YSP at SemEval-2024 Task 1: Enhancing Sentence Relatedness Assessment using Siamese Networks

Yasamin Aali
Alzahra University
yasamin.aali01@gmail.com

Sardar Hamidian
GWU
sardar@gwu.edu

Parsa Farinneya
University of Toronto
parsa.farinneya@mail.utoronto.ca

Abstract

In this paper we present the system for Track A in the SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian Languages (STR). The proposed system integrates a Siamese Network architecture with pre-trained language models, including BERT, RoBERTa, and the Universal Sentence Encoder (USE). Through rigorous experimentation and analysis, we evaluate the performance of these models across multiple languages. Our findings reveal that the Universal Sentence Encoder excels in capturing semantic similarities, outperforming BERT and RoBERTa in most scenarios. Particularly notable is the USE's exceptional performance in English and Marathi. These results emphasize the importance of selecting appropriate pre-trained models based on linguistic considerations and task requirements.

1 Introduction

Semantic relatedness is a fundamental measure in natural language processing, providing a detailed assessment of how closely related two pieces of text are on a semantic level. This metric has wide-ranging significance in NLP tasks such as information retrieval, question answering, and text summarizing, contributing to the understanding of textual connections and improving algorithmic performance.

The importance of determining meaning through quantifying relatedness has been acknowledged within linguistic discussions for many years. Automating the determination of connected meanings holds significant value across diverse applications like evaluating sentence representation methods or supporting question-answering systems as well as summarizing processes.

However, the application of measuring relatedness to non-English languages presents substantial challenges due to disparities in linguistic resources and annotated datasets. In contrast to English many languages lack comprehensive lexical or syntactic resources which creates challenges in inaccurately capturing semantic subtleties and establishing cross-language associations. Additionally, variations across different languages such as morphological, syntactic, and semantic also add complexity to developing language-independent models for expressing relatedness Overcoming these obstacles requires extensive research and effective resource development tailored to different linguistic contexts ensuring that measures offered by the semantic similarity can be effectively applied across different languages. Here are some examples of the score of semantic relatedness of two sentences in three languages:

English: *"You figure this out all by yourself, did you?*
did you find all this on your own?" Score: 0.88

Spanish: *"Jean Hebb Swank es una astrofísica conocida por sus estudios sobre agujeros negros y estrellas de neutrones.*
Bajo la supervisión de Steve Frautschi, obtuvo su doctorado en física en 1967." Score: 0.52

Kinyarwanda: *"East Africa's Got Talent ku nshuro yayo ya mbere u Rwanda ni kimwe mu bihugu byemerewe kuyitabira aho rwahuriye n'ibindi birimo Uganda, Tanzania na Kenya.*
Iya mbere ni umubano mwiza uri hagati ya Mali n'u Rwanda, ndetse u Rwanda ni kimwe mu bihugu bifite abapolisi bagiye kugarura amahoro muri Mali." Score: 0.09

The rest of this paper is structured as follows: Section 2 introduces the problem statement and provides a summary of related works. Sections 3 and 4 detail the system description and experimental setup, respectively. The evaluation results are outlined in Section 5, followed by our conclusion

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 959–963

Language	Train	Dev	Test
English	0.75	0.79	**0.82**
Amharic	0.63	0.57	**0.64**
Algerian Arabic	0.44	**0.53**	0.4
Spanish	0.65	**0.66**	0.64
Hausa	**0.39**	0.38	**0.39**
Kinyarwanda	0.27	0.1	**0.31**
Marathi	0.58	0.65	**0.69**
Telugu	0.61	**0.75**	0.64

Table 1: The table provides a summary of the model's performance across different languages, for train, dev and test set, highlighting its strengths and weaknesses. It mentions the highest scores achieved in English, as well as the performance in Marathi and Telugu compared to English.

in Section 6.

2 Background

2.1 SemEval Task Description

We perform our experiments on data from the first subtask (supervised) of task 1 of SemEval-2024 (Ousidhoum et al., 2024b). We used 5,500 samples with 8 language pairs in the official training set (Ousidhoum et al., 2024a). The goal of the task is to predict the semantic textual relatedness between sentence pairs in different languages. The similarity score of pairs of articles in the provided dataset ranged from 0 to 1, with higher scores indicating higher semantic relatedness. To address the challenges of measuring semantic relatedness in non-English languages, research efforts need to focus on expanding resources and developing language-specific models.

2.2 Related work

Previous approaches to semantic relatedness have been categorized into knowledge-based and corpus-based methods. Knowledge-based methods use lexical resources like WordNet (Miller, 1995) to measure definitional overlap (Lesk, 1986), term distance within taxonomies, and term depth as specificity measures, among others. Knowledge-based methods are widely used in NLP applications, including word sense disambiguation and automatic summarization. The sources of knowledge utilized in these methods encompass various elements such as fuzzy logic, domain knowledge, Knowledge Graphs, ontologies, The Wikipedia among others. WordNet is an English language lexicon that arranges ideas into a conceptual structure. Its purpose is to represent the meaning of English words by categorizing synonyms and various relationships, both taxonomic and non-taxonomic. While semantic similarity quantifies specific likeness, relatedness provides a broader measure that encompasses connectedness as well.

On the other hand, corpus-based measures utilize probabilistic approaches such as Latent Semantic Analysis (Landauer et al., 1997), Explicit Semantic Analysis (Gabrilovich et al., 2007) and Salient Semantic Analysis (Hassan and Mihalcea, 2009), to decode word semantics based on contextual information observed in raw text. Corpus-based methods involve statistically analyzing large text corpora to quantify semantic similarities employing distributional semantics principles that capture contextual information within the text itself. Among these approaches is Latent Semantic Analysis, which uses a singular value decomposition technique to minimize word co-occurrence pattern matrix dimensionality within a corpus successfully applied among various natural language processing tasks including text classification and information retrieval. Another notable method utilizes random projection for mapping words onto high-dimensional spaces with cosine similarity vectors; termed Random Indexing, it outperforms LSA in particular tasks showing utility across diverse NLP applications like word sense disambiguation. In addition, researchers are exploring techniques leveraging the web as a corpus leading toward a branch known as web intelligence.

Neural networks have gained increasing significance in evaluating semantic relatedness due to their ability to comprehend intricate connections and subtleties in meaning from extensive data. Traditional approaches for assessing semantic relatedness often depend on manually crafted features or knowledge-based methods, which may have limitations in capturing the complete spectrum of semantic relationships between words and sentences. In contrast, neural networks can undergo training using large datasets to acquire contextualized representations of words and sentences. These representations capture the subtle meanings typically overlooked by traditional methods, enabling neural networks to achieve cutting-edge performance on tasks related to semantic relatedness. Furthermore, neural networks can be fine-tuned for specific purposes, enhancing their precision and efficiency. Furthermore, recent advancements in deep learning and

neural network technology have demonstrated encouraging outcomes when it comes to measuring semantic relatedness. For example, several pretrained language models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) have been tailored for semantic relatedness applications with state-of-the-art performance achieved on standard benchmark datasets.

Semantic relatedness in neural networks has been greatly impacted by the emergence of transformers, leading to a significant transformation in natural language processing (Gagliardi and Artese, 2023). Unlike traditional methods relying on manual features and statistical models, transformers utilize self-attention to dynamically allocate attention across input elements, effectively capturing long-term dependencies within language data. Current approaches for semantic relatedness mainly involve using powerful models like transformers to encode sentences into embeddings and then computing their similarity score using metrics such as cosine similarity. These advancements in deep learning and neural network technology have enabled the development of powerful models that can accurately measure semantic similarity and relatedness between sentences, surpassing the capabilities of traditional methods.

3 System Overview

The measurement of semantic relatedness involves assessing the connection or correlation between words or phrases, regardless of their meanings. Recently, deep learning models like convolutional neural networks and recurrent neural networks have garnered attention in measuring semantic relatedness. These models excel at capturing intricate patterns and interdependencies in textual data, thereby enhancing performance across a range of natural language processing tasks. For instance, Siamese networks make use of CNN or RNN structures to compare embeddings and gauge the relatedness between sentences or short texts. Attention mechanisms have also been integrated into these models to improve focus on crucial semantic elements (Sharma, 2023).

Siamese architectures are effective because they use the same model to handle similar inputs, and makes it easier to compare sentence pairs and reduces the number of parameters that need training, requiring less data and making them less susceptible to overfitting (Ranasinghe et al., 2019).

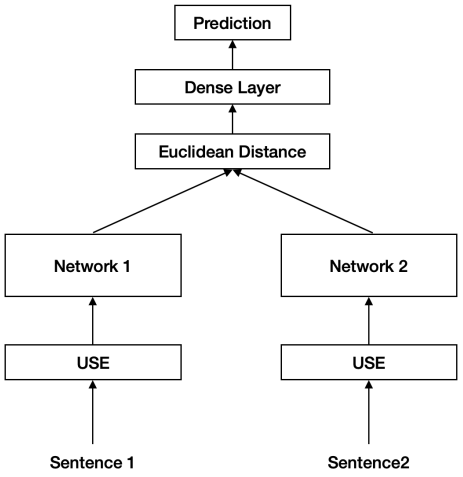

Figure 1: Basic structure of our system.

We implement a comprehensive system for evaluating sentence relatedness using a Siamese Network architecture coupled with Universal Sentence Encoder (USE) embeddings (Cer et al., 2018). The system comprises several distinct components, each contributing to the overall functionality and effectiveness of the model.

The Universal Sentence Encoder model is loaded from TensorFlow Hub [1], enabling the generation of high-quality embeddings for the input sentences. These embeddings capture semantic information, thereby facilitating the comparison and analysis of sentence similarity. Both the training and validation sentences are encoded into USE embeddings, which are then converted to NumPy arrays for seamless integration with the Siamese Network architecture.

The core of the system lies in the construction and training of the Siamese Network model (Figure 1). Utilizing TensorFlow's functional API, the model is designed to accept two input embeddings corresponding to pairs of sentences. It computes the Euclidean distance between these embeddings and passes the result through a Dense layer with sigmoid activation to predict the similarity score between the sentences. By employing Mean Squared Error (MSE) loss and the Adam optimizer, the model is trained on the training data, with performance monitored using the validation set.

Our code is available on GitHub [2]

[1] https://tfhub.dev/google/universal-sentence-encoder/4

[2] https://github.com/yasaminaali/Enhancing-Sentence-Relatedness-Assessment-using-Siamese-Networks

Models	English	Amharic	Alg Arabic	Spanish	Hausa	Kinyarwanda	Marathi	Telugu
USE	**0.75**	0.63	0.44	0.65	0.39	0.27	0.58	0.61
Bert	0.42	0.1	0.4	**0.5**	0.04	0.01	0.32	0.21
Roberta	0.38	0.2	0.31	0.46	0.01	0.1	**0.51**	0.24

Table 2: This table summarizes the key findings of the evaluation, highlighting the varied performances of different models across multiple languages. It specifically mentions the strengths of BERT in English and Spanish, RoBERTa's excellence in Marathi and Spanish, and the consistently exceptional results of the Universal Sentence Encoder across most languages in the training set.

4 Experimental Setup

4.1 Data Split

For all of our experiments, we split the task's training set using an 80/20 train/dev split, and we used the official development set as a test set.

4.2 Pre-processing

Pre-processing improves data quality, eliminates irrelevant information, and makes data more suitable for the calculation of the semantic relatedness score. This entails removing punctuation, numbers, and special characters such as # and $. Contractions are expanded to their full forms, and all text is converted to lowercase for consistency in processing.

4.3 Evaluation Metrics

The evaluation metric for task 1 is the Spearman Correlation between the predicted similarity scores and the human-annotated gold scores, with a range from 0 to 1 (from least to most correlated), which helps to determine how well the predicted scores align with human judgments.

5 Results

Our rankings show that in certain languages, our method closely matches the baseline performance. For example, in English, our rank is 0.82 compared to the baseline of 0.83, and in Spanish, our rank stands at 0.64 compared to the baseline of 0.7. Therefore, our method demonstrates significant effectiveness across different languages.

The evaluation of models across different languages showed varied performances. BERT demonstrated strong performance in English and Spanish, while RoBERTa excelled in Marathi and Spanish. However, the Universal Sentence Encoder consistently delivered exceptional results across most languages in the training set (Table 2).

Specifically, the USE model achieved the highest scores in English with 0.75 on the training set and 0.82 on the test set, indicating remarkable performance. The top overall score reached 0.86, showcasing its effectiveness in this task as well. Following English, Marathi exhibited the second-best performance at a score of 0.69 (Table 1).

5.1 Error Analysis

The model struggled with Kinyarwanda, Hausa and Algerian Arabic, hinting at relatively poor performance for this languages (Table 1).

To gain a deeper understanding of our model's performance, we compare its predictions with the test labels in English. Upon observation, it is evident that the model struggles when calculating the lowest semantic relatedness between sentences.

6 Conclusion

The system presented offers a strong framework for evaluating sentence similarity by integrating a Siamese Network architecture with Universal Sentence Encoder embeddings. A comprehensive overview of the system's components and processes, including data preprocessing, model construction, and training, demonstrates that the system effectively utilizes advanced techniques in natural language processing to make accurate similarity predictions. The evaluation results reveal the superior performance of the Universal Sentence Encoder across multiple languages, outperforming pre-trained models like BERT and RoBERTa in most scenarios. Notably, the system excelled in English and Marathi, demonstrating its versatility and effectiveness across diverse linguistic contexts. Further optimization and refinement of the system may enhance its performance in under-performing languages as well as broaden its applicability in real-world scenarios, ultimately advancing the field of natural language processing and facilitating a wide range of practical applications.

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Isabella Gagliardi and Maria Teresa Artese. 2023. Ensemble-based short text similarity: An easy approach for multilingual datasets using transformers and wordnet in real-world scenarios. *Big Data and Cognitive Computing*, 7(4):158.

Samer Hassan and Rada Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1192–1201.

Thomas K Landauer, Darrell Laham, Bob Rehder, and Missy E Schreiner. 1997. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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. *Preprint*, arXiv:2402.08638.

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.

Tharindu Ranasinghe, Constantin Orăsan, and Ruslan Mitkov. 2019. Semantic textual similarity with siamese neural networks. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1004–1011.

Kabir Sharma. 2023. 30 years of research on semantic similarity measurement.