BERTScore

Harpreet Singh

Computer Science and Engineering (CSE)
Roll No - 2022101048
International Institute of Information Technology Hyderabad

February 18, 2024

Outline

Summary

BERTScore[2]

- BERTScore, a language generation evaluation metric based on pretrained BERT contextual embeddings.
- The most common metrics(ex. *n* gram) evaluate only the surface level similarity of a sentence.
- The pitfalls in such approaches are:
  - First, such metrics often fail to robustly match paraphrases.
  - Example, existing metrics BLEU and METEOR incorrectly give high score to *people like visiting places abroad* over *consumers prefer imported cars* for the reference sentence *people like foreign cars*.
  - Second, they often don't capture the distant dependencies and penalize the semantically critical ordering changes.
  - BLEU for example, given a small window size of *n* gram approch uses 2 sized window, won't capture difference between phrases A *leads to* B vs B *leads to* A.

BERTScore contd..

- In paraphrasing detection, we need to check the accuracy of the existing metrics on their ability to evaluate semantic similarity.

Metric

Formally put, a reference setence is tokenized x is tokenized to k tokens and candidate $\langle x_1, x_2, ...x_k \rangle$ is tokenized to l tokens $\langle \hat{x_1}, \hat{x_2}, ...\hat{x_l} \rangle$, an evaluation metric is a function $f(x, \hat{x})$ which maps to $\mathbb{R}$.

Examples

- The BLEU evaluation metric is based on n-grams, typically up to the window of four.
- Other examples include n-Gram Matching Approaches, Edit-Distance-Based metrics, Embedding-Based metrics.

- BERTScore's architecture is built around several components.

BERTScore contd..

- Token Representation: BERTScore uses pre-trained BERT contextual embeddings. Embeddings to represent tokens (words or subwords) in the text. These embeddings capture the semantic meaning of the tokens in their context.
- Cosine Similarity: After obtaining the embeddings, BERTScore calculates the cosine similarity between the embeddings of the tokens in the generated text and the reference text.
- Precision, Recall, and F1 Score: BERT contextual embeddings. calculates the precision, recall, and F1 score, which are measures of the quality of the generated text.
- Importance Weighting(Optional): Different weights are assigned to different words based on their importance. As some rare words can also determine the context of the text. Value of 1 is assigned to important tokens.
- Baseline Rescaling: This is another optional step. It is done to adjust BERTScores with baseline of 3 to make them more human readable.

Advantages and Disadvantages

Advantages and Disadvantages

Advantages:

- It is more accurate as it leverages state-of-the-art Google BERT's contextual embeddings.
- Captures deep semantic relationships which is benefitial over other metrics which capture syntactic similarity.

Disadvantages:

- There is no one configuration of BERTScore that clearly outperforms all others.
- It relies on BERT's embeddings, therefore it often requires advanced hardware and time for translation.

While the differences between the top configurations are often small, it is important for the user to be aware of the different trade-offs, consider the domain and language specific state-of-the-art configuration to use

3 Strengths of the paper

3 Strengths of the paper

Authors have done a commendable job in their research and field of expertise.

- They have presented their own implementation of the topic on Github. They have diligently cited all the reference papers to facilitate further research and citation. They have kept the tone of paper formal and mathematically oriented, with inclusion of the formulas, on such a difficult topic.
- They have judged popular existing metrics and provided where they think is the scope of further improvement. They introduced their own metric along with mathematical formulations on the subject and suggested ways in which it is better and human interpretable.
- They have put considerable effort in experimentally verifying their research, tabulating their findings and comparing them across the different state-of-the-art metrics.

3 Weaknesses of the paper

3 Weaknesses of the paper

Following are the weaknesses that I believe exist in the paper

- Limitation and Bias: It is biased towards certain model's underlying structure hence inherits certain drawbacks too. Also this leads to the fact that reference-free metrics are not unique to the model.
- ADC Data[1]: It remains unclear how BERTScore will perform on adversarial datasets which attempt to produce examples that elicit incorrect predictions, throughout the paper.
- Resource Consumption: One of the major drawbacks faced while trying testing on paws(labeled-final) dataset, includes more memory consumption and higher time complexity as compared to its predecessor BLEU.

3 Improvements in the paper

3 Improvements in the paper

- Paper should include testing of Adversarial datasets and any other datasets which tests the BERTScore more than syntactic structure such as idiomatic expressions or cultural references.
- In the future, the authors should consider scaling the model for a pair of languages where the words are not directly comparable.
- Also, the model should be able to compare between a bad and the worst output and clearly classify the best output from the available options.

References


Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih.
On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study.
In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6618–6633, Online, August 2021. Association for Computational Linguistics.


Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi.
Bertscore: Evaluating text generation with bert, 2020.


Thank
you