Large Language Models Project Report BARTScore: Evaluating Generated Text as Text Generation

Sixtine Sphabmixay
sixtine.sphabmixay@dauphine.eu

Juliana Carvalho de Souza
juliana.carvalho-de-souza@dauphine.eu

Mengfei Li
mengfei.li@dauphine.eu

Mater year 2 - Mathematics, Machine Learning Sciences, and Humanities
Université Paris Dauphine-PSL
Paris, France

1 Introduction

Text generation is central to various NLP applications, such as machine translation and summarization. However, a persistent challenge in this field is evaluating the quality of generated texts, and determining whether they are fluent, accurate, and effective. In this project, we explored the paper "BARTSCORE: Evaluating Generated Text as Text Generation" [10] This paper introduces a novel evaluation metric, BARTScore, which reportedly outperforms several top-scoring metrics across various tasks and test settings.

This report is structured as follows. First, we provide a concise paper summary, highlighting its key contributions and methodology. Next, we present the results of our experiments on a summarization task using BARTScore. Finally, we offer an analysis of the paper, discussing its implications and potential limitations. The code of the reproduced experiments is available at `https://github.com/JuCarv-bit/bartscore-evaluation.git`.

2 Paper summary

2.1 Motivation for BARTScore

Recent advancements in language models for NLP rely on training with large amounts of raw text, using techniques like predicting missing words. These models have been very successful for tasks like summarizing, translations, and extracting information. They're also used to evaluate how good machine-generating texts are, with metrics like BERTScore and MoverScore. However, these frameworks don't fully use the model's capabilities, since they do not align well with how the models were originally trained. BARTScore [10] was proposed as a new way to evaluate generated text by treating evaluation as a *text generation* task. This means that instead of just comparing the generated text to a reference, BARTScore measures how likely one text could generate another. For instance:

- How likely is a machine-translated sentence (hypothesis) to be generated from the original text (source)?
- How well can a reference text (human-annotated) generate a machine-generated hypothesis?

In particular, BARTScore operates by leveraging parameters learned during the pre-training phase of BART [6], an encoder-decoder, sequence-to-sequence model, to evaluate the quality of the generated text.

The authors claimed that this approach is a better match with the underlying pre-training tasks and allows to take advantage of the parameters learned during the pre-training phase.

2.2 Introducing the BARTScore metric

BARTScore calculates the score using the log probabilities of generating tokens in a target sequence y, conditioned on a source sequence x. This is mathematically expressed as:

$$\texttt{BARTScore} = \sum_{i=1}^{m} w_t \log p(y_t | y_{<t}, x, \theta)$$

where $p(y_t|y_{<t}, x, \theta)$ corresponds to the probability of generating token y_t at step t, given the source x, the preceding tokens $y_{<t}$, and the model parameters θ. Additionally, w_t is the weight assigned to each token, which is typically uniform in the default implementation, and θ represents the BART's pre-trained parameters.

2.3 Advantages of BARTSore

As claimed by the authors of this article, using BARTScore over other metrics presents several advantages. First, BARTScore is an unsupervised measure, which eliminates the need for additional labeled data or human annotations for training. Moreover, unlike traditional metrics like BLEU or ROUGE, which rely on rigid token overlaps, BARTScore evaluates text based on its generative likelihood, capturing more nuanced qualities like fluency, coherence, and factual accuracy. Additionally, of BARTScore is that it can adapt to different evaluation tasks by varying the direction of generation:

- Faithfulness (Source → Hypothesis): Measures how well the hypothesis can be generated from the source. Useful for assessing factuality and coherence.
- Precision (Reference → Hypothesis): Evaluates how closely the hypothesis aligns with the reference text.
- Recall (Hypothesis → Reference): Checks how much of the reference content can be reconstructed from the hypothesis, reflecting semantic coverage. This is particularly useful when we want to perform a summarization task.
- F-Score (Bidirectional): Combines Precision and Recall to provide an overall measure of semantic overlap.

Furthermore, BARTScore fully leverages the pre-trained capabilities of BART, enabling it to provide robust and reliable scores without the need for complex fine-tuning.

Finally, its performance can be further enhanced through simple techniques like fine-tuning task-specific data or prompting, which align the evaluation task with BART's pre-training objectives.

3 Experiment design

We chose to replicate only the **summarization** task because vanilla BERTScore demonstrated better performance compared to other standard NLP metrics. In contrast, for tasks like Machine Translation, BARTScore showed weaker performance relative to standard metrics, with improvements only observed in its modified versions (e.g., BARTScore+prompt, BARTScore+CNN, etc.). This suggests that the authors introduced these modifications to present the method in a more favorable light.

There are other concerns about the variants. We assumed that the modified versions incorporated additional information and adjustments, shifting the focus away from the original goal of introducing the BARTScore metric. Furthermore, if any modification is applied to vanilla BARTScore (e.g., prompting or CNN), the same modification should also be applied to other metrics to ensure a fair comparison.

Next, we attempted to replicate the results presented in this paper concerning a text summarization task over the three datasets investigated in the paper:

- NewsRoom dataset NER18 [3] which contains 60 journal articles with summaries generated by seven different methods and annotated with human scores in terms of coherence, fluency, informativeness, and relevance. The article also presented promising results for machine translation, factuality, and data-to-text tasks.
- REALSumm [1], a metaevaluation dataset for text summarization which measures *pyramid recall* of each system-generated summary.
- SummEval [2], a collection of human judgments of model-generated summaries on the CNNDM dataset [4] annotated by expert judges and crowd-source workers.

Moreover, as our main goal was to access the performance of the newly introduced method of BARTScore, we decided to keep only the vanilla method, and skip the other variants, that also demonstrated increased performance in the results. The variants are described below:

- BARTSCORE, which uses the vanilla BART
- BARTSCORE-CNN, which uses the BART fine-tuned on the summarization dataset
- BARTSCORE-CNN-PARA, where BART is first fine-tuned on CNNDM, then fine-tuned on ParaBank2.
- BARTSCORE-PROMPT, which is enhanced by adding prompts.

3.1 Overview of Spearman correlation and other metrics

As mentioned in the article, our goal was to compare the BARTScore with other existing metrics. For the summarization task, we compared our vanilla BARTScore with ROUGE-1, ROUGE-2, ROUGE-L, MoverScore, and BERTScore metrics. Here is a overview of how these metrics operate:

- ROUGE-1[5]: Measures the overlap of single words between the reference and generated text.
- ROUGE-2[5]: Measures the overlap of pairs of consecutive words between the reference and generated text.
- ROUGE-L[5]: Focuses on the longest common subsequence between the reference and generated text, capturing the longest sequence of words that appear in the same order in both texts.
- MoverScore[8]: Evaluates text generation by comparing word embeddings using word movers distance.
- BERTScore[7] : Utilizes BERT embeddings to compare the similarity of individual words based on their contextual meaning, rather than just surface-level overlap.

A tool that we use to compare the results provided with these metrics is the Spearman correlation. Specifically, it is given by the following formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where d_i is the difference between the ranks of corresponding values for each data point (the rank assesses the ordering of the values rather than their absolute magnitudes), and n is the number of data points.

3.2 Results

Next, we compare our implementation for summarization with reference metrics reported in the paper.

3.2.1 Baseline for summarization

Table 2 displays the results that we obtained. Note that we used already existing packages to use the ROUGE, MoverScore, and BERTScore metrics.

Table 1: Baseline Spearman correlation scores from the analysed paper [1] for reference.

	REALSumm	SummEval				NeR18				Avg.
Metric	COV	COH	FAC	FLU	INFO	COH	FLU	INFO	REL	Avg.
ROUGE-1	0.498	0.167	0.160	0.115	0.326	0.095	0.104	0.130	0.147	0.194
ROUGE-2	0.423	0.184	0.187	0.159	0.290	0.026	0.048	0.079	0.091	0.165
ROUGE-L	0.488	0.128	0.115	0.105	0.311	0.064	0.072	0.089	0.106	0.164
BERTScore	0.440	0.284	0.110	0.193	0.312	0.147	0.170	0.131	0.163	0.217
MoverScore	0.372	0.159	0.157	0.129	0.318	0.161	0.120	0.188	0.195	0.200
BARTScore	0.441	0.322	0.311	0.248	0.264	0.679	0.670	0.646	0.604	0.465

Table 2: Our implementation of Spearman correlation of different metrics on three human judgment datasets.

	REALSumm	SummEval				NeR18				Avg
Metric	COV	COH	CONS	FLU	REL	COH	FLU	INFO	REL	AVG
ROUGE-1	0.454	0.169	0.156	0.095	0.326	0.089	0.049	0.137	0.112	0.176
ROUGE-2	0.468	0.135	0.141	0.085	0.255	0.072	0.040	0.150	0.119	0.163
ROUGE-L	0.431	0.172	0.142	0.113	0.270	0.052	0.017	0.107	0.074	0.153
BERTScore	0.441	0.293	0.134	0.149	0.375	0.169	0.154	0.131	0.176	0.225
MoverScore	0.430	0.149	0.146	0.119	0.313	0.165	0.107	0.196	0.184	0.201
BARTScore	0.297	0.413	0.341	0.268	0.384	0.624	0.593	0.598	0.567	0.435

3.2.2 Remarks on the datasets

An important remark is that there are some divergences of the columns presented on the SummEval and REALSumm datasets downloaded on the code from [9]. Firstly, besides being expected REALSumm did not have the column **COV**, the measure of Covariance, which the paper does not explicit further on how it was computed or provides references explaining the concept. Instead, the score column provided in the dataset was only "pyramid-recall". Also, SummEval provides scores for *coherence* (COH), *consistency* (CONS), *fluency* (FLU), *relevance* (REL), although those scores are reported as COH, FAC, FLU INFO, instead. Therefore, we assumed that CONS corresponds to FAC and REL corresponds to INFO.

3.3 Results Analysis

BARTScore demonstrates remarkable robustness compared to traditional metrics such as ROUGE and MoverScore, particularly in capturing semantic coverage and coherence of the generated text. For instance, on the NeR18 dataset, BARTScore achieves a significantly higher Spearman correlation in between coherence (COH) and informativeness (INFO) than ROUGE and MoverScore. The text generation perspective of BARTScore, which is based on text generation probabilities, enables it to surpass traditional surface-level matching metrics across multiple dimensions, offering a better reflection of the semantic depth and contextual relationships in the text.

The results obtained from our implementation show minimal differences from the reference values across multiple datasets and dimensions, demonstrating the reliability of our reproduction of BARTScore. A comparision between established baseline table and our results shows that our **average** correlation with respect to the three datasets is close to the original measures. This also highlights the reproducibility and stability of using the BART pre-trained model.

4 Personal remarks on the paper

After reading this article, we feel that the BARTScore method proposed by the author is very promising and inspiring. It transforms the problem of evaluating generated text into a generation task itself, which is a completely new perspective for us. Compared with traditional BLEU or ROUGE, BARTScore not only focuses on surface vocabulary matching but also uses pre-trained models to capture deep semantics and language structures, which makes us feel that its results will be more convincing.

However, we also found some confusing points in reading. For example, although the article mentions the use of prompts to improve the evaluation effect, it does not discuss in depth how to optimize these prompts, which is more of an empirical operation. In addition, BARTScore does a good job of evaluating high-quality text, but for low-quality text or some examples with poor generation quality, it seems that the performance is not so ideal. Another thing that makes us regretful is that most of its experiments are based on English data, and the article does not elaborate on its applicability in multilingual or low-resource environments, which makes us a little worried about its universality.

Furthermore, the paper omits relevant details of the experiment. As stated on session 3.2.3, it fails to explain relevant details of dataset columns, as well as some measures reported.

Overall, we feel that this article not only shows an innovative idea but also proves the potential of this method through a large number of experiments. We learned from it how to redefine complex problems and solve them more intuitively. At the same time, we also feel that this article leaves us a lot of room for thinking, such as how to further optimize this method in different tasks and language environments. If there are more practices to verify its effectiveness in the future, we believe that BARTScore will become a very influential tool.

5 Usage of AI Assistants

We used Open AI ChatGPT in part of the experiments' code, mainly in debugging and providing additional suggestions on how to make the code more readable. We also used it for an introductory understanding of concepts presented in the reviewed paper.

References

[1] Manik Bhandari et al. "Re-evaluating Evaluation in Text Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 9347–9359. DOI: 10.18653/v1/2020.emnlp-main.751. URL: https://aclanthology.org/2020.emnlp-main.751.

[2] Alexander R. Fabbri et al. "SummEval: Re-evaluating Summarization Evaluation". In: *Transactions of the Association for Computational Linguistics* 9 (2021). Ed. by Brian Roark and Ani Nenkova, pp. 391–409. DOI: 10.1162/tacl_a_00373. URL: https://aclanthology.org/2021.tacl-1.24.

[3] Max Grusky, Mor Naaman, and Yoav Artzi. "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies". In: *arXiv preprint arXiv:1804.11283* (2018).

[4] Karl Moritz Hermann et al. "Teaching Machines to Read and Comprehend". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf.

[5] Frederic Kirstein, Terry Ruas, and Bela Gipp. "Is my Meeting Summary Good? Estimating Quality with a Multi-LLM Evaluator". In: 2025, p. 14. URL: https://arxiv.org/abs/2411.18444v1.

[6] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". In: *CoRR* abs/1910.13461 (2019). arXiv: 1910.13461. URL: http://arxiv.org/abs/1910.13461.

[7] Subash Neupane et al. "CLINICSUM: Utilizing Language Models for Generating Clinical Summaries from Patient-Doctor Conversations". In: *arXiv preprint arXiv:2412.04254v1* (2024). URL: https://arxiv.org/abs/2412.04254v1.

[8] Tohida Rehman, Debarshi Kumar Sanyal, and Samiran Chattopadhyay. "Can pre-trained language models generate titles for research papers?" In: *arXiv preprint arXiv:2409.14602v2* (2024). URL: https://arxiv.org/abs/2409.14602v2.

[9] Weizhe Yuan, Graham Neubig, and Pengfei Liu. "BARTScore: Evaluating Generated Text as Text Generation". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 27263–27277. URL: https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

[10] Weizhe Yuan, Graham Neubig, and Pengfei Liu. “Bartscore: Evaluating generated text as text generation”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27263–27277.