NLP_Team1@SSN at SemEval-2024 Task 1: Impact of language models in Sentence-BERT for Semantic Textual Relatedness in Low-resource Languages

Senthil Kumar B Aravindan Chandrabose Gokulakrishnan B Karthikraja TP

Department of Information Technology
Sri Sivasubramaniya Nadar College of Engineering
Chennai, Tamilnadu, INDIA
{*senthil*, *AravindanC*, *gokulakrishnan*2010598, *karthikraja*2010588} @*ssn.edu.in*

Abstract

Semantic Textual Relatedness (STR) will provide insight into the limitations of existing models and support ongoing work on semantic representations. Track A in Shared Task-1, provides pairs of sentences with semantic relatedness scores for 9 languages out of which 7 are low-resources. These languages are from four different language families. We developed models for 8 languages (except for Amharic) in Track A, using Sentence Transformers (SBERT) architecture, and fine-tuned them with multilingual and monolingual pre-trained language models (PLM). Our models for English (eng), Algerian Arabic (arq), and Kinyarwanda (kin) languages were ranked 12, 5, and 8 respectively. Our submissions are ranked 5^{th} among 40 submissions in Track A with an average Spearman correlation score of 0.74. However, we observed that the usage of monolingual PLMs did not guarantee better than multilingual PLMs in Marathi (mar), and Telugu (tel) languages in our case.

1 Introduction

Prior NLP work has largely focused on semantic similarity, a subset of relatedness, because of a lack of relatedness datasets. The first dataset for Semantic Textual Relatedness, STR-2022 was introduced by Abdalla et al., 2023, which has 5,500 English sentence pairs manually annotated using a comparative annotation framework, resulting in fine-grained scores. The semantic relatedness of two units of language is the degree to which they are close in terms of their meaning (Mohammad and Hirst, 2012). The linguistic units can be words, phrases, sentences, etc.

The most semantic similarity datasets were annotated using coarse rating labels such as integer values between 1 and 5 representing coarse degrees of closeness. These datasets suffer from issues arising due to the fixed granularity which intuits fuzzy boundaries between related and unrelated notions. The following subsection describes the difference between similarity and relatedness which is crucial in understanding the textual semantics.

1.1 Similarity versus Relatedness

As discussed in Abdalla et al., 2023, the following are the characteristics of similarity versus relatedness:

1. Two terms are considered semantically similar if there is a synonymy, hyponymy, or troponymy relation between them whereas for semantic relatedness, it's enough to have any lexical semantic relation at all between them. (example: money-cost is related whereas price-cost is similar)

2. All similar pairs are also related, but not all related pairs are similar. For example, `surgeon-scalpel`, and `tree-shade` are related, but not similar.

3. If units are sentences, then the similarity between sentence pairs exhibits paraphrase or entailment property whereas the relatedness does not support that property since it accounts for all of the commonalities that can exist between two sentences.

The analysis showed that the presence of proper nouns (PROPN), nouns, and other coarse-grained POS categories in a sentence pair impact semantic relatedness much more than any other POS. We evaluated the semantic textual relatedness of 8 languages (Algerian Arabic (arq), Moroccan Arabic (ary), Kinyarwanda (kin), Hausa (hau), Marathi (mar), Telugu (tel), English (eng) and Spanish (esp)) in Track A of SemEval Task 1: Semantic Textual Relatedness for African and Asian Languages (Ousidhoum et al., 2024b).

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1854–1859

2 Related Work

Similarity task is originally proposed to mimic human perception of the similarity level between word or sentence pairs. The first, word similarity dataset was collected in Rubenstein and Goodenough (1965), which consisted of 65-word pairs with human annotations. In general, the datasets consist of pairs of words (w1, w2) (or sentences) and human-annotated similarity scores S_h.

Abdalla et al. (2023) measured the semantic relatedness using Contextual versus Static embeddings and Unsupervised versus Supervised approach to sentence representation. In an unsupervised approach, the embedding of a sentence is derived from that of its constituent tokens. They used Word2Vec, GLoVe, and Fasttext static embeddings in unsupervised settings and the majority of the static embedding models failed to obtain better correlations with human annotation scores. The contextual embeddings from BERT and RoBERTa do not perform better than the Word2vec embeddings.

Finally, the supervised approach by finetuning the SBERT with the STR-2022 dataset captured high semantic relatedness and the Spearman correlation is 0.82 and 0.83 for BERT-based and RoBERTa-based respectively. The supervised approach using the SBERT framework by formulating a regression task leads to a better correlation score of 0.20 than the unsupervised approach.

This motivated us to use the SBERT framework to score the semantic relatedness between the pairs of sentences across 6 low-resource languages and English, and Spanish in the Track A dataset. We used 2 multilingual pre-trained language models ($LABSE, pp-mpnet-v2$) and language-specific monolingual LM for each of the languages. The following subsections describe the reason behind the selection of particular pre-trained LMs that are used in our models.

2.1 LaBSE

Multilingual pre-trained models such as mBERT (Devlin et al., 2019) and XLM-R (CONNEAU and Lample, 2019) have led to exceptional gains across a variety of cross-lingual natural language processing tasks. However, without a sentence-level objective, they do not directly produce good sentence embeddings.

Language-agnostic BERT Sentence Embedding (Feng et al., 2022) is a multilingual BERT embedding model, called LaBSE, that produces language-agnostic cross-lingual sentence embeddings for 109 languages. The model is trained on 17 billion monolingual sentences and 6 billion bilingual sentence pairs using MLM and TLM pre-training, resulting in a model that is effective even on low-resource languages for which there is no data available during training.

It was trained on parallel sentence pairs from 109 languages using a Siamese network based on the BERT architecture. The model's ability to support 109 languages makes it a powerful tool for multilingual applications and cross-lingual natural language processing tasks. This multilingual PLM is used across all the 8 models in our experiment.

PLM Type	Language Model
Monolingual	MahaSBERT, TeluguSBERT
	DziriBERT
Multilingual	Sentence-T5, LaBSE
	AfroXLMR, IndicSBERT
	pp-mpnet-v2

Table 1: Types of pre-trained LM

2.2 paraphrase-multilingual-mpnet-base-v2

This is based on the multi-lingual model of `paraphrase-mpnet-base-v2`, extended to 50+ languages by Reimers and Gurevych 2020. It uses a multilingual knowledge distillation method that allows extending existing sentence embedding models to new languages. It has achieved state-of-the-art performance on the paraphrase identification task on several benchmark datasets.

2.3 AfroXLMR

Alabi et al. (2022) proposed multilingual adaptive fine-tuning (MAFT) as a method for simultaneously adapting multilingual pre-trained language models (PLMs) on 17 of Africa's most resourced languages and three other high-resource languages widely spoken on the African continent to encourage cross-lingual transfer learning. This approach was more competitive than the AfriBERTa (Ogueji et al., 2021) pre-trained LM on various NLP tasks. We used this pre-trained LM for Kinyarwanda (kin) and Hausa (hau) languages.

2.4 IndicSBERT

The IndicSBERT exhibits strong cross-lingual capabilities and performs significantly better than

Pre-trained LM	English	Spanish
LaBSE	0.802	**0.68**
pp-mpnet-v2	0.805	0.63
sentence-t5-large	**0.824**	-
sentence-similarity-spanish-es	-	0.66

Table 2: Evaluation of Indo-European languages during development

Pre-trained LM	Marathi	Telugu
LaBSE	0.82	0.797
pp-mpnet-v2	0.77	0.747
IndicSBERT	0.58	0.61
MahaSBERT	**0.84**	-
TeluguSBERT	-	**0.811**

Table 3: Evaluation of Marathi, Telugu languages during development

alternatives like LaBSE, LASER, and paraphrase-multilingual-mpnet-base-v2 on Indic cross-lingual and monolingual sentence similarity tasks.

The authors Deode et al. (2023) proposed a simple strategy to train cross-lingual sentence representations using a pre-trained multilingual BERT model and synthetic NLI/STS data. This is the first multilingual SBERT model trained specifically for Indian languages. However, monolingual models are typically found to be performing better than multilingual ones. Hence publicly released monolingual SBERT models for 10 Indic languages. We used `MahaSBERT` for Marathi(mar), and `TeluguSBERT` for Telugu(tel) in evaluating the STR score in Track A.

2.5 DziriBERT

The Algerian dialect is mainly inspired by standard Arabic but also from Tamazight, French, Turkish, Spanish, Italian, and English. Thus the Algerian dialect has several specificities that make the use of Arabic or multilingual models inappropriate. To address this issue the authors (Abdaoui et al., 2022) collected more than one million Algerian tweets and pre-trained the first Algerian language model: DziriBERT.

DziriBERT is a BERT-based model for the Algerian dialect which was trained using the Masked Language Modeling (MLM) task. It handles Algerian text contents written using both Arabic and Latin characters. We used this model for evaluating the semantic relatedness score for the Semitic languages group - Algerian Arabic(arq), and Moroccan Arabic(ary).

3 System Overview

Given a human-annotated dataset for semantic textual relatedness, the participants are allowed to submit systems that have been trained using the labeled training datasets. Apart from that, the participating teams are also allowed to use any other publicly available datasets. We restrict the use of only the dataset provided by the task organizers so that the impact of pre-trained language models on Sentence Transformers can be analyzed for the semantic relatedness task across different low-resource languages. We used the plain vanilla SBERT architecture for fine-tuning with pre-trained LMs for text processing.

In our experiment, predicting semantic relatedness is treated as a regression task, where each sentence is represented as a vector. We use the cosine similarity between the vectors to predict their semantic relatedness, S_p, the Spearman score predicted by the system. Finally, the correlation between S_h, the Spearman score manually annotated by humans, and S_p is computed, and a higher correlation suggests good alignment with human annotations and a better embedding model. Usually, the Spearman correlation between the prediction and gold relatedness scores is used to measure the goodness of the relatedness predictions.

3.1 Dataset

The authors Ousidhoum et al. (2024a) presented SemRel2024 dataset - the first benchmark on semantic distance (similarity or relatedness) that includes low-resource African and Asian languages from five different language families. We used the sentence pairs of 8 languages from the dataset for Track A. Refer Ousidhoum et al. (2024b) to the dataset split size for training, development, and test instances for Track A. The dataset contains semantic relatedness scores for each of the pairs of sentences of 8 languages.

3.2 Training and Testing

During the development phase, only the training and development datasets are given to construct the model for each language. The training data is used to fine-tune the model and development data is used to evaluate the model performance. We report the results using the default hyperparameters

set in the sentence transformer. The PLMs are fine-tuned on training data using cosine similarity loss with batch size as 8, and number of epochs as 20. The official evaluation metric is the Spearman correlation between the predicted similarity scores and the human-annotated gold scores.

During the test phase, we combined the training data + development data to fine-tune the model, and the unseen test data was used to predict the semantic relatedness score. Models using various pre-trained LMs are evaluated using Spearman correlation during the development phase. The model with the maximum Spearman correlation score is used during the testing phase to submit our results. The table 1 lists the types of pre-trained LMs and the corresponding LMs used in our study.

4 Experimental Setup

We aim to focus on the impact of the Sentence-BERT deep neural network in semantic textual relatedness scoring tasks, and the benefit of multilingual/monolingual pre-trained LMs over the task especially for the low-resource languages.

4.1 SentenceBERT

Unlike BERT, SentenceTransformer or SBERT by Reimers and Gurevych (2020) uses a Siamese architecture, where it contains two BERT architectures that are essentially identical and share the same weights. It processes two sentences as pairs during training. This neural network architecture is appropriate for pair-wise semantic sentence tasks such as Sentence Textual Similarity (STS), Semantic Textual Relatedness (STR), Natural Language Inference (NLI), and paraphrase identification tasks. This network leverages the two BERT architectures in parallel to compute/score the similarity/relatedness of pair-wise sentences.

Consider a pair of sentences S1 and S2 that are to be fed into the network. Feed a sentence S1 to BERT A and S2 to BERT B in the SBERT network. Each BERT outputs pooled sentence embeddings u and v respectively. The cosine similarity between these two embeddings (u, v) is computed by using mean-squared error loss as the objective function. This outputs the regressive score between 0 to 1. This is the predicted semantic relatedness score by the model between a pair of sentences S1 and S2. We developed all the models using SBERT for each of the 8 languages (except for Amharic) in Track A.

Pre-trained LM	Algerian Arabic	Moroccan Arabic
LaBSE	0.58	**0.799**
pp-mpnet-v2	0.53	0.73
DziriBERT	**0.67**	0.64

Table 4: Evaluation of Semitic languages during development

Pre-trained LM	Kinyarwanda	Hausa
LaBSE	0.579	0.715
pp-mpnet-v2	0.58	0.67
AfroXLMR	**0.61**	**0.73**

Table 5: Evaluation of African languages during development

4.2 Evaluation during development phase

The train ad development split data for each of the languages are as mentioned in the Ousidhoum et al. (2024a). We used two multilingual pre-trained LMs: LaBSE[1] and paraphrase-multilingual-mpnet-base-v2[2] (in short `pp-mpnet-v2`) across all the models. The idea behind using multilingual PLM for all 8 languages is primarily to check the performance of MLM for semantic textual relatedness tasks in low-resource languages. Apart from that, language-specific monolingual pre-trained LMs are also used in each of the models. During the development phase, the model that scored the maximum Spearman correlation is selected and applied during the testing phase. The models developed for each of the languages along with the pre-trained LMs used and its score are discussed below.

The table 2 shows that `sentence-t5-large` [3] (Ni et al., 2022), a text-to-text model showed better performance for the `English` language. The model using `LaBSE` scored higher than the other multilingual and monolingual LM for `Spanish` during evaluation in the development phase.

Table 3 shows that the monolingual models such as `MahaSBERT`[4] and `TeluguSBERT`[5] perform well than the multilingual models. The interesting fact to note is that even the `IndicSBERT`[6] - one of the popular multilingual models pre-trained on 14 Indian languages, scored poorly than the

[1] sentence-transformers/LaBSE

[2] sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[3] sentence-transformers/sentence-t5-large

[4] l3cube-pune/marathi-sentence-similarity-sbert

[5] l3cube-pune/telugu-sentence-similarity-sbert

[6] ai4bharat/indic-bert

Language	Model	Predict	Rank	baseline	LM type	diff.
English (eng)	SBERT-T5	**0.8352**	12	0.83	**MultiLM**	+0.0052
Spanish (esp)	SBERT-LaBSE	**0.7045**	9	0.7	**MultiLM**	+0.0045
Marathi (mar)	SBERT-MahaSBERT	0.8711	10	0.88	MonoLM	-0.0089
Telugu (tel)	SBERT-TeluguSBERT	0.7889	17	0.82	MonoLM	-.0.0311
Algerian Arabic (arq)	SBERT-DziriBERT	**0.6226**	5	0.6	**MonoLM**	+0.0226
Moroccan Arabic (ary)	SBERT-LaBSE	0.7446	16	0.77	MultiLM	-0.0254
Kinyarwanda (kin)	SBERT-AfroXLMR	**0.7233**	8	0.72	**MultiLM**	+0.0033
Hausa (hau)	SBERT-AfroXLMR	0.6281	11	0.69	MultiLM	-0.0619

Table 6: Evaluation of our SBERT-based models during the test phase. Boldface highlights the score more or equal to the baseline

LaBSE and pp-mpnet-v2 generic multilingual LM. IndicSBERT is one of the regional multilingual LMs trained in Indian languages.

Similarly for the Semitic languages such as Algerian Arabic and Moroccan Arabic, DziriBERT[7] PLM performed better than the generic multilingual LM in the Algerian Arabic language as shown in Table 4. The DziriBERT was specifically pre-trained on Algerian dialects. For Moroccan Arabic, a model with LaBSE had a better score than the model using DziriBERT LM. As per our knowledge, we do not find any monolingual pre-trained LM for Moroccan Arabic that improves the score than the LaBSE. This is one of the major drawbacks of low-resource languages. The availability of good pre-trained LM for task-specific or generic purposes is scarce in low-resource languages.

For African languages, the performance of the model using AfroXLMR[8] pre-trained LM scored better than the other models using generic pre-trained LMs as shown in table 5. This indicates that the use of appropriate pre-trained LMs is more important for semantic relatedness tasks than the generic pre-trained multilingual language models.

5 Result

During the testing phase, we combined the training + development data as training data to fine-tune the model that yielded the maximum score during the development phase. Then the model is tested with the test dataset of the corresponding language. The predicted sentence relatedness score by the models is submitted as a result and is evaluated using the Spearman coefficient. The results are shown in the Table 6. It is evident from the table 6 that almost 4 models had reached a score equal to or above the baseline score which is highlighted using boldface. The difference between the baseline and the model prediction is indicated with the + and - sign. The difference in Spearman correlation value with '+' indicates the improvement whereas the '-' sign indicates the poor performance of the model.

SBERT-based models for the languages English (eng), Algerian Arabic (arq), and Kinyarwanda (kin) performed more than the baseline Spearman score. SBERT-LaBSE model for Spanish (esp) scored almost equal to the baseline system. Even though the monolingual models SBERT-MahaSBERT for Marathi and SBERT-TeluguSBERT for Telugu showed better performance during the development phase, failed to score above the baseline during testing in respective languages. Similarly, SBERT-based models trained using multilingual pre-trained LM for Moroccan Arabic (ary) and Hausa (hau) languages scored lesser than the baseline model in the test phase.

5.1 Conclusion

Table 6 depicts the impact of pre-trained language models (LM) in SBERT for the various low-resource languages. The usage of monolingual LM in Marathi (mar) and Telugu (tel) did not guarantee a greater performance than the baseline system. This shows the limitations of existing state-of-the-art monolingual pre-trained LM MahaSBERT, TeluguSBERT for the STR task.

Apart from that, the multilingual pre-trained LM such as LaBSE, AfroXLMR did not perform well for Moroccan Arabic (ary) and Hausa (hau) which are from Afro-Asiatic language family. This shows the existence of poor resources such as pre-trained LM in those languages. By default, the monolingual LM did not guarantee better performance than the multilingual pre-trained LM, especially for the low-resource languages.

[7] alger-ia/dziribert

[8] Davlan/afro-xlmr-large

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2022. Dziribert: a pre-trained language model for the algerian dialect.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi, and Raviraj Joshi. 2023. L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 154–163, Hong Kong, China. 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Saif M. Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.