MultiLingual News Similarity

Rudra Dhar, Padigala Lakshman Sai
Srijith Padakanti, Yash Sharma
Mentor: Sagar Joshi
Team INFO

Major project for Information Retrival and Extraction course

Abstract

This report describes our Major project for the course IRE. Here we made a system to determine multilingual news article similarity. Specifically, we tackled task 8 of SemEval-2022. The task of multilingual news article similarity entails determining the degree of similarity of a given pair of news articles in a language-agnostic setting. We used a Siamese Architecture, as it's a well-known method for text comparison tasks. We used multilingual encoder representations from our models, specifically, distilbert-base-multilingual and xlm-roberta-base. We performed various experiments and found out that the model which uses distilbert-base-multilingual as encoder, and title, main text, and metadata of the news as inputs, performed the best. It achieved an MSE of 1.3885 and a PCC of 0.2973.

1 Introduction

Numerous media outlets publish thousands of new news articles every day. Understanding which articles refer to the same topic not only improves applications such as news aggregation, but also allows for cross-linguistic research of media consumption and attention. Due to the various ways in which a story might differ, such as when two pieces have a lot of textual overlap but depict the same events that occurred years apart, or when there is very little textual overlap but the news stories talk about the same event, determining how similar two news articles are can be difficult. Therefore we have chosen to determine the similarity between news articles in multilingual setting. Perticualrly we have attempted *SemEval 2022 Task 8: Multilingual News Article Similarity* (2022) organised by Chen et al. (2022) . Here they have more than 7000 news article pairs in 18 different languages which we had to score based on there similarity. We referred papers of Joshi, Taunk, and Varma (2022) and Xu, Yang, Cui, and Chen (2022) and came up with a baseline and several improved models. The model which uses distilbert-base-multilingual as encoder, and title, main text, and metadata of the news as inputs, performed the best.

We have made our code available on GitHub.[1]

2 Task and Dataset description

2.1 Task

This is a Sem-Eval 2022 task. In this task the input will be a pair of newspaper articles in the same or different languages and, the similarity had to be determined as to whether the news articles are of the same topic i.e. to calculate the similarity scores between the news articles on a 4-point scale from least to most similar.

2.2 Dataset

For the training dataset, we are given the language of the news articles, a pair id for the news articles, the corresponding links of the news articles to extract the data, & 7 types of similarity scores based on the different information that can be extracted from the data(ex: tone of the articles, location where the incident took place, the time when the incident happened). Then there is the overall similarity score which is based on all the factors in the article. The training dataset contains the language pairs ar-ar, ar-en, de-de, de-en, en-en, en-es, en-fr, en-pl, es-es, fr-fr, pl-pl, tr-tr and the size of the dataset is 4918 out

[1] https://github.com/sharma18yash/ire_project

of which only 2857 were with the valid links from which the data could be extracted. The test dataset contained the language pairs ar-ar, de-de, de-en, en-en, en-es, en-pl, es-es, fr-fr, pl-pl, and tr-tr, and in addition, the language pairs de-fr, de-pl, es-it, fr-pl, it-it, ru-ru, zh-en, zh-zh which are not present in the training dataset. The size of the test dataset is 4902 out of which 4316 were with valid links from which the data could be extracted. With the links given in the data, the information was extracted, and then it was subsequently cleaned so that it can be used for training and testing purposes.

3 System Description

From the reference cited, we used Siamese architecture for pair of news articles in generating encoding. A linear layer is then applied to the representation. These representations are then combined to create a single representation of the two articles before being sent through fully connected layers to provide the similarity score, which ranges from 1 to 4. If x1 and x2 are the encoding representation, then in the aggregation step we used different strategies like x1-x2 and x1+x2 and other combinations, then passed to a next fully connected layer to compute the similarity score. We used MSE as loss function. We saved model checkpoints at every quarter epoch, and used the model instance which had highest PCC in validation set. Then on this architecture we performed various experiments by passing the data with different attributes.

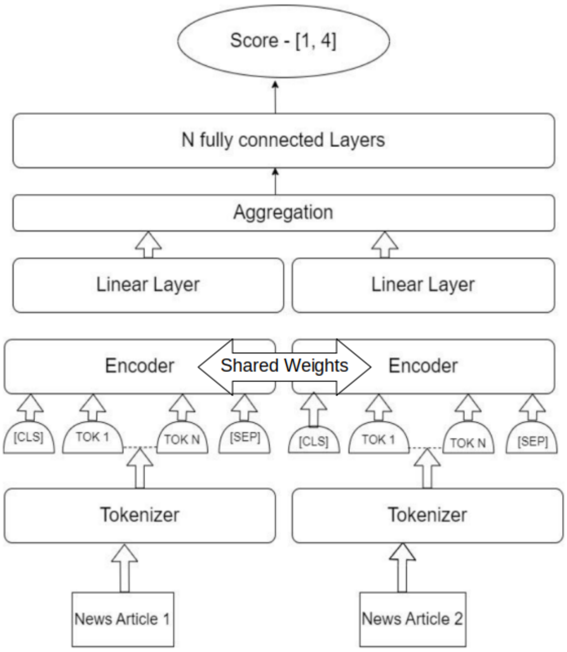

Figure 1: Main architecture for the experiments conducted

3.1 Attributes passed to the encoder in different experiments:

- News text: The plain text extracted from the article is used as input. The length of text in the dataset varies from 205 to 9068 with average text length with 472 words.
- Metadata: In the news article, we can able to extract data like title of the article, tags describing the news story and a short description for some of the article is found.

3.2 Base Encoder Model:

1. **XLM-RoBERTa:** This model, which is built on a transformer-encoder, was pretrained on 100 languages using masked language modelling objectives. As a result, it significantly improved on a variety of cross-lingual comprehension tasks.
2. **Multilingual DistilBERT:** The base encoder additionally employed a condensed version of the multilingual cased BERT-base model trained on Wikipedia data from 104 languages. Despite having

somewhat lower performance than the original model, the distilled version was chosen because it had a smaller number of parameters, which is typically a good fit for low-data scenarios.

3.3 Data Augmentation:

From the links provided, we were able to extract only 2857 out of 4918 pairs in train set and in test set of 4902 pairs, we were able to extract only 4316 pairs. Here the train data size is less when compared to test data, so we created synthetic data by randomly shuffling the text sentences in the news article in other 4 permutations and increased the train data size by 4 times to 14285 pairs.

3.4 Multilable Learning

The Training data had extra labels like geography, time, which we didn't need to predict. But these labels can be used for training. So we designed a experiment where we trained with 7 labels, but tested only with the OVERALL score as given in the task.

3.5 New Split

After extracting the data from the links provided in the task , the size of train data of 2857 pairs and test data of 4902 pairs. To make good split between train and test, we combined the total extracted pairs in to 7759 pairs and then split into 70-15-15 proportion into train dataset, validation dataset and test dataset.

4 Experiments and Results

4.1 Evaluation Criteria

Since the task is regression, Metrics used are

- **PCC** The Pearson Correlation Coefficient measure the linear relationship between two variables (test and pred). It varies between -1 and +1
  $\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$
- **MAPE** Mean absolute percentage error
  $\frac{100\%}{n} sum_{t=1}^{n} |\frac{A_t - F_t}{A_t}|$
- **MSE** Mean Squared Error

4.2 Experiments Performed:

1. **DB :** Multilingual DistilBERT Base as encoder
2. **XRB :** XLM-RoBERTa Base as encoder
3. **DB_M :** Multilingual DistilBERT Base with Title and Metadata
4. **DB_A :** Multilingual DistilBERT Base with Augmented Data
5. **DB_ML :** Multilingual DistilBERT Base with Multilable Learning
6. **DB_S :** Multilingual DistilBERT Base with new split of train and test data pairs

4.3 Results:

Table 1: Results of all the experiments performed on validation data and test data

EXPERIMENT	DEV. PCC	TEST MSE	TEST MAPE	TEST PCC
DB	0.4257	1.3172	0.5575	0.2817
XRB	0.4278	1.3942	0.6774	0.2958
DB_M	0.4705	1.3885	0.5161	0.2973
DB_A	0.9608	1.5156	0.575	0.2458
DB_ML	0.4310	1.4209	0.5557	0.2389
DB_S	0.449	1.36	0.44	0.405

5 Conclusion

We used a variety of techniques to determine the similarity between news articles in multilingual setting as we attemped task-8 of SemEval-2022. We used textual content, meta data, news title to train our model. We tried different embeddings like multilingual distilBERT and XLM-RoBERTa¿ we also tried multilabel learning to increase performance. We got the best result using DistilBERT along with meta data to train our model which achieved a test PCC of 0.2973.

References

Chen, X., Zeynali, A., Camargo, C., Flöck, F., Gaffney, D., Grabowicz, P., . . . Samory, M. (2022, July). SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)* (pp. 1094–1106). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.semeval-1.155 DOI: 10.18653/v1/2022.semeval-1.155

Joshi, S., Taunk, D., & Varma, V. (2022, July). IIIT-MLNS at SemEval-2022 task 8: Siamese architecture for modeling multilingual news similarity. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)* (pp. 1145–1150). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.semeval-1.161 DOI: 10.18653/v1/2022.semeval-1.161

Semeval 2022 task 8: Multilingual news article similarity. (2022). Retrieved from https://competitions.codalab.org/competitions/33835

Xu, Z., Yang, Z., Cui, Y., & Chen, Z. (2022, July). HFL at SemEval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity. In *Proceedings of the 16th international workshop on semantic evaluation (semeval-2022)* (pp. 1114–1120). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.semeval-1.157 DOI: 10.18653/v1/2022.semeval-1.157