

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 dataset [1]

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>.