Evaluating Multilingual Sentence Representation Models on a Real Case Scenario

Rocco Tripodi, Rexhina Blloshmi and Simon Levis Sullam

Paraphrase detection is the task of analyzing two segments of text and determining if they have the same meaning despite the differences in structure and wording. It is hard to apply in real case scenarios, since each lexical form has a definite connotation.

Definition 1.

(quasi-) paraphrases, i.e., sentences or phrases that convey approximately the same meaning using different words (Bhagat, and Hovy, 2013).

Definition 2.

Paraphrase detection

Datasets

- are mainly in English
- Area mainly constructed automatically
- comprise short sentences
- consist of paraphrase pairs with high word overlap
- have simple syntactic structures
- are mostly formulated as binary classification
- do not provide retrieval tasks

Experimental Setup

Tasks

Correlation: evaluate the correlation between the paraphrase similarity scores calculated using different representation models, and that assigned by human annotators.

Classification: given an input sentence from the Protocols marked as a paraphrase by the annotators, find the corresponding sentence in the full text of the Dialogue book.

Retrieval: using the full text of both books to find how many paraphrases a model is able to detect

Models

- Static embeddings: mean pooling of word2vec and ConceptNet
- Contextualized embeddings: mean pooling of BERT, Camem-BERT and XLM-R embeddings (cosine similarity)

Resources

- https://github.com/roccotrip/protocols

Conclusion

- Dataset for qualitative evaluation
- Protocol for evaluation
- We conducted the evaluation on two languages, and we plan to extend it to German, Italian, and French.
- It would be difficult to use the existing models in real-world scenarios, since they can mainly detect highly similar paraphrase sentences.
- As much as our presented evaluation represents a real task, it is rather simplified. This is because we already presented in the results the two books in which the translation artifact actually exist.
- In a real real-world scenario, the task of finding paraphrases would be broader including an extensive detection of heterogeneous tasks to search for.

Drawbacks of current benchmarks

- Conversion (a) to AUC measure for each model on the French dataset.
- Conversion (b) to AUC measure for each model on the English dataset.