Improving Spoken Language Understanding by Wisdom of Crowds

Koichiro Yoshino[2,1,3]**, Kana Ikeuchi**[1]**, Katsuhito Sudoh**[1,3] **and Satoshi Nakamura**[1,3]
[1]Nara Institute of Science and Technology (NAIST)
[2]RIKEN Robotics Project
[3]RIKEN Center for Advanced Intelligence Project (AIP)
koichiro.yoshino@riken.jp, {sudoh, s-nakamura}@is.naist.jp

Abstract

Spoken language understanding (SLU), which converts user requests in natural language to machine-interpretable expressions, is becoming an essential task. The lack of training data is an important problem, especially for new system tasks, because existing SLU systems are based on statistical approaches. In this paper, we proposed to use two sources of the "wisdom of crowds," crowdsourcing and knowledge community website, for improving the SLU system. We firstly collected paraphrasing variations for new system tasks through crowdsourcing as seed data, and then augmented them using similar questions from a knowledge community website. We investigated the effects of the proposed data augmentation method in SLU task, even with small seed data. In particular, the proposed architecture augmented more than 120,000 samples to improve SLU accuracies.

1 Introduction

Recent advances in speech applications running on smartphones and smart speakers increase the importance of spoken language understanding (SLU). SLU is a task to predict an appropriate system function with its arguments, given a user request written or spoken in natural language. Various SLU benchmarks have been proposed: Air Travel Information Services (ATIS) (Dahl et al., 1994), restaurant information navigation (Williams et al., 2014), and other speech applications (Hori et al., 2019).

Adaptation of SLU to newly defined tasks is an important problem (Henderson et al., 2014). The number of training data directly affects the SLU accuracy because most of the recent SLU systems are based on statistical machine learning approaches. Some existing work tackled this problem based on transfer learning (Wu et al., 2019), which uses a pre-trained model on different domain data. However, it is still challenging to make the SLU accurate with no or fewer data. Data augmentation has been applied to solve this problem, which generates pseudo training samples (Hou et al., 2018; Yoo et al., 2019). However, such methods often generate unnatural training samples that will decrease SLU accuracy. Another problem is the ambiguity of user utterances; it is difficult to generate such ambiguous examples with generative approaches. Some existing work tackled this problem by using paraphrasing models (Saha et al., 2018; Ray et al., 2018).

On the other hand, collecting text data from the Web is a widely used approach to building language models of automatic speech recognition systems (Bulyko et al., 2003; Sarikaya et al., 2005; Ng et al., 2005; Tsiartas et al., 2010). Web texts are expected to be more natural than generated pseudo sentences because users handcraft most of them. However, Web texts contain diverse domain texts; thus, we need some criteria to select appropriate texts to be used for the augmented training data. Test-set perplexity (Misu and Kawahara, 2006) or semantic similarity (Hakkani-Tur and Rahim, 2006; Yoshino et al., 2013) were widely used as criteria to select appropriate sentences for the training data augmentation. Such a selective approach using large-scale Web data has been applied to the data augmentation of dialogue systems (Du and Black, 2018; Henderson et al., 2019).

Proceedings of the 28th International Conference on Computational Linguistics, pages 2606–2612
Barcelona, Spain (Online), December 8-13, 2020

Crowdsourcing is a common way to collect human-annotated data at low-cost (Zhao et al., 2011; Mozafari et al., 2014). However, accurate SLU systems based on neural networks require a large-scale dataset. It is not easy to collect sufficient amount of training data only using crowdsourcing even if the cost of crowdsourcing is less than normal annotators.

In this paper, we utilize two sources of the "wisdom of crowd" for collecting a large-scale dataset to train the SLU system. We firstly collect a small amount of seed data by using crowdsourcing and then augment the dataset with similar texts extracted from the Web. As the target Web texts for the extraction, we focus on the online knowledge community website as another "wisdom of crowds." Online knowledge community websites often contain qualified question-style sentences. We choose sentences similar to the seed data from the qualified sentences and use them for SLU training. We conducted experiments to investigate the relationship between the accuracies of SLU systems and their training dataset augmented by sentences from the knowledge community website. As the result, over 120,000 sentences were extracted from the Web, and we got 35 points improvement in accuracy on domain selection of SLU.

2 Spoken language understanding based on crowdsourcing

Our task is to develop an SLU system for a new domain with no available resources. We build a small amount of training data via crowdsourcing, then use the collected data as the seed of data augmentation. We describe the task definition of SLU and the seed data collection using crowdsourcing, in this section.

2.1 Task definition

The task of SLU is defined as the prediction of a dialogue frame F given a user request X with a word sequence $x_1, x_2, ..., x_n$. The dialogue frame expresses user intent with information of domain, category, and query (Williams et al., 2014). Domain indicates a dialogue topic corresponding to the system function. We defined six domains in this work: "video", "weather", "news", "map", "shop" and "recipe". The category is a refined dialogue topic, which shares possible queries. For example, "movie" and "live" are defined as a part of categories belonging to the "video" domain. Query contains slot-values required for predicted domain and category; in our case, "keyword", "date", "location", "state", "from", "to", "use" and "not_use". We defined these domains, categories and queries by selecting typical applications of Yahoo!JAPAN speech assistant[1].

2.2 Seed data collection via crowdsourcing

We used crowdsourcing to collect various user request expressions to the given user intention. We gave a description of the intention in three sentences to crowdworkers without any example queries for collecting diverse variations. The instruction and example intention follows:

Instruction: You will see three sentences as your "intention," which describe your situation. Please input what you will say in that situation.
Intention: You would like to watch cats on video. Your mother has the remote control of the TV in the living room. What will you ask your mother?

We prepared 120 intentions to collect variations on six domains: 24 for "video", 19 for "weather", 18 for "news", 18 for "map", 18 for "shop", and 19 for "recipe" domains. 100 crowd workers worked for one intention; finally, we collected 12,000 user utterance variations.

3 Data augmentation using online knowledge community website

Crowdsourcing is a promising way to collect initial data; however, it costs linearly to the amount of data. It is not realistic to collect all the data using crowdsourcing. Thus, we propose to use the crowdsourced data as a seed data to augment the dataset. Another "Wisdom of Crowds," knowledge community website has the potential to be the augmented training data, because it consists of questions and answers in

[1] `https://v-assist.yahoo.co.jp/`, (October 23rd, 2020)

Table 1: Training data size as results of data augmentation.

	w/o augmentation	+crowd	+crowd+Web (th =)					
			0.90	0.88	0.86	0.84	0.82	0.80
$\#t$	74	7,400	8,048	9,868	23,312	129,093	773,874	3,592,743

Table 2: Accuracies and L2 scores of SLU results (D=domain, C=category, Q=query).

Metric	Method	Dev.			Test		
		D	C	Q	D	C	Q
Accuracy: (larger is better)	w/o augmentation	0.28	0.40	0.84	0.30	0.46	0.83
	+crowd	0.70	0.56	0.90	0.47	0.66	**0.89**
	+crowd+Web (th=0.84)	**0.89**	**0.57**	**0.92**	**0.82**	**0.72**	**0.89**
L2: (smaller is better)	w/o augmentation	0.81	0.82	0.25	0.82	0.75	0.28
	+crowd	0.47	0.72	**0.15**	0.47	0.53	**0.16**
	+crowd+Web (th=0.84)	**0.24**	**0.61**	**0.15**	**0.26**	**0.43**	**0.16**

spoken language style, which are similar to user queries. It is reported that such website data is useful for dialogue systems (Yoshino et al., 2013); we expect that the data also will be useful for SLU. We calculate similarities between any seed queries and any question sentences extracted from the knowledge community website for finding the best alignment between augmented sentences and user intents.

The ith user intent f_i has corresponding $q_{i,j}$ $(1 \leq j \leq J)$, which is a possible query collected in crowdsourcing (J is the number of queries assigned to intent f_i). We calculate similarities between each $q_{i,j}$ and c_k, which is a question sentence extracted from the knowledge community website, for finding similar sentence $\hat{c_k}$ to $q_{i,j}$. $\hat{c_k}$ which will be assigned as an agumented training sample of f_i.

Converting sentences to vector representations is an common approach to calculate similarities: vector space model (Salton et al., 1975), means of distributed representation of words (Mikolov et al., 2013; Le and Mikolov, 2014) and bi-directional long short-term memory neural networks (Bi-LSTM) (Cross and Huang, 2016; Yang et al., 2019). Recently, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is known as a better sentence encoder, which is based on masked word prediction in surrounding sentences.

We used the BERT model trained using Japanese Wikipedia (Sakata et al., 2019) on the task of masked word prediction because we would like to extract semantically similar sentences to seed queries. The task of masked word prediction is based on the distributional hypothesis (Harris, 1954); thus, the resultant model trained in the task can embed semantically similar sentences into close points on the latent space. We note the vector of sentence $q_{i,j}$ as $\mathbf{q}_{i,j}$. Because both vectors $\mathbf{q}_{i,j}$ and $\mathbf{c}_k$ have the same vector size, we define their similarity as the cosine between them as,

$$sim(\mathbf{q}_{i,j}, \mathbf{c}_k) = \cos(\mathbf{q}_{i,j}, \mathbf{c}_k) = \frac{\mathbf{q}_{i,j} \cdot \mathbf{c}_k}{|\mathbf{q}_{i,j}||\mathbf{c}_k|}. \quad (-1 \leq sim(\cdot) \leq 1) \tag{1}$$

4 Experiments in spoken language understanding

We investigated the effect of each data augmentation method in experiments in this section. We describe the SLU system that we used, experimental setting, and the results.

4.1 Spoken language understanding system

As described in Section 2.1, the task is estimating domain, category, and query that compose SLU output frames, given user requests. We used an incremental dialogue state tracker (Coman et al., 2019)[2], which showed a good performance in DSTC2 shared task (Williams et al., 2014). We trained this incremental dialogue state tracker on our dataset and then used the final results of the tracker as our SLU results. Note that our defined SLU task is hierarchical; however, the used dialogue state tracker predicts domain, category, and slot independently.

[2] `https://github.com/ahclab/iDST_iTTD`, (March 1st, 2020)

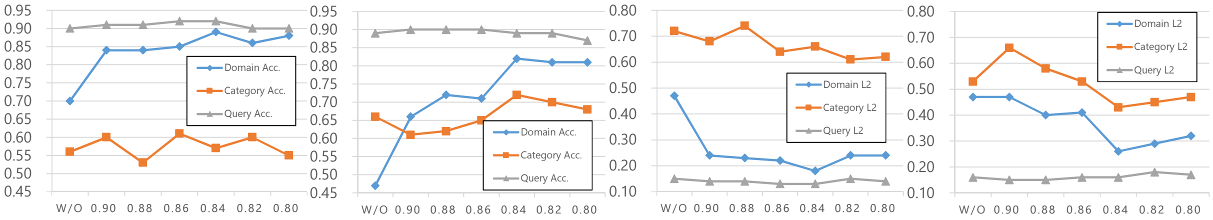

Figure 1: Accuracies and L2 on both dev. set and test set with augmentation thresholds. From left, each figure shows acc. on dev. set, acc. on test set, L2 on dev. set, and L2 on test set, respectively.

Table 3: Examples of data augmentation by each method. Score means $\mathrm{argmax}_j sim(\mathbf{q}_{i,j}, \mathbf{c}_k)$.

Added on:	Sentence	Score
Original	君の名は。を見たいんだけど (I'd like to watch "Your Name.") domain="video", category="movie", keyword="Your Name."	-
+Crowd	君の名は。ってもう見た？ (Have you ever seen "Your Name."?)	-
	君の名は。面白いんだってよ。 (I heard that "Your Name." is interesting)	-
+ Web	君の名は。は見ましたか？ (Did you watch "Your Name."?)	0.88
	君の名は。は面白かったですか？ (Was "Your Name." interesting?)	0.87
Original	東京の天気を教えて？ (Would you give me the weather in Tokyo?) domain="weather", category="forecast", location="Tokyo"	-
+Crowd	東京って今日雨降りそう？ (Is Tokyo likely to rain today?)	-
	東京の降水確率教えて (What is the chance of rain in Tokyo?)	-
+ Web	明日って横浜市雨降りますか？ (Will Yokohama likely to rain tomorrow?)	0.87
	今札幌で雪降ってるって本当ですか？ (Is it true that it is snowing in Sapporo now?)	0.84

4.2 Experimental setting

We compared three settings: without augmentation (single utterance sample is assigned to each user intent), using some crowdsourced data, and using some augmented data from the knowledge community website. We used Yahoo! QA website[3] as the target knowledge community website. We divided our dataset into three portions: training, development, and test sets. Note that samples with the same intent are put together in the same set. We used 2,000 samples as our development set and 2,600 samples as our test set, which was collected in crowdsourcing. We used two metrics according to existing work (Williams et al., 2014; Coman et al., 2019): Accuracy and L2. Accuracy means the accuracy of predicted labels (larger is better), and L2 means the squared error from the one-hot representation of the correct answer (smaller is better). They were calculated for domain, category and query, respectively.

4.3 Experimental results

First, we show the data size for each setting on Table 1: w/o augmentation, using crowdsourced data (+crowd), and using extracted online knowledge community queries in addition to them (+crowd+Web). $\#t$ indicates the number of samples, and th indicates the sample selection threshold. We also show accuracy and L2 for each setting on Table 2. The result showed that scores were improved by using crowdsourced data, and there were additional improvements if we also used the data augmented from the knowledge community website. The method using data augmentation from Web showed large improvements, 19 points on development set, and 35 points on the test set, than using crowdsourced data.

The threshold to select data from a knowledge community website is important. We investigated the threshold by grid search, as shown in Figure 1. These results indicate that we can select the threshold th as 0.84 from the development set, and the threshold achieved the best scores in the test set.

When we compare scores of each threshold to the score of the model only trained by crowdsourced data (w/o), their score decreased in some cases. From Table 1, scores were improved when the method can extract more samples from Web ($th \leq 0.86$). This result indicates that some queries augmented from the knowledge community website do not contribute to the SLU accuracy; however, the method

[3] https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html, (March 5th, 2020)

can improve scores with increasing the training data size even some of them are noisy. We show some augmentation examples in Table 3. We can see that the crowdsourcing (+Crowd) collected variations of a query in the same meaning with different expressions. In contrast, data augmentation results (+Web) contain some examples of different dialogue frames in similar expressions or different entities. This result suggests that the crowdsourcing is useful to collect query variations for SLU system, and the data augmentation from the knowledge community web site can improve the system with the amount of data. This result also indicates that all of the data on the Web is not useful for the training of statistical models; we have to establish a method to select appropriate data to be used for the training as mentioned in existing works (Misu and Kawahara, 2006; Yoshino et al., 2013; Akama et al., 2020).

5 Conclusion

In this paper, we proposed to use two sources of the "wisdom of crowds," crowdsourcing and knowledge community website, to augment the training data for SLU. We used BERT to calculate similarities between seed queries and queries extracted from a knowledge community website, as a converter from utterances to vectors. Experimental results showed that using both crowdsourcing and knowledge community website for data augmentation will improve the accuracy and robustness of the SLU system at low-cost.

The proposed architecture is evaluated only on written texts collected by crowdsourcing; thus, experimental evaluation with real recognized speech is essential for future work. In recent end-to-end SLUs, acoustic features are also important; however, it is difficult to collect such acoustic features from Web texts. One way to solve the problem is using a machine speech chain (Tjandra et al., 2020), which can generate pseudo acoustic features for the augmented query texts. Disambiguation to queries that have several meanings, and using other embedding methods such as RoBERTa (Liu et al., 2019) will be another future work.

References

Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Filtering noisy dialogue corpora by connectivity and content relatedness. In *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Ivan Bulyko, Mari Ostendorf, and Andreas Stolcke. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pages 7–9.

Andrei C Coman, Koichiro Yoshino, Yukitoshi Murase, Satoshi Nakamura, and Giuseppe Riccardi. 2019. An incremental turn-taking model for task-oriented dialog systems. In *Proceedings of Interspeech 2019*, pages 4155–4159.

James Cross and Liang Huang. 2016. Incremental parsing with minimal features using bi-directional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 32–37.

Deborah A Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *Proceedings of the workshop on Human Language Technology*, pages 43–48. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Wenchao Du and Alan Black. 2018. Data augmentation for neural online chats response selection. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 52–58, Brussels, Belgium, October. Association for Computational Linguistics.

D Hakkani-Tur and Mazin Rahim. 2006. Bootstrapping language models for spoken dialog systems from the world wide web. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy, July.

Chiori Hori, Julien Perez, Ryuichiro Higashinaka, Takaaki Hori, Y-Lan Boureau, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, Koichiro Yoshino, and Seokhwan Kim. 2019. Overview of the sixth dialog system technology challenge: Dstc6. *Computer Speech & Language*, 55:1–25.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Teruhisa Misu and Tatsuya Kawahara. 2006. A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts. In *Ninth International Conference on Spoken Language Processing*.

Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden. 2014. Scaling up crowdsourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment*, 8(2):125–136.

Tim Ng, Mari Ostendorf, Mei-Yuh Hwang, Manhung Siu, Ivan Bulyko, and Xin Lei. 2005. Web-data augmented language models for mandarin conversational speech recognition. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–589. IEEE.

Avik Ray, Yilin Shen, and Hongxia Jin. 2018. Robust spoken language understanding via paraphrasing. In *Proceedings of Interspeech 2018*, pages 3454–3458.

Amrita Saha, Rahul Aralikatte, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1116.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Ruhi Sarikaya, Agustin Gravano, and Yuqing Gao. 2005. Rapid language model development using external resources for new spoken dialog domains. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–573. IEEE.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989.

Andreas Tsiartas, Panayiotis Georgiou, and Shrikanth Narayanan. 2010. Language model adaptation using www documents obtained by utterance-based queries. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5406–5409. IEEE.

Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5370–5378. AAAI Press.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7402–7409.

Koichiro Yoshino, Shinsuke Mori, and Tatsuya Kawahara. 2013. Incorporating semantic information to selection of web texts for language model of spoken dialogue system. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8252–8256. IEEE.

Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. 2011. Incremental relabeling for active learning with noisy crowdsourced annotations. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 728–733. IEEE.