IIIT-MLNS at SemEval-2022 Task 8: Siamese Architecture for Modeling Multilingual News Similarity

Sagar Joshi*, **Dhaval Taunk***, **Vasudeva Varma**
IIIT Hyderabad, India
{sagar.joshi, dhaval.taunk}@research.iiit.ac.in
vv@iiit.ac.in

Abstract

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. This task aims to determine the extent to which the articles deal with the entities and events in question without much consideration of the subjective aspects of the discourse. Considering the superior representations being given by these models as validated on other tasks in NLP across an array of high and low-resource languages and this task not having any restricted set of languages to focus on, we adopted using the encoder representations from these models as our choice throughout our experiments. For modeling the similarity task by using the representations given by these models, a Siamese architecture was used as the underlying architecture. In experimentation, we investigated on several fronts including features passed to the encoder model, data augmentation and ensembling among our major experiments. We found data augmentation to be the most effective working strategy among our experiments.

1 Introduction

News articles from the web covering the same event tend to differ on regional and political biases, style of writing, conciseness and preciseness of coverage and the intended audience of the news outlet. Misleading and confusing articles might lead to unnecessary confusion, chaos and tensions potentially exacerbating or even creating non-existent issues. Often, news articles covering the same event in different languages from varied regional sources are readily available. This can apply for happenings of local, regional as well as international significance. In such a scenario, it is pertinent to able to identify or cluster together news articles covering the same story in different languages and different view points while also being able to distinguish between articles similar in style but covering different happenings.

The significance of the problem of multilingual news article similarity (Chen et al., 2022) lies towards the applicability in this area by scoring a pair of news articles based on their similarity in coverage of the event without focusing on the subjectivity of the content.

In our approach, we model the task as a regression problem by making use of a Siamese neural network (Bromley et al., 1993) as the base architecture. We perform experiments in feature engineering by making use of metadata such as title, description, keywords and tags in addition to the news text. We also experimented with artificially augmenting the data by document permutation. Also, we tried with ensembling of two models - one trained using only textual features, the other using metadata information.

While data augmentation strategy did give a decent improvement in performance, considering the length of the news articles and the coverage of a variety of entities or events throughout the article, globally modeling the text representation might not be the best way to model for this task. In the subsequent sections, we elaborate on our experiments performed and the results obtained and analyze the same. We have open-sourced our code[1] for ease of replicability and improvement over our techniques.

2 Task Definition

The task of multilingual news similarity consists of predicting the similarity score between two news articles on a scale of 1 to 4, with a higher score indicating a higher degree of similarity. The news articles may belong to either the same or different

*These authors contributed equally to this work.

[1] https://github.com/sagarsj42/multilingual-news-article-similarity

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1145 - 1150

Lang	#Train
ar_ar	274
de_de	857
de_en	577
en_en	1800
es_es	570
fr_fr	72
pl_pl	349
tr_tr	465

Table 1: Stats of different languages available in train set

languages, with there being no specified set of languages in which the articles might be written in. Most of the pair of news articles were annotated by 1-3 annotators, with the maximum no. of annotations per data sample being 8. In case of multiple annotations per sample, the scores were averaged.

Statistics of the data created for the task are shown in table 1 and 2. The training and evaluation (test) splits were prepared such that the languages appearing in either of the sets might not occur in the other, which shows the necessity of a good multilingual representation for modeling a solution. The news articles were downloaded from the provided URLs following which the dataset was prepared for the task. There were 18 and 22 articles not accessible by the provided URLs in the train and test sets respectively, which were replaced by a placeholder dummy text. 10 % of the training data was used as the validation set. The model with the best performance on the validation split was evaluated on the test data.

3 System Description

Figure 1 shows the architecture we adopted for modeling the task. In the Siamese architecture (Bromley et al., 1993), an encoder representation is taken for each of the news articles following which a linear layer is used for reducing the dimensionality of the representation. An aggregation of these representations is then performed for having a unified representation of the two articles before passing them through fully connected layers that finally output the similarity score from 1 to 4. The entire architecture was trained end-to-end by minimizing the mean squared error (MSE) loss between the actual scores and model predictions. We experimented with various strategies in this generalized architecture which we elaborate in the remainder of this section.

Lang	#Test	Lang	#Test
ar_ar	78	es_it	247
ar_en	11	fr_fr	98
de_de	494	fr_pl	10
de_en	152	hu_hu	1
de_fr	85	it_en	1
de_pl	27	it_it	371
de_ru	1	ja_en	2
el_el	1	ja_ja	5
en_ar	11	ja_zh	2
en_de	9	nl_fr	2
en_en	268	pl_en	58
en_es	6	pl_pl	179
en_fr	3	ru_de	2
en_it	17	ru_en	3
en_pl	5	ru_ru	196
en_ru	1	tr_tr	240
en_zh	10	zh_en	63
es_da	1	zh_nb	1
es_en	380	zh_zh	164
es_es	194		

Table 2: Stats of different languages available in test set

1. **Features passed to the encoder.**
   (a) **News text.** The plain text content of the news article is used as the input. The distribution of the news content lengths is shown in Figure 2. The articles have a mean length of 589 tokens and a standard deviation of 954, with about 60.5% of the documents having length greater than 512 tokens, which is the maximum tokenization output size for the encoders used. Considering, however, that most of the salient information of a news article is contained in its initial section, the token lengths were capped to 512 for all news texts before feeding to the encoder.
   (b) **Metadata.** Information such as title, description, keywords and tags was present for the many of the news articles, the statistics of which are shown in table 3. We ran experiments using only these features as well as concatenating them before the news text, capping the overall length to 512 tokens.
   (c) **NER-extracted features.** Since the task entails determining the similarity based on objective features such as entities, date/time

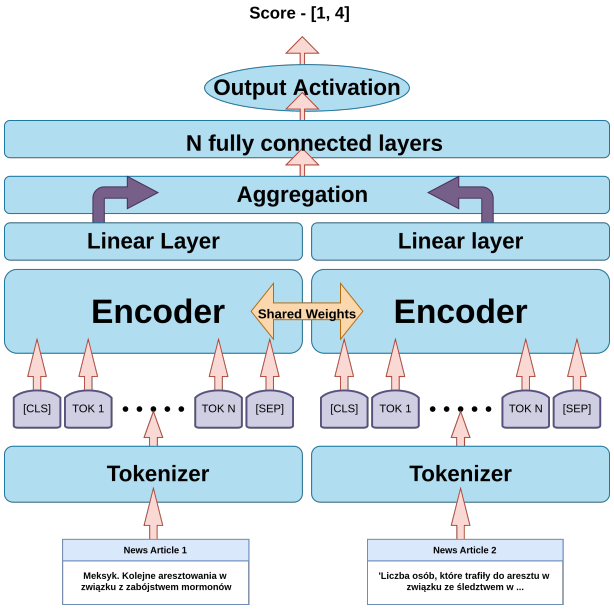

Figure 1: Underlying architecture used across the experiments performed

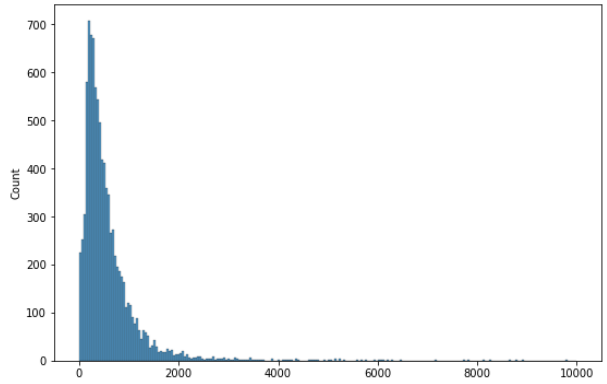

Figure 2: Distribution of number of tokens per news article

values, places, NER features for all the available tags were extracted from text using the Multilingual BERT model from DeepPavlov (Burtsev et al., 2018).

2. **Base encoder model.**

   (a) **XLM-RoBERTa.** (Conneau et al., 2020) This transformer-encoder (Vaswani et al., 2017) based model used was pretrained on 100 languages using masked language modeling objective, achieving remarkable improvements on various cross-lingual understanding tasks.

   (b) **Multilingual DistilBERT.** (Sanh et al., 2019) A distilled version of the multilingual cased BERT-base (Devlin et al., 2018) model trained on Wikipedia data from 104 languages was also used as the base encoder. While the performance of the distilled version was slightly lesser as compared to the original model, this version was chosen on account of it having relatively less no. of paramaters which usually suits well for low-data settings.

Attribute	Count
meta_keywords	4430
tags	4430
title	4421
meta_description	4121

Table 3: Count of the non-null metadata attributes present in the data for the ones used as additional features.

3. **Concatenating encoder representations.** The following concatenation strategies were tried out for building an aggregated representation of the two news articles:

   (a) $[\,|\,x_1 - x_2\,|\,;\,(x_1 + x_2)/2\,]$

   (b) $[\,x_1\,;\,x_2\,;\,|\,x_1 - x_2\,|\,]$

   Here, $x1$ and $x2$ are the encoder representations of the two news articles after passing through the linear layer for reduced dimensionality.

4. **Data augmentation.** Synthetic data samples were created by randomly permuting the sentence order in the news articles to create an augmented data of size 3, 4 and 5 times the original size.

5. **Output activation.** We tried using Sigmoid and ReLU activation functions at the output along with also trying out simple linear output without any activation to determine the best one.

6. **Ensembling.** An ensemble of two models was created by combining the prediction scores of the models by simple average as well as weighted average, in the latter case of which the scores were determined by first training a linear regression model based on the prediction scores on the train data.

4 Experimentation & Results

4.1 Experiments

In this section, we describe the set of experiments performed based on the strategies described in sec-

tion 3. Unless otherwise specified, sigmoid activation was used at the output and the final score was scaled in the range of 1 to 4 as $3 * sigmoid(.) + 1$.

1. **XLM-TXT**: News text feature passed as input to XLM-RoBERTa (XLMR).
2. **DB-TXT**: News text feature passed as input to the DistilBERT model.
3. **XLM-MTD**: Concatenated metadata features passed as input to XLMR.
4. **XLM-MTD-TXT**: Concatenation of metadata features with text passed to XLMR.
5. **XLM-NER-MTD**: Concatenation of extracted NER features and metadata passed as input to XLMR.
6. **DB-NER-MTD**: Concatenation of extracted NER features and metadata passed as input to DistilBERT.
7. **DB-CAT3**: The second concatenation strategy for feature aggregation as described in section 3 used in the same setting as DB-TXT.
8. **XLM-REL**: ReLU activation used at the output in the same setting as XLM-TXT.
9. **DB-CAT3-LIN**: A simple linear output without any activation kept in the same setting as DB-CAT3.
10. **DB-DA3**: Data augmented to thrice the original size, fed to setting same as DB-TXT.
11. **DB-DA4**: Data augmented to four times the original size, fed to setting same as DB-TXT.
12. **XLM-DA5**: Data augmented to five times the original size, fed to setting same as XLM-TXT.
13. **SA**: Simple averaging of the predictions of XLM-MTD-TXT and XLM-DA5.
14. **WA**: Weighted average of the predictions of XLM-MTD-TXT and XLM-DA5.

4.2 Training Setup

The training was done with the number of epochs ranging from 5 to 10. The batch size for train set was kept to be 4 and gradients were accumulated over 8 steps giving an effective batch size of 32. Adam (Kingma and Ba, 2014) optimizer with a weight decay (Loshchilov and Hutter, 2019) of 0.01 was used and the learning rate was kept to a constant value of 5e-6. Validation was performed four times per epoch and the model performance was evaluated using Pearson's Correlation Coefficient (PCC) and Mean Absolute Percentage Error (MAPE) along with the MSE loss value. The best checkpoint was saved based on the PCC score on validation set, following which the predictions on test set were sent for evaluation.

4.3 Results

Results for all the experiments performed on the validation set are shown in table 4 and the PCC reported on some of the experiments on test set are in table 5.

Experiment	Validation set		
	PCC	MAPE	MSE
XLM-TXT	0.53	0.39	0.98
DB-TXT	0.55	0.41	0.93
XLM-MTD	0.46	0.47	1.03
XLM-MTD-TXT	0.52	0.41	0.94
XLM-NER-MTD	0.45	0.43	1.05
DB-NER-MTD	0.47	0.43	1.04
DB-CAT3	0.42	0.47	1.06
XLM-REL	0.45	0.43	1.05
DB-CAT3-LIN	0.42	0.46	1.10
DB-DA3	0.52	0.38	0.99
DB-DA4	**0.58**	**0.41**	**0.94**
XLM-DA5	0.54	0.37	0.99
SA	0.49	0.42	1.02
WA	0.51	0.53	0.95

Table 4: Results of all the experiments performed on validation set.

Experiment	Test PCC
DB-DA3	0.436
DB-DA4	**0.441**
SA	0.43

Table 5: PCC for three of the experiments performed on test set.

4.4 Analysis

The major insights that can be derived out of the results on the validation set are:

- **Metadata as features.** Using plain text (XLM-TXT, DB-TXT) turned out to be the best way

to capture the representation among the experiments we tried. The metadata information in itself (XLM-MTD) was insufficient to provide the representation. Even concatenating the metadata information with text (XLM-MTD-TXT) resulted in a suboptimal solution.

- **NER output as features.** The NER output concatenated with metadata features (XLM-NER-MTD, DB-NER-MTD) did not result in a great feature modeling, with the metadata in combination with text and only text input performing better.

- **Encoder model.** DistilBERT-based multilingual model performed consistently better than its XLMR counterpart across similar experiments (DB-TXT v/s XLM-TXT, DB-NER-MTD v/s XLM-NER-MTD). This might be due to DistilBERT having lesser no. of parameters, thus suiting better against overfitting.

- **Concatenation strategy.** The aggregation strategy of concatenating the absolute difference and average of the two news article representations (DB-TXT) worked better than the other strategy (DB-CAT3) tried out.

- **Output activation.** Sigmoid activation[2] achieved the best training trajectory and results as compared to ReLU (XLM-REL) and linear (DB-CAT3-LIN). During training, the loss from ReLU and linear activations started off with very high values before achieving convergence at values subpotimal as compared to that on sigmoid output.

- **Ensembling.** Simple (SA) and weighted averaging (WA) for ensembling performed competitively, with the validation PCC on simple averaging having an edge over the weighted averaging one.

- **Effectiveness of data augmentation.** Data augmentation turned out to be the best strategy so far, as obvious from the results on validation and test sets. Augmenting the data to 4 times the original size turned out to the best among the tried values $\epsilon \{3, 4, 5\}$ (DB-DA3, DB-DA4, XLM-DA5).

As an overall insight, it seems the method of globally representing the news article by a single representation after capping the length to a smaller, fixed size is not the best way in modeling a solution for this problem. A solution exploiting the information present in the articles at a more granular level - either by explicit feature extraction or implicit detection of these features through techniques suitable across multiple languages can be experimented with in pursuit of better results.

Another possible conjecture we speculate based on the poorer performance on the test dataset is that finetuning the underlying multilingual transformer encoder models might have hampered the effectiveness of the multilingual representation for languages that were not present in the training or validation sets, but were present at the time of evaluation. It is to be noted that since the validation dataset was a split taken from the original training data itself, there was not much of a difference between these two distributions. Hence, if this conjecture holds, a solution that trains the model parameters for the task without deranging their multilinguality aspect should provide a scalable solution across multiple languages.

5 Conclusion

We presented our base architecture adopted for the task of multilingual news article similarity and explained the various experiments performed on the same. An analysis of the results gave insights on what strategies worked best on our underlying Siamese architecture, of which we determined data augmentation to be the most effective one. Finally, we provided some insights on the modeling efficiency of the architecture adopted along with directions for possible improvement. This problem being very relevant in the face of diverse, multilingual news content being continually generated from diverse sources, improvements achieved in this task would be of value in industry and social benefit.

References

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras

[2] Apart from XLM-REL and DB-CAT3-LIN, all the experiments used sigmoid activation at the output.

Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Q. Camargo, Fabian Flöck, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.