Assignment 2

Davide Femia, Riccardo Paolini, Sfarzo El Husseini and **Alessandro D'Amico**
Master's Degree in Artificial Intelligence, University of Bologna
{ davide.femia, riccardo.paolini5, sfarzo.elhusseini, alessandro.damico5 }@studio.unibo.it

Abstract

In this paper we address the problem of Conversational Question Answering, with reference to the CoQA dataset introduced by Stanford University in (Reddy et al., 2018). One of the major challenges of this tasks is the management of the conversation history to answer the current question. We propose a solution based on Transformers such as BERTTiny and DistilRoBERTa. In particular, our architectures consist of two networks, the first deals with the rationale (the sentence which contains the answer) extraction, while the latter identifies the correct answer and reformulates it. We tested each of the models both with history and without history and achieved the following results:

1 Introduction

Previous works are based on two main approaches:

- Encoder-only architectures that aim at extracting the answer directly from the passage.
- Encoder-Decoder architectures in which the decoder attempts to generate a response dependent on the input provided by the encoder.

(Qu et al., 2019) showed how effective an encoder-only architecture that responds by imitating the text can be. On the other hand, encoder-decoder models have been proved to be really effective for causal language modelling thanks to their ability to generate free-form text conditioned on the context.

Our approach aims at combining the advantages of both architectures. In particular, we used the two models by stacking them together, so that the encoder-only model filters the information that will be passed to the encoder-decoder.

We conducted our experiments on CoQA, a conversational question answering dataset, to show the effectiveness of our method. Our BERTTiny-based methods with and without history achieved respectively an F1 score of XXX.X and XXX.X, while DistilRoBERTa-based methods with and without history achieved respectively an F1 score of 50.3 and XXX.X.

Figure 1: Data Flow

2 System description

Our implementation is based on two networks that have been combined, as show in Figure 1. The first one is an encoder-only model (*Span Extractor*) which precisely extracts the salient sentences (*rationale*) of the passage, depending on the given question and the history of the conversation (Devlin et al., 2018). The second one is an encoder-decoder model (*Answer Generator*) that identifies the answer lying in the rationale and refines it. The generated answer will be ~~free-form text~~ as required by the CoQA paper (Reddy et al., 2018), this also makes the answers seem more humane (Egonmwan and Chali, 2019).

In particular, we created new model classes for our BERTTiny-based / DistillRoBERTa-based models. Each of these classes contains its own *forward* and *generate* methods. These functions in turn recall the *forward / generate* method of the underlying models. In addition, they include the necessary code for generating the input of the answer generator. In this part we combine the current question (and the history) with the rationale predicted by the span extractor.

3 Experimental setup and results

Accordingly with Transformers' requirements we pass two tokenized sequences to the model, the question (plus history) and the passage. We decided

Exp	Seed	History	F1-Score Val | Test
1	42	YES	0.2 | 0.2
2	42	NO	0.2 | 0.2
3	2022	YES	0.2 | 0.2
4	2022	NO	0.2 | 0.2
5	1337	YES	0.2 | 0.2
6	1337	NO	0.2 | 0.2

Table 1: Results show the performance of **BERT-Tiny** based model on the **validation** and **test** dataset.

Exp	Seed	History	F1-Score Val | Test
1	42	YES	0.5 | 0.5
2	42	NO	0.5 | 0.5
3	2022	YES	0.5 | 0.5
4	2022	NO	0.5 | 0.5
5	1337	YES	0.5 | 0.5
6	1337	NO	0.5 | 0.5

Table 2: Results show the performance of the **DistilRoBERTa** based model on the **validation** and **test** dataset.

not to use sliding windows or other techniques to split the passage since we noticed that about 95% of the dialogues can be tokenized using 512 tokens, even when we prepend 4 QA pairs as history to the current question. For the few dialogues requiring more than 512 tokens we simply apply truncation.

As regards history management, we prepend the history of the conversation to the current question instead of creating a special embedding. In particular, we considered *n* past question-answer pairs, such that R_0 and A_0 depend on $(Q_{-n}, A_{-n}, ..., Q_{-1}, A_{-1}, Q_0)$.

The model based on *DistilRoBERTa* has been trained for 3 epochs, with batches of 4 samples. The optimizer is AdamW with learning rate 5e-5 and weight decay 0.01 (default betas and epsilon).

Instead, the model based on *BERTTiny* has been trained for 3 epochs, with batches of 32 samples. The optimizer is AdamW with learning rate 5e-5 and weight decay 0.01 (default betas and epsilon).

For both models we use a teacher forcing approach to ease the training of the answer generator in the early steps, since the span extractor initially provides imprecise rationales. The teacher forcing probability decays linearly from 1.0 to 0.3 during training.

The results obtained on CoQA dataset are summarized in Tables 1 and 2.

4 Discussion

The CoQA dataset directly provides a training set and a test set. The training set was further divided into training set and validation set, taking 80% and 20% of the dialogues respectively. Results show that the seed does not affect significantly the performance of the models, meaning that they are robust to changes in class distributions and do not overfit the data.

It is possible to observe that *BERTTiny* performs better when we do not consider the history of the conversation. In contrast, *DistillRoBERTa*'s performance is similar in both cases whether history is considered or not. This is ~~probably~~ due to the fact that *DistillRoBERTa* with its 82 Million parameters and 6 layers has a greater "model capacity" with respect to *BERTTiny* that only contains 4.4 Million parameters and 2 layers. This additional capacity allows *DistillRoBERTa* to properly extract useful information from the history.

5 Conclusion

In our work, we proposed models that mainly introduce changes in the network architecture. We focused on separating the tasks of rationale extraction and answer generation, thus making prediction easier for both networks. These models were then used for question answering on the CoQA dataset.

The results showed that our BERT-Tiny model achieves performances in line with standard Seq2Seq models (Reddy et al., 2018), while DistillRoBERTa improves the overall F1-Score of about 25%, proving to be quite good even in answers that require rephrasing the text. However, its computational cost was considerably higher than BERT-Tiny, requiring three times more training time.

Many other techniques can be used to further improve results, such as sliding windows that divide the passage into smaller sequences, fully differentiable span extractor that can also benefit from training the answer generator following it, or also fancier ways of modeling the history.

6 Links to external resources

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255, Hong Kong. Association for Computational Linguistics.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with history answer embedding for conversational question answering. *CoRR*, abs/1905.05412.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.