Insurance Question Answering via Single-turn Dialogue Modeling

Seon-Ok Na and **Young-Min Kim** * and **Seung-Hwan Cho**
Department of Industrial Data Engineering
Hanyang University, Seoul, Republic of Korea
{nso94,yngmnkim,shcho95}@hanyang.ac.kr

Abstract

With great success in single-turn question answering (QA), conversational QA is currently receiving considerable attention. Several studies have been conducted on this topic from different perspectives. However, building a real-world conversational system remains a challenge. This study introduces our ongoing project, which uses Korean QA data to develop a dialogue system in the insurance domain. The goal is to construct a system that provides informative responses to general insurance questions. We present the current results of single-turn QA. A unique aspect of our approach is that we borrow the concepts of intent detection and slot filling from task-oriented dialogue systems. We present details of the data construction process and the experimental results on both learning tasks.

1 Introduction

Although there has been significant progress in single-turn question answering (QA), it cannot cover complex questions of realistic scenarios (Fu et al., 2020). Recently, multi-turn (conversational) QA has emerged as an alternative to address this problem by clarifying the questions via conversation (Qu et al., 2019a, 2020; Li et al., 2019; Reddy et al., 2019). Conversational QA is a category of dialogue systems which are divided into task-oriented, chitchat, and QA systems (Deriu et al., 2020; Zaib et al., 2021). However, QA is not always distinct from the other two categories.

In this study, we are interested in building a dialogue system in a restricted domain, insurance. The system aims to provide users with general descriptions of cancer insurance. We assumed the task is not pre-defined and should be specified from the data. A significant difficulty is that complete conversational data does not exist. Therefore, we needed to find other types of source data similar to the dialogues between users and experts on cancer insurance. The first is Q&A data from a Korean online QA service.

Although our goal is to construct a multi-turn dialogue system, this study covers only the single-turn QA corresponding to the target system's front part. The novelty of the present study is that we designed the system considering the further extension to multi-turns. Therefore, unlike the existing KB-based or neural QA systems, we borrow the concept of intent detection and slot filling from task-oriented dialogue systems (Gao et al., 2018).

The Transformer-based pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT) achieved excellent performance for NLP tasks. BERT is one of the pioneers of the pre-trained language representation models (Devlin et al., 2018). Since it was proposed in 2018, a paradigm shift has taken place in the NLP domain. Most NLP tasks now are based on pre-trained language models. Meanwhile, there are also previous studies using directly BERT embeddings directly to express queries for conversational QA or FAQ retrieval (Qu et al., 2019b; Mass et al., 2020; Qu et al., 2020; Sakata et al., 2019). We use a Korean version of Electra (Clark et al., 2020), a variant of BERT, for intent detection and slot filling.

This study introduces intermediate results of our ongoing project on dialogue system construction in the insurance domain. We encountered many challenging situations from the first stage, data collection. We designed the system to be constructed using single-turn QA data but to finally serve as a multi-turn dialogue system. In the remainder of this paper, we describe the process of constructing training data for insurance QA in Section 2. Then the methods used for intent detection, slot filling, and the other approaches for the answer retrieval are presented in Section 3. Section 4 presents the experimental results for both learning tasks, including a quantitative analysis of the answer retrieval

[*] Corresponding author: Young-Min Kim

Proceedings of the 2nd Workshop on When Creative AI Meets Conversational AI (CAI2), COLING 2022, pages 35–41
October 12–17, 2022.

result. Then we conclude with some future works in Section 5.

2 Insurance Data for Question Answering

2.1 Data Collection and Preprocessing

Consumer counseling data in the insurance domain does not come to the public because of information privacy. Therefore, we collected single-turn Q&A data from an online QA service called Naver Knowledge iN (KiN)[1], which is part of the biggest Korean web portal, Naver. Login portal users can ask questions and the answerers voluntarily participate in the service. There are various subject sections in which tasks are allocated according to their nature. We scraped the Q&A pairs answered by 25 insurance experts in the insurance sector. The number of scraped Q&A pairs is 12,734.

For a realistic system, we limited our target to cancer insurance. We filtered out the pairs that did not include "cancer" in the title. The remaining data had three main issues for constructing a dialogue system. First, it is not conversational because Naver KiN consists of single-turn Q&A data that are inappropriate for dialogue systems. Second, both task-oriented and QA types of questions were mixed. Although the questions sought relevant information, some were relatively close to the task completion, such as recommendations or buying. Third, the intents and slots were not explicit. Though there are rough sub-categories of the questions, it is challenging to specify user intentions. Moreover, the primary entity types were unclear because the task was undefined. Considering all these issues, we began by defining the user intents and main tasks, unlike the general QA system. We then defined the appropriate slots to complete the task. This is also intended for further extension to multiple turns.

2.2 Topic Modeling

The user intentions in our data are not explicit, unlike typical task-oriented systems. Moreover, the questions are usually not represented clearly because general users do not know the insurance terminologies that precisely describe their situations. We also found that some sentences were literal questions, while the others provided information.

All of the above characteristics make defining intents challenging. Topic modeling can be an appropriate solution to address this problem. It identifies latent topics from a set of documents in an unsupervised manner. We applied the following LDA (Blei et al., 2003), on the question data to extract common themes of user questions.

Several preprocessing such as stopwords elimination and noun extraction have been applied before training the model. The number of topics was fixed to 30, considering perplexity, coherence, and manual validation of topic model results. We recategorized the extracted topics as six different upper topics: *recommendation*, *specific cancers*, *money-related*, *special contracts*, *particular insurance companies*, and *insurance terminologies*.

Then we manually classified each question into one of the six topics. To facilitate the process, we distributed keywords to each topic that represent the topics well. The candidate keywords were high-rank words of topic modeling results. A question with several keywords of a particular topic can be classified as the topic. Finally, we had the Q&A pairs with the topic label. The pairs of two topics, *recommendation* and *particular insurance companies*, were eliminated because the answers related to these two topics can be too subjective. For convenience, we excluded samples with more than 200 syllables in the question. Finally, we ahd 2,295 Q&A pairs as source data.

2.3 User Intents and Main Task

The Q&A pairs with the topic label were the source data for the system construction. We preprocessed the questions to imitate multi-turn conversations. A question was first separated into sentences, and we supposed that each corresponded to an utterance. Each sentence was a data instance in terms of intent detection.

We manually annotated each sentence with a user intent label considering the pre-annotated topic. A user intent here means a detailed purpose of the utterance. Therefore, it is different from the higher-level user intention that can be interpreted as a task. The finally defined intent types are listed in Table 1. In addition to the 2,295 Q&A pairs, we added manually generated 892 pairs to handle the class imbalance and data insufficiency issues. As an utterance can be a question or an information-offering one, the intents were also classified into two different categories: *Request* and *Inform*.

The high-level categories of *Request* intents may be interpreted as tasks for the dialogue system. For

[1] https://kin.naver.com/

Table 1: Intent type definition

Intent type	Definition	Action Type	Count
Personal information	Provide personal information for consultation	Inform	503
Subscription information	Subscribed insurance policy	Inform	480
Emphasis	Emphasis user request	Inform	147
Insurance options required	Options added to insurance policies	Inform	197
Cancer diagnosis details	A history of cancer diagnosis	Inform	149
Approximate premium or claim	Premium or claim that cannot be categorized into the others	Request	576
Claim availability	Questions about claim availability	Request	189
Claim process	Queries the insurance claims payment process	Request	63
Claim	Questions about claim with a stated amount	Request	75
Duplicate coverage	Questions about duplicate coverage availability	Request	85
Premium	Questions about premium with a stated amount	Request	67
Non-payment	Payment of unpaid insurance premiums	Request	41
Considerations	Questions to consider when subscribing to insurance	Request	80
Subscribe	Insurance policy subscription request	Request	288
Terminology Meaning	Request explain on Terminology terms	Request	95
Termination	Request for termination of insurance	Request	62
Greeting	Greeting	Greeting	94

example, *Claim availability*, *Claim process*, and *Claim* can be classified into a high-level category, *Claim-related*. Although the high-level category does not correspond to a task to complete like task-oriented systems; it can serve to reduce the scope of the QA. In other words, we borrowed the concepts of “task” and “slot” from task-oriented systems, expecting they could contribute to the clarification of user requirements.

2.4 Slots

The slots necessary for filling can be defined using concrete examples. We assumed several hypothetical conversation scenarios because the source data did not include conversational situations. Several slots can be defined from these scenarios. Moreover, we examined the frequent nouns extracted from the source data to determine whether they could be used as slot values. Finally, we obtained 11 slots, as presented in Table 2.

2.5 KB for Answers

Once the system recognizes what the user asks, it returns an appropriate answer. To this end, we constructed a KB as an FAQ, apart from the source Q&A pairs. We preferred choosing an operator from the KB over the source data because the real solutions are diverse for the same questions. The KB was constructed using FAQs provided by nine insurance companies and included various techniques, from common insurance sense to insurance products. There were 817 FAQ pairs. We also constructed an insurance terminology dictionary using term lists provided by four insurance companies.

3 Methods

3.1 Intent Detection and Slot Filling

We used a Korean ELECTRA version for intent detection and slot filling. ELECTRA is an efficient model which modified Masked LM in BERT to achieve performance similar to BERT with a lower computing power. Multilingual versions are also available, but a language-specific model generally outputs a better result.

Intent detection is interpreted as a classification problem, and slot filling corresponds to a sequence labeling task. The two models are trained separately using the same pre-trained model learned using Korean Wikipedia data. The selected pre-trained model is *KoElectra-base_v3* developed by monologg[2]. The model has been fine-tuned for both tasks.

3.2 Sentence BERT for FAQ mapping

Even if we finished the construction of the KB and the training data, we had a critical issue with building a dialogue system. We did not know which questions in the source data were answerable by the FAQ in KB. In other words, we needed the mappings between the source and KB questions. This process was for making a golden standard. Therefore, manual mapping is ideal; however, it is time-consuming.

Sentence-BERT(SBERT) can be an effective labor-saving tool. SBERT is a derivative model of BERT and is mainly used to calculate sentence expressions (Reimers and Gurevych, 2019). It has

[2] https://github.com/monologg/KoELECTRA

Table 2: Slot definition and the examples

Slot	Definition	Example	Count
GENDER	User's gender	여성(woman)	127
INSURANT	Family relation of the insurant	아버지(father)	189
COMPANY	Insurance company name	삼성생명(Samsung Life Insurance)	117
PAYMENT	Costs paid by users, including premium	보험료(premium)	773
CANCER_TYPE	Cancer to be covered by insurance	위암(cancer of the stomach)	151
DISEASE_LOG	User's disease history	위암(cancer of the stomach)	679
JARGON	Insurance-specific terms	고지의무(duty to notify)	907
PLAM_NAME	Name of insurance policy	내인생플러스보험(My Life Plus Insurance)	51
BODY_PART	Body parts subject to disease	갑상선(Thyroid)	157
OPERATION_LOG	User's surgical history	갑상선 수술(thyroid surgery)	288
INSURANCE_TYPE	Types of insurance	실손보험(indemnity insurance)	1,555

a pooling layer that is added to the existing BERT and uses Siamese Network and Triplet Network architectures. SBERT provides better sentence embeddings, especially when computing sentence similarity. Therefore, we used a Korean version of SBERT to map the source and KB questions.

The mapping process is as follows: 1) compute the SBERT embeddings of the source and FAQ questions, 2) for each source question, find the three most similar FAQ questions using cosine similarity, 3) manually map the source question and a FAQ question if the questions are semantically similar.

Two annotators carried out the mapping for cross-validation. After the annotation, approximately half of the source questions were mapped to the FAQ questions Q.

3.3 Symbol replacement in slot filling data via dictionary mapping

Among the slots, *PLAN_NAME* is difficult to detect because of its low occurrence and high diversity in values. Moreover, the slot values usually consist of multiple words and have descriptive phrases. These characteristics make recognizing the slot challenging. Another problem is that the newly-coined plan names continuously occur.

To enhance the detection performance of the slot, we invented a simple but effective heuristic method. The method uses a dictionary of insurance product names. The dictionary was constructed using real plan names scraped from insurance company websites. In addition to the original training data for slot filling, we added sentences with the masked product names. The added sentences had product names identified by dictionary mapping and predefined special symbols replace the product names.

This method has three advantages. First, it is effective for the complex slot values, including words from other slots. Many existing plan names include the words signifying *INSURANCE_TYPE* or *COMPANY*. We could handle this issue by replacing the plan names with symbols. Second, it can contribute to solving the imbalanced dataset problem. This method showed similar effects to synonym substitution, one of the text data augmentation methods. As a result, the prediction performance was improved by 3-5% compared to the previous one. Third, post-processing was unnecessary, even though we use a dictionary as important external information. We got both the benefits of dictionary matching and language model at a time. In this way, we could enhance the recall value of *PLAN_NAME*. Figure 1 shows the sentence embedding architecture when applying our approach.

3.4 Answer Retrieval

There are two types of QA: Knowledge-based QA and IR-based QA (Jurafsky and Martin, 2009). The former requires a well-structured KB, whereas the latter the large quantities of texts. However, as our case does not apply to either, we take another approach similar to the FAQ retrieval. Even though the further goal of this study is a conversational dialogue system, we aim at a single-turn QA for now. Therefore, we propose a transitionary retrieval approach to select the most similar FAQ given a user question.

After manual FAQ mapping in Section 3.2, we got a set of source questions mapped to the most similar FAQs. Given a source question (utterance), the mapped FAQ can be the correct response that our QA system should return. We devised a simple but effective method to retrieve the most similar FAQ for a user utterance.

We used the detected slot values and the TF-IDF-based keywords in the proposed method to retrieve a proper FAQ. There were three different types of

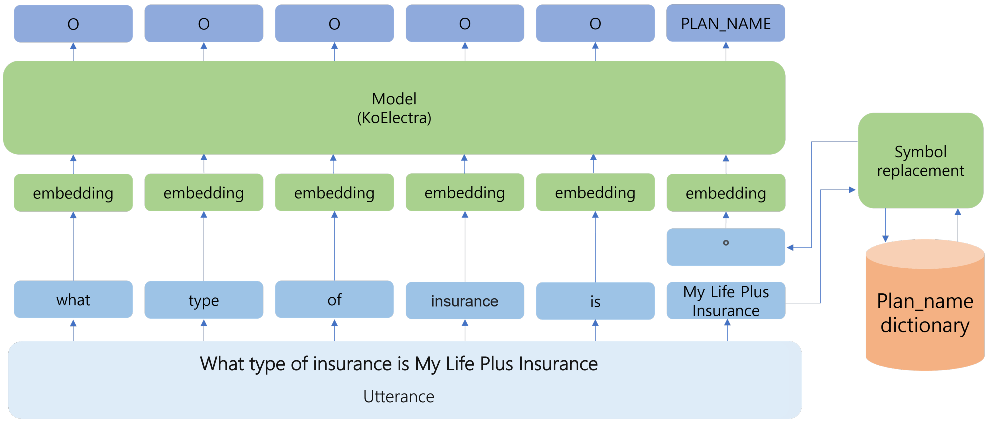

Figure 1: Sentence embeddings by symbol replacement via dictionary mapping

information to retrieve a good FAQ: 1) **set A** - the slot values and keywords detected from the user utterance; 2) **set B** - the slot values and keywords existing in the question part of each FAQ; and 3) **set C** - the slot values and keywords existing in the source questions that were mapped to each FAQ.

The overlapping score between sets A and B was computed for a given utterance and FAQ pair. A similar score was also computed for A and C. The weighted sum of these scores was the similarity score of the utterance-FAQ pair. We selected the FAQ that had the highest similarity score given an utterance.

4 Experiments

First, we present the experimental results for intent detection and slot filling. The former is a typical classification problem, and the latter can be interpreted as a sequence labeling task, as introduced in Section 3. KoElectra was for the training in both tasks. Second, we described the performance of answer retrieval from the FAQs. The weighted sum of the three scores was the similarity score of the utterance-FAQ pair: We selected the FAQ that had the highest similarity score given an utterance.

4.1 Intent detection and slot filling

Table 3 presents the experimental results of intent detection. The micro-averaged f1-score for 17 different intent types was 0.71. The result can be regarded as good given insufficient training data and many classes.

We had unsatisfactory results in the *Claim process* and *Premium* because some categories were similar to them, such as *Approximate premium or claim* and *Claim availability*. If the question was precisely for the insurance premium amount, the model classified the query into the class *Premium*. If not, the result was usually the class *Approximate premium or claim*. There are also a contextual similarity between the classes *Claim availability* and *Claim process*.

Table 3: Intent detection result

Intent type	precision	recall	f1-score
Personal information	0.83	0.79	0.81
Subscription information	0.84	0.83	0.83
Emphasis	0.97	0.89	0.93
Insurance options required	0.51	0.56	0.54
Cancer diagnosis details	0.59	0.79	0.68
Approximate premium or claim	0.58	0.57	0.58
Claim availability	0.67	0.83	0.74
Claim process	0.17	0.08	0.11
Claim	0.55	0.79	0.65
Duplicate coverage	0.88	0.68	0.77
Premium	0.30	0.25	0.27
Non-payment	0.75	0.50	0.60
Considerations	0.78	0.44	0.56
Subscribe	0.64	0.69	0.67
Terminology meaning	0.64	0.64	0.64
Termination	0.86	0.92	0.89
Greeting	1.00	0.92	0.96
macro average	0.68	0.66	0.66
micro average	0.71	0.71	0.71

For further verification, we show T-SNE visualization of 12th layer of the trained KoELECTRA in Figure 2. In the area marked with “1”, there is a mix of the instances from two different classes, *Premium* (light green) and *Approximate premium or claim* (yellow). There is also another area, marked with “2”, where the instances from *Claim availability* (dark green) *Claim process* (cyan). This result signifies that we further need to modify the category definition to separate well these confusable ones.

Table 4 lists the slot filling results. The micro-averaged f1-score is 0.95, which is a high value

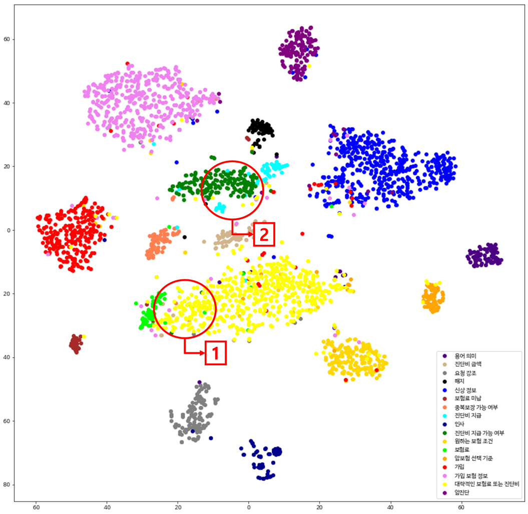

Figure 2: The T-SNE visualization of 12th layer of the trained KoELECTRA.

considering the number of slots. The worst f1-score, 0.58, is observed in the class *PLAN_NAME*, as we can easily guess from the discussion in Section 3.3. However, the result is enhanced compared to the other models trained without dictionary mapping, introduced in Section 3.3. Considering the class imbalance and insufficient training data, we found that the result was satisfactory.

Table 4: Slot filling result

Entity type	precision	recall	f1-score
GENDER	1.00	0.96	0.98
INSURANT	0.97	0.97	0.97
COMPANY	0.81	0.80	0.81
PAYMENT	0.93	0.99	0.96
CANCER_TYPE	0.93	0.83	0.88
DISEASE_LOG	0.89	0.93	0.91
JARGON	0.90	0.93	0.91
PLAN_NAME	0.46	0.76	0.58
BODY_PART	0.84	0.89	0.86
OPERATION_LOG	0.95	1.00	0.97
INSURANCE_TYPE	0.97	0.96	0.97
macro average	0.88	0.91	0.89
micro average	0.95	0.95	0.95

4.2 Answer Retrieval

We evaluated the FAQ retrieval results using the gold standard described in section 3.2. The accuracy was 70% on the test data. We also obtained acceptable results when generating random questions.

Table 5 presents two examples of the FAQ retrieval. Patterns exist in the mappings between the query and the FAQ. The two questions in the table show the representative ways. For the first query, the word “다시” (again) is the primary keyword enabling the mapping, which came from the set A introduced in Section 3.4. The second FAQ corresponds to a vast range of queries. Therefore, the FAQ is mapped to the appropriate queries especially conditioned on the slots and keywords of mapped source questions in the training data (set C in Section 3.4). Thus, the reason for mapping varies such that our strategy using the weighted sum score is proven effective for the answer retrieval.

Table 5: FAQ retrieval examples

	example
Query	이전에 위암으로 보장을 받았는데, 다시 암에 걸리면 보장 받을 수 있나요? (I had previously guaranteed stomach cancer. Can I get it if it recurs?)
1. Retrieved FAQ	보험금은 한번만 보장되나요? (Is this insurance guaranteed only once?)
2. Retrieved FAQ	암 진단 확정 시 보험금 청구서류 및 절차가 어떻게 되는 지 궁금합니다. (What are the procedures and documents required to claim insurance when diagnosed with cancer?)
3. Retrieved FAQ	지급기준이 어떻게 되나요? (What are the claim requirements for customer insurance?)
Query	오늘 보험에 가입했는데 언제부터 보장을 받을 수 있나요? (I bought insurance today, when I could get a guarantee?)
1. Retrieved FAQ	지급기준이 어떻게 되나요? (What are the claim requirements for customer insurance?)
2. Retrieved FAQ	가입하면 바로 보장을 받을 수 있나요? (Can I get a guarantee right away if I subscribe?)
3. Retrieved FAQ	보험금 청구를 하면 언제쯤 보험금이 지급되나요? (How long does it take to claim insurance?)

5 Conclusions

In this study, we built a single-turn dialogue system corresponding to the front part of our target system for the insurance domain. Our final goal is to construct a multi-turn dialogue system that can return informative counselors about insurance. For future scalability, the concept of intention detection and slot filling was borrowed, therefore, for this purpose, training data and KB was constructed on their own. We obtained an encouraging result for both tasks despite the limited quantity of the source. To enhance the performance of the slots with low occurrence and high-value diversity, we proposed a slot replacement method through dictionary mapping. The method also provided a good result. Future work first includes modifying category definitions and improving answer retrieval performance. Furthermore, we will re-design the system for the multi-turn dialogues. We expect the extracted intents and slot values to be effectively used for the multi-turn system.

Acknowledgement

This work was supported by two projects, AI-based ScienceOn User Behavior Prediction Technology, funded by KISTI (202200000001712) and Development of Next Generation Artificial Intelligence Assistant System Technology, funded by AIITONE (202200000000124).

References

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bin Fu, Yunqi Qiu, Chengguang Tang, Yang Li, Haiyang Yu, and Jian Sun. 2020. A survey on complex question answering over knowledge base: Recent advances and challenges. *CoRR*, abs/2007.13069.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–7.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. *CoRR*, abs/1905.05529.

Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. 2020. Unsupervised FAQ retrieval with question generation and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812, Online. Association for Computational Linguistics.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 539–548, New York, NY, USA. Association for Computing Machinery.

Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019a. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 25–33, New York, NY, USA. Association for Computing Machinery.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019b. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1133–1136, New York, NY, USA. Association for Computing Machinery.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*, SIGIR'19, page 1113–1116, New York, NY, USA. Association for Computing Machinery.

Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational question answering: A survey. *CoRR*, abs/2106.00874.