FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance

Wataru Sakata*
LINE Corporation
wataru.sakata@linecorp.com

Tomohide Shibata
Kyoto University
shibata@nlp.ist.i.kyoto-u.ac.jp

Ribeka Tanaka
Kyoto University
tanaka@nlp.ist.i.kyoto-u.ac.jp

Sadao Kurohashi
Kyoto University
kuro@nlp.ist.i.kyoto-u.ac.jp

ABSTRACT

Frequently Asked Question (FAQ) retrieval is an important task where the objective is to retrieve the appropriate Question-Answer (QA) pair from a database based on the user's query. In this study, we propose a FAQ retrieval system that considers the similarity between a user's query and a question computed by a traditional unsupervised information retrieval system, as well as the relevance between the query and an answer computed by the recently-proposed BERT model. By combining the rule-based approach and the flexible neural approach, the proposed system realizes robust FAQ retrieval. A common approach to FAQ retrieval is to construct labeled data for training, which takes a lot of costs. However, a FAQ database generally contains a too small number of QA pairs to train a model. To surmount this problem, we leverage FAQ sets that are similar to the one in question. We construct localgovFAQ dataset based on FAQ pages of administrative municipalities throughout Japan. In this research, we evaluate our approach on two datasets, localgovFAQ dataset and StackExchange dataset, and demonstrate that our proposed method works effectively.

ACM Reference Format:
Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance. In *SIGIR19: 42nd Intl ACM SIGIR Conference on Research and Development in Information Retrieval, July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

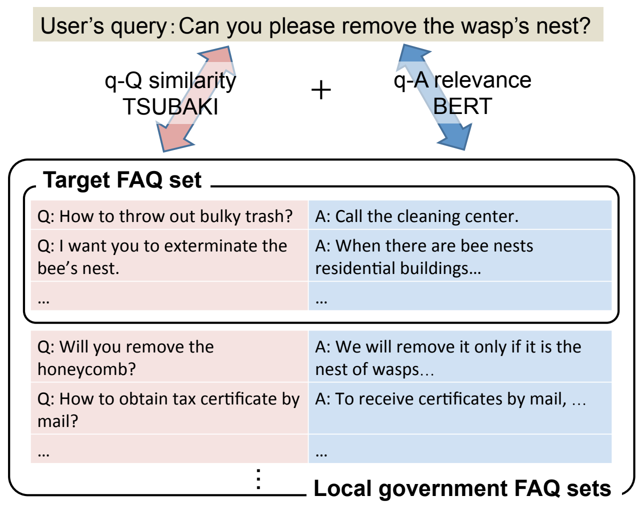

Figure 1: An overview of our proposed method.

1 INTRODUCTION

There are often frequently asked questions (FAQ) pages with various information on the web, like manufactures and administrative municipalities. A FAQ retrieval system, which takes a user's query and returns relevant QA pairs, is useful for navigating these pages.

In FAQ retrieval tasks, it is standard to check similarities of user's query (q) to a FAQ's question (Q) or to a question-answer (QA) pair [2, 3]. Many FAQ retrieval models use the dataset with the relevance label between q and a QA pair. However, it costs a lot to construct such labeled data. Another promising approach is to check the q-A relevance trained by QA pairs, which shows the plausibility of the FAQ answer for the given q. Studies of community QA use a large number of QA pairs for learning the q-A relevance [4, 10, 11]. However, these methods do not apply to FAQ retrieval task, because the size of QA entries in FAQ is generally too small to train a model.

We address this problem by collecting other similar FAQ sets to increase the size of available QA data. It is a reasonable assumption that one can find many similar FAQs provided in the target field.

In this study, we propose a method that combines the q-Q similarity obtained by unsupervised model and the q-A relevance learned from the collected QA pairs. Figure 1 shows the proposed model. Previous studies show that neural methods (e.g., LSTM and CNN) work effectively in learning q-A relevance. Here we use the recently-proposed model, BERT [1]. BERT is a powerful model that applies to a wide range of tasks and obtains the state-of-the-art results on many tasks including GLUE [9] and SQuAD [5]. An unsupervised retrieval system achieves high precision, but it is difficult to deal with a gap between the expressions of q and Q. By contrast, since BERT validates the relevance between q and A, it can retrieve an appropriate QA pair even if there is a lexical gap between q and Q. By combining characteristics of two models, we achieve a robust and high-performance retrieval system.

SIGIR, July 21-25, 2019, Paris, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

We conduct experiments on two datasets. The first one is the localgovFAQ dataset, which we construct to evaluate our model in a setting where other similar FAQ sets are available. It consists of QA pairs collected from Japanese local government FAQ pages and an evaluation set constructed via crowdsourcing. The second one is the StackExchange dataset [2], which is the public dataset constructed for FAQ retrieval tasks. We evaluate our model on these datasets and show that the proposed method works effectively in FAQ retrieval.

2 PROPOSED METHOD

2.1 Task Description

We begin by formally defining the task of FAQ retrieval. Here, we focus on local government FAQ as an example. Suppose that the number of local government FAQ sets is N. Our target FAQ set, T_t, is one of them. When the number of QA entries in T_t is M, T_t is a collection of QA pairs $\{(Q_1, A_1), (Q_2, A_2), ..., (Q_M, A_M)\}$. The task is then to find the appropriate QA pair (Q_i, A_i) from T_t based on a user's query q. We use $T_1, T_2, ..., T_N$ as our training data, including the FAQ set T_t of the target local government.

2.2 q-Q similarity by TSUBAKI

We use TSUBAKI [?] to compute q-Q similarity. TSUBAKI is an unsupervised retrieval engine based on OKAPI BM25 [?]. TSUBAKI accounts for a dependency structure of a sentence, not just its words, to provide accurate retrieval. For flexible matching, it also uses synonyms automatically extracted from dictionaries and Web corpus. The similarity of the given document d to search query q, $Similarity(q, d)$, is computed according to the following equation. $BM(t, d)$ follows the Okapi BM25 formula.

$$Similarity(q, d) = (1 - \beta) \sum_{t \in T_{qw}} BM(t, d) + \beta \sum_{t \in T_{qd}} BM(t, d)$$

T_{qw} and T_{qd} are a set of search words in q and a set of search dependency relation in q, respectively. Parameter β is to regulate the extent of dependency relations used in the scoring, and we use $\beta = 0.2$. Here we regard Q in each QA as a document and compute $Similarity(q, Q)$ for the q-Q similarity.

2.3 q-A relevance by BERT

We use BERT to compute q-A relevance. BERT is based on the Transformer [8] that effectively encodes an input text. It is designed to be pre-trained using a language model objective on a large raw corpus and fine-tuned for each specific task including sentence classification, sentence-pair classification, and question answering. As it is pre-trained on a large corpus, BERT achieves high accuracy even if the data size of the current task is not large enough. We apply BERT to a sentence-pair classifier for questions and answers. By applying the Transformer to the input question and answer, it effectively captures the relevance between the pair.

The training data we use is the collection of QA pairs from FAQ sets (see Sec. 2.1). For each positive example (Q, A), we randomly select $\bar{A}$ and produce negative training data $(Q, \bar{A})$. On this data, we train BERT to solve the two-class classification problem: $Relevance(Q, A)$ is 1 and $Relevance(Q, \bar{A})$ is 0, where $Relevance(Q, A)$ stands for the relevance between Q and A.

- I'd like you to issue a copy of family register, but how much does it cost?
- I'd like you to publish a maternal and child health handbook, but what is required for the procedure?
- I'm thinking of purchasing a new housing, so I want to know about the reduction measure.
- From which station does the pick-up bus of the Center Pool come out?

Figure 2: Examples of queries collected via crowdsourcing.

At the search stage, we compute $Relevance(q, A_i)$ $(i = 1, \cdots, M)$ for every QA pair in the target T_t and the user's query q. QA pairs in a higher rank are used as search results.

2.4 Combining TSUBAKI and BERT

In order to realize robust and flexible matching, we combine the q-Q similarity by TSUBAKI and the q-A relevance by BERT.

When TSUBAKI's similarity score on a given QA pair is high, the pair is probably a positive case because the words in q and Q highly overlap with each other. However, it is difficult to cope with the lexical gaps between q and Q. On the other hand, since BERT validates the relevance between q and A, it can retrieve an appropriate QA pair even if there is a lexical gap between q and Q. To make use of these characteristics, we combine two methods as follows. First, we take the ten-highest results of BERT's output. For QA pairs whose TSUBAKI score gets a higher score than α, we rank them in order of TSUBAKI's score. For the others, we rank them in order of the sum of the TSUBAKI's score and the BERT's score.

TSUBAKI's score tends to be higher when the given query is longer. Hence, before taking the sum, we normalize TSUBAKI's score by using the numbers of content words and dependency relations in the query. We divide the original score by the following value.[1]

$$Count(ContentWords) \times k_1 + Count(DependencyRelations) \times k_2$$

3 EXPERIMENTS AND EVALUATION

We conducted our experiments on two datasets, *localgovFAQ* and *StackExchange*. We constructed localgovFAQ dataset, as explained in Sec 3.1. StackExchange is constructed in the paper [2], which consists of 719 QA pairs. Each Q has paraphrase queries, and the total number of queries is 1,250. All the models were evaluated using five-fold cross validation. In each validation, all the queries were split into training (60%), development (20%) and test (20%). The task is to estimate an appropriate QA pair for each query q among 719 QA pairs.

3.1 LocalgovFAQ Evaluation Set Construction

Amagasaki-city, a relatively large city in Japan, was chosen as a target government, whose Web site has 1,786 QA pairs. First, queries to this government were collected using a crowdsourcing. Example queries are shown in Figure 2. We collected 990 queries in total.

[1] We do not normalize the BERT's score because it takes a value between 0 to 1.

TSUBAKI and BERT output at most five relevant QA pairs for each query, and each QA pair was manually evaluated assigning one of the following four categories:

A Contain correct information.
B Contain relevant information.
C The topic is same as a query, but do not contain relevant information.
D Contain only irrelevant information.

In general, information retrieval evaluation based on the pooling method has inherently a *biased* problem. To alleviate this problem, when there are no relevant QA pairs among the outputs by TSUBAKI and BERT, a correct QA pair was searched by using appropriate different keywords. If there are no relevant QA pair found, this query was excluded from our evaluation set. The resultant queries were 784. Since 20% of queries were used for the development set, 627 queries were used for our evaluation.

3.2 Experimental Settings

For the localgovFAQ dataset, MAP (Mean Average Precision), MRR (Mean Reciprocal Rank), P@5 (Precision at 5), SR@k (Success Rate)[2] and nDCG (normalized Discounted Cumulative Gain) were used as our evaluation measures. The categories A, B, and C were regarded as correct for MAP, MRR, P@5, and SR@k, and the evaluation level of categories A, B, C was regarded as 3, 2, 1, respectively for nDCG. For the StackExchange dataset MAP, MRR and P@5 were used, following Karan et al. [2] .

The pre-training of BERT was performed using Japanese Wikipedia, which consists of approximately 18M sentences, and the fine-tuning was performed using FAQs of 21 Japanese local governments. It consists of approximately 20K QA pairs. The morphological analyzer Juman++[3] was applied to input texts for word segmentation, and words were broken into subwords by applying BPE [6]. For English BERT pre-trained model, a publicly-available model was used[4]. For the fine-tuning for StackExchange dataset, the training set (q, Q, A) was divided into (q, A) and (Q, A).

In the localgovFAQ dataset, Bi-LSTM with attention [7] was adopted as our baseline. A question and an answer were encoded using Bi-directional LSTMs (word embeddings were initialized as word vectors obtained using word2vec), and the query embedding was obtained using the forward and backward LSTM outputs of the answer with an attention mechanism, and the answer embedding was obtained in the same way. Then, the concatenation of the query and answer embeddings were input to an MLP to output a binary vector (relevant or not). An unsupervised method TSUBAKI was applied to only Q as well as the concatenation of Q and A. In the StackExchange dataset, CNN-rank for q-Q and q-A settings was used, whose scores were from Karan et al. [2]. Furthermore, BERT (w/o query paraphrases) was adopted, where (q, A) pairs were not used for BERT training, to see the performance when no manually-assigned query paraphrases were available.

For both BERT and Bi-LSTM models, 24 negative samples for one positive sample were used. For the coefficients explained in Sec. 2.4, k_1, and k_2 were set to 4 and 2, respectively, and α was set to 0.3 using the development set.

[2]Success Rate is the fraction of questions for which at least one related question is ranked among the top k [4].
[3]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN++
[4]https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-12_H-768_A-12.zip

	Model	MAP	MRR	P@5	SR@1	SR@5	NDCG
q-Q	TSUBAKI	0.558	0.598	0.297	0.504	0.734	0.501
q-A	Bi-LSTM	0.451	0.498	0.248	0.379	0.601	0.496
	BERT	0.576	0.631	0.333	0.509	0.810	0.560
q-QA	TSUBAKI	0.395	0.422	0.220	0.348	0.511	0.357
	Proposed	**0.646**	**0.705**	**0.376**	**0.611**	**0.837**	**0.619**

Table 1: Evaluation result on the localgovFAQ dataset.

	Model	MAP	MRR	P@5
q-Q	CNN-rank	0.79	0.77	0.63
	TSUBAKI	0.698	0.669	0.638
q-A	BERT (w/o query paraphrases)	0.631	0.805	0.546
	BERT	0.887	0.936	0.770
q-QA	CNN-rank	0.74	0.84	0.62
	Proposed	**0.897**	**0.942**	**0.776**

Table 2: Evaluation result on the StackExchange dataset.

3.3 Evaluation Results and Discussion

Table 1 shows an experimental result on localgovFAQ dataset. In q-A setting, BERT was better than the Bi-LSTM baseline, which indicates BERT was useful for this task. Although the performance of TSUBAKI in q-Q setting and BERT (in q-A setting) is almost the same in terms of SR@1, the performance of BERT was better than TSUBAKI in q-Q setting in terms of SR@5, which indicates BERT could retrieve a variety of QA pairs. The proposed method performed the best. This demonstrated the effectiveness of our proposed method. The performance of TSUBAKI in q-QA setting was worse than one of TSUBAKI in q-Q setting, which indicates that simply using both Q and A in the unsupervised information retrieval system did not work well.

Table 2 shows an experimental result on StackExchange. In the same as the result on localgovFAQ, BERT performed well, and the proposed method performed the best in terms of all the measures. The performance of BERT was better than one of "BERT (w/o query paraphrases)", which indicates that the use of various augmented questions was effective.

Figure 3 shows the performance of TSUBAKI and BERT according to their TOP1 scores. From this figure, it can be found that in the retrieved QA pair whose TSUBAKI score is high, its accuracy is very high. On the otherhand, there is a relatively loose correlation between the accuracy and BERT score. This indicates TSUBAKI and BERT have different characteristics, and our proposed combining method is reasonable.

Table 3 shows an example of system outputs and their manual evaluations. In the first example, although TSUBAKI retrieved the wrong QA pair since there is a word "consultation" and "counseling" in the query and Q, BERT and the proposed method could retrieve a correct QA pair.

In the second example, the proposed method could retrieve a correct QA pair on the first rank although the first rank of TSUBAKI and BERT was wrong one.

Query	TSUBAKI		BERT		Proposed method	
Is there a consultation desk for workplace harassment?	1 ×	Q: I'd like to have a career counseling. A: Consultation place: Amagasaki-city, ...	1 ✓	Q: I'd like to consult a lawyer for work-related problems. A: On specialized and sophisticated labor issues such as wages, dismissal, occupational accidents, ...	1 ✓	Q: Can we get a lawyer's labor counselor? A: On specialized and sophisticated labor issues such as wages, dismissal, occupational accidents, ...
Where should I renew my license?	1 ×	Q: Where should I apply for medical staff licenses (new, corrected / rewritten, re-issued)? A: License application for doctors, dentists, public health nurses ...	1 ×	Q: To update a bus ticket, do I have to go myself? A: In principle, please apply for the application by yourself. ...	1 ✓	Q: Please tell me about the procedure of updating your driver's license. A: Regarding the renewal procedure of your driver's license ...
	2 ✓	Q: Please tell me about the procedure of updating your driver's license. A: Regarding the renewal procedure of your driver's license ...	2 ×	Q: Can I file an agent application to renew my bus ticket? A: As a general rule, please apply for yourself. ...	2 ×	Q: Can I file an agent application to renew my bus ticket? A: As a general rule, please apply for yourself. ...
Is there a place that we can use for practicing instruments?	1 ×	Q: Where is the location of the polling place before the election's due date? A: There are three polling stations before the date in the city. ...	1 ×	Q: Please tell me about Amagasaki City boys music corps. A: "Amagasaki City Boys Music Club" includes a choir corps, a brass band, ...	1 ×	Q: Please tell me about Amagasaki City boys music corps. A: "Amagasaki City Boys Music Club" includes a choir corps, a brass band, ...

Table 3: An example of system outputs and their manual evaluations. (✓ and × in the table mean correct and incorrect, respectively, where the evaluation categories A, B, and C are regarded as correct.)

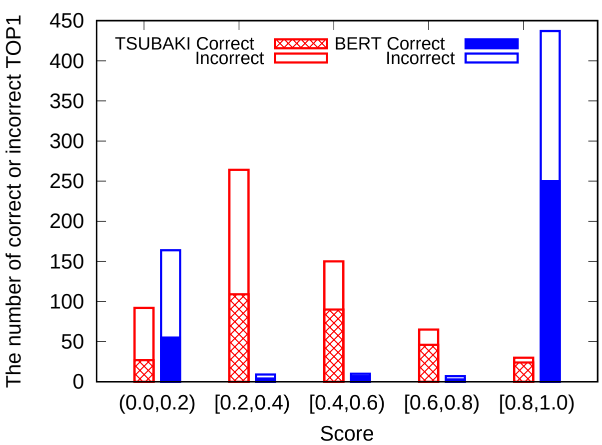

Figure 3: The relationship between the score and the number of queries whose TOP1 outputs of TSUBAKI/BERT were correct or incorrect.

In the third example, no methods could retrieve a correct QA pair. Although BERT could capture the relevance between a word "instruments" in the query and "music" in *A*, the retrieved QA pair was wrong. In an example of correct QA pair, *Q* is "Information on the facility of the youth center, hours of use, and closed day", and *A* part includes the information that the youth center has a music room, and the citizens can use facilities in the center. To retrieve this correct QA pair, the deeper understanding of QA texts is necessary.

4 CONCLUSION

This paper presented a method for using query-question similarity and BERT-based query-answer relevance in the FAQ retrieval task. By focusing on the fact that there are other FAQ sets in the same field, the size of available QA data can be increased. BERT, which has been recently proposed, was applied to capture the relevance between queries and answers. This method realized the robust and high-performance retrieval. The experimental results demonstrated that our combined use of query-question similarity and query-answer relevance was effective. We are planning to make the constructed dataset localgovFAQ publicly available.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805

[2] Mladen Karan and Jan Snajder. 2018. Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. *Expert Systems with Applications* 91 (2018), 418–433.

[3] Alejandro Moreo, Eduardo M. Eisman, Jorge Luís Castro, and José Manuel Zurita. 2013. Learning regular expressions to template-based FAQ retrieval systems. *Knowledge-Based Systems* 53 (2013), 108–128.

[4] Deepak P, Dinesh Garg, and Shirish Shevade. 2017. Latent Space Embedding for Retrieval in Question-Answer Archives. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 855–865.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392.

[6] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1715–1725.

[7] Ming Tan, Bing Xiang, and Bowen Zhou. 2015. LSTM-based Deep Learning Models for Non-factoid Answer Selection. *CoRR* abs/1511.04108 (2015). arXiv:1511.04108

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of Neural Information Processing Systems 2017*. Long Beach, CA, USA, 5998–6008.

[9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop*

BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, Brussels, Belgium, 353–355.

[10] Wei Wu, Xu SUN, and Houfeng WANG. 2018. Question Condensing Networks for Answer Selection in Community Question Answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1746–1755.

[11] Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots. *CoRR* abs/1805.02333 (2018). arXiv:1805.02333