

BERTScore

Marek Kadlčík, 485294

Problem statement

Given: machine translation output and reference translation

Compute: reasonable similarity score between the two

(This problem appears also in automatic image captioning, generative question answering...)

Reminder of existing solutions

- BLEU score
- word error rate
- precision and recall (or f1) of individual words
- METEOR
- ...

What makes a metric good?

- agreement with human judgement
- computational speed

BERTScore algorithm

1. `embeddings_1` \leftarrow BERT(reference translation)
2. `embeddings_2` \leftarrow BERT(machine-translated sentence)
3. `C` \leftarrow cosine similarity matrix, i.e.:
`C[i, j] = cos_similarity(embeddings_1[i], embeddings_2[j])`
4. `recall` \leftarrow take max in each row and compute average
5. `precision` \leftarrow take max in each column and compute average
6. return `F1(recall, precision)`

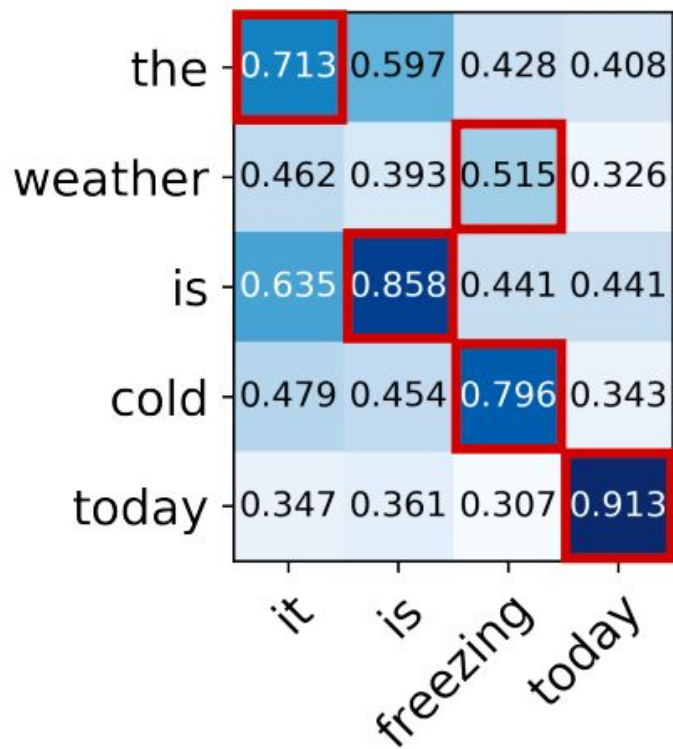
Example

Reference translation:

The weather is cold today.

Machine translation:

It is freezing today.



`recall = avg(0.713, 0.515, 0.858, 0.796, 0.913)`

`precision = avg(...)`

(Authors also try a variant with word weighting - not all words are equally important)

Properties

- not as fast as simple metrics (BERTScore requires evaluating BERT)
- has high agreement (~ 0.95 correlation) with human judgement

For detailed analysis of agreement with human judgement see the original paper, sections *Experimental setup* and *Results*.

Implementations

Author's implementation (pytorch):

- github: https://github.com/Tiiiger/bert_score
- pypi: <https://pypi.org/project/bert-score/>

Huggingface transformers:

- <https://huggingface.co/metrics/bertscore>

Sources

<https://arxiv.org/pdf/1904.09675.pdf>

<https://jlibovicky.github.io/2019/05/01/MT-Weekly-BERTScore.html>