Optimizing the Design and Cost for Crowdsourced Conversational Utterances

Phoebe Liu
Figure Eight Technologies
San Francisco, USA
phoebe.liu@figure-eight.com

Joan Xiao
Figure Eight Technologies
San Francisco, USA
joan.xiao@figure-eight.com

Tong Liu
Rochester Institute of Technology
Rochester, USA
tl8313@rit.edu

Dylan F. Glas
Futurewei Technologies
San Francisco, USA
dylan.f.glas@gmail.com

ABSTRACT

As more industries adopt chatbot technology, there is a growing demand for crowdsourcing utterances for training dialogue systems. Yet, it is well-known that crowdsourcing has a high rate of noisy and low-quality data. While techniques such as "golden answers" help with filtering noise for certain kinds of data collections, these techniques are difficult to use with utterance collection, due to its open-ended nature, resulting in large amounts of data that must be discarded. Thus, obtaining high quality utterance data often requires careful design and multiple iterations of the crowdsourcing task, which can lead to increase in cost of the crowdsourcing task. In this paper, we consider several variations of commonly-used workflows for utterance collection in a crowdsourcing platform and their effect on the utterance quality. In addition, we propose a strategy to adaptively terminate the data collection process based on utterance coverage and evaluate its effect on the performance of the downstream model. Finally, we examine the cost considerations for crowdsourced tasks and provide suggestions for future utterance collection procedures.

KEYWORDS

Crowdsourcing, Utterance collection, Data Collection, Noisy Data

1 Introduction

Recently, there is a growing demand for crowdsourcing utterance data to bootstrap and rapidly prototype machine learning models. The performance of these models often depends not only on the quantity, but also the quality of the collected data. However, crowdsourcing typically yields noisier data than traditional practices of in-house annotation or in-person data collection [7, 11]. The common practice of human-in-the-loop verification may filter noisy data and improve data quality, but it can also be costly. Thus, it is valuable to consider strategies that are cost-effective without compromising data quality or downstream model performance. However, the trade-offs between cost and data quality or model performance still remains an open research question [16].

While some studies focus on the actual prompts or instructions presented to crowdworkers [6, 8], few have studied cost-effective methodologies for influencing data quality or downstream model performance. Strategies such as using a two-tier payment scheme to incentivize crowd workers or using automatic verification may reduce cost, but their impact on the data quality is still debatable [5, 11, 19]. In this paper, we first examine two questions:

- How much does utterance quality improve when using two-tier payment, compared with paying workers in full up front?
- Since a majority of the cost lies in human verification, how does data quality improve using automated verification as compared to no automatic verification?

Following that, we then investigate the other aspect of the cost efficiency problem – since there is cost associated with data collection, is there a strategy that can be used as a stopping criterion for data collection without impacting model performance? In the second part of our paper, we examine such adaptive termination strategy during the data collection phase. Our findings demonstrate that it is possible to reduce costs without sacrificing data quality.

2 Related Work

While many studies examine how we can use crowdsourcing for different applications, e.g. translation [1, 22], transcription of text from audio [9, 13], annotations for images and videos [4, 18], some studies have consider quality control in crowdsourcing tasks. Lease [11] shared some insights that while using automation to detect spammers may improve quality, such techniques rely on the preconception that workers may be by-and-large responsible or irresponsible. Vaughan [17] surveyed studies that show performance-based payments help improve quality, but this also largely depends on how salient the difference between payment and the wages for that particular region is. We are also interested in the correlation between cost and crowdsourced utterance quality, and we aim to evaluate utterance quality with different workflow variants.

3 Utterance Collection Workflow

We gathered information about best practices for utterance collection tasks by interviewing task designers employed at a crowdsourcing platform, whose responsibilities involved assisting customers ("requesters") in designing crowdsourcing tasks. The design and deployment of utterance data collection tasks includes roughly the following steps: data collection, data verification, and the design of monetary incentives for that particular workflow.

1. Data Collection: The task designer first obtains details about the purpose of data collection from the requester. Then, she creates a task on the crowdsourcing platform, writes scenario-specific instructions, provides examples of "good" and "bad" utterances, and chooses the appropriate crowd as per the requester's requirements. For example, a requester may wish to summarize a long paragraph of text into a short sentence, where fluent English-speaking crowd workers should be requested, while another requester interested in training data for an intent classifier may want ungrammatical wordings from non-fluent English speakers.

2. Data Verification: The output from the data collection task is often used as an input for the data verification task, which serves as a form of quality control for obtaining reliable data [10, 12, 14]. For each utterance collected, three distinct crowd workers are asked to judge its validity. The instructions for this task are similar to the data collection task, but rather than asking workers to generate utterances, they are asked to accept or reject an utterance based on their judgment on the validity of an utterance.

Depending on the scenario, a requester may consider an utterance to be valid based on one or more of the following metrics:

- Semantic Equivalence: The meaning of the collected utterance must match the original intent of sentence.
- Grammaticality: The sentence should be well formed to follow the rules of the target language.
- Gibberish: An utterance is considered gibberish if it is nonsense (e.g. "iivouwerweioh") or not in the target language.

To illustrate, a requester collecting utterance variants to train an intent recognition model for a chatbot might permit grammatical errors but require semantic equivalence to a target utterance, whereas another person collecting utterances to train a text summarization model might require grammatically correct utterances.

The ratings are aggregated using an algorithm following the concept described in [3], where quality of workers, annotations, and input utterances are taken into consideration to calculate a score.

3. Monetary Incentives: Due to the open-ended nature of utterance collection tasks, a common practice is to institute a two-tier payment system: a low to moderate base pay followed by a bonus [5, 19] upon successful completion of the task. This ensures that workers are incentivized to perform well on the current task and continue to do so for future tasks.

Example Utterance Collection Workflow: A requester is interested in collecting 100 utterance variants for training a chatbot, and he decides to pay \$0.03 per utterance collected, using a two-tier payment system where the worker is paid \$0.01 upfront and \$0.02 when an utterance is validated. Each utterance collected is then sent to three workers during the verification step, for \$0.01 per judgment. Thus, if 86 utterances were considered as valid after the verification task, this would be cost a total of \$5.72, broken up into:

Data Collection: $100\ utt. \times\ \$0.01 = \$1.00$
Data Verification: $3\ workers\ \times 100\ utt. \times\ \$0.01 = \$3.00$
Subsequent payment: $86\ valid\ utt. \times\ \$0.02 = \$1.72$

For this particular example, 52% of the cost lies in the human verification step. This trend of the human verification task being the costliest in the overall workflow was observed in many of the utterance collection tasks within the platform.

4 Case Study

4.1 Experimental Design

To investigate the question of how cost correlates with utterance quality, we varied two factors in the overall workflow and evaluated the quality of the collected data as follows:

Single-tier (1T) vs. two-tier payment (2T): Since there is a cost associated with human-in-the-loop verification, we evaluated whether paying the workers upfront versus giving a two-tier payment would make a difference to the utterance quality. In the 1T condition, we paid the total cost upfront (\$0.03). In the 2T condition, we paid one third of the cost (\$0.01) upfront, followed by the rest (\$0.02) upon successful validation of an utterance.

With (WS) and without (NS) smart validator: We investigated whether invalid utterances could be prevented during the data collection phase, thus mitigating the cost of human verification. To do so, we created an automated text validator that prevents workers from submitting gibberish and non-target language utterances. This validator estimates how likely it is to generate an utterance based on the character to character transitions of that utterance. An utterance is marked as gibberish when it has low probability. While we acknowledge that other metrics exist for validating utterances, and validation needs are scenario-dependent, we observed gibberish and non-target language utterances to comprise a significant fraction of noisy utterances.

4.2 Scenario

We set up a scenario to collect data using these four workflow variants. In this study, we asked workers to provide utterances that a customer might ask to a real-estate agent. We defined 29 intents (e.g. the safety of a neighborhood, and services the agent provides). For each intent, we collected 40 different utterance variations.

For each workflow variant, we collected 1160 utterances. We then obtained the number of valid utterances for each workflow by aggregating ratings from three distinct workers, at a cost of \$0.01 for each worker judgment. Figure 1 shows the experimental conditions as well as the cost for each condition.

4.3 Hypotheses

We evaluated the data collected using four workflow variants: (1) 1T+NS, (2) 1T+WS, (3) 2T+NS, (4) 2T+WS, based on the payment scheme and whether we used the smart validator or not. We made the following hypotheses:

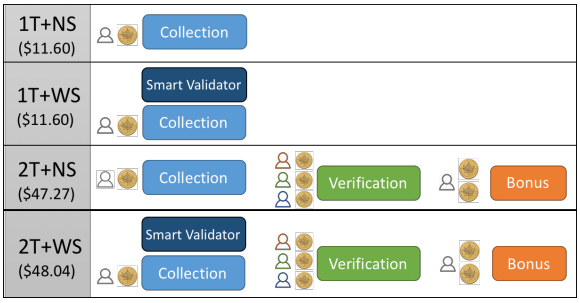

Figure 1: Experimental conditions

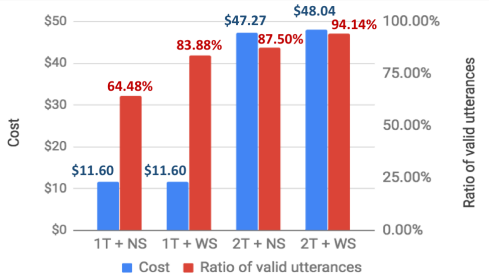

Figure 2: Ratio of valid utterance vs cost for each condition

Hypothesis H1: We predict that a two-tier payment scheme will incentivize workers to provide better utterance quality overall, compared to paying the full amount upfront. Therefore, utterance quality will be better when comparing the 2T conditions with the 1T conditions. However, the overall cost will also increase.

Hypothesis H2: Since as many as one-third of the crowdsourced utterances can be gibberish or is not in the target language, we expect that using the smart validator will filter out these utterances. Therefore, we predict that utterance quality will be better in the WS conditions vs. the NS conditions. As there is no additional cost using WS, the cost will remain the same as NS.

Hypothesis H3: We predict that the benefits of the smart validator and two-tier payment can be combined, so that 2T+WS will result in better utterance quality than 2T+NS.

4.4 Analysis and Results

To evaluate utterance quality, we computed the ratio of the valid utterances for each workflow. For 2T+WS and 2T+NS, we already have these results from the verification task. For 1T+WS and 1T+NS, we ran separate evaluation tasks to obtain the number of valid utterances. The cost for these evaluation tasks is not accounted in the cost of the total crowdsourcing task, as they are used for the purpose of this study and to measure the quality of utterances among different conditions.

Figure 2 shows the ratio of valid utterances in each condition. A chi-square test revealed significant differences among conditions ($\chi 2(1) = 8.457$, $p = .004$). The result supports our hypothesis H1, that using the two-tier payment achieved better utterance quality than using single-tier payment. In addition, the result also supports our hypothesis H2, that using WS achieved better utterance quality than NS. Finally, the result also supports our hypothesis H3, that the benefits of the two-tier payment and smart validator can be combined to improve utterance quality, as it has the highest ratio of valid utterances among all conditions.

Our findings demonstrated that using smart validator, as compared to not using the smart validator, is a way to significantly increase the data quality with either no increase in cost or only marginally increasing the cost (e.g. 2T+WS costs 1.6% more than 2T+NS). We also found that human verification is still more effective than using smart validator but can increase the cost. Finally, the two-tier payment scheme and smart validator can be combined to increase utterance quality.

4.5 Discussion

For the two-tier payment with smart validator (i.e. 2T+WS), we had expected that the workers would have been motivated by the payment structure and naturally wanted to perform well, resulting in similar utterance quality as compared to 2T+NS. However, the increase in utterance quality in 2T+WS led us to speculate that some workers who initially were not attentive realized that they needed to be more attentive after the smart validator prevented them from submission.

In terms of cost, using the smart validator did not incur additional cost for 1T conditions. In 2T conditions, there was a small increase in the cost due to the fact that there were more subsequent payout to the crowdworkers as a result of more valid utterances generated from the data collection phase. Interestingly, the cost was very different, yet the utterance quality was quite similar between 1T+WS and 2T+NS conditions. Compared with the 1T+WS condition, which had 83.88% valid utterances, 2T+NS achieved 87.50% valid utterances, but at a four-fold increase in cost. From a cost-effectiveness perspective, one should consider whether the ratio of utterance acceptability using the smart validator is sufficient and that adding human-in-the-loop is really necessary for their application domain.

5 Adaptive termination strategy

While we have considered the tradeoffs of cost efficiency and utterance *quality* in the previous section, the other aspect of the problem is the impact of data *quantity* on cost. Cost increases linearly with the amount of collected data, and it often may not be obvious how much training data is needed to continue improving the model performance [21, 23]. Thus, we are interested in whether a stopping criterion can be applied during the data collection phase, without hindering the downstream model performance.

5.1 Dataset

For this analysis, we collected a separate dataset according to the scenario described in Sec 4.2. For each of the 29 intents, we collected 748 utterances for a total of 21692 utterances. We then split the data into train (598 utterances per intent) and test (150 utterances per intent) sets.

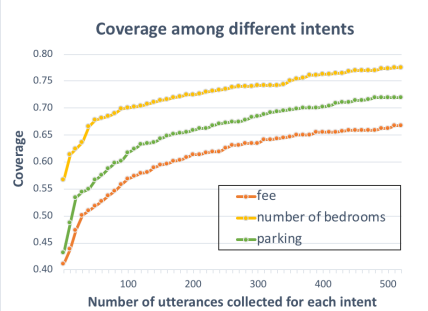

Figure 3: Coverage among different intents

5.2 Coverage

To investigate an optimal stopping criterion for data collection, Kang et al. suggested using the metric of *coverage* to inform the choice of when to stop collecting more data [8]. *Coverage* models how well a training dataset covers the complete space of ways an intent can be expressed. It is also positively correlated with model accuracy and is an independent way to evaluate the quality of training data without training a model. For a given test set, we would want the training set to have as high coverage as possible.

To calculate coverage, we first identify, for each test sentence, the most similar training sentence with the same intent, according to the pairwise sentence cosine distance measure $D(a,b)$. We then derive coverage by averaging the shortest distances for all sentences in the test set. For a given intent, the coverage is:

$$coverage = \frac{1}{|Y_i|}\sum_{b}^{Y_i} max_{a}^{X_i}(1 - D(a,b)) \quad (1)$$

where X_i and Y_i are the sets of utterances collected for intent i. Figure 3 shows the coverage for example intents as we vary the number of utterances.

From the graph, we observe that the coverage curves differ by intent. For example, utterances collected for the intent of asking about the number of bedrooms reached a higher coverage with fewer utterances than the ones collected for intent "fee". This suggests that there is a difference in the number of natural phrasing variations among intents, for example, someone can only thank another person in a few ways (e.g. "Thanks a lot" or "thank you") while he can talk about his budget for apartment hunting in many ways (e.g. "I am looking for an apartment that is not too expensive" or "I want to find a cheap apartment").

5.3 Adaptive strategy for data collection

As observed in Figure 3, given the same amount of training data, different intents may reach different coverage due to the difference in the number of phrasing variations among intents. This prompts us to wonder whether we can use coverage to adaptively terminate data collection for each intent. The idea is that, rather than the common practice of crowdsourcing a constant number of utterance examples for an intent (a "fixed strategy"), we can adaptively terminate the data collection of each intent when the coverage of that intent exceeds a threshold (an "adaptive strategy").

To explore the effect of using this adaptive data collection strategy based on coverage, we performed a post-analysis to create a partial dataset from the dataset in Sec. 5.1 by using the adaptive strategy. For each intent, we first divided the entire data into batches of 10 utterances. For a given batch, we then calculated the coverage for that intent. When the coverage of that intent exceeded a threshold, θ, we no longer added that to the partial dataset. For this study, we used a θ of 0.69.

Condition	Training examples	Model accuracy
Fixed strategy	15167	82.76%
Adaptive strategy	9027	81.33%

Table 1: Model performance trained with fixed strategy versus adaptive strategy

In order to examine the relationship between coverage and model performance, we trained an intent classifier using a multi-layer perceptron (MLP), which has been demonstrated to provide competitive results on text classification task [2, 15, 20]. For both cases, we applied a consistent set of text pre-processing steps: we tokenize and lowercase the text, remove punctuation, and lemmatize. For the MLP, the input dimension to the neural network was 4672, followed by two leaky rectified hidden layers, with 200 hidden neurons each. The output layer was a softmax with the number of neurons equal to the number of possible intents.

Using the test set, we compared the accuracies of the intent classifier trained using the fixed strategy vs. the adaptive strategy. Table 1 shows the result. Interestingly, the model performance is similar in both cases. Yet, the adaptive strategy trains the model with only 59.5% of the total collected data (e.g. 9027 out of 15167 utterances), suggesting that model performance may already have asymptoted. This indicates that coverage is an effective stopping criterion that can be applied during the data collection process without training a model. Given our scenario, using this adaptive strategy would have been 40% more cost-efficient compared with the common approach of predefining a constant number of utterances during data collection. Our analysis serves as a recommendation for future utterance data collection efforts, as adaptive data collection based on per-intent coverage can be a cost-effective method that examines each intent independently.

6 Conclusion

Training data is the key to building machine learning models, and finding the most cost-effective way to collect training data via crowdsourcing still remains an open question. In this paper, we considered several variations of commonly-used workflows for utterance collection. Our findings demonstrated that using a smart validator can significantly increase data quality with no increase in cost. Though human verification was found to be slightly more effective than the smart validator, it also increased the cost substantially. We then demonstrated that using the metric of *coverage* can be a cost-effective way to adaptively terminate the data collection process while maintaining model performance. Our findings provide clear guidance for future data practitioners and demonstrate that it is possible to reduce costs without sacrificing data quality.

REFERENCES

[1] Ambati, V. et al. 2012. Collaborative workflow for crowdsourcing translation. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. (2012), 1191–1194. DOI:https://doi.org/10.1145/2145204.2145382.

[2] Bhatt, S. et al. 2017. An Access Control Framework for Cloud-Enabled Wearable Internet of Things. *IEEE 3rd International Conference on Collaboration and Internet Computing*. April (2017), 328–338. DOI:https://doi.org/10.1109/CIC.2017.00050.

[3] Dumitrache, A. et al. 2018. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. *arxiv.org*. (2018).

[4] Gebru, T. et al. 2017. Scalable Annotation of Fine-Grained Categories Without Experts. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), 1877–1881.

[5] Ho, C.-J. et al. 2015. Incentivizing High Quality Crowdwork. *Proceedings of the 24th International Conference on World Wide Web* (Republic and Canton of Geneva, Switzerland, 2015), 419–429.

[6] Jiang, Y. et al. 2017. Understanding Task Design Trade-offs in Crowdsourced Paraphrase Collection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2017), 103–109.

[7] Kajino, H. et al. 2013. Clustering Crowds. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. (2013), 1120–1127.

[8] Kang, Y. et al. 2018. Data Collection for a Production Dialogue System : A Clinc Perspective. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)* (2018), 33–40.

[9] Kushalnagar, R.S. et al. 2012. A readability evaluation of real-time crowd captions in the classroom. *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*. (2012), 71–78. DOI:https://doi.org/10.1145/2384916.2384930.

[10] Lane, I. et al. 2010. Tools for Collecting Speech Corpora via Mechanical-Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Stroudsburg, PA, USA, 2010), 184–187.

[11] Lease, M. 2011. On quality control and machine learning in crowdsourcing. *AAAI Workshop - Technical Report*. WS-11-11, (2011), 97–102.

[12] McGraw, I. et al. 2010. Collecting Voices from the Cloud. *International conference on Language Resources and Evaluation*. (2010), 1576--1583. DOI:https://doi.org/10.1007/978-1-4842-3417-4_11.

[13] Naim, I. et al. 2013. Text Alignment for Real-Time Crowd Captioning. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013), 201–210.

[14] Parent, G. and Eskenazi, M. 2010. Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data. *2010 IEEE Spoken Language Technology Workshop* (Dec. 2010), 312–317.

[15] Rosario, B. and Hearst, M.A. 2004. Classifying semantic relations in bioscience texts. *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*. (2004). DOI:https://doi.org/10.3115/1218955.1219010.

[16] Sabou, M. et al. 2014. Corpus Annotation through Crowdsourcing : Towards Best Practice Guidelines. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. (2014), 859–866. DOI:https://doi.org/10.1056/NEJMp058231.

[17] Vaughan, J.W. 2018. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*. 18, (2018), 1–46.

[18] Vijayanarasimhan, S. and Grauman, K. 2009. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*. 2009 IEEE, (2009), 2262–2269. DOI:https://doi.org/10.1109/CVPRW.2009.5206705.

[19] Ye, T. et al. 2017. When Does More Money Work? Examining the Role of Perceived Fairness in Pay on the Performance Quality of Crowdworkers. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (2017).

[20] Yoav Goldberg 2016. A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*. 57, (2016), 345–420. DOI:https://doi.org/10.1162/COLI_r_00312.

[21] Yogatama, D. et al. 2017. Generative and Discriminative Text Classification with Recurrent Neural Networks. *arxiv.org*. (2017). DOI:https://doi.org/10.1109/SLT.2016.7846260.

[22] Zaidan, O.F. and Callison-Burch, C. 2011. Crowdsourcing Translation: Professional Quality from Non-professionals. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (Stroudsburg, PA, USA, 2011), 1220–1229.

[23] Zhu, X. et al. 2012. Do We Need More Training Data or Better Models for Object Detection? *Procedings of the British Machine Vision Conference 2012*. (2012), 80.1-80.11. DOI:https://doi.org/10.5244/C.26.80.