of the other words.

$$\hat{\theta} = \arg\max_{\theta} \prod_{k=1}^{K} \sum_{A} p_{\theta}(s_k, A|t_k) \tag{4.4}$$

The parameters of the model above are estimated as $\hat{\theta}$, using the EM algorithm. We can also extract the *Viterbi alignment* ,$\hat{A}$, for all the sentence pairs, which is the alignment with the highest probability under the current model parameters θ:

$$\hat{A} = \arg\max_{A} p_{\hat{\theta}}(s_1^J, A|t_1^I) \tag{4.5}$$

The alignment models are asymmetric and differ with the choice of translation direction. We can therefore perform the above after switching the direction of the language pair and obtain models and Viterbi alignments for the corpus as represented below:

$$\hat{\theta} = \arg\max_{\theta} \prod_{k=1}^{K} \sum_{a} p_{\theta}(t_k, a|s_k) \tag{4.6}$$

$$\hat{A} = \arg\max_{A} p_{\hat{\theta}}(t_1^I, A|s_1^J) \tag{4.7}$$

Given the Viterbi alignment for each sentence pair in the parallel corpus, we can also compute the word-level alignment probabilities using simple relative likelihood estimation for both the directions. The alignments and the computed lexicons form an important part of our link selection strategies.

$$P(s_j/t_i) = \frac{\sum_s count(t_i, s_j; \hat{A})}{\sum_s count(t_i)} \tag{4.8}$$

$$P(t_i/s_j) = \frac{\sum_s count(t_i, s_j; \hat{A})}{\sum_s count(s_j)} \tag{4.9}$$

We perform all our experiments on a symmetrized alignment that combines the bidirectional alignments using heuristics as discussed in Koehn et al. [2007]. We represent this alignment as $A = \{a_{ij} : i = 0 \cdots J \in s_1^J; j = 0 \cdots I \in t_1^I\}$.

4.1.2 Semi-Supervised Word Alignment

We use an extended version of MGIZA++ [Gao and Vogel, 2008] to perform the constrained semi-supervised word alignment. Manual alignments are incorporated in the EM training

phase of these models as constraints that restrict the summation over all possible alignment paths. Typically in the EM procedure for IBM models, the training procedure requires for each source sentence position, the summation over counts from all words in the target sentence. The manual alignments allow for one-to-many alignments and many-to-many alignments in both directions. For each position i in the source sentence, there can be more than one manually aligned target word. The restricted training will allow only those paths, which are consistent with the manual alignments. Therefore, the restriction of the alignment paths reduces to restricting the summation in EM.

4.2 Active Learning Setup

We discuss our active learning setup for word alignment in Algorithm 5. We start with an unlabeled dataset $U = \{(S_k, T_k)\}$, indexed by k, and a seed pool of partial alignment links $A_0 = \{a_{ij}^k, \forall s_i \in S_k, t_j \in T_k\}$. This is usually an empty set at iteration $t = 0$. We iterate for T iterations. We take a pool-based active learning strategy, where we have access to all the automatically aligned links and we can score the links based on our active learning query strategy. The query strategy uses the automatically trained alignment model M_t from the current iteration t for scoring the links. Re-training and re-tuning an SMT system for each link at a time is computationally infeasible. We therefore perform batch learning by selecting a set of N links scored high by our query strategy. We seek manual corrections for the selected links and add the alignment data to the current labeled data set. The word-level aligned labeled data is provided to our semi-supervised word alignment algorithm for training an alignment model M_{t+1} over U.

Algorithm 5 AL FOR WORD ALIGNMENT

1: Unlabeled Data Set: $U = \{(S_k, T_k)\}$
2: Manual Alignment Set : $A_0 = \{a_{ij}^k, \forall s_i \in S_k, t_j \in T_k\}$
3: Train Semi-supervised Word Alignment using (U, A_0) $\rightarrow M_0$
4: N: batch size
5: **for** $t = 0$ to T **do**
6: L_t = LinkSelection(U,A_t,M_t,N)
7: Request Human Alignment for L_t
8: A_{t+1} = A_t + L_t
9: Re-train Semi-Supervised Word Alignment on (U, A_{t+1}) $\rightarrow M_{t+1}$
10: **end for**

We can iteratively perform the algorithm for a defined number of iterations T or until a certain desired performance is reached, which is measured by alignment error rate (AER) [Fraser and Marcu, 2007a] in the case of word alignment. In a more typical scenario, since

reducing human effort or cost of elicitation is the objective, we iterate until the available budget is exhausted.

4.2.1 Query Selection Strategies

We propose multiple query selection strategies for our active learning setup. The scoring criteria is designed to select alignment links across sentence pairs that are highly uncertain under current automatic translation models. These links are difficult to align correctly by automatic alignment and will cause incorrect phrase pairs to be extracted in the translation model, in turn hurting the translation quality of the SMT system. Manual correction of such links produces the maximal benefit to the model. We would ideally like to elicit the least number of manual corrections possible in order to reduce the cost of data acquisition. In this section we discuss our link selection strategies based on the standard active learning paradigm of 'uncertainty sampling'[Lewis and Catlett, 1994]. We use the automatically trained translation model θ_t for scoring each link for uncertainty, which consists of bidirectional translation lexicon tables computed from the bidirectional alignments.

Uncertainty Sampling: Bidirectional Alignment Scores

The automatic Viterbi alignment produced by the alignment models is used to obtain translation lexicons. These lexicons capture the conditional distributions of source-given-target $P(s|t)$ and target-given-source $P(t|s)$ probabilities at the word level where $s_i \in S$ and $t_j \in T$. We define certainty of a link as the harmonic mean of the bidirectional probabilities. The selection strategy selects the least scoring links according to the formula below which corresponds to links with maximum uncertainty:

$$Score(a_{ij}|s_1^I, t1^J) = \frac{2 * P(t_j|s_i) * P(s_i|t_j)}{P(t_j|s_i) + P(s_i|t_j)} \tag{4.10}$$

Confidence Sampling: Posterior Alignment probabilities

Confidence estimation for MT output is an interesting area with meaningful initial exploration [Blatz et al., 2004, Ueffing and Ney, 2007]. Given a sentence pair (s_1^I, t_1^J) and its word alignment, we compute two confidence metrics at alignment link level – based on the posterior link probability as seen in Equation 4.10. We select the alignment links that the initial word aligner is least confident according to our metric and seek manual correction of the links. We use $t2s$ to denote computation using higher order (IBM4) target-given-source models and $s2t$ to denote source-given-target models. Targeting some of the uncertain parts of word alignment has already been shown to improve translation quality in SMT

[Huang, 2009]. We use confidence metrics as an active learning sampling strategy to obtain most informative links. We also experimented with other confidence metrics as discussed in [Ueffing and Ney, 2007], especially the IBM 1 model score metric, but it did not show significant improvement in this task.

$$P_{t2s}(a_{ij}, t_1^J|s_1^I) = \frac{p_{t2s}(t_j|s_i, a_{ij} \in A)}{\sum_i^M p_{t2s}(t_j|s_i)} \tag{4.11}$$

$$P_{s2t}(a_{ij}, s_1^I|t_1^J) = \frac{p_{s2t}(s_i|t_j, a_{ij} \in A)}{\sum_i^N p_{s2t}(s_i|t_j)} \tag{4.12}$$

$$Conf(a_{ij}|S,T) = \frac{2 * P_{t2s} * P_{s2t}}{P_{t2s} + P_{s2t}} \tag{4.13}$$

Query by Committee

The generative alignments produced differ based on the choice of direction of the language pair. We use A_{s2t} to denote alignment in the source to target direction and A_{t2s} to denote the target to source direction. We consider these alignments to be two experts that have two different views of the alignment process. We formulate our query strategy to select links where the agreement differs across these two alignments. In general query by committee is a standard sampling strategy in active learning[Freund et al., 1997], where the committee consists of any number of experts, in this case alignments, with varying opinions. We formulate a query by committee sampling strategy for word alignment as shown in Equation 4.14. In order to break ties, we extend this approach to select the link with higher average frequency of occurrence of words involved in the link.

$$Score(a_{ij}) = \alpha \tag{4.14}$$

$$where \quad \alpha = \begin{cases} 2 & a_{ij} \in A_{s2t} \cap A_{t2s} \\ 1 & a_{ij} \in A_{s2t} \cup A_{t2s} \\ 0 & otherwise \end{cases}$$

Margin Sampling

The strategy for confidence based sampling only considers information about the best scoring link from Eq 4.13. However we could benefit from information about the second best scoring link as well. Earlier work has shown success in multi-class classification problems using such a 'margin based' approach, where the difference between the probabilities assigned by the underlying model to the first best and second best labels is used as a sampling criteria [Scheffer et al., 2001]. We adapt such a margin-based approach to link-selection using the $Conf1$ scoring function discussed in the earlier sub-section. Our $margin$ technique

Language	Sentences	Words	
		Src	Tgt
Chinese-English	21,863	424,683	524,882
Arabic-English	29,876	630,101	821,938

Table 4.1: Corpus Statistics for Chinese-English and Arabic-English parallel data released by LDC. These datasets also have complete manual alignment information available.

is formulated below, where $\hat{a}_{ij_1}$ and $\hat{a}_{ij_2}$ are potential first best and second best scoring alignment links for a word at position i in the source sentence S with translation T. The word with minimum margin value is chosen for human alignment. Intuitively such a word is a possible candidate for mis-alignment due to the inherent confusion in its target translation.

$$Margin(i) = \quad Conf(\hat{a}_{ij_1}|S,T) - Conf(\hat{a}_{ij_2}|S,T)$$

4.3 Experiments

4.3.1 Data Setup

To run our active learning and semi-supervised word alignment experiments iteratively, we simulate the setup by using a parallel corpus for which the gold standard human alignment is already available. We experiment with two language pairs - Chinese-English and Arabic-English. Corpus-level statistics for both language pairs can be seen in Table 4.1 and their alignment link level statistics can be seen in Table 4.2. Both datasets were released by LDC as part of the GALE project.

Chinese-English dataset consists of 21,863 sentence pairs with complete manual alignment. The human alignment for this dataset is much denser than the automatic word alignment. On an average each source word is linked to more than one target word. Similarly, the Arabic-English dataset consisting of 29,876 sentence pairs also has a denser manual alignment. Automatic word alignment in both cases was computed as a symmetrized version of the bidirectional alignments obtained from using GIZA++ Och and Ney [2003] in each direction separately.

From Table 4.2 it is clear that the manually aligned data has more links for the Chinese-English language pair and a comparable number of links for the Arabic-English case when compared with the automatic symmetrized alignments.

Alignment	Automatic Links	Manual Links
Chinese-English	491,887	588,075
Arabic-English	786,223	712,583

Table 4.2: Alignment link statistics for the data released by LDC for Chinese-English and Arabic-English. Note a higher density of links from manual alignment for Chinese-English

4.3.2 Results

For word alignment we performed experiments to show that active learning can help select the most informative alignment links that have high uncertainty according to a given automatically trained model. We then show that fixing such alignments leads to reduction of error in word alignment, as measured by AER [Fraser and Marcu, 2007a]. The hypothesis is that, such careful selection of links can ensure least human effort involved in correcting them as well. We compare this with a baseline where links are selected at random for manual correction.

We first perform an unsupervised word alignment of the parallel corpus. We then use the learned model in running our link selection algorithm over the entire alignments to determine the most uncertain links according to each active learning strategy. The links are then looked up in the gold standard human alignment database and corrected. In scenarios where an alignment link is not present in the gold standard data for the source word, we introduce a NULL alignment constraint, else we select all the links as given in the gold standard. The aim of our work is to show that active learning can help in selecting informative alignment links, which if manually labeled can reduce the overall alignment error rate of the given corpus. We, therefore measure the reduction of alignment error rate (AER) of a semi-supervised word aligner that uses this extra information to align the corpus. AER requires a gold standard manually annotated set of ”Sure” links and ”Possible” links which are used to compute recall and precision respectively. In our case we use the manually aligned data as sure links.

We plot performance curves for both Chinese-English, Figure 4.1 and Arabic-English, Figure 4.2, with number of manual links elicited on x-axis and AER on y-axis. In each iteration of the experiment, we gradually increase the number of links selected from gold standard and make them available to the semi-supervised word aligner and measure the overall reduction of AER on the corpus. We compare our link selection strategies to a baseline approach, where links are selected at random for manual correction.

All our approaches perform comparably or better than the baseline for both language pairs. Query by committee (qbc) performs similar to the baseline in Chinese-English and only slightly better for Arabic-English. This could be due to our committee consisting of two alignments that differ only in direction and so are not sufficient in deciding for uncertainty.

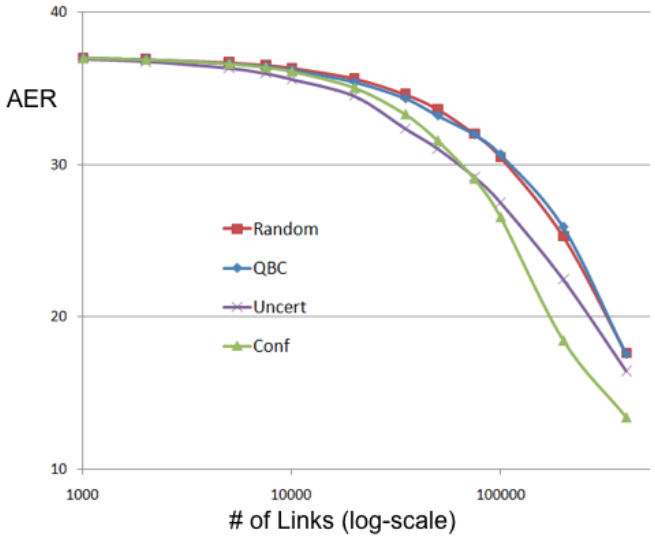

Figure 4.1: Performance of Link Selection Algorithms for Chinese-English. Effort is computed as #links aligned on x-axis and AER of the resulting semi-supervised word on y-axis. Actively selecting links for human alignment outperforms random selection baseline

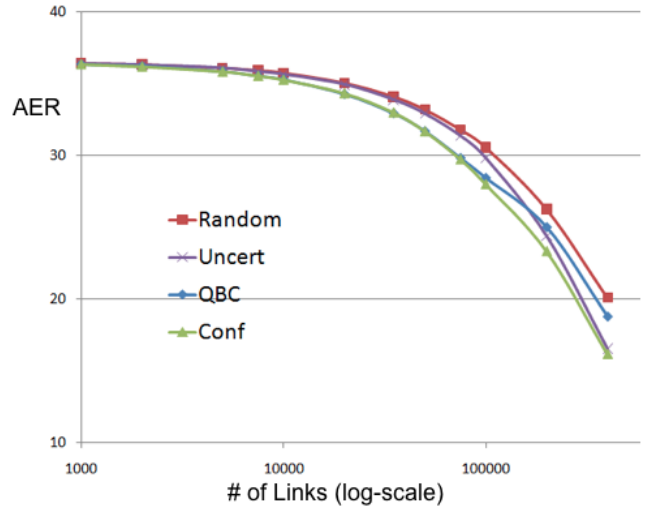

Figure 4.2: Performance of Link Selection Algorithms for Arabic-English. Effort is computed as #links aligned on x-axis and AER of the resulting semi-supervised word on y-axis. Actively selecting links for human alignment outperforms random selection baseline

Confidence based (conf) and uncertainty based (uncert) methods perform significantly better than the baseline in both language pairs.

We observe that confidence based metrics perform significantly better than the baseline. We can interpret the improvement in two ways. For the same number of manual alignments elicited, our selection strategies select links that provide higher reduction of error when compared to the baseline. An alternative interpretation is that assuming a uniform cost per link, our best selection strategy achieves similar performance to the baseline, at a much lower cost of elicitation. From the scatter plots in Figure 4.1 we can say that using our best selection strategy one achieves similar performance to the baseline, but at a much lower cost of elicitation assuming cost per link is uniform.

4.3.3 Batch Selection vs Decay Approach

Re-training the word alignment models after eliciting every individual alignment link is infeasible. In our data set of 21,863 sentences with 588,075 links, it would be computationally intensive to re-train after eliciting even 100 links in a batch. We therefore sample links as a discrete batch, and train alignment models to report performance at fixed points. Such a batch selection is only going to be sub-optimal as the underlying model changes with every alignment link and therefore becomes 'stale' for future selections. We observe that in some scenarios while fixing one alignment link could potentially fix all the mis-alignments in a sentence pair, our batch selection mechanism still samples from the rest of the links in the sentence pair. We experimented with an exponential decay function over the number of links previously selected, in order to discourage repeated sampling from the same sentence pair. We performed an experiment by selecting one of our best performing selection strategies ($conf$) and ran it in both configurations - one with the decay parameter ($batchdecay$) and one without it ($batch$). As seen in Figure 4.3, the decay function has an effect in the initial part of the curve where sampling is sparse but the effect gradually fades away as we observe more samples. In the reported results we do not use batch decay, but an optimal estimation of 'staleness' could lead to better gains in batch link selection using active learning.

4.3.4 Translation Results

We also perform end-to-end machine translation experiments to show that our improvement of alignment quality leads to an improvement of translation scores. For Chinese-English, we train a standard phrase-based SMT system Koehn et al. [2007] over the available 21,863 sentences. We tune on the MT-Eval 2004 dataset and test on a subset of MT-Eval 2005 dataset consisting of 631 sentences. The language model we use is built using only the English side of the parallel corpus. We understand that this language model is not the optimal choice, but we are interested in testing the word alignment accuracy, which

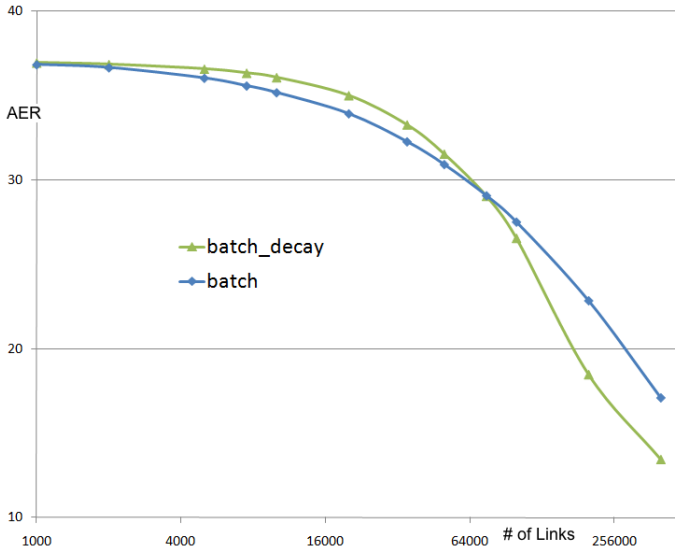

Figure 4.3: Introducing decay parameter to combat batch selection effect has improved performance benefits on top of our best performing conf sampling strategy

MT System built with/	BLEU	METEOR
Complete Automatic Alignment	18.82	42.70
Complete Human Alignment	19.96	44.22
Active Annotation of 20% links	19.34	43.25

Table 4.3: Chinese-English MT system trained on LDC data and performance on MT03 test sets. Selectively aligning only 20% of the links, achieves 40% of the possible gain obtainable from using complete human alignment. In this research we do not explore why complete human alignment only yields 1 BLEU point in translation quality

primarily affects the translation model. We first obtain the baseline score by training in an unsupervised manner, where no manual alignment is used. We also train a configuration, where we substitute the final word alignment with gold standard manual alignment for the entire parallel corpus. This is an upper bound on the translation accuracy that can be achieved by any alignment link selection algorithm for this dataset. We now take our best link selection criteria, which is the confidence based method and re-train the MT system after eliciting manual information for only 20% of the alignment links. We observe that at this point we have reduced the AER from 37.09 to 26.57. The translation accuracy reported in Table 4.3, as measured by BLEU Papineni et al. [2002] and METEOR Lavie and Agarwal [2007], also shows significant improvement and approaches the quality achieved using gold standard data.

Type	# Total Entries	# Unique Source	Avg. Fanout
Automatic Alignment	151,744	29,029	5.2
Human Alignment	173,573	29,034	5.9

Table 4.4: Some statistics over the lexicons extracted from automatic alignment vs. manual alignment for Chinese-English MT System. Human alignment is dense, yielding larger lexicons

Type	# Total Phrase	# Unique Source	Avg. Fanout
Automatic Alignment	103,896	21,783	4.7
Human Alignment	85,402	25,568	3.3

Table 4.5: Some statistics over phrase tables trained using automatic alignment vs. manual alignment for Chinese-English MT System. Human alignment yields smaller lexicons with less ambiguity

Perhaps as an artifact of the instruction set given to the annotators, the manual alignment links are denser than the automatic alignment. Therefore, it is not surprising that the resulting lexicon which is an accumulation of the word-to-word links is larger for the human alignment case when compared to the automatic case. And as a resulting artifact the phrase table is larger in the case of automatic alignments due to the phrase extraction heuristics Koehn et al. [2003] which prefer recall over precision. Table 4.4 shows statistics of the final lexicons after training a moses system from both cases of complete automatic alignment and human alignment. Similarly, statistics about phrase tables can be seen in Table 4.5.

4.4 Summary

Word Alignment is a particularly challenging problem and has been addressed in a completely unsupervised manner thus far Brown et al. [1993]. While generative alignment models have been successful, lack of sufficient data, model assumptions and local optimum during training are well known problems. Semi-supervised techniques use partial manual alignment data to address some of these issues in low-resource scenarios. We have shown that active learning strategies can reduce the effort involved in eliciting human alignment data for training semi-supervised word alignment. Our approaches show a significant reduction in effort due to careful selection of maximally uncertain links that provide the most benefit to the alignment model. Experiments on Chinese-English and Arabic-English have shown considerable reduction in AER at lower efforts. The work discussed in this section also appears in our recent publications Ambati et al. [2010b,c].

Chapter 5

Multi-Type Annotation Active Learning

Where as in chapter 3 we explored building general purpose MT systems by successfully constructing parallel corpora via active learning, in chapter 4 we showed that cost of annotating word alignment links can be reduced via active learning as well. However, in both cases we target a single type of annotation and optimize the cost involved in labeling data for a particular annotation. In this chapter, we argue that the traditional framework of active learning for eliciting a single kind of annotation needs to be extended to work with multiple types of annotation. We explore this in the context of two tasks in translation: comparable corpora classification and building domain specific translation systems.

5.1 Introduction

The traditional setup for active learning applied to a supervised learning algorithm, assumes the selection of unlabeled instances and elicitation of annotation of a single kind. Given the nature of supervised learning tasks of the current day, we can say that this is a strong assumption. When multiple tasks are being trained together in novel setups like joint-learning or multitask learning Caruana [1997], there is also a need for new active learning strategies that can elicit informative samples that improve the performance for these learning scenarios. We will discuss some of the related work in the following section.However, results in multitask learning have not been applied to the task of building translation systems. In fact, active learning for MT has not yet been explored in its full potential. Much of the literature has explored one task - optimizing the selection of monolingual sentences for translation elicitation [Ambati et al., 2010a, Haffari et al., 2009]. In our work, so far we have explored a second annotation, completely independent from the first task, in

the form of word alignment. However, an MT system has a variety of annotations that result in potential improvement of translation quality. In context of an SMT system, some types of annotations that are of value are - translation of individual words or phrases, named entity tagging, syntactic annotation of an existing or a new sentence, post-editing of system-generated translations, evaluation of system generated output among others.

In this chapter we first discuss two different problem setups in machine translation that involve more than one kind of annotation. We will then devise active learning strategies for each of these scenarios exploring the benefit of combining the different annotations.

- *Comparable Corpora Classification:* Given a set of sentence pairs as input, the task is to predict which of the pairs are translation equivalents of each other. We will use a supervised classification algorithm for this task, which requires labeled training data. We will also discuss two different kinds of annotations that can be collected to train the classifier - class labels vs. parallel translation segments.
- *Focused domain Translation:* Given a monolingual corpus with mixed domains in the source-language, the task is to build a domain specific translation system for a specific domain in the target-language. We address this by combing two different tasks - a text classification task for identifying focused source-language data and a translation task for building an MT system for the specific domain. In such a setup, the output of the text classification phase constraints the choice of the sentences for translating and training an MT system in the second phase. Therefore it is important that any active learning solution pays attention to both the tasks.

5.1.1 Background: MultiTask Learning

Many problems in information extraction, text mining, natural language processing and other fields involve setups that involve sub-problems that are trained separately from one another. For example, building a relation extraction system, may require part-of-speech, named entity identification before we can extract relationships. Recent work is moving away from training tasks individually to combining the tasks in ways that benefit each other either by sharing the labeled data space, feature space or sometime both. Such setups are increasingly becoming popular in NLP and are called joint learning or multitask learning [Caruana, 1997]. Multitask learning is a case of transfer learning [Pan and Yang, 2010] and can be seen as transfer of knowledge from one task to the other in order to bridge difference in assumptions across both the tasks and in the process reduce the effort or increase the performance of either or both the tasks. For instance recent work in parsing shows that named entity recognition task and parsing tasks can be combined to improve overall performance of a parser [Finkel and Manning, 2010].

There has been some recent interest in applying active learning to multitask learning.

[Reichart et al., 2008a] propose an extension of the single-sided active elicitation task to a multitask scenario, where data elicitation is performed for two or more independent tasks at the same time. Settles et al. [2008] and Vijayanarasimhan and Grauman [2008] propose elicitation of annotations for image segmentation under a multi-instance learning framework. Reichart et al. [2008b] proposes an extension of the single-sided active elicitation task to a multitask scenario, where data elicitation is performed for two or more independent tasks at the same time. Settles et al. [2008] propose elicitation of annotations for image segmentation under a multi-instance learning framework. Active learning with multiple annotations also has similarities to the recent body of work in learning from instance feedback and feature feedback Melville et al. [2005]. Druck et al. [2009] propose active learning extensions to the gradient approach of learning from feature and instance feedback. Attenberg et al. [2010] also present their approach, called active dual supervision: determining which feature or example a classifier is most likely to benefit from labeling next. Roth and Small [2008] discuss active learning for pipeline models, which uses error at various phases of the pipeline to combine local active learning strategies into one that minimizes the annotation requirements for the overall pipeline. They use this to build a named entity and relation extraction system that involves three separate phases.

Harpale and Yang [2010] propose an Active Learning framework for the Multitask Adaptive Filtering problem. They explore AL approaches to rapidly improve their system, based on Dirichlet Process priors, with minimal user/task-level feedback and benefit from learning across multiple tasks simultaneously due to the shared prior. [Zhang, 2010] propose an active learning framework exploiting relations where multiple tasks are related in the output spaces by constraints. They utilize not only the uncertainty of the prediction in a single task but also the inconsistency of predictions across tasks to propose active learning strategies for the multitask learning problem.

5.1.2 Multitask Learning vs. Multi-Type Annotation Learning

The comparable corpora classification task involves training a classifier with annotations at two levels and similarly, the focused domain translation task involves two different kinds of annotations to train the text classifier and translation system. While in traditional multitask learning the goal is to train two different tasks that share either the input space or the output space, in this scenario our goal is to train only a single classifier but with help from two different annotations, which do not have any overlap in the feature space. Also for multitask learning, usually, the goal is to transfer knowledge between the tasks to improve performance on all tasks simultaneously. However, in our scenarios we are only interested in the performance of a single task. For comparable corpora classification this is the classifier accuracy, and for focused domain translation we are interested in the final MT system performance. Therefore, in the rest of the chapter, we do not refer to these problem setups as multitask learning, but as learning with multiple annotations or ‘multi-type annotation

learning'. We will propose active learning strategies for each of these scenarios to jointly reduce the effort involved in obtaining the multiple annotations with a goal of improving performance of the focus task. Our multi-type annotation active learning strategies can be extended and used for other annotation tasks that fit these scenarios, but in this thesis we do not explore the general applicability outside these two scenarios.

5.1.3 Cost Models

We will frame and study the multi-type annotation active learning under a cost-sensitive framework. More formally, consider multiple annotation tasks that elicit different types of annotation $i \in 1...n$ from the unlabeled dataset U. The goal of multi-annotation active learning is to select the optimal set of instances for each annotation i. Let the number of instances elicited for each annotation per iteration be $k_1...k_i....k_n$ and the cost models for be $c_1...c_i....c_n$. However, instead of optimizing the number of instances selected, we optimize the selection under a provided budget B per iteration of the active learning algorithm, and $|U|$ is the size of the total unlabeled data set. Therefore our cost based formulation will be as seen below and therefore our evaluation of the active learning will also be geared towards reducing overall cost across the multiple annotation tasks.

$$\begin{aligned} B_i = &\quad c_i * k_i \\ where \sum_i^n B_i \leq &\quad B, \forall k_i \leq \quad |U| \end{aligned}$$

The cost-sensitive framework will make it possible to run interesting experiments.

- *Real-world costs:* We can obtain real-world cost information from online platforms like Mechanical Turk, and plug them into our setup to study the effect of multiple annotations.
- *Skewed costs:* We can simulate scenarios under extreme parametrization of the cost models, and study the benefit of using multiple annotations. For, instance one interesting configuration is where the document/sentence classification is 10 times lower than the actual translation task.

5.1.4 Evaluation

As discussed earlier, the success criteria in a multitask learning setup can typically measured in two different ways:

Individual performance:

Although in a multitask learning setup, we combine different tasks in order to improve the overall performance, we may still want to evaluate individual performance of each task separately. We can define the specific evaluation criteria for individual tasks and measure the individual task performance guided by their respective evaluation criteria and observe performance on the task of relevance. In all our tasks we will measure effectiveness of multi-type annotation active learning by performance of the relevant individual task. We will use BLEU or METEOR automatic metrics to evaluate the performance of the individual MT system on their respective test sets. For classification tasks we will report classifier accuracy or F-scores.

Cumulative performance

When the goal is to improve not one individual task, but all the tasks involved in the multitask learning, the evaluation criteria needs to also reflect the cumulative aspect of the learning problem. One way to compute such a cumulative metric is to do a weighted combination of the scores from evaluating individual metrics and normalize the final score as shown below. λ_i can also be seen as the importance of the task to our final evaluation.

$$multieval = \sum_{i=1}^{N} \frac{\lambda_i T_i}{N} \tag{5.1}$$

In this thesis however, our primary aim is to improve the quality of translation systems and we only use the second annotation as appropriate in the context of the translation system pipeline. We will therefore not evaluate on the performance of the first annotation, but only report the end translation quality. However, that said, the second task may still improve due to re-training on the newly available annotated data, although the data may have been selected in a suboptimal manner.

5.2 Comparable Corpora Classification Task

The state-of-the-art Machine Translation (MT) systems are statistical, requiring large amounts of parallel corpora. Such corpora needs to be carefully created by language experts or speakers, which makes building MT systems feasible only for those language pairs with sufficient public interest or financial support. With the increasing rate of social media creation and the quick growth of web media in languages other than English makes it relevant for language research community to explore the feasibility of Internet as a source

for parallel data. Resnik and Smith [2003] show that parallel corpora for a variety of languages can be harvested on the Internet.

There are multiple challenges in building comparable corpora for consumption by the MT systems. The first challenge is to identify the parallelism between documents of different languages which can be reliably done using cross lingual information retrieval techniques. Once we have identified a subset of documents that are potentially parallel, the second challenge is to identify comparable sentence pairs. This is an interesting challenge as the availability of completely parallel sentences on the internet is quite low in most language-pairs, but one can observe very few comparable sentences among comparable documents for a given language-pair. Our work tries to address this problem by posing the identification of comparable sentences from comparable data as a supervised classification problem. Unlike earlier research Munteanu and Marcu [2005] where the authors try to identify parallel sentences among a pool of comparable documents, we try to first identify comparable sentences in a pool with dominantly non-parallel sentences. We then build a supervised classifier that learns from user annotations for comparable corpora identification. Training such a classifier requires reliably annotated data that may be unavailable for low-resource language pairs. Involving a human expert to perform such annotations is expensive for low-resource languages and so we propose active learning as a suitable technique to reduce the labeling effort.

There is yet one other issue that needs to be solved in order for our classification based approach to work for truly low-resource language pairs. As we will describe later, our comparable sentence classifier relies on the availability of an initial seed lexicon that can either be provided by a human or can be statistically trained from parallel corpora Och and Ney [2003]. Experiments show that a bigger lexicon provides us with better coverage for effective identification of comparable corpora. However, availability of such a resource can not be expected in very low-resource language pairs, or even if present may not be of good quality. This opens an interesting research question - Can we also elicit such information effectively at low costs? We propose active learning strategies for identifying the most informative comparable sentence pairs which a human can then extract parallel segments from.

The first form of annotation provides us with sentence pairs labeled with class labels (comparable or not-parallel) and we can use them in tuning the feature weights of our classifier. The second form of supervision, parallel translation segments, can be used as a seed lexicon to instantiate the feature space for the classifier. For the comparable sentence classifier to perform well, we show that both forms of supervision are needed and we introduce an active learning protocol to combine the two forms of supervision under a single joint active learning strategy. Our work on application of multi annotation active learning for comparable corpora classification has also been published Ambati et al. [2011a]. Our contribution in this part of the thesis is the application of active learning for acquiring comparable data in the low-resource scenario, especially relevant when working

with low-resource languages.

There has been a lot of interest in using comparable corpora for MT, primarily on extracting parallel sentence pairs from comparable sources Zhao and Vogel [2002], Fung and Yee [1998]. Some work has gone beyond this focussing on extracting sub-sentential fragments from noisier comparable data Munteanu and Marcu [2006], Quirk et al. [2007]. Munteanu and Marcu [2005] propose bootstrapping using an existing classifier for collecting new data. However, this approach works when there is a classifier of reasonable performance. In the absence of parallel corpora to train lexicons human constructed dictionaries were used as an alternative which may, however, not be available for a large number of languages. Our proposal of active learning is suitable for highly impoverished language scenarios where it is expensive to obtain expert annotations.

5.2.1 Supervised Comparable Sentence Classification

In this section we discuss our supervised training setup and the classification algorithm. Our classifier tries to identify comparable sentences from among a large pool of noisy comparable sentences. We define comparable sentences as being translations that have around fifty percent or more translation equivalence. In future we will evaluate the robustness of the classifier by varying levels of noise at the sentence level.

Training the Classifier

Following Munteanu and Marcu [2005], we use a Maximum Entropy classifier to identify comparable sentences. The classifier probability can be defined as:

$$Pr(c_i|S,T) = \frac{1}{Z(S,T)} exp\left(\sum_{j=1}^{n} \lambda_j f_{ij}(c_i,S,T)\right)$$

where (S,T) is a sentence pair, c_i is the class, f_{ij} are feature functions and $Z(S)$ is a normalizing factor. The parameters λ_i are the weights for the feature functions and are estimated by optimizing on a training data set. For the task of classifying a sentence pair, there are two classes, $c_0 = comparable$ and $c_1 = non\ parallel$. A value closer to one for $Pr(c_1|S,T)$ indicates that (S,T) are comparable.

To train the classifier we need comparable sentence pairs and non-parallel sentence pairs. While it is easy to find negative examples online, acquiring comparable sentences is non-trivial and requires human intervention. Munteanu and Marcu [2005] construct negative examples automatically from positive examples by pairing all source sentences with all target sentences. We, however, assume the availability of both positive and negative

examples to train the classifier. We use the GIS learning algorithm for tuning the model parameters.

The features are defined primarily based on translation lexicon probabilities. Rather than computing word alignment between the two sentences, we use lexical probabilities to determine alignment points as follows: a source word s is aligned to a target word t if $p(s|t) > 0.5$. Target word alignment is computed similarly. Long contiguous sections of aligned words indicate parallelism. We use the following features:

- Source and target sentence length ratio
- Source and target sentence length difference
- Lexical probability score, similar to IBM model 1
- Number of aligned words
- Longest aligned word sequence
- Number of un-aligned words

Lexical probability score, and alignment features generate two sets of features based on translation lexica obtained by training in both directions. Features are normalized with respect to the sentence length.

In our experiments we observe that the most informative features are the ones involving the probabilistic lexicon. However, the comparable corpora obtained for training the classifier cannot be used for automatically training a lexicon. We, therefore, require the availability of an initial seed parallel corpus that can be used for computing the lexicon and the associated feature functions. We notice that the size of the seed corpus has a large influence on the accuracy of the classifier. Figure 5.1 shows a plot with the initial size of the corpus used to construct the probabilistic lexicon on x-axis and its effect on the accuracy of the classifier on y-axis. The sentences were drawn randomly from a large pool of Urdu-English parallel corpus and it is clear that a larger pool of parallel sentences leads to a better lexicon and an improved classifier.

5.3 Active Learning for Comparable Corpora Classification Task

Our selection strategies for obtaining class labels for training the classifier uses the model in its current state to decide on the informative instances for the next round of iterative training. We propose the following two sampling strategies for this task.

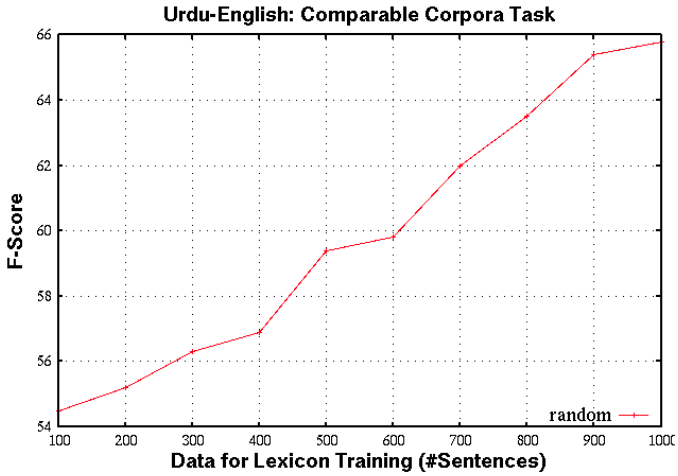

Figure 5.1: The size of seed parallel corpora for training lexicons vs. classifier performance for Urdu-English language pair. Lot of parallel data is required for cleaner lexicons and better performance of classifier, but obtaining such data is expensive

5.3.1 Framework for Multi-Type Annotation Active Learning

We now discuss our active learning framework for building comparable corpora as shown in Algorithm 7. We start with an unlabeled dataset $U_0 = \{x_j =< s_j, t_j >\}$ and a seed labeled dataset $L_0 = \{(< s_j, t_j >, c_i)\}$, where $c \in 0, 1$ are class labels with 0 being the non-parallel class and 1 being the comparable data class. We also have $T_0 = \{< s_k, t_k >\}$ which corresponds to parallel segments or sentences identified from L_0 that will be used in training the probabilistic lexicon. Both T_0 and L_0 can be very small in size at the start of the active learning loop. In our experiments, we tried with as few as 50 to 100 sentences for each of the datasets.

We perform an iterative budget motivated active learning loop for acquiring labeled data over k iterations. We start the active learning loop by first training a lexicon with the available T_k and then using that we train the classifier over L_k. We, then score all the sentences in the U_k using the model θ and apply our selection strategy to retrieve the best scoring instance or a small batch of instances. In the simplest case we annotate this instance and add it back to the tuning set C_k for re-training the classifier. If the instance was a comparable sentence pair, then we could also perform the second annotation conditioned upon the availability of the budget. The identified sub-segments (ss_i , tt_i) are added back to the training data T_k used for training the lexicon in the subsequent iterations.

Algorithm 6 JOINT INSTANCE SELECTION FOR MAAL LEARNING SETUP

Given Unlabeled Comparable Corpus: U_0
Given Seed Parallel Corpus: T_0
Given Tuning Corpus: L_0
for $k = 0$ to K **do**
    Train Lexicon using T_k
    θ = Tune Classifier using C_k
    while $Cost < B_k$ **do**
        i = Query(U_k,L_k,T_k,θ)
        c_i = Human Annotation-1 (s_i, t_i)
        $(ss_i$,$tt_i)$ = Human Annotation-2 x_i
        $L_k = C_k \cup (s_i, t_i, c_i)$
        $T_k = T_k \cup (ss_i, tt_i)$
        $U_k = U_k$ - x_i
        $Cost = Cost_1 + Cost_2$
    end while
end for

5.3.2 Cost Model

In the previous section we have shown that our classifier requires two kinds of annotated data: class labels for identifying comparable vs. non-parallel data and secondly clean parallel segments within the comparable sentences. Lack of existing annotated data requires reliable human annotation that is expensive and effort-intensive. We propose active learning for the problem of effectively acquiring multiple annotations starting with unlabeled data. In active learning, the learner has access to a large pool of unlabeled data and sometimes a small portion of seed labeled data. The objective of the active learner is then to select the most informative instances from the unlabeled data and seek annotations from a human expert, which it then uses to retrain the underlying supervised model for improving performance.

A meaningful setup to study multi annotation active learning is to take into account the cost involved for each of the annotations. In the case of comparable corpora we have two annotation tasks, each with cost models $Cost_1$ and $Cost_2$ respectively. The goal of multi annotation active learning is to select the optimal set of instances for each annotation so as to maximize the benefit to the classifier. Unlike the traditional active learning, where we optimize the number of instances we label, here we optimize the selection under a provided budget B_k per iteration of the active learning algorithm.

5.3.3 Query Strategies for Comparable Corpora Classification

We first propose two sampling strategies for eliciting the first kind of annotation, which is class label annotation for a given sentence-pair. Our goal is to select instances that could informative in tuning the weights for the parameters for our classifier.

Certainty Sampling

This strategy selects instances where the current model is highly confident. While this may seem redundant at the outset, we argue that this criteria can be a good sampling strategy when the classifier is weak or trained in an impoverished data scenario. Certainty sampling strategy is a lot similar to the idea of unsupervised approaches like boosting or self-training. However, we make it a semi-supervised approach by having a human in the loop to provide affirmation for the selected instance. Consider the following scenario. If we select an instance that our current model prefers and obtain a contradicting label from the human, then this instance has a maximal impact on the decision boundary of the classifier. On the other hand, if the label is reaffirmed by a human, the overall variance reduces and in the process, it also helps in assigning higher preference for the configuration of the decision boundary. Melville et al. [2005] introduce a certainty sampling strategy for the task of feature labeling in a text categorization task. Inspired by the same we borrow the name and also apply this as an instance sampling approach. Given an instance x and the classifier posterior distribution for the classes as $P(.)$, we select the most informative instance as follows:

$$x* = arg\,max_x P(c = 1|x)$$

Margin-based Sampling

The certainty sampling strategy only considers the instance that has the best score for the comparable sentence class. However we could benefit from information about the second best class assigned to the same instance. In the typical multi-class classification problems, earlier work shows success using such a 'margin based' approach Scheffer et al. [2001], where the difference between the probabilities assigned by the underlying model to the first best and second best classes is used as the sampling criteria.

Given a classifier with posterior distribution over classes for an instance $P(c = 1|x)$, the margin based strategy is framed as $x* = arg\,min_x P(c_1|x) - P(c_2|x)$, where c_1 is the best prediction for the class and c_2 is the second best prediction under the model. It should be noted that for binary classification tasks with two classes, the margin sampling approach reduces to an uncertainty sampling approach Lewis and Catlett [1994].

Our multi-type annotation active learning technique for the comparable corpora classification task is more closely related to the work of feature vs. instance labeling Druck et al. [2009]. While in this line of work, the authors decides whether to annotate features directly or provide annotations for entirely new instances, our approach jointly selects a single instance for both kinds of annotation.

5.3.4 Query Strategies for Acquiring Parallel Segments for Lexicon Training

We now propose two sampling strategies for the second annotation. Our goal is to select instances that could potentially provide parallel segments for improved lexical coverage and feature computation.

Diversity Sampling

We are interested in acquiring clean parallel segments for training a lexicon that can be used in feature computation. It is not clear how one could use a comparable sentence pair to decide the potential for extracting a parallel segment. However, it is highly likely that if such a sentence pair has new coverage on the source side, then it increases the chances of obtaining new coverage. We, therefore, propose a diversity based sampling for extracting instances that provide new vocabulary coverage . The scoring function $tc\_score(s)$ is defined below, where $Voc(s)$ is defined as the vocabulary of source sentence s for an instance $x_i =< s_i, t_i >$, T is the set of parallel sentences or segments extracted so far.

$$tc\_score(s) = \sum_{s=1}^{|T|} sim(s, s') * \frac{1}{|T|} \tag{5.2}$$

$$sim(s, s') = |(Voc(s) \cap Voc(s')| \tag{5.3}$$

Alignment Ratio

We also propose a strategy that provides direct insight into the coverage of the underlying lexicon and prefers a sentence pair that is more likely to be comparable. We call this *alignment ratio* and it can be easily computed from the available set of features discussed in Section 5.2.1 as below:

$$a\_score(s) = \frac{\#unalignedwords}{\#alignedwords} \tag{5.4}$$

$$s* = arg\,max_s a\_score(s) \tag{5.5}$$

This strategy is quite similar to the diversity based approach as both prefer selecting sentences that have a potential to offer new vocabulary from the comparable sentence pair. However while the diversity approach looks only at the source side coverage and does not depend upon the underlying lexicon, the alignment ratio utilizes the model for computing coverage. It should also be noted that while we have coverage for a word in the sentence pair, it may not make it to the probabilistically trained and extracted lexicon.

5.3.5 Joint Selection Strategy for Multiple Annotations

Finally, given two annotations and corresponding sampling strategies, we try to jointly select the sentence that is best suitable for obtaining both the annotations and is maximally beneficial to the classifier. We select a single instance by combining the scores from the different selection strategies as a geometric mean. For instance, we consider a margin based sampling ($margin$) for the first annotation and a diversity sampling (tc_score) for the second annotation, we can jointly select a sentence that maximizes the combined score as shown below:

$$total\_score(s) = margin(s) * tc\_score(s) \tag{5.6}$$

$$s* = arg\,max_s total\_score(s) \tag{5.7}$$

5.4 Experiments

5.4.1 Data Set Creation

This research primarily focuses on identifying comparable sentences from a pool of dominantly non-parallel sentences. To our knowledge, there is a dearth of publicly available comparable corpora of this nature. We, therefore, simulate a low-resource scenario by using realistic assumptions of noise and parallelism at both the corpus-level and the sentence-level. In this section we discuss the process and assumptions involved in the creation of our datasets and try to mimic the properties of real-world comparable corpora harvested from the web.

We first start with a sentence-aligned parallel corpus available for the language pair. We then divide the corpus into three parts. The first part is called the 'sampling pool' and is set aside to use for drawing sentences at random. The second part is used to act as a non-parallel corpus. We achieve non-parallelism by randomizing the mapping of the target sentences with the source sentences. This is a slight variation of the strategy used in Munteanu and Marcu [2005] for generating negative examples for their classifier. The

third part is used to synthesize a comparable corpus at the sentence-level. We perform this by first selecting a parallel sentence-pair and then padding either sides by a source and target segment drawn independently from the sampling pool. We control the length of the non-parallel portion that is appended to be lesser than or equal to the original length of the sentence. Therefore, the resulting synthesized comparable sentence pairs are guaranteed to contain at least 50% parallelism.

We use this dataset as the unlabeled pool from which the active learner selects instances for labeling. Since the gold-standard labels for this corpus are already available, which gives us better control over automating the active learning process, which typically requires a human in the loop. However, our active learning strategies are in no way limited by the simulated data setup and can generalize to the real world scenario with an expert providing the labels for each instance.

We perform our experiments with data from two language pairs: Urdu-English and Spanish-English. For Urdu-English, we use the parallel corpus NIST 2008 dataset released for the translation shared task. We start with 50,000 parallel sentence corpus from the released training data to create a corpus of 25,000 sentence pairs with 12,500 each of comparable and non-parallel sentence pairs. Similarly, we use 50,000 parallel sentences from the training data released by the WMT 2008 datasets for Spanish-English to create a corpus of 25,000 sentence pairs. We also use two held-out data sets for training and tuning the classifier, consisting of 1000 sentence pairs (500 non-parallel and 500 comparable).

5.4.2 Results

We perform two kinds of evaluations: the first, to show that our active learning strategies perform well across language pairs and the second, to show that multi annotation active learning leads to a good improvement in performance of the classifier.

Performance of Active Learning for Single Annotation

In earlier section we proposed multiple active learning strategies for both eliciting both kinds of annotations. A good active learning strategy should select instances that contribute to the maximal improvement of the classifier. The effectiveness of active learning is typically tested by the number of queries the learner asks and the resultant improvement in the performance of the classifier. The classifier performance in the comparable sentence classification task can be computed as the F-score on the held out dataset. For this work, we assume that both the annotations require the same effort level and so assign uniform cost for eliciting each of them. Therefore the number of queries is equivalent to the total cost of supervision.

Figure 5.2 shows our results for the Urdu-English language pair, and Figure 5.3 plots the

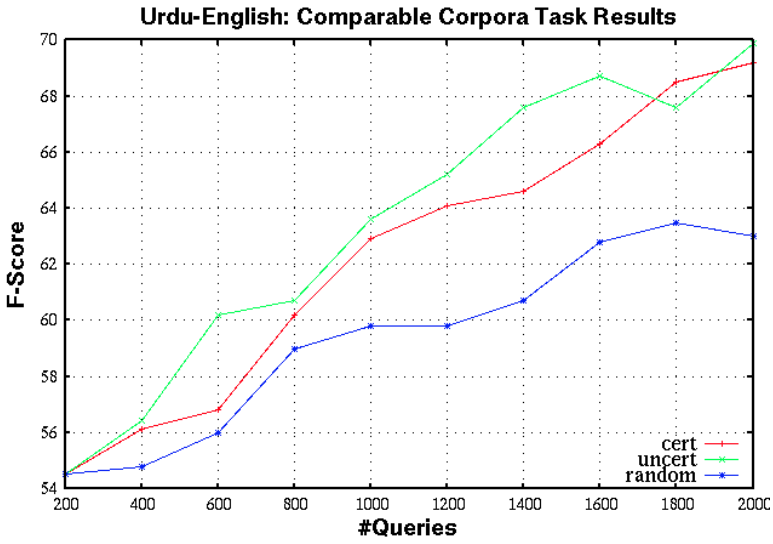

Figure 5.2: Comparable corpora classifier performance curve for Urdu-English language-pair. Number of labeled instances (# of queries) on x-axis to train the classifier and the classifier f-score on y-axis. Both our strategies (cert,uncert) beat a random selection baseline

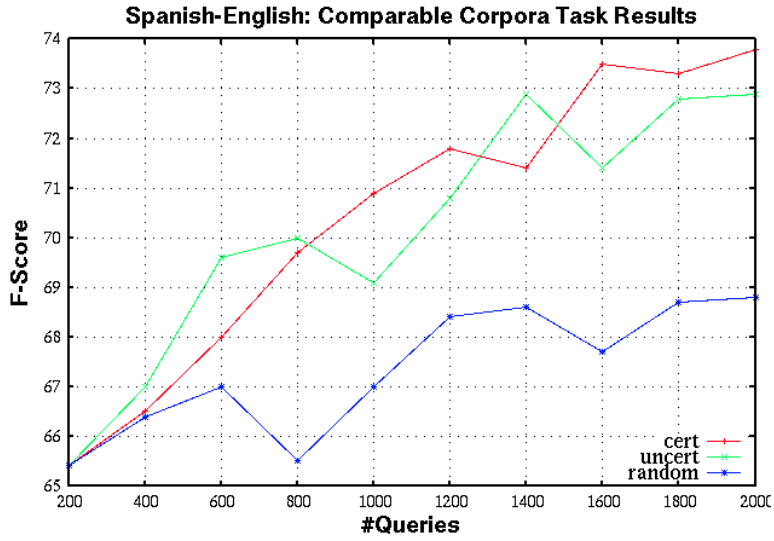

Figure 5.3: Comparable corpora classifier performance curve for Spanish-English language-pair. Number of labeled instances (# of queries) on x-axis to train the classifier and the classifier f-score on y-axis. Both our strategies (cert,uncert) beat a random selection baseline

Spanish-English results with the x-axis showing the total number of queries posed to obtain annotations and the y-axis shows the resultant improvement in accuracy of the classifier. In these experiments we do not actively select for the second annotation but acquire the parallel segment from the same sentence. We compare this over a random baseline where the sentence pair is selected at random and used for eliciting both annotations at the same time.

Firstly, we notice that both our active learning strategies: certainty sampling and margin-based sampling perform better than the random baseline. For the Urdu-English language pair we can see that for the same effort expended (i.e 2000 queries) the classifier has an increase in accuracy of 8 absolute points. For Spanish-English language pair the accuracy improvement is 6 points over random baseline. Another observation from Figure 5.3 is that for the classifier to reach an fixed accuracy of 68 points, the random sampling method requires 2000 queries while the from the active selection strategies require significantly less effort of about 500 queries.

Performance of Joint Selection with Multiple Annotations

We now evaluate our joint selection strategy that tries to select the best possible instance for both the annotations. Figure 5.4 shows our results for the Urdu-English language pair, and Figure 5.5 plots the Spanish-English results for active learning with multiple annotations. As before, the x-axis shows the total number of queries posed, equivalent to the cumulative effort for obtaining the annotations and the y-axis shows the resultant improvement in accuracy of the classifier.

We evaluate the multi annotation active learning against two single-sided baselines where the sampling focus is on selecting instances according to strategies suitable for one annotation at a time. The best performing active learning strategy for the class label annotations is the certainty sampling (annot1) and so for one single-sided baseline, we use this baseline. We also obtain the second annotation for the same instance. By doing so, we might be selecting an instance that is sub-optimal for the second annotation and therefore the resultant lexicon may not maximally benefit from the instance. We also observe, from our experiments, that the diversity based sampling works well for the second annotation and alignment ratio does not perform as well. So, for the second single-sided baseline we use the diversity based sampling strategy (annot2) and get the first annotation for the same instance. Finally we compare this with the joint selection approach proposed earlier that combines both the annotation strategies (annot1+annot2). In both the language pairs we notice that joint selection for both annotations performs better than the baselines.

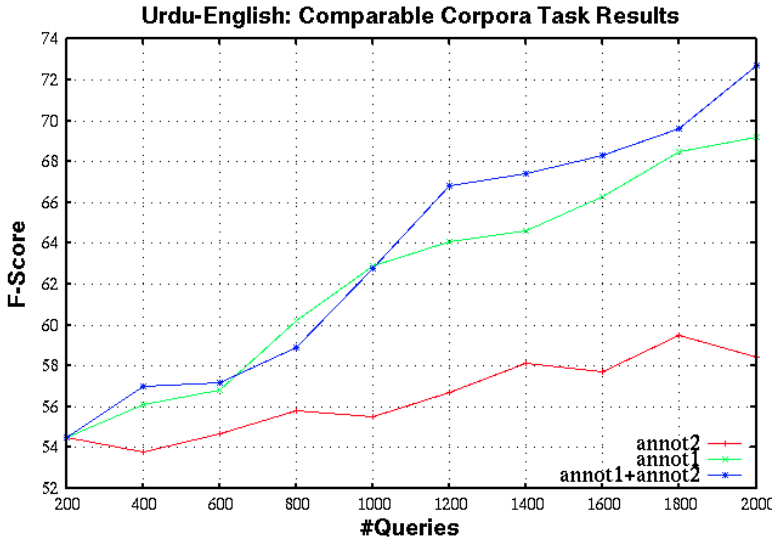

Figure 5.4: Comparable corpora classifier performance curve for Urdu-English language-pair. Number of queries (either class-labels:annot1 or parallel-segments:annot2 for lexicon training) on x-axis and the classifier f-score on y-axis. A joint-selection strategy (annot1+annot2) outperforms the best active selection algorithms (annot1 or annot2) for individual annotations.

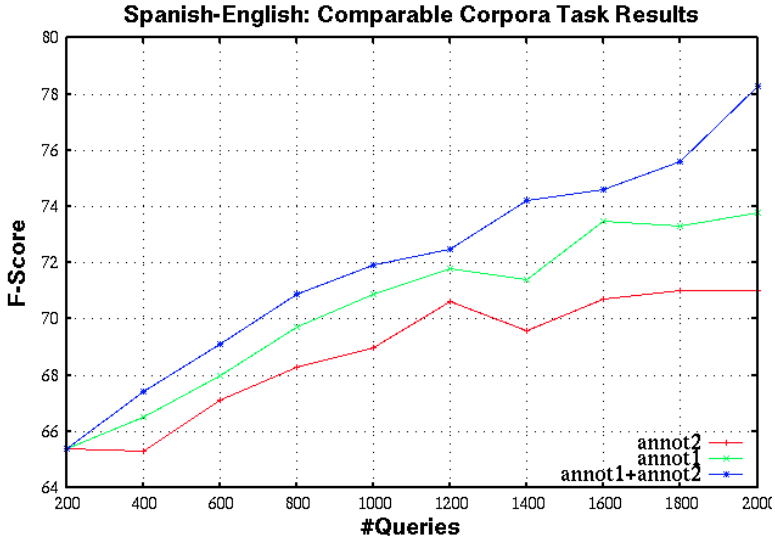

Figure 5.5: Comparable corpora classifier performance curve for Spanish-English language-pair. Number of queries (either class-labels:annot1 or parallel-segments:annot2 for lexicon training) on x-axis and the classifier f-score on y-axis. A joint-selection strategy (annot1+annot2) outperforms the best active selection algorithms (annot1 or annot2) for individual annotations.

5.4.3 Summary

We proposed active learning with multiple annotations for the challenge of building comparable corpora in low-resource scenarios. In particular, we identified two kinds of annotations: class labels (for identifying comparable vs. non-parallel data) and clean parallel segments within the comparable sentences. We implemented multiple independent strategies for obtaining each of the above in a cost-effective manner. Finally we also proposed a joint selection strategy that selects instances that are beneficial to both the annotations. Our active learning experiments in simulated low-resource scenarios show significant results over strong baselines for two language pairs.

5.5 Focused Domain Machine Translation

5.5.1 Introduction

In real world, low-resource language translation systems are often domain specific, targeted to address an immediate need of a small group of users. For example, the recent earthquake in Haiti, required translation of Haitian Creole into English in order to help the relief volunteers communicate with the local people and also disseminate information. The need was to quickly build MT systems that could address medical, travel, emergency domains. There are other recent projects that build MT systems for African languages in order to aid the rehabilitation of refugees in the United States. Such domain specific MT systems are of importance for humanitarian causes and are very time critical and typically have pre-specified, limited budgets. One can not afford the translation of a million sentences to train high accuracy systems, neither can we wait for the time taken for data entry. Hence the need of the day is to build algorithms that provide usable translation systems at a very low budget and take less time for development.

Domain specificity is also required for translation in majority languages. With the explosion of Internet, current translation service providers have a need to cater to the demands of users by adapting to new trends in data, new unseen domains and even genres like social media. It is unlikely to find parallel corpora for every such scenario, and creation of such resources needs more resources than we can expend in time and money.

In order to build an MT system from a source language S to a target language T, we first need to obtain domain specific sentence level data. It is easier to obtain a pool of mixed domain data by crawling the web or other forms of social media. However, if we are interested in a specific domain, we will need to then sample from this pool of mixed domain data. Our approach to building a domain specific MT system can be seen in figure 5.6. We treat this as a sentence classification task where the classifier can be trained on labeled data with human provided domain tagged sentences. The second phase of the approach is to

collect all the sentences classified in the first phase as a particular domain and have humans translate them. This is an instance of multitask learning, where the output of the first task, constraints the input for the second task. In this setup, the two learning problems differ in both feature space, input and output spaces and so there is little incentive in training these tasks together. However, the output of one task constraints the input space of the other.

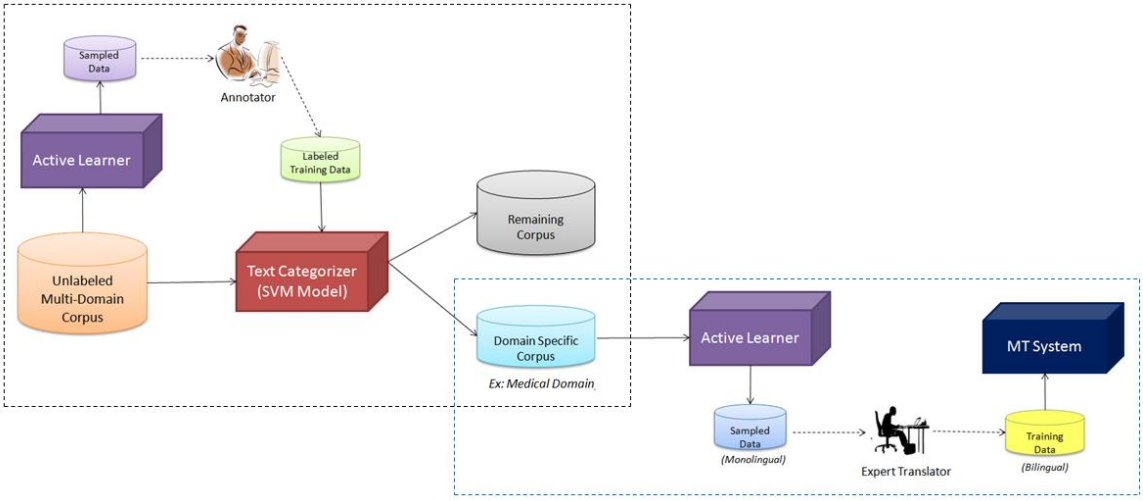

Figure 5.6: An ideal framework for building a Domain Specific Machine Translation that requires a highly accurate text classifier for obtaining in-domain data. The data can then be translated to build an MT system for the domain. The framework also shows the active learning modules embedded in each of the tasks

5.5.2 Task 1: Sentence Classification

Given a monolingual corpus U_c consisting of sentences from a set of domains D, our task is to take as input a source sentence $s_i \in U_c$ and predict whether it belongs to a particular domain $d \in D$. This task is very similar in nature to document classification, but with sparse text. Support vector machines (SVM) have proven successful for the document classification task. Given human annotated data L_c where each sentence $s_i \in L_c$ is domain tagged, we can train a classifier to predict the domain of a new test sentence. We frame this as a binary class classification problem. We use SVMs for training our sentence classification. SVMs have a nice property that make it a preferable binary classifier when classifying text Joachims [2001]. We can also use the SVM to output a posterior probability for the classification as $P(D = d|s)$.

The performance of the classifier depends to a large extent on the features used. We tried

using higher order features in the form of bigrams and trigrams and observe that unigrams or bag of words are a better tradeoff given the feature sparsity at sentence level. Motivated by document classification, we also tried more sophisticated features like term-frequency and inverse sentence frequency, and observe very minor improvements over binary features. So we stick to a very basic form of the features and use unigrams in the sentence as binary features. We obtain two sets of data - in domain $L_{c=d}$ and out of domain $L_{c=d'}$.

5.5.3 Task 2: Sentence Translation

The second task is build a machine translation system. The input to train a statistical MT system is a set of monolingual sentences U_t, which are then translated by a human to create a set of parallel sentences $L_t = \{< s, t >\}$. As seen in figure 5, we use the in-domain data obtained from the first phase of sentence classification as input to the MT system in the second phase ($L_c = U_t$).

5.6 Active Learning for Focussed Domain Translation Task

The main challenge in building domain specific translation systems in such cases are two-fold. Firstly the non-availability of domain specific monolingual source language data and the non-availability of access to bilinguals speakers that are experts in the domain. We therefore propose a two-step workflow for building a domain specific translation system. We first acquire focused unlabeled monolingual data in the source language. This will be done by gathering text from the Internet and then applying a classification algorithm that was trained on data labeled from the relevant domain. This can help us identify the relevant domain specific text in the source language from among a mixed domain data. The second phase is to provide this data to a language translator with expertise in the particular domain to translate from the source language into the target language. Both the phases require human effort for labeling which can become expensive based on the nature of the language and the domain. We resort to Active Learning (AL) techniques for building MT systems to cost-effectively elicit these annotations in low-resource and impoverished languages.

Our hypothesis is that the task of active sentence selection can be improved by using domain membership knowledge from a first task of sentence categorization. We improve the sentence categorization by eliciting a second annotation in the form of domain tagging. However, the traditional setup for active learning assumes a single model for which the selection of data is optimized. Given the nature of supervised learning tasks of the current day, we can say that this is a strong assumption. For instance recent work in parsing shows that named entity recognition task and parsing tasks can be combined to improve overall performance of a parser Finkel and Manning [2010]. Such setups are increasingly

becoming popular in NLP and are called joint learning or multitask learning Caruana [1997]. Building a domain specific MT system is also a multitask learning problem with a pipeline architecture where the two tasks are non-linearly dependent upon each other. We therefore extend the traditional active learning setup that is suitable for eliciting a single annotation to work with multiple annotations, more specifically for tasks in a pipeline model.

In the rest of the section we will try to establish notation corresponding to both the tasks and provide independent active learning strategies for each of the tasks.

5.6.1 Active Learning for Text Classification

In our work we have tried a density based selection of sentences for improving our sentence classifier. While this has the advantage that we can use this with any classification model, we can do better with information of the model and its uncertainty. Active learning for document classification task has been explored well for a variety of classification models like SVM Tong and Koller [2002] etc. We will explore similar query selection strategies for selecting samples for the sentence classifier.

Sampling instances around the decision boundary of the classifier helps refinement of the model quicker. Given a classifier with posterior distribution over classes for an instance $P(D = d|s)$, the margin based strategy is framed as below, where d_1 is the best prediction for the domain and d_2 is the second best prediction under the model. Intuitively, we select sentences where the model only selects a winner by a narrow margin. We will call this approach as margin-based sampling.

$$tScore(s) = P(D = d_1|s) - P(D = d_2|s)$$

Our experiments have shown that we can improve the quality of a domain specific MT system in the multitasking scenario. We propose work along the following multiple lines:

Uncertainty sampling:

We also experiment by sampling sentences where the model is highly uncertain about the prediction. In order to be able to do this, we need the classifier to provide normalized classification scores for all the output classes. This information can be obtained by proper calibration of the posterior distribution of the SVM classifier.

$$s* = \arg\min_{s} P(D = d|s)$$

Margin Sampling:

Sampling instances around the decision boundary of the classifier helps refinement of the model quicker. Given a classifier with posterior distribution over classes for an instance $P(D = d/s)$, the margin based strategy is framed as below, where d_1 is the best prediction for the domain and d_2 is the second best prediction under the model. Intuitively, we select sentences where the model only selects a narrow winner.

$$s* = \arg\min_{s} P(D = d_1|s) - P(D = d_2|s)$$

Density/Vocabulary Coverage:

Our active learning algorithm for this particular task takes the following approach. We first improve performance of the sentence classification task by actively selecting training sentences. We use a density based diversity approach in order to favor sentences that provide a larger coverage of the feature space, unigrams in our case. The scoring function $tc\_score(s)$ is defined below, where $Voc(s)$ is defined as he vocabulary of sentence s in the unlabeled, mixed domain corpus U.

$$tc\_score(s) = \sum_{s=1}^{|U|} sim(s, s') * \frac{1}{|U|} \tag{5.8}$$

$$sim(s, s') = \|(Voc(s) \cap Voc(s')\| \tag{5.9}$$

5.6.2 Active Learning for Sentence Translation

For sentence selection we use the density weighted diversity ensemble strategy as discussed in Ambati et al. [2010a]. However, we extend it to include information about the domain in the form of posterior class membership probability $P(D|s)$, obtained from the classifier above. The strategy is to select sentences that have the most representative n-grams and have not yet been seen in the labeled bilingual corpus. Representativeness or the 'density' of a sentence is computed as a function of the unlabeled monolingual data as can be seen in equation below, where U_t is the unlabeled domain specific data and L_t is the domain specific parallel data and $Phrases(s)$ is the same as before, the phrases in a sentence up to a certain length $n = 3$. . Novelty or 'uncertainty' is computed as the number of new phrases that a sentence has to offer. We compute the final score of a sentence as the harmonic mean of both these metrics with a tunable parameter 'β', that helps us balance the novelty or density factor of the selected sentence.

$$d(s) = \frac{\sum_{x \in Phrases(s)} P(x|U_t) * e^{-\lambda count(x|L_t)}}{\|Phrases(s)\|}$$

$$u(s) = \frac{\sum_{x \in Phrases(s)} \alpha}{\|Phrases(s)\|}; \alpha = \begin{cases} 1 & x \in Phrases(L_t) \\ 0 & \end{cases}$$

$$V(s) = \frac{(1+\beta^2)d(s) * u(s)}{\beta^2 d(s) + u(s)}$$

$$s* = \arg\max_{s \in U} P(D|s)V(s)$$

5.6.3 Multi-Type Annotation Active Learning for Focused Domain Translation

In this section we discuss our multiple annotation active learning strategies for selecting the task and the instance for the tasks in the pipeline model.

For the first task, we start with an unlabeled dataset U_t^0 and a seed labeled dataset $L_t^0 = \{(< s_j >, d_i)\}$, where $d \in 0, 1$ are class labels with 0 being one domain and 1 for the second. We describe the algorithm for a binary classification setup, but it should generalize for a multi-class classification as well. We will use the subscript c to represent data related to the first phase of classification and a subscript t to represent all variables related to the second phase of translation. We perform an iterative budget motivated active learning loop for acquiring labeled data over k iterations. We implement a batch learning setup where in each iteration selection of instances and annotation continues a fixed budget B_k for iteration k.

We study the multi annotation active learning under a cost-sensitive framework. Consider multiple annotation tasks that elicit different types of annotation $i \in 1...n$ from the unlabeled dataset U. The goal of multi-annotation active learning is to select the optimal set of instances for each annotation i. Instead of optimizing the number of instances selected, we optimize the selection under a provided budget B per iteration of the active learning algorithm. Therefore our cost based formulation should include costs for each annotation type. In our case we have two annotations each costing $Cost1$ and $Cost2$.

One-sided Active learning

One of the approaches proposed by Reichart et al. [2008b] is to focus the selection strategy on a single task, but obtain annotations for both the tasks. In our case, we have the two different annotations in the form of domain labels and translation. We therefore have two baselines, one focusing on the text classifier performance (a1) and the other focusing on the

translation quality (a2), where we use the active learning strategies for each of the individual annotations respectively, as discussed in the previous section. Therefore, although this approach may be optimal for one annotation, it may be suboptimal for the other annotation. The cost, however, will be the sum of the costs of acquiring both annotations, separately. Our goal in this section is to now use these as baselines but implement approaches that perform better than them.

Joint Instance Selection Approach

In the previous section, we proposed a joint instance selection approach for the comparable corpora selection task in which we combine the values of each individual one-sided active learning strategies to select the instances for labeling. We have shown that the combined score may select an unlabeled instance that is suboptimal for each individual annotation, but is optimal for the combined score. We use the same approach for this task as well. We score the unlabeled sentences individually and create two scored lists, one list of sentences that are deemed of high value for improving the sentence classification task and the second list of sentences that are considered maximally beneficial to the classifier. Similar to our earlier approach, we combine these scores using a geometric mean to obtain a total score for the sentence.

For instance, if the value function of the active selection strategy for first annotation is $V_1(s)$ and the value function for second annotation is $V_2(s)$, then the we compute the joint value as shown below. We then sort the sentences according to this score to select the top scoring instance and obtain both annotations for the same instance.

$$total\_score(s) = V_1(s) * V_2(s) \tag{5.10}$$

$$s* = arg\,max_s total\_score(s) \tag{5.11}$$

Task Selection Approach

From the framework described for building domain specific MT systems in Figure 5, it is evident that for the selection of good in-domain sentences for translation, the domain information available for the sentence needs to be reliable. In other words, the accuracy of the classifier needs to be high. Although we can not pre-compute the level of accuracy that is desirable of the sentence classifier for an effective MT system, one thing we can be sure of is that the performance of the second task is dependent on that of the first task. In view of the above we propose a multi-type annotation strategy that initially focuses on improving the sentence classifier but then switches to the second task when we observe saturation in performance on the first task.

The task selection strategy suggests that we decide a point on the learning curve of the first task where we switch from eliciting one kind of annotation to the other. Our active learning algorithm is described in Algorithm 7. We start the active learning loop by first training a classifier for the first phase using the available seed data. We, then score all the sentences in the U_c^k using the model θ and apply our selection strategy to retrieve a small batch of instances. The most important part of the algorithm is the "taskFlag" variable that decides which annotation we will elicit from the humans. If the taskFlag value is set to 0, we elicit supervised training data for sentence classification, and if it is 1, we elicit translations for the in-domain data that is classified in the first phase.

Algorithm 7 TASK SELECTION FOR MAAL LEARNING SETUP

Unlabeled Mixed Domain Data: U_c^0
Seed Labeled Corpus: L_c^0
taskFlag: 0
θ_0 = Train Classifier L_c^0
for $k = 0$ to K **do**
    while $Cost < B_k$ **do**
        if taskFlag : 1 **then**
            i = Query1 (U_c^k, θ_k)
            d_i = Human Annotation-1 (s_i)
            $L_c^k = L_c^k \cup (s_i, d_i)$
            $U_c^k = U_c^k$ - s_i
            $Cost = Cost + Cost_1$
        else
            j = Query2 (U_t^k)
            t_j = Human Annotation-2 (s_j)
            $L_t^k = L_t^k \cup (< s_j, t_j >)$
            $U_t^k = U_t^k$ - s_i
            $Cost = Cost + Cost_2$
        end if
    end while
    θ_{k+1} = Train Classifier L_c^k
    U_t^{k+1}= In-Domain as per θ_{k+1}
    O_t^{k+1}= Out-of-Domain as per θ_{k+1}
    taskFlag: AnnotationSelector(U_t^{k+1},O_t^{k+1}, θ_{k+1})
end for

Once the crossover point is reached we do not switch back to first annotation through the rest of the learning process. Therefore it is important that we select the appropriate crossover point, as switching earlier or later would hurt the performance of the second task. So, how do we compute this reliably?. Ideally, we would want the text classifier that

is trained on the labeled data to perform well with reduced error on the unlabeled data. However, we do not have true labels for the unlabeled data and so we can not accurately compute the performance saturation. We propose to estimate an approximation for the future error and use that for detecting the diminishing returns on training the task with more data. We discuss two ways to do this:

- **Held-out dataset** We can use a development set and use the performance of the trained classifier on that dataset as a surrogate for its performance on future data. In our case, the sentence classification task is evaluated using accuracy of classification and therefore we compute the slope of this curve plotted across iterations to observe the diminishing effect of further tagging of sentences. We decide to stop when the difference in slopes across different iterations drops below a threshold.
- **Expected error** We can approximate the future error by calculating the average expected error of on the unlabeled data. We can use the posterior class distribution of the classifier over the unlabeled data at every iteration to compute the same.

Using performance on a held-out dataset as a predictor for future accuracy is typically what supervised machine learning algorithms try to achieve. This method however requires us to have an initial gold-standard held-out dataset that we can use and also requires that it belongs to the same distribution as the labeled data and future unlabeled data. We, therefore, do not use this approach in this thesis.

We compute the expected error over the unlabeled data at end of every iteration after training the classifier and use the posterior class distribution as shown in Equation 5.13. We can then compute a derivative of the error to observe when the performance gains from further training sentences starts to plateau. The future error estimate of the sentences selected by the value function $V(.)$ is represented by $\epsilon(V_1(s))$, and the value δ is the threshold we use to decide the diminishing returns. We select a very low value for $\delta = 0.02$. The derivative estimation is not going to be exact as the classifier in that iteration would not have seen all the possible training data, but it is only going to be indicative of the relative performance over iterations and useful in detection of diminishing returns.

$$\epsilon(V_1(s)) = \frac{\sum_{s_i \in U_t} max(P(D=1|s), P(D=-1|s))}{|U|} \tag{5.12}$$

$$\frac{\partial \epsilon(V_1(s))}{\partial s_t} < \delta \tag{5.13}$$

Our multi-annotation active learning strategy described in this section is inspired by the switching strategy technique discussed in Donmez et al. [2007]. In that work, the authors identify strategies that work well for a certain operating range and propose switching to

another strategy for a different operating range. However, in that work, while the strategies are different, the annotation and the task remain the same. While our work is inspired by the switching strategy technique, we switch between annotations for training different tasks. In a relevant piece of work, Roth and Small [2008] discuss active learning for pipeline models. While in their approach they assume that the unlabeled data and the seed data are the same and accessible across all the phases in the pipeline model, it is clear that in our approach the unlabeled pool for selection is constrained by the previous phase. Therefore the distribution of data for subsequent phases changes based on the accuracies of the initial phases. Also, in their work once the instance is selected, the annotation is performed for the entire pipeline, whereas we only seek annotation for a particular phase.

5.7 Experiments

We perform two sets of experiments. Firstly, we are interested in learning the accuracy of the sentence classification model and how the active learning component affects its performance. Then we also experiment to see the effect of the first task on the overall translation quality.

5.7.1 Data Sets

Spanish-English

We perform this experiment with simulated data, where we have domain information for the dataset a priori. We selected Spanish-English as our language pair, where we have data from two different domains - political and travel. The political data, consisting of 240K sentences is a subset of the Europarl data and the travel data consisting of 121K sentences was released as part of the BTEC [Takezawa et al., 2002] corpus from IWSLT. The unlabeled data was created by combining the two datasets and randomizing the sentences.

Haitian Creole-English

The recent disaster in Haiti triggered a massive relief effort from around the world. Haitian Creole is spoken by a significant portion of the local population, but is not a common language outside of Haiti. Machine Translation was perceived to be immediately useful in such situations and a MT system can be useful not only as a communication device, but also in relief efforts like translation of medical documents, relief information documents, travel and safety information and more importantly help related SMS text messages pouring in from various parts of Haiti. Unfortunately, no readily available Machine Translation engine

Domain	Sentences	Ht tokens	En-Tokens
SMS messages	16,676	351K	324K
Newswire text	13,517	336K	292K
Medical dialog	1,619	10K	10K
Dictionaries	42,178	97K	92K
Bible + Others	41,872	939K	865K
Wikipedia	8,476	77K	90K

Table 5.1: Haitian Creole-English Datasets

existed for Haitian Creole at the time. This situation highlighted the need for low-resource translation systems and several research groups responded to work on building MT systems and participated in data collection efforts as a result. After the immediate urgency subsided, the data collected was released and many others used the data to understand building rapid MT systems.

Our group at CMU participated in the Haitian Creole-English translation system that was built as part of the Featured Translation Task of the WMT11 Hewavitharana et al. [2011]. The task involved translating text (SMS) messages that were collected during the humanitarian operations in the aftermath of the earthquake in Haiti in 2010. Due to the circumstances of this situation, the SMS messages were often noisy, and contained incomplete information. Additionally they sometimes contained text from other languages (e.g. French). As is typical in SMS messages, abbreviated text (as well as misspelled words) were present.

The WMT11 organizers provided us with several datasets (some unrelated to the SMS domain) of Haitian Creole-English parallel data from a variety of sources, including the medical dialog data released by CMU. A summary of the data is given in Table 5.1. The primary in-domain data comprises the translated (noisy) SMS messages. The additional data contains newswire text, medical dialogs, the Bible, several bilingual dictionaries, and parallel sentences from Wikipedia.

We use the SMS data as the primary in-domain data for our experiments and attempt to build a focused domain SMS translation system. For the second domain we will use data available from the Bible. We will not use the additional data consisting of newswire text, medical dialogs and Wikipedia. We combine the SMS data and Bible data and create a mixed domain data and use this as our starting point in the process of building a focused domain SMS translation system from Haitian Creole into English. The total SMS data consists of 16,676 sentence pairs and the Bible consists of around 30K parallel data therefore making the in-domain vs. out of domain ratio around 1:3 for this task.

5.7.2 Active improvement of sentence categorization

For both the language pairs, we evaluate our active learning based text categorization setup. In order to simulate the annotation phase of active learning we withhold the domain information and only reveal for the instances that we select as part of the query selection strategies. The provided labels are used to train our supervised sentence classifier. For both language pairs we also held out a 2000 sentence test set prepared by drawing 1000 sentences each from both the domains. We compare our active selection strategy to a random baseline, where the sentences for training the classifier were drawn randomly from the unlabeled data. Figure 5.7 shows the performance on Spanish-English dataset and Figure 5.8 shows the performance on Haitian Creole-English dataset.

The plots show accuracy curves for the sentence classifier with the y-axis showing accuracy on the 2000 sentence test set and the x-axis showing the number of training samples drawn from the unlabeled data. The active selection outperforms a strong random baseline in a significant manner. Observing the graph for Spanish-English, we note that while the random selection strategy requires 2000 sentences to reach 78% accuracy on the test set, the active sampling approach reaches the same accuracy with only 600 sentences, i.e a 75% reduction in the amount of domain tagged training data. Although similar observations can be made on the Haitian Creole-English graph, given the sparse nature of the dataset, the margins obtained by active learning are relatively smaller, still significant.

5.7.3 Multiple Annotation Active Learning and Translation Performance

Does categorization accuracy influence translation? : Oracle Experiment

We then ran experiments for building a focused domain translation system. We wanted to understand the effect of sentence categorization quality on the active selection strategy. In other words, how well does the classification accuracy transfer into translation accuracy or how much does noise effect the AL strategy and resulting MT system.

We test it on the Spanish-English data discussed above and perform end-to-end translation experiments. The evaluation setup is similar to the sentence selection evaluation setup in chapter 3. Where as for the sentence selection task we had in-domain unlabeled monolingual data U to choose sentences from, here we use the U_d sentences tagged by the sentence classifier as being in travel domain. We also use a step function for the posterior distribution $P(D = d|s)$ and therefore every sentence tagged by the classifier is used in the sentence selection phase with uniform weight.

Figure 5.9 shows the results from the oracle experiment on Spanish-English language pair. We are interested in building a travel domain system and so the ideal scenario is where we have access to the entire BTEC data tagged as belonging to travel domain (InDomain

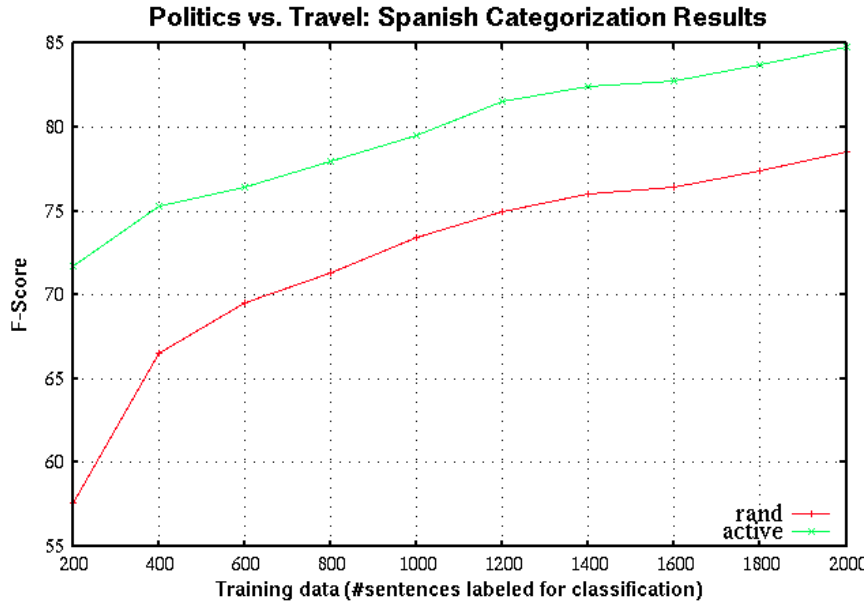

Figure 5.7: Spanish Travel data classification task from mixed domain data (Travel + Politics) with classifier performance on y-axis and # sentences labeled on x-axis. Active selection of sentences (active) for class label annotation outperforms a random selection baseline

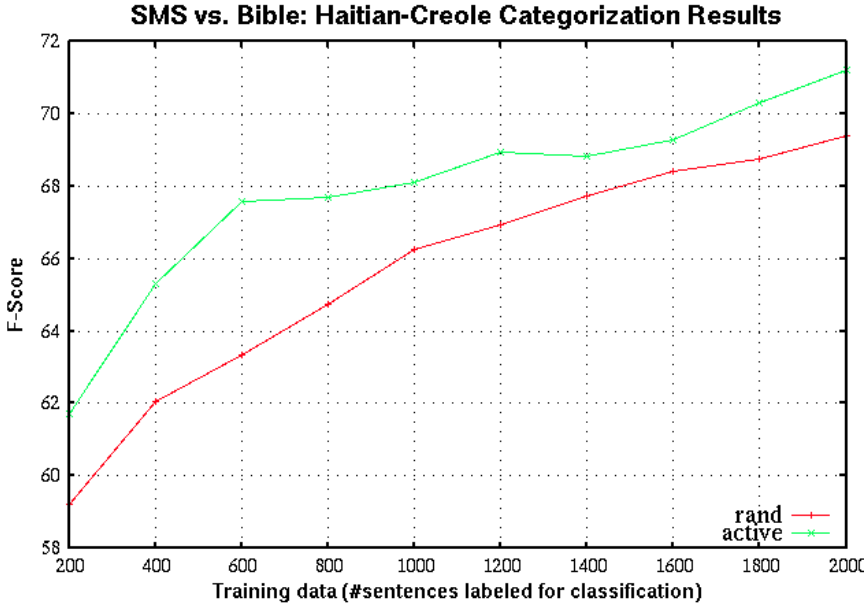

Figure 5.8: Haitian Creole SMS classification task from mixed domain data (SMS + Bible) with classifier performance on y-axis and # sentences labeled on x-axis. Active selection of sentences (active) for class label annotation outperforms a random selection baseline

Corpus). We also have a baseline where we do not have an extra categorization phase and pass the mixed domain corpus to the active sentence translation phase (Mixed Corpus). We also tried two points on the accuracy curves for the first task where we switch between the two tasks. 'SwitchA' curve corresponds to switching at an accuracy level of 75% and the second 'SwitchB' corresponds to when the curve starts to plateau which in this case was at about 87% accuracy level. As expected the best results are observed in the 'InDomain' curve, where there was zero noise in the data. However, in reality such quality can only be achieved by marking all the sentences by a human which is very expensive. We show that using our two-stage active learning strategy we do much better than the baseline that was trained on the wild corpus. We also approach the best case scenario by automatically obtaining domain information for a very small portion of the MT training data.

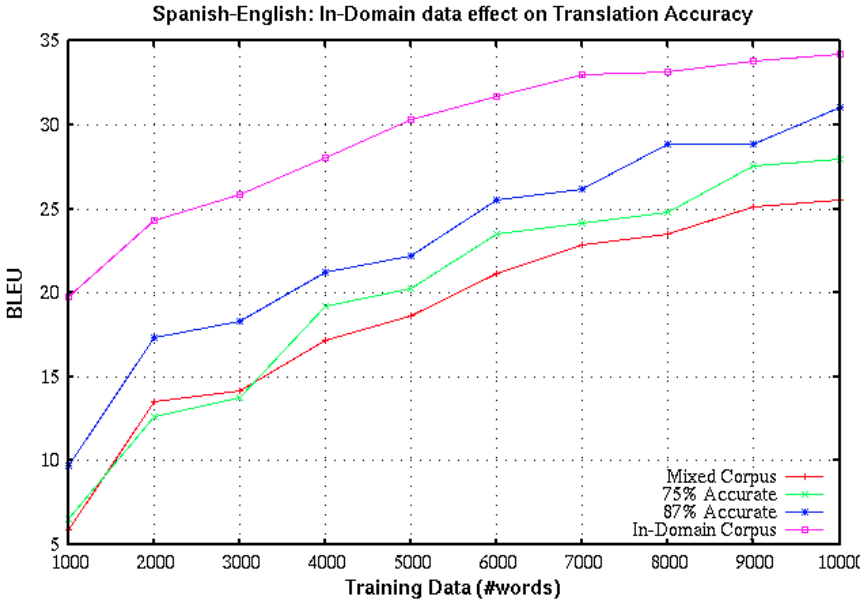

Figure 5.9: Performance curves of the MT system where the parallel data it was trained on was created by actively selecting and translating sentences from a pool of monolingual in-domain data, categorized by a text-classifier. The accuracy of the classifier has an effect on the sentences translated and hence the final MT system quality, in this case a Spanish-English MT system for Travel domain. Complete in-domain data is the best achievable result, but an 87% accurate classifier already provides good support for building a travel domain MT system

Results

We then ran real-world experiments for building a focused domain translation system for two language pairs. We perform our experiments on both Spanish-English and Haitian Creole-English language pairs. The translation framework we use is Moses Koehn et al. [2007]. We use the standard Moses pipeline for extraction, training and tuning our system. We built an SRILM language model using English-side of the Europarl parallel corpus distribution which consists of 1.6 million sentences or 300 million words. While experimenting with varying data sets we do not vary the language model. The weights of the different translation features were tuned using standard MERT Och [2003] techniques released as part of Moses. For Spanish-English language pair the development set for tuning consisted of 500-sentences and the test set used was released by BTEC for the IWSLT task of 343 sentences. In case of Haitian Creole-English language pair we used the development and test sets released by WMT 2011 for the shared task. We evaluate the performance of the end-to-end systems using BLEU. In the case of Spanish-English we are interested in building a travel domain translation system. Similarly in the Haitian Creole-English case we are building a SMS domain translation system. The ideal scenarios in both cases is to have access to a text classifier system that is 100% accurate about its judgments but given that we are building such a classifier along with training a translation system, we will evaluate on how well we perform under a given budget for obtaining annotations for both the tasks together.

The baselines we report are single-sided selection for text classifier (a1), single-sided selection for translation (a2), and a joint selection for both the classifier and translation (a1a2). In all the baselines, irrespective of the selection strategy, we annotate the instance for both tasks. The iterative active learning setup is as follows: we first select a batch of size 'N' sentences each according to each selection strategy and obtain annotations. We then re-train both the tasks - classifier and translation system by adding the data back into their respective training corpora. We use a batch size of N=500 for Haitian-Creole and English system, and a batch size of N=500 for the Spanish-English system in order for the performance curves to be more smooth and pronounced across iterations. Finally, for the cross-over strategy (a1a2-crossover), we select a specific task and use a selection strategy for selecting the instance and therefore, we only obtain one type of annotation at a given point for the entire batch. Finally we also report the oracle selection (in-domain) which is equivalent to having access to the accurate classification labels. This is the upper bound on what the system can achieve as all the data is manually domain tagged and the sentence selection for second phase is performed using this pure data.

The learning curves for Spanish-English can be seen in Figure 5.10 and for the Haitian Creole-English can be seen in Figure 5.11. Our active learning strategy for multiple annotations,'cross-over', switches from eliciting annotations for first task to second. For Spanish-English curve this happens at an accuracy level of 87% for the sentence categorization task and for the Haitian-Creole task we see that the switching point is at 91% accuracy.

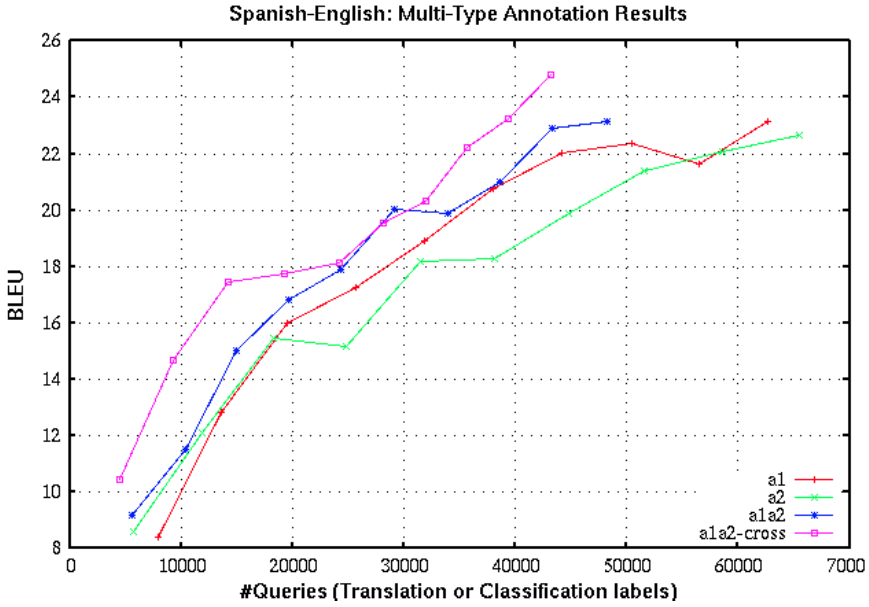

Figure 5.10: Building a Spanish to English Travel domain MT system, starting with politics+travel mixed corpus. Our a1a2-cross approach switches from actively training a text-classifier to actively training a translation system. It outperforms a joint active selection approach (a1a2) that selects instances for both tasks together and also two other baselines that focus only on the individual tasks (a1) and (a2)

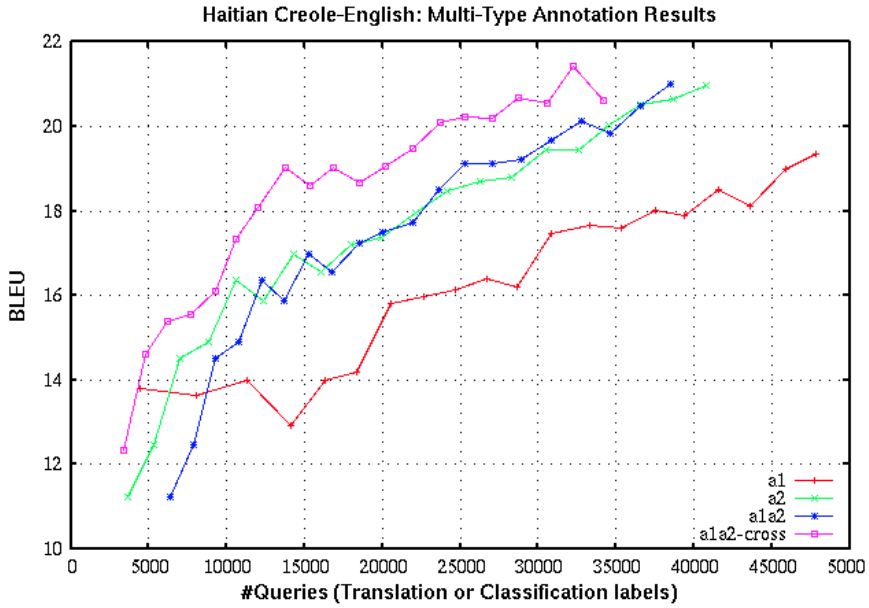

Figure 5.11: Building a Haitian Creole to English SMS translation system, starting with SMS+Bibles mixed corpus. Our a1a2-cross approach switches from actively training a text-classifier to actively training a translation system. It outperforms a joint active selection approach (a1a2) that selects instances for both tasks together and also two other baselines that focus only on the individual tasks (a1) and (a2)

As expected the best results are observed in the 'In-Domain' curve, where there is zero noise in the categorization data. However, in reality such quality can only be achieved by marking all the sentences by a human which is very expensive. We show that using our two-stage active learning strategy we do much better than the baseline that was trained on corpus. In case of Haitian-Creole although the available data is very less, the effect of categorization accuracy still bears improvements on the overall translation performance. However, in this case the gap between in-domain and the others is smaller and so is the room for exploration.

5.8 Summary

In this chapter, we have discussed active learning with multiple annotations in low-resource scenarios. We proposed two different problem setups relevant to building MT systems, namely - comparable corpora classification task and building focused domain translation systems.

For the comparable corpora classification task, we identified two kinds of annotations: class labels (comparable vs. non-parallel) and clean parallel translation segments within the comparable sentences. We implemented multiple independent strategies for obtaining each of these annotations in a cost-effective manner. Our active learning experiments in a simulated low-resource comparable corpora scenario across two language pairs show significant results over random baseline. Finally we also proposed a joint selection strategy that selects a single instance which is beneficial to both the annotations. The results indicate an improvement over single strategy baselines.

For the focused domain translation task, we discussed a pipeline framework involving building of a sentence classification system and a translation system, both of which require annotated data. We then proposed query strategies for switching from one task to another as a viable strategy when involving multiple tasks where the first task influences the effectiveness of the second. We evaluated our approaches on datasets from two different language pairs - Spanish and English, Haitian Creole and English. We showed that our strategies provide a cost-effective way of building domain specific translation systems for these language pairs.

In this thesis we experiment with a cost-sensitive multiple annotation active learning framework enabling a plugin of different cost factors for each annotation. In future, we wish to estimate true real-world cost information using online platforms like Amazon's Mechanical Turk, and plug them into our setup to study the effect of multiple annotations. This will give a more realistic trade-off between the two annotation tasks.

Chapter 6

Crowdsourcing Translation

In all our experiments in Active Learning, we have so far made an implicit assumption that expert annotators are available. We assume that an expert translator is present, and provides high quality annotations. However, for some learning tasks it may be difficult to find experts for annotation. In this chapter we will explore this dimension. More importantly, we will study the 'expert vs. non-expert' annotator problem and propose crowdsourcing as a suitable alternative for eliciting annotation.

The data provided by an expert is of high quality and most desirable for improvement of a translation system. For some problems in NLP like parsing, syntactic annotation, where linguistic expertise is required, there is no substitute for an expert. Translation on the other hand, is a slightly less demanding task, which can perhaps be completed by non-linguists or native speakers of the language, who have some familiarity with a second language. For example, to translate the English sentence "Is there a hotel near by?" into Chinese, we can make do with the availability of a native Chinese speaker who understands English. And to create a translation system for a travel domain, where we will need a few thousands of such English sentences translated into Chinese, we should be able to make use a non-linguist's help. It is this aspect of translation that we want to explore. The following research questions will be pursued:

- How do we collect data from the crowd? What are some of the challenges involved in bringing humans in the loop of building translation systems?
- Our experiment with non-experts highlight the need for quality assurance of crowd data which is an onerous task if done manually. We will explore ways of automatically exploiting overlapping labels from multiple non-experts to improve quality.
- Eliciting data from a large number of non-experts may improve quality, but the associated cost will quickly become prohibitive. We will explore the cost vs. quality

trade-off and devise techniques to elicit high quality data at reduced costs.

6.1 Introduction

Crowdsourcing can be treated as a distributed problem-solving model where tasks that are challenging, difficult or time-consuming for computers are passed to human crowds over the web. These tasks broadly belong to the language or vision community, where for a number of tasks it is still impossible for computers, but only requires a few seconds for a human to complete. For example, identifying a person in a photograph, tagging a video for a particular event, flagging an email for spam, identifying the sentiment of a written text, spotting characters in an image are still some of the challenge research problems to computers. A few variations of the concept of crowdsourcing exists today, and the reader is encouraged to refer to [Law and von Ahn, 2011] for a detailed discussion of the problems and challenges with each format of crowdsourcing.

6.1.1 Crowdsourcing and Amazon Mechanical Turk

Paid crowdsourcing has become popular with support from third-party platforms like Amazon Mechanical Turk, where each task is highly granular and is completed by a person. In our work we use Amazon Mechanical Turk (MTurk) for crowdsourcing. Amazon Mechanical Turk (Mturk) is an online marketplace that enables computer programs to coordinate with humans via crowdsourcing. A typical workflow for data collection on Mturk is depicted in Figure 6.1. Requesters can pose tasks known as HITs (Human Intelligence Task), and workers, also known as turkers, can then browse among existing tasks and complete them for a payment provided by the requester. Payments are processed by their micro-payment system which is available both in dollar (USD) and Indian rupee (INR). In order to ensure quality, MTurk provides screening based on worker parameters like 'completion ratio', 'abandonment rates' and geographic location. As we will see later, these parameters although useful are not sufficient in eliciting quality data from the workers.

Crowdsourcing compensates for the lack of experts with a large pool of expert/non-expert crowd. However, crowdsourcing has thus far been explored in the context of eliciting annotations for a supervised classification task, typically monolingual in nature [Snow et al., 2008]. We test the feasibility of eliciting parallel data for Machine Translation (MT) using Mechanical Turk (MTurk). MT poses an interesting challenge as we require turkers to have understanding/writing skills in both the languages.

Recent efforts in MT include feasibility studies for using crowdsourcing techniques for MT Evaluation; users are provided with translations from multiple systems and asked to select the correct one [Callison-Burch, 2009], [Zaidan and Callison-Burch, 2009]. One

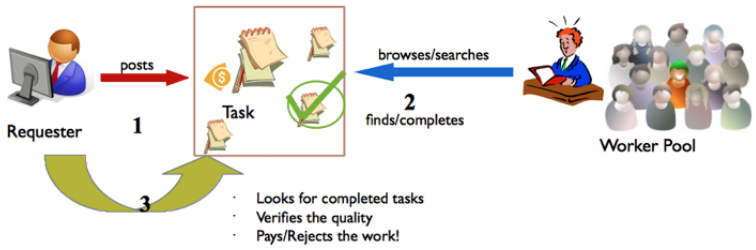

Figure 6.1: Amazon Mechanical Turk Workflow with requesters posting tasks online and turkers selecting and completing tasks for a payment

observation that [Callison-Burch, 2009] make is about the availability of bilingual speakers for annotation tasks in MT. They observe that it is relatively more difficult to find translators for low-resource languages like Urdu, Thai, etc. than it is to find for Chinese, Arabic, Spanish, etc. With the increasing pervasiveness of the Internet, and more and more people in the developing world gaining computer literacy, the situation should ameliorate.

In case of Machine Translation a HIT on Mturk is a task where a turker is provided with one or more sentences in the source language to be translated to a target language. Quality assurance is always a concern with an online crowd that has a mixture of experts and non-experts. Making sure that the workers understand the task is the first step towards quality. We provide detailed instructions on the HIT for both completion of the task and its evaluation. We also set the workers qualification threshold to 85%, which guarantees only those workers who have had a success rate of 85% or above in the past hits.

6.1.2 Language Landscape of MTurk

We first conducted a pilot study for a variety of language pairs in order to probe the reception on MTurk Ambati and Vogel [2010]. Our pilot study helped us calibrate the costs for different language pairs as well as helped us select the languages to pursue further experiments. We then selected 3 language pairs which we explored in greater detail during the course of the project. The language pairs are Telugu-English, Urdu-English and Spanish-English. Ideally, we would like a turker who is native speaker of the language which he/she is translating into. But as we will see in our data analysis, in almost all the cases we observe that is not the case. We sampled 100 sentences for each language-pair and requested three translations for each sentence. The Spanish data was taken from BTEC [Takezawa et al., 2002] corpus, consisting of short sentences in the travel domain. Telugu data was taken from the sports and politics section of a regional newspaper. For Urdu, we used the NIST-Urdu Evaluation 2008 data. We report results in Table 6.1. The goal of the experiment

Language Pair	Cost	#Days	#Turkers
Spanish-English	$0.01	1	16
Telugu-English	$0.02	4	12
Urdu-English	$0.03	2	13
English-Spanish	$0.01	1	19
English-Telugu	$0.02	3	35
English-Urdu	$0.03	2	21

Table 6.1: Statistics from a sentence translation task conducted on Mechanical Turk for various languages, both translating into English and out of English

was to study the task design and feasibility of switching translation direction.

The first batch of HITs were posted to collect translations into English. The ideal target population would be native speakers of English who also understand the source language. We noticed from manual inspection of the quality of translations that most of our translators were non-native speakers of English. This calls for adept and adequate methods for evaluating the translation quality. For example more than 50% of the Spanish-English tasks were completed in India, and in some cases a direct output of automatic translation services.

The second set of experiments were to test the effectiveness of translating out of English. The ideal target population for this task were native speakers of the target language who also understood English. Most participant turkers who provided Urdu and Telugu translations, were from India and USA and were non-native speakers of English. However, one problem with enabling this task was the writing system. Most turkers do not have the tools to create content in their native language. We used 'Google Transliterate' API [1] to enable production of non-English content. This turned out to be an interesting HIT for the turkers, as they were excited to create their native language content. This is evident from the increased number of participant turkers. Manual inspection of translations revealed that this direction resulted in higher quality translations for both Urdu and Telugu and slightly lower quality for Spanish.

6.1.3 Challenges for Crowdsourcing and Machine Translation

Low Quality

Quality assurance of crowd sourced data is necessary due to two reasons. Firstly, there are gamers on the web who would could either trick the task for money and thus introduce

[1] http://www.google.com/transliterate/

random noise into the annotation. Problems like blank annotations, mis-spelling, copy-pasting of input are prevalent, but easy to identify. Turkers who do not understand the task but attempt it anyway are the more difficult ones to identify, but this is to be expected with non-experts. Secondly, annotators available on the web may mis-interpret the task description and could complete the task incorrectly.

Turking Machines

We also have the problem of machines posing as turkers – 'Turking machine' problem. With the availability of online translation systems like Google translate, Yahoo translate (Babelfish) and Babylon, translation tasks on MTurk become easy targets to this problem. Turkers either use automatic scripts to get/post data from automatic MT systems, or make slight modifications to disguise the fact. This defeats the purpose of the task, as the resulting corpus would then be biased towards some existing automatic MT system. It is extremely important to keep gamers in check; not only do they pollute the quality of the crowd data, but their completion of a HIT means it becomes unavailable to genuine turkers who are willing to provide valuable translations. We, therefore, collect translations from existing automatic MT services and use them to match and block submissions from gamers. We rely on some gold-standard to identify genuine matches with automatic translation services.

Output Space

Due to the natural variability in style of turkers, there could be multiple different, but perfectly valid translations for a given sentence. Therefore it is difficult to match translation outputs from two turkers or even with gold standard data. We therefore need a fuzzy matching algorithm to account for lexical choices, synonymy, word ordering and morphological variations. This problem is similar to the task of automatic translation output evaluation and so we use METEOR [Lavie and Agarwal, 2007], an automatic MT evaluation metric for comparing two sentences. METEOR has an internal aligner that matches words in the sentences given and scores them separately based on whether the match was supported by synonymy, exact match or fuzzy match. The scores are then combined to provide a global matching score. If the score is above a threshold δ, we treat the sentences to be equivalent translations of the source sentence. We can set the δ parameter to different values, based on what is acceptable to the application. In our experiments, we set $\delta = 0.7$. We did not choose BLEU scoring metric as it is more oriented towards exact matching and high precision, than towards robust matching for high recall.

Batch	Sentences	Types	#Translations	#Turkers	Cost	#Hours
btec1	1000	5,631	3	71	$0.01	16
btec2	1000	6,810	3	83	$0.01	14

Table 6.2: Completion and cost statistics for translating two batches of Spanish-English data via Crowdsourcing

6.2 Datasets

With experience gained in crowdsourcing translation, we then pursued large-scale parallel data creation in the crowd. We chose the following three language pairs.

6.2.1 Spanish-English

Spanish is a majority language spoken in large parts of the world. We selected two batches of thousand sentences from the travel domain and a third batch of 1000 sentences from the Europarl corpus which was in the political domain. The second batch was only slightly longer than the first batch of sentences and the third batch was the longest of all. Detailed description of datasets can be seen in Table 6.2.

We performed our experiments on the Spanish-English language pair on two batches. In each batch, we selected 1000 Spanish sentences that were crowd-sourced for translation via MTurk. Each sentence was presented to three different turkers for translation. The first batch of tasks were completed by a total of 71 turkers, to provide 3000 translations. This batch consisted of sentences that were less than 7 words. A total of 17 man hours was spent among these turkers. The second batch which consisted of sentences less than 12 words, was completed by 101 turkers. The total cost for both batches was within 100 USD.

6.2.2 Urdu-English

For Urdu to English translation we used three different batches of data as seen in Table 6.3. The first batch (ldc1) was a subset of the training data released by LDC for the NIST 2008 evaluation. The sentences were chosen using a length criteria of 8 to 15 words. The second batch of sentences were the test set released from the NIST 2009 evaluation. This dataset was provided to us by Chris-Callison Burch, and was also collected using Mechanical Turk. Details of the dataset can also be seen in their paper Zaidan and Callison-Burch [2011]. The last dataset was a set of thousand sentences from BBC Urdu News portal. The first two batches of Urdu sentences also had accompanying expert translations in the form of references, but the BBC dataset set did not have any equivalent expert provided translations and so we do not use this in gold-standard evaluations. We also have four translations

Batch	Sentences	Types	#Translations	#Turkers	Cost	#Hours
ldc1	1000	11,022	3	51	$0.03	46
ldc2	1792	39,923	4	51	$0.03	NA
bbc	1000	14,089	3	51	$0.03	34

Table 6.3: Completion and cost statistics for translating three batches of Urdu-English data via Crowdsourcing

Batch	Sentences	Types	#Translations	#Turkers	Cost	#Hours
chanda	1000	13,606	3	143	$0.03	72
news	1000	7,073	3	109	$0.03	68

Table 6.4: Completion and cost statistics for translating two batches of Telugu-English data via Crowdsourcing

provided for the ldc2 sentence set, while the other two have only three translations available from crowdsourcing.

6.2.3 Telugu-English

Datasets collected for Telugu-English can be seen seen in Table 6.4. The first batch of sentences were sampled from a children's magazine called 'Chandamama' and consists primarily of short stories. The second batch of sentences were sampled from a Telugu news daily and were in the length range of 7 to 12 words per sentence. For this language pair there is no parallel corpus available publicly, and so we do not have references for any of the batches. However since I speak this language, it was useful in conducting manual inspection and analysis and draw useful observations. We do not have any equivalent expert provided translations or additional parallel data and so we do not include Telugu-English in our gold-standard evaluations nor do we attempt to build MT systems using these datasets.

6.2.4 Domain and Crowdsourability

One of the observations from all the experiments was that the difficulty of the domain was a factor that decided the success of the crowdsourcing task. A few others have studied the crowdsource-ability of a task for their applications [Eickhoff and de Vries, 2011]. While it is difficult to have a precise understanding of the difficulty of the sentences and words, it can be safe to assume that the longer a task takes to be completed and the number of untranslated words in the final output are an indication of the difficulty of the task. For the Spanish-English datasets we observe that while travel domain was easy to translate in the crowd, Europarl was a difficult domain. Similarly for the Telugu-English datasets,

while short stories domain was interesting for the crowd, the medical domain was harder to translate.

6.3 Quality in Crowd

Quality control is a concern with an online crowd where the expertise of the turkers is unknown. We also notice from the datasets we receive that there exist a lot of consistently poor and noisy translators lacking the necessary minimal expertise. Therefore it is important to not only cleanup the data to get rid of noisy annotations, but also to have an estimate of the quality of the data. These estimates of reliability can help us use the annotation accordingly in down-stream tasks.

Earlier work has addressed the modeling of turker reliability in an expectation-maximization (EM) framework [Raykar et al., 2010, Ipeirotis et al., 2010]. The authors also discuss modeling of turker reliability jointly with task difficulty in a generative framework. Although we break down the problem into two modeling phases, we do not explicitly pose this as an EM-algorithm due to the following reasons. First, since it is a low-resource scenario we will only be collecting less number of translations for each sentences, typically three to five. EM-algorithms may not be required in such data scenarios unless the noisy data has a large number of labels per input available [Ipeirotis et al., 2010]. Secondly, we do not have gold-standard data that we could use in order to execute the maximization phase of EM. And finally, the number of sources or turkers providing data on Mechanical Turk is very large, with a long-tail of turkers providing very few (sometimes one) labels. It is infeasible to accurately estimate the reliability of such a distribution of pool of workers in sparse data scenarios.

In this section we discuss how we plan to use the technique of redundancy effectively to collect such reliability estimates of annotation (translation) and annotator (translator) separately. We emphasize on annotation reliability only using annotator reliability when available. In this section, we discuss our approach to estimating these as well as selection strategies for computing the best available translation from among the multiple crowd translations available per source sentence.

But before we discuss the two methods, we will try to address the two implicit assumptions that we will make in crowdsourcing translation:

- In most crowdsourcing efforts it is advisable to get multiple annotations in the crowd. Is repeated labeling important?
- Given redundant labels for the input, computing agreement among labels is an effective solution to identify quality labels. Is non-agreement or non-overlap with peer annotations indicative of poor quality?

Batch1: btec1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**LPenalty**
First Translation	44.95	7.40	38.65	46.14	57.83	46.31	0.78
Diverse Translation	38.95	6.75	44.23	43.25	55.85	43.44	0.77
Batch2: btec2							
First Translation	37.80	7.03	43.95	48.50	56.74	50.48	0.89
Diverse Translation	32.35	6.62	47.22	46.00	55.62	47.60	0.85

Table 6.5: Spanish-English: Selecting the first available translation or selecting a translation that is most different from all other translations both result in low-quality data in comparison with Gold-Standard expert translations

Batch1: ldc1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**LPenalty**
First Translation	15.91	4.81	72.41	37.47	48.92	38.77	0.87
Diverse Translation	12.09	4.13	75.99	33.24	45.95	34.29	0.82
Batch2: ldc2							
First Translation	17.10	5.59	68.79	42.04	48.71	46.08	1.04
Diverse Translation	9.89	4.18	76.69	32.94	41.22	35.98	0.97

Table 6.6: Urdu-English: Selecting the first available translation or selecting a translation that is most different from all other translations both result in low-quality data in comparison with Gold-Standard expert translations

With the datasets collected for Spanish-English and Urdu-English where we have expert quality reference translations available for the sentences, we tried to empirically address the assumptions made. As can be seen from Table 6.5 and Table 6.6, for both the language pairs we see that selecting the first incoming translation has low scores in comparison with the expert provided gold-standard translation according to a variety of translation evaluation metrics. Interestingly, when we select the translation that is the most divergent, in terms of word coverage, among the multiple translations we see much poor quality numbers for both Urdu and Spanish.

This suggests that redundancy of translations for the input is a good idea and computing majority consensus translation is an effective solution to identify and prune low quality translation. However we would like to do better than these two baselines in our selection strategies.

6.3.1 Annotation Reliability

We use inter-annotator agreement as a metric to compute translation reliability. The assumption here is that the more number of times a sentence is translated or annotated similarly by two or more annotators, the more likely it is to be a correct annotation.

In our experiments, when using exact match between sentences, we notice a relatively low degree of agreement between turkers on a sample of 1000 sentences selected from a Spanish corpus that is translated by three different translators. About 21.1% of the time all three annotators agree with each other, 23.8% only two translators agree, and 55.1% there was no agreement between translators at all.

Given the structural output space of translations, it is a strong assumption to assume that two different translators will provide exact matching translations. We would like to extend the exact matching algorithm to support fuzzy matching between strings that is not robust to variations in spelling or other language phenomenon. Such a fuzzy matching algorithm also needs to be more flexible to accommodate edit-distance or n-gram overlap for matching. In this regard, automatic MT evaluation metrics like BLEU [Papineni et al., 2002] and METEOR [Lavie and Agarwal, 2007] are promising as they try to solve the same issues for automatic translation evaluation.

We choose METEOR [Lavie and Agarwal, 2007] to score matches between crowd translations as METEOR is a robust evaluation metric that correlates well with human evaluations. A challenging task is to perform matching when there could be more than one semantically valid translations for a given sentence and METEOR addresses to some extent using semantic resources like WordNet [Miller, 1995] and a large English paraphrase table [Denkowski and Lavie, 2011]. METEOR is also configurable to customize the matching towards higher recall or higher precision. In our work we select the hyper parameters in METEOR to prefer the match with a higher recall instead of precision.

$$score(s_i, t_i^j) = \frac{\sum_{k=1}^{K} METEOR(t_i^j, t_i^k)}{K}$$

6.3.2 Annotator Reliability

The above approach of seeking multiple annotations from turkers and using inter-annotator agreement works great in accounting for natural variability of translators and reducing occasional human error. However, this is expensive and may not be a viable long-term strategy. We would therefore like to identify reliable translators who are good at the given task of translation. This can help us vary our strategies and amortize the cost in future translations. Reliability of a translator is also useful in selecting a best fit translation for a sentence when there is no agreement between multiple turkers.

Given a worker w_k and a set of translations $T_k = \{t_{kj}\}$ that he/she translated, we estimate reliability based on translations from other users $U_n = \{t_{nj}\}$ as shown in equation below.

$$rel(w_k) = \frac{\sum_{t_j \in T_k} \sum_{n_i \in U} \alpha}{\|T_k\|}$$
$$\alpha = \begin{cases} 1 & t_{kj} \equiv t_{nj} \\ 0 & \end{cases}$$

[Zaidan and Callison-Burch, 2011] model turker reliability using agreement based features and extra demographic information collected for each turker. They also train a classifier with these features and tune weights on gold-standard data to see improvements in the overall reliability estimates. In our work, we only model the reliability of translator based on agreement based features and do not use extra information for individual translators. Given that we do not assume gold-standard data we borrow weights for combining these features from a machine translation evaluation task [Denkowski and Lavie, 2011] and re-apply them for the crowd data evaluation task.

6.3.3 Translation Selection Strategies

To ensure quality of translation output, each translation is requested from multiple turkers, in our case from three different translators. Translation Selection, therefore, is the task of selecting a best fit translation from among multiple translations received from the crowd. We propose three different strategies based on the reliability estimation proposed in the previous section.

Reliable translation

In this selection strategy we prefer the translation that is most reliable among the set of crowd translations for a given sentence, where reliability is measured as defined in previous section. This approach does not use any extra information apart from the set of translations, but one can imagine using other interesting features like Language Model scores on the target sides to ensure grammatical sentences.

$$k* = \arg\max_k score(s_i, t_k) \tag{6.1}$$

Reliable translator

In this selection strategy we select a translation that was provided by the most reliable translator, where reliability is measure on the entire set of translations for the entire batch of sentences.

$$k* = \arg\max_k rel(w_k, t_k) \tag{6.2}$$

We do not require all the turkers to work on all the tasks, but use data available from the tasks completed by each of them. We observe that the reliability estimates for each turker start off as uniform and are refined as they participate and complete more tasks. As we will see later in some of our experiments, this approach works well when we have more data available to estimate a reliable turker from a low-quality one. We also aggregate reliability scores for each turker across multiple batches and so this selection strategy has additional advantage of the global context, unlike the previous approach.

Weighted Majority Voting

While both the previous approaches work well in different scenarios and have their own advantages, we observe two major drawbacks with each of the them.

- While the majority agreement is a good strategy, we observe that in some cases we do not have any agreement among the crowd translations for a sentence. This makes it hard to compute translation reliability for each of the individual translations.

- Due to a long-tail of turkers seen for most tasks on Mechanical Turk, the reliability estimates for the turkers may be over-estimated or under-estimated depending upon the success seen on a single task. Therefore it not prudent to only rely on the translator reliabilities.

We therefore propose using both translation reliability and translator reliability to select the one best translation, so that we can combine the best from both scenarios. We use a naive selection strategy that works well as seen in our results. We select the translation with highest translation reliability and solve ties by preferring translator with highest reliability.

$$k* = \arg\max_k rel(w_k) * score(s_i, t_k) \tag{6.3}$$

Pair	Cost/sen	Days	#Turkers
Spanish-English	$0.01	1	101
Telugu-English	$0.02	4	12
English-Haitian Creole	$0.06	-	6
Urdu-English	$0.03	2	11
Chinese-English	$0.02	1	14

Table 6.7: Statistics from a pilot study conducted to study the reception of translation on Mechanical Turk. Spanish is a language pair that is done faster and cheaper when compared to other languages

6.4 Cost Effective Elicitation

Crowdsourcing has been used to reduce the cost of annotation for NLP [Callison-Burch, 2009, Snow et al., 2008]. However, our experiments in MT have shown that this may not necessarily be the case. We found that at lower pay rates, it is difficult to find a sufficient number of annotators to complete even a single assignment of the task. For example, we could not find turkers to complete the translation of 100 Haitian-Creole sentences into English even after a period of 10 days. Haitian-Creole is spoken by a small population and it seems that only a very small portion of that was on MTurk. Understandably, the recent disaster in Haiti, may also have been a factor in the low activity for this language. For a few other languages pairs, while we could find a few turkers completing the task, the price had to be increased to attract any attention. We selected 100 sentences from each language pair (with length under 10 words) and posted on MTurk to collect translations. Table 6.7 shows the minimum cost at which we could start getting turkers to provide translations and the number of days they took to complete the task. This scenario highlights not only the need to obtain quality translation but also to do so within a limited budget constraint. MTurk has so far been a suppliers' market, and tasks like translation show how one can only get a few turkers making it a demand-driven market.

Therefore, although the approach of seeking multiple translations from turkers works great in accounting for natural variability of translators and reduces human error, it may not be a viable long-term strategy as it increases the cost of annotation. In this section, we first motivate the need for minimizing cost even in a low-cost setting like MTurk and then we discuss the application of two cost minimization approaches for data collection.

From the datasets we receive, we also notice that there are consistently poor and noisy translators. The problem with these translators is that not only do they make the average cost of the task expensive, but they also make it difficult to estimate the quality of translations by skewing the distribution of the majority consensus translation. To test this, we perform an experiment with the Spanish-English language pair, where we collect three

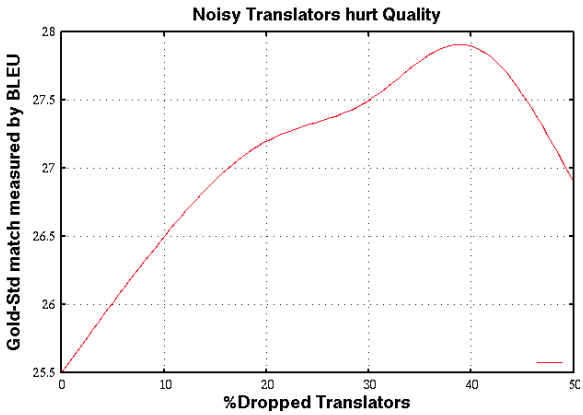

Figure 6.2: Dropping data provided by consistently low quality translators results in a better computation of reliable translation and results in higher translation quality in comparison with Gold-standard expert quality data

batches of 1000 sentences each. We observe that by identifying and rejecting translations from the low-quality translators, the overall quality of our data set measured in terms of match with gold standard improved. We model reliability as discussed in Section 6.3.3 and drop all the translations provided by unreliable translators. We then select randomly from among the remaining translations for each input. In the graph shown in Figure 6.2, we show percentage of total turkers dropped on the x-axis and their agreement with the gold standard on the y-axis. We could weed out about one-fourth of the translations before it starts to affect the quality. In fact there was an initial improvement in quality, possibly due to the removal of poor quality submissions.

6.4.1 Exploration vs. Exploitation

Given that low-quality translators can be dropped without deteriorating the quality, we propose a strategy to drop translators and reduce the total cost of elicitation. The general approach is to identify effective turkers early on (exploration) and get them to do more tasks (exploitation). Typically a gold standard data G can be used to compute this. Consider we have a set of tasks T and set of workers W, and $O \subseteq W$ is a set of workers who are suitable to complete T, and $|G| << |T|$. As a general rule, the exploration vs. exploitation approach is used only when - $(|G| * |O| + |T|) << (|T| * |W|)$, where $|.|$ represents the size of each set. We observe that for language-pairs where only a few translators are available, this approach may not be justified.

Using the exploration phase we identify reliable translators as early as possible. If the translations provided by the worker significantly deviate from the gold standard, we can

suspect incompetency of the worker for the task. In MTurk, there is no provision to direct HITs to a specific turker, but we can attract desired turkers by providing incentives in the form of a bonus. We can also maintain a local log of preferred turkers for our task and favor the annotation from a preferred turker, when the quality is higher than that of others who attempt the same task.

[Donmez et al., 2009] propose an approach called 'IEThresh' to select, at every stage of annotation, the right annotator for annotating an instance. They estimate a confidence interval for the reliability of each annotator and filter out ones that are below a certain threshold. We implement their approach and adapt to the task of translation using fuzzy matching to identify the majority vote. If none of the translators are filtered out, this approach is equivalent to selecting the majority vote translation, which could produce high quality data, but at a high cost. This approach works better when a large number of data points can be obtained for each translator in order to establish a good confidence interval for the initial reliability. Therefore we set a very low threshold on the confidence interval in the initial stages in order to not filter out too many translators. As we obtain a better estimate, the threshold can be made much stricter to select good translators only.

6.4.2 Selective Re-Labeling

Instead of selecting the best translator based on exploration vs. exploitation, we also propose another strategy to selectively request for a translation only when no majority vote exists between already collected translations. This is a greedy approach, inspired by [Sheng et al., 2008], but does not have any information about the turker but only looks at the incoming sequence of translations to decide when to stop obtaining more translations.

For every source sentence s_i, we have a set of translations $T_i = \{t_j, 0 < j < k\}$. In selective re-labeling we start with an empty set and starting obtaining translations from the crowd. We then compute a novelty score for the incoming translation t_k as show below. We stop collecting further translations when the novelty of t_k is below a certain threshold, indicating that there is no further value from a new translation at that point. We add the new translation to the set of translations T_i and move to the the novelty score of the new if the last translation collected t_k agrees with the set of existing translations T_i. We compute agreement score as matches of all n-grams in the new translation with the n-grams in the set of translations T_i.

$$novelty(s_i, t) = \frac{\sum_{j=1}^{k} \delta(t_i^j, t)}{k} \frac{\sum_{x}^{Phrases(t)} \alpha}{\sum_{j=1}^{k} |Phrases(t_i^j)|} \alpha = \begin{cases} 1 & x \notin Phrases(s) \\ 0 & \end{cases} \tag{6.4}$$

6.5 Experiments

We conduct all our experiments on both Spanish-English and Urdu-English language pairs. For this set of experiments our goal is to produce quality translations at cheaper cost in the crowd. We measure quality by comparing these translations with those provided by an expert language translator.

6.5.1 Quality

We evaluate several different selection strategies on these two data sets. For each sentence in the batch, we apply our selection strategy to select one translation from among all the submitted translations for the sentence. We then compare the selected translations with the gold standard data and report results from METEOR [Lavie and Agarwal, 2007] and BLEU [Papineni et al., 2002] automatic metrics. Tables 6.8, 6.9 show results of quality maximization on both btec1 and btec2 batches for Spanish-English and ldc1 and ldc2 batches for Urdu-English respectively.

Our baseline is a random selection strategy, where a translation was selected at random from among the multiple translations. We observe that all our selection strategies work similar to or better than the baseline. 'Maj-Translator' approach always prefers the output of that translator who agrees the most with his peers over the entire batch. Similarly 'Maj-Translation' always prefers the consensus translation, while 'Weighted Voting' is the weighted majority voting approach discussed in Section 6.3.3. We observe that in the first batch of sentences, as they are short we find that the quality of translations were good, with a strong baseline performance. Our selection strategies perform better with weighted majority voting, providing a better selection. In batch 2 however, we find a drop in quality showing that there were low quality translations which skewed the majority vote approach. In such a scenario, using worker reliability estimates works better over pure majority voting.

6.5.2 Cost

Assuming the cost of annotation of each input is uniform and the cost of eliciting a label from the annotator is uniform, we can equate cost to the number of queries or tasks posted for each strategy. As we can not direct HITs at a particular turker on MTurk, we perform these experiments offline on the data collected. We experiment with Explore-Exploit strategy by not requesting multiple translations when we have already received a translation from an explored worker. An explored or tested worker is one who has participated in at least 20 tasks and has a reliability score of 0.5 or above.

By doing so, we reduce the number of queries, but we risk relying upon a turker who only does well initially. IE-Thresholding approach which relaxes this by continuously checking

Batch1: btec1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**LPenalty**
Rand	44.95	7.40	38.65	46.14	57.83	46.31	0.78
Maj-Translation	48.53	7.71	36.30	47.35	58.78	47.63	0.79
Maj-Translator	47.89	7.60	35.56	45.84	57.65	45.89	0.78
Weighted Voting	48.92	7.73	35.78	46.97	58.22	47.18	0.79
Batch2: btec2							
Rand	37.80	7.03	43.95	48.50	56.74	50.48	0.89
Maj-Translation	38.23	7.02	44.06	48.31	56.02	50.58	0.90
Maj-Translator	43.18	7.45	39.78	48.61	56.83	50.16	0.88
Weighted Voting	43.94	7.49	39.16	49.05	57.16	50.57	0.88

Table 6.8: Expert match: Translation selection strategies for obtaining quality data in the crowd on two batches of Spanish-English

for performance as a confidence interval, was tested in a similar manner. We observe that, for low-resource language translation, where number of translators available are less, the Explore-Exploit strategy works better than IEThresh [Donmez et al., 2009] by balancing the number of queries and quality.

Tables 6.10, 6.11 show results of cost reduction on the Spanish-English and Urdu-English language pairs respectively.

6.6 Collaborative Workflow for Crowdsourcing Translation

Crowdsourcing is becoming popular with researchers for cost-effective elicitation of annotations. However, quality of crowd data is a common concern in crowdsourcing approaches to data collection. We will investigate collaborative methods for translation, where the participants are working with the output produced by each other Ambati et al. [2012]. We believe that this setup may reduce the overall effort and improve the quality by reducing cheating and incorrect translations. We will then compare the effectiveness of this 'wiki-style' collaborative translation output to the output from paid crowdsourcing model on Amazon Mechanical Turk, by evaluating how well it agrees with gold-standard expert translations.

From the earlier sections in this chapter it is clear that four main challenges in crowdsourcing translations are - the large output space, low quality turkers, non-availability of turkers and finally cost of elicitation. We will design our workflow trying to address as many challenges as possible.

Batch1: ldc1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**LPenalty**
Rand	15.26	4.77	72.83	37.61	49.28	38.97	0.86
Maj-Translation	15.73	4.78	73.47	37.32	48.50	38.78	0.87
Maj-Translator	17.21	5.07	70.35	40.19	52.08	41.42	0.87
Weighted Voting	18.11	5.27	69.81	41.32	52.49	42.82	0.89
Batch2: ldc2							
Rand	17.10	5.59	68.79	42.04	48.71	46.08	1.04
Maj-Translation	17.29	5.64	68.61	42.63	48.83	46.96	1.06
Maj-Translator	22.55	6.42	62.39	47.67	53.07	52.22	1.07
Weighted Voting	22.74	6.44	62.22	48.05	53.15	52.80	1.08

Table 6.9: Expert match: Translation selection strategies for obtaining quality data in the crowd on two batches of Urdu-English

6.6.1 Our Workflow

In this section we will first discuss the desirable characteristics of a collaborative workflow for translation and then discuss our three-phased collaborative workflow that addresses some of the challenges discussed above. The desired characteristics of our collaborative workflow are three-fold:

- Verifiable: We want to improve the verifiability of crowdsourcing for complex outputs like translation. We want to achieve this by breaking down the complex task into meaningful sub-tasks. For example, while it is difficult for multiple translators to agree upon a sentence translation, consensus can be reached upon when translating at word level, which may in turn be used to check validity of sentence translations.

- Diverse users: We want users of monolingual and bilingual nature to be part of our workflow, as it is relatively easier to find the former. For example, while there are more than a billion Chinese speakers, only a very small portion of them may be able to translate from English into Chinese.

- Work with non-experts: We want our workflow to not only be robust to low quality inputs, but also be able to assist inexpert translators in providing better translations. Bilinguals efficient in translation of entire sentences are few in number, but a major portion of speakers can translate individual words with high accuracy.

Figure 6.3 shows our pipeline collaborative model. In this workflow the translators are working in phases where output from earlier phases can be enhanced in the subsequent phases.

Batch1: btec1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**#Queries**
Rand	44.95	7.40	38.65	46.14	57.83	46.31	1000
Weighted Voting	48.92	7.73	35.78	46.97	58.22	47.18	3000
IE-Thresholding	46.60	7.53	37.30	46.20	58.33	46.30	1447
Explore-Exploit	48.83	7.70	34.63	46.25	58.06	46.21	1950
Batch2: btec2							
Rand	37.80	7.03	43.95	48.50	56.74	50.48	1000
Weighted Voting	43.94	7.49	39.16	49.05	57.16	50.57	3000
IE-Thresholding	35.87	6.87	46.09	46.57	55.14	48.56	1482
Explore-Exploit	41.02	7.24	41.80	47.97	56.01	49.82	1948

Table 6.10: Cost Savings: Strategies for obtaining quality data while saving cost by selective repeated labeling in the crowd on two batches of Spanish-English

Phase 1: Context-sensitive Lexical Coverage

In the first phase we focus on translations at the word/phrase level. We first identify content words in the provided sentence and post hits for collecting their translations. This task can be repeated a large number of times as it is cheaper to translate a single word than an entire sentence. Unlike translation at the sentence level, this has the additional benefit of being able to be verified by a simple lexical comparison.

We designed the hit so that the user can also see the sentence that the word is in and translate the word in-context. We also observe that turkers converge on a translation much faster when required to translate within the context of an input sentence than out-of-context. We also conducted experiments which show that such translations also happen to be of higher quality.

Phase 2: Assistive Translation by Weak Bilinguals

In the second phase, we collect sentence level translations. The turker is required to translate the entire sentence into a target language by preserving the meaning. However, in this phase, we require the translators to use vocabulary gathered from phase 1 in order to translate certain words in the sentence. As part of a post-verification process, we ensure that the turkers indeed use one of the potential translations for the words in the sentence.

We observed that breaking the translation task into two different phases enables us to not only control spammers, but also engage non-expert translators, who may require some guidance in completing a translation. As translating a word is not an expensive task when compared to entire sentence translation, we can repeat the word translation task more

Batch1: ldc1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**#Queries**
Rand	15.26	4.77	72.83	37.61	49.28	38.97	1000
Weighted Voting	18.11	5.27	69.81	41.32	52.49	42.82	3000
IE-Thresholding	14.11	4.53	74.74	36.02	47.74	37.39	1494
Explore-Exploit	15.73	4.78	73.47	37.32	48.50	38.78	2841
Batch2: ldc2							
Rand	17.10	5.59	68.79	42.04	48.71	46.08	1792
Weighted Voting	22.74	6.44	62.22	48.05	53.15	52.80	7168
IE-Thresholding	22.27	6.39	62.64	47.43	52.86	51.99	3318
Explore-Exploit	22.81	6.34	63.62	47.92	52.22	53.17	6290

Table 6.11: Cost Savings: Strategies for obtaining quality data while saving cost by selective repeated labeling in the crowd on two batches of Urdu-English

number of times until reach a sufficient level of inter-translator agreement and only then, proceed to phase 2.

Phase 3: Target Synthesis by Monolingual Speakers

In the final phase we do not require bilingual speakers, but only monolingual speakers of the target language. The task in this phase is to construct a new translation by synthesizing a translation from among the multiple translations produced in phase 2. We also allow for post-editing of the translation for spelling and grammar errors.

For example, consider the multiple translations for the Spanish sentence below. We observe that typically there are missing words, mis-spelt words, non-translated words and incorrect grammatical usage. We also notice that while there is no evidently better translation among the multiple translations, a meaningful and complete translation can be synthesized from them. This is similar in spirit to multi-engine machine translation Nirenburg and Frederking [1994].

Spanish: lo tomar desde la parada de taxis

- i'll climb it from the taxi stop
- i'll take it from the taxi rank
- i will have it from the taxi rank

In this phase, a turker is only shown the three translations and is required to guess the correct translation. In cases where one of the translations is actually the perfect translation,

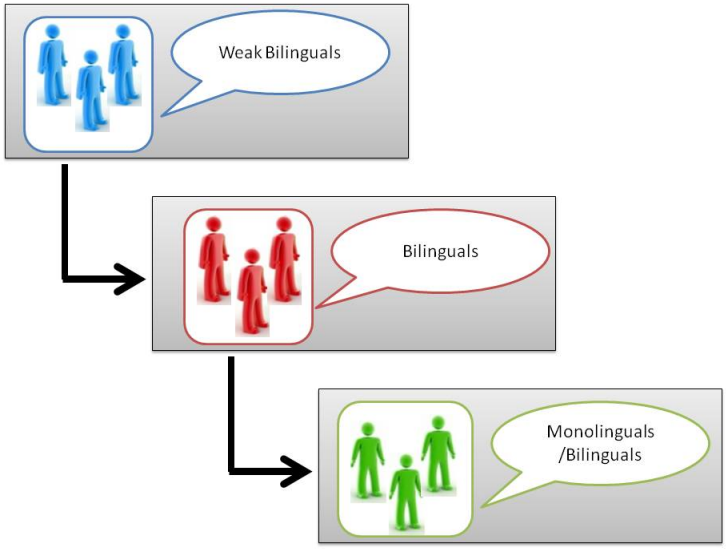

Figure 6.3: Our three phased collaborative workflow for translating in the crowd. The three stage benefits by involving bilingual and monolingual speakers for completion of the translation

the user can select the translation as the correct one. The redundancy among the translations gives sufficient evidence to the monolingual speaker to guess the correct intent of the source sentence, although he does not speak the language. Once the context and meaning of the sentence is understood, it is easier for the turker to synthesize a new translation from the alternatives. One can also obtain multiple translations in this phase, although we observe that usually a single solution is sufficient as the monolingual speakers do a good job of constructing the right sentence from the alternatives.

6.6.2 Evaluation

Our baseline is the traditional setup for crowdsourcing translation as described in the introduction section, where sentences were translated by independently working turkers and a majority agreement was conducted to select the best translation. For the baseline, we obtained translations from 5 different turkers and, similar to [Ambati et al., 2010a], use a fuzzy matching algorithm for comparing two sentences and computing majority agreement. The fuzzy matcher has an internal aligner that matches words in the sentences given and scores them separately based on whether the match was supported by the exact match or the fuzzy match. The scores are then combined to provide a global matching score. If the score is above a threshold, we treat the sentences to be equivalent translations of the source

Method	Language	BLEU
Baseline	Telugu-English	21.22
Collaborative	Telugu-English	27.82
Baseline	Telugu-Hindi	18.91
Collaborative	Telugu-Hindi	20.9

Table 6.12: Evaluation of quality under different workflows

sentence.

Results

We obtain translations for two language pairs using our workflow and compare it with the traditional crowdsourcing workflow. Firstly we pick Telugu-English language pair, where the source language is a minority and the target language is a majority language. This represents a scenario where availability of bilingual speakers is scarce, but obtaining users for target language is quite easier. We also try a new language pair Telugu-Hindi, where it is extremely difficult to find experts that speak both the languages, although it is relatively easier to find weak bilingual speakers, who, given enough word level assistance, can perform a decent job of translation. The target language, Hindi, is linguistically a richer language than English, with greater scope for grammatical errors due to gender, number inflections on nouns and verbs. For both language-pairs we translated 100 sentences each using both workflows and compared with the available gold-standard translations that were obtained from experts. As shown in Table 6.12, we compare the gold-standard match using automatic translation evaluation metric: BLEU Papineni et al. [2002].

From Table 6.12, we can see that the collaborative workflow proposed in this chapter performs much better for obtaining translations in a crowdsourcing paradigm, when compared to a traditional crowdsourcing setup of farming the task to multiple independent turkers. We also observe that our collaborative workflow enables quicker turn-around time for translations as it fosters participation of weak-bilinguals and monolingual speakers, which is a greater portion of the population than pure bilingual speakers.

6.7 Summary

Crowdsourcing is becoming popular with researchers for cost-effective elicitation of annotations. However, quality of crowd data is a common concern in crowdsourcing approaches to data collection. When working with crowd data, the objectives are two-fold - maximizing the quality of data from non-experts, and minimizing the cost of annotation by pruning noisy

annotators. In this section we addressed the above two aspects in the context of creating parallel corpora for building automatic MT systems. We proposed selection techniques by explicitly modeling annotator reliability based on agreement with other turker submissions. Crowdsourcing is a cost-effective solution. However, in order to improve quality, the same task is completed by multiple users, which may not be cost effective as a long-term strategy. We propose a novel technique inspired by exploration vs. exploitation idea to reduce cost further. We conducted experiments in two language pairs: Spanish-English and Urdu-English and showed effectiveness of our approaches for crowdsourcing translation.

Finally we introduced a novel collaborative workflow for language translation. Our three-phase workflow involves breaking the atomic task of sentence translations into three stages - word translation, assisted sentence translation and translation synthesis. We showed that collecting translations using our collaborative workflow has several advantages over the traditional crowdsourcing approach of independently obtaining multiple translations. We evaluated our approach on two language pairs and showed that the overall quality of translations has improved at lowered costs.

Chapter 7

Active Crowd Translation

As part of this thesis we have explored two major research directions - Active Learning and Crowdsourcing. We believe the these areas are important for the benefit of low-resource language translation. However, integrating the two components under a tidy setup demands re-usable frameworks and poses interesting challenges. In this chapter we discuss our active crowd translation framework

7.1 Active Crowd Translation Framework

Given infinite resources and time we could embark on a continuous mission of a MT system that improves every day. But, unfortunately we can not do this for any language-pair let alone low-resource language-pairs. Every MT system and project is ultimately limited by a fixed budget and effort levels. Cost therefore becomes an integral part of our utility functions in active learning Arora et al. [2009]. This poses interesting challenges for integration in our project.

We propose and implement an end-to-end framework for training in a low-resource scenario Ambati et al. [2010a]. We call this 'Active Crowd Translation' (ACT). In our ACT framework,as seen in Figure 7.1, the ACT module first selects sentences from a huge monolingual source corpus, which it thinks are the most informative sentences to be translated next. The sentences are then posted to a crowd-sourcing online marketplace like Mechanical Turk, where multiple translators can translate a sentence. Translations from the crowd are then compared with each other and possibly with external sources for quality assurance. Best fit translations are then selected from the multiple translations. The MT system is re-trained with this new parallel data. Feedback from the MT system then drives further iterations of the active data collection. The key component of the system is a 'Crowd' of non-experts and experts actively participating in the translation of sentences as deemed

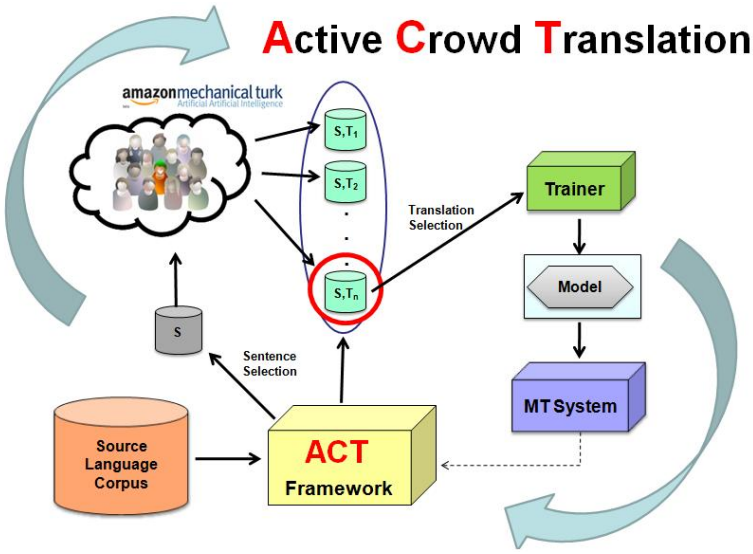

Figure 7.1: ACT Framework: A cost-effective way of building MT systems by combining Active learning for selecting informative instances to annotate and Crowdsourcing for reaching out to non-experts for low-cost annotation

useful by an 'Active Learning' module.

In a traditional Machine Translation scenario, an expert is provided with a defined set of source language sentences which are then translated into the target language. These sentences are not provided in any natural order, and the data is then used to train an MT system. The expert uses his linguistic expertise in both the languages to ensure the quality of the corpus. In the ACT framework we use the knowledge of the crowd in an active manner to build the entire system. This thesis shows that it is indeed possible to build a preliminary translation system using the crowd in a cost effective manner.

7.2 Crowd Data and Building MT Systems

One of the crucial components in the ACT framework is the manner is which the translation selection is done to build parallel data that can be used to train MT systems. In this section we discuss some of the strategies for this parallel data creation and argue its effect on the downstream MT building process.

7.2.1 Select-Best

In the previous section on crowdsourcing we discussed several different ways to select the most reliable translation from the set of translations provided by the crowd. We propose to use the best performing strategy of weighted majority voting and use the translation in building the translation system.

7.2.2 Select-All

Translations from humans vary considerably. We observe that the same sentence can be translated into different meaning-preserving variations, and there is not a consensus in the word choice or the ordering of the sentence in the target language. As an example consider the translations provided by three turkers.

Spanish: Me alegra mucho que hayas podido venir

- I am so glad you could come
- I 'm very happy you were able to come
- I am glad you could make it

We notice that all the translations are valid and yet there can not be a 'majority translation' by any fuzzy match. Such sentences are quite helpful to capture the paraphrasing phenomenon in MT that has proven to be useful [Callison-Burch et al., 2006]. We observe that this does occur quite often in our experiments.

Our approach is to use all the translations to build a translation model. We assume that the maximum likelihood training of the statistical phrase translation models would accentuate the good fragments of the translations and demote the less agreed parts of the translations. In case of good translations that are just paraphrases, the availability of alternatives at decoding time will ensure better coverage as well.

7.2.3 Weighted Select-All

In the previous section we also discuss methods to compute reliability of a translation and we do not want to lose this information when combining all the translations from the crowd together. Therefore, we also tried weighting the translations by their reliability score as computed from the previous section.

One of the weighting approaches we tried was to exploit the GIZA++ framework to weight individual sentence pairs during training of word alignment. This can be achieved by

modifying the "*.snt" files produced as part of the GIZA++ framework. The way GIZA++ uses the weights of individual sentences is to scale the co-occurrence counts of word pairs during the EM algorithm phase of word alignment, the counts are scaled based on the reliability of the sentence.

This has a very small impact on the overall translation quality, as the effect of weighting multiple translations only influences the word-alignment and the lexical probabilities and does not significantly affect the coverage of the phrase table produced. We therefore use a simpler approach that carries all the way to the phrase-table creation which is to duplicate the top-best translation. A second approach can be to re-duplicate better translations multiple times when compared to the low-quality translations.

7.3 Experiments

7.3.1 Spanish-English

We perform our experiments on the Spanish-English language pair. We use the standard Moses pipeline [Koehn et al., 2007] for extraction, training and tuning our system. We built an SRILM language model using English-side of the Europarl corpus, consisting of 1.6M sentences or 300M words. While experimenting with data sets of varying size, we do not vary the language model. The weights of the different translation features were tuned using standard MERT [Och, 2003]. Our development set consists of 343 sentences. The test set used consists of 500 sentences. Both these are from the IWSLT04 workshop data sets. We evaluate performance as measured by BLEU [Papineni et al., 2002] and other automatic translation metrics on a held out dataset test set. We also conduct end to end MT experiments in order to evaluate the effectiveness of our translation selection strategies in producing high quality parallel corpus using crowd data.

We start with the first 1000 sentences from Spanish-English BTEC corpus that produces a BLEU score of 10.64 when tuned and tested on the test set. We create different versions of parallel corpora by using our selection strategies on the crowd translated short sentences. We then train and test an MT system on each of the parallel corpora versions separately. In Table 7.1 we show translation results as scored by several automatic MT evaluation metrics. We only compare our best performing selection strategies. Two main observations are that, selection strategies that explicitly model reliability of workers based on agreement, in fact produce good quality translation systems that mimic systems trained on expert data. A surprising result in our experiments is when we use all the human data with multiple translations for each sentence, we outperform the results from expert data. This could be due to the paraphrases in the non-expert translation.

Batch1: btec1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**LPenalty**
Rand	18.98	4.42	58.48	48.38	57.10	52.49	0.94
Weighted Voting	20.39	4.56	57.10	49.35	57.86	53.49	0.94
ALL	21.64	4.68	55.82	50.81	59.44	54.95	0.94
Weighted-ALL	20.95	4.62	56.57	50.19	58.86	54.30	0.94
Expert	19.93	4.38	58.62	47.92	55.15	52.54	0.97
Batch2: btec2							
Rand	15.92	4.06	60.67	45.92	54.43	50.67	0.95
Weighted Voting	16.43	4.15	60.46	47.05	55.23	51.99	0.96
ALL	18.14	4.32	58.48	47.99	56.63	52.64	0.95
Weighted-ALL	18.06	4.32	58.69	48.16	56.66	52.93	0.95
Expert	16.81	4.13	60.32	46.35	54.51	51.11	0.95

Table 7.1: Expert match: Translation selection strategies for obtaining quality data in the crowd on two batches of Spanish-English

7.3.2 Urdu-English

We also tried a true low resource language pair: Urdu-English. Language Data Consortium has recently released parallel data for the Urdu-English Machine Translation track at the NIST Evaluation 2008.

We used this dataset to conduct our active learning simulation for Urdu-English. We use the Moses Koehn et al. [2007] translation framework for the MT system. The language model used was the English side of the Europarl parliamentary dataset used for training the Spanish-English MT system in the previous section.System tuning was done with minimum error-rate training on a subset of 450 sentences, selected from the NIST DEV08 data set with one reference translations available. We use the rest of the 450 sentences from the NIST DEV'08 for the test set. Our post-processing included a script to reattach as much punctuation as possible to the preceding or following word. Ambiguous marks, such as straight double quotes, were left as separate tokens.

Table 7.2 shows results from training an MT system on the Urdu-English language pair. The first thing to observe is that the scale of the translation scores are much lower considered to the Spanish-English language pair above. This is to be expected as the Spanish-English is a limited domain experiment, travel domain in this case, and therefore has high coverage for the dev and test sets. In the case of Urdu, we are working in the news domain which is very open and a low-resource scenario assumption hurts the coverage, thereby affecting the translation quality. However, we notice that even for this language pair translation selection provides better quality parallel data and relatively higher translation results when compared

Batch1: ldc1							
Strategy	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**LPenalty**
Rand	2.20	2.53	92.59	31.35	35.62	38.88	1.09
Weighted Voting	3.10	2.72	91.02	33.38	37.74	41.15	1.09
ALL	4.53	3.00	88.16	35.36	40.10	43.12	1.08
Weighted-ALL	4.02	2.90	88.96	34.54	39.32	42.25	1.08
Expert	2.91	2.46	92.68	30.63	34.73	38.08	1.09
Batch2: ldc2							
Rand	2.87	2.45	95.00	29.09	32.55	36.29	1.13
Weighted Voting	3.54	2.63	94.35	30.72	34.16	38.17	1.12
ALL	3.94	2.76	93.55	32.10	35.70	39.77	1.12
Weighted-ALL	3.79	2.73	93.89	31.89	35.36	39.59	1.12
Expert	3.09	2.53	97.54	29.43	32.33	37.07	1.15

Table 7.2: Expert match: Translation selection strategies for obtaining quality data in the crowd on two batches of Urdu-English

to random selection. Using all the translations from the crowd produces better MT system overall, indicating that the benefit of coverage and recall overweights the disadvantage due to the noise in translations.

7.4 Analysis

7.4.1 Operating Ranges: Does Crowd Data Help MT System Initially?

From the experiments above we see that for building an MT system translation selection approaches do not work as well as using all the available translations. The natural question following that is- ”Is this true for the entire evolution of the MT system?”. When the MT system does not have any data initially, we would expect any data to provide a better performance over no-data. However as we start collecting more data the margin for improvement using low-quality data is very low.

We conduct experiment to answer this issue by evaluating the translation performance of an MT system that is trained at various points in the evolution process. We start with varying seed for parallel data and add a 1000 sentence crowd parallel data and re-train the MT system and test its performance. collected from the data at various to measure the difference in benefit from the crowd. Figure 7.2 shows performance of using a translation selection strategy (wvote) vs. using all the crowd translations to train a Spanish-English translation system with varying seed sizes : 0k, 1k, 5k and 10k parallel data. Similarly

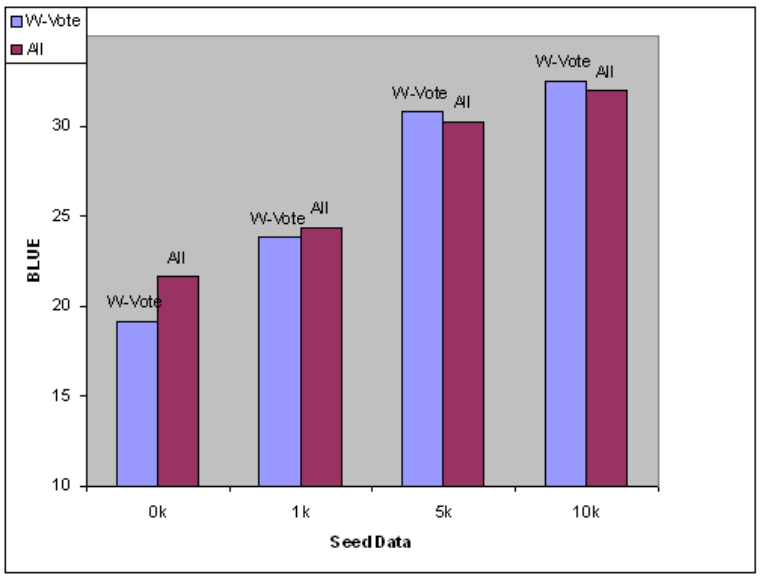

Figure 7.2: Crowd data impact at different operating ranges of a Spanish-English MT System. For each operating range, we use varying amounts of seed data and add additional data collected from the crowd to re-train the system

From the results it is clear that the benefit of using all non-expert translations is more pronounced in the early stages of an MT system where the lexical coverage of the system is very poor. As the data collection progresses and we have a reliable translation model, we see that more than recall/coverage it is important to have quality data from the crowd. This holds true for the Spanish-English language pair, although we do not test the phenomenon for Urdu-English language pair due to the very low performance.

7.4.2 Training a Full Urdu-English System

Finally, we experiment with Urdu-English case to see if new data can be collected via crowdsourcing methods to augment the complete available data for a low-resource language-pair. We select Urdu-English language-pair for which is an example of a low-resource scenario. In this section, we describe how we train a complete SMT system for Urdu-English utilizing the very little available parallel data and then show how we enhance it by additional data obtained from the crowd.

Language Data Consortium has recently released parallel data for the Urdu-English Machine Translation track at the NIST Evaluation 2008. It consists of web data segments accumulated from multiple news websites and other manually created sentences. The data released by NIST consists of 92K parallel segments which we re-segmented to obtain 62K parallel sentence pairs. We notice that some sentences in Urdu are not translations of the English counterpart. So these segments need to be further divided into sentences on both the source and target which are mappable translations of each other. Our Urdu-English dataset consisting of 95K sentential segments on both sides, which are re-segmented and sentence aligned to obtain 54K parallel sentence pairs. We also observe that out of 35170 Urdu types in the lexicon, 6209 vocabulary types were marked to be transliteration of some English word. For preprocessing, we used the tokenization scripts provided by AFRL team for both English and Urdu sides. From the total set of constrained data, we removed parallel sentences where either the Urdu or the English side was blank, where either side had more than 100 tokens, and where either side was longer than the other by a factor of more than 5:1. We use this dataset to train a Moses system for Urdu-English. The language model used is the English side of the Europarl parliamentary dataset used for training the Spanish-English MT system. System tuning was done with minimum error-rate training on a subset of 450 sentences, selected from the NIST DEV08 data set with one reference translations available. We use the rest of the 450 sentences from the NIST DEV'08 for the test set. These datasets belong to a Newswire domain.

We then augment these parallel sentences with the three batches of data as described in Table 6.3 to create 58K sentence pairs. As shown in Table 7.4.2, we observe modest improvement in the automatic metric scores all across the board, indicating improvements from crowd data.

Training with All Available Data							
System	**BLEU**	**NIST**	**TER**	**METEOR**	**Precision**	**Recall**	**LPenalty**
All Data	11.53	4.82	77.78	50.28	54.73	60.22	1.10
All Data+ Crowd Data	11.71	4.82	77.46	50.33	54.61	60.26	1.11

Table 7.3: Results from training an Urdu-English MT system with all publicly available data, and then improving it with about 3k new data collected from crowdsourcing

7.5 Summary

In recent years, corpus based approaches to machine translation have become predominant. Success of these approaches depends on the availability of parallel corpora. In this section we discussed Active Crowd Translation (ACT), a new paradigm where active learning and crowd-sourcing come together to enable automatic translation for low-resource language pairs. Active learning aims at reducing cost of label acquisition by prioritizing the most informative data for annotation, while crowd-sourcing reduces cost by using the power of the crowds to make do for the lack of expensive language experts. We experimented and compared our active learning strategies with strong baselines and show significant improvements in translation quality even with less data. Similarly, our experiments with crowdsourcing on Mechanical Turk have shown that it is possible to create parallel corpora using non-experts and with sufficient quality assurance, a translation system that is trained using this corpus approaches one trained using expert data.

Chapter 8

Conclusions and Contributions

Building Machine Translation systems for low-resource languages has a noble intention, useful for many. Corpus based approaches to automatic translation like Example Based and Statistical Machine Translation systems are the state-of-the-art. However, these approaches require significantly large amounts of resources, primarily in the form of parallel data to train mathematical models. This thesis covers new approaches to building language translation systems for low-resource languages, where the amount of parallel data available is limited and the paucity of language experts is a discernible problem.

Low density resource scenario requires effective usage of resources - money and people to collect data. This thesis has explored Active Learning strategies to collect relevant and informative data for several subtasks in MT. We have also focused on getting these annotations fast, cheap and effectively from non-experts online via crowdsourcing. We will now summarize the conclusions and observations made in the three major directions of the thesis - active learning, multi-type annotation active learning and crowdsourcing.

8.1 Active Learning for MT

- We proposed active learning strategies for the sentence selection task. We formulate strategies as either 'data-driven' or 'model-driven' and show that our approaches for prioritizing monolingual sentences lead to faster improvement of MT systems.

- We experimented with multiple languages including Spanish-English, Japanese-English and Urdu-English and show improvement across the board.

- We also proposed a hybrid ensemble technique that effectively switches between active learning strategies for different operating ranges in the evolution of an MT

system. We show that the hybrid approach work better than the best performing active learning approach for Spanish-English language-pair.

- We are also the first to identify the problem of alignment link selection and proposed several active learning strategies for eliciting alignment links that are informative and need manual attention for improvement of a semi-supervised word alignment model.
- We apply our techniques to two language pairs: Chinese-English and Arabic-English and show significant reduction of alignment error rate even with elicitation of manual links that are fewer than when compared to strong baselines.

8.2 Multi-Type Annotation Active Learning

- We identify the comparable corpora classification task and propose active learning extensions to reduce cost of building such classifiers for low-resource language-pairs. We show improvement for both Urdu-English and Spanish-English language-pairs.
- We also propose improving a comparable corpora classifier via multiple annotations - providing a class label vs. parallel segment extraction. We then extended the traditional single annotation focused instance selection strategies to a joint instance selection strategy that focuses on both these annotations. We show improvements for both the language pairs over the best single-annotation focused active selection strategies.
- We showed improvement in building focused domain MT systems by first formulating it as a combination of sentence classification and sentence translation tasks. We then experiment several query selection strategies and show cost reduction for both the tasks, individually.
- We then show that a joint instance selection approach can further reduce the cost of eliciting annotated data for both the tasks. We experiment with building a focused MT system for both Spanish-English and Haitian Creole-English language-pairs.
- Finally, we also show that a selective approach for choosing and annotating data for one task at a time shows significant improvement in the quality of a focused domain MT system.

8.3 Crowdsourcing for MT

- We have designed the translation task to collect data via crowdsourcing. We identify challenges in conducting translation on Amazon's Mechanical Turk and explore

solutions by collecting data for a variety of language-pairs.

- We design and experiment with algorithms to select high quality data from among the crowd generated data that is typically a mix of non-expert translators.
- We observed that cost can be an important factor for low-resource translation and proposed techniques for minimizing the cost of data collection, by reducing the redundancy in repeated labeling where possible.
- We also devise a collaborative three-stage work-flow for crowdsourcing translation, with a goal of improving overall quality of translation. The approach produces significant improvement in quality over traditional crowdsourcing when evaluated against expert produced translations.
- Finally, we are the first to show that active learning and crowdsourcing can be combined seamlessly to obtain significant cost reductions for building MT systems for low-resource language-pairs.
- We also run experiments for end-to-end active crowd translation using our ACT framework for Spanish-English and Urdu-English language-pairs. We also experiment with various techniques for using the crowdsourced data for building and MT system. Our results from Spanish-English language pair show that with effective methods we can build an MT system that approaches the quality of expert translation

8.3.1 Contributions

The major contribution of this thesis is the development and application of a new framework for building machine translation systems for low-resource languages. We call this Active Crowd Translation (ACT). Along the path, we also make the following contributions:

- Improvement of active learning algorithms for the parallel data creation task with significant improvement on low-resource languages.
- Designing active learning setup for the task of word-alignment and implementing strategies that reduced alignment error rates with significant cost reduction.
- Application of active learning to building a comparable corpora classifier and extending the traditional single-annotation driven active learning to select instances for eliciting multiple types of annotations.
- Extension of active learning setup to also jointly select an annotation type and an instance in the context of building domain specific translation systems for low-resource languages where topic classification and translation are two inherent tasks.

- Designing several techniques for quality-effective and cost-effective application of crowdsourcing to the general problem of language translation.
- Implementation of a novel framework called Active Crowd Translation (ACT), that combines active learning and crowdsourcing for building MT systems for different language pairs in low-resource scenarios.

8.4 Broader Impact of Thesis

Machine Translation is needed for any language pair. Many low-resource languages are in danger or extinction, and the ability to translate to and from may help to preserve their active use and thus preserve linguistic and cultural diversity. Low-resource but not-so-rare languages are typically spoken by the economically disadvantaged, and thus beyond the reach of international aid agencies and governmental services. Some low-resource translation projects have immense humanitarian benefit. For instance, projects that build MT systems for African languages in order to aid the rehabilitation of refugees that have cultural and communication problems in the beginning of settlement in the United States. Such projects are very time critical and typically have pre-specified, limited budgets. One can not afford the data entry of a million sentence translations to train high accuracy data-driven systems, neither can we wait for the time taken to summon them. Our work through this project will help build algorithms that address such scenarios. Our algorithms and proposed techniques will help create a framework to provide usable translation systems faster and cheaper by effectively channeling the efforts non-expert bilingual speakers dispersed all over the globe.

Going beyond low-resource languages, even in scenarios where large parallel corpora are available, translation performance is very different or different languages. The results of this project will lead to significant improvements for such languages, where even a few million sentence pairs are insufficient to achieve satisfactory translation quality. The project will therefore contribute to making machine translation technology more usable and more broadly applicable.

Chapter 9

Future Work

This thesis has explored three interesting research directions - active learning, crowdsourcing and machine translation. In this section we will discuss some of the interesting problems that have come up during the course of the thesis and we have either not solved completely in the interest of time, or have solved partially due to their peripheral nature with respect to the thesis. We will also discuss some of the new directions going forward that can build on top of our work already done in this thesis.

9.1 Active Learning and Machine Translation

9.1.1 Model-Based Approaches for Active Learning and SMT

In this thesis we have observed that for the low-resource scenarios data-driven selection strategies like *diversity* and *density* work better than model driven strategies like *uncertainty* and *decoding-score*. Although we observed this phenomena in the initial stages of a translation system where the training data is low, we have not explored all the operating ranges of a translation system, especially the large-data scenario. We hypothesize that in the large data scenarios, when an MT system needs to prioritize the next data to be annotated, model-based approaches would play a better role.

We have identified the challenges in exploring active learning for large scale data scenarios and have some understanding of the solutions. The first issue is the pool-based active learning setup where we require to iteratively train the MT system in order to update the model parameters that are needed by the active learning strategy. In such a setup the cost of re-training the MT system for every new incoming batch of sentences may become infeasible. While small batches provide for smooth performance curves and also error recovery, we can not afford to reduce the batch sizes in view of the increased computational

cost. Some researchers are already exploring methods to allow online updation of MT models with minimal performance loss compared to full re-training Ortiz-Martínez et al. [2010] and we will need to apply similar techniques here.

Another direction to reducing costs of training large data MT scenarios is to explore lower granularity annotations like phrasal translation instead of complete sentence translation. When an MT system is already trained on a large parallel data and has arguably seen most of the contexts in the language, the additional value from a new sentence may be captured succinctly if we identify the novel sub-sentential fragment in the sentence that can be translated without need for a complete translation. Besides phrasal translation also fits well into the paradigm of crowd-sourcing where small tasks can be provided to a lot of translators. Bloodgood and Callison-Burch [2010] have explored translation of phrases for reducing the cost and effort of eliciting translations from a crowd. Such methods can be combined more closely with active learning strategies and we propose to build an overall strategy for deciding when to translate a complete sentence vs. a single phrase at every phase of the MT system.

9.1.2 Syntax-Based Machine Translation and Active Learning

While this thesis has tried to combine active learning and crowdsourcing for Machine Translation, the techniques are general and applicable to any linguistic annotation like - morphology, syntax, search personalization, sentiment analysis etc. However, in some cases the annotation may be complex and not amenable for crowdsourcing. For example, while it may be reasonable to assume crowd being knowledgeable in identifying prefixes and suffixes in a word, category of the word or going further and requesting a treebank-style syntactic parse is far fetched.

Future work in this direction will involve creative ways of breaking down the syntactic annotation tasks into subtasks or questions or game-style surveys , which when answered can be combined to reveal the complex annotation. E.g. A dependency parser needs to know if a link exists between two words. Can we show a human the two words and ask him to construct a sentence using them? Or show substitutability , or show a similar sentence and ask them to rephrase etc? For relative clauses, we can make it a co-reference problem.

In another related work we perform rule extraction from parallel corpora and word alignments [Ambati et al., 2009]. Similar extraction also assumes availability of annotated treebanks for atleast one-side of the parallel corpus. We highlight some of the issues in building syntax-based MT systems and posit that active learning can be applied to such scenarios for reducing effort involved [Ambati and Carbonell, 2009]. Going forward, as semantics starts to be incorporated into current translation systems, we hypothesize that bilingual informants can be used to provide high-level semantic information which is easy for humans but still significantly harder for computers. [Llitjos, 2007, Monson et al., 2008]

have already shown that bilingual informants can be involved in correcting MT translation output which can be used to help refine syntax translation systems and grammar models.

9.2 Crowdsourcing

9.2.1 Computer Supported Collaborative Crowdsourcing

Collaborative content creation activities have recently gained interest both in the real-world and the research community. Interesting studies about the functioning of Wikipedia and the crowds behind it, is an evidence to this effect. In this thesis we have shown that crowd can provide translations that are comparable to expert quality translations. We have also conducted preliminary experiments on how non-experts can be used in an effective workflow that fosters collaboration on their outputs. We believe that there are interesting future directions in taking crowdsourcing to a collaborative endeavor and draw communication channels between crowds that can increase the innovation.

I have started work in this direction by experimenting with the following two tasks and setting them up in a collaborative environment, where they can be studied closely to highlight and necessitate future research.

- *Collaborative translation:* I recently conducted an experiment where I uploaded a document to be translated collaboratively, where collaboration means multiple remotely placed translators conducting translation on the same document, in the same time frame. I also had a professional quality translation for the same against which I could compute the results. The quality of the output from collaborative translation outperforms the regular MTurk approach. I tried the same with three different language - Spanish, Telugu, French, and the results were similar across all of them.

- *Essay Reviewing:* I uploaded five student essays using Etherpad, a collaborative editing environment, and paid users for reviewing (finding at least two or more errors). It was really interesting to see that they were completed quicker than usual and when restricted to workers in the US, the quality was higher than normal. While the broader goal of this work is to create data for an automatic essay reviewer, the fact that collaborative correction broadly enhances the quality by enabling many eyes for correction is interesting.

Research for future will depend on extensive experimentation and data collection for both these tasks. But, even the preliminary data has thrown some interesting questions for future. some interesting questions for research.

- While collaboration enables complementing the skillsets of the participants, it also leads to redundancy where users may ”redo” and ”undo” other’s work. How can computers be integrated seamlessly into such environments to mitigate this?
- One of the challenges in collaboration is the ”entry-point” barrier, where an incoming worker does not know where to contribute (e.g a novice editor of Wikipedia). How can we use active learning strategies to select sub-tasks and highlight them for users?

9.2.2 Task Recommendation in Crowdsourcing

As researchers embrace micro-task markets for eliciting human input, the nature of the posted tasks moves from those requiring simple mechanical labor to requiring specific cognitive skills. On the other hand, increase is seen in the number of such tasks and the user population in micro-task market places requiring better search interfaces for productive user participation. In our recent work [Ambati et al., 2011c], we posit that understanding user skill sets and presenting them with suitable tasks not only maximizes the over quality of the output, but also attempts to maximize the benefit to the user in terms of more successfully completed tasks. We also implemented a recommendation engine for suggesting tasks to users based on implicit modeling of skills and interests. We conducted a preliminary evaluation of our system using publicly available data gathered from a variety of human computation experiments recently conducted on Amazon’s Mechanical Turk. Our results are encouraging and suggest that when users are provided with tasks that are closer to their interests, the completion rate improves. Although, this work is preliminary, the direction is beneficial for improving overall quality in crowdsourcing markets.

9.2.3 New Setups for Crowdsourcing Linguistic Annotations

Crowdsourcing is becoming popular for cost-effective elicitation of annotations. However, in order to obtain large scale annotations, which are often the requirement in building MT systems, paid crowdsourcing model may not be scalable. For instance, in order to obtain few millions of Chinese-English word alignment links from bilingual speakers, it would cost us about a few thousand dollars. Therefore we will investigate and adopt other forms of crowdsourcing like the social games Ahn [2006] and wikipedia style collaborative methods for translation and other related annotations. We believe that this setup may reduce the overall effort and improve the quality by reducing cheating and incorrect translations. We will then compare the effectiveness of our new approaches to the output from paid crowdsourcing models.

We have already started working on a word alignment game (WAG) in the spirit of the Games with a PurposeAhn [2006]. This WAG will display a sentence pair and highlight a word or short phrase in one sentence. The task of the players is to select the corresponding

word(s) in the other sentence. If they agree, both get credit. Using this agreement adds some quality control, as it is far easier to get a match when trying to get the right words, then when clicking randomly at words. This makes the two player mode preferable over a single player mode, in which the system simulates a partner, by relying on previously aligned words. However, if the player appears to be competent, then new words can be provided for alignment ant he answer can be accepted, unless there is strong evidence from the statistical dictionary, automatic alignment and word frequencies that the answer is not correct. To make this game entertaining and engaging, it will be developed in collaboration with CMUs Edutainment department in the form of a class project. We will collect manual alignments for the low resource languages, for which we also solicit translations, but also for t least two languages, for which we have already large parallel corpora, e.g. Chinese and one of the European languages. Our goal for these languages is to get for a 10 million word corpus up to 1 million word alignments. This will provide us with a wide range of unannotated and annotated data to run systematic experiments, in which we can study the impact of manual word alignments on word alignment quality, phrase alignment, and end-to-end MT performance. The backend of this WAG will be connected to the active learning component, which selects sentences and words within these sentences for annotation, and also performs quality estimates for the gamers and the collected annotations.

9.3 Multi-Type Annotation and Multi-Task Active Learning

9.3.1 Multi-Domain Machine Translation

Given unlabeled data sets for multiple domains and a fixed budget to build translation systems for multiple domains, we propose to build multi-domain translation systems simultaneously. As shown in Figure 9.1, we have a set of domains D each with monolingual unlabeled data U_i, where i is an index over the set of domains D. The active learning module will then draw labeled sentences to be annotated by an expert translator. We assume that the translator is also provided with the domain information D_i of the sentence s and therefore he provides a translation t within the context of the chosen domain. The labeled data is now available to train all the individual MT systems, which in turn will be tuned and tested on their respective domain specific development and test sets.

This is a classic case of multitask learning that requires eliciting multiple annotations for the individual MT system building tasks, where providing translation for a domain specific sentence can be treated as a different kind of annotation. Given multiple kinds of annotations, a natural question to ask is when to seek which annotation for what instance. This is a case of multi-task active learning where selection strategies need to consider transfer of knowledge across domains and reduction of annotation effort for a cumulative improvement of all the MT systems simultaneously.

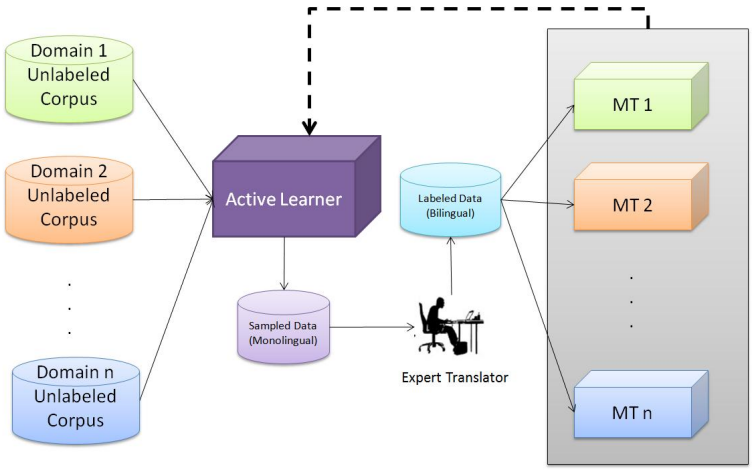

Figure 9.1: Multi-domain MT using Active transfer learning

9.3.2 Active Transfer Learning

A final scenario we wish to study in the context of transfer learning is the standard problem of adapting an existing system to an unseen domain. The goal of active transfer learning in domain adaptation is to maximize the transfer of knowledge from the existing task and optimally acquire new knowledge relevant to the the new task, genre or domain.

Domain adaptation has been studied in the context of NLP for several years now [Blitzer, 2008]. For statistical MT, domain adaptation has been addressed in one of three ways. Firstly, the data is weighted to account for the difference in distributions of input data in the domains or use information retrieval techniques to sample data that is closer to the target domain. A second approach is to interpolate the phrase tables resulting from individual training of two systems on the data from different domains. Special features are introduced to learn the weights for global preferences for each phrase table [Koehn and Schroeder, 2007]. They also outline a lattice based combination of separate hypotheses from the different translation models. A third approach is to preserve the translation models learnt from a source domain, but then influence translations using a language model trained from the monolingual data from target domain. Our approach will be based on building a sentence classifier similar to as seen before, but now to identify the common parts and the uncommon parts between the two domains and carefully transferring knowledge without hurting the performance of the new domain.

9.4 Combining Active and Semi-Supervised Learning

Semi-supervised learning is a broader area of Machine Learning focusing on how to use unlabeled data in conjunction with labeled data to improve the learning process. Many semi-supervised learning algorithms fall under a co-training[Blum and Mitchell, 1998] or boosting/ self-training framework. For a more general overview of semi-supervised learning we refer to [Zhu, 2005].

Co-training assumes the data set has multiple views and training different classifiers on a non-overlapping subset of these features provides additional labeled data. Bootstrapping algorithms have applied this framework to the tasks of named entity recognition [Collins and Singer, 1999], statistical parsing [Steedman et al., 2003a] and text classification [Miller and Uyar, 1997] among others. [Miller et al., 2004] learn features from unlabeled data to use them in a supervised learning algorithm. They achieved significant performance gains in named-entity recognition by using word clustering features and active learning.

Self-training uses the supervised learning algorithm trained on the labeled data to annotate available unlabeled data. High confidence data points are then added as extra labeled data to train the learner. Self-training was initially used to learn word statistics for a statistical parser [Charniak, 2000]. It was later applied to improve performance of a parsing by introducing a self trained parser in a discriminative re-ranking framework [McClosky et al., 2006]. [Yarowsky, 1995] takes a bootstrapping approach to using labeled and unlabeled data for the task of word sense disambiguation task. Similar bootstrapping techniques were also applied to induce part-of-speech taggers [Cucerzan and Yarowsky, 2002]. [Nigam et al., 2000] show improvement in the text classification task, by using simple generative models like the naive bayes to learn from labeled data in conjunction with the expectation maximization algorithm that can provide class probabilities for unlabeled data.

While active learning looks at prioritizing the unlabeled data to annotate the next set of examples is one way of learning, semi-supervised learning looks at how to re-use labeled data for labeling the unlabeled instances. We propose to combine both these approaches more closely, along the lines below:

- Active learning can be used to seed labeled data required for semi-supervised learning methods. This direction has also been addressed to some extent in this thesis, but obtaining seed data for different tasks, sometimes including different types of annotations will provide interesting challenges. For instance, consider the problem of building syntactic resources for low-resource languages by projecting syntax via word-alignment links. In such a setup one can imagine combining the different annotations more closely and actively by addressing questions such as - ”Do we involve a human to correct word-alignment links or correct a syntactic annotation?”. A Co-Training setup

for such a problem also renders the benefit of combining the seemingly different tasks under a single framework and will make way to study the multi-type active learning strategies we explored in this thesis.

- We can also explore semi-supervised approaches like Self-training, similar to [Schwenk, 2008]. Translations can be filtered based on their quality of translation [Specia et al., 2010] or confidence of the models [Blatz et al., 2004] and only high quality translation pairs can be added to the labeled dataset. However, interesting research direction lies in identifying the translations that are closer to the boundary. We propose to use crowdsourcing as an alternative, as such translations are very difficult to judge for the metrics but easier to judge for a human.

- We propose a graph-based learning framework that unifies the semi-supervised learning with active learning. Recently semi-supervised learning has come to be seen as combining labeled and unlabeled data in a graph or a network datastructure for studying label propagation. An interesting proposal is to also study active learning strategies within the same setup and formulate the active learning strategies as new kernel functions for a Graph based semi-supervised learning framework. This has the advantage of closely coupling active and semi-supervised learning and gives a more formal treatment to Active Learning under the graph theoretic framework. We also propose then extending this to work on proactive learning with multiple imperfect oracles. We believe that within this framework we can pose repeated labeling, a strategy to improve quality of the output labels in crowdsourcing, as yet another query function over the graph. We also envision that informative priors can be easily be incorporated into our model. We are working on testing this on two language learning tasks - a standard part of speech tagging problem and a word-alignment problem for the machine translation.

Bibliography

Luis von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006. 132

Vamshi Ambati and Jaime Carbonell. Proactive learning for building machine translation systems for minority languages. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics, 2009. 130

Vamshi Ambati and Stephan Vogel. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010. 93

Vamshi Ambati, Alon Lavie, and Jaime Carbonell. Extraction of syntactic translation models from parallel data using syntax from source and target languages. In *Proceedings of the Machine Translation Summit XII*. Association for Machine Translation in the Americas (AMTA), 2009. 130

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Active learning and crowd-sourcing for machine translation. In *Proceedings of the LREC 2010*. European Language Resources Association (ELRA), 2010a. 18, 57, 78, 111, 115

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Active learning-based elicitation for semi-supervised word alignment. In *Proceedings of the ACL 2010*. Association for Computational Linguistics, 2010b. 56

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Active semi-supervised learning for improving word alignment. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing (ALNLP)*. Association for Computational Linguistics, 2010c. 56

Vamshi Ambati, Sanjika Hewavitharana, Stephan Vogel, and Jaime Carbonell. Active learning with multiple annotations for comparable data classification task. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Portland, Oregon, 2011a. Association for Computational Linguistics. 62

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Multi-strategy approaches to active learning for statistical machine translation. In *Proceedings of the 13th Machine Translation Summit*. Machine Translation Summit, 2011b. 37

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Towards task recommendation in micro-task markets. In *Proceedings of the AAAI 2011 Workshop on Human Computation (HCOMP)*. Association for the Advancement of Artificial Intelligence (AAAI), 2011c. 132

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Collaborative workflow for crowdsourcing translation. In *Proceedings of 2012 ACM Conference on Computer Supported Cooperative Work (CSCW)*. Association for Computing Machinery (ACM), 2012. 107

Shilpa Arora, Eric Nyberg, and Carolyn P. Rosé. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics, 2009. 115

Josh Attenberg, Prem Melville, and Foster Provost. A unified approach to active dual supervision for labeling features and examples. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases*. Springer-Verlag, 2010. 59

Jason Baldridge and Miles Osborne. Active learning for hpsg parse selection. In *Proceedings of the HLT-NAACL 2003*, pages 17–24, Morristown, NJ, USA, 2003. Association for Computational Linguistics. 15

Yoram Baram, Ran El-Yaniv, and Kobi Luz. Online choice of active learning algorithms. *J. Mach. Learn. Res.*, 5:255–291, 2004. ISSN 1532-4435. 36

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. Confidence estimation for machine translation. In *Proceedings of Coling 2004*, Geneva, Switzerland, 2004. Coling 2004 Organizing Committee. 49, 136

John Blitzer. *Domain Adaptation of Natural Language Processing Systems*. PhD thesis, University of Pennsylvania, 2008. 134

Michael Bloodgood and Chris Callison-Burch. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010. Association for Computational Linguistics. 130

A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of Computational Learning Theory*, 1998. 135

Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992. 14, 15

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. 5, 45, 46, 56

Chris Callison-burch. *Active learning for statistical machine translation*. PhD thesis, Edinburgh University, 2003. 15

Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of EMNLP 2009*. Association for Computational Linguistics, 2009. 18, 92, 93, 103

Chris Callison-Burch and Mark Dredze, editors. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, Los Angeles, 2010. 17

Chris Callison-Burch, David Talbot, and Miles Osborne. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of ACL 2004*. Association for Computational Linguistics, 2004. 46

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL*. Association for Computational Linguistics, 2006. 117

Rich Caruana. Multitask learning. In *Proceedings of Machine Learning*, 1997. 57, 58, 77

Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of NAACL*. Morgan Kaufmann Publishers Inc., 2000. 14, 15, 135

Colin Cherry and Dekang Lin. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006. 46

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of 43rd ACL*. Association for Computational Linguistics, 2005. 13

M. Collins and Y. Singer. Unsupervised Models for Named Entity Classification. In *Proceedings of EMNLP*, 1999. 135

Josep M. Crego and Nizar Habash. Using shallow syntax information to improve word alignment and reordering for SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2008. 14

S. Cucerzan and D. Yarowsky. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. 135

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. What can syntax-based MT learn from phrase-based MT? 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, 2007. 13

John DeNero and Dan Klein. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th ACL*, Prague, Czech Republic, 2007. Association for Computational Linguistics. 13

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011. 100, 101

Pinar Donmez and Jaime G. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *CIKM '08*. Association for Computing Machinery, 2008. 18

Pinar Donmez, Jaime G. Carbonell, and Paul N. Bennett. Dual strategy active learning. In *ECML*, pages 116–127, 2007. 10, 36, 39, 44, 82

Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of KDD 2009*. Association for Computing Machinery, 2009. 18, 105, 107

Pinar Donmez, Jaime Carbonell, and Jeff Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proceedings of the SIAM Conference on Data Mining*, 2010. 18

Gregory Druck, Burr Settles, and Andrew McCallum. Active learning by labeling features. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*. Association for Computational Linguistics, 2009. 59, 68

Matthias Eck, Stephan Vogel, and Alex Waibel. Low cost portability for statistical machine translation based on n-gram coverage. In *Proceedings of MT Summit X*, 2005. 16, 21

Carsten Eickhoff and Arjen de Vries. How crowdsourcable is your task? In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*. Association for Computing Machinery (ACM), 2011. 97

Jenny Rose Finkel and Christopher D. Manning. Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data. In *Proceedings of ACL 2010*. Association for Computational Linguistics, 2010. 58, 76

Victoria Fossum, Kevin Knight, and Steven Abney. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of of the Third Workshop on Statistical Machine Translation,ACL*. Association for Computational Linguistics, 2008. 14

Alexander Fraser and Daniel Marcu. Semi-supervised training for statistical word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006. 45

Alexander Fraser and Daniel Marcu. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303, 2007a. 48, 52

Alexander Fraser and Daniel Marcu. Getting the structure right for word alignment: LEAF. In *Proceedings of the 2007 Joint Conference on EMNLP-CoNLL*. Association for Computational Linguistics, 2007b. 46

Yoav Freund, Sebastian H. Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine. Learning.*, 28(2-3):133–168, 1997. 25, 36, 50

Pascale Fung and Lo Yen Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 414–420, Montreal, Canada, 1998. 63

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. What's in a translation rule? In Susan Dumais; Daniel Marcu and Salim Roukos, editors, *Proceedings of HLT-NAACL 2004*. Association for Computational Linguistics, 2004. 13

Rashmi Gangadharaiah, Ralf Brown, and Jaime Carbonell. Active learning in example-based machine translation. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*, 2009. 16, 36

Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, Columbus, Ohio, 2008. Association for Computational Linguistics. 47

Gholamreza Haffari and Anoop Sarkar. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 2009. 15

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. Active learning for statistical phrase-based machine translation. In *Proceedings of HLT NAACL 2009*, Boulder, Colorado, June 2009. Association for Computational Linguistics. 15, 21, 35, 36, 57

Dilek Hakkani-tr, Giuseppe Riccardi, and Allen Gorin. Active learning for automatic speech recognition. In *Proceedings of the ICASSP*. Institute of Electrical and Electronics Engineers (IEEE), 2002. 14

Abhay Harpale and Yiming Yang. Active learning for multi-task adaptive filtering. In *Proceedings of ICML*, 2010. 59

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel. Cmu haitian creole-english translation system for wmt 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011. 84

Robin Hewitt and Serge Belongie. Active learning in face recognition: Using tracking to build a face model. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*. Institute of Electrical and Electronics Engineers (IEEE), 2006. 14

J Howe. Crowdsourcing: A definition, 2006. URL `http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html`. 3

Fei Huang. Confidence measure for word alignment. In *Proceedings of the Joint ACL and IJCNLP*, Suntec, Singapore, 2009. Association for Computational Linguistics. 50

Rebecca Hwa. Sample selection for statistical parsing. *Comput. Linguist.*, 30(3):253–276, 2004. 15

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*. Association for Computing Machinery (ACM), 2010. 19, 98

T. Joachims. A statistical learning model of text classification with support vector machines. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Association for Computing Machinery (ACM), 2001. 75

R.S.M. Kato and E. Barnard. Statistical translation with scarce resources: a south african case study. In *SAIEE Africa Research Journal*, 2007. 16

Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*. Association for Computing Machinery (ACM), 2008. 17

Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007. Association for Computational Linguistics. 134

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the HLT/NAACL*, Edomonton, Canada, 2003. Association for Computational Linguistics. 13, 56

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL Demonstration Session*. Association for Computational Linguistics, 2007. 28, 42, 47, 54, 88, 118, 119

Andreas Krause and Eric Horvitz. A utility-theoretic approach to privacy and personalization. In *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*. Association for the Advancement of Artificial Intelligence (AAAI), 2008. 16

Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of WMT 2007*. Association for Computational Linguistics, 2007. 23, 55, 95, 100, 106

Edith Law and Luis von Ahn. *Human Computation*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011. 92

David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufmann, 1994. 26, 32, 36, 49, 67

Ariadna Font Llitjos. *Automatic Improvement of Machine Translation Systems*. PhD thesis, LTI SCS, Carnegie Mellon University, 2007. 130

Adam Lopez. Statistical machine translation. *ACM Computational Surveys*, 40:8:1–8:49, August 2008. 4

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. Spmt: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on EMNLP*, Sydney, Australia, July 2006. Association for Computational Linguistics. 13

Andrew McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. In *Proceedings of ICML*. Morgan Kaufmann, 1998. 15, 21, 36

David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, 2006. 135

Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 74, New York, NY, USA, 2004. ACM. ISBN 1-58113-828-5. 25, 36

Prem Melville, Foster Provost, Maytal Saar-Tsechansky, and Raymond Mooney. Economical active feature-value acquisition through expected utility estimation. In *UBDM '05: Proceedings of the 1st international workshop on Utility-based data mining*. Association for Computing Machinery (ACM), 2005. 16, 59, 67

D. J. Miller and H. S. Uyar. A Mixture of Experts Classifier with Learning Based on Both Labelled and Unlabelled Data. In J. D. Cowan, G. Tesuaro, and J. Alspector, editors, *Proc of NIPS-1997*, San Fransisco, CA, 1997. Morgan Kaufmann. 135

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38: 39–41, 1995. 100

Scott Miller, Jethran Guinness, and Alex Zamanian. Name tagging with word clusters and discriminative training. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. 135

Behrang Mohit and Rebecca Hwa. Localization of difficult-to-translate phrases. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2007. 16

Christian Monson, Ariadna Font Llitjs, Vamshi Ambati, Lori Levin, Alon Lavie, Alison Alvarez, Roberto Aranovich, Jaime Carbonell, Robert Frederking, Erik Peterson, and Katharina Probst. Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages. In *Proceedings of the LREC 2008*. European Language Resources Association (ELRA), 2008. 130

Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005. 62, 63, 69

Dragos Stefan Munteanu and Daniel Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia, 2006. 63

Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *ICML*, 2004. 35, 36, 37

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 30(3), 2000. 135

Sergei Nirenburg and Robert Frederking. Toward multi-engine machine translation. In *HLT '94: Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994. 110

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the ACL*, Sapporo, Japan, 2003. Association for Computational Linguistics. 28, 42, 88, 118

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, pages 19–51, 2003. 45, 51, 62

Daniel Ortiz-Martínez, Ismael García-Varea, and Francisco Casacuberta. Online learning for interactive statistical machine translation. In *HLT-NAACL*. Association for Computational Linguistics, 2010. 130

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. 58

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*. Association for Computational Linguistics, 2002. 23, 28, 55, 100, 106, 112, 118

Chris Quirk, Raghavendra U. Udupa, and Arul Menezes. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of the Machine Translation Summit XI*, pages 377–384, Copenhagen, Denmark, 2007. 63

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, April 2010. 19, 98

Roi Reichart, Katrin Tomanek, and Udo Hahn. Multi-task active learning for linguistic annotations. In *Proceedings of ACL*. Association for Computational Linguistics, 2008a. 59

Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. Multi-task active learning for linguistic annotations. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, 2008b. Association for Computational Linguistics. 16, 59, 79

Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3): 349–380, 2003. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120103322711578. 62

D. Roth and K. Small. Margin-based active learning for structured output spaces. In *Proceedings of the European Conference on Machine Learning (ECML)*. Springer, 2006. 15, 21

Dan Roth and Kevin Small. Active learning for pipeline models. In *Proceedings of AAAI*. Association for the Advancement of Artificial Intelligence (AAAI), 2008. 59, 83

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *IDA '01: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*. Springer-Verlag, 2001. 50, 67

Holger Schwenk. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation*, 2008. 136

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 7, 14

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS*. MIT Press, 2008. 59

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*. Association for Computing Machinery (ACM), 2008. 105

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources. In *Proceedings of the LREC 2008 Workshop on Collaboration: interoperability between people in the creation of language resources for less-resourced languages*. European Language Resources Association (ELRA), 2008. 2

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the EMNLP 2008*. Association for Computational Linguistics, 2008. 17, 18, 92, 103

Lucia Specia, Dhwaj Raj, and Marco Turchi. Machine translation evaluation versus quality estimation. *Machine Translation*, 24:39–50, March 2010. ISSN 0922-6567. 136

M. Steedman, M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaier, P. Ruhlen, S. Baker, and J. Crim. Bootstrapping statistical parsers from small datasets. In *11th Conference of the European Association for Computational Linguistics: EACL 2003*, Budapest, Hungary, April 2003a. 135

Mark Steedman, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. Example selection for bootstrapping statistical parsers. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Association for Computational Linguistics, 2003b. 15, 21

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Towards a broad-coverage bilingual corpus for speech translation of travel conversation in the real world. In *Proceedings of LREC 2002, Las Palmas, Spain*, 2002. 28, 42, 83, 93

Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. Active learning for natural language parsing and information extraction. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 1999. 15

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning*, pages 45–66, 2002. 77

Nicola Ueffing and Hermann Ney. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40, 2007. 27, 49, 50

Ashish Venugopal, Andreas Zollmann, and Vogel Stephan. An efficient two-pass approach to synchronous-CFG driven statistical MT. In *Proceedings of HLT-NAACL*. Association for Computational Linguistics, April 2007. 13

Sudheendra Vijayanarasimhan and Kristen Grauman. Multi-level active prediction of useful image annotations for recognition. In *Proceedings of NIPS'08*. Springer, 2008. 59

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, 1997. ISSN 0891-2017. 13

Hua Wu, Haifeng Wang, and Zhanyi Liu. Boosting statistical word alignment using labeled and unlabeled data. In *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006. 46

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *Proceedings of ACL '01*. Association for Computational Linguistics, 2001. 13

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, pages 189–196, 1995. 135

Omar F. Zaidan and Chris Callison-Burch. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of EMNLP*. Association for Computational Linguistics, 2009. 18, 92

Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2011. 96, 101

John M. Zelle and Raymond J. Mooney. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996. 15

Yi Zhang. Multi-task active learning with output constraints. In *Proceedings of AAAI*. Association for the Advancement of Artificial Intelligence (AAAI), 2010. 59

Bing Zhao and Stephan Vogel. Full-text story alignment models for chinese-english bilingual news corpora. In *Proceedings of the ICSLP '02*, September 2002. 63

X. Zhu. Semi-Supervised Learning Literature Survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005. 135