Assessing Multilingual News Article Similarity

Chan Zhuo Hua Leo Ly Reema Kumari William Nafack

1. Abstract

Multilingual text processing poses new challenges to the standard NLP techniques. In this paper, we employ various pre-trained architectures to understand text similarity between news articles from multiple languages. The target is a human annotated overall similarity on a scale of [1,4]. We looked into XLM-Roberta for assessing multilingual news article similarity. We employed different methods to enrich the training data - combining head-tail text data and meta data, and augmenting data using Google Translate API.

2. Introduction

Document similarity is used in many text processing tasks such as text classification, clustering and information retrieval. It is usually measured as the distance between the vector representations of text documents under the assumption that terms are independent and unordered, thus the structural information in text is lost. Since the association between terms in text contributes towards the meaning of the text document, considering the structural information in measuring similarity can potentially improve the accuracy of document classification. Furthermore, systems that make use of document similarity are often restricted to one language. They do not take into account multi-lingual sourced documents making it impossible to track similar news coverage / articles between different outlets or regions. In this paper, we aim at tackling this particular task and describe a comparison of multiple Natural language processing (NLP) models to score overall similarity of news articles from multiple languages. In this case, we are interested in the real-world happenings covered in the news articles, not their style of writing, political spin, tone, or any other more subjective "design decision" imposed by a medium/outlet. We use the 2022 SemEval Task 8 data for training and evaluation where human annotation for comparison of a pair of news articles on various parameters such as geolocation, shared entities, shared narratives are present. Overall score (out of 4) represents the similarity between articles considering all parameters. We use a correlation between our predicted overall score and human annotated overall score between a pair of articles to assess the models.

3. Literature Review

Bär et al. (2011) explored the notion of similarity between texts in the context of NLP. They demonstrate that the similarity between documents is often ill-defined and used as a term that handles a completely different subject. They formalize text similarity and suggest content, structure, and style as the major dimensions inherent to texts. However, structure and style are not actively being accounted due to the multi-facet articles that we treat and hence we only focus only on the content. Most similarity approaches operate on word-distribution-based document representations. They do not take into account the semantic relations between documents which can be problematic when the documents differ in language, vocabulary or type. There has been little to no work handling the cases of multilingual documents hence we refer to some methods used mainly with English sourced documents. Paul et al. (2016) exploits the knowledge available in Knowledge Graphs to leverage information gathered from relations between entities in research articles. They use this to establish a semantic similarity between documents based on hierarchical and transversal relations. Transformer-based models have been the state of the art in many NLP tasks. Their introduction enabled a transition from the context-free word embeddings such as Word2Vec Mikolov et al. (2013) and GloVe Pennington et al. (2014). Ostendorff et al. (2020) worked on predicting specific topic relations between wikipedia articles. They implemented two architectures for extracting sentence embeddings via a BERT Devlin et al. (2019) and Siamese BERT Reimers & Gurevych (2019a), which was then feed into a mutli-layer perceptron for their classification task. Referring back to the SemEval task, (Xu et al., 2022) which won the competition used data augmentation on top of XLM-Roberta Liu et al. (2019) to get the best results.

4. Methodology

In this section, we describe the dataset and investigated systems to facilitate the reproduction of our results.

4.1. News Dataset

Given dataset from SemEval has the links for a pair of news articles, languages in which these articles were written and multiple scores to assess the similarity between the articles

on various dimensions and overall similarity. We have 4964 records in the train set and 4902 records in the test set.

The links can be used to scrape details such as title, text, description, keywords etc from the articles using semeval_8_2022_ia_downloader. These details are processed and used for the score prediction task. The score on each of the dimensions is a continuous number on the scale of [1,4].

The training data contains the averaged scores from several annotators that assess pairs of news stories on the following aspects: "Geography", "Entities", "Time", "Narrative", "Overall", "Style", and "Tone". The competition, however, evaluates participants' performance by the ability to estimate the Overall similarity between a pair of news stories, not from any other features.

There are in total 8 different language pair combinations in the training data where as we have 18 different pair of languages in the testing data. Also, English to English article pairs constitute 36% of training data but a very small part of testing data.

4.2. Data Cleaning

Since the articles belong to multiple languages, a unified data cleaning approach across all the articles could not be created. We removed tabs, line breaks, and hyperlinks from all articles as part of the data cleaning process. We explored various methods to combine article components to feed into models. Particularly, we appended a combination of article features and then created a vector of tokens with truncation.

4.3. Data Augmentation

During inspection of the data, we discovered that there were combinations of language news pairs found in the evaluation set that are not in the training set. Therefore, we augmented the training data by applying Back Translation and Translate Train to add the missing pairs data. We used the Google Translate API (Googletrans) `https://py-googletrans.readthedocs.io/`. Due to some API call errors, parts of the resulting data did not get translated properly. Consequently, we filtered out those data points before adding them to the training set. The results are displayed on Table 1.

4.3.1. Back Translation

We enriched the data using Back translations. This procedure consists of translating the data from its source language to English and then back to its source language. The resulting data is then appended to the training set.

4.3.2. Translate Train

Since the training set did not have a few language pairs presented in the test dataset (e.g. de-fr, ru-ru, zh-zh), translate train was applied to complete the missing pairs. This procedure translated data from its source language(s) (one of the pair or both) to a destination language(s). Then, we appended this new data to the training set.

Lang. pairs	Train	Eval	Train+DA
AR-AR	274	298	518
DE-DE	857	608	857
DE-EN	577	185	577
DE-FR	0	116	103
DE-PL	0	35	73
EN-EN	1800	236	1800
ES-EN	0	496	796
ES-ES	570	243	1085
ES-IT	0	320	278
FR-FR	72	111	137
FR-PL	0	11	62
IT-IT	0	411	488
PL-EN	0	64	76
PL-PL	349	224	581
RU-RU	0	287	378
TR-TR	465	275	860
ZH-EN	0	213	632
ZH-ZH	0	769	192

Table 1. Data sets distributions

4.4. Architecture

BERT is a bi-directional encoder representation for transformers. Using bi-directional transformers, it is able to capture the semantic relationship from long-term context words in both directions of the key word to create the representation (Devlin et al., 2019). RoBERTa is a robustly optimized BERT pre-training approach which is trained on more data and epochs (Liu et al., 2019). As these approaches learn a good representation of the text, they perform well on contextual NLP tasks such as text classification.

5. Models

5.1. XLM-RoBERTa

XLM-RoBERTa (Conneau et al., 2019) is a transformer based multilingual language model which learns word representations in 100 separate languages using an architecture on masked language tasks. The training procedure is the same as the traditional RoBERTa model for English language. It has outperformed monolingual language models on tasks involving single language as well such as GermEval18 or GermEval14.

We modified our RoBerTa architecture to handle our use

case of feeding both documents into the Transformer. The documents were separated by a [SEP] token, indicating the split between the seed document and target document as shown in Figure 1 below. The encoded representation was then passed into our fully connected classification layer to perform the regression over the identified class labels.

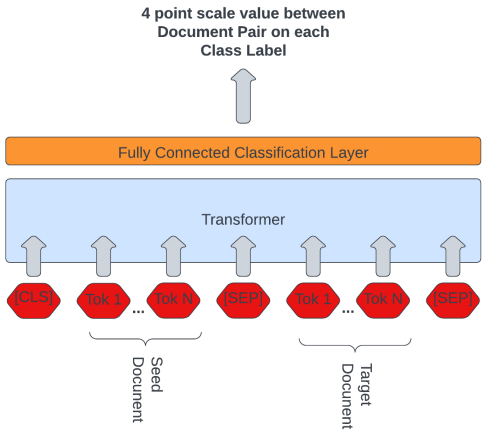

Figure 1. RoBERTa Architecture for Document Pair Similarity

For this architecture, we created 10% validation dataset to tune parameters. For the text, we used a combination of title and text. We then took 200 tokens from the head and 56 tokens from the tail as we believe that the head and tail part would encode most of the relevant information. We used the encoding of the [CLS] token from the model to encode all the information from the sequence to further perform regression for the different similarity dimension scores such as style, tone, geolocation, and overall. We used the Pearson Correlation Coefficients (Pearson CC) between predicted scores for overall and actual scores on the validation dataset to obtain the best model.

5.2. Weighted Loss

Since each pair of news stories' similarity is annotated on 7 dimensions, with the emphasis on the Overall score, we postulated that the dynamics between Overall scores and the remaining 6 target variables can be helpful in building a robust model. As the Overall score is the most important variable, for each attempt, we assigned the highest weight to the Overall loss, while other aspects share equal weights to each other but less than the Overall. For example, if the Overall gets 70% weight, then the remaining 6 dimensions constitutes 30% weight, with each sub-dimension occupies a weight of 5%.

5.3. Siamese XLM-RoBERTa

Sentence embeddings using Siamese Networks was introduced by Reimers & Gurevych (2019b). Sentence-BERT fine-tunes a pre-trained BERT network using Siamese and triplet network structures and adds a pooling operation to the output of BERT to derive a fixed-sized sentence embedding vector. The produced embedding vector is more appropriate for sentence similarity comparisons within a vector space. We created a Simaese XLM-RoBERTa inspired by the Siamese architecture. The motivation to use this architecture was to exploit longer sequence length for our documents and improve our results on document similarity comparisons. By simultaneously passing the pair of documents as two input text into the same transformer architecture, we can use **512** tokens for each document instead of **256** tokens. The embeddings from both pair of documents are then concatenated. In the literature there is no widely accepted method of concatenation. Conneau et al. (2017) uses $[u, v|u - v|u * v]$ for sentence embeddings while Sentence BERT Reimers & Gurevych (2019b) presents $[u, v|u - v|]$ as the best method. We experimented with the following concatenations:

- $[u, v]$ Concatenation of the two vectors u and v.
- $[u, v|u - v|]$ and absolute value of element-wise difference.
- $[u, v, |u - v|u * v]$ and element-wise product.

We proceeded to use the simple concatenation process which performed the best, which was then fed into a single linear layer for our Regression with dropout. We trained for 5 epochs using a batch size of 4 and a learning rate of 2×10^{-5}. The modified architecture came at the cost of increasing the computation time so we only trained on the Overall class label without weighted loss. Figure 2 below shows a description of the architecture.

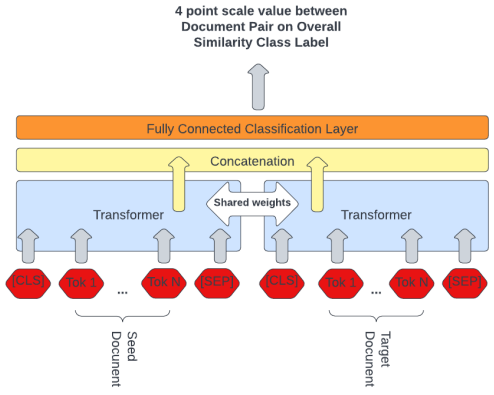

Figure 2. Siamese RoBERTa Architecture for Document Pair Similarity

6. Results

We trained our model gradually to evaluate the importance of each added feature. First, our model was trained on text data only (Baseline). Meta-data, comprising of meta keywords and meta description, and data augmentation (DA) was added in the next iteration. Single label loss, solely based on Overall similarity, was then substituted with various multi-label loss with the following weights for Overall scores: 25%, 50%, 60%, 70%, 80%, and 90%. Finally, the best model after previous iterations was hyperparameter-optimized. Table 2 summarizes our results.

Table 2. Results on various models

Model	Pearson CC
Baseline	67.0
+ Meta + DA	72.3
Siamese XLM-RoBERTa	
+ Meta + DA ($[u, v]$)	29.18
Siamese XLM-RoBERTa	
+ Meta + DA ($[u, v\|u - v\|]$)	23.27
Siamese XLM-RoBERTa	
+ Meta + DA ($[u, v, \|u - v\|u * v]$)	24.34
Meta + DA + Weighted loss	
25% Overall	72.6
50% Overall	70.2
60% Overall	72.1
70% Overall	72.4
80% Overall	72.3
90% Overall	71.7

7. Discussion

The baseline model which solely utilizes text data performs the worst on our data. It is expected as most of new stories are much bigger than 256 tokens. Even when using a combination of both head and tail portions of only text data, the model did not greatly improve.

With the support of meta data, however, there was a noticeable improvement from 67% to 72.3%. We chose to include meta data because they seem to well capture the content of news stories with just a few words. We ordered data in a way to prioritize meta data over title and text. However, many documents lack meta data, preventing it from becoming a reliable feature.

In the next set of iterations, we tried with various Overall weighted loss, ranging from 25% to 100%. The Pearson CC scores range from 70.2% to 72.6%. With little deviation in terms of performance between different weighted losses, we believe that weighted loss does not contribute to the model's performance. It also means that other 6 sub-dimensions are highly correlated with the Overall target variable.

The Siamese XLM-RoBERTa did not perform as we expected. This could be due to the concatenation approach that was used. There is still no clear distinction of which concatenation operations fits better for the sentence embeddings especially when it comes to capturing the context of the text. In general, the element-wise difference measures the distance between the dimensions of the two document vectors and, thus, ensures that similar pairs are closer to each other than dissimilar pairs. We would expect this effect to be evident in our correlation between the pairs of documents, but that was not the case. Furthermore, while the Single model treats every document as pair, the Siamese model treats every document in a pair independently. This could result in missing out some key relationships between terms in the documents which the single model can capture, especially for the Multilingual context. Moreover, for being computationally expensive, the model does not allow us to perform a weighted loss using the other sub-dimensions defining the document similarity. Neither does it allow us to train extensively with different parameters which could improve the results, as shown above with the single model.

7.1. Negative Effects

In many attempts during the quest to optimize the models, we found a few factors negatively affect the model's performance. We attempted to address overfitting by using Dropout layers as regularization method. However, the Pearson CC score on testing data decreased to 71.4%.

We then used the best model from the Overall weighted loss and introduced more training data by using Google Translate API noted in section 4.3. We optimized hyperparameters and increased the number of training epochs from 8 to 20 times. Cosine warmup rate is also included. The training loss significantly reduced but not the testing loss, with Pearson score fluctuates between 72% and 73%. We concluded that it was not worth the extra training time. It signified that overfitting remains a serious issue. One possible explanation is due to many surprising languages in testing dataset, the model is unable to make good distinction, even with the help of data augmentation using Google Translate API.

8. Conclusion

Assessing similarity between pair of documents in a multilingual context is a challenging task. In this project we made use of language models to capture important information that could help in calculating the similarity between the documents. Another area that could be explored is the use of Graph Networks. Knowledge graphs are very useful for extracting and linking entities from documents. This could be used to explore the similarity between documents based on their knowledge graph representations.

References

Bär, D., Zesch, T., and Gurevych, I. A reflective view on text similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 515–520, Hissar, Bulgaria, September 2011. Association for Computational Linguistics. URL https://aclanthology.org/R11-1071.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL https://aclanthology.org/D17-1070.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Ostendorff, M., Ruas, T., Schubotz, M., Rehm, G., and Gipp, B. Pairwise multi-class document classification for semantic relations between wikipedia articles, 2020. URL https://arxiv.org/abs/2003.09881.

Paul, C., Rettinger, A., Mogadala, A., Knoblock, C. A., and Szekely, P. A. Efficient graph-based document similarity. In *ESWC*, 2016.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019a. URL https://arxiv.org/abs/1908.10084.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Xu, Z., Yang, Z., Cui, Y., and Chen, Z. Hfl at semeval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity. *arXiv preprint arXiv:2204.04844*, 2022.

A. Appendix

A.1. Example of Translate train

Text2 of pair with id 1559133520_1558214611 translated from fr to pl.

Key	Value
pair_id	1559133520_1558214611
url2_lang	fr
text2	Stocks réduits ou distribués au "compte-gouttes", les soignants français manquent toujours de masques dans leur lutte contre le coronavrius. Ils bénéficient de dons de particuliers en attendant les livraisons. Pour les soignants le problème reste entier. Le personnel médical en France fait toujours face à un manque criant de masque de protection face au Covid-19. La semaine dernière le sujet a pris l'ampleur d'une polémique tant les médecins, en première ligne dans la lutte contre l'épidémie de coronavirus, manquaient de cet objet devenu emblématique de la crise sanitaire que traverse le pays et alors que les premiers soignants sont morts du coronavirus ces derniers jours. Samedi dernier, le ministre de la Santé affirmé que les autorités avaient commandé "250 millions de masques", qui seront livrés "progressivement" dans les jours qui viennent. A l'heure actuelle, il y a un "stock d'État" de 86 millions d'unités, dont 5 millions de masques FFP2 plus protecteurs, le reste étant des masques chirurgicaux, a rappelé le ministre de la Santé. Il ajoutait : "nous prévoyons une consommation de 24 millions de masques par semaine", priorité sera donnée pour ces masques aux personnels de santé en ville comme à l'hôpital et aux personnes intervenant auprès des personnes âgées.

Key	Value
pair_id	1559133520_1558214611
url2_lang	pl
text2	Zapasy zredukowane lub dystrybuowane do kroplowania", francuskich opiekunów wciaż brakuje masek w walce z Coronavriusem. Korzystaja z darowizn od osób fizycznych, czekajac na dostawy. W przypadku opiekunów problem pozostaje cały. Personel medyczny we Francji wciaż stoi w obliczu rażacego braku maski ochronnej przeciwko Covid-19. W ubiegłym tygodniu podmiot zyskał wielkość kontrowersji jako lekarzy, na pierwszej linii w walce z epidemia Coronawirusa, brakowało tego obiektu, który stał sie symbolem kryzysu zdrowotnego, przez który przechodzi kraj i podczas gdy pierwsi opiekunowie zmarli Coronawirus w ostatnich dniach. W ubiegła sobote minister zdrowia powiedział, że władze dowodziły 250 milionów masek", które zostana dostarczone stopniowo" w najbliższych dniach. Obecnie istnieje akcje państwowe" w wysokości 86 milionów sztuk, w tym o 5 milionów bardziej ochronnych masek FFP2, a reszta to maski chirurgiczne, przywołała ministra zdrowia. Dodał: Planujemy konsumpcje 24 milionów masek tygodniowo", priorytetem zostanie te maski dla personelu medycznego w mieście, jak w szpitalu i osób pracujacych z osobami starszymi.

A.2. Model hyperparameters

XLM-Roberta model:

- Max token length: 512
- Batch size: 5
- Learning rate: $2e^{-5} - 5e^{-6}$
- Weight decay rate: $1e^{-4}$
- Number of epochs: 8 - 15
- Warm up rate: 0.0 - 0.1
- Dropout: 0.00 - 0.15

Single Roberta model:

- Batch size: 8
- Learning rate: $2e^{-5}$

- Weight decay rate: $1e^{-3}$
- Number of epochs: 5
- Dropout: 0.4

Siamese Roberta model:

- Batch size: 6
- Learning rate: $2e^{-5}$
- Weight decay rate: $1e^{-3}$
- Number of epochs: 2
- Dropout: 0.4