UPB at GermEval-2020 Task 3: Assessing Summaries for German Texts using BERTScore and Sentence-BERT

Andrei Paraschiv
University Politehnica of Bucharest, Romania
Computer Science and Engineering Department
andrei.paraschiv74@stud.acs.upb.ro

Dumitru-Clementin Cercel
University Politehnica of Bucharest, Romania
Computer Science and Engineering Department
clementin.cercel@gmail.com

Abstract

The overwhelming amount of online text information available today has increased the need for more research on its automatic summarization. In this work, we describe our participation in GermEval-2020, Task 3: German Text Summarization. We compare two BERT-based metrics, Sentence-BERT and BERTScore, to automatically evaluate the quality of summaries in the German language. Our lowest error rate achieved was 31.9925, ranking us in 4th place out of 6 participating teams.

1 Introduction

The objective of the text summarization task is to generate a condensed and coherent representation of the input text, with the important ideas from it, as well as maintaining the meaning of the original (Allahyari et al., 2017). The task of automatic summarization is a hard problem since the system must understand the content, context, and meaning of the text. Most often, additional word-level knowledge is required to complete the task (Malviya and Tiwary, 2016).

In this task, a major issue is evaluating the quality of summaries that were automatically generated. Since human evaluation is expensive, time-consuming, and prone to subjective biases, automatic metrics have sparked the interest of researchers. Sharing similarities with the evaluation of Machine Translation (MT), many evaluation metrics originate in this area of research (Papineni et al., 2002).

Summarization skill assessment is often used to test the reading proficiency and the cognitive acquisitions for learners (Grabe and Jiang, 2013). In addition, the automated scoring tools of summaries can help students to improve their reading comprehension and also lead to improvements in educational applications.

There are two kinds of evaluation methods of summaries: extrinsic evaluation, where the candidate summary is judged on how useful it is for a specific task, and intrinsic evaluation based on a deep analysis of the candidate summary, for instance, a comparison with the original text, with a reference summary, or with the text generated by another automated system (Jones and Galliers, 1995).

The shared task 3 proposed by the organizers of Germeval 2020 encouraged participants to suggest a metric for an intrinsic evaluation of candidate summaries for the German text data against reference summaries. The quality of each candidate summary will be indicated by a score between 0 and 1, where 0 and 1 are a "bad summary" and a "good summary", respectively. Our approaches rely on two newly introduced measures for evaluating summary quality, Sentence-BERT (Reimers and Gurevych, 2019) and BERTScore (Zhang et al., 2019) and we assess their performance on the competition dataset to observe how well they correlate with human judgment.

In the next section, we cover the relevant work to the goal of this research task. Section 3 presents the methodology used in our case. Then, Section 4 presents the results from the experiments. Finally, we discuss the conclusions of the paper.

2 Related Work

For almost twenty years, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR

BERT Model	BERT Version	Corpora used for training
Deepset.ai[1]	Cased	Wikipedia, Legal data, news
bert-base-german-europeana-uc [2]	Uncased	Europeana newspapers
bert-base-german-uc[2]	Uncased	Wikipedia, Subtitles, News, Commoncrawl
literary-german-bert[3]	Uncased	German Fiction Literature
bert-adapted-german-press[4]	Uncased	Newspapers

Table 1: Collection of pre-trained BERT models for the German language used in our study.

(Banerjee and Lavie, 2005) are the most used metrics to assess summaries. These measures based on n-grams matching stand out through simplicity and a relatively good correlation with human evaluations. Although these metrics and their variants are widely used, there are valid objections to their limitations (Reiter, 2018).

In recent years, metrics based on word embeddings as well as measures based on deep learning models have gained more attention from researchers. Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) are dense representations for words in a vector space. Using these representations rather than the n-gram decomposition of the texts, researchers computed summary similarity scores. Either by enhancing existing metrics like BLEU (Wang and Merlo, 2016; Servan et al., 2016) or by using an adapted version of Earth Mover's Distance proposed by Rubner et al. (1998) (Li et al., 2019; Echizen-ya et al., 2019; Clark et al., 2019), these representations proved to be more in tune with human judgment than traditional measures such as ROUGE, METEOR, and BLEU.

Another application of deep learning in evaluation metrics to score summaries are measures learned by the model. For instance, models like ReVal (Gupta et al., 2015) or RUSE (Shimanaka et al., 2018) learn sentence-level embeddings for the input sentences and then compute a similarity score between them. A common architecture in summary scoring is the siamese neural network (Bromley et al., 1994). Ruseti et al. (2018) used a siamese BiGRU neural network to score candidate summaries against the source text. Further, Xia et al. (2019) proposed three architectures (i.e., CNN, LSTM, and attention mechanism-based LSTM) to assess the students for reading comprehension by scoring their summaries against the source text.

Pre-trained language models based on Transformers (Vaswani et al., 2017), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have improved performance in many tasks in the field of natural language processing in the last year. In contrast to previous word embeddings, these contextual embeddings can produce different vector representations for the same word in distinct sentences, depending on the neighboring words. Since the contextual embeddings capture also the context of the words in the token representations for the input sentences, evaluation metrics based on them tend to be more correlated with human evaluations. For instance, both the BERT adaptation for RUSE and BERT with an appended regressor did outperform the individual RUSE model (Shimanaka et al., 2019). Also, Zhao et al. (2019) shows that MoverScore, the Word Mover's Distance (Kusner et al., 2015) over contextualized embeddings, can achieve state-of-the-art performance.

3 Methodology

In our case, we adopt two novel BERT-based metrics, Sentence-BERT (Reimers and Gurevych, 2019) and BERTScore (Zhang et al., 2019) to automatically asses pairs of German candidate-reference summaries. Specifically, for the two metrics, we evaluate five different pre-trained BERT models as listed in Table 1. In each experiment, we have generated a score between 0 and 1 for every candidate-reference summary pair and then submitted the resulting file to the competition website for error evaluation.

Sentence-BERT In order to derive fixed embeddings for two input summaries (i.e., the candidate and reference summary, respectively), Sentence-BERT uses a siamese network architecture that has a pooling layer on the top of BERT. There are three scenarios available for using the

[1] https://deepset.ai/german-bert
[2] https://github.com/dbmdz/berts
[3] https://huggingface.co/severinsimmler/literary-german-bert
[4] https://huggingface.co/severinsimmler/german-press-bert

pooling layer, as follows: using the output corresponding to the [CLS] token, the mean of the vectorial representations over all 12 BERT headers, as well as the max-over-time of these output vectors. Our experiments indicated that only the mean vector scenario delivers optimal scores.

Through fine-tuning, Sentence-BERT will produce summary-level embeddings that capture both the semantic and context of these texts in a powerful way. By exploiting the cosine similarity measure, the two summary embeddings can then be compared.

BERTScore In contrast to Sentence-BERT, BERTScore is a token-level matching metric. Since BERT-based models use a Wordpiece tokenizer (Schuster and Nakajima, 2012), both the candidate (s^c) and reference (s^r) summaries are split into k and m tokens, respectively. The vector space representations v^c and v^r for s^c and s^r respectively, are then computed through 12 Transformer layers (Vaswani et al., 2017). Using a greedy matching approach, the resulting tokens are paired and the precision, recall and F1 scores are determined:

$$R_{BERT} = \frac{1}{k} \sum_{v_i^c \in v^c} \max_{v_j^r \in v^r} (v_i^c)^\top v_j^r$$

$$P_{BERT} = \frac{1}{m} \sum_{v_j^r \in v^r} \max_{v_i^c \in v^c} (v_i^c)^\top v_j^r$$

$$F1_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Additionally, we compute the inverse document frequencies (idf) based on the source text of the summaries, for each word from all candidate-reference summary pairs and use them for importance weighting in BERTScore, as described in the original paper. Also, we tested the re-scaling strategy of the scores as suggested by the authors, but the performance did not improve.

4 Performance Evaluation

4.1 Corpus

The experimental data consisted of 216 German language source texts, their reference summary, and summaries proposed for evaluation. More specifically, there were 24 distinct source texts, each with one reference summary and nine summaries proposed for evaluation. All texts were provided in lower case, with punctuation and quotations intact. The length of the source texts varied from around 2000 characters to 12000, averaging around 5800 characters. Also, the length of the reference summaries varied from 3% to 13% of the source text length with an average of 6%. Moreover, the candidate summaries varied from 0.6% length of the source text to 21%, having an average around 6%.

4.2 BERT Fine-tuning

We fine-tune the aforementioned BERT models (see Table 1) using the Opusparcus corpus (Creutz, 2018) that introduced 3168 human-annotated paraphrase pairs, sourced from the OpenSubtitles2016 thesaurus consisting of parallel corpora (Lison and Tiedemann, 2016). The paraphrase pairs are scored on a scale from 1 to 4, in 0.5 increments, where 4 is a good match and 1 is a bad match. For our purposes of fine-tuning, we translated the scores in the [0, 1] interval according to Table 2.

In order to train SentenceBERT, we used the Opusparcus dataset with the modified scores for 5 training-epochs, with a mean squared loss. Further, we use the fine-tuned BERT models as the basis for computing the BERTScore.

Opusparcus Rating	Similarity Score
4	0.85
3.5	0.70
3	0.50
2.5	0.30
2	0.20
1.5	0.10
1	0.05

Table 2: Mapping from the Opusparcus ratings to the similarity scores for each paraphrase pair used for fine-tuning of Sentence-BERT and BERTScore via BERT.

4.3 Results

In Table 3, we show the results for our experiments. First of all, we find that training SentenceBERT with the literary-german-bert and bert-adapted-german-press models and using a score translation from the Opusparcus to the [0, 1] interval delivered a more accurate evaluation.

For BERTScore, after trying out the vectors from several attention heads, we concluded that using the last layer for the token representations yields the best performance. Using the fine-tuned BERT models with Sentence-BERT as basis for

BERT Model	Sentence-BERT	BERT-Score	BERT-Score with idf	BERT-Score with fine-tuning and idf
Deepset.ai	37.2916	35.6950	35.3121	**31.9925**
bert-base-german-europeana-uc	**35.2817**	**32.9403**	**32.2169**	32.0194
bert-base-german-uc	42.7792	34.1719	33.4136	40.5780
literary-german-bert	36.5822	44.7441	43.2454	35.5773
bert-adapted-german-press	36.5098	33.1080	32.2967	35.3199

Table 3: Results for comparing the metrics: Sentence-BERT trained on Opusparcus, BERT-Score without fine-tuning, BERT-Score without fine-tuning and with idf weighting, and BERT-Score with both fine-tuning and idf weighting, considering different pre-trained BERT models of the German language.

BERTScore did improve the error rate for all pre-trained BERT models, but had a significant impact on the case sensitive version from deepset.ai, which delivered the best result of 31.9925. The fine-tuning of the uncased BERT versions with Sentence-BERT before applying BERTScore did add some improvement, but the small decrease in error may not be justified by the computational effort. On the other hand, for the cased BERT version, the increase in performance was significant.

Overall, BERTScore did perform more closely correlated with the human evaluators, regardless of the used pre-trained BERT model. The impact of the idf-weighting on the final result amounted to about 1 percentage point improvement.

As expected, since the provided summaries had no capitalization and since the importance of capitalization in the German language is significant, the case sensitive version, without fine-tuning, performed worse for both metrics. Also, the BERT model pre-trained with the Europeana Newspaper corpus performed the best for both metrics.

As seen in Table 4, the scores obtained by our best model, compared to the baselines are at least 10 percentage points better. Surprisingly, from all the baseline scoring methods, BLEU performed the best.

Baseline	Score
BLEU	41.4299
ROUGE-1	42.6328
ROUGE-2	55.7044
ROUGE-L	43.7750
METEOR	48.0823

Table 4: Results using the baseline scoring methods: BLUE, three variants of ROUGE (i.e., ROUGE-1 using unigram overlap, ROUGE-2 using bigram overlap, and ROUGE-L using the Longest Common Subsequence), and METEOR.

5 Conclusions

In this paper, we analyzed the robustness of two different metrics (i.e., Sentence-BERT and BERTScore) based on the pre-trained BERT language model, with application to automatic assessment of summary quality. Intuitively, Sentence-BERT learns embeddings for the two input summaries whereas BERTScore focuses on the token-level embeddings in each summary and computes an average score from them. Compared to classical scoring methods, like BLEU, ROUGE, or METEOR, these metrics are more compute-intensive and lack the simple explainability that classical scores provide. Also, as seen in our experiments, the scores can differ depending on the pre-trained BERT model is used.

Since BERT embeddings are context-dependent, this simpler approach, BERT-Score, proves to be more in tune with the human evaluators. Also, computationally, BERTScore is much easier to streamline since it does not require an additional training dataset. Due to the lack of qualitative and manually annotated datasets of paraphrases in German, the easiest use in production would be BERTScore with an appropriate cased model. We also showed that BERTScore applied on a BERT model fine-tuned using a paraphrase dataset and the SentenceBERT similarity objective can lead to a higher correlation between human assessments and the automatic scores.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved

correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Hiroshi Echizen-ya, Kenji Araki, and Eduard Hovy. 2019. Word embedding-based automatic mt evaluation metric using word position information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1874–1883.

William Grabe and Xiangying Jiang. 2013. Assessing reading. *The companion to language assessment*, 1:185–200.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. Reval: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072.

Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*, volume 1083. Springer Science & Business Media.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Pairui Li, Chuan Chen, Wujie Zheng, Yuetang Deng, Fanghua Ye, and Zibin Zheng. 2019. Std: An automatic evaluation metric for machine translation based on word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1497–1506.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shrikant Malviya and Uma Shanker Tiwary. 2016. Knowledge based summarization and document generation using bayesian network. *Procedia Computer Science*, 89:333–340.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE.

Stefan Ruseti, Mihai Dascalu, Amy M Johnson, Danielle S McNamara, Renu Balyan, Kathryn S McCarthy, and Stefan Trausan-Matu. 2018. Scoring summaries using recurrent neural networks. In *International Conference on Intelligent Tutoring Systems*, pages 191–201. Springer.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. 2016. Word2vec vs dbnary: Augmenting meteor using vector representations or lexical resources? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1159–1168.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with bert regressor. *arXiv preprint arXiv:1907.12679*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Haozhou Wang and Paola Merlo. 2016. Modifications of machine translation evaluation metrics by using word embeddings. In *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, pages 33–41.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2019. Automatic learner summary assessment for reading comprehension. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2532–2542.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.