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

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 [16].

In ConvQA task, the system is asked a sequence of questions in a conversational manner. And every question after the first one is built on what was asked before [14]. The dependence is often of the form of coreference or ellipsis [2]. ConvQA task can be based on a retriever 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 [10].

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?" 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] [9].

1.2 Goal

As part of the research project, we will build a framework that consists of two modules. One module will use T5 [12] for question rewriting, and the second will use BERT [7] for question answering. The framework is shown in Figure 1 and described in more detail in the next sections.

1.3 Datasets

For question rewriting, we use CANARD [8] and CoQAR [11], which are adapted from QuAC [4] and CoQA [14], respectively.

QuAC dialogs often switch topics while CoQA dialogs include more queries for details. QuAC focuses on information that could plausibly be in context material, and CoQA does not significantly cover unanswerable questions. Our analysis strongly implies that beyond yes/no questions, abstractive behavior is not a significant component in either QuAC or

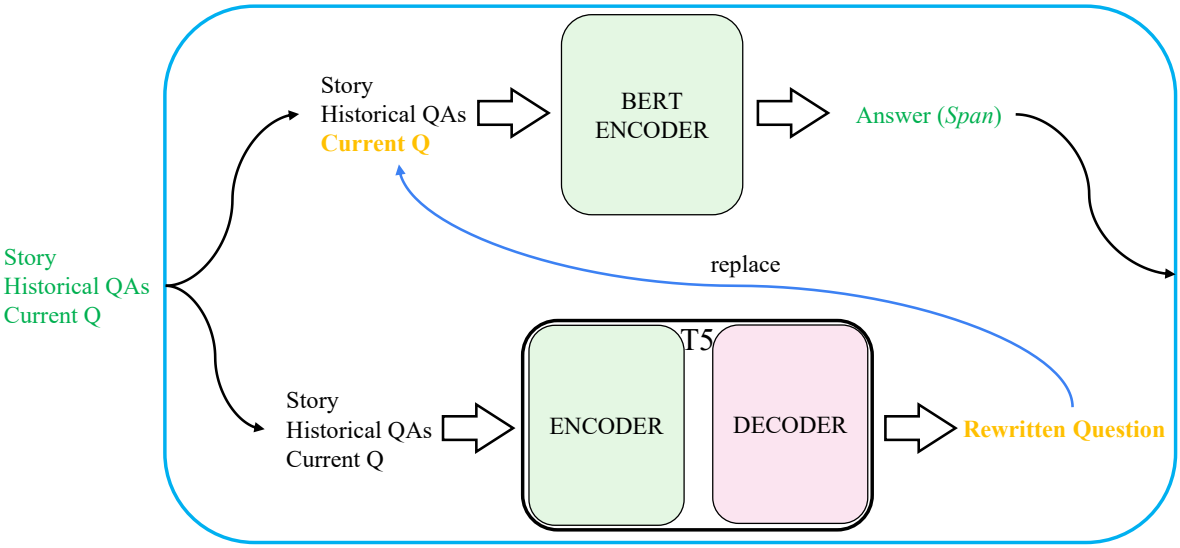

Figure 1. Our Framework

CoQA. As such, QuAC models can be trivially adapted to CoQA [15]. We will have deeper analysis on the predictions of these datasets in the following sections.

Evaluation for CoQA is the same as SQuAD [13], exact match and F1 score. We will analyze the predictions based on these metrics.

2 Our Framework

Our framework is a combination of BERT [7] and T5 [12]. As you can see in Figure 1, the common input to both models is the story and conversations (including the historical question answers and current question). These inputs are first sent to a T5 small model to generate the rewritten question. Then, we can replace the original question or even the original conversation with the rewritten question to form updated inputs for BERT. The final output from BERT is the answer, and we expect an improvement of more than 5% according to the analysis in section 4.

3 Fine-Tuning Models

We first fine-tuned BERT base on CoQA dataset and T5 small on a mixed dataset of CANARD and CoQAR. The dev set of both CoQAR and CoQA is the same, which makes it easier to build our framework.

The BERT version with the best performance in our case is BERT base trained on CoQA training dataset for 4 epochs. For T5 question rewriting model, we did several experiments because the performance of question rewriting is the key in our framework. The result is shown in Table 1

It is interesting that even without the story, the model can generate a good rewritten question. This is because one can do coreference using only the historical conversation. It is also worth mentioning that the last two experiments show that with the additional training samples from CoQAR, the BLEU score on CANARD is not improved. Indeed, CANARD is adapted from QuAC and CoQAR is adapted from CoQA, and these two datasets have different focus.

According to this result, we select the last model as the component in our framework.

model	story	batch size	hist size	dataset	BLEU CoQA	BLEU CANARD
$T5_{small}$	w/o	16	20	canard	0.3068	0.4665
$T5_{small}$	w/	16	20	canard	0.3304	0.4866
$T5_{base}$	w/	4	20	canard	0.3563	**0.5065**
$T5_{base}$	w/o	4	20	canard	0.3276	0.4953
$T5_{small}$	w/	4	3	canard	0.3261	0.4855
$T5_{small}$	w/	16	3	canard	0.3281	0.4881
$T5_{small}$	w/	16	3	mixed	**0.4039**	0.4756

Table 1. T5 experiments

4 Model Prediction Analysis

4.1 Distribution of Answers

After fine-tuning BERT, we generated its predictions on the CoQA dev set. For evaluating answers, we used the official script provided by the authors of the dataset. The answers are evaluated by comparing them to human answers. The evaluation metric is exact match score and F1 score of word overlap.

We split the answers of BERT into three categories according to the evaluation metrics:

- *correct* answers are those with exact match with one of human answers,
- *partially* correct answers do not have exact match but have non-zero F1 score,
- *incorrect* answers have zero F1 score.

The dataset is conversational with the maximum turn of 25. We analyzed the distribution of turn for questions with correctness of our prediction. And the results suggest we can consider only the data with turns less equal than 20, since the depth of conversations in the dataset decrease after turn 10 and is rare at depth larger than 20. Then we normalized the types of answers of the first 20 turns. The result is Figure 2. We see that in the first turn the percent of correct answers is 68.8%, and for turns 2-20 it is on average 58.3% (95% confidence interval = 57.5% − 59.1% [1]). In the first turn, incorrect answers comprise 10.4% of answers. In turns 2-20, they comprise on average 17.3% (95% confidence interval = 16.4% − 18.3% [2]).

Such a difference between the first turn and the rest means BERT underperforms in the multi-turn setting. This allows us to experiment with question rewriting from T5 to decrease the gap between single-turn and multi-turn QA.

Even [1] suggest that with all history, the open retrieval QA model performs better than with only the rewritten question. We argue here that with the correct rewritten question, we can turn multi-turn setting into single-turn and improve the performance.

4.2 Failure Case Analysis

As a next step, we looked closer into the answers to find patters connected to incorrect answers. Since we want to get an improvement in the conversational setting, we will only consider examples from turn 2 onward.

We found the ratio of correct/partial/incorrect answers for answers composed of 1 up to 10 words (longer answers comprised very small part) and presented it in Table 2. For answers that are single words it is much easier to get an exact match with the ground truth. As the answers start to be longer, it is less common to get an exact match but more common to have a word overlapping with ground truth

[1] Computed with t-test, but we also obtained almost identical results with bootstrapping.
[2] See footnote 1.

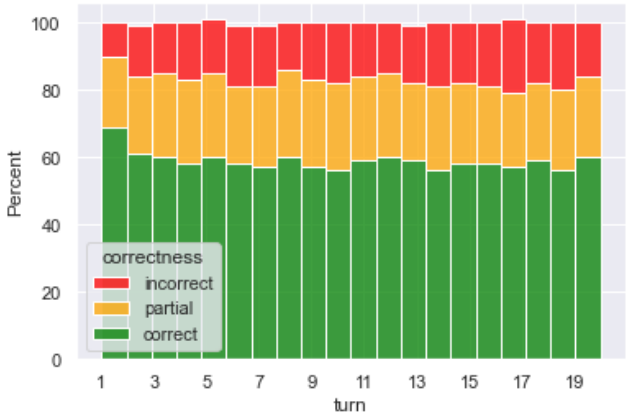

Figure 2. Normalized correctness for turns ≤ 20.

answers, thus being partially correct. It is worth noting that ratio of incorrect answers is much bigger for single-word answers (0.22). For comparison, there are 205 single-word answers to a question in the first turn and their incorrect ratio is 0.13. Given that single words have the highest count among the generated answers, they constitute a significant bottleneck hurting performance of the model.

#words	count	correct	partial	incorrect
1	**3518**	**0.70**	0.08	**0.22**
2	1360	0.66	0.21	0.13
3	880	0.56	0.29	0.14
4	511	0.45	0.46	0.09
5	326	0.37	0.51	0.13
6	227	0.34	0.56	0.10
7	138	0.26	0.61	0.13
8	114	0.21	0.66	0.13
9	91	0.26	0.59	0.14
10	46	0.15	**0.74**	0.11

Table 2. Distribution vs length of the answer. (turn ≥ 2)

Finally, we checked the correctness with respect to the domain of a question. The public part of CoQA consists of five different domains, every domain appears in approximately similar number of questions. Again, we compared the number of incorrect answers in the first turn against the next turns. The result is shown as Table 3. We observed that 4 domains have very similar gap between the first turn and other turns. However, for the *gutenberg* part, we got approximately the same ratio. We plan to examine questions in this domain more closely as our goal is to minimize the gap for other four domains.

4.3 Limitations of CoQA

During exploration of the answers, we noted that CoQA under-evaluates some of our answers. That happens when

turns	domain				
	cnn	gutenberg	mctest	race	wikipedia
turn = 1	0.08	0.19	0.10	0.10	0.05
turn ≥ 2	0.15	0.19	0.18	0.20	0.14
abs diff	**0.07**	**0.00**	**0.08**	**0.10**	**0.09**

Table 3. Ratio of incorrect answers.

the generated answer is more detailed than ground truth or vice versa. For example:

- model: *"in a barn near a farm house"*, CoQA: *"in a barn"* (EM: 0.0, F1: 0.71)
- model: *"Joe Fontana"*, CoQA: *"Detective Joe Fontana"* (EM: 0.0, F1: 0.8)

Fortunately, this effect is mitigated by a high F1 score. However, final evaluation might be noisy due to the preference of human contributors to shorter or more detailed answers.

The worse problem is that F1 score does not capture sematics of the answer. This can lead perfect answers to be classified as incorrect (e.g. model: *"fifty"*, CoQA: *"50"*, (EM: 0.0, F1: 0.0)), and adds more noise to the final evaluation.

5 Future Work

Currently, we keep working on connecting the two modules of the framework. The inputs for BERT and T5 are processed differently, and we need to figure out how to unify them.

Once we have it working, we will try appending the rewritten question as well as replacing the original question with the rewritten one.

We plan to examine our conclusions from section 4 more closely. We will find out to which aspects the model does not pay enough attention. Improving the weakest parts of the model can potentially lead to big performance gain.

We summarize the future work as following:

- Build a working framework
- Probe potential vulnerabilities of the model

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] 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

[9] 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

[10] 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.

[11] Gwenole Lecorve Quentin Brabant and Lina Rojas-Barahona. 2022. CoQAR Question Rewriting on CoQA. To be published in LREC2022. https://github.com/Orange-OpenSource/COQAR

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[13] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. https://doi.org/10.48550/ARXIV.1806.03822

[14] 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.

[15] Mark Yatskar. 2018. A qualitative comparison of coqa, squad 2.0 and quac. *arXiv preprint arXiv:1809.10735* (2018).

[16] 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).