Team dina at SemEval-2022 Task 8: Pre-trained Language Models as Baselines for Semantic Similarity

Dina Pisarevskaya, Arkaitz Zubiaga
Queen Mary University of London, UK
dinabpr@gmail.com, a.zubiaga@qmul.ac.uk

Abstract

This paper describes the participation of the team "dina" in the Multilingual News Similarity task at SemEval 2022. To build our system for the task, we experimented with several multilingual language models which were originally pre-trained for semantic similarity but were not further fine-tuned. We use these models in combination with state-of-the-art packages for machine translation and named entity recognition with the expectation of providing valuable input to the model. Our work assesses the applicability of such "pure" models to solve the multilingual semantic similarity task in the case of news articles. Our best model achieved a score of 0.511, but shows that there is room for improvement.

1 Introduction

The Multilingual News Article Similarity Task[1] (SemEval 2022 Task 8) is designed as a shared task to encourage participants to build systems that check if a monolingual or cross-lingual pair of news articles belong to the same story (Chen et al., 2022). The task consists in providing a similarity score from 1 to 4 for a pair of news articles.

In this paper, we describe the participation of the "dina" team in the shared task. In our participation, we took an exploratory approach focusing on language features, while trying several multilingual language models as baselines. These language models had been pre-trained for semantic similarity and are "pure" (without any fine-tuning conducted for the task), which we enhance to use jointly with state-of-the-art packages for machine translation and named entity recognition.

Our best-performing submission to the task is based on a "pure" paraphrase-xlm-r-multilingual-v1 model combined with the presence of overlapping named entities and overlapping dates, which achieved a Pearson's correlation score of 0.511 in the final evaluation.

2 Related Work

Multiple prior SemEval semantic similarity tasks (2012-2017) focused on estimating the degree of semantic equivalence between two text fragments. In 2014 and 2015, two subtasks were proposed for semantic textual similarity in English and Spanish (Agirre et al., 2015). It was found that aligning words between sentences worked best for English, using features such as WordNet, word embeddings, or paraphrase databases. In 2016, a new cross-lingual subtask was added for English and Spanish. In 2017, participants were instructed to predict the degree of semantic similarity (namely, a continuous valued similarity score on a scale from 0 to 5) between monolingual and cross-lingual sentences in Arabic, English and Spanish (Cer et al., 2017). State-of-the-art deep learning models and feature engineered systems were implemented (Tian et al., 2017), and machine translation was widely used for cross-lingual and non-English setups, in order to convert two sentences into the same language. Based on the corpus of English tasks data (2012-2017), the STS benchmark was presented for training and evaluation[2]. Its extended version is often used to evaluate the performance of multilingual pre-trained models[3].

Currently, state-of-the-art methods are based on pre-train transformer language models, which are fine-tuned for downstream tasks. Multilingual pre-trained language models for 50+ languages are freely available[4], such as distiluse-base-multilingual-cased models,

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

[2] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

[3] https://www.sbert.net/examples/training/multilingual/README.html

[4] https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1196 - 1201

paraphrase-multilingual-MiniLM-L12-v2, paraphrase-multilingual-mpnet-base-v2, LaBSE (Feng et al., 2020), and other various models, such as paraphrase-xlm-r-multilingual-v1[5] (Reimers and Gurevych, 2019). All these models are applicable to the sentence similarity task.

Event clustering can be considered as a task that is close to the task of semantic similarity between news articles. Miranda et al. (2018) used similarity metrics and ranking for clustering documents into monolingual and cross-lingual story clusters. Linger and Hajaiej (2020) used multilingual DistilBERT and Sentence-BERT for multilingual document representation. In general, BERT-like models with different averaging and pooling are widely used for document representation in the clustering task too.

In the system proposed in our paper, we build on this line of research of leveraging large multilingual language models, which we further experiment with by incorporating machine translation and named entity recognition components.

3 Task and Data

The aim of the task was to check, at the document level, if a pair of news articles, which can be written in the same or different languages, provide similar information. The training set provided by the organisers includes 4,964 monolingual and cross-lingual pairs of news articles: 1800 English-English pairs, 857 German-German pairs, 577 German-English pairs, 570 Spanish-Spanish pairs, 465 Turkish-Turkish pairs, 349 Polish-Polish pairs, 274 Arabic-Arabic pairs, and 72 French-French pairs. For test data, the organisers added three new languages: Italian, Russian, Chinese.

For each entry in the dataset, different similarity scores are given for "Geography", "Entities", "Time", "Narrative", "Style" and "Tone", in addition to the main "Overall" similarity score. These scores are based on a 4-point scale, from most (1) to least (4) similar. The aim of the task is to predict the "Overall" score, but the participants can make use of the other auxiliary scores. For the final evaluation, the organisers proposed to use Pearson's correlation between the predicted similarity score and the gold "Overall" similarity score. Correlation scores for different article pairs are then averaged to compute the final score.

A script[6] provided by the task organisers enabled downloading news article pairs pertaining to train and test subsets. This script retrieves the URLs of news articles, using the Internet Archive[7] to support this retrieval.

4 Data Preprocessing

Upon downloading the news articles with the script provided by the organisers, we noticed that some texts were not correctly parsed: they contained only non-relevant texts (e.g. error messages), without the main content. Therefore, we removed such problematic text pairs from the training set, based on the manually collected list of strings and heuristic rules for them, mostly for English (e.g. texts that started with "Get full access to" or "TAKE A FREE TRIAL"); pairs with short texts (< 400 characters) that contained mostly metadata were also removed from the training set. Finally, our training set consists of 4,228 text pairs. From texts both in the training and test sets, we also applied heuristic rules based on the manually collected list of strings (mostly for English) to remove non-relevant text fragments (e.g. "Your browser does not support the audio element" or "Share this item on Twitter") from texts.

All texts (in the training and test sets) were split into sentences using the stanza[8] python package, which supports all the languages under consideration. We relied on the assumption that news texts usually contain the most important information at the beginning, while the remainder of the story contains other less important details, according to the 'inverted pyramid' principle (Pöttker, 2003). According to this principle, the first paragraph or sentences of a news article are generally expected to cover the main information addressing the who, what, when, where, and why of a story. Therefore, aiming to improve both the efficiency and effectiveness of our system, we only take the first 10 sentences of all non-English texts and we translate them into English using m2m100 models (Fan et al., 2021): 1.2B model[9] for the training set and 418M[10] model for the test set (the latter model was

[5] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

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

[7] http://web.archive.org

[8] https://github.com/stanfordnlp/stanza

[9] https://huggingface.co/facebook/m2m100_1.2B

[10] https://huggingface.co/facebook/

Model	Train set	En-En data
(1)		
4 sent	0.51	0.65
5 sent	0.51	0.67
3 sent merged	0.64	0.70
(2)		
4 sent	0.61	0.72
5 sent	0.61	0.73
3 sent merged	0.66	0.73
(3)		
3 sent merged	0.63	0.70

Table 1: Pearson's correlation scores for train set for pre-trained language models: LaBSE (1), paraphrase-xlm-r-multilingual-v1 (2), and distiluse-base-multilingual-cased-v1 (3). Scores are presented for the whole train set and only for the English-English pairs of the train set.

chosen, as it is faster). Given that similar events are expected to mention the same or related named entities, we used the spacy[11] python package on English texts (original and translated) for named entity recognition.

All the preprocessing, fine-tuning and evaluation tasks were performed using Google Colab and two NVIDIA GeForce GTX 1080 Ti Graphics Cards (22 GB RAM).

5 Description of Models and Experimental Setup

In order to detect similarity at the document level, we focused on sentence-transformer models. All similarity scores were calculated based on cosine similarity. During the development stage, we conducted all our experiments on the training set, with a held-out subset reserved for evaluation.

We tested different transformer models, including LaBSE (Feng et al., 2020) and paraphrase-xlm-r-multilingual-v1[12]. Based on the assumption that the main information about a story is expected to appear in the beginning of news articles, different setups for both models were taken: mean semantic similarity (cosine similarity) for all pairwise sentence-to-sentence similarities for the first 4 sentences in both articles (`4 sent`) and the first 5 sentences in both articles (`5 sent`); and the semantic similarity between the merged set of first 3 sentences from each article (`3 sent merged`). To come up with the best approach to submit to the shared task, we conducted three sets of experiments.

Experiment set 1. Table 1 provides performance scores for the different "pure" pre-trained language models, measured based on Pearson's correlation values between predicted and gold similarity scores. Among the sentence selection approaches, the best-performing setup is the one based on the merged combination of the first 3 sentences (`3 sent merged`). Among the pre-trained language models, the paraphrase-xlm-r-multilingual-v1 model performs better than the other one. This finding confirms the notion that LaBSE works less well in detecting similarity of sentence pairs that are not translations of each other other[13] (Reimers and Gurevych, 2020). Multilingual distiluse-base-multilingual-cased-v1[14] model, one of the models recommended[15] for semantic similarity tasks, was also included in experiments, namely, for the best setup (first 3 sentences merged), but performed worse than paraphrase-xlm-r-multilingual-v1 model.

Experiment set 2. In addition, we also conducted some baseline fine-tuning experiments. We took XLSum dataset as the biggest dataset with available news texts that contains most of the languages present in our train and test sets (Arabic, Chinese, English, Spanish, Russian, and Turkish are included; it does not include German, Italian, and Polish). XLSum consists of 1.35 million article-summary pairs from the BBC in 44 languages (Hasan et al., 2021). We selected a smaller sample of 98,697 news texts with as balanced representation of 6 languages as possible (15,000 or fewer random examples for each language). Unsupervised domain-specific fine-tuning of the paraphrase-xlm-r-multilingual-v1 model on the news dataset (1 epoch with MultipleNegativesRankingLoss) did not improve the results: Pearson's correlation score was 0.63 for all text pairs from the preprocessed

m2m100_418M

[11] https://spacy.io/

[12] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

[13] https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models

[14] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[15] https://www.sbert.net/docs/pretrained_models.html#multi-lingual-models

training set, and 0.71 for only English-English pairs. Masked Language Model approach (1 epoch, batch size 16) also yielded worse scores (0.41 and 0.51 respectively).

Experiment set 3. We also conducted more model fine-tuning experiments. For internal evaluation, 10% of the training set was selected as an internal test set (with a balanced representation of English, Spanish, German, and Polish languages), and training was made on the remaining 90% (internal train set). The models distiluse-base-multilingual-cased-v1 and paraphrase-xlm-r-multilingual-v1 were fine-tuned on the internal train set (with MultipleNegativesRankingLoss, 3, 5, 8 epochs with 5 epochs as the best option, the latter one better). The model distiluse-base-multilingual-cased-v1 was also fine-tuned on the aforementioned XlSum dataset part (3 epochs with MultipleNegativesRankingLoss). On the internal test set, the models ensemble of distiluse-base-multilingual-cased-v1 fine-tuned on train set and distiluse-base-multilingual-cased-v1 fine-tuned on news dataset, with addition of named entities intersection ratio, achieved the best score, but still lower than the baseline "pure" models.

Other models considered. Apart from the three main sets of experiments above, we also considered other options. Our experiments with baseline experiments with generative pre-trained models did not lead to competitive results. Experimental results with GPT-Neo 2.7B[16] on English texts yielded only 0.13 correlation score between the perplexity scores for the first sentences and the gold scores. Basic fine-tuning for the mt5-base model[17] (setups up to 5 epochs, the task was handled as a multi-class classification task with 7 labels) did not give relevant results (all texts were misclassified for the highest scores).

Other features considered. Another possibility we considered was to include the dates mentioned in the news articles into our model. This was based on the assumption that news articles that are related to each other are supposed to be published in similar dates. However, our attempts at parsing dates from the articles (using the dateparser and num2words python packages) led to noisy outcomes, so we ended up using a rule-based approach to extract months and days from texts in English (original or translated). We check if both texts contain the same months and days (handled as intersections with two different sets).

[16] https://huggingface.co/EleutherAI/gpt-neo-2.7B

[17] https://huggingface.co/google/mt5-base

6 Results

6.1 Analysis of Results

On the test set, among our submitted results, the setup leading to the best performance score was a "pure" paraphrase-xlm-r-multilingual-v1 model combined with named entities intersections and dates intersections (0.511 Pearson's correlation). This setup performed better than a "pure" paraphrase-xlm-r-multilingual-v1 model (0.502 Pearson's correlation) or a "pure" paraphrase-xlm-r-multilingual-v1 model combined with named entities intersections (0.508 Pearson's correlation), so the features helped improve the model's scores. Named entity and date intersection scores were added to the model's scores using rule-based coefficients selected manually.

The best model also performed slightly better than the ensemble of a "pure" paraphrase-xlm-r-multilingual-v1 model and a distiluse-base-multilingual-cased-v1 model fine-tuned on news (0.502 Pearson's correlation). It shows that more domain-specified models and more complicated fine-tuning techniques should be used for the task. Among the setups we considered, we observe that "pure" models can be deemed stronger baselines.

6.2 Error Analysis

We performed an error analysis to understand why our proposed models yielded moderate performance scores. We identified two main reasons that can inform future directions of our research in improving these models:

1. Text parsing errors in the test set: some texts include, or sometimes solely consist of, meta-information that was not excluded by rules. Our rules did not cover all cases for all languages, as they required language knowledge, and should ideally be further pre-processed to reduce the noise and improve model performance. Examples of these cases include:

- German: *"OK Wir setzen auf unserer Website Cookies und andere Technologien ein, um Ihnen den vollen Funktionsumfang unseres Angebotes anzubieten"*

- German: *"Um die Funktion unserer Website zu verbessern und die relevantesten Nachrichten und zielgrichtete Werbung anzuzeigen, sammeln wir technische anonymisierte Informationen über Sie, unter anderem mit Instrumenten unserer Partner. Ausführliche Informationen zur Datenverarbeitung finden Sie in den Datenschutzrichtlinien. Ausführliche Informationen zu den von uns genutzten Technologien finden Sie in den Regeln der Cookies-Nutzung und des automatischen Einloggens. Indem Sie „Akzeptieren und schließen" anklicken, stimmen Sie ausdrücklich der Verarbeitung Ihrer persönlichen Daten zu, damit das beschriebene Ziel erreicht wird. Ihre Zustimmung können Sie auf die Weise widerrufen, wie in den Datenschutzrichtlinien beschrieben."*

- Italian: *"Informativa Privacy Questo sito utilizza cookies per migliorare servizi ed esperienza dei lettori. Le informazioni raccolte dai cookies sono conservate nel tuo browser e hanno la funzione di riconoscere l'utente quando ritorna sul nostro sito web e aiutare il nostro team a capire quali sono le sezioni del sito ritenute più interessanti ed utili."*

2. Translation errors. It produced model "hallucinations" and repetitions, such as "Chief Executive Officer of the Ministry of Foreign Affairs and Foreign Affairs of the Ministry of Foreign Affairs" or "WASHINGTON-SANA WASHINGTON-SANA WASHINGTON-SANA" instead of the correct English translations. We noticed this to be the case particularly for articles originally written in Chinese and Arabic (translations of more than 30% of test samples in these languages contained such translation errors). Therefore, other translation approaches, i.e. using Google Translate API, might be used to obtain better English translations to detect named entities intersections.

Table 2 provides the best model setup scores, broken down by language. For this analysis, we selected only monolingual language pairs. The best results are for Spanish, French and English, while Arabic, German and Chinese yield the lowest scores. It can be caused by parsing errors, as well as by translation errors. Separate monolingual language models for these languages can be applied in further research, as well as models with better transfer learning techniques for these languages. In the future, further experiments could be conducted for the multilingual XLSum dataset, using different sampling techniques and sample size.

Language	Score	No. of pairs
French	0.692	111
Spanish	0.678	243
English	0.625	236
Turkish	0.610	275
Italian	0.573	411
Russian	0.530	287
Polish	0.512	224
Chinese	0.426	769
German	0.250	608
Arabic	0.154	298

Table 2: Pearson's correlation scores for paraphrase-xlm-r-multilingual-v1 model combined with named entities intersections and dates intersections, for different languages from the test set (monolingual text pairs).

7 Conclusion

This paper presents the participation results of our team "dina" in the Multilingual News Similarity shared task held as part of SemEval 2022. We tested a range of state-of-the-art pre-trained multilingual transformer models, which were further tested by incorporating features based on dates, machine translation and named entity recognition. Our best model achieved a Pearson's correlation score of 0.511. It can be considered as a moderate performance score with substantial room for improvement, based on performance scores from other participants in the task, where the best system achieved a score of 0.818. In future experiments on cross-lingual semantic similarity of news texts, we aim to focus on more sophisticated fine-tuning techniques for domain adaptation on a further preprocessed and cleaned dataset.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 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, Denver, Colorado. Association for Computational Linguistics.

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.

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.

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*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Mathis Linger and Mhamed Hajaiej. 2020. Batch clustering for multilingual news streaming. In *Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020, April 14th, 2020. Vol. 2593 of CEUR Workshop Proceedings*, pages 55–61, Lisbon, Portugal.

Sebastião Miranda, Artūrs Znotiņš, Shay B. Cohen, and Guntis Barzdins. 2018. Multilingual clustering of streaming news. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4535–4544, Brussels, Belgium. Association for Computational Linguistics.

Horst Pöttker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 191–197, Vancouver, Canada. Association for Computational Linguistics.