Optimized Methods for Extracting Information from Conversation History for Question Answering

Yuanzhi Zhu
yuazhu@ethz.ch
ETH Zurich

Bartosz Dzionek
bdzionek@ethz.ch
ETH Zurich

Hongyi Liu
hongliu@ethz.ch
ETH Zurich

Abstract

Conversational question answering (ConvQA) has been one of the most important question answering tasks. In most of the state-of-the-art models, a passage and the past Q&A session are sent as an input to a neural network like BERT. However, only few models are concerned how to combine and use the input as additional information to the current question. In this task, we would try to propose better methods for extracting information from the conversation history as input to a BERT model to generate answers. Our project would first build our framework using the existing models. Then, we will extend the model by exploring potential methods or we will combine existing methods in a smart way. It is not necessary for us to rewrite the whole model. We plan to focus more on distilling information from conversation history to increment the probability of correct answer.

1 Introduction

1.1 Background

Question Answering (QA) has been a challenging task in natural language understanding. QA is a task of automatic answer extraction for questions asked in a natural language [5]. Thus the key components in QA require the capability of understanding the question and the passage in which the question is generated.

ConvQA techniques form the building blocks of QA dialog systems and the idea behind it is to let the machine generate an answer to a question based on the provided passage and historical conversation [19].

In ConvQA task, the system is asked a sequence of questions in a conversational manner. As shown in Figure 1, every question after the first one is built on what was asked before [13]. The dependence is often of the form of coreference or ellipsis [2]. ConvQA task can be based on a retrieval using a large collection of documents of diversified topics, which make it an open domain conversational question answering task [3]. But the task was also explored in the simplified setting where the system is directly given a reference text [12].

One potential solution to the ConvQA task is question rewriting (QR). QR is a technique that allows to convert a conversational QA task into non-conversational QA. This is done by generating questions that are semantically equivalent to the original one and contain the whole contextual information [2]. For instance, the question "How old would **she** be?" in Figure 1 can be rewritten as "How old would **Jessica** be?" based on the conversation history. ConvQA task is also important as it is the base for many modern chatbots [6] [11].

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1. A sample of dataset for training (*CoQA*) [13]

1.2 Historical Development

The development of systems allowing a natural langauge conversation with computers can be traced back to *ELIZA* in 1966. It was a simple program based only on keywords and fixed rules [18]. In the next decades, various architectures were developed but they were still based on pattern matching. A notable example is *ALICE* from 2007, working on XML-based AIML files with patterns and templates [14]. The significant progress came around the year 2015 with the advent of large datasets and deep learning that outperformed traditional rule-based methods [19].

In 2018, the invention of BERT revolutionized many NLP tasks including ConvQA[8]. Most of the state-of-the-art models in the leaderboard of both *CoQA* and *QuAC* are variants of BERT [13] [4]. In [2], researchers adapted different models for QR including *AllenAI Coref* [10], *Transformer++*[15] and others. In [16], the researchers compared and combined several QR methods like *Transformer++*[15] and *QuReTeC* [17].

Except QR, there are also other methods to extract information from the conversational history. In [1], the topic for each round of QA is extracted as additional information to improve the overall model performance.

Vakulenko et al. (2021) combined QR methods of different types (sequence generation or term classification) and found it improved upon individual QR methods and achieved state-of-the-art retrieval performance on *CAsT 2019* [16]. However, they simply append terms from *QuReTeC* to the rewritten question produced by one of the generative models like *Transformer++* [15], which left huge improvement space for modification to better combine different ideas.

In this task, we would try to combine state-of-the-art QR approaches and evaluate their performances on ConvQA task. Then we would modify the QA model to further improve the performance.

2 Dataset

The main dataset for this project would be *CoQA*[13] and *QuAC*[4], and we use them to evaluate the whole model.

Conversational **Q**uestion **A**nswering (*CoQA*) is a dataset designed to make a model understand a text passage and answer questions on that passage. It is formed of question/answer sequences drawn from crowdworker conversations. The questions are conversational and the dataset comes with highlighted evidence text.

Question **A**nswering in **C**ontext (*QuAC*) is a large-scale dataset that consists of around 14K crowdsourced Question Answering dialogs with 98K question-answer pairs in total. Data instances consist of an interactive dialog between two crowd workers: a student and a teacher.

For question rewriting, we may also test our method on datasets like *CANARD*[9] and *CAsT 2020*[7]. *CANARD* is derived from *QuAC* for extractive conversational QA. In *CAsT 2020*, the current turn question depends on both the question and the answer passage to the previous turn question.

3 Goal

3.1 Minimum Goal

The minimum goal of our project would be first building a framework using the existing models which can perform conversational question answering task on a dataset like *CoQA*. The key challenge of this part is to find suitable models/methods that are accessible to us, since many of the existing models are not publicly available.

Then, we would like to assess the combination of different state-of-the-art QR approaches on the impact of ConvQA task performance. Considering the huge amount of combination of QR models, we would like to first evaluate each combined model using the ROUGE1-R for question rewriting. We plan to achieve better performance during this stage.

3.2 Ideal Goal/Target Goal

After trying out several possibilities, we could pick up the models with the highest scores. Simple way of concatenating output from existing models is not satisfactory for our research, and we would modify the picked method to further improve the performance. For example, we can throw away all the historical questions because using only historical answers is enough to make the current question understandable to the model [12].

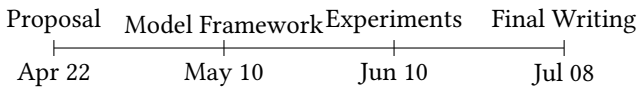
4 Evaluation & Baselines

The word-level F1 and the human equivalence score(HEQ) are two metrics provided by the *QuAC* challenge. In this project, we use the word-level F1 score as our main evaluation metric.

Our baseline in this project is to beat fine-tuned BERT base model and our ultimate goal would be to beat the models from some recent literature.

Our ideal goal will be evaluated by comparing performance to the simple concatenation approach.

5 Timeline



References

[1] Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2021. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. https://doi.org/10.48550/ARXIV.2110.00768

[2] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2021).

[3] Danqi Chen and Wen-tau Yih. 2020. Open-Domain Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, Online, 34–37. https://doi.org/10.18653/v1/2020.acl-tutorials.8

[4] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. arXiv:1808.07036 [cs.CL]

[5] Philipp Cimiano, Christina Unger, and John McCrae. 2014. *Ontology-Based Interpretation of Natural Language*. https://doi.org/10.2200/S00561ED1V01Y201401HLT024

[6] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A Customer Service Chatbot for E-commerce Websites. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 97–102. https://aclanthology.org/P17-4017

[7] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. In *TREC*.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[9] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Empirical Methods in Natural Language Processing* (Hong Kong, China). http://umiacs.umd.edu/~jbg//docs/2019_emnlp_sequentialqa.pdf

[10] Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 687–692. https://doi.org/10.18653/v1/N18-2108

[11] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 498–503. https://doi.org/10.18653/v1/P17-2079

[12] Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. 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*. 1133–1136.

[13] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.

[14] Bayan Shawar and Eric Atwell. 2007. Chatbots: Are they Really Useful? *LDV Forum* 22 (01 2007), 29–49.

[15] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question Rewriting for Conversational Question Answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 355–363. https://doi.org/10.1145/3437963.3441748

[16] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A Comparison of Question Rewriting Methods for Conversational Passage Retrieval. https://arxiv.org/pdf/2101.07382.pdf

[17] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. *Query Resolution for Conversational Search with Limited Supervision*. Association for Computing Machinery, New York, NY, USA, 921–930.

[18] Joseph Weizenbaum. 1966. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168

[19] Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational question answering: A survey. *arXiv preprint arXiv:2106.00874* (2021).