DartmouthCS at SemEval-2022 Task 8: Predicting Multilingual News Article Similarity with Meta-Information and Translation

Joseph Hajjar[1], Weicheng Ma[2], and Soroush Vosoughi[3]

[1,2,3]Department of Computer Science, Dartmouth College
[1]joseph.h.hajjar.gr@dartmouth.edu
[2]weicheng.ma.gr@dartmouth.edu
[3]soroush.vosoughi@dartmouth.edu

Abstract

This paper presents our approach for tackling SemEval-2022 Task 8: Multilingual News Article Similarity. Our experiments show that even by using multi-lingual pre-trained language models (LMs), translating the text into the same language yields the best evaluation performance. We also find that stylometric features of the text and meta-information of the news articles can be predicted based on the text with low error rates, and these predictions could be used to improve the predictions of the overall similarity scores. These findings suggest substantial correlations between authorship information and topical similarity estimation, which sheds light on future stylometric and topic modeling research.

1 Introduction

Given a pair of news articles in arbitrary languages, the objective of SemEval-2022 Task 8 is to predict whether the two articles cover the same story (Chen et al., 2022). While the task falls in the category of topical similarity estimation, traditional statistical topic models may not be appropriate due to the large vocabulary size and the difficulty of matching representative "topic words" across languages. For example, determining whether two articles are about the same news story requires answering questions such as what the article is talking about, who is involved, and where did it occur. We demonstrate this by applying an LDA model (Blei et al., 2003) on the original dataset, and using the LDA representations of news articles in a logistic regression (LR) model yields a Pearson correlation of 0.193 on the evaluation dataset.

Due to the difficulty of obtaining gold standard topical similarity annotations, prior research on this topic is mostly conducted under an unsupervised setting. For example, Bisandu et al. (2018) model text using n-gram features and cluster articles using an improved square root cosine similarity measure. Singh and Singh (2021) use the similarity of news headlines to approximate similarities of news articles. However, their approaches rely primarily on simple statistical features on the word level, which is not proper for our task. Rupnik et al. (2016) use an external event list as an alias of "topic words" with greater granularity to solve the topical similarity estimation task. While this idea could be well adapted to our task, engineering the list of events is effort-consuming and lacks generalizability to unseen data, since the news articles may not always talk about a singular, distinguishable event.

Our models are based on the pre-trained multilingual BERT (mBERT) model (Devlin et al., 2018), one of the most competitive models on multi-lingual natural language processing (NLP) tasks. Though mBERT could handle all the languages in the challenge dataset, we find that translating both news articles in each instance into the same language improves the performance of the model by 0.024 in Pearson correlation coefficients. This is counter-intuitive, since translating the text into another language unavoidably harms its syntactic quality. We hypothesize that the improvements result from the match of named entities across documents when they are translated into the same language.

In addition to text translation, we find that the intermediate scores of each instance that are related to stylometric features of the text and meta-information of the news articles could be predicted by the an mBERT model with high Pearson correlation scores (0.71 to 0.88). With a simple LR model, the six predicted intermediate scores (i.e., geography, entities, time, narrative, styles, and tone) can be aggregated to make more accurate predictions of the overall similarity scores. We speculate from this result that stylometric features are important for aligning the events and named entities across

July 14-15, 2022

news articles, while the meta-information of these articles provides hints about whether the two articles are talking about completely irrelevant events. We additionally try injecting the predicted intermediate scores into the text encodings produced by mBERT. However, the performance of the ensemble model drops below the pure mBERT-based approach. This is expected since the predicted scores are not dimensionally compatible with text encodings. We leave for future research the problem of efficiently boosting the text encodings with stylometric features and meta information predicted from the text.

Our approach ranked the 10th in the all-language data challenge and the 6th most successful results in the English-only challenge. The main areas which impacted our model's performance were the languages and the lengths of the articles. Shorter articles and certain languages proved to confuse the model.

2 Dataset

Each sample in the training data for the task included a url and language code for each article and a score for the pair's overall similarity, along with a score for each of six stylistic features and news meta-information (i.e., geography, time, shared entities, shared narratives, style, and tone). The feature scores and overall similarity were calculated by averaging several annotators' scores and were rated on a scale of 1 to 4, where 1 indicated low similarity and 4 implied high similarity. Each pair of news articles is annotated by 1 to 8 annotators, with the majority of instances annotated by 1 to 3 annotators. In carrying out experiments, we split the released training data into two sets, 80% for training and the remaining 20% for validation. The evaluation dataset is left out for testing purposes only.

2.1 Scraping Article Contents

Since the dataset provides links to news articles, we scrape the text using the newspaper3k library [1]. Two links were provided for each article, one pointing to the article's most up to date link and another pointing to an Internet Archive site containing the earliest available version of the article. The analysis was meant to be carried out on the content of the earliest version of the article, however if that link was unable to be scraped, then the most up to date version of the article was used as the article's content. It is noteworthy that in 854 training instances, at least one news article cannot be retrieved from either link. Removing these data which were not usable resulted in a training dataset containing 7049 instances. Similarly, in the evaluation data, there were 31 instances where at least one news article could not be retrieved from either link.

[1] https://github.com/codelucas/newspaper

2.2 Training Data and Evaluation Data

Training data were released in two batches, where the first batch contained 2,939 pairs of news articles and the second batch contained 4,964 pairs of news articles.

A key difference between the first and second training datasets was that the second dataset included 577 pairs with articles in different languages (we refer to these as bilingual pairs in the rest of the paper) whereas in the first dataset, articles were paired with other articles in the same language. Additionally, the evaluation data had 1440 bilingual pairs, and the bilingual pairs in question had more variety in the languages appearing in the pairs.

The languages used in the training data were, in order of most to least prevalent: English, German, Spanish, Polish, Turkish, French, and Arabic. The only bilingual pairs in the training data were a set of article pairs where one article was in German and the other was in English. The evaluation data had the same languages as the training data, along with Russian, Italian, and Chinese (zh: ISO 639-1).

3 Experimental Setup

All our models are implemented based on the Multi-Task Deep Neural Networks for Natural Language Understanding (MT-DNN) (Liu et al., 2019) model. Specifically, we use the pre-trained mBERT model with 12 layers and 12 attention heads on each layer. For data pre-processing, we separately truncate or pad the two news articles in each instance to 512 words, join them with a "[SEP]" token, and tokenize the text with the mBERT tokenizer. The regression head of MT-DNN is applied on top of the last-layer output of mBERT to generate predictions.

For each approach we present, we fine-tune the mBERT model for 5 epochs with early stopping. A batch size of 8 and a learning rate of 5×10^{-5} are used for all the settings. These are default hyperparameters from the MT-DNN library. Additionally, we implement a bidirectional LSTM (BiLSTM)

(Hochreiter and Schmidhuber, 1997) model as our baseline, using exclusively the text of the articles as input. We set the dimension of hidden states to 100 for this model, and we train the model for 20 epochs with a learning rate of 0.001 and a batch size of 512 in all the experiments. Our BiLSTM model achieves a Pearson Correlation of 0.277 on the validation set.

We repeat all the evaluations three times with different random seeds (i.e., 0, 1, and 2), and we report the mean Pearson correlation coefficient and the standard deviation of scores. All the experiments are run on a single RTX-6000 graphics card.

3.1 Translating Articles

As mentioned in Section 2, there were a substantial number of bilingual article pairs in the training and particularly in the evaluation data. Since some of our approaches require translating both articles in each instance into the same language, we adopt googletrans[2] to carry out the translations. We experimented with settings where the articles are translated (1) into English, (2) into the language of the first article, or (3) into the language of the second article, and in the evaluations we used the setting under which each model performs the best on the validation set.

3.2 Approaches

Figure 1 visualizes the three approaches we tried in the challenge.

3.2.1 LR-Based Feature Ensemble Model

In our best-performing model, the mBERT model is used to encode the text and predict six intermediate scores provided in the training dataset. An LR model with the intermediate scores as input is then used to predict the overall similarity scores. The scikit-learn (Pedregosa et al., 2011) implementation of the LR model is used in our model.

3.2.2 Text-Based Regression Model

For this approach, we rely purely on the text encoding ability of mBERT to make topical similarity predictions. Under this setting, we remove all the meta-information or stylistic clues of the documents from the input and directly fine-tune the vanilla mBERT model to predict the overall similarity scores.

[2]https://github.com/ssut/py-googletrans

3.2.3 Feature-Injected mBERT Model

Since the meta-information and stylistic features of the news articles have proven to be useful in Section 3.2.1, we attempt to combine these intermediate predictions with text encodings produced by mBERT. Specifically, in each instance, we predict the six intermediate scores and concatenate these to the last-layer hidden state of mBERT to make predictions on feature-enhanced text representations. Other parts of the mBERT model remain unchanged. Different from the ensemble model, the feature-injected model is end-to-end.

4 Results and Analysis

The efficacy of each approach was judged on the Pearson Correlation between the predicted similarity and the ground truths (Freedman et al., 2007). With this metric, scores closer to 1 and -1 imply the strongest positive and negative correlations possible, and scores close to 0 imply poor to no correlation.

A summary of the scores can be found in Figure 2. The three approaches worked most successfully when used in tandem with translations. Here we focus on the best performing translation setting for each of the three approaches described in Section 3.2 and compare it to the same approach applied to non-translated data.

4.1 Impact of Language Unification

For our LR-based model, the best performing setting was translating the bilingual pairs in the training data and the evaluation data to the first language of the pair. As discussed in Section 3, we ran this setting with three different seeds, and the approach yielded a mean of 0.746 with a standard deviation of 0.00210. This represented an increase in the correlation by 0.024 when compared to the same setting with no translation.

The next best performing approach was our text-based regression model, with translations to the first language in the bilingual pair. The vanilla mBERT model yielded a mean correlation of 0.737 and a standard deviation of 0.00398. Translation in this setting proved to increase the correlation of the scores by 0.020.

Lastly, the injected features approach performed best when the bilingual pairs were translated to English, and yielded a mean score of 0.737 with a standard deviation of 0.00642. Although this approach scored the lowest among the three, trans-

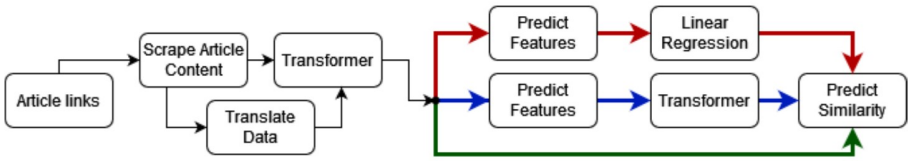

Figure 1: A visualization of three different models we explore in the challenge. The red line shows the approach outlined in Section 3.2.1, the green line shows the approach outlined in Section 3.2.2 and the blue line shows the approach outlined in Section 3.2.3

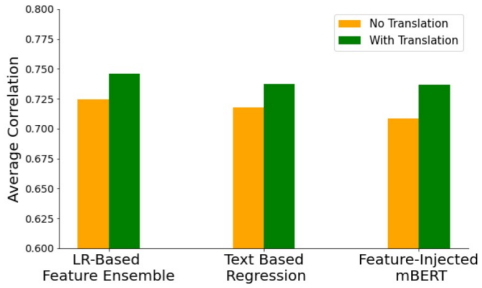

Figure 2: Pearson correlation coefficients achieved by our models on the evaluation dataset. For each model, the green bar represents the score achieved when both documents in each instance are translated to the same language.

Langs	Overall (%)	Outliers (%)	Short (%)
ar - ar	6.08	15.85	18.07
zh - zh	15.69	22.95	48.27
ru - ru	5.85	11.47	4.70
pl - pl	4.57	7.10	1.65

Table 1: Percentage of same-language instances in the entire evaluation data (Overall), in the subset of evaluation data where our best model performs at least 1.5 points off the gold standard labels (Outliers), and in the subset of evaluation data where at least one article in a pair is at most 80 characters in length (Short).

lation proved to be the most useful when applied in this setting, with an increase of 0.29 when compared to injecting the features without translation.

Due to the nature of the task, i.e., determining whether articles speak about the same event, it is reasonable to infer that translating articles into the same language boosted performance because of the alignment of named entities in the same language. Although mBERT can manage all of the languages used in the dataset, named entities may be harder to represent between languages, and given their significance in determining article similarity, it is likely that their translation into the same language was the operative factor in enhancing performance.

4.2 Error Analysis for the LR-Based Feature Ensemble Model

When examining the samples where our LR-based model's predictions were furthest from the ground truth annotations, the length in the pair of articles proved a significant factor. The shorter the length of the articles, the more difficult it was for the model to produce accurate predictions, and conversely, samples with longer article lengths tended to yield higher correlation. The shorter articles on which the model was failing tended to be articles where we had only been able to scrape the headline of the article.

For example, our model incorrectly predicted that the following two articles were dissimilar: (1) "释放激发制造业活力稳定经济增长-商会频道-长城网" and (2) "稳外贸工作座谈会释放利好一揽子新举措待发–财经–人民网" (translated: "Release to stimulate the vitality of the manufacturing industry to stabilize economic growth - Chamber of Commerce Channel - Great Wall Network", "Symposium on Stabilizing Foreign Trade Unleashes Favorable Package of New Measures to Be Launched"). In this case, the articles are lacking in substance to investigate their similarity. Instead, the model has to rely on the articles' headlines which contain far less information.

Another likely influence on the model's prediction was the language on which it was predicting. Although our approach leveraged multilingual models, there were certain languages which tended to yield poor outcomes. The most notable were Chinese-Chinese, Arabic-Arabic, Polish-Polish, and Russian-Russian pairs. Some of these could be explained by the prevalence of certain language pairs in the subset of evaluation data where article lengths were particularly short. This is detailed in Table 1. Thus, it is difficult to determine whether the length of the article or the language is the operative factor in the poor predictions. It

Instances	LEN-A1	LEN-A2
Good preds	1427	1354
Poor preds	394	507

Table 2: The median number of characters in Article 1 (LEN-A1) and Article 2 (LEN-A2) for each instance of the evaluation datasets where the error margin achieved by our best model is small (Good preds) or large (Poor preds). We regard the instances where the predictions are within an error margin of 0.2 points as good predictions and those whose error margins are above 1.5 points as poor predictions.

is important to note, however, that in the case of articles written in Chinese, each character encodes substantially more information than a character in the Latin, Arabic, or Cyrillic alphabets. Nevertheless, it is noteworthy that two of the four languages which confused the model were those which it had not seen in the training data.

4.3 Inter-Approach Error Analysis

Our LR-Based model tended to outperform the text-based regression and feature-injected mBERT models when the article pairs spoke about the same news story but with different entities as subjects. For example, in one pair of articles, both articles referred to tableau proposals for India's Republic Day celebration being rejected, however one speaks about the proposal from the Maharashtra and the other about the West Bengal government proposal. Due to the articles being about a similar event, they contain similar phrases, such as "Twenty-two proposals, 16 from states and union territories and six from central ministries — out of a total 56 have been short-listed for this Republic Day parade" and "the Ministry of Defence has selected 22 tableaux out of 56 proposals for the Republic Day parade." This pairing was given a similarity score of 1.5; our LR-based model predicted 1.48, our text-based regression model 2.42, and our feature-injected mBERT 2.67. Because the latter two models rely more heavily on textual data than the former, the articles' similar phrases likely inflated their scores while the LR-based model was more greatly influenced by the pair's features. In this case, both articles were written in English.

Conversely, our LR-based model underperformed when the textual similarity correlated more with the score than the features did. In one case, one article refers to Austrian MPs who are sick with COVID-19 and the other compares Austria and the EU's approach to dealing with MPs who are sick with COVID-19. Similar phrases and words appear in both articles, such as the pair "That will depend above all on whether the government submits corresponding requests to Parliament", "Now the President wants to discuss Parliament's timetable", and the pair "how MPs who have tested positive could also take part in votes", "Ten MPs of the ÖVP ... are in self-quarantine." This pairing was scored 3.0 for overall similarity, however the feature scores for this pair are 1.0, 2.0, 1.0, 4.0, 1.0, 1.0 for geography, entities, time, narrative, style, and tone, respectively. Our LR-based model predicted 1.95, our text-based regression 2.63, and our feature-injected mBERT 2.87. It is reasonable to conclude that the overwhelmingly low features scores contributed to the large difference in our LR-based model's prediction and the annotated score, while the textual similarity of the pair improved the scores for the latter two models.

5 Conclusion and Future Work

This paper describes the model we use for tackling SemEval-2022 Task 8: Multilingual News Article Similarity. Our work shows that while the vanilla mBERT model could greatly outperform shallower baseline models (e.g., BiLSTM and LDA) on this task, introducing intermediate prediction objectives (e.g., stylometric features and meta-information of news articles) helps improve the performance of mBERT noticeably. Additionally, we find that unifying the languages of news articles in each training and evaluation instances has a positive effect on the ensemble model's performance. We speculate that translating both articles in an instance into the same language helps the model align similar named entities across articles, which is important for assessing whether a pair of articles focus on the same set of events or entities.

Since it is shown by our experiments that stylometric features are contributive for topical similarity estimations, future work can extend our ensemble model by involving a richer list of writing-style-related linguistic features.

References

Desmond Bala Bisandu, Rajesh Prasad, and Musa Muhammad Liman. 2018. Clustering news articles using efficient similarity measure and n-grams. *International Journal of Knowledge Engineering and Data Mining*, 5(4):333–348.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Xi Chen, Ali Zeynali, Chico Q. Camargo, Fabian Flock, Devin Gaffney, Przemyslaw A. Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 Task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

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.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jan Rupnik, Andrej Muhic, Gregor Leban, Primoz Skraba, Blaz Fortuna, and Marko Grobelnik. 2016. News across languages-cross-lingual document similarity and event tracking. *Journal of Artificial Intelligence Research*, 55:283–316.

Ritika Singh and Satwinder Singh. 2021. Text similarity measures in news articles by vector space model using nlp. *Journal of The Institution of Engineers (India): Series B*, 102(2):329–338.