NLP Term Project

Multilingual News Article Similarity

Group Name: MocktaiL

Ravi Karthik (18EC3AI19), Nitin Bisht (21MM61R08), Dharmana Jaswanth (18CS10014), Ananya Das (20EC10103)

April 15, 2022

1 Task Description

The task is to find out if a given pair of news articles cover the same story or not regardless of the political spin, tone or the style of writing. The pair of documents could be multilingual i.e.; one document could be in one language and other documents could be in another language. It is a document level similarity task in the domain of new articles and we needed to label the articles on the similarity scale of '0' to '4'. Here '0' means least similar and '4' means highly similar.

2 Individual Contributions

The sub-parts of the project were contingent on one another, none of the work can be done independently. Each one of us worked together to tackle a particular sub-problem and then moved on to the next task. The key sub-problems done by us are:

1. Scraping the data from the JSON files.
2. Cleaning the data obtained from the JSON files.
3. Searched, experimented and trained different models like BERT (Devlin et al., 2018) and distil-use base multilingual models (Reimers and Gurevych, 2019) to generate embeddings for our data.
4. Tried and experimented with various machine learning models and trained the model on the dataset to predict the similarity between two articles.
5. Created a generic feed-forward network to compare the predictions of it with the machine learning models.(Bebis and Georgiopoulos, 1994)

3 Approaches Used

3.1 Baseline Approach

Since our data was in JSON format we had to write a program in order to extract data from it. To do that we imported the modules **os, json** and used them. During this process, we made sure that the data is properly cleaned. i.e., to make sure that there is not any sample where there is no file corresponding to an id if there is no text data in some samples etc.

For implementing models, we used the **mBERT** model (Wolf et al., 2019). Here 'm' stands for multilingual. We need this multilingual model because our dataset contains multiple languages. More specifically we used the **'bert-base-multilingual-cased'** model. We can get this model from the **sentence_transformers** module offered by Hugging Face.

This module truncates the text to **512** tokens. It then encodes this into a **768**-feature dense embedding. Hence using this module, we can encode our text data into **768** feature dense embeddings. This encoding process took us around **50 min**.

Following this, we implemented various machine learning techniques to make predictions on the similarity between the 2 articles. If the prediction value is greater than 4 we truncated it to 4 and if it was less than 0 we truncated it to 0 (Because the value should be in between [0,4]).

3.2 Improved Approach

Initially in our baseline approach, we used **Random Forest Regressor** (*Sklearn Randomforest Classifier* n.d.) in order to make predictions using the features of the 2 news articles obtained from the pre-trained **mBERT** model. Now we used a **Gradient Boosting Regressor** (*Sklearn Gradient Boosting Regressor* n.d.) to make predictions.

The rationale behind using gradient boosting regressors is they can be more accurate than random forests as they're capable of capturing complex patterns in the data. Gradient boosting trees are trained one at a time, each one correcting the faults of the one before it. As a result, if we have less data that is not prone to overfitting, gradient boosting is always preferable to random forests (*Sklearn Randomforest Classifier* n.d.). We performed hyperparameter tuning and we came up with a good set of hyperparameters.

3.3 Final Approach

Here we completely changed the pre-trained model. Initially, we used **'bert-base- multilingual-cased'**. There were some issues in using this pre-trained model. Multilingual BERT is quite worse at mapping sentences of similar meaning to the same vector. It is noticed the performance is further reduced when different languages are mixed to calculate similarity. The reason behind this is Individual token vector values are predicted by models like mBERT, not sentences. Due to lexical variances in languages, this leads sentence aggregations to misalign the vectors. Means, out of the box, mBERT and other pre-trained multilingual transformers are not ideal for cross- language phrase similarity (Pires et al., 2019). Their output is finetuned to the specific language and when we compare words of different languages, they may diverge.

To overcome this problem, we decided to use the Distillation method (Hinton et al., 2015). It is based on a teacher-student setup. Here we force the multilingual transformers to produce the same vectors across the language. Figure 1 depicts how teacher-student learning is used in multilingual texts.

The teacher model predicts English phrases all of the time, but the student model is multilingual and is ”forced” to output vectors that look like English when given foreign language inputs. This means two things:

1. Across languages, vector spaces are aligned, which means that identical phrases in various languages are mapped to the same location.
2. The teacher model M vector space qualities are accepted and transferred to other languages in the original source language.

Here we are using **‘distiluse-base-multilingual-cased-v2’**. We can get this model from the **sentence_transformers** module offered by **Hugging Face**. This model maps sentences and paragraphs to a 512-dimensional dense vector space.

4 Metrics Used

We used the Pearson correlation coefficient as the metric to get the similarity score between two articles.

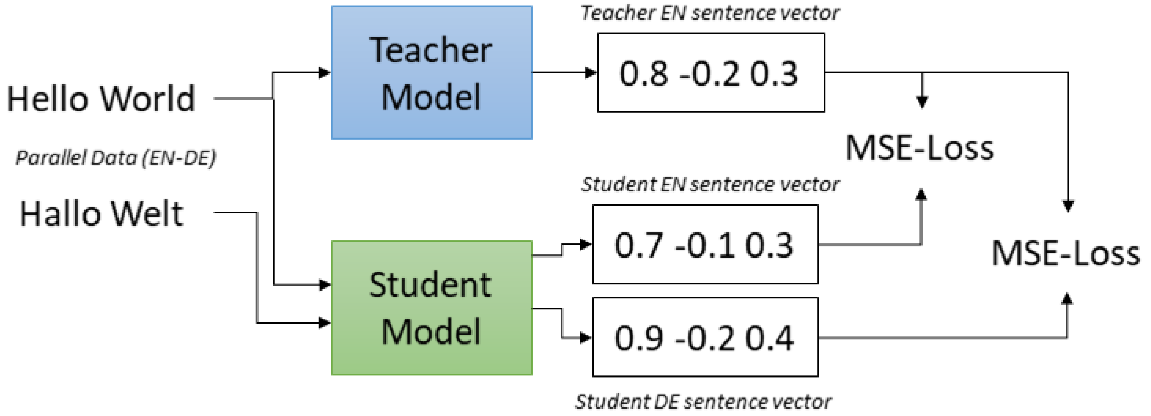

Figure 1: Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector (Reimers and Gurevych, 2020)

5 Experiments

1. For creating the embeddings, we began with Multilingual BERT (mBERT) (Wolf et al., 2019). However we found out that the output of mBERT is fine tuned to the specific language and when comparing words of different languages, their vectors diverge. Then we used a model based on distillation where we "force" the transformer to produce the same vector across languages by teacher-student setup. This gave substantial improvement over the baseline mBERT model.

2. We used different machine learning models to find the similarity score. We started with a Decision Tree regressor, then trained the Random Forest model. Then we experimented with different Gradient Boosting algorithms. Out of different boosting techniques experimented, Gradient Boosting Machine (GBM) gave best results compared to XGBoost and LightGBM (*Sklearn Gradient Boosting Regressor* n.d.) .

3. We tried various ways to combine the feature vectors of the news article pair into a single feature vector. Out of all the possibilities we found out that element wise multiplication is giving better results.

4. We also experimented with a deep neural network. But it did not give good results. The metrics corresponding to this deep neural network are shown in the Section 6.3. A possible reason can be lack of a large enough dataset which may have led to inadequate training and hence poor results.

6 Results

6.1 Results of Baselines

The baseline model based on distillation- Multilingual BERT with Random Forest Regressor gave an accuracy of **0.6057** on test dataset.

Architecture Used	Pearson Coefficient of Train Data	Pearson Coefficient of Test Data
Multilingual BERT with Cosine Similarity	0.27	0.2691
Multilingual BERT with Decision Tree Regressor	0.9993	0.3405
Multilingual BERT with Random Forest Regressor	0.9835	0.6057

Table 1: Comparison of baseline methods over train and test data in terms of Pearson Coefficient.

6.2 Results of Improved model

Initially the Pearson coefficient was **0.60** for our baseline model. We got an improvement of **0.04** giving a coefficient value of **0.64**.

Improved Architecture Used	Pearson Coefficient of Train Data	Pearson Coefficient of Test Data
Multilingual BERT with Gradient Boosting Regressor	0.99785	0.64147

Table 2: Performance of Improved model over train and test data in *terms of Pearson Coefficient.*

6.3 Results of Final model

New Architecture Used	Pearson Coefficient of Train Data	Pearson Coefficient of Test Data
Distiluse Multilingual Base with Random Forest Regressor	0.98729	0.73008
Distiluse Multilingual Base with Deep Neural Network	0.24933	0.00991

Table 3: Performance of Final model over train and test data in terms of Pearson Coefficient.

The final model based on distillation- Distiluse Multilingual Base with Random Forest Regressor gave an accuracy of **0.73008** on test dataset.

7 Difficulty Faced

1. Writing the code for data extraction as a part of Data Processing was a bit challenging as scraper was unable to distinguish between missing data and the data present.
2. While we were training the deep neural network the results were pretty poor. No matter what the hyperparameters are there is only a slight improvement on train data ,but not on test data.
3. Coming with right set of hyperparameters was a bit challenging while training various ML models.
4. It took some time in understanding various concepts related to transformers.

8 Path not taken

One of the models which we could have tried is Facebook's LASER-Language-Agnostic Sentence Representations (Artetxe and Schwenk, 2019). It was released by Facebook on Jan 22, 2019. It supports around 90 distinct languages written in 28 different alphabets. LASER does this by integrating all languages into a single common environment (rather than having a separate model for each).This approach is particularly useful for delivering numerous NLP features including text similarity, in a single language that can be quickly implemented in over 100 different languages without requiring further language knowledge.

It maps a sentence in any language to a point in a high-dimensional space with the goal that the same statement in any language will end up in the same neighborhood closely placed. This representation could be seen as a universal language in a semantic vector space.

Laser architecture is similar to neural machine translation in that it is based on an encoder/decoder system that employs a single common encoder for all input languages and a single shared decoder to create the output language. A five-layer bidirectional LSTM network serves as the encoder. It does not employ an attention mechanism and instead represents the input sentence using a 1,024-dimension fixed-size vector. It is generated via max-pooling over the BiLSTM's (Chiu and Nichols, 2015) final states, which allows for direct comparison of sentence representations and feeding them into a classifier.

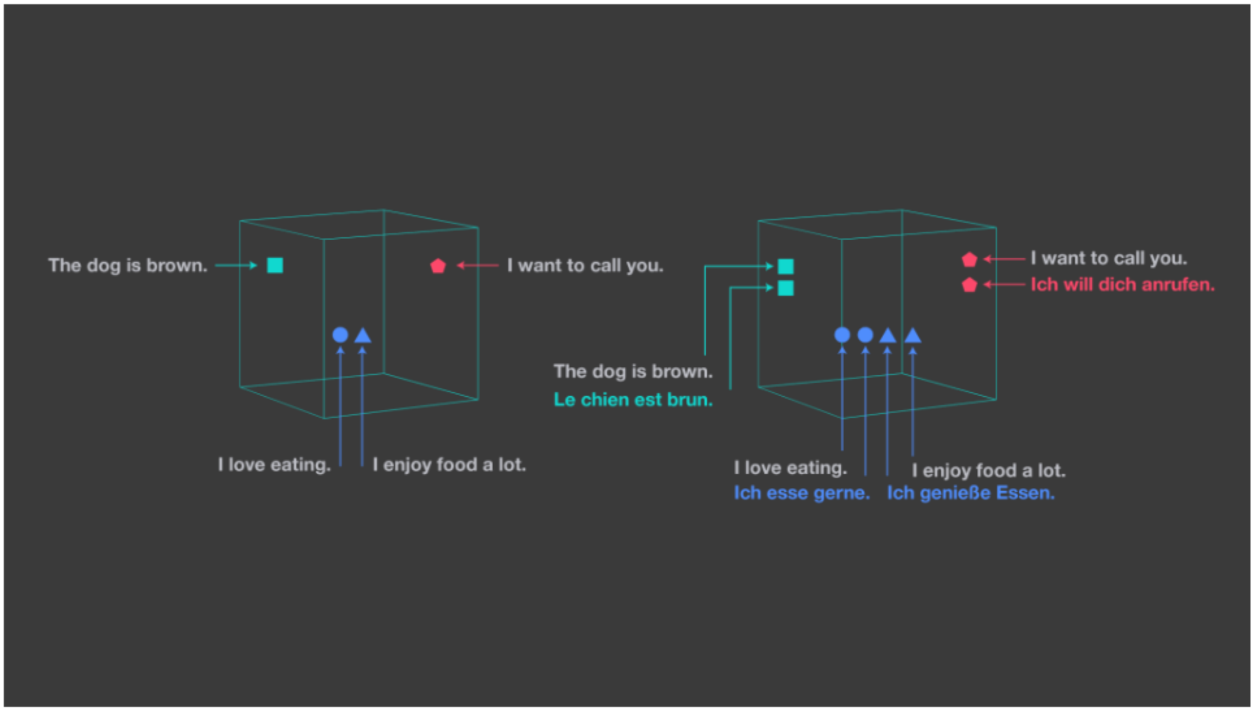

Figure 2: A depiction of LASER in higher dimensions (Schwenk and Schwenk, 2020)

9 Code

1. **Baseline Model**
2. **Improved Model**
3. **Final Model**
4. **Github Repository**

References

Artetxe M. and Schwenk H. (2019). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". *Transactions of the Association for Computational Linguistics* 7, pp. 597–610. DOI: 10.1162/tacl_a_00288. URL: https://doi.org/10.1162%2Ftacl_a_00288.

Bebis G. and Georgiopoulos M. (1994). "Feed-forward neural networks". *IEEE Potentials* 13.4, pp. 27–31. DOI: 10.1109/45.329294.

Chiu J. P. C. and Nichols E. (2015). *Named Entity Recognition with Bidirectional LSTM-CNNs.* DOI: 10.48550/ARXIV.1511.08308. URL: https://arxiv.org/abs/1511.08308.

Devlin J., Chang M.-W., Lee K., and Toutanova K. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805.*

Hinton G., Vinyals O., and Dean J. (2015). "Distilling the knowledge in a neural network (2015)". *arXiv preprint arXiv:1503.02531* 2.

Pires T., Schlinger E., and Garrette D. (2019). "How Multilingual is Multilingual BERT?" *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* DOI: 10.18653/v1/p19-1493.

Reimers N. and Gurevych I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. URL: http://arxiv.org/abs/1908.10084.

– (2020). "Making monolingual sentence embeddings multilingual using knowledge distillation". *arXiv preprint arXiv:2004.09813.*

Schwenk H. and Schwenk H. (2020). *Zero-shot transfer across 93 languages: Open-Sourcing Enhanced Laser Library.* URL: https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/.

Sklearn Gradient Boosting Regressor (n.d.). URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

Sklearn Randomforest Classifier (n.d.). URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

STS Benchmark (2012). URL: https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark#STS_benchmark_dataset_and_companion_dataset.

Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., Platen P. von, Ma C., Jernite Y., Plu J., Xu C., Scao T. L., Gugger S., Drame M., Lhoest Q., and Rush A. M. (2019). *HuggingFace's Transformers: State-of-the-art Natural Language Processing.* DOI: 10.48550/ARXIV.1910.03771. URL: https://arxiv.org/abs/1910.03771.