Team 26: Text summarization

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

March 8, 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 yeras 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

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.

Written by humans with purpose of summarization. Design new benchmark.

Document Understanding Conference: Can be used as test set. It has multiple reference summaries for each article. High quality. Small data.

Previously simulated summaries were used as training data, e.g. headlines.

Other mentionable dataset is new york times corpus which is the largest summarization dataset currently available. All articles from The New York times. Several hunder thousand articles written between 1987-2007. Summaries written by library scientists. Biased towards extractive strategies.

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.

In the paper [1] the dataset is also analysed to understand what kind of summarization techniques are used.

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`.