2nd German Text Summarization Challenge

Dominik Frefel, Manfred Vogel, Fabian Märki
University of Applied Sciences Northwestern Switzerland
Institute of Data Science, Bahnhofstrasse 6, 5210 Windisch
dominik.frefel@fhnw.ch
manfred.vogel@fhnw.ch
fabian.maerki@fhnw.ch

Overview

Automatic text summarization has made tremendous progress in recent years. However, the rating of a summary is still an open research topic. Especially when it comes to measuring the abstractiveness, existing evaluation metrics like ROUGE, BLEU or METEOR show severe shortcomings.

In the 2nd German Text Summarization Challenge we aimed to explore new ideas and solutions regarding an automatic quality assessment of German text summarizations. For the challenge, we provided a text corpus together with several summaries per text. The goal was to assign a quality measure in the range from 0 (bad) to 1 (excellent) to each summary. We asked the participants to consider aspects such as correctness in content and grammar as well as facets like compactness and abstractiveness. The participants were able to submit (and resubmit) their solution to our evaluation board. The solution was evaluated automatically, and the achieved rank published on the leaderboard.

Data

The dataset provided consists of 24 distinct source texts from our German summarization corpus (Frefel, 2020). It contains one reference summary and 9 summaries proposed for evaluation for each source text. The summaries are generated by various summarization algorithms and humans. Each summary is evaluated and given a score between 0 to 1 by the task organizers. All texts are provided in lower case, with punctuation and quotations intact. The source texts are on average 786 tokens long. The reference summaries contain on average 46 and the generated summaries 38 tokens. The average compression ratio is 6%.

Rank	**Participant**	**Error**
1	David Biesner	29.037
2	UPB	31.993
3	ROUGE-1 Baseline	32.098
4	Inovex	34.630

Table 1: Challenge results

Evaluation

The participants' submissions are ranked by the mean squared error of their score predictions. We use our own German ROUGE-1 implementation as a baseline (Frefel, 2020). It scores an error of 32.098. Refer to table 1 for the results of all participants.

References

Dominik Frefel. 2020. Summarization corpora of wikipedia articles. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6653–6657, Marseille, France. European Language Resources Association.