scaLAR SemEval-2024 Task 1: Semantic Textual Relatednes for English

M Hemanth Kumar and **Anand Kumar M**
Artificial Intelligence, Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India
mogilipalemhemanthkumar@gmail.com,m_anandkumar@nitk.edu.in

Abstract

This study investigates Semantic Textual Related- ness (STR) within Natural Language Processing (NLP) through experiments conducted on a dataset from the SemEval-2024 STR task. The dataset comprises train instances with three features (PairID, Text, and Score) and test instances with two features (PairID and Text), where sentence pairs are separated by '/n' in the Text column. Using BERT(sentence transformers pipeline), we explore two approaches: one with fine-tuning (Track A: Supervised) and another without fine-tuning (Track B: UnSupervised). Fine-tuning the BERT pipeline yielded a Spearman correlation coefficient of 0.803, while without fine-tuning, a coefficient of 0.693 was attained using cosine similarity. The study concludes by emphasizing the significance of STR in NLP tasks, highlighting the role of pre-trained language models like BERT and Sentence Transformers in enhancing semantic relatedness assessments.

1 Introduction

Semantic Textual Relatedness (STR) is a crucial concept in natural language processing (NLP), focusing on determining the degree of similarity between linguistic units like words or sentences based on their meaning. This measure plays a vital role in evaluating the effectiveness of Large Language Models (LLMs) and aids in various NLP tasks. At its core, STR delves into understanding the closeness in meaning between two pieces of text. It examines different dimensions of relatedness, including sharing the same viewpoint, originating from the same context, or complementing each other's content. For instance, if two sentences convey similar ideas through paraphrasing or entailment, they might be considered semantically similar. However, relatedness en- compasses all possible commonalities between them. In NLP, researchers and practitioners leverage STR to enhance textual coherence, refine narrative structures, and tackle diverse language understanding challenges. By quantifying semantic relatedness, NLP systems can better comprehend and generate human-like responses, ultimately advancing the capabilities of language models.

The concept of semantic relatedness between language units has been recognized as foundational in understanding meaning. The automatic determination of relatedness has found numerous applications, including the evaluation of sentence representation methods, question answering, and summarization. Semantically similar sentences are those that exhibit either a paraphrasal or entailment relationship. In contrast, relatedness encompasses a broader spectrum of commonalities between two sentences. This includes considerations such as whether they pertain to the same topic, convey the same perspective, emerge from the same temporal context, or if one sentence elaborates on or logically follows from the other. Despite the significance of relatedness, much of the prior work in natural language processing has predominantly fo- cused on semantic similarity, particularly within the context of English.

We Explored SBERT. Sentence-BERT(Reimers and Gurevych, 2019) builds upon the architecture of BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2018), leveraging transformer-based models to encode contextual information from input sentences. Unlike BERT, which focuses on token-level representations, Sentence-BERT aims to generate fixed-size representations for entire sentences. To achieve this, Sentence-BERT employs siamese or triplet network architectures, which are trained on sentence pairs or triplets with similar or dissimilar semantic meanings. Through contrastive loss functions, Sentence-BERT learns to map semantically similar sentences closer together in the embedding space while pushing dissimilar sentences farther

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 902–906

apart.

2 Background

The exploration of semantic relatedness in language finds its roots in seminal works by (Halliday and Hasan, 1976) and (Miller and Charles, 1991), which laid early foundations for understanding the subtleties of meaning in text. Initially, these efforts primarily focused on semantic similarity, assessing the likeness between linguistic units through techniques like paraphrasing or entailment. However, as research progressed, scholars began recognizing the necessity of considering a broader array of connections between text segments, thereby giving rise to the concept of semantic relatedness.

Semantic similarity denotes the extent of resemblance in meaning between two linguistic units, while semantic relatedness encompasses a wider spectrum of connections, encompassing elements such as topical relevance, viewpoint alignment, temporal coherence, and logical sequence. While semantic similarity often relies on paraphrasing or entailment, relatedness factors in various nuances contributing to the overall coherence and cohesion of text.

Traditional methodologies for measuring semantic relatedness relied on lexical and syntactic features, including word overlap, syntactic parse trees, and semantic networks. However, the emergence of deep learning techniques has ushered in a paradigm shift towards leveraging neural embeddings and transformer-based models to capture richer semantic representations. These modern approaches have demonstrated superior performance across various Semantic Textual Relatedness (STR) tasks, such as semantic similarity estimation and semantic textual entailment.

In the field of natural language processing (NLP), assessing the relatedness between pairs of sentences is a fundamental task. The paper(Hany et al., 2023) addresses this challenge by proposing an innovative approach that combines two key techniques. First, the authors leverage embedding similarity techniques, utilizing seven different transformers to generate sentence vectors. These vectors capture the semantic content of sentences, allowing for more accurate relatedness assessment. Second, a classical machine learning regressor is trained on these sentence vectors. By integrating these methods, the study achieved impressive results on the SICK dataset. Specifically, the mean square error is reduced to 0.0481, and high Pearson's and Spearman's correlations of 0.978 and 0.9696, respectively, demonstrate the effectiveness of this approach. Overall, this research highlights the potential of combining embedding similarity techniques with machine learning for improving relatedness score assessment and advancing NLP algorithms. The zero-shot text classification (0SHOT-TC) has garnered significant attention. This task involves detecting classes that the model has never encountered during training. The emergence of pre-trained language models has transformed 0SHOT-TC into a binary classification problem, akin to textual entailment. Specifically, the model learns whether there is an entailment-relatedness (yes/no) between a given sentence (premise) and each category (hypothesis). However, existing approaches struggle with fully expressing the category space using labels or label descriptions. In contrast, humans can effortlessly extend a set of words to describe the categories to be classified. To bridge this gap, the paper(Liu et al., 2023) introduces a novel method called Semantically Extended Textual Entailment (SETE). Inspired by human knowledge extension, SETE enriches category representations using a combination of static knowledge (e.g., expert knowledge, knowledge graphs) and dynamic knowledge (e.g., language models).

Early methods, such as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), fell short in capturing nuanced word meanings and context. To address this, the paper(Abdalla et al., 2021) surveys the evolution of semantic similarity techniques, categorizing them into knowledge-based, corpus-based, deep neural network-based, and hybrid approaches. By examining the strengths and limitations of each method, the survey provides a comprehensive overview for researchers navigating the complex landscape of semantic similarity research. Understanding the degree of semantic relatedness between two language units is fundamental. However, prior research has primarily focused on semantic similarity, a subset of relatedness, due to the scarcity of relatedness datasets. To address this gap, the authors(Chandrasekaran and Mago, 2021) introduce the Semantic Textual Relatedness (STR-2022) dataset, comprising 5,500 English sentence pairs manually annotated using a comparative annotation framework. Human intuition regarding sentence relatedness proves highly reliable, with a repeat annotation correlation of 0.84. The dataset not only

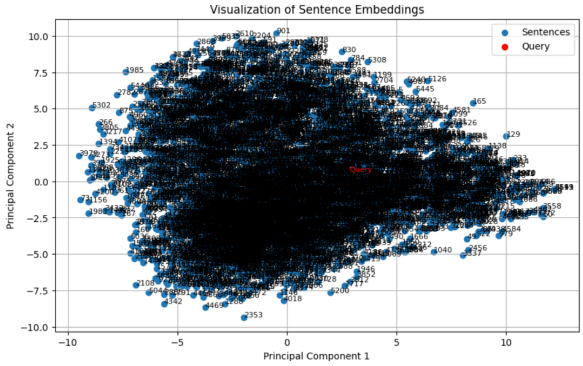

Figure 1: Visualization of Sentence Embeddings

facilitates exploration of what makes sentences semantically related but also serves as a valuable resource for evaluating automatic sentence representation methods and various downstream NLP tasks.

3 System Overview

A model in the Sentence-Transformers library is the bert-base-nli-mean-tokens collection. It is especially made for the purpose of Semantic Textual Similarity (STS). Sentences and paragraphs are mapped to a 768-dimensional dense vector space by this paradigm. The pre-training phase Bert-base-nli-mean-tokens are pre-trained using the conventional BERT architecture. This model is pretrained on the tasks of Modeling Masked Languages (MLM): Tokens in the training data are masked with a unique token [MASK] or randomly substituted with a small percentage, Bidirectional Contextualization: By analyzing both left and right context, BERT generates bidirectional contextualized embeddings as it learns to forecast masked tokens and Next Sentence Prediction (NSP): In the original text, BERT further forecasts if two sentences will come after one another. In learning sentence relationships, this aids the model. Fig -1 shows the visualization of Sentence embeddings from this model.

4 Experimental Setup

we considered the dataset(Ousidhoum et al., 2024a) from codalab SemEval-2024 STR task (Ousidhoum et al., 2024b). There are 5500 samples as train instances and three features namely PairID, Text ans Score . There are 250 samples as test instances and two features PairID and Text. In the Text column of the data, the pair of sentences are separated by '/n'. We conducted a couple of experiments

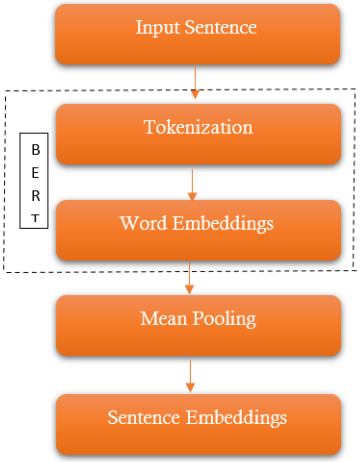

Figure 2: Methodology Used

for the Semantic Textual Relatedness on the above mentioned data using BERT. In our study, we utilized the bert-base-nli-mean-tokens model, a variant of the BERT architecture tailored for the Semantic Textual Similarity (STS) task. This model, an integral component of the Sentence Transformers (sbert) library, plays a pivotal role in assessing the semantic relatedness between pairs of text sentences. The designation 'base' signifies a medium-sized version of the BERT model, balancing computational efficiency with performance. Pre-trained on Natural Language Inference (NLI) tasks, the model captures intricate semantic relationships between text segments, essential for STS tasks. Additionally, employing mean pooling, it generates fixed-length sentence embeddings efficiently, providing a comprehensive representation of semantic content. Our utilization of bert-base-nli-mean-tokens in our research ensures robust and nuanced analysis of semantic similarity, contributing to advancements in natural language understanding and related fields.The fig- 2 shows the methodology we followed to extract sentence embeddings.

4.1 Without Fine Tuning

At First, we separated the two sentences by the delimited '/n'. we have the pair of sentences. Now, our task is to get the sentence embeddings for these sentences. We used Sentence Transformers pipeline, with the "bert-base-nli-mean-tokens" as the model.This particular model is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture and is trained to generate sentence embeddings by taking the mean of the token embeddings.These sentence

embeddings are used to calculate Semantic Textual Relatedness by using custom defined Cosine Similarity function. The custom function computes the cosine similarity between two input vectors u and v(sentence Embeddings) . Utilizing NumPy's dot product and Euclidean norm functions, the function calculates the cosine similarity by dividing the dot product of the input vectors by the product of their Euclidean norms. Commonly employed in Natural Language Processing tasks, cosine similarity serves as a fundamental metric for comparing the semantic similarity between word embeddings or Sentence embeddings, facilitating various applications such as information retrieval and document clustering.
Every word in the text was mapped to a word embedding space using the model. The cosine distance between the two sentences was computed after the embeddings. Equation 1 illustrates how the cosine similarity between the two embedding vectors is computed.
$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2}$

The requested forecast for the two sentences under consideration was then given as the cosine similarity value.

4.2 Fine Tuning

Similar to the previous experiment, we separated the two sentences by the delimited '/n'. We used the same pipeline to generate sentence embeddings. But this time, instead of directly evaluating the performance of the model using cosine similarity. We first fine-tuned the model with the following parameters Table 1 and then we evaluated it's performance.

Table 1: Parameters Used

Parameter	Value
Batch size	16
Epochs	1
Loss	CosineSimilarityLoss
Optimizer	Adam

5 Results

On Google Colab, we put our approach into practice. Sentence Transformer was the library that we used. Pythorch7 (>=1.11.0) and Python 6 (>= 3.8) are required by the library. The Official Competition website provides the dataset that was provided for each phase. The Spearman rank correlation coefficient, which assesses how closely the rankings predicted by the system match human assessments, is the official evaluation statistic for this activity. The GitHub page8 dedicated to the competition has the assessment script for this common job, which offers a uniform process for rating the effectiveness of competing solutions. The Spearman correlation coefficient can be calculated using the formula found in
$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

where n is the number of samples and d is the pairwise distances between the ranks of the variables xi and yi.

Table 2: Performance of Models Used

Model	BERT-BASE
Supervised	0.803
UnSupervised	0.693

Results for our experiments are shown in the table-2. It is important to remember that the bert-base-nli-mean-tokens model is no longer in use because of its poorer sentence embeddings.But In Supervised Approach, after finetuning the model with training data , the sentence embeddings are more meaningful as shown in fig-1. We were able to achieve better results than baseline in Unsupervised Approach using this model. In Supervised Approach , the baseline score is 0.830 and our proposed approach score is 0.803, with further improvements to our approach, we might achieve better results than baseline.

6 Conclusion

In order to address Task 1 at SemEval-2024, this paper presents the use of a BERT-BASE model embedding. For our submission, we chose a Supervised (finetuning) and unsupervised (not finetuning) approach, utilizing pre-trained Transformers that are already tailored to the domain. Based on this strategy, we used the contextual embeddings generated by the Sentence Transformer and used cosine similarity to measure the similarity between pairs of sentences, thereby quantifying the similarity between them. Although our method was successful, there is still room for improvement, as evi-

denced by the final ranking. Possible alternate approaches include utilizing the zeroshot capabilities of models like GPT , increasing the training data size by adding more datasets. There is some space for improvement in our straightforward method when compared to the top-performing models. It is noteworthy, nonetheless, that the assignment could be completed with a reasonable computational cost and no further pre-training thanks to Google Colab's free online tools.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M. Mohammad. 2021. What makes sentences semantically related: A textual relatedness dataset and empirical study. *arXiv preprint arXiv:2110.04845*.

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

Jacob Devlin et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

M A K Halliday and R Hasan. 1976. *Cohesion in English*. Longman.

Mena Hany et al. 2023. Enhancing textual relatedness assessment with combined transformers-embedding similarity techniques and machine learning regressors. In *Intelligent Methods, Systems, and Applications (IMSA)*. IEEE.

Tengfei Liu et al. 2023. Zero-shot text classification with semantically extended textual entailment. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.