Use of a QA-based method for the comparison of two Text Summarization models

Caterina Borzillo
borzillo.1808187@studenti.uniroma1.it

Abstract

The aim of this project is to use a Question Answering model to compare two (or more) Text Summarization models. That's because in recent years NLP evaluation metrics such BLEU or ROUGE have been shown to be inappropriate with regard to the evaluation of machine-made summaries, both for the purely syntactic assessment and the length of the output that they have to measure. For this purpose, I implemented a two-stage method that is able, given a summary, to answer human-annotated questions; the basic idea is that if the model succeeds in answering the questions correctly, it means that the model-generated summary is good since it contains the most important information, and vice versa. A detailed explanation of the implemented method along with the final results will be discussed below.

1 Introduction

Going into the details of the implementation, I introduced a new method (Stage 1) which aims to extract the context where the answer to a specific question can be found: given a summary and a query (question), the selection of the 3 most relevant sentences with respect to that query is done by averaging the scores found through the use of the BM25 ranker function together with the use of the cosine similarity function between different word-embeddings. The reason why I decided to employ 2 different methods to extract the context for a question is that, if from one side with the BM25 ranker I'm adopting a purely syntactic approach, on the other side using a word embedding approach (in particular I used GloVe embeddings) I'm taking into account semantics and thus the words' meaning in order to compute the final score. The formula used to compute the final scores is the following:
$final\_scores = 0.5 * bm25\_scores + 0.5 * glove\_similarity\_scores$
In the formula each variable is a tensor (of length n = number of sentences in a summary) consisting of a sequence of scores in which each score tells how relevant the i-th sentence of the summary is with respect to the query. Different types of experiments could be conducted by changing the weight associated to the 2 different score computation methods; for example, we might give more importance to semantics with respect to exact term-matching. Concerning Stage 2, once I extracted contexts from summaries, I took the pre-trained encoder-decoder model T5 (in particular T5ForConditionalGeneration) and then I fine-tuned it on the dataset I used for this project (transfer learning).

2 Dataset

The dataset I used for developing this project is the Narrative QA dataset (Kočiský et al., 2017) which contains a big collection of books and movie scripts (I worked only on books) each one having 30 question-answers pairs generated by human annotators in free-form natural language. I used both the (human-annotated) summary and the full-story setting provided by the dataset: the first one for fine-tuning the Question-Answering model and the second one to apply two different text summarization models that in the end I will compare.

3 Stage-1

In stage 1 I implemented a method for extracting a context, given a summary (which can be either human-annotated or machine-generated) and a query (which can be either the question or the answer - for a supervised approach -). This "context" is represented by the set of the 3 most relevant sentences that somehow contain the answer to the query (or have to do with the answer itself if we give the answer as query). To extract this context, as I explained earlier, I averaged the scores computed for each sentence in the summary by the BM25 ranking function and the scores found

through a similarity system based on word embeddings.

4 Stage-2

In Stage 2 the fine-tuning of the T5ForConditionalGeneration model is performed. I decided to take this model because according to the documentation (T5h) and the paper (Raffel et al., 2020) I was able to specify during the training phase the 'context' and the 'question' fields (one of the T5 task prefixes, see this one refers to Appendix D.15 in Raffel et al., 2020) for which the model had already been trained.

One important thing to note is that the fine-tuning phase is done only by extracting contexts (from Stage 1) related to the human-annotated summaries in the Narrative QA dataset (and not to the machine-generated summaries). The reason is that in this way the model learns to answer questions based on contexts taken from human-written texts. So the machine-generated summaries will eventually only be given to the already fine-tuned model with the purpose of being able to compare them based on the model-generated answers. Therefore if the answers generated by the model based on the contexts taken from the BERT summary will be closer to the gold answers with respect to the answers generated by the model based on the contexts taken from the GPT2 summary, then we can observe that the BERT model for the Text Summarization task is better because it contains the most important information, and vice versa.

4.1 Some training details

After conducting an evaluation regarding the length of each answers and each context-question pair within the training dataset, I set the MAX_SOURCE_LENGTH and MAX_TARGET_LENGTH to 1500 and 200 respectively. Due to some problems related to CUDA's memory on Google Colaboratory, I put BATCH_SIZE = 16 (and not 32, as I originally planned); finally, for the same reason, the QA model has been trained for 5 epochs only.

5 Models comparison and results

As I mentioned earlier, the summarization systems used for the comparison are the BERT model and GPT2 model (sum). These State-Of-The-Art models are very well known, in particular BERT (Bidirectional transformer) is a transformer used to overcome the limitations of RNN and other neural networks as Long term dependencies and GPT2 is a transformer which has the peculiarity of including in its architecture an Attention layer that implement the well know attention mechanism. In order to compare them, I give in input to the QA model the contexts (Stage 1) - taken from the BERT/GPT2 summaries and the questions, as I previously had done with the summaries written by humans. Then to measure the output I computed the ROUGE score between the answers based on the human-written contexts and the gold answers, the answers based on BERT context and the gold answers and the answers based on GPT2 context and the gold answers. The results are:

- *Human-written summaries*:
  - RougeL_fmeasure: 46.91,
  - RougeL_precision: 50.22,
  - RougeL_recall: 49.06
- *BERT*:
  - RougeL_fmeasure: 4.70,
  - RougeL_precision: 5.73,
  - RougeL_recall: 4.84
- *GPT2*:
  - RougeL_fmeasure: 4.89,
  - RougeL_precision: 5.97,
  - RougeL_recall: 5.16

As we can see from the table, GPT2 model obtains slightly higher results with respect to BERT model, but the results actually are very close to each other.

6 Final considerations

The use of Question-Answering models for evaluating AI systems that aim to summarize texts has been in recent years well studied and explored, stated that their purpose is to try to overcome NLG metrics limitations. For example Scialom et al., 2021 proposed a new approach to evaluate summarization systems based on a QA framework called *QuestEval* which, in contrast to metrics such as ROUGE, does not require any ground-truth reference. Less recent pubblications concerning approaches who try to find QA based-solutions to this problem are Chen et al., 2018, Scialom et al., 2019 and Eyal et al., 2019.

This alternative and simpler method I'm proposing here naturally contains some critical points, such as for example the final use of the ROUGE metrics for the comparison of the answers generated by the model given different summaries. Also, the number of epochs (5) used for training are not sufficient to carry out a good transfer learning along with the small number of samples involved in the training. However, I think that the mechanism I implemented to extract context from a summary given a query, which accounts for both syntactic and semantics aspects, could be a starting point for future projects. An idea could be to use this mechanism to compute a score which represents the similarity between the gold answers and the model predicted answers in place of ROUGE score measure. For this purpose, I tried to implement this function to assess both the syntactic and semantic similarity between answers, but BM25 ranker cannot find any interesting scores in this case to quantify the similarity between gold answers and predicted answers and therefore I used only the similarity between the (GloVe) word embeddings associated to the answers. This similarity score I accumulated for each answer for both models gives the following results:

- *Similarity score for BERT model*: 128.59
- *Similarity score for GPT2 model*: 107.66

As we can notice, it seems that the answers given by the QA model based on the summaries generated by the BERT model are "closer" to the gold answers with respect to those given by the GPT2 model. Obviously one could use one among the multiple existing pre-trained word embeddings or one might also want to train word embeddings based on the language used in the books.

References

State-of-the-art text summarization models. https://medium.com/analytics-vidhya/text-summarization-using-bert-gpt2-xlnet-5ee80608e961.

T5 hugging face documentation. https://huggingface.co/docs/transformers/model_doc/t5.

Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2018. A semantic qa-based approach for text summarization evaluation.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrative qa reading comprehension challenge.

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.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models.