GateNLP-UShef at SemEval-2022 Task 8: Entity-Enriched Siamese Transformer for Multilingual News Article Similarity

Iknoor Singh, Yue Li, Melissa Thong and **Carolina Scarton**
Department of Computer Science, University of Sheffield (UK)
{i.singh, yli381, mlthong1, c.scarton}@sheffield.ac.uk

Abstract

This paper describes the second-placed system on the leaderboard of SemEval-2022 Task 8: Multilingual News Article Similarity. We propose an entity-enriched Siamese Transformer which computes news article similarity based on different sub-dimensions, such as the shared narrative, entities, location and time of the event discussed in the news article. Our system exploits a Siamese network architecture using a Transformer encoder to learn document-level representations for the purpose of capturing the narrative together with the auxiliary entity-based features extracted from the news articles. The intuition behind using all these features together is to capture the similarity between news articles at different granularity levels and to assess the extent to which different news outlets write about "the same events". Our experimental results and detailed ablation study demonstrate the effectiveness and the validity of our proposed method.

1 Introduction

News article similarity measures could facilitate various important tasks such as the clustering of news (Montalvo et al., 2007; Azzopardi and Staff, 2012), duplicate news detection (Alonso et al., 2013; Gibson et al., 2008; Singh et al., 2021a; Theobald et al., 2008), fact-checking (Hassan et al., 2017; Jiang et al., 2021) and tracking of the spread of news (Zhai and Shah, 2005). SemEval-2022 task 8 (Chen et al., 2022) assesses the similarity between news articles in terms of the real world happenings. Therefore, it mainly considers the location, time, entities and narratives, instead of writing style, political spin, or tone of the article. The training set comprises of monolingual and cross-lingual news articles pairs in 10 different languages.

The main contributions in this paper are: 1) We propose an entity-enriched Siamese Transformer whose general idea is to enable the model to explicitly learn from entity features (geolocation, organization, date and quantity) that are crucial to determine the similarity between news events but difficult to extract directly from the language models during fine-tuning. 2) We explore different data augmentation approaches through semi-supervised learning in order to tackle the imbalanced data problem. 3) We compare our proposed model with strong baselines and conduct an ablation study to analyse the contribution of each component in our model. We also present an error analysis at the end of this paper. Our best system which exploits Language Agnostic BERT Sentence Representations (Feng et al., 2020) ranks 2nd in the competition.

2 Background

The goal of the task is to predict similarity scores ranging from 1 (most similar) to 4 (least similar) for a give news article pair. The training data consists of 4,964 article pairs in seven distinct languages (English, German, Spanish, Turkish, Polish, Arabic and French), with seven groups of monolingual pairs and only one group of cross-lingual pair (German and English). The 4,902 pairs in test data feature three more languages (Chinese, Russian and Italian) and seven groups of new cross-lingual pairs. Appendix A.1 (Figure 3) shows the distribution of the train and test data. Most of the news pairs exhibit low level of similarity. However, the test set contains more pairs with the same news stories. The similarity based on location, time, entities, narratives, style and tone is also annotated respectively and given for the training data.

Previous studies on textual similarity have investigated different approaches based on corpus, knowledge or deep neural network (Chandrasekaran and Mago, 2021; Gomaa et al., 2013; Devlin et al., 2019; Reimers and Gurevych, 2019; Thakur et al., 2021a; Singh et al., 2021b), but they mainly experiment with short text pairs (e.g., the Semantic Textual Similarity (STS) benchmark (Cer et al., 2017), which is a sentence-level task). Mea-

1121

surement of the document similarity is arguably more challenging than that of short text since the information in the document is sparse and the model is easier to be misled by non-essential content. This paper tries to tackle the above mentioned challenges to build robust models for document similarity (Section 3).

3 System Overview

We propose an entity-enriched Siamese Transformer model which exploits multiple multilingual pre-trained Transformers (MPT) for multilingual news document similarity. The main sub-dimensions of similarity as annotated by the annotators are geolocation, time, shared entities and the shared narratives between the news articles (Chen et al., 2022)[1]. All this information is encoded in our model to capture different dimensions of news articles. Figure 1 shows the architecture of our proposed model and section 3.1 discusses its details.

3.1 Model Details

Reimers and Gurevych (2019) propose a Siamese and triplet network training methodology for the BERT-based (Devlin et al., 2019) models to derive semantically meaningful sentence embeddings. Instead of just learning the sentence representations, we use the Siamese (or bi-encoder) network architecture to learn document-level representations for the purpose of capturing the narrative of the article.

In this work, we test multiple MPT models as the backbone Transformer (Vaswani et al., 2017) encoder for our Siamese network and train the model using a regression objective function (Reimers and Gurevych, 2019). In this, the model is trained such that it forces contextual representations from similar documents close to each other in contrast to dissimilar documents by reducing the mean squared error loss between the overall similarity score and the cosine similarity between document representations. First, we concatenate title and news article text together to generate a single document. For the documents that lie beyond the maximum input sequence length for the model, we truncate their length and use the initial part of the document as a proxy for its fundamental narrative that is usually stated in the title and the lead paragraphs of news articles.

In an effort to explicitly capture other dimensions of the news article such as mentions of geolocation, event date, organisations and other named entities, we use SpaCy NER[2] to extract entities from the documents. For the non-English documents, we use their machine translated English versions to extract the entities (Fan et al., 2021).[3] All the extracted entities are aggregated into four different types of entities. The entity types along with their SpaCy labels (in brackets) are:

- **Geolocation (GEO)**: Location (LOC) and geopolitical entities (GPE) which includes mentions of countries, cities, states etc.

- **Organization (ORG)**: Name of organization (ORG), person (PERSON) and mentions of other important named entities (FAC, EVENT, NORP, PRODUCT, WORK_OF_ART) which can include names of airports, highways, sports events, religious or political groups etc.

- **Date (DATE)**: Date of publish from training set and date (DATE) and time (TIME) extracted from the news articles.

- **Quantity (QTY)**: Mentions of numerical quantity (QUANTITY) which can include both ordinal and cardinal measurements (ORDINAL, CARDINAL).

After extracting the list of all the above mentioned entities for both the documents, we lower-case all the entities and compute the cosine similarity to get a single similarity score which we use as feature for further training. The complete algorithm to compute similarity between entities is presented at Algorithm 1. We use the cosine scores of these four additional entity features along with the shared narrative score from Siamese Transformer to train a multilayer perceptron (MLP) with sigmoid activation to get the final news article similarity score. Overall, only the Siamese Transformer model is differentiable and all other are non-differentiable static features before the MLP layer. It is worth noting that both the Siamese Transformer and the MLP layer are trained separately and not jointly. The intuition behind using all these features together is to capture the similarity between news articles with different granularity and to assess the extent to which outlets write about same events.

---

[1]https://competitions.codalab.org/competitions/33835

[2]Pre-trained model en_core_web_trf is used for our experiments. Ref. https://spacy.io/models

[3]Pre-trained model m2m_100_418M is used for our experiments. Ref. https://github.com/pytorch/fairseq
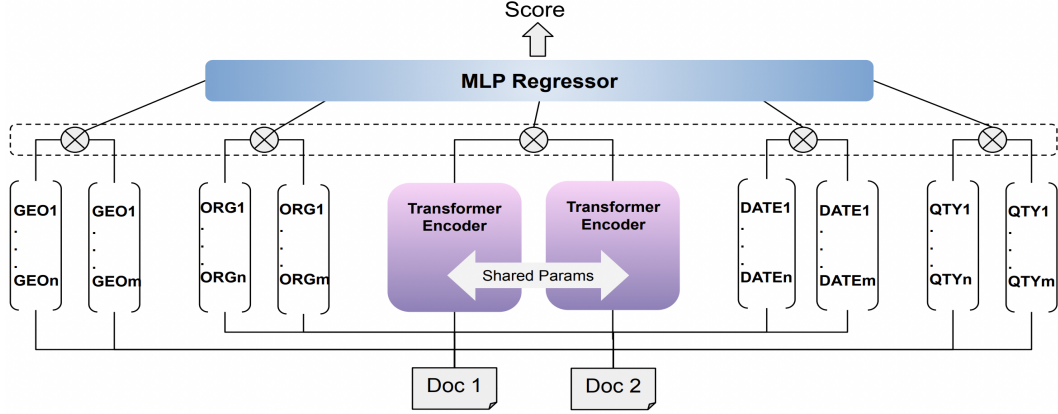
Figure 1: Entity-enriched Siamese Transformer model

Algorithm 1 Similarity between list of entities

function SIMILARITY(A, B)

     ▷ where A and B are list of entities

 Entities ⟵ Union (A, B)

 Product ⟵ Sum(A.count(ent)*B.count(ent) for ent in Entities)

 MagnitudeA ⟵ Sqrt(Sum(A.count(ent)^2 for ent in Entities))

 MagnitudeB ⟵ Sqrt(Sum(B.count(ent)^2 for ent in Entities))

 Score ⟵ Product / (MagnitudeA * MagnitudeB)

 return Score

end function

3.2 Semi-Supervised Learning

One of the issues with the training dataset is its imbalanced nature, i.e. there are considerably more document pairs with low similarity scores as compared to pairs with high similarity scores (Figure 3). In order to upsample the instances with high similarity, we first augment the training set and then adopt a semi-supervised training methodology to train the model. For this, we first explore randomly sampling document pairs and label them using our previous best model trained on the training set. However, we found that random sampling does not generate good quality augmented training data as most of the generated pairs still lie in the similarity range of [0.0,0.2]. Therefore, we employ the augmentation strategy similar to one proposed by Thakur et al. (2021b), i.e. BM25 sampling, for which we use the Elasticsearch[4] implementation of BM25 to generate augmented article pairs. How-

ever, our method is slightly different where we use the title of a news article as a query for retrieving other news article to get the document pairs. This helped the generation of documents pairs with similarity score that are more evenly distributed. A total of 59,943 additional document pairs were generated using the BM25 sampling and 2,845 document pairs were generated using machine translation. The complete details of the generation of augmented data is mentioned in Appendix A.2.

Semi-supervised learning (Zhu and Goldberg, 2009) is a widely known training paradigm where a model is first trained on a human labelled dataset and the model is further used to extend the training set by automatically annotating the unlabelled dataset. Following previous studies (Thakur et al., 2021b; Jurkiewicz et al., 2020), we initially start with training on the original training set and then for all the generated unlabelled document pairs, we use the previously trained model for inference to get the similarity scores for the new synthetic document pairs. Finally, we train our entity-enriched Siamese Transformer in a semi-supervised fashion on both the complete augmented training set.

4 Experimental Setup

The training set contain 4,964 article pairs, however, after removing pairs with empty documents or if one of the document has less than ten tokens, we get a total of 4,544 article pairs for training. For the experiments, the training set is split into train and development sets using a 80:20 split ratio. The data is split in a stratified fashion to keep equal proportion of all languages, except Arabic document pairs that only appear in the development set, keeping it as a unseen language for the training set We get a total of 3,635 document pairs for the training

---

[4]https://www.elastic.co/elasticsearch/

set and 909 document pairs for the development set. In the end, we train models on the complete train and development set for final evaluation on the test set consisting of 4,953 document pairs.

The training set has different types of labels which take into account different aspects of similarity among news articles (e.g., geolocation, narrative) (Chen et al., 2022). For our experiments, we use the "Overall" similarity as the true labels. However, the overall similarity label is in a range [1,4] where 1 signifies highest similarity and 4 means lowest similarity. In order to normalise these overall similarity label, we subtract all values from 4 to bring values in range [0,3], followed by min-max normalisation. The formula for this is as follows,

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} = \frac{4 - x}{3} \qquad (1)$$

where x denotes the overall similarity value and x_{norm} is the normalised value in range [0,1] from least similar to most similar. We use these as labels for training our models.

Since there are no baselines provided by task organisers, we use strong baseline methods for comparison. First, we use separate entity-based cosine similarity scores (Section 3.1) to find the Pearson correlation with the official test set labels. Second, we use Siamese multilingual BERT (Devlin et al., 2019), XLM-RoBERTa base variant (Conneau et al., 2020), Universal Sentence Encoder (USE)[5] (Yang et al., 2020) and LaBSE (Feng et al., 2020). Third, we also use the pre-trained MPNet[6] (Song et al., 2020) and fine-tuned it on the machine translated English training set. Please refer to Appendix A.3 for training details and the hyperparameters used in our experiments.

5 Results and Discussion

Table 1 shows the Pearson correlation coefficient scores on the official test data. The first part of the Table 1 presents the results of the baseline methods and the second part presents the results of our proposed model. For baselines, we find that if we just consider the entity-based features without training, organisation (ORG) achieves the highest Pear-

son correlation with the official test set. Furthermore, for the trained Siamese Transformer models, the LaBSE works best and the XLM-RoBERTa achieves the lowest score of around 0.70.

The second part of Table 1 shows the results of our entity-enriched Siamese Transformer trained on the augmented data using the semi-supervised learning paradigm. Training the Siamese Transformer on the augmented data brings significant improvements for all the models except USE when compared with the ones trained on the training set without augmentation. A potential explanation for LaBSE's impressive performance is the MLM and TLM pre-training objective on a dataset of 109 different languages, followed by training on a translation ranking task (Feng et al., 2020). We also test the Siamese Transformer both with and without the entity-enrichment in order to study how significant the improvements are statistically with entity features. Furthermore, we use Williams test (Graham and Baldwin, 2014) to test the statistical significance of increased correlation with the added auxiliary entity features. As shown in the Table 1, for all Siamese Transformers trained on the augmented data, entity-enrichment brings improvements in results to a statistically significant degree (p-value<0.01). We also find that the results of entity-enriched Siamese Transformers are statistically significant (p-value<0.01) when compared with the baseline Siamese Transformers for all the models. Overall, our entity-enriched Siamese LaBSE model trained on augmented data achieves the highest Pearson correlation of 0.80164.

Figure 2 presents the detailed analysis of the results for the entity-enriched Siamese LaBSE trained on augmented data model (best model). We observe that model performance varies in different language settings. The model performs the worst over German-French pairs (0.619), while it achieves the highest score on French-Polish articles (0.866). Overall, the model performs better on mono-lingual pairs than on cross-lingual pairs.

We further analyse the "serious mistakes" that our best model tends to make: giving high/low similarity scores for dissimilar/similar article pairs. We identify these instances based on the difference between the true similarity score and the predicted similarity score from the model. We examine the cases when the absolute values of the differences are larger than 2.0 (total 46 samples found), and observe that: 1) News pairs that cover different

---

[5]We employ a custom variant of USE model which supports 50+ languages. Ref. `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2`

[6]We employ a custom MPNet model which already fine-tuned on a 1 billion sentence pairs dataset with contrastive training objective. Ref. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

Experiments	Test Results	P-value
Baselines		
Geolocation (GEO)	0.26696	
Organisation (ORG)	0.45503	
Date (DATE)	0.43482	
Quantity (QTY)	0.35791	
Siamese mBERT	0.73069	
Siamese XLM-RoBERTa	0.70775	
Siamese USE	0.73324	
Siamese MPNet (English MT)	0.74536	
Siamese LaBSE	0.79187	
Proposed Method		
Siamese mBERT (Augmented data)	0.76649	
Entity-enriched Siamese mBERT (Augmented data)	0.76771	<0.0001
Siamese XLM-RoBERTa (Augmented data)	0.77669	
Entity-enriched Siamese XLM-RoBERTa (Augmented data)	0.77781	<0.0001
Siamese USE (Augmented data)	0.73320	
Entity-enriched Siamese USE (Augmented data)	0.73590	<0.0001
Siamese LaBSE (Augmented data)	0.80089	
Entity-enriched Siamese LaBSE (Augmented data)	**0.80164**	0.0022

Table 1: Pearson correlation coefficient score on the official test data. The best performance is in bold.
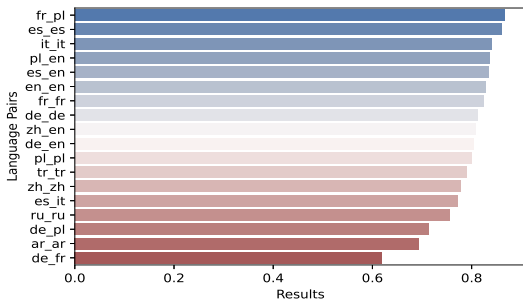

Figure 2: Performance over different language pairs in official test data. Average score of mono-lingual and cross-lingual pairs is 0.79833 and 0.78205 respectively.

stories around the same entities or topics are more challenging for the model. For example, the model outputs a similarity score of 1.77 for the following dissimilar news pair: an article about COVID-19 quarantine policy in Azerbaijan and another one regarding sanitary rules for preventing COVID-19 in Azerbaijan. 2) Web scraping could introduce noises in the model evaluation and obscure the true performance. We find that the scraping tool provided by the organizer fails to extract the real news content for all the Arabic news from Ahewar news website. The banner of the news website is returned instead. If we ignore the 25 pairs involving news from this website in the official test set, the

Pearson correlation coefficient score over the Arabic data increases significantly from 0.69425 (2nd worst in Figure 2) to 0.84067 (4th best). 3) The model tends to overestimate the degree of similarity. Among the 46 "serious mistakes", only 15 of them are similar news pairs with predictions indicating dissimilarity. And these wrong predictions are all caused by scraping-related issues (the real content of news is not returned). However, we argue that our entity features may potentially increase the robustness and decrease the level of overestimation. We compare the performance between models with and without entity features. The results show that the differences between true label and prediction decrease for 22 of the 31 dissimilar news pairs.

6 Conclusion

We introduce an entity-enriched Siamese Transformer for SemEval-2022 Task 8: Multilingual News Article Similarity. The error analysis shows that the entity-enrichment leads to statistically significant improvement and make models more robust for computing similarity between news articles in both monolingual and cross-lingual setting. Our entity-enriched LaBSE model achieves a Pearson Correlation of 0.802, ranking 2nd in the competition. The code is available at https://github.com/iknoorjobs/semeval-code.

7 Acknowledgement

References

Omar Alonso, Dennis Fetterly, and Mark Manasse. 2013. Duplicate news story detection revisited. In *Asia Information Retrieval Symposium*, pages 203–214. Springer.

Joel Azzopardi and Christopher Staff. 2012. Incremental clustering of news reports. *Algorithms*, 5(3):364–378.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37.

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

John Gibson, Ben Wellner, and Susan Lubar. 2008. Identification of duplicate news stories in web pages. Technical report, MITRE CORP BEDFORD MA.

Wael H Gomaa, Aly A Fahmy, et al. 2013. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Ye Jiang, Xingyi Song, Carolina Scarton, Ahmet Aker, and Kalina Bontcheva. 2021. Categorising fine-to-coarse grained misinformation: An empirical study of covid-19 infodemic. *arXiv preprint arXiv:2106.11702*.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Soto Montalvo, Raquel Martínez, Arantza Casillas, and Víctor Fresno. 2007. Bilingual news clustering using named entities and fuzzy similarity. In *International Conference on Text, Speech and Dialogue*, pages 107–114. Springer.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Iknoor Singh, Kalina Bontcheva, and Carolina Scarton. 2021a. The false covid-19 narratives that keep being debunked: A spatiotemporal analysis. *arXiv preprint arXiv:2107.12303*.

Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2021b. Multistage bicross encoder for multilingual access to covid-19 health information. *PloS one*, 16(9):e0256874.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021a. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021b. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Martin Theobald, Jonathan Siddharth, and Andreas Paepcke. 2008. Spotsigs: robust and efficient near duplicate detection in large web collections. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 563–570.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Yun Zhai and Mubarak Shah. 2005. Tracking news stories across different sources. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 2–10.

Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

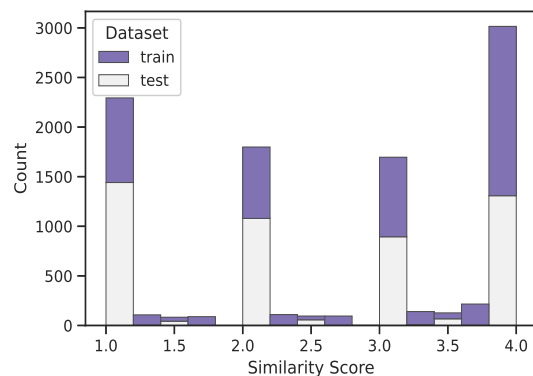# A Appendix

A.1 Data Distribution

Figure 3: Distribution of training and test data

A.2 Augmented Data Generation

The details of augmented data used in our experiments is as follows,

Augment Data 1. In this, we used the training set provided by organisers where the title of the article is used as a query to all the other articles in the training set that are indexed in Elasticsearch. If title or article are not in English, then the English machine translated version is used to perform the retrieval. Also, we exclude the cases of document pairs that are already present in the training set. We generated 25,125 by using the top 5 retrieved documents for each case.

Augment Data 2. We machine translated the English document pairs with overall similarity score in range of [0.5-1] to other languages. We only choose to translate document pairs that are English to get a good quality machine translated augmented training data. For this we randomly sampled five different language pairs from the list of language pairs that follow the same distribution as that of language pairs found in the test set and machine translated the documents[7] (Fan et al., 2021). We generated 2845 additional document pairs using this augmentation strategy.

---

[7]Pre-trained model m2m_100_418M is used for our experiments. Ref. https://github.com/pytorch/fairseq

Augment Data 3. Here, we utilised "All the news" dataset available on Kaggle[8] which contains 143,000 articles from 15 American publications. Here, title from the SemEval training set is used as query to all the articles in Kaggle new datasets. We generated 34,818 document pairs using the top 5 retrieved documents.

In our experiments, we use augment data 1, 2 and 3 which constitute a total of 62,788 additional document pairs and use them all together along with the training set to train the models.

A.3 Hyperparameters

The Transformer encoder of our Siamese Transformer (shared parameters) model is trained for 4 epochs with a batch size of 8, learning rate of 2e-5 and maximal input sequence length of 512. The output of model along with entity features is passed to the MLP layer which is composed of 32 hidden units trained with learning rate of 1e-3. The output of this MLP layer is used to get the final similarity score for the news articles. Adam (Kingma and Ba, 2015) is used to optimize the model parameters. For all the experiments, the model which works the best on the dev set is submitted and evaluated on the test set provided by the task organisers. All experiments are conducted on a machine with NVIDIA GeForce RTX 3090.

---

[8]https://www.kaggle.com/snapcrack/all-the-news