NLP_STR_teamS at SemEval-2024 Task1: Semantic Textual Relatedness based on MASK Prediction and BERT Model

Lianshuang Su
School of Information
Science and Engineering,
Yunnan University
su_sls@163.com

Xiaobing Zhou
School of Information
Science and Engineering,
Yunnan University
zhouxb@ynu.edu.cn

Abstract

This paper describes our participation in the SemEval-2024 Task 1, "Semantic Textual Relatedness for African and Asian Languages ." This task detects the degree of semantic relatedness between pairs of sentences. Our approach is to take out the sentence pairs of each instance to construct a new sentence as the prompt template, use MASK to predict the correlation between the two sentences, use the BERT pretraining model to process and calculate the text sequence, and use the synonym replacement method in text data augmentation to expand the size of the data set. We participate in English in track A, which uses a supervised approach, and the Spearman Correlation on the test set is 0.809.

1 Introduction

We participated in the English language of track A in Task 1, "Semantic Textual Relatedness for African and Asian Languages." Track A uses a supervised approach where systems are trained on labeled training datasets. This task detects the degree of semantic relatedness between pairs of sentences for African and Asian Languages (Ousidhoum et al., 2024b).

Semantic Textual Relatedness (STR) is an important measure of the relationship between texts. It is considered to be the basis for understanding meaning(Miller and Charles, 1991) and is crucial for many natural language processing tasks. By computing semantic textual relatedness, we can perform applications such as text matching(Xu et al., 2013), information retrieval(Wagh and Kolhe, 2011), text categorization(Alsamurai, 2017), and question answering systems(Das and Saha, 2022).

However, previous NLP work has focused on semantic similarity (a small subset of semantic relatedness), in large part due to the lack of datasets on relatedness. For example, SemEval-2015 task1 is paraphrase and semantic similarity in twitter(Xu et al., 2015). And SemEval-2016 task1 is semantic textual similarity, monolingual and cross-lingual evaluation(Agirre et al., 2016).

Semantic relatedness and semantic similarity are two ways to explore the closeness of meaning. Two terms are considered semantically similar if there is a synonym, contextual, or modal relation relationship between them. Two terms are considered semantically related if there is any lexical semantic relation between them. Thus, all similar pairs are also related, but not all related pairs are similar(Abdalla et al., 2021). In semantic textual relatedness, we focus on the meaning and semantic information of the text, not just the surface word or sentence structure. Thus, the semantic relatedness between two texts can relate to their themes, intentions, emotions, etc.

The semantic relatedness of texts can be computed using the content and links of hypertext encyclopedias(Yazdani and Popescu-Belis, 2013). Semantic relatedness between texts can also be measured by calculating the similarity between text representations using a pre-trained language model.

In the following, we describe in detail the methods we used and give the evaluation results and conclusions.

2 Background

In this section, we present important details about the task setup. Each instance in the train set, dev set, and test set is a sentence pair, and these two sentences are separated by a newline character. The instance is labeled with a score representing the degree of semantic textual relatedness between the two sentences(Ousidhoum et al., 2024a). As shown in Table 1, there are two sentence pairs examples to present the semantic textual relatedness.

The scores can range from 0 (maximally unrelated) to 1 (maximally related), which are obtained using a comparative annotation framework. The

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 337–341

sentence1	sentence2	STR score
A girl is communicating with sign language.	A young girl is using sign language.	0.83
You should have respect for your mother.	Even if this is your own mother!	0.41

Table 1: Sentence pairs examples

	Train	Dev	Test
before text data augmentation	5500	250	2600
after text data augmentation	11000	250	2600

Table 2: Size of the data set

train and dev sets give sentence pairs and semantic textual relatedness scores, and the test set only gives sentence pairs. The train set was enlarged by using text data augmentation. The size of the dataset is shown in Table 2. The task we participated in was the English in track A. The task is a regression task whose input is a sentence pair and the output is the semantic textual relatedness score for that sentence pair.

3 System Overview

In this section, we present our approach applied to the task of predicting STR. We use the BERT pre-training model(Devlin et al., 2018) for text sequence processing and computation, and also employ text data augmentation to improve the training results. We adopted prompt tuning(Liu et al., 2023) to construct a new sentence, "The correlation of the next two sentences sent0 and sent1 is [MASK].", and used this constructed new sentence as a prompt template, where [MASK] is used to predict the correlation between the two sentences.

3.1 Model

We use the BERT pre-training model, designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both the left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. STR tasks are related to the semantics of the text, so using the case-insensitive English BERT pre-training model works better than the case-sensitive English BERT pre-training model.

We use the BERT model for encoding and feature extraction of text sequences. The structure of the system is shown in Figure 1(Mutinda et al., 2021). The two special tokens [CLS] and [SEP] are added to the model's input data to convert the text into the format expected by BERT. The forward function accepts a batch as input. It extracts $input\_ids$ and $attn\_mask$ from the batch, where $input\_ids$ is a sequence that converts the input text into a numeric representation acceptable to the model, $attn\_mask$ is a sequence of binary masks used to indicate which tokens are real input and which tokens are padded. Then it encodes the $input\_ids$ and $attn\_mask$ to obtain the $enc\_outputs$, which are hidden states of the model's output. Next, the corresponding embedding representation is extracted from the hidden state based on the mask position. These embedding representations are processed through a linear transformation to end up with a scalar value $logits$. Sigmoid activation is performed on $logits$ to get the output score.

3.2 MASK Prediction

Since the labels in the train set are continuous, we modeled this task as a regression problem. We adopted prompt tuning and used the Pattern-Exploiting Training (PET) method(Schick and Schütze, 2021) to construct a new sentence "The correlation of the next two sentences sent0 and sent1 is [MASK]." as a prompt template. In this prompt template, sent0 and sent1 are two sentences, and [MASK] is used to predict the correlation between the two sentences. Thus, it could convert the downstream task into a Complete Fill-in-the-Blank(cloze) task(Ding et al., 2021), and Masked Language Modeling(MLM)(Wettig et al., 2023) BERT can be used for prediction. Since the language of our participation is English, this prompt template is constructed in English. If we want to evaluate the semantic textual relatedness in other languages, we need to modify this template to the corresponding language. The constructed prompt template is fed into the model using the pre-training

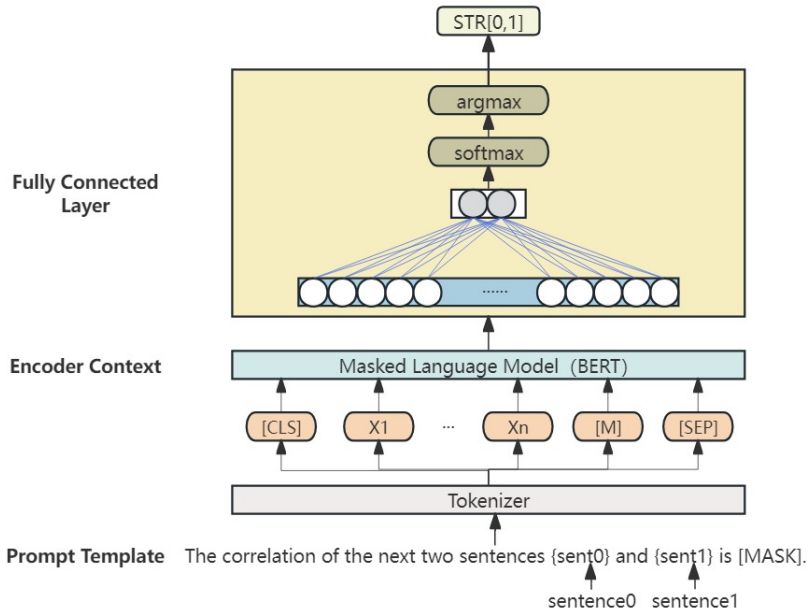

Figure 1: System model structure

model BERT, and the model predicts the representation of the correlation based on the context and the position of [MASK].

3.3 Text Data Augmentation

Through text data augmentation, more training samples can be generated to expand the size of the train set. Besides, text data augmentation can improve the generalization ability and robustness of the model. For example, Connor Shorten and others used a CNN model combined with text data augmentation EDA when the training set size was 5000, and the result improved from 87.7 to 88.3(Shorten et al., 2021). This task is to find out the degree of semantic correlation between two sentences, considering the task requirements and data characteristics, the data samples after performing text data augmentation can change the expression of the sentences, but the overall semantics of the sentences should remain unchanged.

Therefore, we used the synonym replacement method in the text data augmentation method in our experiments instead of random insertion, deletion, and other methods. After using this method changes the number of samples in the train set from 5500 to 11000.

4 Experimental Setup

The data set is given in CSV file format by the SemEval 2024 shared task organizer. It has three columns: PairID, Text, and Score, where Text is a sentence pair. We take out the two sentences in the sentence pair and use these two sentences to construct a new sentence: "The correlation of the next two sentences sent0 and sent1 is [MASK].". This new sentence is then fed into the model for processing and training. When performing text data augmentation, we replace the two sentences with synonyms and then insert the newline character in the middle of the replaced sentence pairs to ensure that the data format is consistent with the original data set.

We use the BERT pre-training model to process and calculate text sequences. The text data augmentation method is synonym replacement. Since this task is a regression task, we use the mean squared error(MSE) loss function:

$$L_{mse} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{'})^2 \quad (1)$$

where y_i is the ground truth for sample i, $y_i^{'}$ is the prediction score for sample i, n is the number of samples.

The batch size is set to 64, the number of training iterations is 6, and the learning rate is 2e-2. At the same time, in order to help the model converge better and achieve better performance, we set up a learning rate scheduler.

	Dev Score	Test Score
before text data augmentation	0.819	0.820
after text data augmentation	0.832	0.809

Table 3: Score on the dev set and test set

5 Results and Analysis

5.1 Results

This section shows the results of our system on the English STR task in track A of SemEval-2024 task 1. We use the Spearman correlation between system output and human annotation as an evaluation metric. Under the premise that other conditions are the same, we use the data set after text data augmentation for training. As shown in Table 3, the Spearman Correlation obtained on the dev set increased from 0.819 to 0.832, but the Spearman Correlation obtained on the test set dropped from 0.820 to 0.809.

5.2 Analysis

As shown in the results, after text data augmentation, the Spearman Correlation obtained on the dev set has improved, but the Spearman Correlation obtained on the test set has declined. Because before text data augmentation, the dev set was around 4.5% size of the train set and the test set was around 47% size of the train set. The size gap between the data sets is large. In addition, relying solely on semantic synonym replacement in sentences for data augmentation will have certain inaccuracies which leading to biased estimates. At the same time, text data augmentation doubled the size of the train set, resulting in a larger difference in the size of the train set, dev set, and test set.

6 Conclusion

This paper describes our participation in the SemEval 2024 competition in the Semantic Textual Relatedness for African and Asian Languages task. We participated in the English task in track A. Our approach is to use the BERT pre-training model for text sequence processing and computation, employing text data augmentation to enlarge the size of the train set, and adopting prompt tuning to construct a prompt template "The correlation of the next two sentences sent0 and sent1 is [MASK].", where [MASK] is used to predict the correlation between two sentences. The final Spearman Correlation obtained on the test set was 0.809.

In the future, we will use methods such as context awareness and manual intervention to address errors caused by text data augmentation to ensure their accuracy and rationality. At the same time, we will expand the size of the dev set, reduce the size difference between data sets, and try to use other more powerful pre-trained models.

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*.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Ather Abdulrahem Mohammedsaed Alsamurai. 2017. Text categorization based on semantic similarity with word2vector. Master's thesis, Fen Bilimleri Enstitüsü.

Arijit Das and Diganta Saha. 2022. Deep learning based bengali question answering system using semantic textual similarity. *Multimedia Tools and Applications*, 81(1):589–613.

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*.

Minjie Ding, Mingang Chen, Wenjie Chen, and Lizhi Cai. 2021. English cloze test based on bert. In *International Conference on Knowledge Science, Engineering and Management*, pages 41–51. Springer.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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

Faith Wavinya Mutinda, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. Semantic textual similarity in japanese clinical domain texts using bert. *Methods of Information in Medicine*, 60:e56–e64.

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.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Kishor Wagh and Satish Kolhe. 2011. Information retrieval based on semantic similarity using information content. *International Journal of Computer Science Issues (IJCSI)*, 8(4):364.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2977–2992. Association for Computational Linguistics.

Jiaming Xu, Pengcheng Liu, Gaowei Wu, Zhengya Sun, Bo Xu, and Hongwei Hao. 2013. A fast matching method based on semantic similarity for short texts. In *Natural Language Processing and Chinese Computing: Second CCF Conference, NLPCC 2013, Chongqing, China, November 15-19, 2013, Proceedings 2*, pages 299–309. Springer.

Wei Xu, Chris Callison-Burch, and William B Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Majid Yazdani and Andrei Popescu-Belis. 2013. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence*, 194:176–202.