An Adaptive Framework for Conversational Question Answering

Lixin Su,[1,2] Jiafeng Guo,[1,2] Yixing Fan,[2] Yanyan Lan,[1,2] Ruqing Zhang,[1,2] Xueqi Cheng[1,2]
[1]University of Chinese Academy of Sciences
[2]CAS Key Lab of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences
sulixinict@gmail.com,{guojiafeng, fanyixing, lanyanyan, zhangruqing, cxq}@ict.ac.cn

Abstract

In Conversational Question Answering (CoQA), humans propose a series of questions to satisfy their information needs. Based on our preliminary analysis, there are two major types of questions, namely verification questions and knowledge-seeking questions. The first one is to verify some existing facts, while the latter is to obtain new knowledge about some specific object. These two types of questions differ significantly in their answering ways. However, existing methods usually treat them uniformly, which may easily be biased by the dominant type of questions and obtain inferior overall performance. In this work, we propose an adaptive framework to handle these two types of questions in different ways based on their own characteristics. We conduct experiments on the recently released CoQA benchmark dataset, and the results demonstrate that our method outperforms the state-of-the-art baseline methods.

Introduction

Conversational Question Answering is an effective way for humans to gather information. In CoQA, humans accomplish their information need through a series of questions over a given passage. Without loss of generality, these questions can be divided into two categories , namely *verification question* and *knowledge-seeking question*. The verification question is to confirm some facts contained in the question based on the passage. Questions in this type often begin with words such as *did* or *is*, and its answer is *yes/no*. The knowledge-seeking question is to obtain some new knowledge which remains unknown to the questioner. For verification question, the answer is a boolean value indicating the true or false of the question. For knowledge-seeking questions, the answer is often a meaningful text span extracted from the original passage. Due to these differences, an ideal model should be able to handle the two different types of questions accordingly.

There is few work considered this problem in previous study on CoQA (Reddy, Chen, and Manning 2018). In CoQA, the authors applied three different methods for the task. The first one is to directly apply seq2seq model to all questions. It performed poorly as verification questions would dominate the model learning. The second one is to utilize reading comprehension model to locate answers, which cannot answer the verification questions. The last one is to combine both of the previous two models. However, it is still inefficient in modeling the characteristic of different type questions.

In this paper, we propose an adaptive framework to overcome the difficulty of the multiple types of questions. The key is to extract rationale for the question from the passage, then apply different answering components for each type of questions correspondingly. The framework consists of four stacked components. Firstly, a rationale extraction component is used to extract related evidences for all questions. Then, a query gating component is applied to distinguish each question and distribute it to the corresponding answering component. Finally, there are two answering components for verification question and knowledge-seeking question respectively, namely, MatchNet and DistillNet. Experimental results on CoQA dataset demonstrate the effectiveness of our framework.

Adaptive Framework

Given a passage p, the previous conversation history $\{q_1, a_1, \ldots, q_{i-1}, a_{i-1}\}$ and current question q_i, the CoQA task is to predict the answer a_i. In this section, we describe the components of our framework in detail.

Rationale Extraction As the passage is usually long and redundant for the current question, we first apply a reading comprehension model to extract related text span in passage, namely rationale. We combine the conversation history and the current question as $q_{combine} = q_{i-1} \langle q \rangle a_{i-1} \langle a \rangle q_i$, and feed it to the DrQA (Chen et al. 2017) model. Hence, the rationale r_i is extracted as

$$r_i = DrQA(p, q_{combine}) \tag{1}$$

We train the model with the human annotated rationale in all questions.

Query Gating The query gating component distinguishes the type of the question and dispatches it to its suitable processing component. We apply a GBDT model acting as the query gating component, which classifies a question as a verification or knowledge-seeking question based on the content of each query.

$$g_i = GBDT(q_i) \tag{2}$$

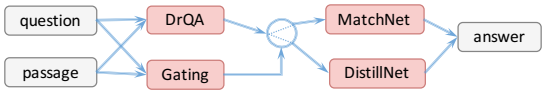

Figure 1: Adaptive framework for CoQA.

where $g_i \in \{0, 1\}$ denotes the type of questions. Specifically, 0 denotes the verification questions and 1 denotes the knowledge-seeking questions. The feature used for the model is the word feature. We obtain the golden label according to the ground-truth, e.g., the question with yes/no as the answer is labeled as 0. As shown in Figure 1, the gating component is a switcher to control operation on the questions.

MatchNet As the answer for the verification questions is yes or no, we treat them as a text matching problem. In this component, MatchNet (Min et al. 2018) is used to match the rationale extracted by the DrQA model and the question to verify that the fact described in the question is true or false,

$$p_v = MatchNet(q_{combine}, r_i). \quad (3)$$

DistillNet For the knowledge-seeking question, the corresponding answer is usually the entity or phrase from the passage. For the fluency and naturalness of the answer, we apply a seq2seq model (See, Liu, and Manning 2017) as the DistillNet to refine the final answer based on the rationale r_i and question $q_{compose}$,

$$p_a = DistillNet(r_i, q_{combine}) \quad (4)$$

Experiments and Conclusion

In this section, we describe the experimental settings, results and conclusion.

Experimental Settings We conduct experiments on the recently released CoQA dataset (Reddy, Chen, and Manning 2018). This dataset contains 116,630 question-answer pairs over 7,699 passages. Among all the questions, 20,375 questions belong to the type of verification questions. We adopt the official F measure to evaluating the performance via official evaluation script.

Experimental Results We compare our method with several baselines: 1)the PGNet (See, Liu, and Manning 2017) which takes the passage and the conversation history as input and generate the final answer; 2) the DrQA (Chen et al. 2017) model which takes the passage and the conversation history as input and locates the final answer in the passage; 3) DrQA+PGNet which combines the DrQA and PGNet to extract the rationale and generate answer.

A summary of the results are shown in Table 1. Apart from the overall results, we have also divide datasets into two parts according to the question type(i.e., verification question, denoted as $A \in yes/no$. knowledge-seeking question, $A \notin yes/no$). In this way, we can gain deep understanding of each models. From the Table, we can see that, PGNet is more effective than DrQA in verification questions. While DrQA performs better than PGNet in knowledge-seeking questions. This is not surprising since DrQA is to identify text spans from the original passage, which is more suitable to the knowledge-seeking questions. The DrQA+PGNet obtained the best among the three baseline models, since it combined the advantage of the two models. Our proposed model, which automatically distributes all questions to the right component, obtains comparable results in verification questions compared with DrQA+PGNet. As for the knowledge-seeking questions, our model achieves the best performance compared with all the baseline models. Finally, from the overall results, we can see that our model achieves the state-of-the-art results. Our framework is more flexible as each component can be replaced with other models.

	$A \in yes/no$	$A \notin yes/no$	overall
PGNet	65.1	42.54	45.4
DrQA	13.15	67.26	54.7
DrQA + PGNet	**65.7**	68.46	66.2
Ours	65.6	**69.3**	**66.9**

Table 1: Performance comparison, where *overall* denotes all data, $A \in yes/no$ denotes the verification questions, $A \notin yes/no$ denotes the knowledge-seeking questions.

Conclusion In this work, we analyze the questions under the conversational question answering tasks, and find that these questions can be categorized into two types. We proposed an adaptive framework to tackle the difficulty derived from the variety of question types. The proposed framework adapt each question to the right answering component according to its type. The experimental results demonstrate that our method can outperform the state-of-the-art baselines on the CoQA benchmark dataset.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 61425016, 61472401, 61722211, 61872338, 61773362 and 20180290, the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2016102, and the National Key R&D Program of China under Grants No. 2016QY02D0405.

References

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Min, S.; Zhong, V.; Socher, R.; and Xiong, C. 2018. Efficient and robust question answering from minimal context over documents. *arXiv preprint arXiv:1805.08092*.

Reddy, S.; Chen, D.; and Manning, C. D. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.