A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation

Jiyi Li
University of Yamanashi, Kofu, Japan
RIKEN AIP, Tokyo, Japan
jyli@yamanashi.ac.jp

Fumiyo Fukumoto
University of Yamanashi, Kofu, Japan
fukumoto@yamanashi.ac.jp

Abstract

The target outputs of many NLP tasks are word sequences. To collect the data for training and evaluating models, the crowd is a cheaper and easier to access than the oracle. To ensure the quality of the crowdsourced data, people can assign multiple workers to one question and then aggregate the multiple answers with diverse quality into a golden one. How to aggregate multiple crowdsourced word sequences with diverse quality is a curious and challenging problem. People need a dataset for addressing this problem. We thus create a dataset (CrowdWSA2019) which contains the translated sentences generated from multiple workers. We provide three approaches as the baselines on the task of extractive word sequence aggregation. Specially, one of them is an original one we propose which models the reliability of workers. We also discuss some issues on ground truth creation of word sequences which can be addressed based on this dataset.

1 Introduction

For many tasks in NLP area, the target outputs are word sequences. To train and evaluate the models, the ground truth in the form of word sequences are required. Instead of the oracle which is expensive and has an insufficient number, the crowd which is cheaper and easier to access is a good alternative for collecting the gold standard data.

Because the ability of crowd workers is diverse, to guarantee the quality of the collected data, one solution is to generate redundant data by assigning multiple workers to one instance and then aggregate the multiple answers into golden ones. How to aggregate multiple word sequences with diverse quality is a research problem. In NLP areas such as machine translation, although a few evaluation metrics (Liu et al., 2016) such as BLEU (Papineni et al., 2002) can use multiple golden answers for an instance, because the multiple crowdsourced answers are not golden ones, the aggregation approach for generating a golden one based on these crowdsourced answers is indispensable.

In crowdsourcing area, there are many existing work on answer aggregation for labels (Dawid and Skene, 1979; Whitehill et al., 2009; Zheng et al., 2017). Snow et al. (2008) evaluated crowdsourced label annotations for some NLP tasks and used majority voting for label aggregation. However, there is little work on answer aggregation for word sequences. Nguyen et al. (2017) proposed an aggregation method based on HMM for a sequence of categorical labels and needs to be improved for aligning sparse and free word sequences. If treating a word as a category, there are tens of thousands categories and a sequence only contains a small number of them. To address the problem of answer aggregation for word sequences, people need the datasets which contain multiple word sequence answers provided by different crowd workers for one instance. However, we find that most of the existing datasets in NLP area only contain a single golden answer for one instance.

In this paper, we create a dataset with several crowdsourced word sequence collections for the purpose of solving this problem through a real-world crowdsourcing platform. It contains the translated sentences of the target language by multiple workers from the sentences of the source language. The source sentences are extracted from several existing machine translation datasets. The raw target sentences in these existing datasets can be utilized for evaluating the quality of the crowdsourced data and the performance of the answer aggregation approaches. Our exploration study gives an analysis of worker quality in this dataset.

We provide several approaches on this dataset for the task of extractive sequence aggregation on crowdsourced word sequences, which extracts the

Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, pages 24–28
Hong Kong, China, November 3, 2019.
https://doi.org/10.18653/v1/P17

good word sequence from the candidates. One of them is our original approach which models the reliability of workers, because worker reliability is regarded as an important factor in label aggregation approaches (Zheng et al., 2017).

2 Datasets

2.1 Data Collections: CrowdWSA2019

A number of NLP tasks have the target outputs in the form of word sequences, e.g., machine translation, text summarization, question and answering and so on. In different tasks, the properties of the word sequences, e.g., text length and syntax, can be different from each other. In this paper, without loss of generality, we create a dataset[1] based on the machine translation task which uses short and complete sentences.

To collect the crowdsourced data, we first chose some collections of raw sentence pairs from the existing bilingual parallel corpora. The corpora we utilized are Japanese-English parallel corpora, i.e., JEC Basic Sentence Data[2] (one collection extracted, named as J1) and Tanaka Corpus[3] (two collections extracted, named as T1 and T2). We utilized Japanese as the source language and English as the target language.

We uploaded the sentences in the source language (denoted as *question*) to a real world crowdsourcing platform[4]. We asked the crowd workers to provide the translations in the target language (named as *answer*). Each crowdsourcing micro-task contained ten random source sentences in random order. A worker completed the sentences in a micro-task each time and can answer several random micro-tasks. For the evaluation based on this dataset, we can utilize the original sentences in the target language (named as *true answer*) of these raw sentence pairs to compare with the crowdsourced data and the aggregated word sequences (named as *estimated true answer*).

For the quality of the collected data, because the purpose of creating this dataset is to verify the word sequence aggregation methods, it would be better if the answers of word sequences have diverse quality. The crowd workers on the crowdsourcing platform are mainly Japanese native speakers and non-native speakers of English. Their English abilities are diverse. In the task description, we also encouraged the English beginners to join and provide answers so that the collected answers have diverse quality. Note that if the purpose is collecting high-quality annotation data for specific NLP tasks such as machine translation, using experts or native speakers would be better for improving the data quality.

data	#que.	#wor.	#ans.	#apq	mmr
J1	250	70	2,490	9.96	0.1423
T1	100	42	1,000	10	0.5929
T2	100	43	1,000	10	0.5791

Table 1: Number of questions, workers and answers. #apq: average number of Answers Per Question. mmr: worker-question answer Matrix Missing Rate.

2.2 Exploration Study

We explore some properties of these collections. First, Table 1 lists the statistics of the three collections. Besides the number of questions, workers and answers, it also shows two measures. $\#apq$ is the average number of Answers Per Question. It shows the redundancy of the answers. mmr is the worker-question answer Matrix Missing Rate. It shows the sparsity of the answers. Our collections follows the practical scenario, i.e., when the number of questions is huge, it is impossible that each worker can answer all questions. The redundancy and sparsity may influence the results of answer aggregation approach. For example, it has been shown that the performance of some aggregation approaches for categorical labels may degrade when the mmr is low (Li et al., 2017).

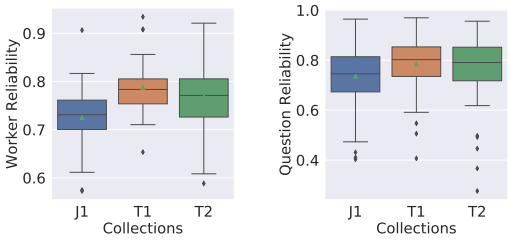
Second, we show the answer quality in the data. Figure 1 illustrates the distribution of answer reliability by embedding similarity. We measure the similarity between a worker answer and the true answer of a question to evaluate the quality. We use the universal sentence encoder to encode the sentences (Cer et al., 2018) into embeddings, and compute the cosine similarity between the embeddings of two sentences. The reliability of a worker is the mean similarity for all answers of this worker. This reliability of a question is the mean similarity of all answers of this question. Both the mean of two types of reliability are in the range of [0.7, 0.8]. The quality on both T1 and T2 is higher. One possible reason is that the size

[1] https://github.com/garfieldpigljy/CrowdWSA2019

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JEC%20Basic%20Sentence%20Data

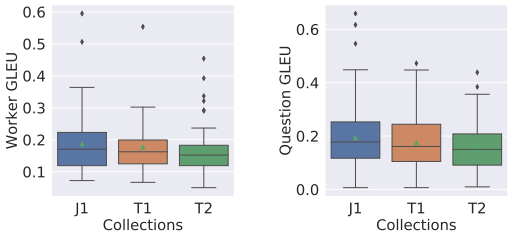
[3] https://github.com/odashi/small_parallel_enja

[4] https://www.lancers.jp/



(a) Worker-Wise (b) Question-Wise

Figure 1: Answer Reliability by Embedding Similarity



(a) Worker-Wise (b) Question-Wise

Figure 2: Answer Reliability by GLEU

of J1 is larger and more low quality workers join the task, while the high quality workers on a non-native English crowdsourcing platform is limited.

Figure 2 shows the distribution of answer reliability measured by GLEU (Wu et al., 2016). To be consistent with the reliability of workers and questions computed by embedding similarity, we use the mean of sentence-wise GLEU of all answers of a worker (question), in contrast to corpus-wise BLEU measure. In contrast to Figure 1, the quality on J1 is higher. One possible reason is that the low quality workers judged by embedding similarity can provide good words or phrases translations which the GLEU focuses on, but cannot provide good word orders and syntax on the sentence level which the DAN (Iyyer et al., 2015) model of universal sentence encoder considers.

3 Extractive Answer Aggregation

When we obtain multiple answers for a given question, we need to aggregate them into one answer which can be used as the golden data in the collected dataset. There are at least two alternatives of answer aggregation approaches for the case of word sequence, i.e., extractive and abstractive answer aggregation. Extractive aggregation methods extract the potential optimal one from multiple worker answers; abstractive aggregation methods generate a new answer by analyzing and understanding all of the worker answers. In the research area of crowdsourcing, most of the existing work of answer aggregation focus on categorical or numerical labels (Zheng et al., 2017). They estimate a pre-defined category or value and thus are extractive approaches. In this paper, we focus on the baselines of extractive answer aggregation.

We define question set $\mathcal{Q} = \{q_i\}_i$, worker set $\mathcal{W} = \{w_k\}_k$, answer set $\mathcal{A} = \{a_i^k\}_{i,k}$, true answer set $\mathcal{Z} = \{z_i\}_i$ and estimated true answer set $\hat{\mathcal{Z}} = \{\hat{z}_i\}_i$. The answer set of a question is $\mathcal{L}_i$; the answer set of a worker is $\mathcal{V}_k$. The encoder is $e(\cdot)$. We use cosine similarity for sim computation.

3.1 Sequence Majority Voting

Majority voting is one of the most typical answer aggregation approaches. For the specific data type of word sequences, we adapt it into a Sequence Majority Voting (SMV) approach. For each question, it first estimates the embeddings of the true answers by $\hat{e}_i = mean(e(\mathcal{L}_i))$; after that it extracts the worker answer $\hat{z}_i = \arg\max_{a_i^k} sim(e(a_i^k), \hat{e}_i)$ as the true answer.

3.2 Sequence Maximum Similarity

We adapt the method in Kobayashi (2018), which is proposed as a post-ensemble method for multiple summarization generation models. For each question, it extracts the worker answer which has largest sum of similarity with other answers of this question. It can be regarded as creating a kernel density estimator and extract the maximum density answer. The kernel function uses the cosine similarity. This Sequence Maximum Similarity (SMS) method can be formulated as $\hat{z}_i = \arg\max_{a_i^{k_1}} \sum_{k_1 \neq k_2} sim(e(a_i^{k_1}), e(a_i^{k_2}))$.

3.3 Reliability Aware Sequence Aggregation

Both SMV and SMS do not consider the worker reliability. In crowdsourcing, worker reliability is diverse and is a useful information for estimating true answers. Existing work in the categorical answer aggregation strengthen the influences of answers provided by the workers with higher reliability. Therefore, we also propose an approach which models the worker reliability, named as Reliability Aware Sequence Aggregation (RASA).

The RASA approach is as follows. (1). ENCODER: it encodes the worker answers into embeddings; (2). ESTIMATION: it estimates the embeddings of the true answers considering worker

reliability; (3). EXTRACTION: for each question, it extracts a worker answer which is most similar with the embeddings of the estimated true answer.

For estimating the embeddings of the true answers, we adapt the CATD approach (Li et al., 2014) which is proposed for aggregating multiple numerical ratings. We extend it into our sequence case by adapting it to the sequence embeddings. We define the worker reliability as β. The method iteratively estimates β_k and $\hat{e}_i$ until convergence, $\beta_k = \frac{\chi^2_{(\alpha/2,|\mathcal{V}_k|)}}{\sum(e(a_i^k)-\hat{e}_i)^2}$, $\hat{e}_i = \frac{\sum \beta_k e(a_i^k)}{\sum \beta_k}$, where χ^2 is the chi-squared distribution and the significance level α is set as 0.05 empirically. We initialize $\hat{e}_i$ by using the SMV approach. SMS does not estimate $\hat{e}_i$ and cannot initialize $\hat{e}_i$.

3.4 Experimental Results

The evaluation metric is GLEU and the average similarity between the embeddings of the estimated true answers and the true answers (the original target sentences in the corpus) on the all questions. For the extractive answer aggregation, there exists theoretical optimal performance. It is the performance of selecting the worker answer with largest embedding similarity (or GLEU) with the true answer. Table 2 lists the results.

First, both SMS and RASA outperform the naïve baseline SMV. RASA is better than SMV because it considers the worker reliability. SMS is better than SMV as it is based on kernel density estimation which is more sophisticated than majority voting. Second, SMS performs best on J1 collection and RASA performs better on T1 and T2 collection. One of the possible reasons is that J1 has more low quality workers. RASA tends to strengthen the influences of major workers. The estimated embeddings are near to the answers of "good" workers and the "good" workers are the ones that the embeddings of their answers are near to the estimated embeddings. If there are many low-quality workers, it is possible to mistakenly regard a low-quality worker as a high-quality worker because this worker may provide more similar answers with other (low-quality) workers. RASA thus may strengthen the answer by a low-quality worker. Third, the results on both embedding similarity and GLEU are consistent. Forth, in the theoretical optimal results, the quality of J1 is higher than T1 and T2 on GLUE but lower on embedding similarity. This observation is consistent with that in Figure 1 and 2 in Section 2.2. Finally, all methods still cannot be close to the theoretical optimum. The performance is still possible to be improved.

data	SMV	SMS	RASA	Optimal
J1	0.7354	**0.7969**	0.7914	0.8853
T1	0.7851	0.8377	**0.8451**	0.9047
T2	0.7696	0.8288	**0.8339**	0.8986

(a) Embedding Similarity

data	SMV	SMS	RASA	Optimal
J1	0.1930	**0.2627**	0.2519	0.4990
T1	0.1740	0.2194	**0.2296**	0.3698
T2	0.1616	0.2170	**0.2345**	0.3637

(b) GLEU

Table 2: Results of extractive answer aggregation. The optimal result is the theoretical optimal performance of the collection for extractive answer aggregation.

4 Conclusion

In this paper, we proposed a dataset for the research of crowdsourced word sequence aggregation. We also provided three approaches on these datasets for the task of extractive aggregation for crowdsourced word sequences. One of them considers the worker reliability. There are some future work on this topic of answer aggregation.

First, for abstractive answer aggregation approach, an option is that we can train an encoder-decoder model to decode the estimated embeddings of the true answer into a word sequence which can be different from worker answers. Therefore, the abstractive approaches are possible to reach better results than the optimal results of the extractive approaches shown in Table 2.

Second, we can collect additional pairwise comparisons on the preferences of the worker answers by using another round of crowdsourcing tasks and extract the preferred answers. It is similar to a creator-evaluator framework (Baba and Kashima, 2013). Otani et al. (2016) proposed an approach for aggregating the results of multiple machine translation systems with pairwise comparisons. The typical approach of aggregating the pairwise comparison into a rank list was Bradley-Terry model (Bradley and Terry, 1952); CrowdBT model (Chen et al., 2013) extended it in crowdsourcing settings; Zhang et al. (2016) summarized more existing work.

Acknowledgments

This work was partially supported By JSPS KAKENHI Grant Number 17K00299 and 19K20277.

References

Yukino Baba and Hisashi Kashima. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 554–562, New York, NY, USA. ACM.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 193–202, New York, NY, USA. ACM.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Hayato Kobayashi. 2018. Frustratingly easy model ensemble for abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4165–4176, Brussels, Belgium. Association for Computational Linguistics.

Jiyi Li, Yukino Baba, and Hisashi Kashima. 2017. Hyper questions: Unsupervised targeting of a few experts in crowdsourcing. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1069–1078, New York, NY, USA. ACM.

Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A confidence-aware approach for truth discovery on long-tail data. *Proc. VLDB Endow.*, 8(4):425–436.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 299–309, Vancouver, Canada. Association for Computational Linguistics.

Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. IRT-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Austin, Texas. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 2035–2043, USA. Curran Associates Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiaohang Zhang, Guoliang Li, and Jianhua Feng. 2016. Crowdsourced top-k algorithms: An experimental evaluation. *Proc. VLDB Endow.*, 9(8):612–623.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proc. VLDB Endow.*, 10(5):541–552.