Team Performance Pressure@Task01: Evaluating Multilingual Text Similarity leveraging Pre-trained Language Models

Adarsh Kumar, Animesh Jain, Tanuj Saraf, Abhinav Bohra, and Ishan Sharma
18MA20003, 18CS10004, 18IE10039, 18CS30049 and 18CS30020
{adarshkumar712, animesh99, saraft20, abhinavbohra}@iitkgp.ac.in
ishan996@iitkgp.ac.in

Abstract

In this paper, we describe our baseline approaches along with their results on Shared Task 01: SemEval Task 8 on Multilingual News Article Similarity. We further added our proposed next steps towards the completion of this project. So far, we experimented with two pretrained models, mBert and XLM-Roberta on two baselines: **Zero-shot** and **Vanilla Fine-tuning**. We assessed the model performance using the Pearson Correlation score. We observed a significant improvement in Pearson correlation score with finetuning over the zeroshot score, with >**0.3** improvement in score for both the models.

1 Introduction

As news articles are written in various languages and by various people, it becomes very difficult to keep track of articles which contain news about the same events. We try to solve this problem by computing the similarity between a pair of news articles - as a score on a 4-point scale (with 4 being most similar) - irrespective of the language they are written in.

2 Task and Dataset Description

The problem statement can be summarized as follows: ”Given a pair of news articles, rate them on a 4-point scale from most to least similar”.

The training dataset provided by the SemEval Task 8 consisted of 4,964[1] pairs of news articles (cross-lingual and monolingual pairs included) in the following languages: English(en), Spanish(es), German(de), Polish(pl), Turkish(tr), Arabic(ar) and French(fr). The news articles were provided as urls to the publisher's websites which were downloaded using a publicly available python script. Distribution of pairs is as follows: en-en: 1800, de-de: 857, de-en: 577, es-es: 570, tr-tr: 465, pl-pl: 349, ar-ar: 274, fr-fr: 72.

The evaluation data, consisting of 4954 pairs[2] of news articles, is given in a similar format without the additional information of Geography, Entities, time, Narrative, Overall, Style and Tone columns. It also contains news articles in unseen languages. Some more information on dataset can be seen in Section A.

3 Baseline Approaches

We implemented two different baselines, namely zero shot and vanilla fine-tuning. For each of the baseline, we experimented with two models, Multilingual-Bert (Devlin et al., 2018) and XLM-Roberta (Conneau et al., 2019). Let's discuss the baseline approaches in detail.

Zero Shot

In this approach, we evaluated the zero shot (no training / fine-tuning) performance of large pre-trained models. Here, we treated the output corresponding to [CLS] token (or pooled output) as contextualized embeddings. These embeddings, evaluated for each document text(truncated to max length) of a given pair, were further used to calculate cosine similarity between the pair. Since the cosine similarities lie between -1 and 1. , the result was then further scaled between 1 to 4, which was then used as the predicted similarity score for given pair.

Vanilla Fine-tuning

Here, we evaluate the performance of pre-trained models after fine-tuning them on dataset. The model in this approach consist of a pre-trained langauge model followed by regressor layer (i.e., a

[1] Out of these, 24 pairs were inaccessible while downloading, so we excluded those pairs for our assessment A

[2] Out of these, 243 pairs were inaccessible while downloading, so we excluded those pairs for our assessment A

Dropout and Linear layer). The contexualized embeddings for each sentence of a given pair, obtained from model [CLS] token output (or pooled output) were concatenated before feeding into the regressor layer. We finetuned the entire model along with regressor layer using the MSE loss.

4 Experiment and Results

All the experiments were performed using Pytorch (Paszke et al., 2019) and HuggingFace (Wolf et al., 2019) Transformers library. As the implementation environment, we used Google Colab which is a free research tool with a Tesla K80 GPU and 12GB RAM. Optimization was done using Adam (Kingma and Ba, 2017) with a learning rate of 5e-5. As discussed earlier in section 3, in our experiments, we used pre-trained mBert and XLM-Roberta models available in Hugging-Face library. Input sentences were tokenized using respective tokenizers for each model, with maximum sequence length restricted to 512 tokens for mBert and 1024 for XLM-Roberta. We trained each regression model with a batch size of 16. In all the approaches, we used only the first token output provided by each model as input to regression layer. Each regression layer has 1 dropout layer with dropout of 0.2 and 1 fully connected layer. For the evaluation, we have used weighted pearson correlation score and mean squared error (MSE) as a metric for measuring the performance. Results for the baselines are illustrated in Tables 1 and 2.

Baseline Model	Pearson
mBERT	-0.0047
XLM-Roberta	-0.2885

Table 1: Results for ZeroShot baseline models.

Baseline Model	Pearson
mBERT	0.3035
XLM-Roberta	0.2879

Table 2: Results for Vanilla Finetuning Baseline models.

As we can see from the result tables, the Vanilla Finetuning Baseline models clearly outperformed the zero-shot baselines. However, we observed slight overfitting in the finetuned models, which we plan to address in our next step.

5 What's next?

While our finetuned baseline approach did outperform the zero shot performance, there are still several aspects on which we can work on. As our next step, we plan to focus on the following aspects:

- Currently the model is trained on entire training dataset, which might have caused overfitting. For the next set of experiments, we have created a train validation split (with pairs in Validation set), with all the pairs from languages labelled as 'ar' and 'fr' as unseen languages in validation set (similar to evaluation data which also have unseen languages).

- As of now, we haven't used extra information provided with main text of dataset (i.e. similarity scores for Geographies, Time, Narrative etc). As next step, we would like to experiment whether inculcating those into pipeline will improve the performance or not. We plan to do so by adopting a Multi-Task approach (Zhang and Yang, 2017).

- We plan to further experiment with some different loss functions along with MSE Loss, like Cosine Similarity Loss, Constrastive loss to improve learnt sentence embeddings (Gao et al., 2021).

- As we have already mentioned, the dataset text is larger than the maximum length allocated for that model in most of the cases. Our current approach of truncation leads to loss of information. Instead of truncating the sentences to maximum length, we would like to experiment with Multilingual-Summarisers to shorten the text length (Thanh et al., 2015).

6 Contribution Details

All members of our team generally contributed to many aspects of the project when needed, but some members focused on specific tasks more than others. Tanuj Saraf mainly worked on the data handling, and some other exploratory analysis related to data. Adarsh Kumar worked on Zeroshot Baseline and literature review for next steps. Animesh Jain and Abhinav Bohra implemented the second baseline i.e. fine-tuning using mBERT and XLM-Roberta. Ishan Sharma worked on the literature review and writing the report. Further for the next steps, we plan to distribute the different aspect points we mentioned, across members equally.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Vo Duy Thanh, Vo Trung Hung, Ho Khac Hung, and Tran Quoc Huy. 2015. Text classification based on svm and text summarization. *International journal of engineering research and technology*, 4.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Yu Zhang and Qiang Yang. 2017. A survey on multitask learning. *CoRR*, abs/1707.08114.

A Appendix

Please note that we have ignored all those pairs for which atleast one of the files were inaccessible, and assessed our model for all the available files only.

Furthermore, here is the distribution of Token lengths across the dataset using tokenizers of mBert and XLM-Roberta:

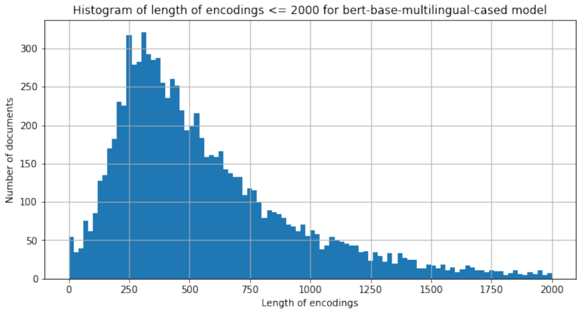

Figure 1: Token Length Distribution with mBert tokenizer

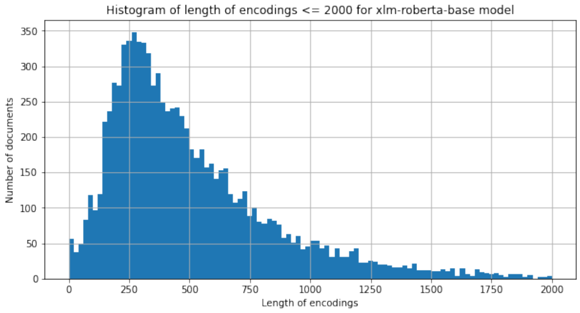

Figure 2: Token Length Distribution with XLM-Roberta tokenizer

B Supplemental Material

Links to the Huggingface models used for our baseline:

- mBert: `https://huggingface.co/bert-base-multilingual-cased`
- XLM-Roberta: `https://huggingface.co/xlm-roberta-base`