Semantic Textual Similarity Analysis of Clinical Text in the Era of LLM

Yeli Feng
Amplify Health Asia, Singapore

***Abstract*—Semantic textual similarity analysis finds its practical usage in many real-world natural language processing applications, including clinical and biomedical settings. Many occurrence frequency based techniques have been proposed in the past, from n-grams to term frequency-inverse document frequency. Semantic relation based methodologies, such as the WordNet and Wu and Palmer measure, have flourished too. Recently, vector embedding based approaches that leverage large language models lifted analysis performance significantly. However, many existing semantic textual similarity analyses in the clinical and biomedical domains only exploited the early generation of large language models, whereas the latter field is advancing rapidly. This paper aims to fill the gap by investigating the potential of the latest foundational language models. This paper proposed MLTE, a feature fusion approach that leverages multi-layer Transformer embeddings for semantic similarity analysis. The effectiveness of the proposed method was evaluated with public clinical datasets in English, Chinese, and Japanese. Experimental outcomes showed that the proposed method is competitive and outperformed related works over two out of four datasets.**

***Index Terms*—Large Language Models, Semantic Textual Similarity, Bi-encoders, Multi-layer Transformer Embeddings, Multilingual clinical text**

I. Introduction

Semantic Textual Similarity (STS) analysis predicts the relationship between pairs of short texts. It is considered the most essential technique for various applications such as information retrieval, text categorization, question-answering systems, intelligent search engine, and so on [1]. In the field of clinical Natural Language Processing (NLP), STS has also been actively studied in academia [2] as well as healthcare organizations [3].

Research in STS analysis has progressed significantly, thanks to the recent technology breakthrough in Large Language Models (LLM). Techniques from deep learning to the Bidirectional Encoder Representation from Transformers (BERT) [5] and GPT-2 are systematically reviewed in [4]. State-of-the-art (SOTA) task-specific LLM and comprehensive benchmarks have been contributed to accelerate research in the biomedical domain as well [6].

However, the current artificial intelligence (AI) arms race among global research powerhouses is pushing foundational LLMs advancing at lightning speed. For example, many new architectures emerged after BERT, the backbone in most clinical NLP literature as of today [4], [6], was proposed. It elicits questions like whether there is still a need for specialized clinical LLM. Agrawal et al. argued that general-purpose LLMs such as GPT-3 have great promise for diverse clinical extraction tasks [7]. On the other hand, Lehman et al. concluded that with significantly fewer computational resources than GPT-3, relatively small size but specialized clinical LLMs, such as Text-to-Text Transfer Transformer (T5) [8], can achieve comparable or better performance for clinical language understanding tasks [9].

Similarly, the machine translation field has benefited tremendously from the LLM arms race [10]. We raise the question of whether there is still a need to develop specialized LLMs [11], [12] catered for each non-English clinical text. We investigate further the potential of LLM beyond BERT generation for multilingual clinical text similarity analysis. The main contributions of this paper are:

- Inspired by the ELMo [13] approach, this paper proposes a new method that exploits LLM's multi-layer Transformer embeddings as semantic-rich feature backbone and pairwise distances and kernel metrics for feature compression. Experimental results showed that the proposed method performed on par with fine-tuned LLMs over two out of four datasets in evaluation. Further enhancement was realized when fine-tuned LLMs served as the feature backbone.
- This paper experimentally shows that the recent advance in machine translation makes it feasible to take a language-unified approach to the STS task for the multilingual clinical text. Despite a bit of performance sacrifice in one of the non-English datasets in evaluation, practical MLOps benefits from the simplicity of such an approach.
- Furthermore, this paper shows that the semantic representation power of embeddings from BERT models fine-tuned with biomedical or clinical corpora don't naturally exhibit a performance advantage over embedding produced by more recent T5 Transformer trained for the general domain.

With a focus on advancing large-scale clinical NLP applications in the real world as future work, the proposed method takes a bi-encoder approach. Although the cross-encoder way often tops the STS task performance in the literature. Details are discussed in the next section. It is followed by the proposed method and experimental evaluation sections.

Author correspondence: yfeng002@e.ntu.edu.sg

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00227

II. Related Work

In [1], [4], the authors systematically reviewed STS analysis methods utilizing convolutional neural networks, recurrent neural networks, and prior non-neural techniques. Among them, the Embeddings from Language Models (ELMo) method exploited the hidden states of intermediate layers in bidirectional LSTM as part of features and created pre-LLM era SOTA for the biomedical language understanding domain [14]. The ELMo method inspired our feature design, which will be explained in section III.

In the LLM era, fine-tuning pre-trained LLMs and zero or few shots learning with Generative Pre-trained Transformers (GPTs) are the popular approaches that have been actively studied. However, with 100 more times model parameters, the latter approach has yet to show a decisive performance advantage [7]. From a real-world MLOps deployment perspective, this translates to much higher inference costs. Therefore, we adopted the fine-tuning LLMs approach. The rest of this section discusses related works in this direction, including those used for performance comparison in area IV.

Devlin et al. proposed BERT and demonstrated that a pre-trained BERT model can be fine-tuned to achieve excellent performance over a wide range of language understanding tasks [15]. The authors argued that BERT's bidirectional attention mechanism is more optimal than GPT's left-to-right unidirectional attention mechanism for sentence-level tasks. Since then, BERT has been adopted as the backbone in many works of the clinical NLP literature. Lee et al. proposed BioBERT [16] where the BERT model trained with general domain corpora (English Wikipedia and BooksCorpus) was fine-tuned on biomedical domain corpora (PubMed abstracts and PMC full-text articles). Alsentser et al. proposed Clinical-BioBERT [17] where the authors used around 2 million clinical notes from the MIMIC-III [1] database to fine-tune the BioBERT further.

To improve BERT for biomedical and clinical NLP tasks, Gu et al. challenged the prevailing assumption that pre-train with more text from other domains benefits the task at hand always [6]. The authors argued that one disadvantage of such continual fine-tuning is that the vocabulary from the general domain is not representative of the target biomedical domain, and subsequently showed that the PubMedBERT model trained from scratch with PubMed vocabulary only outperformed fine-tuning original BERT with its general domain vocabulary, overall on the BLURB benchmark [2]. Whereas, Yasunaga et al. continued along the fine-tuning path but exploited the rich dependencies between pairs of documents, such as references and hyperlinks. The author proposed LinkBERT [18] where a new objective, document relation prediction, was introduced to jointly train the model with BERT's original masked language modeling objective. BioLinkBERT remains at the top of the BLURB benchmark.

In the BERT architecture, a sentence pair is concatenated by a special token $[SEP]$ into a single input for the model to produce a joint embedding representation. This paradigm is called cross-encoder. It allows the attention mechanism to capture the finer relationship between a sentence pair, which is not feasible when inputs are independent sentences. The latter is called a bi-encoder. Although cross-encoders often outperform bi-encoders in prediction accuracy, they are computationally intensive. They can be excessively slower at inference time, especially for text classification and information retrieval applications over large-scale prior knowledge. Therefore, bi-encoders are more suitable for real-world applications when scaling and speed matter.

Along the bi-encoder direction, Reimers et al. proposed Sentence-BERT (SBERT) [19], which utilizes siamese network structures to generate embeddings that can be subsequently compared for semantic relationships using cosine-similarity. SBERT reduced the computing time for finding the most similar pair among 10,000 sentences from 65 hours with cross-encoding BERT to about 5 seconds on a modern GPU. The author reported that the prediction power of SBERT is competitive to BERT over the STS benchmark[3] in the general domain. In [20], Guo et al. proposed Sen-SCI-CORD19-BERT for STS analysis of COVID-19. It is an SBERT-based model fine-tuned with the CORD19STS dataset, including 13,710 sentence pairs curated from the COVID-19 open research dataset challenge.

In [21], Gao et al. introduced a simple contrastive learning approach, SimCSE, that uses "entailment" pairs as positives and "contradiction" pairs as negatives in a supervised setting and dropout noise as contrastive objective. Using the STS benchmark, the authors showed that SimCSE further advanced the SOTA of SBERT for STS tasks in the general domain.

While English NLP tasks in general and medical specialized domains have been actively studied using BERT and its variants. Whether these models and techniques generalize well in other languages, one common approach in the literature is extending English corpora trained models to a target language through knowledge distillation. In [11], Reimers et al. proposed a teacher-student framework where the teacher and student models are trained in parallel. The teacher is a pre-trained SBERT in charge of generating embeddings for text translated from a target language into English. At the same time, the student is fed with the original text and their English translation. This allows the student model to learn from embeddings produced by the teacher, hence knowledge distillation. Tan et al. further improved this framework in the general domain by incorporating contrastive learning [12].

Another common approach is utilizing BERT architecture but training models from non-English corpora. In [23], Cui et al. introduced MacBERT, trained from clinical corpora in the Chinese language. In [24], Chen et al. presented JCSE, a Japanese BERT model where contrastive learning was also exploited to overcome insufficient data in Japanese by gen-

[1]MIMIC-III Clinical Database at https://physionet.org/content/mimiciii/

[2]Biomedical Language Understanding and Reasoning Benchmark at https://microsoft.github.io/BLURB/

[3]Semantic textual similarity benchmark at https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

erating synthetic sentences. The effectiveness of JCSE and data augmentation was evaluated over a Japanese Clinical STS dataset. Unlike [11] [12], both MacBERT and JCSE adopted a cross-encoding architecture.

This paper explores the potential of a more recently proposed, less studied LLM model, T5, in a bi-encoding way for clinical STS tasks.

III. Proposed Method

The ELMo [13] method proposed by Peters et al. stacks hidden states from multiple intermediate layers of forward and backward paths of a bidirectional LSTM with word embeddings as features to train downstream language understanding tasks. The authors argued that the higher-level LSTM hidden states capture the context-dependent aspects of word meaning, while lower-level states capture the syntax aspects. Inspired by the feature design approach of ELMo, this paper looks into a fusion of multi-layer embeddings from LLM intermediate encoder blocks, specifically Transformer-based encoders, for more semantic representation power.

A. MLTE Feature Fusion

Given a text with multiple words, LLM first tokenizes and encodes the input text into a sequence of embedding vectors $X^0 = \{x_1, ..., x_n, p\}$, as illustrated in Fig. 1, where p refers to a positional encoding vector captures the order of words in the input text. X^0 is then served to the first encoder block; its output is then forwarded to the next encoder block, and so on. Each encoder block consists of a multi-head attention layer, followed by Add&normalize and fully connected linear layers. It ends with another Add&normalize layer. This structure constitutes the self-attention mechanism, the bedrock of modern LLM architectures.

Large LLM has more encoder blocks than its basic version, and the dimension of embedding x_n is also higher. For example, there are 12 encoder blocks in the Sentence-T5 model [22], and the dimension of its embedding is 768. Its large version consists of 24 encoder blocks with the embedding extent extended to 1024. However, in each LLM, the input and output dimensions of all encoder blocks are identical. Our MLTE feature proposal extracts embeddings from all encoder blocks. Then it fuses them into input feature as described by (1), where $\overline{X^i}$ refers to a mean pooling of encoder block l's output, which is a sequence in $n+1$ length.

$$F = [\ \overline{X^1},\ \overline{X^2},\ ...\ \overline{X^i},\ ...\ \overline{X^L}\ ] \tag{1}$$

Dimension of the MLTE feature F equals (D, L), where D is the dimension of embedding vector x and L represents the number of encoder blocks to utilize. The dimension of the fused feature is very high. Given MLTE features of a pair of clinical texts A and B, the prediction model in Fig. 1 first utilizes pairwise metrics to compress the input feature pairs into $(L, 5)$ dimension as defined in (2). The combination of the five chosen metrics complements each other. The cosine similarity compares the orientation of a pair of vectors while Euclidean and Manhattan distances compare their magnitudes. The polynomial and sigmoid kernels are helpful to measure the affinity of the vector pair in their interacted space. Finally, the model learns a regressor or classifier from the compressed feature z, optionally concatenated with other features available from data, e.g., the category of a sentence pair.

$$z = \begin{bmatrix} f_{cosine-similarity}(F_A^i,\ F_B^i) \\ f_{manhattan-distance}(F_A^i,\ F_B^i) \\ f_{euclidean-distance}(F_A^i,\ F_B^i) \\ f_{polynomial-kernel}(F_A^i,\ F_B^i) \\ f_{sigmoid-kernel}(F_A^i,\ F_B^i) \end{bmatrix}^T , i \in [1, 2, ...L] \tag{2}$$

B. Language Unified Approach

As reviewed in the related work section, the teacher-student framework has been widely adopted in the literature. The training process of this framework involves two BERT or similar models. The teacher is pre-trained with domain knowledge in English. Together with the English translation of the original corpora, the teacher helps the student model to learn from corpora in the target language. Adopting this framework in practical MLOps means an independent model must be deployed for each language. We argue this is not an optimal operation practice if a single English model deployment could achieve on-par performance.

LLM for machine translation has made tremendous progress lately, which has given rise to commercial AI translation services being available at a meager cost. For example, Table I shows examples of Chinese and Japanese to English

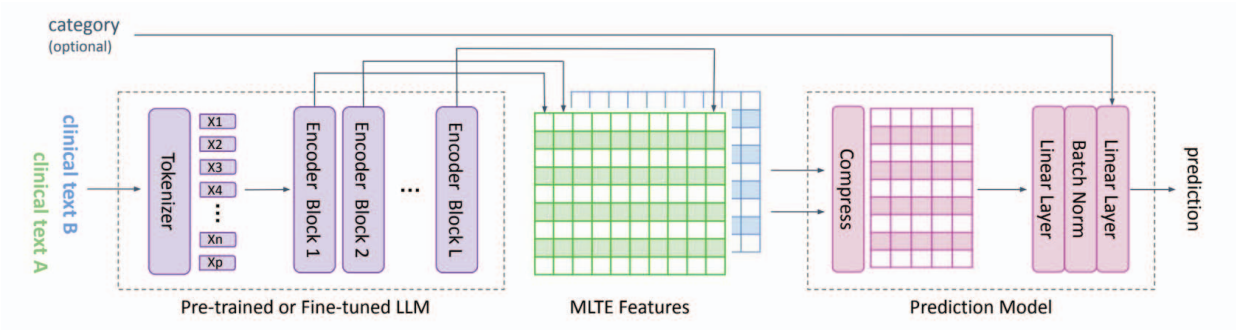

Fig. 1: Schematic Diagram of the Proposal Method with Encoder-only Transformer as Feature Backbone

translations done by Google translate API [4]. Riding on this success, we propose a language-unified approach, where a single English-only domain specialized LLM goes for MLOps. In this unified approach, non-English corpora are translated into English, and subsequent LLM fine-tuning and inference are completed in English only. Thereby, a single model for multilingualism is achieved straightforwardly. Its effectiveness will be analyzed in subsection IV-D.

TABLE I: Examples of Clinical Text Translation

Original clinical text	English translation
糖尿病和尿毒症有什么区别?	What is the difference between diabetes and uremia?
レントゲン所:全的に30-50%の水平的な骨吸をめた	X-ray findings: 30-50% horizontal bone resorption was observed in the whole jaw

C. Learning Objective

The prediction model in Fig. 1 will be trained with MLE loss. For the STS task, Pearson linear correlation and Spearman rank correlation metrics are widely used to evaluate correlations between predictions and truth in a linear and monotonic fashion. Therefore, we introduce a non-monotonic error into the model training objective.

Specifically, let a batch of N training sentence pairs be sorted in ascending order by their actual similarity values; a penalization is introduced if the predicted value of pair i is larger than its adjacent pair j, where $i < j$ and $y_i \leq y_j$, as shown in (3). The hyperparameter β determines the strength of penalizing non-monotonic error. The optimal hyperparameter β value is dataset dependent and will be explained in the experimental section.

$$\widehat{y}_i = \underset{\theta}{\mathcal{L}}^2\big(\underset{\theta}{\mathcal{L}}^1(z_i) \; + \; c_i\big) \tag{3a}$$

$$\min_{\theta} \; \frac{1}{N}\sum_{i}^{N}\big(\widehat{y}_i - y_i\big)^2 + \beta \sum_{i}^{N-1} max\big(0,\, \widehat{y}_i - \widehat{y}_j\big) \tag{3b}$$

IV. Experimental Results

This section first describes the experiment setup and then presents a performance comparison of off-the-shelf LLMs trained for the general domain or fine-tuned for biomedical or clinical domains. Finally, evaluation results of the proposed method are presented and compared with related works.

A. Setup

1) Data sets: Four datasets from clinical and biomedical domains were used for the experiments. As shown in Table II, BIOSSES [26] includes 100 English sentence pairs in the biomedical domain, where the semantic similarity between pairs is labeled from the least (0) to the most (4) similar. EBMSASS [27] was constructed for quickly finding relevant prior knowledge in evidence-based medicine practice. It includes 1000 pairs of English sentences from biomedical publications. The Japanese clinical STS dataset, JACSTS [29], consists of 2670 Japanese sentence pairs extracted from clinical reports and electronic medical records. CHIP-STS includes 30,000 Chinese sentence pairs from the Chinese Biomedical Language Understanding Evaluation (CBLUE) benchmark [30], where the corpus was extracted from clinical trials, electronic healthcare records, medical books, and search logs from real-world search engines.

TABLE II: Datasets in Experiment

Name	Language	Number of samples			Label
		Train	Dev	Test	
BIOSSES [26]	English	64	16	20	[0., 4.]
EBMSASS [27]	English	600	200	200	[1., 5.]
JACSTS [29]	Japanese	2202	734	734	0,1,2,3,4,5
CHIP-STS [30]	Chinese	16000	4000	10000	0, 1

In all experiments, default train, dev, and test sets were followed except for JACSTS, where a random splitting was applied. For Chinese and Japanese sentences, Google Translate API was used.

2) Evaluation Metrics: The sentence pairs in CHIP-STS are labeled as either similar (1) or different (0). We followed the literature, treated it as a classification task, and measured predictions with accuracy and macro F1 score. Following the literature, Pearson linear correlation and Spearman rank correlation were used to measure performance over all the other datasets.

B. Off-the-Shelf Pre-trained LLMs

As discussed in section II, most prior works for various clinical and biomedical NLP tasks fine-tuned the general domain BERT model. We are interested in whether these domain-specific models have any performance advantage over T5 Transformer, a more recent LLM architecture, for the clinical STS task. Three BERT and two T5 models were selected for this experiment. Domain and model size are listed in Table IIIa. The performance of selected off-the-shelf LLMs in the STS task was evaluated using test sets of BIOSSES, EBMSASS, and JACSTS, and the dev set of CHIP-STS, as its test set doesn't include ground truth.

The experiment results in Table IIIb show the general domain T5 performed much better than all four domain-specific LLMs for the STS task. For example, the BioLinkBERT-large [18] performed the worst over CHIP-STS, JACSTS, and EBMSASS, although it has 60% model parameters than the general T5. One possible reason could be these off-the-shelf BERT models were fine-tuned in a cross-encoder manner with a linear neural layer for a specific task.

The literature attributed better performances obtained by cross-encoder LLM models to it allowing the attention mechanism to capture the finer relationship between a sentence pair, which is not feasible when inputs are independent sentences [17], [18]. When they are used as bi-encoders here, the result infers that the semantic representation power of their embeddings is weak compared to the general domain T5 model

[4]https://cloud.google.com/translate?hl=en
[5]https://pubmed.ncbi.nlm.nih.gov
[6]Stanford Natural Language Inference (SNLI) Corpus

TABLE III: Off-the-Shelf LLMs

Model Name	Year	Model Size	Domain	Corpus
Bio-ClinicalBERT [17]	2019	415M	Biomedical + Clinical	PubMed[5] + MIMIC-III
PubMedBERT [6]	2021	441M	Biomedical	PubMed
BioLinkBERT-large [18]	2022	1024M	Biomedical	PubMed
Clinical-T5-large [25]	2023	770M	Clinical	MIMIC-III
Sentence-T5-large [22]	2022	638M	General	web forums + SNLI[6]

(a) Model Information

Model Name	BIOSSES	EBMSASS	JACSTS	CHIP-STS
	Pearson	Pearson	Pearson	Macro-F1
Clinical-BioBERT	0.6064	0.4774	0.7168	0.6596
PubMedBERT	0.7785	0.3713	0.6204	0.6060
BioLinkBERT-large	0.7883	0.3658	0.5176	0.3695
Clinical-T5-large	0.3702	0.5648	0.7185	0.4704
Sentence-T5-large	0.8522	0.7631	0.7871	0.7241

(b) Model Performance

not trained in a cross-encoder way. This is reinforced by the similarly poor result from Clinical-T5-large [25], which also took a cross-encoder approach. The reason behind a much weaker semantic representation power of cross-encoders could be that a cross-encoder model training for a specific task has a linear neural layer placed after the encoder. We postulate that the linear neural layer here attributed much semantic representation power. Therefore, although their embeddings alone are weak, cross-encoders perform better when the downstream task is fixed.

C. Outcome of Proposed Method

Given the above analysis, Sentence-T5-large [22] was chosen for evaluating the proposed method. All MLTE prediction models were trained for 30 epochs with Adam optimizer, learning rate 0.02, weight decay 0.001, and batch size 24. Using the respective dev set, a simple grid search between [0, 3] with a step of 0.5 was conducted to find an optimal hyperparameter β value for each MLTE prediction model. For CHIP-STS, the normalized category information was concatenated with the compressed MLTE feature.

In this experiment set, the pre-trained Sentence-T5-large model was further fine-tuned in a bi-encoder way with a respective train set to enhance the semantic representation power of the embeddings. For BIOSSES, EBMSASS, and JACSTS, the fine-tuning objective was maximizing the cosine similarity of embedding sentence pairs. As CHIP-STS is a binary classification task, the contrastive loss [21] was used as the tuning objective. The optimizer was AdmW, learning rate 1e-5, weight decay 1e-4. A dropout rate of 0.1 was introduced to sentence embeddings pooled by a mean function. All fine-tuning processes took ten epochs in a bath size of 8, and the learning rate was scheduled with a 10% linear warmup as in the literature. Model checkpoint from the best epoch according to the corresponding dev set was used for performance evaluation.

The results are listed in Table IV, where "ft" denotes fine-tuning. In rows without "MLTE", the performance was computed with cosine similarities of sentence embedding pairs as in Table IIIb. The proposed method introduced performance improvement to all occasions under evaluation, except for using the multi-layer embeddings from the pre-trained LLM for BIOSSES. However, the performance lift to the other 3 data sets is significant; for example, in Pearson term, as high as 0.09 better for EBMSASS, close to the result produced by corresponding fine-tuned LLM. BIOSSES and CHIP-STS benefited more from fine-tuned LLMs. But we can see further performance improvement across all data sets when applying MLTE to fine-tuned LLMs, as shown by Sentence-T5-large-ft-MLTE in Table IV.

D. Performance Comparison

This section compares the performance of the proposed method with related works that all took a fine-tuned approach. So Sentence-T5-large-ft-MLTE is used for comparison. See in Table V. For English corpora, we can see that our method performed better than PubMedBERT [6] and NCBI-BERT [14] on BIOSSES, although both are cross-encoders, which are denoted by ^{c} in the table. Although BioLinkBERT [18] has 60% more model parameters, its performance is only 0.011 higher than our method in Pearson term. For EBMSASS, our method outperformed BERT-base [28]. And it is far better than non-LLM based G+O+C method [27], where similarity measure was anchored on a group of linguistic components (Generic, Ontology, Concept2vec).

For non-English corpora, our language-unified approach gave a better result on JACSTS than JCSE-large [24], in Spearman term. Although the latter fine-tuned the BERT model with the original Japanese corpus. For CHIP-STS test set, our method resulted in a slightly lower F1 score. Besides the cross-encoder advantage in BERT-base, MacBERT-large, and BERT-wwm-ext-base [23], another possible contributing factor could be these models were trained with original Chinese corpora that are different from the general English corpora sentence-T5-large trained with. In addition, a Chinese word segmentation process was introduced in the tokenization process of MacBERT-large and BERT-wwm-ext-base to ensure whole Chinese word masking.

TABLE IV: Experiment Results of Proposed Method

Method	BIOSSES		EBMSASS		JACSTS		CHIP-STS	
	Pearson	*Spearman*	*Pearson*	*Spearman*	*Pearson*	*Spearman*	*Macro-F1*	*Accuracy*
Sentence-T5-large	0.8522	0.8434	0.7631	0.6620	0.7871	0.7957	0.7241	0.7305
Sentence-T5-large-MLTE	0.8371	0.8442	0.8534	0.6688	0.8306	0.8379	0.7557	0.7565
Sentence-T5-large-ft	0.9147	0.9213	0.8669	0.7250	0.8588	0.8563	0.8495	0.8495
Sentence-T5-large-ft-MLTE	0.9248	0.9335	0.8879	0.7261	0.8652	0.8611	0.8527	0.8528

TABLE V: Performance Comparison

NCBI-BERT (base) (P+M)	PubMedBERT	BioLinkBERT	Sentence-T5-large-ft-MLTE
0.916[c]	0.923[c]	0.936[c]	0.925

(a) BIOSSES *(Pearson)*

G+O+C	BERT-base	Sentence-T5-large-ft-MLTE
0.804	0.867[c]	0.888

(b) EBMSASS *(Pearson)*

JCSE-large	Sentence-T5-large-ft-MLTE
0.824[c]	0.861

(c) JACSTS *(Spearman)*

BERT-base	BERT-wwm-ext-base	MacBERT-large	Sentence-T5-large-ft-MLTE
0.830[c]	0.839[c]	0.856[c]	0.819 [7]

(d) CHIP-STS (test set) *(Macro-F1)*

V. Conclusion

Pre-trained and fine-tuned BERT and similar LLMs have been widely used in STS tasks for English clinical text and other languages. This paper proposed a feature fusion method that first extracts multi-layer Transformer embeddings from the more recent T5 model and then compresses the high dimensional features for learning a clinical text similarity predictor. In addition, a simple unified approach was adopted for non-English corpora. Experimental results over clinical text corpora in three languages showed the proposed method is effective. It outperformed some related work in comparison.

References

[1] Prakoso, D., Abdi, A. & Amrit, C. Short text similarity measurement methods: a review. *Soft Computing*. **25** pp. 4699-4723 (2021)

[2] Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M. & Liu, H. MedSTS: a resource for clinical semantic textual similarity. *Language Resources And Evaluation*. **54** pp. 57-72 (2020)

[3] Wu, H., Wang, M., Wu, J., Francis, F., Chang, Y., Shavick, A., Dong, H., Poon, M., Fitzpatrick, N., Levine, A. & Others A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ Digital Medicine*. **5**, 186 (2022)

[4] Amur, Z., Kwang Hooi, Y., Bhanbhro, H., Dahri, K. & Soomro, G. Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives. *Applied Sciences*. **13**, 3911 (2023)

[5] Kenton, J. & Toutanova, L. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings Of NaacL-HLT*. **1** pp. 2 (2019)

[6] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions On Computing For Healthcare (HEALTH)*. **3**, 1-23 (2021)

[7] Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are few-shot clinical information extractors. *Proceedings Of The 2022 Conference On Empirical Methods In Natural Language Processing*. pp. 1998-2022 (2022)

[8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal Of Machine Learning Research*. **21**, 5485-5551 (2020)

[7] Computed on the CBLUE benchmark platform https://tianchi.aliyun.com/dataset/95414/submission

[9] Lehman, E., Hernandez, E., Mahajan, D., Wulff, J., Smith, M., Ziegler, Z., Nadler, D., Szolovits, P., Johnson, A. & Alsentzer, E. Do We Still Need Clinical Language Models?. *ArXiv E-prints*. pp. arXiv-2302 (2023)

[10] Rivera-Trigueros, I. Machine translation systems and quality assessment: a systematic review. *Language Resources And Evaluation*. **56**, 593-619 (2022)

[11] Reimers, N. & Gurevych, I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *Proceedings Of The 2020 Conference On Empirical Methods In Natural Language Processing (EMNLP)*. pp. 4512-4525 (2020)

[12] Tan, W., Heffernan, K., Schwenk, H. & Koehn, P. Multilingual Representation Distillation with Contrastive Learning. *Proceedings Of The 17th Conference Of The European Chapter Of The Association For Computational Linguistics*. pp. 1469-1482 (2023)

[13] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. Deep Contextualized Word Representations. *Proceedings Of The 2018 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pp. 2227-2237 (2018,6)

[14] Peng, Y., Yan, S. & Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *Proceedings Of The 18th BioNLP Workshop And Shared Task*. pp. 58-65 (2019)

[15] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. (2018)

[16] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. & Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. (2019)

[17] Alsentzer, E., Murphy, J., Boag, W., Weng, W., Jindi, D., Naumann, T. & McDermott, M. Publicly Available Clinical BERT Embeddings. *Proceedings Of The 2nd Clinical Natural Language Processing Workshop*. pp. 72-78 (2019)

[18] Yasunaga, M., Leskovec, J. & Liang, P. LinkBERT: Pretraining Language Models with Document Links. *Proceedings Of The 60th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers)*. pp. 8003-8016 (2022)

[19] Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv Preprint ArXiv:1908.10084*. (2019)

[20] Guo, X., Mirzaalian, H., Sabir, E., Jaiswal, A. & Abd-Almageed, W. Cord19sts: Covid-19 semantic textual similarity dataset. *ArXiv Preprint ArXiv:2007.02461*. (2020)

[21] Gao, T., Yao, X. & Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv Preprint ArXiv:2104.08821*. (2021)

[22] Ni, J., Abrego, G., Constant, N., Ma, J., Hall, K., Cer, D. & Yang, Y. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. *Findings Of The Association For Computational Linguistics: ACL 2022*. pp. 1864-1874 (2022)

[23] Cui, Y., Che, W., Liu, T., Qin, B. & Yang, Z. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions On Audio, Speech, And Language Processing*. **29** pp. 3504-3514 (2021)

[24] Chen, Z., Handa, H. & Shirahama, K. JCSE: Contrastive Learning of Japanese Sentence Embeddings and Its Applications. *ArXiv E-prints*. pp. arXiv-2301 (2023)

[25] Lehman, E. & Johnson, A. Clinical-t5: Large language models built using mimic clinical text. (PhysioNet,2023)

[26] Soğancıoğlu, G., Öztürk, H. & Özgür, A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*. **33**, i49 (2017)

[27] Hassanzadeh, H., Nguyen, A. & Verspoor, K. Quantifying semantic similarity of clinical evidence in the biomedical literature to facilitate related evidence synthesis. *Journal Of Biomedical Informatics*. **100** pp. 103321 (2019)

[28] Wang, Y., Wang, M. & Nakov, P. Rethinking STS and NLI in Large Language Models. *ArXiv Preprint ArXiv:2309.08969*. (2023)

[29] Mutinda, F., Yada, S., Wakamiya, S. & Aramaki, E. Semantic Textual Similarity in Japanese Clinical Domain Texts Using BERT. *Methods Of Information In Medicine*. **60**, e56 (2021)

[30] Zhang, N., Chen, M., Bi, Z., Liang, X., Li, L., Shang, X., Yin, K., Tan, C., Xu, J., Huang, F. & Others CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. *Proceedings Of The 60th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers)*. pp. 7888-7915 (2022)