LSX_team5 at SemEval-2022 Task 8: Multilingual News Article Similarity Assessment based on Word- and Sentence Mover's Distance

Stefan Heil and **Karina Kopp**
{stefan.heil,karina.sultangaleeva}@stud-mail.uni-wuerzburg.de
Konstantin Kobs and **Albin Zehe** and **Andreas Hotho**
{kobs,zehe,hotho}@informatik.uni-wuerzburg.de
University of Würzburg

Abstract

This paper introduces our submission for the SemEval 2022 Task 8: Multilingual News Article Similarity. The task of the competition consisted of the development of a model, capable of determining the similarity between pairs of multilingual news articles. To address this challenge, we evaluated the *Word Mover's Distance* in conjunction with word embeddings from *ConceptNet Numberbatch* and term frequencies of *WorldLex*, as well the *Sentence Mover's Distance* based on sentence embeddings generated by pretrained transformer models of *Sentence-BERT*. To facilitate the comparison of multilingual articles with *Sentence-BERT* models, we deployed a Neural Machine Translation system. All our models achieve stable results in multilingual similarity estimation without learning parameters.

1 Introduction

The assessment of similarity between documents is a central challenge in the context of information retrieval. Especially the evaluation of similarities between news articles across different languages opens up opportunities for numerous downstream tasks, such as the analysis of regional differences in news coverage of topics or sentiments towards events and news.

Task 8 of the *SemEval* challenge 2022 (Chen et al., 2022) posed the problem of the assessment of similarity of news articles across a variety of languages and provided a dataset of news article pairs in seven languages, as well as a number of bilingual pairs. Apart from an overall similarity score to be estimated in the challenge, a number of scores for similarity in different categories, such as *narrative* or *entities* were provided.

For our submission for this task, we evaluated the performance of the *Word Mover's Distance* (WMD) and *Sentence Mover's Distance* (SMD) in the context of similarity assessment of multilingual news articles. We participated in all given languages, comparing both document pairs in the same language, as well as pairs in different languages. Our code is available at https://github.com/StefanJMU/SemEval2022_Task_8.

We considered WMD and SMD due to their conceptual simplicity and capability to integrate well with resources such as pretrained word embeddings or sentence embeddings, produced by state-of-the-art language models. Additionally, they offer the appeal of being themselves parameter-free, and hence independent of labelled training data.

The rest of the paper is organized as follows: We introduce the approaches deployed in the challenge submission, as well as the used resources, supplementing the *WMD* and *SMD*. Subsequently, we present the results achieved and conclude with a discussion of the results.

2 System Overview

Figure 1 shows a schematic overview of the methods we investigated for our submission. We evaluated two different methods for this task, namely the *Word Mover's Distance* (WMD) (Kusner et al., 2015) and the *Sentence Mover's distance* (SMD) (Clark et al., 2019). Both approaches take two news articles and calculate a similarity score from the representation of both texts as either *Bag of Words* or *Bag of Sentences*, respectively. Both approaches have been found to constitute a metric exhibiting a pronounced correlation with the human-assessed similarity scores of the text pairs (Kusner et al., 2015). We create the required word and sentence representations using *word embeddings* and *sentence embeddings* generated from state-of-the-art language models.

For the *WMD* approach, we deployed a preprocessing pipeline, involving the tokenization of both news articles, as well as the removal of stopwords and punctuations. Subsequently, the preprocessed texts were transformed into a *Bag of Words* using

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1190 - 1195
July 14-15, 2022 ©2022 Association for Computational Linguistics

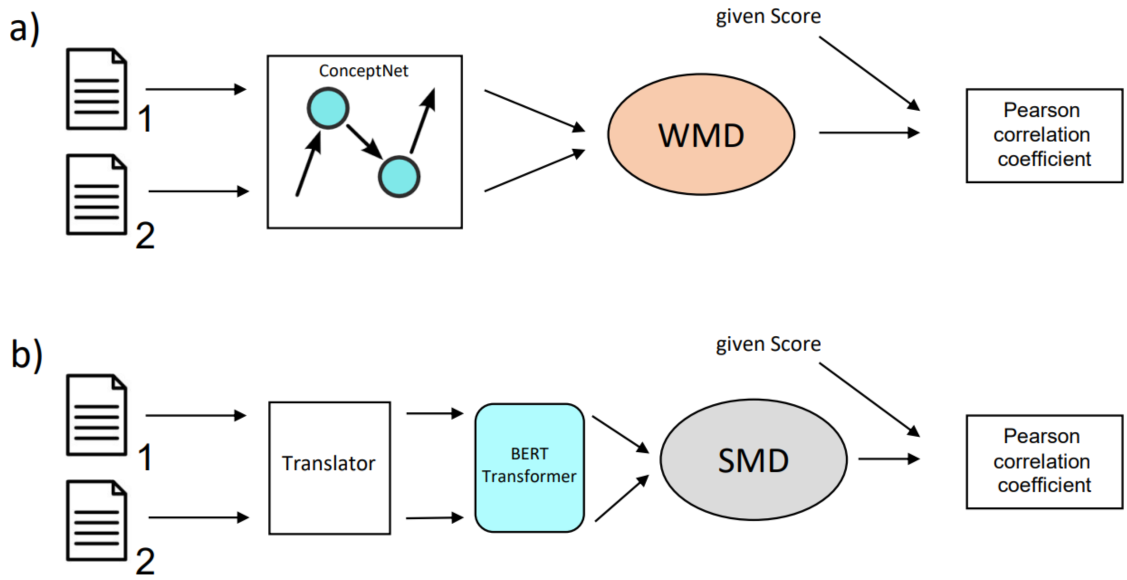

Figure 1: Schematic overview of our experiments in order to determine the best model for submission. Given two news articles, we **a)** use pretrained multilingual ConceptNet Numberbatch word embeddings for both articles and compute the *Word Mover's Distance* (Kusner et al., 2015) between both non-translated texts; also **b)**, we obtain sentence embeddings from a Sentence-BERT model (Reimers and Gurevych, 2019) by first translating both articles to English and then computing the *Sentence Mover's Distance* (Clark et al., 2019) on the translated documents.

the vector representations provided by *ConceptNet Numberbatch*. These vector representations, together with the dot-product similarity, constituted the metric at the core of the linear optimization problem forming the *Word Mover's Distance* between two texts. The *WMD* also allows for a different weighting of the words of texts in the evaluation of similarity. These weights were chosen according to the respective *TF-IDF*, calculated with the help of *WorldLex* (Gimenes and New, 2015).

The *SMD* approach required the decomposition of texts into sentences, which were subsequently encoded with a transformer language model, resulting in a *Bag of Sentences* representation of texts. By interpreting these sentence embeddings as words, a similarity score is readily computable using the *WMD* again.

The following subsections introduce both metrics in more detail, as well as the word embeddings and models involved, which were deployed to extend the application of the metrics to similarity assessment of texts across of different languages.

2.1 ConceptNet Numberbatch

For the calculation of *WMD* for multilingual articles, we deployed ConceptNet Numberbatch word embeddings (Speer and Lowry-Duda, 2017). ConceptNet Numberbatch are embeddings based on the knowledge graph ConceptNet (Speer et al., 2017). Due to its multilinguality (and support of all languages used in the challenge), we selected these embeddings to facilitate the calculation of the *WMD*.

2.2 WorldLex

Evaluating semantic similarity by matching words across texts can be obstructed by the presence of stop words or words, which can occur in many semantic contexts and are therefore no compelling indicators for semantic similarity. Apart from the removal of stopwords, we deployed a weighting of words according to the occurrence frequency in Twitter, blogs and newspapers gathered by Gimenes and New (2015), which are curated in the WorldLex database and were available for all languages used in the SemEval task. To create the WorldLex database, the authors converted all collected documents to lowercase. After that, the frequencies of all of the different words were calculated and lists of words were extracted utilizing spellcheckers to remove words with orthographic and typographic errors, as well as foreign words.

2.3 Word Mover's Distance

In 2015, the *Word Mover's Distance* has been proposed by Kusner et al. (2015) and constitutes a conceptually simple mean of quantifying the distance (and inversely correlated, the similarity) between texts, by optimizing a linear program for a cost minimal mapping between words of two texts with respect to a similarity measure between words. A text with n distinct words is considered as a vector $t \in \mathbb{R}^n$, with $||t||_1 = 1$ and t_i indicating the weight (from the WorldLex Database described in Section 2.2) of the ith distinct word of the text. The rationale of the *WMD* is, that for each word in a text, the weight has to be accounted for by words in the other text, while no word can account for more than its own weight. Considering two vector representations t^1 and t^2 of texts T_1 and T_2 of length n_1 and n_2 respectively, the distance between both texts Ω can be mathematically formulated as

$$\Omega = \min_{M} \sum_{i,j} M_{ij} c(i,j) \tag{1}$$

$$s.t. \sum_{j=1}^{n_2} M_{ij} = t_i^1 \,, 1 \leq i \leq n_1 \tag{2}$$

$$\sum_{j=1}^{n_1} M_{ji} = t_i^2 \,, 1 \leq i \leq n_2 \,, \tag{3}$$

where $c(i,j)$ is the distance between the ith distinct word of T_1 and the jth distinct word of T_2, and M is the accounting matrix, where M_{ij} indicates the amount of weight the ith word of T_1 provides for the accounting of the weight of the jth word of T_2. The resulting Ω can subsequently be used as measure correlated with smililarty or distance of text pairs.

Many options for choosing the weights of words, such as the document frequency of the word, are possible. We deployed a TF-IDF weighting of the words, where the inverse document frequency was derived from the *WorldLex* database introduced in the previous section. The rationale for introducing also the inverse document frequency, instead of only a term frequency weighting as suggested by Kusner et al. (2015), was to have a stronger emphasis of words, which are potentially more semantically meaningful and are therefore more suited for comparing themes and content of two texts.

For the quantification of distances between words, Kusner et al. (2015) propose distance measures such as cosine similarity on dense word embeddings, and use themselves word embeddings generated by *Word2Vec* (Mikolov et al., 2013). To facilitate the comparison of multilingual text pairs, we used multilingual *ConceptNet Numberbatch* embeddings (Speer and Lowry-Duda, 2017).

2.4 Sentence Mover's Distance

For our second line of experiments, we used the *Sentence Mover's Distance*, which was introduced by Clark et al. (2019), who adapted the concept of the *Word Mover's Distance* to calculate the distance of texts based on sentences instead of words, in order to address the typical shortcomings of approaches considering only Bag of Words without the incorporation of any compositional information contained in the word order. For the required representations of sentences, Clark et al. (2019) suggest averaging the word embeddings of the words in a sentence and a subsequent weighting of the sentences within the *Bag of Sentences* according to the number of constituent words.

Apart from the operating mode of the *SMD* proposed by Clark et al. (2019), the Sentence Mover's Distance allows also the incorporation of more rich representations of sentences, such as sentence embeddings produced by complex transformer-based models. Trained implementations of such models are readily available, such as the *Sentence-BERT* proposed by Reimers and Gurevych (2019). For the evaluation of the *Sentence Mover's Distance*, we deployed the pretrained *all-MiniLM-L12-v2* provided on `sbert.net`, which embeds an English sentence into a 384-dimensional embedding vector and supports the dot-product as similarity measure.

2.5 Neural Machine Translation

Since the used pretrained transformer can only embed English sentences, we propose to use English as an intermediary language, into which the multilingual news articles were automatically translated using a Neural Machine Translation system.

For the translation of the articles, we used a model submitted to WMT's 2021 News translation task (Chau et al., 2021). This translator has ranked first in the most language directions the team participated in, introducing the first multilingual model with stronger translation performance than bilingual ones. Using the *any-to-English* model and *DeepL*[1] (for the languages not supported by the first translation model), we translated all news articles

[1] `https://www.deepl.com/translator`

Dataset	Language en	pl	es	de	zh	ru	tr	fr	it	ar
Training	4048	663	1114	2198	-	-	903	144	-	541
Test	1487	555	1291	1536	1728	574	548	345	1127	589

Table 1: Number of used articles for each language in training and test data.

and used them to calculate BERT transformer embeddings. These in turn were used for the *SMD* calculation. We also evaluated possible performance differences for the *WMD*, if texts are translated beforehand, instead of relying on the power of the multilingual word embeddings of *ConceptNet Numberbatch*.

3 Experimental Setup

For preprocessing routines, the Python *nltk*-package was utilized (version 3.6.5). The embeddings have been conducted using *ConceptNet Numberbatch* version 19.08. The *TF-IDF* weighting used *WorldLex*. *WorldLex* is no longer under active development.[2] The transformer model was *all-MiniLM-L12-v2* provided on `sbert.net`.

We employed three experiments to find the best model used for submitting to the task's evaluation:

1. Word Mover's Distance with multilingual ConceptNet Numberbatch embeddings
2. Word Mover's Distance with translated news articles before embedding them using ConceptNet Numberbatch embeddings
3. Sentence Mover's Distance using translated news articles embedded with an English Sentence-BERT model

For the evaluation, the Pearson Correlation Coefficient (Pearson, 1895) was used. The target similarity scores range from one (very similar) to four (not similar), so these scores really correspond numerically to distances. We can thus use the computed distances directly as predictions. Since the calculation of the Pearson Correlation Coefficient normalizes all predicted and actual similarity scores using the mean and standard deviation, we do not need to scale the predicted scores to the range of the labels.

Due to the fact that our employed methods do not need any training labels, we can directly evaluate the results on the training data provided by the task organizers. The model found to perform best was then used for the task submission. The number of articles used for each language is listed in Table 1.

4 Results

Table 2 shows the Pearson Correlation Coefficient of the articles for the provided training data consisting of eight language pairs. Comparing the results of *WMD* applied to original and translated documents shows, that the estimation of text similarity with the proposed model benefits from a translation in a common language, instead of solely relying on a multilingual word embedding (all documents were translated into English). It can also be seen from the table, that for all language pairs except for *ar-ar* and *de-de*, *SMD* shows the best performance. As in the analysis of Kusner et al. (2015), we also find, that using the richer representations of complete sentences, instead of words, is generally beneficial. For the two remaining language pairs, *WMD*, applied on translated articles, achieves higher scores than the other metrics.

Given these results, we used the *Sentence Mover's Distance* with translated input texts and the pretrained language model for sentence embeddings as the model for our task submission. We achieved the 23nd place with an overall Pearson Correlation Coefficient of 0.57. Table 2 also shows the final scores for all 18 language pairs present in the test set. For the most language pairs, the results on the test data are very similar or even slightly better compared to the scores on the training set. We expected this behavior, since the used approach does not use any training labels for optimization and is hence not prone to overfitting. The rigidness of the optimization and straightforwardness of the similarity notion however, allows also for higher performance fluctuations observed for instance in the two language pairs (**en-en** and **de-de**), which perform noticeably worse for the test set than for the training data, while **pl-pl** shows much better performance on the test data.

Also, language pairs including German news articles are typically performing worse than average

[2]The deployed data can be retrieved from `http://www.lexique.org/?page_id=250`.

Model	Training Language Pairs en-en	de-en	es-es	fr-fr	tr-tr	ar-ar	pl-pl	de-de	Test ru-ru
WMD	0.77	0.60	0.67	0.68	0.72	0.55	0.36	0.54	—
WMD translated	0.77	0.66	0.68	0.69	0.76	**0.60**	0.42	**0.63**	—
SMD	**0.78**	**0.68**	**0.71**	**0.73**	**0.77**	0.59	**0.45**	0.59	—
Test Results SMD	0.68	0.69	0.71	0.70	0.71	0.65	0.59	0.31	0.56

	Testing Language Pairs zh-zh	es-en	it-it	pl-en	zh-en	es-it	de-fr	de-pl	fr-pl
Test Results SMD	0.56	0.77	0.69	0.69	0.66	0.65	0.40	0.47	0.75

Table 2: Pearson Correlation Score for all language pairs in the provided training and test data of the task evaluation. Since the *SMD* worked best on the training data overall, we used this method for the final submission.

and thus decrease the overall test score in the challenge. While looking for possible reasons for such poor results, we found some articles in training and test data, which could not be extracted correctly. For instance, some German articles from the test data are identical and contain only information about the issue of opening the web pages due to privacy regulations[3]. Another reason for bad performance in similarity estimation of pairs containing German article (or comparison of two German articles) could be the incompletely loaded content of some German pages, since some articles only contain the beginning of the actual text[4], which is indicated by the spontaneous termination of the article content with the following sentence “read the full article...”[5].
In order to eliminate such externally imposed assessment impediments, the data scraping system[6] would need to be overhauled.

5 Discussion and Conclusion

For our submission, we evaluated multiple approaches, that operate on the word or sentence level and calculate a distance between two texts using a linear program optimized on pretrained word or sentence embeddings. To be able to apply English-only models for the representation of sentences, we used a Neural Machine Translation system that, in our experiments, improved the performance of multilingual word embeddings. The proposed model shows stable performance in similarity estimation between mono- and multilingual document pairs. The usage of state-of-the art pretrained word and sentence embeddings led to a fast system with low computational cost, allowing implementation without use of graphics processing units. The use of an extensively pretrained *Sentence-BERT* transformer for sentence embeddings of documents, that were translated into English, confirmed, that the proposed model is well suited for the similarity comparison of multilingual articles without optimizing any parameters in the model.

[3]we found two groups of articles, each with the same content:
1.Group: 1586615494, 1490686353,1520406037,1524031333, 1525352422
2.Group: 1572312750, 1576180076,1611845398,1612866403 ,1617051090,1619154724,1627621567, 1551767123,1562891463, etc.

[4]articles with IDs 1488265289,1493242324, 1505316713,1516114270,1517039073, 1519376267,1531637961,1549821395, etc.

[5]translation of original German sentence: “Den vollständigen Inhalt lesen...”

[6]`https://github.com/euagendas/semeval_8_2022_ia_downloader`

Acknowledgements

We would like to thank all organizers of the SemEval-2022 Challenge. Our special thanks go to our supervisors from the Data Science Chair and Prof. Dr. Andreas Hotho for the given opportunity, computing resources and fascinating learning experience in natural language processing.

References

Tran Chau, Bhosale Shruti, Cross James, Koehn Philipp, Edunov Sergey, and Fan Angela. 2021. Facebook ai´s wmt21 news translation task submission. arXiv:1503.06733.

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.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover’s similarity: Automatic evaluation for multi-sentence texts. *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 2748–2760.

Manuel Gimenes and Boris New. 2015. Worldlex : Twitter and blog word frequencies for 66 languages. *Behaviour Research methods*, 48:963–972.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. *Proceedings of Machine Learning Research*, 37:957–966.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the royal society of London*, 58(347-352):240–242.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3973–3983.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada. Association for Computational Linguistics.