NLP Assignment 2: Question Answering with transformers on CoQA

Domenico Dell'Olio, Giovanni Pio Delvecchio and **Raffaele Disabato**
Master's Degree in Artificial Intelligence, University of Bologna
{ domenico.dellolio, giovanni.delvecchio2, raffaele.disabato}@studio.unibo.it

Abstract

In this work, we present an approach to Generative Conversational Question Answering involving a seq2seq architecture with a BERT encoder (Tiny BERT and distil-RoBERTa). We tested the models both by inputting them only the Question + Passage pair or the pair followed by part of the conversational history. We reached about 18% SQUAD-F1 on test with our best configuration on a platform with limited resources.

1 Introduction

Generative Conversational Question Answering is a task centered on generating answers to a series of questions about a text passage, where each question can refer to previous QA pairs, as typical of human interaction. Common approaches to the problem are mainly neural, as seq2seq RNN models, Transformer Encoder-Decoder Models and hybrid models mixing KBs with Neural Architectures (Zaib et al., 2021), to introduce expert knowledge. Our approach tests a seq2seq architecture with a light BERT pre-trained frozen encoder (Tiny BERT - TB and Distil-RoBERTa - DR) and a attention LSTM decoder on CoQA dataset (Reddy et al., 2018). These architectures (Fig. 1)were tested both by inputting the question and the passage pair or by concatenating the latter with the dialogue history. Four different configuration were trained, validated and tested three times, using a different seed each time. Moreover, both the greedy and the beam search sampling were tested for sampling answers. Our best model, achieving 18% SQUAD-F1 (Rajpurkar et al., 2018) on test, takes advantage of the bigger capacity of DR encoder w.r.t. TB, the beam search sampling and the addition of previous QA pairs to the input. It shows to have captured semantic relations between question and answers, but still having some problems on some question categories.

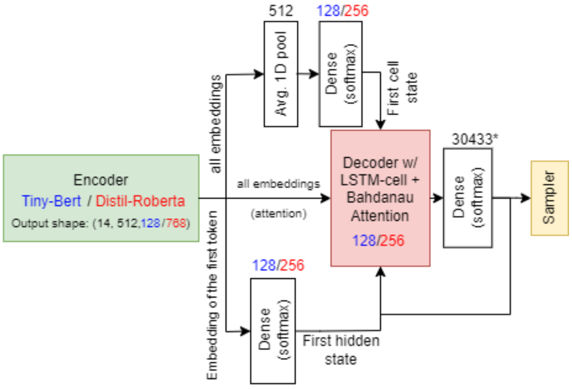

Figure 1: Tested models architecture. Blue text refer to models with Tiny BERT, Red to Distil-RoBERTa. There are no variations between model with or without history. *size of the vocabulary.

2 System description

The dataset is composed of passages with a sequence of QA pairs and additional info. Examples are extracted as ((Question, Passage[:History][1]), Answer). When the History is provided, it is added in reverse order, so to keep more relevant past QA pairs to the current one when truncating. In fact, the question and the passage are tokenized and padded with the appropriate BERT tokenizer (byte-pair based) up to a max of 512 tokens. Differently, answers are tokenized, lowercased and padded with a word-based TF tokenizer, fit on the training set, filtering out special characters except {',<,>}, up to a max of 20 token ($\geq 99^{th}$ percentile of answer lengths). This limits the computational load of the model. Four architectures were tested (two with and two without history) as shown in Figure 1. Actual answers were then generated both with a greedy sampling and beam search approach. The code was adapted from the one provided on GitLab and its original tutorial (see sec. 6). Our solution modifies and tries to optimize the architecture shown in the reference code.

[1] ":" indicates the concatenation operation

Model	SQUAD-F1-gs	EM-gs	SQUAD-F1-bs	EM-bs
		Validation		
Tiny Bert	14.49	11.92	14.58	12.11
Distil-Roberta	**17.39**	**13.95**	**17.50**	**14.17**
TB + History	14.49	11.91	14.68	12.18
DB + History	17.22	13.87	17.37	14.10
		Test		
Tiny Bert	14.51	12.15	14.55	12.28
Distil-Roberta	17.77	14.58	17.88	14.79
TB + History	14.43	12.09	14.69	12.36
DB + History	**17.83**	**14.70**	**17.91**	**14.92**

Table 1: Average results of the runs using seeds 42, 1337, 2022 for reproducibility. "gs" stands for greedy sampling, "bs" for beam search.

3 Experimental setup and results

The "train" part of the dataset was split using 80% of dialogues as training and 20% as validation, while the "validation" part was used as test set. Four configurations of model were devised (two with TB as encoder, two with DR - two with and two without history). Each model was then fine-tuned 3 times (each time with a different seed) for 3 epochs, freezing the encoder, with Adam optimizer. The batch size was fixed to 14 to allow learning under Google Colab restrictions, with the outcome of discarding a single example. Some hyper-parameters were tuned by hand observing the average SQUAD-F1 across seeds on validation:

- Decoder cells: 128 for TB models, 256 for the others, chosen from {32, 64, 128, 256, 512}.
- Learning rate: 1e-3 chosen from {1e-3, 1e-5}. It was kept low as typical in fine tuning tasks.
- Beam width for the beam search: 3 chosen from {2, 3, 4}.
- Length penalty weight for the beam search: 1.5 chosen from {0 to 1.75 with steps of 0.25}.

Other variations were done on the architectures (stratifying the decoder, unfreezing the encoder...) but resulted less successful. The validation and test sets results of the models are reported in Table 1.

4 Discussion

The models with DR encoder achieve better results w.r.t. those with TB as they have more capacity than the latter. Also the models providing history improve a bit on their counterparts, while the beam search method for generation improves the quality of the prediction w.r.t. greedy sampling, as expected. We can state that the DR + History model is the best configuration, reaching almost 18% SQUAD-F1 on test (the lower score on validation is due to less variance in the scores w.r.t. those of the no-history variation). By analyzing some answers generated with the beam search (the 5 worst and best ones for each available text source - see Appendix A), we can further observe that:

- The models seem to capture the semantic relation between a question and a kind of answer. Wrong predictions are often semantically near and belong to the same class of concept (object, person, "Yes/No"...) of the true answer. This may follow from the expressive power of the embeddings returned by the encoder.
- The models show some problems when answering "No" or to some "Counting" questions. This may follow from the intrinsic difficulty in treating the negation with a LM and from the fact that probably most of the counting questions require counting up to three (Small sign of overfitting).
- The models achieve good results on questions requiring short answers (1-2 words), due probably to the length penalty weight imposed or the intrinsic simplicity of these examples.

There's room for improvement, which could be achieved by trying some variation on the training recipe (bigger batch size, adding an appropriate POS tagging, adding outputs of event extraction methods...) which we couldn't completely try also for Google Colab resource restrictions.

5 Conclusion

In this work we proposed a seq2seq approach taking advantage of BERT encoders to solve a Generative Conversational Question-Answering problem. By testing different lightweight encoders and the effectiveness of previous conversational history as added input, we indicate the best performing model to be the one using the more capable Distil-RoBERTa encoder, sampling with the beam search and considering the history of previous QA pairs, which, as expected, slightly improves the answers of the model. The tested models correctly captures the semantic relation between question and answer but shows some problem on more complex kind of answers like the ones involving negation or long phrases. There's still some room for improvement, which could be achieved by trying different training recipes that are more computationally demanding or including additional input information.

6 Links to external resources

- Link to the Google Drive folder containing the model checkpoints and predictions.
- Link to the seq2seq model provided on GitLab.
- Link to the original TensorFlow seq2seq model with attention tutorial.

A Examples of generated answers

In this appendix we report some interesting examples of wrong answers to corroborate the observations in Section 4.

A.1 Semantically similar answers

Question: *Who is her daughter?*
Passage: [...]*Her daughter, Nicole, speaks fluent English.*[...]
True answer: *nicole*
RB + history: *her baby/ her mother/ her baby*

Question: *What looked like a birds belly?*
Passage: [...]*A bottle floated by over the heads of Asta and his friends. They looked up and saw the bottle. "What is it?" said Asta's friend Sharkie. "It looks like a bird's belly,"* [...]
True answer: *a bottle*
RB + history: *a turtle/ a fish/ a huge boy*

Question: *What worked her way northward?*
Passage: [...]*long the leeward coast of Malaita, the Ariel worked her leisurely way* [...]
True answer: *the ariel*
RB + history: *a ship/ a crafty ship/ a boat*

These three example show how the models (we take the best one as reference), when generating a wrong answer, usually generate something in the same concept class of the true answer or semantically near to it.

A.2 Problematic kind of questions

Question: *Anything recent?* (referring to the movies Farina starred in, e.d.)
Passage: [...]*Farina also had roles, generally as either cops or gangsters, in a number of movies, including "Midnight Run" (1988), "Get Shorty" (1995), "The Mod Squad" (1999) and "Snatch" (2000)* [...]
True answer: *no*
RB + history: *yes/ yes/ yes*

Question: *How many burroughs are there?*
Passage: [...]*Staten Island is one of the five boroughs of New York City in the U.S. state of New York.* [...]
True answer: *five*
RB + history: *two/ two/ three*

These are examples show how the network has some difficulty on tasks requiring reasoning, as acknowledging negation. Also some counting question where the number is more than three or two are problematic. This is probably due to overfitting as "two" or "three" are common true answers.

References

Pranav Rajpurkar, Jian Zhang, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL 2018*.

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

Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational question answering: A survey. *CoRR*, abs/2106.00874.