Understanding of long texts

Jan Flajžík

ČVUT - FIT

flajzjan@fit.cvut.cz

December 31, 2024

1 Introduction

Due to rapid development in the area of natural language processing (NLP) the need for embeddings of high quality is higher than ever before. So, in this semestral project, we created a research of existing libraries for generating embeddings of long texts. After that we performed benchmarks on existing datasets with three selected libraries and finally used two of those libraries that performed the best in the benchmarks for the text classification problem.

2 Libraries

There are numerous existing libraries for long text embeddings. In this project we mainly focused on accessible ones.

2.1 mUSE

Multilingual universal sentence encoder (mUSE) developed by Google uses transformer based encoder and deep averaging network for generating 512-dimensional sentence embeddings. [1] This model only supports 16 languages and does not support Czech language which we worked with in one of the following benchmark. The model is available for TensorFlow on Kaggle[1]

2.2 LaBSE

Language-agnostic BERT Sentence Embedding (LaBSE) was developed at Google [4] and supports over 100 languages. The architecture is based on BERT and uses dual encoder for generating cross-lingual embeddings and the output dimension of the model is 768. The model is available on huggingface[2] or Kaggle[3].

2.3 Info-XLM

Cross-lingual Language Model (XLM) is a multilingual language model developed by Microsoft which extends Cross-lingual Language Model - RoBERTa (XLM-R) [2]. It supports 100 languages and the output dimension of this model is 768. In the benchmarks we used info-xlm base available at Hugging Face[4].

2.4 Unused libraries

There are multiple other libraries for generating multilingual text which we did not cover in this project. We present two examples of them in the following list.

- SONAR or Sentence-level multimodal and language-agnostic representations is a multilingual and multimodal fixed-size sentence embedding space [3] developed by Meta,
- LASER or Language-agnostic sentence representations, also developed by Meta which employs BiLSTMs in its architecture.

3 Benchmark datasets

3.1 Tatoeba

Tatoeba is a multilingual benchmark dataset that contains about 500 000 translation in about 400 languages [5]. For our benchmark we selected 4 languages: German, Russian, French and Czech. For comparasion we used English translation of sentences in those languages. Since the dataset contains a lot of short sentences we removed all below 100 characters. Finally, we were left with dataset of about 13 000 text and their English translation. The dataset is available on Hugging Face[5].

3.2 BUCC

BUCC dataset contains 35 000 sentences in four different languages: French, German, Russian, Chi-

[1] https://www.kaggle.com/models/google/universal-sentence-encoder

[2] https://huggingface.co/sentence-transformers/LaBSE

[3] https://www.kaggle.com/models/google/labse/tensorFlow2/labse/1?tfhub-redirect=true

[4] https://huggingface.co/microsoft/infoxlm-base

[5] https://huggingface.co/datasets/Helsinki-NLP/tatoeba_mt

nese. For each sentence in those languages there is a English translation [6]. We used the whole dataset for The dataset is also available on Hugging Face[6]

4 Benchmarks

The benchmark was done using bitext retrieval which means that we generated embeddings for all texts in the batch in both languages. After that we computed cosine similarity between all embeddings which was followed by calculating accuracy based on the similarity. Finally, we computed average accuracy among all batches. The results are displayed in tables 1 and 2.

model	de-en	fr-en	ru-en	zh-en
infoxlm	0.97	0.97	0.97	0.94
LaBSE	1.00	1.00	0.98	1.00
mUSE	1.00	0.99	0.98	0.99

Table 1: Average accuracy performance of the BUCC dataset

model	fr-en	ru-en	de-en	cs-en
infoxlm	0.93	0.94	0.93	0.92
LaBSE	0.94	0.96	0.95	0.95
mUSE	0.94	0.96	0.96	0.35

Table 2: Average accuracy performance of the Tatoeba dataset

As we can see, LaBSE and mUSE got slightly better results on both examined datasets. This has one exception of Czech language which is not supported by mUSE which led to very poor performance. Nevertheless in the classification task we will employ LaBSE and mUSE since they performed better on both benchmarks.

5 Classification task

Finally, we employed the examined libraries for a text classification task. We chose to classify the overall sentiment of multilingual sentences in Massive Multilingual Sentiment Corpora [7] into three classes. The corpora is available on Hugging Face[7]. This dataset consists of over 6 million texts in 27 languages. We reduced the dataset to those consist only texts of those languages we experimented with in the benchmarks and added the texts in Spanish as it is the second largest spoken language in the world. We also removed all texts shorter than 140 characters as a lot of texts were taken from Twitter. Finally, we reduced the dataset to consist the same amount of text for all languages. This left us with dataset of 9 000 texts.

For classification we used basic model that consisted of embedding, input layer with dropout and output layer. Since the dimension of embeddings generated by the selected libraries vary, the input dimension was adjusted by them. The model architecture is depicted in figure 1. For training of the model we split the dataset to train, validation and test subset in ratio of 60:20:20. The final model was trained on both train and validation dataset. Training was done using batch size of 64. We selected AdamW with cosine scheduler with warm up. The loss function was cross entropy loss.

The only hyperparameters we tuned were the dropout probability which was set to 0.2 and 0.5 and the learning rate which was after first few experiments set to 0.1 and 0.01. The number epochs was originally set to 10, but after few experiments we noticed that there was no significant improvement in performance in later epochs and set the value to 5.

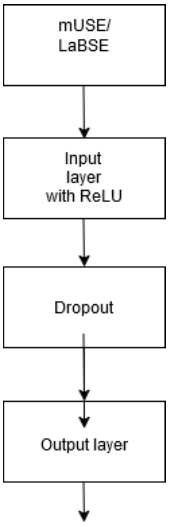

Figure 1: Classification model architecture

Since the performance was relatively poor we also took a look which language from test dataset was mislabeled the most. Figures 2 and 3 depicts the ratio of incorrectly classified data points for both of the library. Although the mUSE does not support the Czech language, therefore performed poorly in the Tatoeba benchmark, the highest number of texts that were incorrectly classified was in Spanish in both cases.

[6] https://huggingface.co/datasets/mteb/bucc-bitext-mining

[7] https://huggingface.co/datasets/Brand24/mms

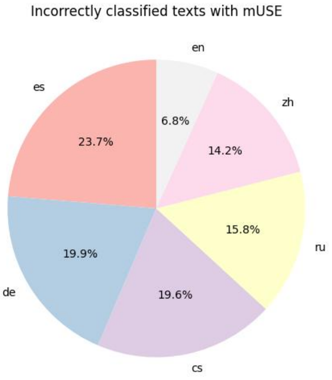

Figure 2: The ratio of incorrectly classified text per language using mUSE

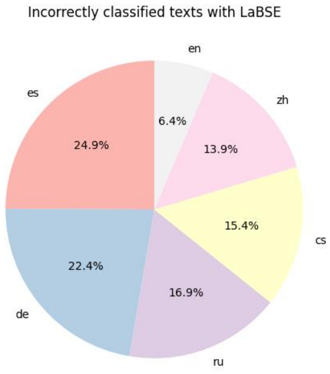

Figure 3: The ratio of incorrectly classified text per language using LaBSE

5.1 Results

The results of the classification model are displayed in table 3. As we see the performance of the model was relatively poor. The reason for this might be the simplicity of the model or the diversity of the dataset.

model	f1 score
LaBSE	0.65
mUSE	0.60

Table 3: F1 classification score per library

6 Future work

This project can be extended by adding other libraries from the unused libraries section or those which we did not discussed at all. Also the libraries can be used in numerous other tasks other than text classification.

7 Conclusion

In this semestral work, we researched existing libraries for generating embeddings of long texts. We performed a benchmark on three of them which resulted in LaBSE outperforming mUSE and Info XLM on all examined languages. Finally, we employed LaBSE and mUSE in the text classification task in which that used LaBSE for embedding generation reached better results.

References

[1] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

[2] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training, 2021.

[3] Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Sonar: Sentence-level multimodal and language-agnostic representations, 2023.

[4] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding, 2022.

[5] Jörg Tiedemann. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online, November 2020. Association for Computational Linguistics.

[6] Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In Serge Sharoff, Pierre Zweigenbaum, and Reinhard Rapp, editors, *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[7] Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark, 2023.