Evaluating Summarization Tasks Using Sentence-BERT

Parthi Shah
pss434@nyu.edu

Praxal Patel
psp334@nyu.edu

Xintong Li
xl3269@nyu.edu

Abstract

ROUGE and BLEU have been used extensively for evaluating text summarization tasks but are biased towards surface lexical similarities. To overcome their drawbacks, we propose a novel technique using embeddings from Sentence-BERT where we produce sentence embeddings for each reference text and summarized text and compute the cosine similarity for evaluating the quality of summarization. We test our approach on the XSum Dataset which contains archived BBC news articles and human generated summaries and summaries generated using various state-of-the-art text summarization models including LEAD, EXT-ORACLE, PTGEN, CONVS2S and T-CONVS2S. Recent evaluation techniques, such as BERTScore, use contextual token embeddings for evaluating machine translation task and correlate better to human evaluations than ROGUE and BLEU. However, BERTScore is computationally expensive for as it computes pairwise cosine similarity for every token. Moreover, our results suggest that BERTScore cannot distinguish between human generated summaries and machine generated summaries. To overcome the drawbacks of BERTScore, we propose using a novel evaluation technique using embeddings from Sentence-BERT.

1 Introduction

Evaluation methods for text generation tasks such as text summarization are crucial measurements of model performance. Automatic evaluation metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation)(Lin, 2004) and BLEU (bilingual evaluation understudy)(Papineni et al., 2002) are widely used. However, these metrics are based on n-gram matching which measures the appearance of n-grams in both human reference text and model generated text. This has some major drawbacks. For example, the metrics fail to account for synonyms. For two sentences which share the similar semantic meanings but are paraphrases of each other such as “Symptoms of Coronavirus include fever and running nose” and “Nasal congestion and raised temperature are indications of potential COVID-19 infections” will have a low score on ROUGE and BLEU metrics. Also, sentences with negations can have a high ROUGE score. For example “He is a good man” and “He is not a good man”. We encounter that a good summarized text might use words that can replace phrases in the reference task. For eg. Using “Although” in place of “Despite the fact that” would lead to more precise summarized texts. ROUGE and BLEU would score these summaries lower which isn’t the ideal scenario.

Therefore, in order to address these challenges of ROUGE and BLEU, BERTScore was introduced. BERTScore computes pairwise cosine similarity between tokens (Zhang et al., 2019). The advantage of this method is that it takes contextual meaning into account and is effective for detecting synonyms and distant dependencies. And in their experiments, Zhang et al. has showed that BERTScore outperformed other evaluation metrics in machine translation tasks. The correlation BERTScore has with human judgement is significantly higher than metrics like BLEU and ROUGE.

However, there are some limitations with BERTScore which we address in this study. BERTScore is computationally very expensive with time complexity of O(n^2). Moreover, it computes pairwise similarity between every word of the summarized text and the reference text and chooses the maximum value obtained for every word. This approach works well for computing the similarity between two sentences but for summarization task, it fails to capture if the most important information of the reference text has been retrieved or not. Hence, we propose summarization evaluation using embeddings obtained from Sentence-BERT (SBERT). SBERT adds a pooling operation to the output of BERT / RoBERTa to derive a fixed sized sentence embedding. SBERT is fine-tuned on SNLI and MultiNLI Datasets (Reimers and Gurevych, 2019). We use the XSum BBC News article dataset for our metric evaluation. In this study, we compare our metric against results from BERTScore and find the correlation between human evaluation and our metric evaluation. We also show how ROUGE score fails to distinguish between human generated summaries and ranks machine generated summaries significantly higher.

As shown in Table 1, SBERT score significantly rates human summaries higher and is able to distinguish between human generated and machine generated summaries which BERTScore and ROUGE score fail at accomplishing. For bolstering our results, we had humans rank the 50 XSum summaries giving ratings from 1-6 (1 being the highest), on the basis of information retrieval, grammatical correctness and brevity. Our metric achieved a Spearman's rank correlation coefficient value of 0.7844 while BERTScore achieved 0.47428 whereas ROUGE achieved 0.16965.

2 Related Works

The existing summarization approaches are mainly of two types: Extractive and Abstractive. Extractive summarization approaches identify important sections of the reference texts and generate them verbatim. Major drawback with Extractive summarization is that the resulting summaries are not always concise. Abstractive summarization allows the model to paraphrase the source document and create concise summaries with phrases not in the source document. Recent works (Zhang et al., 2019) have focused on improving the ROUGE score for state of the art Abstractive Summarization tasks. We argue that ROUGE is not the right metric that should be improved for the Abstractive Summarization Tasks. The study conducted by Kané reviews various limitations of ROUGE and BLEU like high score for opposite meanings, low score for similar but paraphrased sentences and high scores for unintelligible sentences. The study formulates a Metric Scorecard for assessing the evaluation metrics and RoBERTa large pre-trained model fine tuned to predict sentence similarity. RoBERTa model remarkably outperforms BLEU and ROUGE for every criteria in the study. Zhang et al. signify the importance of computing token similarity using contextual embeddings instead of exact n-gram matching for scoring sentence similarity tasks.

Recent studies also shed light on additional scoring limitations of ROUGE and BLEU for negation, pronoun swapping, number swapping and entity swapping. In the study on Evaluating the Factual Correctness for Abstractive Summarization, Yuhui Zhang proposed "Factual score" which evaluates the factual correctness for abstractive summaries. The study extracts facts using OpenIE extractor methods and factual embeddings are generated for reference text and generated summaries using sentence encoder techniques . Cosine similarity is then computed between the factual embeddings of generated and the reference text for computing the final Factual score. The study shows that Factual scores are more correlated to Bert Score. This is quite helpful for our particular approach.

Many recent deep reinforcement learning studies have shown that using contextual embeddings and other distributional semantics as rewards, they have achieved better correlation with human evaluation as compared to using ROUGE as a reward/objective function. Li et al. have suggested in their study that distributional semantic representations are a practical way instead of ROUGE as ROUGE fails to capture the semantic relation between similar words. To verify their hypothesis, they have used Distributional Semantics Representations (DSR) vs ROUGE as a reward function while training their model for summarization tasks. The DSR model performs better than ROUGE on human evaluation.

Another unique study by Kryscinski et al. focuses on improving abstraction in summarization tasks. They proposed a novelty metric which is optimized during the training phase. The novelty metric is defined as the fraction of unique n-grams in the summary that are novel. They compare the relations between the normalized novelty metric and the ROUGE scores which comes out to be inversely proportional. Moreover, they depict the performance for the models that provide a good trade off between novelty metric and ROUGE scores.

There has been significant work to prove that word embeddings work better for evaluation of machine translation tasks. Mathur et al. have proved that the word embeddings created using supervised (BiLSTM + Attention) and unsupervised techniques, and a metric generated using sentence level similarity correlates with human evaluation really well. Gabriel et al. in their study for Abstractive Summarization with Narrative Flow have used BERTScore, ROUGE and Human evaluation for testing their model performance and BERTScore has very similar scores as Human Evaluation.

3 Experiments

3.1 Methodology

For the score generation implementation, the process is divided into two steps:

Generating sentence embedding from SBERT for the reference and the summarized texts: The architecture of SBERT enables fixed-sized vectors for input sentences. SBERT adds a pooling operation to the output of BERT / RoBERTa to generate that fixed sized sentence embedding. Pooling operations are flexible i.e. *MEAN* - computing the mean of all output vectors, *MAX* - computing the maximum-over-time of all output vectors and *CLS* - using the output of CLS token.

SBERT employed various concatenation strategies (depends on the task being performed) during training for sentence pair embeddings, out of which (u, v, || u-v ||) worked the best on the STS benchmark dataset(Cer et al., 2018). || u-v || implies that closer the sentence embeddings are in a vector space, more similar they might be and different pairs will be farther apart. Here, *u* is the embedding after pooling for sentence A and *v* is the embedding after pooling for sentence B. In order to fine-tune BERT / RoBERTa, SBERT creates siamese and triplet networks to update the weights such that the produced sentence embeddings are semantically meaningful and can be compared with cosine-similarity.

Computing the summarization task scores based on similarities in sentence embeddings : We produce the sentence embedding for the reference text and generated summaries by passing the sentences through a SentenceTransformer. WordEmbeddingModel and PoolingModel form the SentenceTransformer. We used ”bert-base-nli-mean-tokens”: BERT-base model with mean-tokens pooling for sentence transformation.

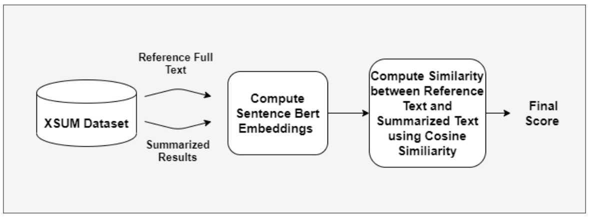
Reference Full Text
XSUM Dataset
Summarized Results
Compute Sentence Bert Embeddings
Compute Similarity between Reference Text and Summarized Text using Cosine Similiarity
Final Score

Figure 1: Model Approach (Contains summaries produced by Humans and State of the Art models such as Oracle, Lead, Ptgen, Convs2s and TopicConvs2s)

3.2 Data and Tools

The dataset we used came from the text summarization experiment conducted by Narayan et al. (2018) where they implemented 5 text summarization models: LEAD, EXT-ORACLE, PTGEN, CONVS2S and T-CONVS2S and compared the model-generated summaries with the human-written golden summaries. And the evaluation was measured using ROUGE scores. The original dataset used in their experiment is the XSum Dataset which consists of 226,711 archived BBC news articles each with a summary professionally written by humans and mostly by the article author. According to Narayan et al., XSum is less biased toward extractive methods compared to other summarization datasets. It has a very high percentage of novel ngram (36% novel unigrams, 83% novel bigrams, 96% novel trigrams, and 98% novel 4-grams) in the gold standard summaries as compared to CNN (17% novel unigrams), DailyMail (17% novel unigrams) and NYTimes (23% novel unigrams) indicating XSum summaries to be more abstractive (Narayan et al., 2018). For our baseline experiment results, we only use the testing dataset from their experiment, which contains 11,334 articles and corresponding human summaries. We also obtained the model generated summaries by the 5 models in their experiment and their ROUGE scores.

Results

Automatic evaluation: The cosine similarity scores for metrics using SBERT and BERTScore are shown in Table 1.

We implemented BERTScore using word embeddings from ‘roberta-large’, ‘bert-large-uncased’ and ‘xlnet-base-cased’. These models

Model	Sentence-BERT	BERTScore
Golden Summary	0.4850	0.8132
LEAD	0.4475	0.8048
EXT-ORACLE	0.4379	0.8152
PTGEN	0.4647	0.8147
CONVS2S	0.4390	0.8172
T-CONVS2S	0.4414	0.8162

Table 1: Cosine Similarity Score using SBERT and BERTScore (Roberta embeddings)

were chosen owing to their high Pearson Correlation with human judgement on WMT16 to-English dataset. We observed that implementation of BERTScore with 'roberta-large' resulted in extremely large values. However, it failed to distinguish between different types of summaries and therefore lacks the evaluation power whereas metrics using SBERT assigned highest score to gold summaries and demonstrated clear difference between the performance of human-generated summaries and model-generated summaries. We suggest that BERTScore is not optimized for summarization task as it takes the highest value of pairwise cosine similarity. This results in high rating of summaries with similar words but actually very low information retrieval. Scores computed using SBERT works better as SBERT is fine tuned for tasks like predicting entailment, contradiction and semantic independence and obtains significant results on Semantic Text Similarity Dataset without prior training.

Human evaluation: In addition to computing cosine similarity scores, we also asked human evaluators to rank the generated summaries based on information retrieval and grammatical errors and obtained 50 human generated rankings. We compared the human generated rankings with our evaluation metrics generated rankings and computed their spearman correlations. The correlation between human ranking and SBERT metrics ranking is 0.78344 while with BERTScore metrics is 0.47428. This result further bolstered our proposal that metrics using SBERT is a better evaluator than BERTScore.
For further comparison with traditional metrics, we computed ROUGE scores for our dataset as shown in Table 2. Thereby, ROUGE scores were not able to distinguish between human generated summaries and machine generated summaries. Human generated summaries had extremely low F1, recall and precision scores compared to machine generated summaries. The reason for low scores is the presence of novel n-grams in the summarized text. Moreover, ROUGE scores resulted in extremely low Spearman Correlation Coefficient (0.16965) when compared to human rankings using ranks produced by F-1 scores. The length of the reference texts in the XSum Dataset made it compelling to paraphrase in order to incorporate high amount information. For abstractive summarization tasks, we advocate using metrics that compare similarity using contextual embeddings rather than exact matching words.

Model	ROUGE-1	ROUGE-2	ROUGE-L
Gold	f :0.04853 p :0.33837 r :0.02785	f : 0.00331 p: 0.02918 r: 0.00184	f : 0.05897 p :0.28556 r :0.03453
LEAD	f :0.05035 p :0.35568 r :0.02892	f : 0.00341 p: 0.03075 r: 0.00189	f : 0.05750 p :0.29136 r :0.03358
EXT-ORACLE	f :0.05605 p :0.37794 r :0.03258	f : 0.00462 p: 0.03909 r: 0.00259	f : 0.06410 p :0.31325 r :0.03780
PTGEN	f :0.05564 p :0.39325 r :0.03168	f : 0.00592 p: 0.05247 r: 0.00328	f : 0.06416 p :0.34056 r :0.03690
CONVS2S	f :0.04826 p :0.38250 r :0.02721	f : 0.00493 p: 0.04895 r: 0.00271	f : 0.05756 p :0.33933 r :0.03273
T-CONVS2S	f :0.04831 p :0.37995 r :0.02725	f : 0.00488 p: 0.04771 r: 0.00269	f : 0.05788 p :0.33780 r :0.03296

Table 2: F-1,Precision and Recall scores of text summaries generated by 5 text summarization models on the test set in the (Narayan et al., 2018) experiment. Here we showed ROUGE-1, ROUGE-2 and ROUGE-L scores.

Github Link

https://github.com/praxalpatel/NLU-Final-Project

Collaboration Statement

Praxal Patel (psp334) : Worked on Related Work, Results for ROUGE score, SBERT score
Parthvi Shah (pss434) : Worked on Introduction and Methodology and Results for BERTScore
Xintong Li (xl3269): Worked on Introduction, Results for BERTScore and Data & Tools
All members contributed equally to formatting the proposal and final edits.

Appendix

Model	XLNET - Base	BERT
Golden Summary	0.3580	0.408
LEAD	0.3710	0.403
EXT-ORACLE	0.3789	0.408
PTGEN	0.3558	0.410
CONVS2S	0.3475	0.408
T-CONVS2S	0.3448	0.407

Table 3: F1 Score of text summaries using XLNET and BERT

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Zihang Dai*, Zhilin Yang*, Yiming Yang, William W. Cohen, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Language modeling with longer-term dependency.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Saadia Gabriel, Antoine Bosselut, Ari Holtzman, Kyle Lo, Asli Çelikyilmaz, and Yejin Choi. 2019. Cooperative generator-discriminator networks for abstractive summarization with narrative flow. *CoRR*, abs/1907.01272.

Hassan Kané. 2019. Towards neural similarity evaluators.

Wojciech Kryscinski, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. *CoRR*, abs/1808.07913.

Siyao Li, Deren Lei, Pengda Qin, and William Yang Wang. 2019. Deep reinforcement learning with distributional semantic rewards for abstractive summarization.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. pages 1797–1807.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. *CoRR*, abs/1707.06875.

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*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

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

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Yuhui Zhang. 2019. Evaluating the factual correctness for abstractive summarization.

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