Automatic Question Generation for Chatbot Development

Ryusei Doi
AAII, Department of Data Science
Musashino University
Tokyo, Japan
s2022048@stu.musashino-u.ac.jp

Thatsanee Charoenporn
AAII, Department of Data Science
Musashino University
Tokyo, Japan
thatsane@musashino-u.ac.jp

Virach Sornlertlamvanich
AAII, Department of Data Science
Musashino University
Tokyo, Japan
Faculty of Engineering
Thammsat University
Pathum Thani, Thailand
ORCID: 0000-0002-6918-8713

***Abstract*— It is a labor intensive task to prepare a list of questions for the intents in creating a chatbot system. It is not so easy to predict the variation of questions to be matched with the answers or what we want to provide the information to the chatbot users. In this research, we aim to solve the problem by applying a transformer framework to generate a question sentence from a list of keywords and explanatory text. In this paper, the question text is generated by using a trained Japanese T5 model by applying a list of keywords extracted from the questions and the explanatory text of the Amagasaki City FAQ database. To expand the coverage of the questions in matching the chatbot intent, we apply WordNet to expand the keywords in the questions. Finally, Semantic Textual Similarity (STS) Sentence-BERT is applied to measure the similarity of the user query and the question list in the chatbot.**

Keywords*— chatbot, T5 model, WordNet, Semantic Textual Similarity (STS), Sentence-BERT

I. INTRODUCTION

In recent years, aiming to reduce the personnel cost, chatbots have been used in many service tasks such as customer service and information desk. However, it is a labor intensive task in preparing a question list for the intents in creating a chatbot. It is a problem of looking up a sentence in the FAQ database. It is not trivial to assume a set of variation of the questions which can be properly matched with the user queries.

Word expansion is one of the common approaches using to expand the matching coverage between user query and the questions in the FAQ database. This approach can be expected to solve the problem of mismatching due to the word variation in the expression or the synonym, for example "What is the price of …?" can be asked by saying "How much is …?" or "What does it cost …?". In our preliminary experiment, we utilize the synset of WordNet [1] to expand the wordform, discarding the multiple word sense problem by including all possible words found in the synsets. But the result does not show much improvement in question matching rate and it consumes a lot time and memory to include all the combination of the words from the synsets.

Instead of expanding the word by its synonyms, we generate other related questions from the question and answer in the FAQ database. The questions can be generated by using Text-to-Text Transfer Transformer (T5) [2, 3], which is a model using the transfer learning model Transformer developed by Google. We use the answer in FAQ as the context and the extracted content words from the question in FAQ as the keywords for T5 to generate a new question. It is expected that based on the large scale pre-trained model, the questions in other variation of expressions can be generated.

We found that the simple cosine similarity measurement between sentences cannot find the proper questions. This is because of the difference in expression and wordform using in the sentences to compare. Actually, the cosine similarity method computes similarity of the sum of word vectors appeared in the sentences. It does not include the word context which is the disambiguation key information for word sense disambiguation. Especially in the case of user free input query, the sentence can be varied a lot in expressing a specific question.

To improve the matching rate between the user query and questions in FAQ, we utilize Semantic Textual Similarity (STS) Sentence-BERT (SBERT) model to measure the semantic similarity between the user query and question. In our experiment, we fine-tune the Japanese Sentence-BERT model[1] which is generated from the base model by Tohoku Univ. NLP Lab[2].

II. LITERATURE REVIEW

ELIZA [4] is the very first chatbot introduced prior to the development of the first personalized computer. In 1966, Joseph Weizenbaum developed ELIZA at the MIT Artificial Intelligence Laboratory. ELIZA receives the keywords as an input to trigger the output process encoded in a set of rules. Along with the progress in natural language processing (NLP) research, natural language understanding (NLU) has been introduced together with a set of NLP methodologies including syntactic parsing, part-of-speech tagging, named entity recognition, topic modeling and so on. Pattern matching based on the results from NLP is the technique commonly used by almost all chatbots. One of the difficulties obstructs the success in pattern matching approach is the variation of user expression in conversation. For instance, Hi Siri, "What time is it now?", "Do you have the time?", "Can you tell me the time?", "Have you got the time?" are the queries with the same intention of asking the time. It is more efficient to set the user's wish to ask the chatbot by an intention [5]. Relevant labeled datasets are needed for training the chatbot intent classification.

T5 (Text-to-Text Transfer Transformer) is an end-to-end transformer based architecture that uses a text-to-text approach. Every task including translation, question answering, and classification is cast as feeding the model text as input and training it to generate some target text. This allows for the use of the same model, loss function,

[1] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

[2] https://github.com/cl-tohoku/bert-japanese

hyperparameters, etc. across the diverse set of tasks [3]. Transfer learning, where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task, such as the question generation to serve chatbot development.

The combination score of query and question similarity, and query and answer relevance is successfully proposed by [6] to achieve the FAQ retrieval system for the Amagasaki FAQ database. Lexical gap is reported to be an issue for matching the query. To relax the matching, synonym based word expansion is introduced in computing the similarity between query and question but consume a high computation resource.

III. DATASET

We use the Japanese administrative municipality domain FAQ database (AmagasakiFAQ) which is prepared by the Amagasaki city local government. It is an FAQ database containing a set of 1,786 questions and the corresponding answers in FAQ page of Amagasaki city. The FAQ dataset is quite large and manually prepared to give the responsive answer about the city.

TABLE I. AN EXAMPLE OF A PAIR OF QUESTION AND ANSWER IN AMAGASAKI FAQ

No.	Question (Q)	Answer (A)
1	How do I get to the Imakita Regional General Center?	Imakita Regional General Center does not have enough parking lots, so please use the city bus. Please come to "Tachibana Station" by the JR line, "Tsukaguchi Station" and "Mukonoso Station" by the Hankyu Line, and "Amasaki Station", "Mukogawa Station" and "Deyashiki Station" by the Hanshin Line, and then use the city bus. Which station are you from? 1. From JR Tachibana Station (location is about a 10-minute walk to the southwest). 2. From Hankyu Tsukaguchi Station (south). 3. From Hankyu Mukonoso Station (south). 4. From Hanshin Amagasaki Station (north). 5. From Hanshin Mukogawa Station. 6. From Hanshin Deyashiki Station (north). <Revised> [Related FAQ] I want to know about the Regional General Center. <Revised> [Inquiry] Imakita Regional General Center 3-14-1 Nishitachibanacho, Amagasaki City. Phone 06-6416-5729.

Table I shows an example of a pair of question and answer. Though there is no detail of how the FAQ is prepared, it can be observed that the questions are manually prepared based on the given answers of the city related information. Almost all the questions are to ask about a part of the information in the given answers.

To test the our proposed method in preparing questions for intent development for a chatbot, we apply our approach to evaluate the accuracy of similarity measure against the test set of 784 queries prepared by Kyoto University from crowdsourcing according to the FAQ explanatory answers [6].

The expression of query is different from the question in FAQ but have exactly the same meaning. However, the answer shows much more information about the detail condition in mailing the resident card.

The test set gives more candidate of answers in three groups of relation, that is relevance, relate, and same group. We group all the related answers into a list of relevant answers to measure the similarity in the evaluation process in the next Section.

TABLE II. AN EXAMPLE OF A PAIR OF USER QUERY AND THE MATCHED QUESTION AND ANSWER IN AMAGASAKI FAQ

No.	Query (q)	No.	Question (Q)	Answer (A)
86	Can you mail me a copy of my resident card?	62	Can I have a copy of my resident card mailed to me?	A copy of the residence certificate can be requested by mail from the person or a person in the same household. In the case of a request from a third party (other than the person or a person in the same household as the person), a power of attorney from the person is required. If you have not been delegated by the person, or if you are requesting mail from a corporation, public service, lawyer, etc., please contact the Citizens Division. However, the resident's card with my number can only be obtained by the person or a member of the same household. Please see the following link for details. [URL]. <Revised> [Related FAQ] What kind of content is included in the copy of the resident's card, and how much is the fee? Can an agent obtain a resident card with my number? <Revised> [inquiry] Citizen

No.	Query (q)	No.	Question (Q)	Answer (A)
				Service Department, Citizen Collaboration Bureau. Citizens Division. Phone 06-6489-6408. Inquiry time. From 8:45 am to 5:30 pm. However, the counter handling hours are from 9:00 am to 5:30 pm. holiday. Saturdays, Sundays, national holidays, year-end and New Year holidays (December 29-January 3).

IV. Question Generation and STS SBERT Fine Tuning

In the question generation process, we extract noun, verb, adverb, and adjective words from the question to create a list of keywords, and use the answer as the context for T5 to generate a corresponding question. We use the default hyperparameter to generate only one output to reduce the complexity in evaluation. In the practical use case in chatbot, some number of questions may needed to extend the possibility to match with other information in the answer. However, list of keywords in concern is needed to prepare corresponding the part of information provided in the answers.

From Table I, the list of content words, ('How', 'get to', 'Imakita' 'Regional', 'General', 'Center'), is extracted from the question (Q) to use as the keyword list, and the answer (A) is used as the context for T5 to generate a question which is shown in the generated question (Q') in Table III.

TABLE III. AN EXAMPLE OF GENERATED QUESTION ACCORDING TO THE QUESTION AND ANSWER IN AMAGASAKI FAQ

No.	Generated Question (Q')	Question (Q)	Answer (A)
1	What bus stops are there for the Imakita Regional General Center?	How do I get to the Imakita Regional General Center?	Imakita Regional General Center does not have enough parking lots, so please use the city bus. Please come to "Tachibana Station" by the JR line, ….

Table III shows the generated question according to the question and answer in the FAQ. The generated question still requests for the same information as in the question but has the different expression. This is because the keywords from the question are provided in the generation process. The result of question generation can be used to serve the variants of question in the intent of the chatbot.

To improve the sentence similarity measure, instead of using word vector cosine similarity, we fine-tune the Japanese SBERT model for Semantic Textual Similarity (STS) measure [7]. The triplet loss function is used to fine-tune the model. The negative sentence is randomly generated from the labeled sentences of the same positive group to make the triplet of (anchor, positive, negative). The loss minimizes the distance between anchor and positive while it maximizes the distance between anchor and negative. Table III shows an example of the labeled sentences of generated question (Q'), question (Q) and the answer (A) of the sentence label number 1.

In each iteration of the fine-tuning process, an anchor vector (v1) is selected to focus on. A positive vector (v2) from the same group as the anchor vector (v1) and a negative vector (v3) from a different group are selected for comparison. The distance between v1 and v2 (anchor and positive) is minimized while the distance between v1 and v3 (anchor and negative) is maximized as shown in (1).

$$Loss = max(distance(anchor, positive) - distance(anchor, negative) + margin, 0) \quad (1)$$

V. Evaluation

We evaluate the feasibility of the generated question (Q') by measuring the STS similarity with the question (Q). In this case, it will show how appropriate to use the generated questions for the intent of a chatbot. By the way, the keywords for questioning must be listed up to guide T5 to generate the proper question out of the provided answer.

Table IV shows the result of the success in matching between Q' and Q. The result of Top5 and Top10 shows the matched sentence found within the top ranked five and ten sentences, respectively. The result also compares the performance of the original SBERT and the fine-tuned SBERT model used in the STS similarity measure.

TABLE IV. ACCURACY IN SIMILARITY MEASURE BETWEEN QUESTION (Q) AND GENERATED QUESTION (Q')

sim(Q, Q')	Top5	Top10
SBERT	0.3947	0.4765
Fine-tuned SBERT	0.5140	0.6165

The fine-tuned SBERT model outperforms the original SBERT in measuring semantic similarity. The result of accuracy shows that the generated question (Q') can somehow be used as a candidate question in the intent for chatbot development.

TABLE V. ACCURACY IN SIMILARITY MEASURE BETWEEN QUERY (Q) AND QUESTION (Q)

sim(q, Q)	mAP@n	Top5
SBERT	0.3600	0.6543
Fine-tuned SBERT	**0.4757**	**0.7577**

In case that the questions are manually prepared, the fine-tuned SBERT also shows its improvement in matching between query (q) in the test set of the 784 queries and question (Q) with the STS measurement. mAP@n is the mean of average precision (AP) when measured throughout the all n output sentences. The precision (P) is counted in only when

the relevant sentence is found. The precision is averaged by the number of relevance of a sentence to produce the average precision (AP). Finally, the mean of average precision (AP) of all sentences in the test set is calculated to produce the mean average precision (mAP). Table V shows the improvement of fine-tuned SBERT in both mAP@n and Top5 measurements.

TABLE VI. ACCURACY IN SIMILARITY MEASURE BETWEEN QUERY (Q) AND GENERATED QUESTION (Q')

sim(q, Q')	mAP@n	Top5
SBERT	0.2510	0.5242
Fine-tuned SBERT	0.3327	0.6518

Up to this point, we can observe that there is a similarity between the question (Q) and the generated question (Q') since we use the content words extracted from the question (Q) to generate the list of keywords for question generation in T5. The variance of the question help in expanding the coverage of the expression. The results of Table V and IV show that the fine-tuned SBERT helps improving the matching between the query (q) and the question (Q). However, the success rate of matching drops when we apply the similarity measure between the query (q) and the generated question (Q'). The human created question can somehow outperform the matching rate comparing to the T5 question generation model.

Furthermore, we conduct some additional experiments to confirm on the contribution of the information from other available sources (Q, A, Q') to the similarity measure of the query (q). Only the similarity measure by STS fine-tuned SBERT is used in this comparison because it already shows its outperformance to the word vector and the original SBERT similarity measure. We combine the similarity score between the query (q) and other combination of available sources (Q, A, Q') and re-rank the result to measure the mean average precision (mAP@n).

TABLE VII. COMPARISON OF SIMILARITY SCORING METHOD

Similarity scoring	mAP@n
sim(q, Q')	0.3317
sim(q, A) + sim(q, Q')	0.4292
sim(q, Q) + sim(q, Q')	0.4609
sim(q, Q)	0.4757
sim(q, A) + sim(q, Q) + sim(q, Q')	0.4821
sim(q, A) + sim(q, Q)	**0.5081**

The sum of similarity between query (q) and answer (A), and between query (q) and question (Q) shows the highest precision by the mAP@n score. Consequently from the result in Table V, the human created question has the highest potential to match with the user query. Table VII also shows that the context from the answer (A) can also be used to find the proper match to the query (q). It is reasonable to use the answer additionally to support the matching. On the contrary, the generated question (Q') has a trend to decrease the precision. However, if there is not enough human created question at hand, the generated question can be alternatively used from the results in Table IV. Generating additional questions from the FAQ can help in terms of data expansion in the development of chatbot. If the accuracy of question generation is improved, better results can be expected.

The weight of similarity score is also observed to consider whether it is good to use answer (A) to match with query (q) directly or not. We change the weight for similarity score between query (q) and question (Q).

TABLE VIII. COMPARISON OF SIMILARITY SCORING METHOD

Similarity scoring	mAP@n
sim(q, A) + 8 x sim(q, Q)	0.4951
sim(q, A) + 4 x sim(q, Q)	0.4966
sim(q, A) + 2 x sim(q, Q)	0.5013
sim(q, A) + sim(q, Q)	**0.5081**

As shown in Table VIII, the weight of similarity score between query (q) and question (Q) is doubled in each step. The result consistently decreased when the weight of similarity score (q, Q) is increased. Therefore, it is confirmed that the answer (A) makes a good contribution in matching with the query, and the sum of the similarity by both question and answer can provide the best match answer to the query.

VI. CONCLUSION

Question generation by T5 is successfully conducted by using content words from the question and context from the answer in the FAQ. The generation question is experimentally validated by observing the results of similarity measure with the question in the FAQ. It is confirmed that the generated question can be served as a candidate question for creating the intent in chatbot development. To improve the matching rate, instead of using the word vector, the results of experiment show that the STS fine-tuned SBERT outperforms the word vector in similarity measure in the high variation of expression task.

ACKNOWLEDGMENT

This work was partially supported by the Thailand Science Research and Innovation Fundamental Fund, Contract Number TUFF19/2564 and TUFF24/2565.

REFERENCES

[1] C. Fellbaum, Ed. *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[2] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in *Proc. NAACL Conf.*, 2021, pp. 483–498, Online.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv preprint, arXiv:1910.10683, pp. 1-67, 2020.

[4] B. R. Ranoliya, N, Raghuwanshi, and S. Singh, "Chatbot for university related FAQs," in *Int. Conf. on ICACCI*, Sep. 2017, pp. 1525–1530.

[5] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," in *Artificial Intelligence Applications and Innovations*, Springer Inter. Publication, 2020, pp. 373-383

[6] W. Sakata, T. Shibata, R. Tanaka, and S. Kurohashi, "FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance," in *SIGIR'19*, Jul. 2019, pp. 1113–1116.

[7] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *EMNLP/IJCNLP*, 2019, pp. 3980-3990.