PRECOG

2024 April 6

BERTScore: Paper Review

Gagan Aryan

Table of Contents

Problem Statement

Given a reference sentence x tokenised to k tokens <x1, x2 …. xk> and a candidate sentence x' tokenised to l tokens <x'1, x'2 … x'l>, devise an evaluation metric f(x, x') ∈ ℝ such that f(x, x') has a high correlation with human judgement.

Proposed Solution

Given the tokenised reference sentence <x1, x2 … xk> and a candidate sentence <x'1, x'2 … x'l> contextual embeddings <**x**1, **x**2 … **x**m> and <**x'**1, **x'**2 .. **x'**n>. The main model to generate embeddings is BERT. The resultant recall, precision and f1 scores are given as below:

$$R_{BERT} = \frac{1}{|x|}\sum_{x_i \in x} \max_{x'_j \in x'} \boldsymbol{x}_i^T \boldsymbol{x'}_j \qquad P_{BERT} = \frac{1}{|x'|}\sum_{x'_j \in x'} \max_{x_i \in x} \boldsymbol{x}_i^T \boldsymbol{x'}_j \qquad F_{BERT} = 2\frac{P_{BERT} \times R_{BERT}}{P_{BERT} + R_{BERT}}$$

Optionally idf weights are also added using the reference sentence and test corpus. So for a given token w the resultant idf score is given below given M reference sentences, the function I is an indicator function i.e., returns 1 if a token w is present in a reference sentence and 0 otherwise. To improve score readability baseline rescaling is done for the BERTScore metrics. Baseline b is computed using 1M random pairs of candidate and reference sentences in the commoncrawl dataset and then averaging the BERTScores obtained by these sentence pairs across a language. The rescaled value for recall is given below:

$$\text{idf}(w) = -\log \frac{i}{M}\sum_{i=1}^{M} I[w \in x^i] \qquad R'_{BERT} = \frac{R_{BERT} - b}{1 - b}$$

Contextual Embedding Models

Twelve contextual embedding models are evaluated in the paper. These include variants of BERT, RoBETa, XLNet and XLM. 24-layer RoBERTs (large) model is used for english tasks, 12-layer BERT (chinese) models for chinese tasks and 12-layer case multilingual BERT (multi) model for other languages

Machine Translation

WMT18 is used as the primary evaluation corpus, this contains predictions of 149 translation models across 14 language pairs and two types of judgement scores. Results of WMT17 and WMT16 are also reported. Absolute Pearson correlation and on top of this Williams test is performed to check statistical significance of the difference with other evaluation mertrics. For the performance of ranking different models, randomly 100 out of 10k systems are selected and the systems are ranked using the automated metrics. The percentage of metric ranking agreeing with human metrics is reported.
While the Pearson correlation coefficient is used to assess the linear relationship between the outputs of human annotations and model outputs (system level), the Kendall correlation coefficient is used to check for agreement in the ranking between human-evaluated metrics and model-generated metrics (segment level).

Image Captioning

Human judgements of twelve submission entries from the COCO 2015 Captioning Challenge is used. From this each participating model generates caption for each image and each image has five reference captions. After this two Pearson correlation metrics are computed: percentage of captions better or equal to human captions and percentage of captions indistinguisable from human captions. Eight task-agnostic and two task specific metrics are compared here.

Machine Translation

BERTscore overall shows a strong correlation with human judgements and the other metrics rarely show significantly better correlation than BERTScore. However, RUSE metric does demonstrate competitive results in to-English translations presumably because of its supervised nature and reliance on specific datasets. However whether applying idf weighing helps for this task is unclear since the improvements in score is neither significant nor consistent in comparison to metrics calculated without applying the weights. It is also demonstrated that the F_BERT performs reliably well across different settings and hence is recommended.

Image Captioning

BERTScore outperforms all the task-agnostic metrics by large margin for the image captioning set up. It is also demonstrated in the results on how weak correlation with the typical n-gram based matching metrics (like BLEU and ROUGE) is. It is also worth noting that the idf importance weighing shows significant benefit for this task. However it is worth nothing that LEIC a metric that is optimised for the COCO Captioning Challenge dataset outperforms all of the other metrics.

Much more holistic way of measuring similarity

BERTScore addresses the primary drawback of other n-gram based metrics, which is comparing words verbatim can be a flawed method of comparing sentences. Since BERTScore takes context into consideration, it is by far a much more nuanced and holistic way of understanding the performance of a system.

A solution to Image Captioning task evaluations

It is has been shown in the paper that none of the task-agnostic metrics correlate well with the human metrics for image captioning tasks and only the fine tuned ones seem to perform well. But even among that LEIC is fine tuned specifically for image captioning challenge and hence may not have performed well on other datasets of image captioning tasks. The introduction of BERTScore serves as a new solution for this.

Comprehensive experimentation

The authors carry out thorough experimentation demonstrating that the BERTScore metric performs better on a number of contextual models. Besides, the experimental setup takes both system-level and segment-level performance individually and analyses whether the difference in the metrics is statistically significant in comparison to the others.

Computationally expensive

Although it has been conclusively demonstrated that BERTScore performs significantly better in comparison to other metrics, it also highly computationally expensive in comparison to n-gram metrics since it involves generating contextual embeddings of the tokens and using those to calculate the scores. This computationally expensive nature could potentially lead to this metric not being used in real-world applications.

Fails to explain IDF behaviour

For the case of machine translation task the applying IDF weights does not show consistent results and it is unclear as to in what cases applying IDF weights is beneficial and in what cases it is not beneficial.

Unclear performance with creative texts

While the experiments have demonstrated BERTScore's ability to capture nuances in machine translation and image captioning tasks and the robustness in paraphrase classification has also been discussed, it is unclear to me whether the metric would continue to give similar results in texts involving creative expressions and figures of speech where even BERT may not be able to capture the context of the text well enough.

Lacks domain specific fine tuning insights

While the experiments carried out are highly comprehensive, they do lack domain specific fine tuning insights. Some set of recommendations or guidelines on how one would modify BERTScore for domain specific problems where a vanilla BERTScore could potentially perform worse than the other metrics that are designed to handle specific tasks would have been a great addition to the paper. This could have even been an addition as simple as performing some sort of fine tuning that could have potentially helped bring the performance of BERTScore better than LEIC for the image captioning task.

Experiment with low-resource languages

It might have been worth investigating how BERTScore performs in cases of language translations tasks involve low-resource languages. This would have helped uncover limitations with BERTScore also could have provided insights on how adaptable the overall metric is on datasets that are not as comprehensive as the ones that have been used in the paper.

Comparison with static embeddings

Given that BERTScore computation is expensively primarily because of the contextual embeddings generation step, it might have been worth exploring how using static embeddings instead of contextual ones would compare with the proposed set up. An answer to the question on whether we can use static embeddings, fine tune it according to the specific task and get better results in comparison to a regular BERTScore could have been interesting.