HuaAMS at SemEval-2022 Task 8: Combining Translation and Domain Pre-training for Cross-lingual News Article Similarity

Sai Sandeep Sharma Chittilla and **Talaat Khalil**
Huawei Technologies R&D
Amsterdam, Netherlands
{sandeep.chittilla, talaat.khalil}@huawei.com

Abstract

This paper describes our submission to SemEval-2022 Multilingual News Article Similarity task. We experiment with different approaches that utilize a pre-trained language model fitted with a regression head to predict similarity scores for a given pair of news articles. Our best performing systems include 2 key steps: 1) pre-training with in-domain data 2) training data enrichment through machine translation. Our final submission is an ensemble of predictions from our top systems. While we show the significance of pre-training and augmentation, we believe the issue of language coverage calls for more attention.

1 Introduction

In the recent few years, there has been a growing interest towards automating news understanding tasks thanks to the continuous demand by downstream applications. One of the most important aspects of news understanding is identifying news articles that cover the same stories. Grouping such similar articles can be useful in multiple application scenarios including news recommendation, news stories analysis, news retrieval and ranking amongst others.

Task 8 in SemEval 2022 (Chen et al., 2022) provides an experimental setup to address and identify the challenges of assessing the similarity between two news articles. Unlike standard document similarity tasks (Agirre et al., 2015), this task is focused on a more challenging multilingual setting where the systems are not only expected to evaluate pairs of long text articles in the same language but in different languages as well. Moreover, the task assumes that accurately identifying articles that share the same story cannot be solely captured by textual similarity since the underlying similarity function is hypothetically a combination of a set of other features like time, geo-location, mentions of named entities and narratives.

Participating teams are required to provide a similarity score for each pair of news articles. The scores should range between 1 and 4 where 1 indicates that two stories are almost identical and 4 indicates no similarity at all. The systems are evaluated based on the Pearson's correlation with ground truth scores which are provided in a test set annotated by human evaluators.

In our submission, we hypothesize that the sub-dimensions (that are assumed to be story similarity predictors by the task organizers) do not have to be independently modeled and present an approach that assumes that such sub-dimensions can be represented in the model's latent space while directly optimizing to learn article similarity scores. Formally speaking, we model the similarity task as a supervised regression problem to optimize the similarity score between two news articles. To test our hypothesis, we experiment with different pre-trained language models and evaluate multiple methods of boosting the model's performance via domain pre-training and data augmentation. Our final submission ranked 8 out of a total of 32 teams in the official rankings with a Pearson correlation of 0.771 - a score difference of 0.047 compared to the top ranked submission.

This paper is organized as follows: In section 2 we describe the official datasets as well as the external datastes that were used. In section 3 we present our baselines and systems' setups. Section 4 presents our data splits and model training specifics. We show and analyse our results in section 5, then we conclude and discuss future work in section 6.

2 Datasets

The used datasets are categorized into two categories: 1) Task Datasets which are provided by the task organizers and 2) External datasets that are used as additional assets for model training.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1151 - 1156

Pair	Training	Evaluation
ar-ar	263	298
de-de	832	611
de-en	532	190
de-fr	-	116
de-pl	-	35
en-en	1689	235
es-en	-	498
es-es	506	243
es-it	-	320
fr-fr	69	111
fr-pl	-	11
it-it	-	442
pl-en	-	64
pl-pl	333	224
ru-ru	-	287
tr-tr	419	275
zh-en	-	223
zh-zh	-	769
Total	4643	4953

Table 1: Count of training and evaluation samples by language-pair. "-" means not present in training data.

2.1 Task Dataset

The task organizers provided a number of 4,964[1] news article pairs along with their "Overall Similarity" scores to be used for training purposes. Multiple trained human evaluators were asked to evaluate pairs of news articles and provide similarity scores for different sub-dimensions as well as an overall score; all in 1-4 range. The final scores are calculated by averaging the individual scores across all the evaluators. The training set contains same language pairs as well as cross-lingual pairs. The evaluation dataset was created in the same way as the training data however it contains new languages that are not present in the training set and the language pair distribution does not match the training set distribution as shown in Table 1.

2.2 External Datasets

We made use of monolingual datasets from multiple news sources to fine-tune the pre-training phase of the language models that we experimented with (see section 3 for the details). NADiA dataset (Al-Debsi et al., 2019) is used for Arabic, CCNEWS dataset (Hamborg et al., 2017) is used for English, Global Voices news data (Tiedemann, 2012) is used for Polish and MLSM (Scialom et al., 2020) dataset's input text is used for German, Spanish, French and Turkish.

[1]Task organizers published article URLs but only 4643 pairs were retrievable by the time we scrapped them

3 System Overview

All of our systems are solely trained on the text descriptions of input articles as features and their similarity score as the target variable. All the presented systems except for the baseline system 3.1, are based on XLM-RoBERTa (XLM-R). This choice is made based on the fact that it's a multilingual model that is pre-trained on large amounts of data spanning 100 languages and performs competitively on several cross-lingual transfer tasks (Conneau et al., 2019). The general architecture of our systems itself is kept relatively simple i.e. a Language Model with a regression head on top.

3.1 Baseline

Our baseline system is a multi-variate linear regression model that uses 2 independent variables: a) count of the named entities[2] that are shared between the two news articles and b) cosine similarity between the sentence embeddings of the news article pair. Formally speaking; We model this as $Y = A + B_1X_1 + B_2X_2$ where X_1 and X_2 are the aforementioned variables and Y is the similarity score. The model is trained to minimize the Mean Square Error (MSE). We also experimented with a Gamma Function for regression on the non-negative similarity values, however no difference was perceived in Pearson's scores. We evaluated sentence embeddings generated using LaBSE (Feng et al., 2020), MPNET (Song et al., 2020) and SBERT (Reimers and Gurevych, 2019) and reported baseline results using LaBSE embeddings since it resulted in the least MSE.

3.2 XLMR

Our first system dubbed *XLMR* is an XLM-R model with a regression head on top which is trained to minimize the MSE on the task dataset 2.1. The input is formed by concatenating the text of the two input articles and placing a special token in between.

3.3 XLMR-Pre

Our second system dubbed *XLMR-Pre* follows the same architecture as *XLMR* however, we continue model pre-training using the Masked Language

[2]Exact lexical matching

Pair	Original	After
ar-ar	263	1607
de-de	832	2176
de-en	532	532
en-en	1689	1689
es-en	-	1344
es-es	506	1850
es-it	-	1344
fr-fr	69	1413
it-it	-	1344
tr-tr	419	1763
pl-pl	333	1677
ru-ru	-	1344
zh-zh	-	1344
zh-en	-	1344
Total	4643	12707

Table 2: Count of training samples by language-pair before and after data augmentation by translation. Please note that only samples from the training split were translated to avoid any *potential* data leakage. "-" means not present in training data.

Modelling (MLM) objective (Devlin et al., Liu et al.) for languages in the training set using the collected external datasets 2.2.

3.4 XLMR-Aug

Our third system dubbed *XLMR-Aug* follows the same setup of *XLMR* however the training data is supplemented with data augmentation. Synthetic data is created in two different ways namely: pair switching and Machine Translation (MT). Pair switching is achieved by switching the text concatenation order of the input pairs to act against pair order bias. Machine Translation is leveraged to address the fact that the majority of the training set pairs are in the same language (4500 pairs) however the systems are evaluated on their ability to score cross-lingual pairs as well. A number of 1344 English pairs are sampled and translated to different languages. Additionally a number of 667 German language examples are translated into English to encourage improvements in the English language pairs. Translated pairs' statistics are reported in Table 2. The used MT system is an in-house general purpose Transformer Big model (Vaswani et al., 2017) that is not adapted to any specific domain.

3.5 Ensemble Systems

Ensemble systems are developed by averaging the individual scores of different combinations of our three systems: *XLMR*, *XLMR-Pre* and *XLMR-Aug*.

4 Experimental Setup

4.1 Data Setup

To be able to run our model selection experiments and validate our hyper-parameter settings, a standard split of 80:10:10 is applied to the task dataset 2.1 to split it into train, validation and held-out sets. This setting resulted in a training set size of 3719 samples, validation set size of 464 and a held-out set size of 460 samples. The validation set was used for hyper-parameter tuning and the held-out set was used for system comparisons since there were no datasets provided for such purposes by the task organizers. Our experiments that leveraged data augmentation techniques made use of a total of 12707 training samples. To avoid any data leakage, we only used samples from the train set to augment the data.

4.2 Hyper-parameters

Model hyper-parameters were initially setup with the recommended values for XLM-R model fine-tuning (Conneau et al., 2019) and were manually fine-tuned based on the correlation scores and the loss on the validation set. *XLMR* and *XLMR-Pre* systems were trained for 4 $epochs$ and *XLMR-Aug* system was trained for 10 $epochs$. In all the experiments, an $AdamW$ optimizer (Kingma and Ba, Loshchilov and Hutter) was used with a linear schedule, a $learning\_rate = 2e-5$, $epsilon = 1e-8$ and training and validation $batch\_size = 8$.

4.3 Training

The Huggingface transformers library[3] was used to conduct all our model training experiments and all our models were initialized using xlm-roberta-large[4] weights. All models were trained using 8 Nvidia Tesla-V100 GPUs.

4.4 Evaluation

We evaluated our models using Pearson's correlations score on the overall test set. Additionally, we conducted a per-language correlation scoring for better reasoning and model development.

[3]https://github.com/huggingface/transformers
[4]https://huggingface.co/xlm-roberta-large

System	Score
Baseline	0.677
XLMR	0.808
XLMR-Pre	0.804
XLMR-Aug	0.790

Table 3: Pearson's scores for different systems during model development on the held-out set

5 Results and Analysis

During the development phase, we used the fixed held-out set to compare the performance of different systems. When the evaluation phase ended, the gold labels were made available for the evaluation set and thus we re-evaluated all our systems using that set as well. The results on the held-out set are shown in Table 3 and the results on the evaluation set are shown in Table 4. A per language breakdown scoring is also provided on both the held-out and the evaluation sets in Table 5 and Table 6 respectively.

Our submitted system is ES_2S_3 (table 4) which is an ensemble of *XLMR-Pre* and *XLMR-Aug*, however our post-evaluation analysis showed that $ES_1S_2S_3$ which an ensemble of our three systems performs slightly better than our official submission with a marginal increase of 0.004 in the correlation score. We attribute this increase to the power of ensembling given that the three systems were competitive to each other in terms of the aggregate performance scores however each system has it's own strengths when it comes to language specific performances as shown in Table 6. A little inconsistency between the scores of the different systems on the held-out and evaluation sets is attributed to the fact that the evaluation set has unseen language pairs and a radically different language distribution compared to the training and held-out distributions.

Our per language evaluation (Table 6) reveals explainable patterns. *XLMR-Pre* performs the best on fr_pl and fr_fr language pairs due to abundance of French pre-training data in this model. *XLMR-Aug* performs the best in 12 out of 18 language pairs due to it's MT augmentation that boosts it's performance on unseen pairs. An ensemble of *XLMR* and *XLMR-Pre* performs the best for en_en pairs due to the bias of the original XLM-R model towards English, the news domain fine-tuning and the lack of translation noise or parameter sharing competition with other languages.

System	Score
Baseline	0.615
XLMR (S_1)	0.752
XLMR-Pre (S_2)	0.755
XLMR-Aug (S_3)	0.753
ES_1S_3	0.768
ES_1S_2	0.767
ES_2S_3	**0.771***
$ES_1S_2S_3$	**0.775**

Table 4: Pearson's scores for different systems on the evaluation set. ES_iS_j is an ensemble of S_i and S_j. * *indicates our best submitted system*

Pair	XLMR	XLMR-Pre	XLMR-Aug
ar_ar	0.603	**0.717**	0.606
de_de	0.810	0.788	**0.838**
de_en	0.862	0.862	**0.899**
en_en	0.818	**0.827**	0.764
es_es	**0.914**	0.861	0.906
fr_fr	**0.812**	0.762	0.682
pl_pl	**0.709**	0.622	0.579
tr_tr	0.823	0.782	**0.841**

Table 5: Language pair wise Pearson's scores for different systems on the held-out set

6 Discussion

In this paper we described our submissions to the news similarity task in SemEval 2022. Our models showed competitive performance by leveraging pre-trained language models and showed that further improvements can be gained by the use of domain pre-training and data augmentation using machine translation. Due to the competition time limits such domain pre-training and translation experiments were conducted on relatively small datasets and we did not manage to experiment with a model that combines both additions. We believe that scaling these approaches by using huge amounts of monolingual data across different languages is potentially a direction that is worth exploring.

We see improvement posibilities when it comes to modeling as well. In the early stages of our experimentation we tried a contrastive learning based approach similar to the works done by (Chopra et al., 2005) though, initial results were not promising and we decided to discard this direction, we believe that further efforts can be fruitful. We've also experimented with explicit modeling of Named Entities within our models without a positive outcome

Pair	XLMR (S_1)	XLMR-Pre (S_2)	XLMR-Aug (S_3)	ES_1S_3	ES_1S_2	ES_2S_3	$ES_1S_2S_3$
ar_ar	0.784	0.790	0.774	0.797	0.805	0.805	**0.809**
de_de	0.752	0.753	0.730	0.757	0.766	0.763	**0.768**
de_en	0.795	0.797	0.741	0.785	**0.809**	0.793	0.802
de_fr	0.559	0.528	0.583	**0.592**	0.564	0.586	0.590
de_pl	0.673	0.713	0.667	0.701	0.721	0.720	**0.725**
en_en	0.780	0.791	0.756	0.779	**0.795**	0.786	0.791
es_en	0.807	0.810	0.794	0.816	0.821	0.819	**0.824**
es_es	0.819	0.813	0.813	0.828	0.826	0.829	**0.833**
es_it	0.718	0.717	0.744	0.752	0.738	0.750	**0.754**
fr_fr	0.834	0.847	0.818	0.837	**0.848**	0.845	0.847
fr_pl	0.853	**0.943**	0.846	0.862	0.911	0.908	0.898
it_it	0.788	0.766	0.763	0.786	0.790	0.781	**0.790**
pl_en	0.632	0.615	**0.712**	0.709	0.659	0.703	0.705
pl_pl	**0.679**	0.655	0.643	0.672	0.678	0.663	0.675
ru_ru	0.704	0.678	0.718	**0.724**	0.703	0.717	0.719
tr_tr	0.810	0.814	0.804	0.824	0.827	0.830	**0.833**
zh_en	0.684	0.689	0.758	**0.763**	0.715	0.763	0.762
zh_zh	0.729	0.725	0.739	0.748	0.741	0.750	**0.752**

Table 6: Language pair wise Pearson's scores for different systems on the evaluation set. ES_iS_j is an ensemble of S_i and S_j

however this could be due to the fact that we used a very simple string matching approach for named entities identification. Another modeling aspect is the train/evaluation language distribution modeling. Given that the distribution of evaluation language pairs are available, one could leverage this to improve the model optimization process.

Finally, in this exploratory work we haven't made use of any available article related meta-data which can have strong predictive power of article similarity. Examples include URL normalization to identify parallel articles in different languages, domain and country information among other features. We leave out these territories to be explored in future works.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Ridhwan Al-Debsi, Ashraf Elnagar, and Omar Einea. 2019. Nadia: News articles dataset in arabic for multi-label text categorization.

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.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.