BERTScore

Ch. Vishnu Sathwik
2024121002

Introduction to BERTScore

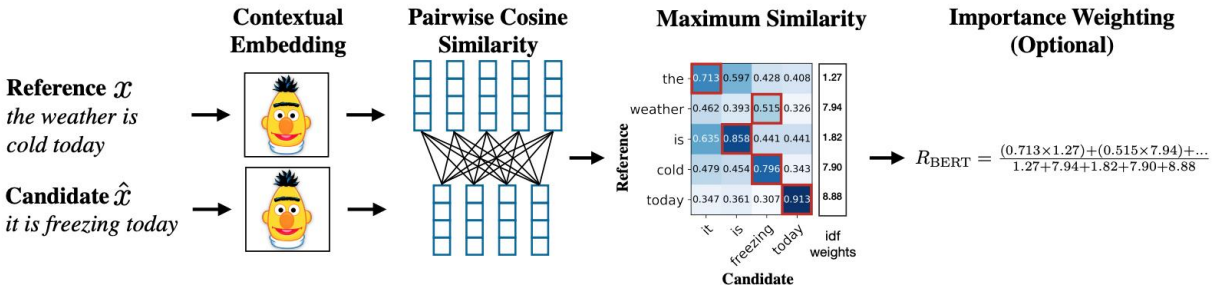
- **Purpose** : BERTScore is a metric introduced to evaluate the quality of text generated by language models.
- **Motivation :** Text evaluation metrics like BLEU, ROUGE, METEOR depends on surface level n-gram match to give scores. This might fail in cases where different words with same meanings are appear in reference text and candidate text.
  - **Ex** : Candidate text : 'people like foreign cars'

    Reference text : 'consumers prefer imported cars.'

    Traditional n-gram based metrics fail to give appropriate score in such cases
- **Working :** BERTScore uses BERT model to generate contextual embeddings to calculate similarity scores.

How BERTScore Works?

- **Token embeddings :** Word embeddings are generated for each token in the candidate and reference texts using pretrained BERT model.
- **Matching :** Cosine similarities are calculated between every token in candidate text to reference text. Maximum similarity score words are matched.



- **Scoring :**
  - Precision measures how well the generated tokens match reference tokens
  - Recall measures how well the reference tokens covers the reference tokens
  - F1 score is the harmonic mean of Precision and recall

Advantages and Limitations

- Advantages :
  - Captures semantic meaning by considering context through BERT's embeddings.
  - More robust to paraphrasing and word order variations than traditional metrics.
  - Applicable to multiple languages and tasks with appropriate pretrained models.

2. Limitations :
   - Generating embeddings from BERT is computationally expensive.
   - Word representations are dependent on the training data. OOV problem!
   - Biases in the model can affect the evaluation score;

Three major strengths:

1. Paper explains the results with a wide range of datasets comparing with results of other metrics too.
2. BERTScore metric is robust to minor variations in word choice and order, which other metrics like BLEU can give low score.
3. As BERT is trained on a huge amount of data across all the domains, BERTScore can be applied to any domain domain.

Three major weakness;

1. Paper fails to address the effect of biases in the model in the evaluation.
2. BERT embeddings can be replaced with next generation embeddings such as LLM2Vec which can capture meanings more effectively. So the metric can be replaced easily.
3. BERTScore cannot evaluate grammatical correctness, fluency, or stylistic quality of the text. So this may give high scores to text which is semantically correct but ungrammatical, unstructured.

Possible improvements

1. Addressing of biases present in BERT.
2. Paper should have addressed about the results when embeddings from other LLMs used.
3. Addressing the draw backs of BERTScore more effectively.