NLP Project (2nd Submission)

Multilingual News Article Similarity

Group Name: MocktaiL

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

March 30, 2022

Improved Approach

Initially, in our baseline approach we used **Random Forest Regressor** to make predictions using the features of the 2 news articles obtained from pre-trained **mBERT** model.Now we are using a **Gradient Boosting Regressor** to make predictions. We performed hyperparameter tuning and we came up with a good set of hyperparameters.

Results of Baselines

Methods Used	Pearson Coefficient of Training Data	Pearson Coefficient of Testing Data
Cosine Similarity	0.27	0.2691
Decision Tree Regressor	0.9993	0.3405
Random Forest Regressor	0.9835	0.6057

*Table 1: Comparison of baseline methods over training and test data (*STS Benchmark *2012) in terms of Pearson Coefficient.*

Results of Improved model

New ML Model Used	Pearson Coefficient of Training Data	Pearson Coefficient of Testing Data
Gradient Boosting Regressor	0.99785	0.64147

*Table 2: Performance of Improved model over training and test data (*STS Benchmark *2012) in terms of Pearson Coefficient.*

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**.

Further Improvements

1. mBERT is bad at mapping sentences with sentences of similar meaning to the same vector , and the performance drops even further when we mix different input languages

to calculate similarity . Models like mBERT predict vector values of individual tokens and not sentences. This causes sentence aggregations to misalign the vectors due to lexical differences in languages. This means mBERT and other pre- trained multilingual transformers are not suitable for cross-language sentence similarity .Because of which the output of the mBERT is fine-tuned to the specific language and when comparing words of different languages, their vectors are diverged.Pires et al., 2019.So we can improve our performance by using a better models like LASER, DISTILUSE-BASE-MULTILINGUAL-CASED.(Liu et al., 2019)

2. The Pearson correlation values have reached a saturation and is showing no major improvements on changing models. There was no further major change using traditional Ensemble machine learning methods. To overcome the issue we will try using deep learning models (Liu et al., 2019) instead of standard ensemble machine learning methods and compare our results.

Link to Model

https://colab.research.google.com/drive/1x6bwFJ9Pi0anT9wyuSrKkh8DioPCiMgI?usp=sharing

NOTE: Only the code for the improved model is shown here.Other code remains same.

References

Liu X., He P., Chen W., and Gao J. (2019). "Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding". *arXiv:1904.09482 [cs]*. URL: https://arxiv.org/abs/1904.09482.

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.

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