# HW-TSC 2024 Submission for the SemEval-2024 Task 1: Semantic Textual Relatedness (STR)

**Mengyao Piao, Chang Su, Yuang Li, Xiaosong Qiao, Xiaofeng Zhao, Yinglu Li, Min Zhang, Hao Yang, Dandan Tu**
Huawei Translation Services Center, China
{piaomengyao1, suchang8, liyuang3, qiaoxiaosong, zhaoxiaofeng14, liyinglu, zhangmin186, yanghao30, tudandan}@huawei.com

## Abstract

The degree of semantic relatedness of two units of language has long been considered fundamental to understanding meaning. In this paper, we present the system of Huawei Translation Services Center (HW-TSC) for Task 1 of SemEval 2024, which aims to automatically measure the semantic relatedness of sentence pairs in African and Asian languages. The task dataset for this task covers about 14 different languages, These languages originate from five distinct language families and are predominantly spoken in Africa and Asia. For this shared task, we describe our proposed solutions, including ideas and the implementation steps of the task, as well as the outcomes of each experiment on the development dataset. To enhance the performance, we leverage these experimental outcomes and construct an ensemble one. Our results demonstrate that our system achieves impressive performance on test datasets in unsupervised track B and ranked first place for the Punjabi language pair [1].

## 1 Introduction

The semantic relatedness of two units of language is the degree to which they are close in terms of their meaning (Abdalla et al., 2021). The linguistic units can be words, phrases, sentences, etc. Though our intuition of semantic relatedness is dependent on many factors such as the context of assessment, age, and socioeconomic status (Harispe et al., 2015), it is argued that a consensus can usually be reached for many pairs (Harispe et al., 2015). In the SemEval 2024 shared task 1 (Ousidhoum et al., 2024b), there are three sub-tracks — Track A: Supervised, Track B: Unsupervised, and Track C: Cross-lingual and each track involves several language pairs. Our team — Huawei Translation Services Center (HW-TSC) — participated in the Track B: Unsupervised one which covers most African and Asian language pairs and has to be developed without the use of any labeled data for semantic relatedness. In this paper, we describe HW-TSC's system for unsupervised semantic relatedness tasks, which leverages multiple pre-trained multilingual language models to capture the semantic relatedness of different language pairs. The main features of our system are as follows:

- **N-gram Chars Method**: We employ the tokenizers of two base models for this method. The first one is XLM-RoBERTa (Conneau et al., 2019), a large unsupervised cross-lingual model that extends Facebook's RoBERTa model with more languages and data. The second one is MultilingualBERT (Devlin et al., 2018), a transformers model that is pre-trained on a large multilingual corpus using self-supervised objectives. To measure the similarity between two sentences, we use their n-gram dictionaries as features and compute a similarity score based on them.

- **BERTScore Method**: This method adopts a metric, named BERTScore (Zhang* et al., 2020). It is a metric to assess the quality of the generated text. BERTScore is mainly based on the idea of computing a score from the cosine similarity of the token-level representations obtained from the BERT model for the generated and reference texts.

- **Pretrained Large Language Model Method**: We use XGLM (Lin et al., 2021), a large-scale auto-regressive language model, as the backbone of this method. XGLM is a pre-trained language model that can handle multiple languages and domains. By leveraging the powerful large language model, we can efficiently

[1] https://docs.google.com/spreadsheets/d/1KGN26MYVlfEOqooq-bzD6EBNnpl-YT5XrY9COKESS-g/edit?usp=sharing



*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1634–1638

obtain the token logits and perform calculations with them.

- **Translate to English and N-gram Chars Method**: This method needs us to process data with a translation system first, which converts the data from various languages into English. After the translation, we follow the same procedure as the N-gram Chars Method, which uses the n-gram character dictionaries of the generated and the reference texts to compute a similarity score.

- **Dataset**: We utilize the original development and test dataset from SemRel2024 (Ousidhoum et al., 2024a), a novel collection of semantic relatedness datasets annotated by native speakers for 14 languages: Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu.

In this paper, we analyze the characteristics of the shared task and describe our solutions, which include the ideas and implementation processes. We use Sentence-BERT (Reimers et al., 2019) as our baseline for the experiment. We conduct various experiments with the base model, large language models, etc. Our model achieves the best performance for the Punjabi language pairs in the unsupervised track B. The results are encouraging for semantic relatedness, although there is still scope for improvement.

## 2 Related Work

Our track is unsupervised, meaning that the systems submitted by participants do not rely on any labeled data for measuring semantic relatedness or similarity between text units longer than two words in any language. Consequently, any pre-trained language models that are further fine-tuned with text similarity data, using methods such as instruct-tuning, classification, or a similarity objective, are disqualified from our track. For our baseline score, we used Sentence-BERT (Reimers et al., 2019) (SBERT), a variant of the pre-trained BERT network that employs siamese and triplet architectures to generate sentence embeddings that are semantically meaningful and comparable by cosine similarity. SBERT has been fine-tuned on natural language inference (NLI) data, resulting in sentence embeddings that surpass other state-of-the-art methods. Hence, we selected SBERT as our baseline model and obtained our baseline score.

We introduce BERTScore (Zhang* et al., 2020) secondary, which is an automatic evaluation metric for text generation. Analogously to common metrics, BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence. What is more, different from other matches, it computes token similarity using contextual embeddings. BERTScore correlates better with human judgments and provides stronger model selection performance than existing metrics.

## 3 Method

### 3.1 N-gram Chars Method

The n-gram method (Kondrak and Grzegorz, 2005) is a statistical method used in natural language processing (NLP) to analyze the co-occurrence of words in a given text. It involves breaking down the text into sequences of words, where each sequence contains a fixed number of words, referred to as n-grams. The most common types of n-grams are bi-grams (2-grams), tri-grams (3-grams), and quadri-grams (4-grams), but n-grams can have any length. The primary purpose of using n-gram models is to capture the statistical dependencies between words in a language. By analyzing these dependencies, n-gram models can be used for various NLP tasks, such as language modeling, text generation, machine translation, and information retrieval. In this task, we first realized this way, for tokenizing, we tried pre-trained model XLM-Roberta (XLMR) and Multilingual-BERT (MBERT) because it is a multilingual task. At last, we calculate the similarity score with two sentences' n-gram dictionary shown as algorithm1.

### 3.2 BERTScore Method

BERTScore is a metric for evaluating the quality of text generation, particularly for tasks like machine translation, summarization, and text completion. BERTScore leverages the pre-trained BERT model (Bidirectional Encoder Representations from Transformers) to measure the semantic and syntactic alignment between the generated text and its reference or target text. The core idea behind BERTScore is to compute a score based on the cosine similarity of token-level representations from the BERT model for the generated text and

**Algorithm 1** N-gram Chars Score Method

**Require:** Word sequences of the two sentences $Sq_a$, $Sq_b$; there length $Len_a \leftarrow len(Sq_a)$, $Len_b \leftarrow len(Sq_b)$; N-gram window width $N$

**Ensure:** $0 < N < min(Len_a, Len_b)$

$Dict_{\{a,b\}} \leftarrow \{\}$
**for** $i \leftarrow 0$ **to** $Len_{\{a,b\}} - N$ **do**
  $W_{\{a,b\}i} \leftarrow Sq_{\{a,b\}}[i : i + N]$
  **if** $W_{\{a,b\}i}$ **not in** $Dict_{\{a,b\}}$ **then**
    $Dict_{\{a,b\}}\}[W_{\{a,b\}i}] = 1$
  **else**
    $state \leftarrow Dict_{\{a,b\}}[W_{\{a,b\}i}] + 1$
    $Dict_{\{a,b\}}[W_{\{a,b\}i}] \leftarrow state$
  **end if**
**end for**
$same \leftarrow 0$
**for all** $key$ **from** $Dict_a$ **do**
  **if** $key$ **is in** $Dict_b$ **then**
    $count \leftarrow min(Dict_a[key], Dict_b[key])$
    $same \leftarrow same + count$
  **end if**
**end for**
$score \leftarrow \frac{2 \times same}{Len_a + Len_b - 2N + 2}$
**return** $score$

the reference text. Additionally, BERTScore does not require training or tuning and is based on a publicly available pre-trained model. This makes it a useful and practical tool for evaluating the quality of generated text in various natural language processing tasks. Therefore, we calculate the score with the leverage of BERTScore.

### 3.3 Pretrained Large Language Model Method

Different from the method above, we take advantage of the pre-trained large language model to obtain the logits of the token and compute the score. XGLM is an open-source general language model pre-training framework[2]. The model architecture is general and can be easily extended, supporting various model scales and task-specific architectures. XGLM uses a Transformer-based architecture, after pre-training XGLM learns language structure and grammatical rules, and can generate high-quality natural language text. All in all, XGLM is a flexible and powerful general language model pre-training framework, that supports only Chinese and English. Therefore, we first use the model on the English task to get the logits of the token. Then try to calculate the sum, mean, and half of the logits in proper order.

[2] https://huggingface.co/docs/transformers/main/en/model_doc/xglm

### 3.4 Translate to English and N-gram Chars Method

Though XLMR and MBERT can support multiple languages, if we look closely at the training data we can see that most of it is in English. Our track mainly faced 14 different African and Asian languages, in order to satisfy our track more, we took advantage of our team to process the data with a translation system to make the data from African and Asian languages into English. And then get the logits of the token as well as the last one.

## 4 Experiments Results

In the beginning, we applied the following three methods to the English development dataset: N-gram Chars Method with XLMR and MBERT, BERTScore Method, and Pre-trained Large Language Model Method with XGLM to calculate the sum, mean, and half of the logits. See Table 1 Different methods on English development dataset. We can see Ngram-XLMR, Ngram-MBERT, and BERTScore got really impressive performance than every method on XGLM, though XGLM is a large language model and can generate high-quality natural language text almost all the methods with XGLM are below 0.5 in the score.

| Method-Model | Score |
|---|---|
| Ngram-XLMR | 0.651 |
| Ngram-MBERT | 0.604 |
| BERTScore | 0.650 |
| sum-XGLM | 0.091 |
| mean-XGLM | 0.314 |
| half-XGLM | 0.211 |

Table 1: Different method on English development dataset

Afterwards, we use these methods on the Afrikaans development dataset. What's more, we add Translate to English and N-gram Chars Based method. See Table 2 Different methods on Afrikaans development dataset. we can conclude that Ngram-XLMR and BERTScore still perform better than other methods. What is more, the Translate to English and N-gram Chars Based method did not bring us too many surprises. The table shows that the methods that translate to English are

| Method-Model | Score |
|---|---|
| Ngram-XLMR | 0.475 |
| Ngram-MBERT | -0.170 |
| BERTScore | 0.102 |
| eng-Ngram-XLMR | -0.171 |
| eng-Ngram-MBERT | 0.014 |
| eng-BERTScore | 0.102 |

Table 2: Different method on Afrikaans development dataset

almost all below the methods that did not.

| Ngram-XLMR ratio | Score |
|---|---|
| 0 | 0.650 |
| 0.1 | 0.689 |
| 0.2 | 0.690 |
| 0.3 | 0.689 |
| 0.4 | 0.685 |
| 0.5 | 0.680 |
| 0.6 | 0.676 |
| 0.7 | 0.674 |
| 0.8 | 0.673 |
| 0.9 | 0.672 |
| 1 | 0.651 |

Table 3: Different ensemble ratio with Ngram-XLMR and BERTScore on English development dataset

| Ngram-XLMR ratio | Score |
|---|---|
| 0 | 0.175 |
| 0.1 | 0.126 |
| 0.2 | 0.106 |
| 0.3 | 0.093 |
| 0.4 | 0.088 |
| 0.5 | 0.084 |
| 0.6 | 0.082 |
| 0.7 | 0.080 |
| 0.8 | 0.080 |
| 0.9 | 0.080 |
| 1 | 0.099 |

Table 4: Different ensemble ratio with Ngram-XLMR and BERTScore on Punjabi development dataset

From the experiments above, we can see N-gram Chars Based with XLMR and MBERT, BERTScore Based can always get better performance in English and Afrikaans. Will they get a better performance in the other 12 languages? See Table 5 this shows three methods and a Baseline on all language development datasets. To compare with the result from the three methods and baseline, we can see Ngram-XLMR and BERTScore always get better scores in all languages.

At last, we make other experiments to ensemble the results of Ngram-XLMR and BERTScore methods to find out if this way can bring us better performance. We make Ngram-XLMR with ratio A, and BERTScore method with ratio (1-A). See Table 3 Different ensemble ratio with Ngram-XLMR and BERTScore on English development dataset. See Table 4 Different ensemble ratio with Ngram-XLMR and BERTScore on Punjabi development dataset. We can see that the ensemble way may or may not improve the performance.

## 5 Conclusion

This paper describes HW-TSC's unsupervised system for Semantic Textual Relatednes shared task held in SemEval 2024 Task 1 and also presents the design, the data, and the results. The participants of the shared task were provided with a collection of unsupervised datasets in multiple languages. The shared task is challenging, partly due to the unsupervised development data, and can not use models that have fine-tuned with text similarity data whether through instruct-tuning (e.g., BLOOMZ (Muennighoff et al., 2022)), classification, or a similarity objective (like SBERT). Our system uses three base models with the dataset and carries out comprehensive experiments with different pre-trained models and methods. Finally, our system achieved the 1st best performance in the Punjabi language. For some of the problems reflected in this task, there is still a lot of research space. In the future, we will investigate the transfer method to transfer the knowledge of one language to multiple languages to improve efficiency and we plan to leverage other multiple languages model's skills.

## References

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

Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, Stoyanov, and Veselin. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

| Language | Ngram-XLMR | BERTScore | Ngram-MBERT | SBERT(Baseline) |
|---|---|---|---|---|
| eng | 0.651 | 0.650 | 0.604 | 0.758 |
| afr | 0.475 | 0.102 | -0.170 | 0.639 |
| amh | 0.630 | 0.085 | 0.630 | 0.650 |
| arb | 0.214 | 0.226 | 0.214 | 0.402 |
| arq | 0.487 | 0.420 | 0.408 | 0.296 |
| ary | 0.565 | 0.524 | 0.462 | 0.460 |
| hau | 0.325 | 0.240 | 0.325 | 0.382 |
| hin | 0.585 | 0.684 | 0.581 | 0.613 |
| ind | 0.490 | 0.468 | 0.501 | 0.445 |
| kin | 0.115 | 0.010 | 0.130 | 0.323 |
| pan | 0.099 | 0.175 | 0.099 | 0.173 |

Table 5: Top three methods and Baseline on all languages development dataset


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

Harispe, Sébastien, Ranwez, Sylvie, Janaqi, Stefan, Montmain, and Jacky. 2015. *Semantic similarity from natural language and ontology analysis*. Springer.

Kondrak and Grzegorz. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Lin, Xi Victoria, Mihaylov, Todor, Artetxe, Mikel, Wang, Tianlu, Chen, Shuohui, Simig, Daniel, Ott, Myle, Goyal, Naman, Bhosale, Shruti, Du, Jingfei, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Muennighoff, Niklas, Wang, Thomas, Sutawika, Lintang, Roberts, Adam, Biderman, Stella, Scao, Teven Le, Bari, M Saiful, Shen, Sheng, Yong, Zheng-Xin, Schoelkopf, Hailey, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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.

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

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.