Team Performance Pressure @SemEval Task 8, 2022

Multilingual News Similarity

Adarsh Kumar, Animesh Jain, Abhinav Bohra,
Ishan Sharma , Tanuj Saraf

IIT Kharagpur

Introduction and Task Overview

- The **SemEval 2022 Task 8 (Multilingual News Article Similarity)** is aimed at developing systems that identify multilingual news articles that provide similar information by rating them on a real-valued [1 – 4] scale, from most to least similar.

- **Application:** The advent of news aggregators that collect thousands of articles from different publishers and websites has increased interest in detecting news stories connected to a specific theme & integrating them into news clusters. It's been urgent & difficult to track the similarity of news coverage across different publications/locations.

- As such news reports may not only be in English, there is also a need for effective multilingual systems. Therefore, recent research has concentrated on detecting similarities between documents, phrases, and stories.

2

Dataset

- The training dataset consisted of 4,9641 pairs of news articles (cross-lingual and monolingual pairs included) 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 table below shows distribution of language pairs across train, validation and test set. It is evident that there are more monolingual pairs in all the sets as compared to crosslingual pairs.
- The three "surprise" languages that were not present in the training data are Chinese, Italian, and Russian

Language Pair	Train Set	Val Set	Test Set
Monolingual	3603	784	3462
Cross-lingual	515	62	1440
Unseen pairs	NA	346	3131

Table 1: Counts of different types of language pairs

Dataset

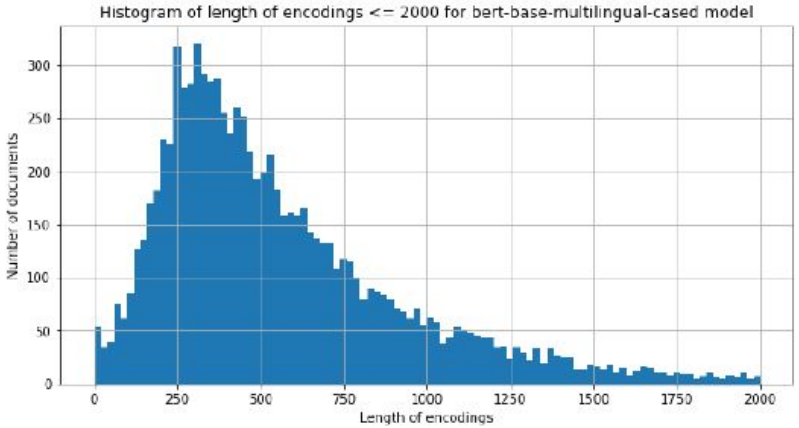


Figure 2: Token Length Distribution with mBert tokenizer

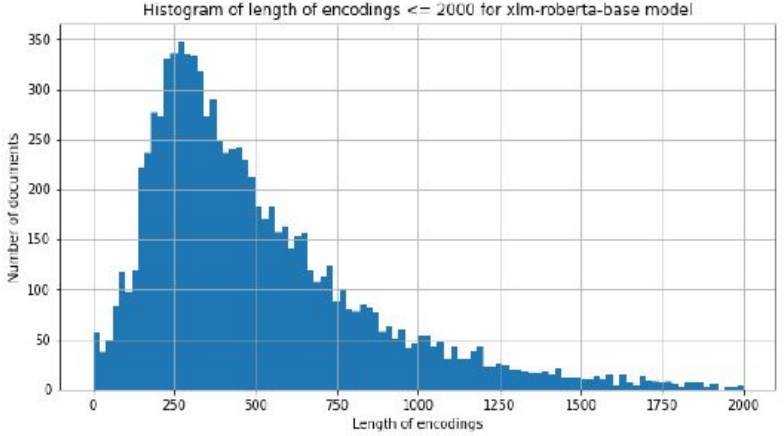


Figure 3: Token Length Distribution with XLM-Roberta tokenizer

Dataset statistics and Label distribution

Methodology

A discussion on Strategies

Strategies

1. Sentence Transformer (ST) based Similarity

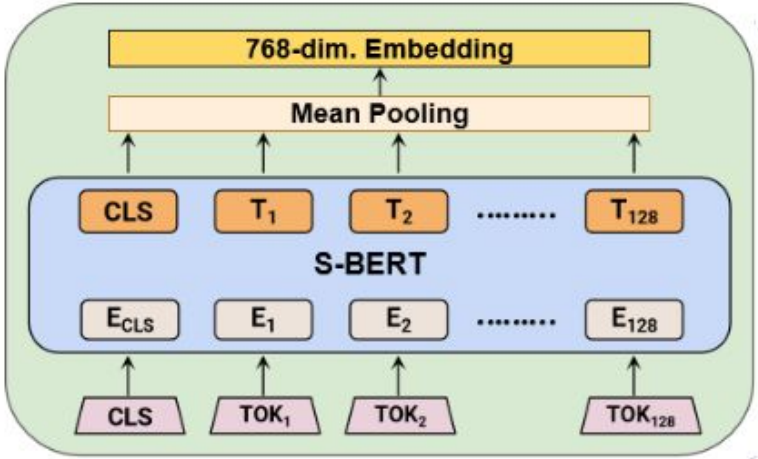
- We used the encoded embeddings from Sentence transformer (ST) model for each of the input text in a given pair.
- Using these embeddings, we evaluated the Cosine similarity score (between 0 to 1, from least to most)
- Renormalised to expected score (from 1 to 4,most to least).

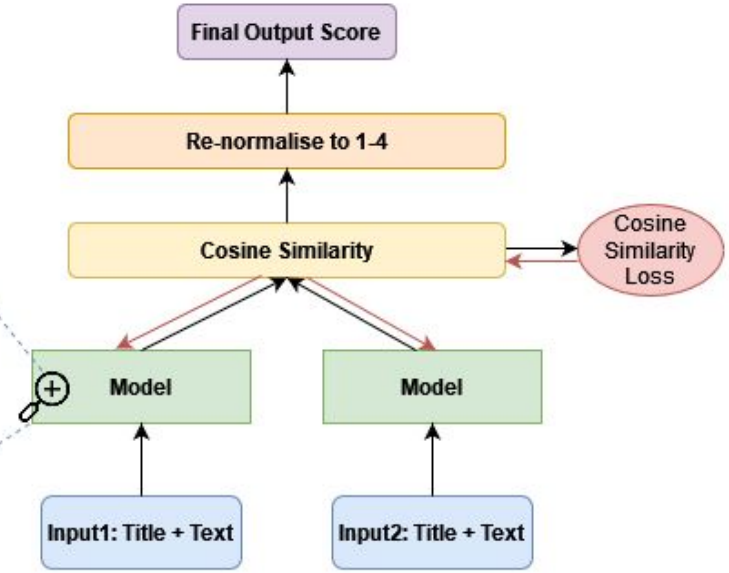
Strategies

2. Finetuning ST with Cosine Similarity Loss

- In this approach, we finetuned the ST models using the cosine similarity loss between cosine similarity evaluated using Embeddings (as described in Previous slide) and the normalised target similarity score.
- For normalisation, we converted the target score (gold-standard) from 1 to 4, to 1 to 0
- Once finetuned, we used the predicted cosine similarity score after renormalisation back to 1 to 4 as our Similarity score.
- We experimented our final approach in two different settings: Multilingual and Translated. For the translated setting, we translated all the text and titles in articles to English using Google Translator3.
- We tried our approach with SBERT4 and BERT models, including mBERT and XLM R

Model Architecture





(These input can either be in their original form or translated form.)

Results

Baseline Performances

Regression			
Approach Category	**Approach**	**Model**	**Pearson Correlation**
Baseline	–	mBERT	0.3089
		XLM-R	0.2852
Baseline	+ Validation	mBERT	0.3164
		XLM-R	0.3246
Baseline + Validation	+ Translation	mBERT	0.3371
		XLM-R	0.3045
Baseline + Validation	+ Title	mBERT	0.3187
		XLM-R	0.3345
Baseline + Losses + Validation	+ CosineEmbeddingLoss	mBERT	0.3346
		XLM-R	0.3336
Baseline + Losses + Validation	+ Cosine Similarity + MSE Loss	mBERT	**0.3621**
		XLM-R	**0.3454**
Sentence Transformer	+ Translation + Title + MSE Loss	S-BERT	0.2864

Table 2: Results with Regressor layer (Dropout layer followed by a feed forward layer) for final prediction as described in Section 4.3. Note that here we feeding the concatenated contexualized embeddings from Model to the regressor layer

Sentence Transformer Variants

Similarity		
Approach	**Model**	**Pearson Correlation**
ST Multilingual CS	S-BERT	0.5866
ST Translated CS	S-BERT	0.6058
ST Multilingual + "Title" CS	S-BERT	0.6778
ST Translated + "Title" CS	S-BERT	0.7086
**ST Multilingual + "Title" + CosSimLoss Finetuned CS	S-BERT	**0.7518**
**ST Translated + "Title" + CosSimLoss Finetuned CS	S-BERT	0.7492
ST(mBERT) + "Title" + Translation + CosSimLoss Finetuned CS	mBERT	0.6874
ST(XLM) +"Title" + Translation + CosSimLoss Finetuned CS	XLM-R	0.6357

Table 3: Results of the Consine Similarity based approach, where final score was evaluated using cosine similarity between embeddings and then scaling to 1 to 4. ** represents the models we used for our final submission on codalab as described in Section 3. Table also represent the ablation study for various elements included in our final approach along with our final approach experimented on mBERT in the ST framework. Note: **CS**: Cosine Similarity

Other Experiments

Approach	Model	Pearson Correlation
Baseline + Validation + Frozen Transformer Model	mBERT XLM-R	0.0376 0.0910
Concat(finetuned, non-finetuned embeds) + Regressor Layer	mBERT XML-R	0.3048 0.2962
Multi-task(Shared Parameter) + Top 3 Correlated params with Overall	mBERT XLM-R	0.3100 0.2556
Multi-task(Shared Parameter) + All 7 extra fields	mBERT XLM-R	0.2830 0.2662

Table 6: Other Experiments

Path Not Taken

Path not taken

- **Dropped Approaches:** We dropped certain approached in our final approach, as they didn't prove much useful in improving the scores of the task. This include, using Multi Task based approach, Summarisation, concatenation with non finetuned model embeddings and few other approaches. Some of the results from these approaches are added in the appendix section, to describe what didn't work out in our case.

- **Didn't try (Time constraint):** Because of time constraint, we were not able to try out some approaches, which might have been a good experiments here. This include Adverserial Finetuning, Hierarchical Modelling of Articles with LSTM adding additional analysis from text like common word count between monolingual pairs or in translated pairs. These approaches could be a great future work for our task, which we would definitely like to see, if they work out in this case or not.

Conclusion

Conclusion

- In this paper, we describe our approach to predict Similarity across Multilingual Article pairs provided in the SemEval Shared Task 8, leveraging the knowledge of pretrained Sentence Transformer Based Models.
- As the evaluation results indicate, our final approach clearly outperforms other approaches we experimented on. Furthermore, our ablation study further suggests the importance of various elements in our final approach.
- However, there is still a lot of scope of improvement in our approach discussed in previous section, We believe that experimenting on those approaches might prove pretty useful in evaluation the final scores and are definitely worth trying in future.

Thanks!

It's been a great learning experience!

Any questions?