Team 26: Text summarization

Teemu Sormunen (teemu.t.sormunen@aalto.fi)
Abdullah Gunay
Vasileios Christoforidis (vasileios.christoforidis@aalto.fi)

March 10, 2021

Abstract

In our project, we have decided to create text summarizer with BERT. There is already a Bert Extractive Summarizer -package. It's based on following article: https://arxiv.org/ftp/arxiv/papers/1906/1906.04165.pdf This package is intended for lecture-summarization, and our goal is to extend and fine-tune this model for news or scientific articles. Even though foundational work is already done for the package, there's still much to customize such as tokenizer and model. Our target language is english. Summarization tools such as Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) are available, and part of the project is to try to evaluate the texts automatically and manually.

1 The dataset

The Cornell University NEWSROOM dataset [1] consists of 1.3 million news articles and summaries between the years 1998 and 2017. It contains articles from 38 major publications and the summarization strategies combine both extraction and abstraction. The Newsroom website (summari.es) contains various different tools for analyzing the large article database. The dataset is written by humans with purpose of summarization. Dataset was created by scraping the website for content and using summaries in the HTML metadata. These metadata summaries were created to be found from search engines and social medias. Top news websites gathered from Alexa.com were used.

We will be filtering the dataset [1] to use extractive summaries, as our model is extractive. Another option for training dataset would be the New York Times corpus, which contains over 650 000 manually summarized news articles. These summarizations were done afterwards by library scientists. [2]

2 State of art

Often, for the baseline precision Lede-3 is used. It's automatic summarization strategy that copies first n words of the first sentence and uses it as summary. In our work we will be using this as a baseline also. Current state of art models in text summarization can be classified to fully extractive, fully abstractive or mixed models. In the study [1] ROUGE-1, ROUGE-2 and ROUGE-L F_1 score variants were used to account for different summary lengths. In our work we will also be using ROUGE and BLEU scores for testing. We can test on both datasets [1] [2]

In the leaderboard it seems that mixed models perform the best, then comes the extractive and abstractive is the least efficient. TextRank was one of the best models in the study [1] and it will also be used as comparison for our model. The very least goal is to achieve better results than the Lede-3, but it will be interesting to see how well the extractive BERT performs ranks compared to TextRank.

References

[1] M. Grusky, M. Naaman, and Y. Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL `http://aclweb.org/anthology/N18-1065`.

[2] E. Sandhaus. The new york times annotated corpus. In *The New York*

Times Annotated Corpus LDC2008T19. Linguistic Data Consortium, 2008. URL https://doi.org/10.35111/77ba-9x74.

NOTE-TO-SELF:

Extractive summarization generates verbatim summarization. It takes subset of the sentences in the original text, and attempts to identify important sections. [1]

Abstractive summarization interprets the original text, and creates a new, shorter text that attempts to have the same amount of relevant information.

Extractive Fragment Coverage explains how many words in the summary are from the article. Many words from article in the summary means high coverage.

Extractive fragment density explains average length of extractive fragment. Coverage can be high, but if the consecutive words (fragments) are small, then it might present new information (be more abstractive). So high density means, that there's long fragments from original text.

compression ratio defines the proportion of article word amount and summary word amount.