Combined Transformers for Conversational Question Answering

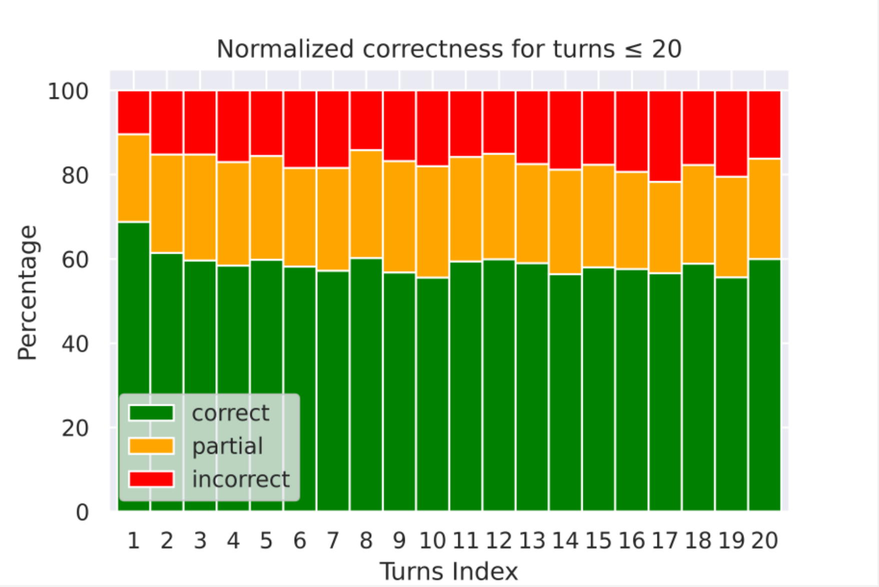
Yuanzhi Zhu, Bartosz Dzionek
ETH Zurich

1 Introduction

It’s now more and more important to enable machines answer sequential questions in a given context and a conversational manner[1]. In this project, we investigate the possibility to combine a generative model (eg. T5) with a baseline Bert model[2] for better conversational question answering.

2 Baseline Analysis

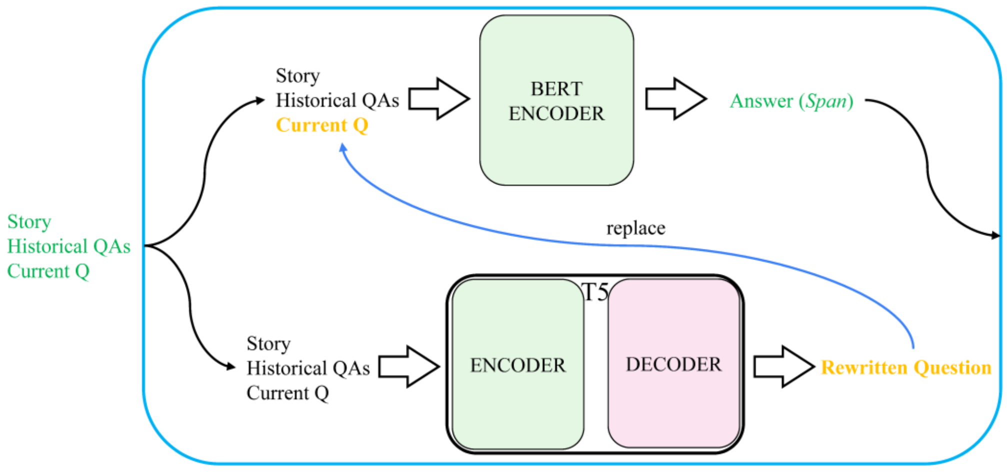
- Bert tends to perform better for the first turn.
- Even with increasing turns, the performance of Bert baseline model is stable.



For most of the examples given in the development dataset of CoQA, the total number of turns of conversation is smaller than 20.

3 Our Method

- Provide additional information from a pre-trained T5, the rewritten questions, to Bert model for more accurate answer



4 Results and Discussion

We compared the final results of different methods that form final inputs to the Bert model in the table below.

As a result, we discovered several methods that have comparable performance.

Inputs	EM	F1
Baseline (2 epochs)	67.2	77.1
Replaced with Rewritten Q (4 epochs)	64.8	74.8
Replaced all (2 epochs)	62.4	71.8
Concatenated w/o history (2 epochs)	63.8	73.5
Concatenated with Rewritten Q (4 epochs)	66.5	76.7
Concatenated with Rewritten Q (2 epochs)	**67.3**	**77.2**
With T5 embedding (4 epochs)	**67.4**	**77.2**
Pooled T5 embedding (3 epochs)	66.9	76.8

EM means the generated answer exactly match the gold answer.
For the last one, we send the token embedding from T5 encoder, together with the raw output from Bert to MLP to predict the answer span.
For more results, please refer to our final report in our Github repository[3] under folder *reports*

5 Conclusion

- We get a comparable results but which is not the desired outcome. It’s challenging to make an improvement without modifying the Bert model itself.
- Still many experiments can be done to fill the gap between first and the rest turns’ performance.
- The power of T5 for question rewriting is limited and the output from T5 is not always helpful.
- In the future we should pay more attention to the alignment of these two models, such that they can understand each other better (T5 sentence embedding should be better) .

References

1. M. Zaib, W. E. Zhang, Q. Z. Sheng, A. Mahmood, and Y. Zhang, *“Conversational question answering: A survey,”* arXiv preprint arXiv:2106.00874, 2021.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *“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), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019
3. yuanzhi-zhu/CSNLP-Project-ETH: This is the repo for our computational semantics natural language processing course project (github.com)