Unpacking DialCrowd's useful functions - an Ablation Study.

Ting-Rui Chiang   Jeffrey P. Bigham   Maxine Eskenazi

CMU-LTI-21-020

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

Unpacking DialCrowd's useful functions - an Ablation Study

Ting-Rui Chiang[1] and **Jeffrey P. Bigham**[1,2] and **Maxine Eskenazi**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Human-Computer Interaction Institute, Carnegie Mellon University
{tingruic,jbigham,max+}@cs.cmu.edu

1 Introduction

The creation of a human intelligence task (HIT) that collects high-quality data touches on many variables. Requesters, those who create a HIT, need to design a completely understandable task and pay the workers a decent wage (Berg, 2015; Marge et al., 2010; Grady and Lease, 2010; Alonso and Baeza-Yates, 2011). In this report, we examine the variables involved in that enumerate the elements involved in task creation that the DialCrowd software offers to the research community [1]. This report systematically examines the role each variable plays. The report also looks at how good vs poor payment effects the quality of a HIT's data. For this, an ablation study has been carried out on each of the DialCrowd variables, including background information, instructions of the task, examples and counterexamples, explanations for the provided examples.

DialCrowd is a toolkit that helps requesters create tasks that get high quality results (Lee et al., 2018). It features an easy-to-use GUI for task creation. This interface provides guidelines such as suggesting decent pay levels or encouraging the addition of items such as high level explanations.

Specifically, in addition to investigating the effect of payment (McGraw et al., 2010; Novotney and Callison-Burch, 2010; Marge et al., 2010), this report also assesses the relative importance of each variable that DialCrowd uses to help requesters create a successful HIT (the task chosen for this HIT is intent categorization, which requires workers to identify the intent of an utterance.):

Background: Requesters should explain the high level goal of the task and its relative importance in their work. Kaufmann et al. (2011); Kaveti and Akbar (2020) show that providing background information provides workers with more intrinsic motivation, and results in higher data quality.

Instruction: Requesters should provide clear, understandable and unambiguous instructions for what they expect the workers to do.

Description of Categories: Requesters should provide a clear description (definition) of each category the worker is to use.

Examples and Counterexamples of Categories: Requesters should provide concrete examples of a correct worker response and of an incorrect worker response (for each category, in the present study). Willett et al. (2012) showed that providing examples can improve the quality of the collected data. Doroudi et al. (2016) found that examples created by experts can be successfully use in worker training.

Explanations: Requesters should also explain the choice of each example and counterexample.

2 Experimental Setups

2.1 Tasks and Datasets

There were two tasks chosen for the present study: an intent categorization task and a system act categorization task. The intent categorization task is the simpler of the two tasks, simply asking workers to categorize the speaker's intent in a given utterance. Each unit task included only one utterance extracted from the dataset proposed by Larson et al. (2019). The topics of the intents were related to daily life, e.g. paying bills, making transactions. The names of the categories are also intuitive, e.g. "pay_bill" for paying bills, and "transaction" for making a transaction. This made it relatively easy for workers to attempt to categorize an utterance without reading each category description. On the other hand, the system act categorization task required workers to categorize an utterance based on its context. Each unit task included a turn pair, one user utterance and one system utterance. We extracted the utterances from the MultiWoz database

[1] https://cmu-dialcrowd.herokuapp.com/

Intent and Description	
Transactions	Request for information about transactions of a bank account.
Transfer	Request to make a transfer from one banking account to another one.
Balance	Ask information about the amount of money in a banking account.
Pay bill	Request for help to pay a bill.
Bill Balance	Request for information about the balance of a bill.

Sys. Act	Description
Request	The system asks the user for more information about what the user want.
Recommend	The system provides a recommendation for the user.
Book	The system confirms with the user that a booking was made successfully.
NoBook	There are options that satisfy the user's requirements, but they can not be booked, sometimes because there are no reservations or the option is not available during the time frame the user requests.
NoOffer	There is no option that satisfies the user's requirements, regardless whether reservations are available.

Table 1: The categories in the intent categorization task (top) and the system act categorization task (bottom).

2.1 (Eric et al., 2020). The categories were based on the ontology defined by conversational AI researchers (Budzianowski et al., 2018) and thus were less intuitive. It was more difficult to categorize an utterance solely based on the name of a given category, such as "offer" or "request". It was necessary to read the descriptions of the categories in order to correctly carry out the task.

A subset of the datasets was used for each of the tasks. There were 5 categories from each of the two datasets respectively. For the system act categorization task, the categories with less ambiguity were chosen. Table 1 shows the categories used in the study. For both of the tasks, we randomly sampled 10 utterances for each category (thus 50 utterances in total). We also randomly sampled 10 instances for each category. We chose the most representative and ambiguous of these 10 instances as the examples and counterexamples provided to the workers.

2.2 Platform and General Settings

Amazon Mechanical Turk and DialCrowd (Lee et al., 2018) were used in the study. Following common practice, worker qualifications were: (1) HIT approve rate > 95%, (2) number of HITs approved > 100. Each HIT contained 5 items to annotate. We also used a customized worker qualification score to prevent any given worker from working on two different conditions.

2.3 Metrics

Accuracy and pairwise Cohen's Kappa (Cohen, 1960) were used for correctness and inter-worker agreement respectively. Accuracy was based on the label provided in the dataset. To ensure the correctness of the labels, after the workers finished the two tasks, the instances that have an error rate greater than 50% were checked manually. We found and corrected two samples with incorrect labels in the system act categorization dataset. Thus the results shown in this paper are on a modified version of MultiWoz 2.1.

2.4 Settings to Compare

For both of the tasks, we first created a HIT that included all of the variables in the study. The ablation consisted of leaving one variable out in each HIT version. The settings were: (1) full, (2) without examples (and counterexamples), (3) without instructions, (4) without descriptions (for each of the categories), (5) without background, (6) without explanations (for each example and counterexample).

Finally, there was a setting with reduced payment. In the full setting, remuneration was $15/hr. The time it took to do a task was estimated by having 5 workers do a pilot study. A HIT took on average 2.4 minutes for the intent classification task and 3.5 minutes for the system act categorization task. Thus the reduced payment version paid

$0.65/HIT and $0.90/HIT respectively.

We tracked the time spent on each part of the task, time spent reading the descriptions, time spent reading examples, etc.

In each HIT, workers were given a box where they could give optional feedback.

3 Results and Discussion

From the results in Table 2 and the visualization in Figure 1, we have the following observations:

3.1 Trends Common to the Two Tasks

Lack of general instructions or category descriptions negatively affects data quality. Both inter-worker agreement (Cohen's Kappa) and accuracy decrease in the "without" conditions for each of these variables, with system act categorization having a greater decrease. A possible explanation is that the system act task instructions contain essential information: the workers are to categorize the individual system response instead of the whole conversation, for example.

Reducing payment may cause workers to spend less time on a task. We examined the full and partial time data. Both time spent reading the instructions and the time spent on the whole task are reduced in the reduced payment condition. This could reflect the worker desiring to maximize their income per time spent. This is underlined by the lower amount of feedback left in the reduced payment condition. However, less time spent on a HIT did not affect the accuracy.

Removing the background explanation of causes the workers to spend less time on reading the instructions. The background is the high level explanation of the importance of the task for research. It provides an intrinsic incentive. It is possible that workers may spend less time on task due to lower motivation.

3.2 Results that Differ in the Two Tasks

Removing examples only affects the data quality of the intent categorization task. In contrast, removing examples did not affect the data quality of the system act categorization task significantly. One possible explanation is that the examples in the system act categorization task are not effective for the data quality. The examples in the system act tasks may present too much information, and thus workers may not read all of them. Compared to the examples in the intent categorization task, the examples in the system act categorization task are much longer. Each of them is a conversation consisting of two utterances, which may make it even harder to understand. On the other hand, reading the description of the categories may be more helpful for understanding than reading the examples. Since the categories in the system act categorization task is more complex, any single example may not be representative enough for each of the categories. Workers may be able to understand the definition of a category by reading its examples and counterexamples. Therefore, examples and counterexamples were more helpful in the intent task.

4 Conclusions and Future Directions

This study examined the relative importance of the variables found in the DialCrowd software for creating crowdsourcing tasks. We have showed that providing instructions and descriptions of categories is important for obtaining data of good quality. Although removing the background and reducing payment did not significantly impact data quality, these actions did cause workers to spend less time on task. If this happens with a more complex task, less time spent on task may result in lower quality data. Removing examples reduces the accuracy of the intent categorization task significantly, but has no significant impact on the system act categorization task (p-value = 0.497). More research is needed to determine what the full implications of this result could be.

Acknowledgements

This work is funded by National Science Foundation grant CNS-1512973. The opinions expressed in this paper do not necessarily reflect those of the National Science Foundation.

References

Omar Alonso and Ricardo Baeza-Yates. 2011. Design and implementation of relevance assessments using crowdsourcing. In *European Conference on Information Retrieval*, pages 153–164. Springer.

Janine Berg. 2015. Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers. *Comp. Lab. L. & Pol'y J.*, 37:543.

setting	Total Time	time spent on the instructions	Kappa	# of Feedback	Acc.	p-val
Full	230 (242)	124 (169)	0.396	10	0.676	
w/o background	151 (161)	64 (68)	0.434	22	0.664	0.849
w/o instructions	243 (184)	137 (143)	0.339	20	0.596	0.077
w/o description	194 (164)	73 (83)	0.268	15	0.6	0.093
w/o example	199 (142)	80.1 (105.1)	0.422	18	0.708	0.497
w/o the explanations	180 (218)	86 (203)	0.305	16	0.628	0.301
change payment	179 (172)	69 (76)	0.391	9	0.692	0.772

setting	Total Time	time spent on the instructions	Kappa	# of Feedback	Acc.	p-val.
Full	168 (243)	100 (232)	0.688	22	0.848	
w/o background	160 (140)	79 (114)	0.627	22	0.82	0.627
w/o instructions	130 (90)	51 (50)	0.547	14	0.78	0.066
w/o description	147 (129)	90 (105)	0.448	19	0.724	0.001
w/o example	88 (46)	35 (25)	0.499	17	0.736	0.003
w/o the explanations	122 (119)	51 (73)	0.545	16	0.78	0.066
change payment	161 (151)	79 (115)	0.642	8	0.824	0.546

Table 2: Results of the system act categorization task (top) and the intent categorization task (bottom). Time is measured in seconds. Standard deviation is in parentheses. The p-value is the significance of accuracy difference compared to the full setting. (Due to technical issues, some time stamps were lost. Time marked with a star (*) is the average of about half of the records.)

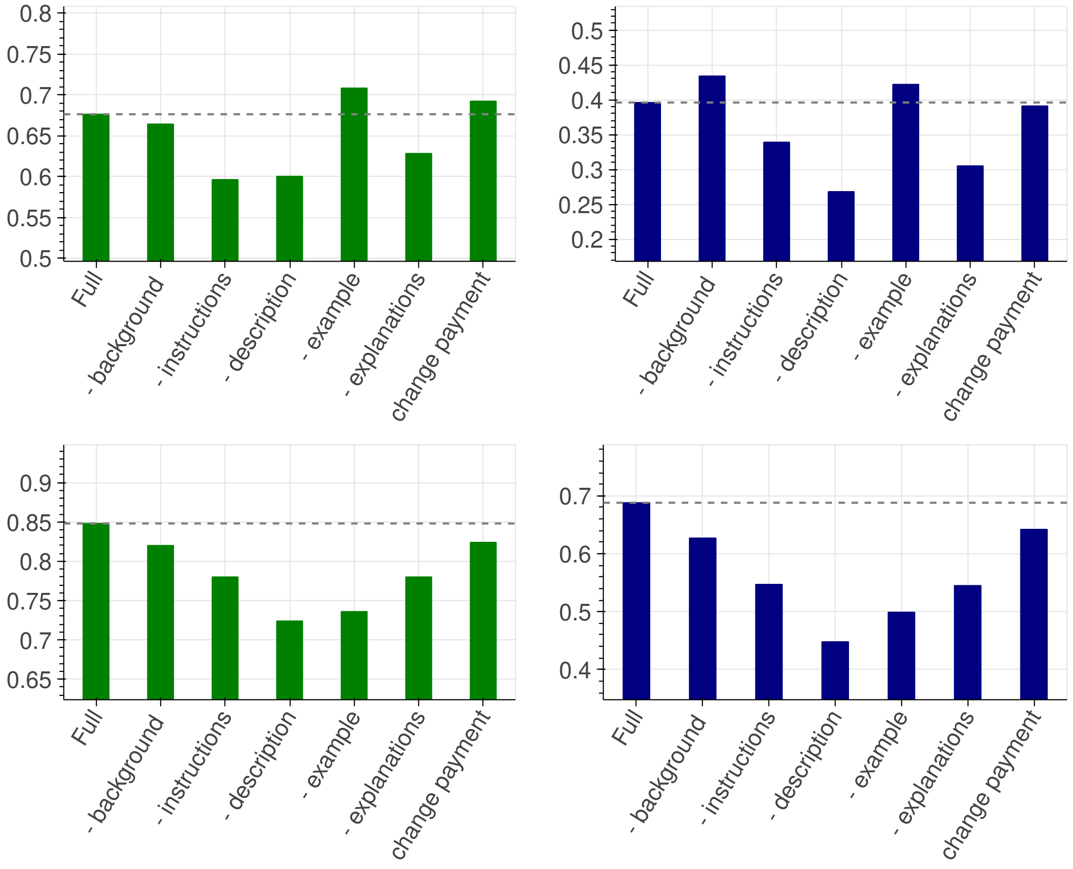

Figure 1: The accuracy (left) and Cohen's Kappa (right) of the system act (top) and the intent (bottom) categorization task.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2623–2634.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 172–179, Los Angeles. Association for Computational Linguistics.

N. Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk.

Pushyami Kaveti and Md Navid Akbar. 2020. Role of intrinsic motivation in user interface design to enhance worker performance in amazon mturk. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '20, New York, NY, USA. Association for Computing Machinery.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Kyusong Lee, Tiancheng Zhao, Alan W. Black, and Maxine Eskenazi. 2018. DialCrowd: A toolkit for easy dialog system assessment. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 245–248, Melbourne, Australia. Association for Computational Linguistics.

Matthew Marge, Satanjeev Banerjee, and Alexander Rudnicky. 2010. Using the Amazon Mechanical Turk to transcribe and annotate meeting speech for extractive summarization. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 99–107, Los Angeles. Association for Computational Linguistics.

Ian McGraw, Chia-ying Lee, Lee Hetherington, Stephanie Seneff, and Jim Glass. 2010. Collecting voices from the cloud. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Scott Novotney and Chris Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215, Los Angeles, California. Association for Computational Linguistics.

Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 227–236.