

Original Paper

Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study

Ying Xiong¹, PhD; Shuai Chen¹, MS; Qingcai Chen^{1,2}, PhD, Prof Dr; Jun Yan³, PhD; Buzhou Tang^{1,2}, PhD, Prof Dr

¹Harbin Institute of Technology, Shenzhen, China

²Peng Cheng Laboratory, Shenzhen, China

³Yidu Cloud Technology Company Limited, Beijing, China

Corresponding Author:

Buzhou Tang, PhD, Prof Dr

Harbin Institute of Technology

HIT Campus, Xili University Town

Shenzhen, 518055

China

Phone: 86 075526033182

Email: tangbuzhou@gmail.com

Abstract

Background: With the popularity of electronic health records (EHRs), the quality of health care has been improved. However, there are also some problems caused by EHRs, such as the growing use of copy-and-paste and templates, resulting in EHRs of low quality in content. In order to minimize data redundancy in different documents, Harvard Medical School and Mayo Clinic organized a national natural language processing (NLP) clinical challenge (n2c2) on clinical semantic textual similarity (ClinicalSTS) in 2019. The task of this challenge is to compute the semantic similarity among clinical text snippets.

Objective: In this study, we aim to investigate novel methods to model ClinicalSTS and analyze the results.

Methods: We propose a semantically enhanced text matching model for the 2019 n2c2/Open Health NLP (OHNLP) challenge on ClinicalSTS. The model includes 3 representation modules to encode clinical text snippet pairs at different levels: (1) character-level representation module based on convolutional neural network (CNN) to tackle the out-of-vocabulary problem in NLP; (2) sentence-level representation module that adopts a pretrained language model bidirectional encoder representation from transformers (BERT) to encode clinical text snippet pairs; and (3) entity-level representation module to model clinical entity information in clinical text snippets. In the case of entity-level representation, we compare 2 methods. One encodes entities by the entity-type label sequence corresponding to text snippet (called entity I), whereas the other encodes entities by their representation in MeSH, a knowledge graph in the medical domain (called entity II).

Results: We conduct experiments on the ClinicalSTS corpus of the 2019 n2c2/OHNLP challenge for model performance evaluation. The model only using BERT for text snippet pair encoding achieved a Pearson correlation coefficient (PCC) of 0.848. When character-level representation and entity-level representation are individually added into our model, the PCC increased to 0.857 and 0.854 (entity I)/0.859 (entity II), respectively. When both character-level representation and entity-level representation are added into our model, the PCC further increased to 0.861 (entity I) and 0.868 (entity II).

Conclusions: Experimental results show that both character-level information and entity-level information can effectively enhance the BERT-based STS model.

(*JMIR Med Inform* 2020;8(12):e23357) doi: [10.2196/23357](https://doi.org/10.2196/23357)

KEYWORDS

natural language processing; deep learning; clinical semantic textual similarity; knowledge graph

Introduction

Background

Electronic health record (EHR) systems have been widely used in hospitals all over the world for convenience to health information storage, share, and exchange [1]. In recent years, EHRs have become a key data source for medical research and clinical decision support. Therefore, the quality of EHRs is crucial. However, copy-and-paste and templates are very common in EHR writing [2,3], resulting in EHRs of low quality in content. How to detect copy-and-paste and templates in different documents has become increasingly important for the secondary use of EHRs. This can be regarded as a clinical semantic textual similarity (ClinicalSTS) task, which is also applied to clinical decision support, trial recruitment, tailored care, clinical research [4-6], and medical information services, such as clinical question answering [7,8] and document classification [9].

In the past few years, some shared tasks on STS, such as Semantic Evaluation (SemEval), have been launched by different organizers [10-14]. These shared tasks mainly focus on general domains, including newswire, tutorial dialog system, Wikipedia, among others. There has been almost no study on STS in the clinical domain. To boost the development of ClinicalSTS, Wang et al [15] constructed a clinical STS corpus of 174,629 clinical text snippet pairs from Mayo Clinic. Based on a part of this corpus, BioCreative/OHNLP organizers held the first ClinicalSTS shared pilot task (challenge) in 2018 [16]. A corpus of 1068 clinical text snippet pairs with similarity ranging from 0 to 5 was provided for this shared task. In 2019, the n2c2/OHNLP organizers extended the 2018 shared task corpus and continued to hold ClinicalSTS shared task [17]. The extended corpus is composed of 2055 clinical text snippet pairs.

In this paper, we introduce our system developed for the 2019 n2c2/OHNLP shared task on ClinicalSTS. The system is based on bidirectional encoder representation from transformers (BERT) [18] and includes the 2 other types of representations besides BERT: (1) character-level representation to tackle the out-of-vocabulary (OOV) problem in natural language processing (NLP) and (2) entity-level representation to model clinical entity information in clinical text snippets. In the case of entity-level representation, we apply 2 entity-level representations: one encodes entities in a text snippet by the corresponding entity label sequence (called entity I) and the other one encodes entities with their representation on Mesh [19] (called entity II). Our system achieves the highest Pearson correlation coefficient (PCC) of 0.868 on the corpus of the 2019

n2c2/OHNLP track on ClinicalSTS, which is competitive with other state-of-the-art systems.

Related Work

A model for STS usually consists of 2 modules: a module to encode text snippet (or sentence) pairs and a module for prediction (classification or regression). According to sentence pair encoding, STS models can be classified into the following 2 categories: sentence encoding models and sentence pair interaction models. The sentence encoding models first use Siamese neural network to individually encode 2 sentences with 2 neural networks of the same structure and shared parameters [20-23], then combine the 2 sentences' representation through concatenation, element-wise product, or element-wise difference operations, and finally make a classification or regression prediction via a specific layer such as multilayer perceptron (MLP) [24]. The main limitation of the sentence pair encoding models is that they ignore word-level interactions. The sentence pair interaction models adopt matching-aggregation architectures to encode word-level interactions [25,26]. These models first build an interaction matrix and then use a convolutional neural network (CNN) [27] and long short-term memory [28] with attention mechanism [29,30] and hierarchical architecture [31] to obtain aggregated matching representation for final prediction.

In recent years, pretrained language models good at capturing sentence-level semantic information, such as BERT [18], XLNet [32], RoBERTa [33], have been proved to significantly improve downstream tasks. However, most pretrained language models are at the token level. In order to tackle the inherent OOV problem of NLP, character-level representation is also considered in various NLP tasks, such as named entity recognition [34-36] and entity normalization [37], and brings improvements. Besides, researchers have started investigating how to use entity-level representation in NLP tasks [38,39].

Methods

Data Set

The n2c2/OHNLP organizers manually annotated a total of 2055 clinical text snippet pairs by 2 medical experts for the ClinicalSTS task, where 1643 pairs are used as the training set and 412 as the test set. The similarity of each clinical text snippet pair is measured by PCC ranging from 0 to 5, where 0 means that 2 clinical text snippets are absolutely different, and 5 means that 2 clinical text snippets are entirely semantically equal. All clinical text snippets are selected from deidentified EHRs. Table 1 gives examples of each score.

Table 1. Examples of ClinicalSTS^a.

Score	Example of clinical text snippet pair
0	<p>The 2 sentences are completely dissimilar</p> <p>S1: The patient has missed 0 hours of work in the past seven days for issues not related to depression.</p> <p>S2: In the past year the patient has the following number of visits: none in the hospital none in the er and one as an outpatient.</p>
1	<p>The 2 sentences are not equivalent but have the same topic</p> <p>S1: There is no lower extremity edema present bilaterally.</p> <p>S2: There is a 2+ radial pulse present in the upper extremities bilaterally.</p>
2	<p>The 2 sentences are not equivalent but share some details</p> <p>S1: I met with the charge nurse and reviewed the patient's clinical condition.</p> <p>S2: I have reviewed the relevant imaging and medical record.</p>
3	<p>The 2 sentences are roughly equivalent but some important information differs</p> <p>S1: I explained the diagnosis and treatment plan in detail, and the patient clearly expressed understanding of the content reviewed.</p> <p>S2: Began discussion of diagnosis and treatment of chronic pain and chronic fatigue; patient expressed understanding of the content.</p>
4	<p>The 2 sentences are mostly equivalent and only a little detail is different</p> <p>S1: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 2 puffs by inhalation every 4 hours as needed.</p> <p>S2: Albuterol [PROVENTIL/VENTOLIN] 90 mcg/Act HFA Aerosol 1-2 puffs by inhalation every 4 hours as needed #1 each.</p>
5	<p>The 2 sentences mean the same thing, they are absolutely equivalent</p> <p>S1: Goals/Outcomes: Patient will be instructed in a home program, demonstrate understanding, and state the ability to continue independently.</p> <p>S2: Patient will be instructed in home program, demonstrate understanding, and state ability to continue independently-ongoing.</p>

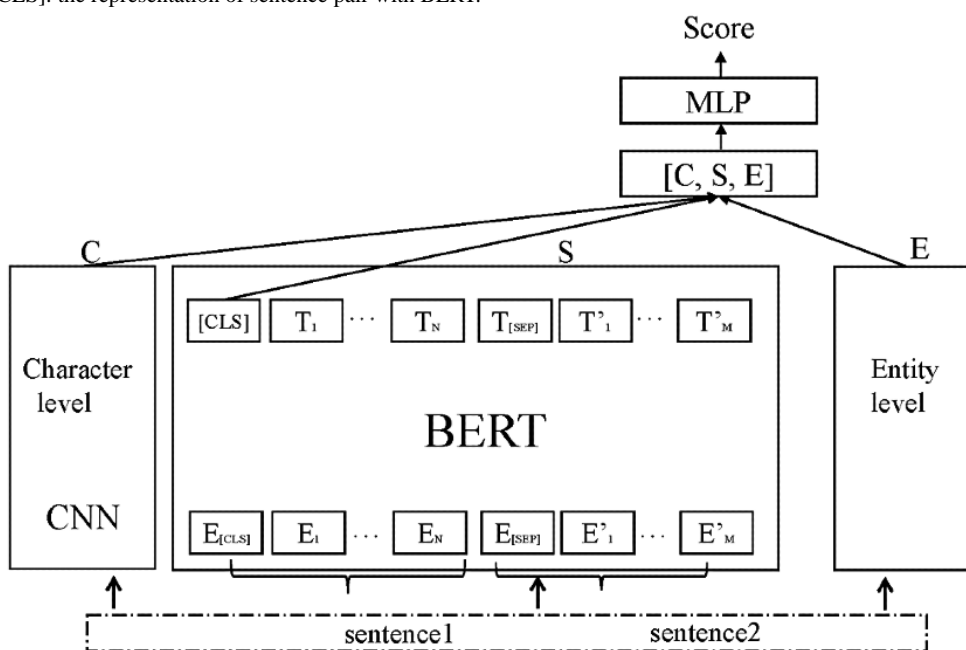
^aClinicalSTS: clinical semantic textual similarity.

Models

Figure 1 presents an overview architecture of our model. In this model, we first use 3 representation modules at different levels

to encode input text snippet pairs, that is, character-level, sentence-level, and entity-level representation modules, and then feed them to MLP for prediction.

Figure 1. Overview architecture of our model for the ClinicalSTS track of the 2019 n2c2/OHNLN challenge. BERT: bidirectional encoder representation from transformers; ClinicalSTS: clinical semantic textual similarity; CNN: convolutional neural network; MLP: multilayer perceptron; PCC: Pearson correlation coefficient; [CLS]: the representation of sentence pair with BERT.



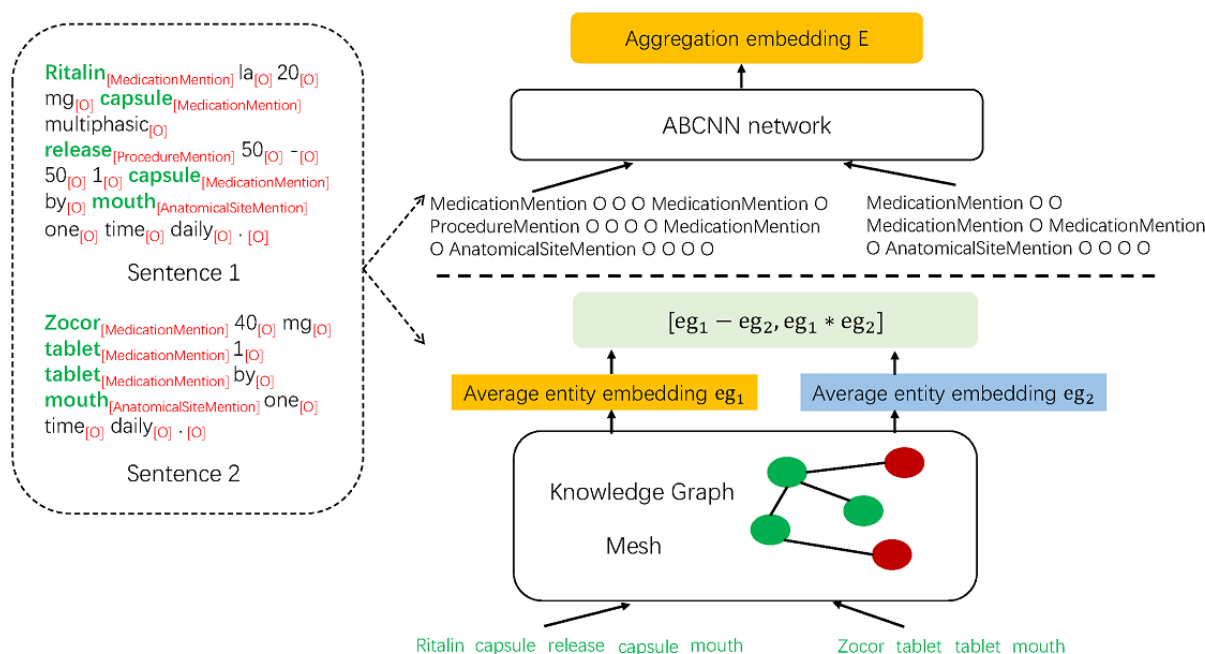
Character-Level Representation

In order to tackle the OOV problem in NLP, following [34-37], given a pair of clinical text snippets (a, b), we first apply character-level CNN on each token to obtain its character-level representation, and then apply max pooling operation on all tokens in a and b to obtain the character-level representation of (a, b), denoted by C. We model the character-level representation with CNN, because there is no significant difference in using CNN and long short-term memory, according to previous studies [40,41].

Sentence-Level Representation

We use BERT to encode the input clinical text snippet pair (a, b) and obtain its sentence-level representation, denoted by S = BERT(a, b).

Figure 2. Entity-level representation.



In the first method for entity-level representation (entity I), we convert text snippet a and b into entity-type sequences corresponding to them, and then deploy attention-based CNN [27] on the pair of the entity-type sequences in the following way:

$$E = BCNN(es_a, es_b) \quad (1)$$

where es_a is the entity label sequence of text snippet a, es_b is the entity label sequence of text snippet b, BCNN is basic bi-CNN, and E is the entity-level representation of (es_a, es_b) . For example, given a text snippet b “Zocor 40 mg tablet 1 tablet by mouth one time daily.” shown in Figure 2, cTAKES first extracts 3 medication mentions {“Zocor”, “tablet”, “tablet”} and 1 anatomical mention {“mouth”}, and then we obtain the entity-type sequence corresponding to text snippet b: “MedicationMention O O MedicationMention O MedicationMention O AnatomicalSiteMention O O O O”. In this entity-type sequence, “O” stands for “Other.”

Entity-Level Representation

We first deploy cTAKES [42], a popular clinical NLP tool, to extract entity mentions from text snippets, and then propose 2 methods to obtain the entity-level representations of the text snippets according to the extracted entity mentions, as shown in Figure 2. cTAKES can extract 9 kinds of entities: AnatomicalSiteMention, DiseaseDisorderMention, FractionAnnotation, MedicationMention, Predicate, ProcedureMention, RomanNumeralAnnotation, SignSymptomMention, and Temporal Information.

The second method for entity-level representation (entity II) first directly adopts entity representation learned by TransE [43] on an external knowledge graph (KG; Mesh in this study), and then applies average pooling operation on all entities individually in sentences a and b to get entity-level representations of a (denoted by eg_a) and b (denoted by eg_b) respectively, and finally aggregates their representations using equation 2.

$$E = \tanh(W_e[eg_a - eg_b; eg_a * eg_b] + b_e) \quad (2)$$

where “[;]” denotes concatenation operation, W_e is a weight matrix, and b_e is a bias vector.

MLP Layer

To aggregate the information of 3 modules, we concatenate them together:

$$f = [S; C; E] \quad (3)$$

Then, we use MLP (as shown in equation 4) to predict the STS score p_{score} of (a, b) as follows:

$$p_{\text{score}} = \text{MLP}(Wf + b) \quad (4)$$

where W is a weight matrix, and b is a bias vector.

The loss function used in our model is the minimum square error (MSE) function:

$$\text{Loss} = \text{MSE}(p_{\text{score}} - g_{\text{score}}) \quad (5)$$

where g_{score} is the gold-standard score.

Experimental Setting

Before conducting experiments, we preprocess the corpus using the following simple rules: (1) convert clinical text snippets into lowercase; (2) tokenize clinical text snippets using special symbols, such as “[”, “]”, “/”, “,”, and “.”, and keep them unstained in some situations such as “.” in decimals. The hyperparameters of our model are shown in Table 2. Other parameters are optimized via fivefold cross validation on the training set. The pretrained BERT model used for text snippet pair representation in our experiments is [BERT-Base, Uncased] [44]. We train all model parameters simultaneously, set epochs as 12, and save the last checkpoints as the final models. The performance of all models is measured by PCC.

Table 2. Hyperparameters setting of our model.

Parameters	Value
Learning rate	2×10^{-5}
Sequence length of BERT ^a	380
Epochs	12
Batch size	20
Knowledge graph embedding dimension d	100
Character-level kernel size	3
Convolution kernels of BCNN ^b	50
Kernel size of BCNN	3
Word embedding dimension of entity I	50

^aBERT: bidirectional encoder representation from transformers.

^bBCNN: Basic bi-CNN.

Results

Table 3 shows the overall results of our proposed model. Our model achieves the highest PCC of 0.868, which is competitive with other state-of-the-art models proposed for the 2019 n2c2/OHNLP track on ClinicalSTS. The model using entity II is better than that using entity I by 0.007 in PCC, indicating that entity II is a better supplement to BERT than entity I. When character-level representation is removed, the PCC of our model decreases to 0.859 (entity I) and 0.854 (entity II). When entity-level representation is removed, the PCC of our model decreases to 0.858. When both types of representations are removed, the PCC of our model further decreases to 0.848. The results indicate that both character-level representation and entity-level representation are supplementary to BERT. Although the improvements individually from entity I and character-level text snippet representation are more remarkable than entity II, the improvement from the combination of entity

I and character-level representation is much smaller than the combination of entity II and character-level representation. It is because both character-level representation and entity I come from text snippets, whereas entity II comes from external KG. The diversity between character-level representation and entity II is much larger than that between character-level representation and entity I. It is interesting that our model is not further improved when both entity I and entity II are considered in our model at the same time, which may be also because of the diversity.

Moreover, we investigate the effect of the domain-specific pretrained BERT models [45,46] on our model. We replace the pretrained BERT model in the general domain, [BERT-Base, Uncased] [44], by the pretrained BERT model in the clinical domain [45] to obtain a new model. The highest PCC of the new model is 0.872, which is slightly better than our previous model, indicating that the domain-specific pretrained BERT model is beneficial to our model.

Table 3. Pearson correlation coefficient of our model on the test set.

Model and setting	PCC ^a
Our model	
Entity I	0.861
Entity II	0.868 ^b
Entity I + Entity II	0.862
Without character -level text snippet representation	
Entity I	0.859
Entity II	0.854
Without entity-level representation	0.858
Without both	0.848

^aPCC: Pearson correlation coefficient.

^bThe highest PCC.

Discussion

Error Analysis

Although the proposed model achieves competitive performance, there are also some errors. To analyze these errors, we look into samples for which the difference between the predicted STS score and gold-standard similarity score is greater than 1.0 and find that the main errors can be classified into 2 types.

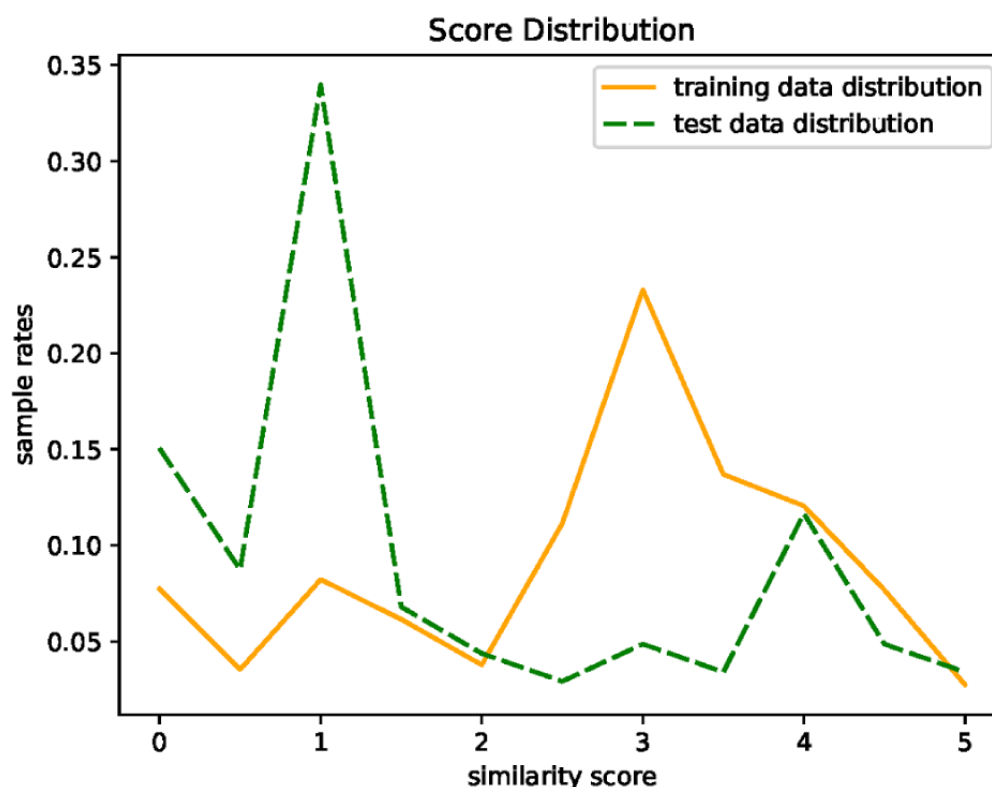
The first type of error is related to polarity of clinical text snippets as our model is insensitive to positive and negative words. For example, as shown in [Table 4](#), because both clinical text snippets in example 1 depict coughing up, their STS score predicted by our model is 2.5, but their gold-standard STS score is 1.0 as the polarity of the first text snippet is positive, whereas that of the second text snippet is negative. The second type of error is related to prescriptions that include medication names, usages, and dosages. For example, the gold-standard STS score of example 2 in [Table 4](#) is 1.0 as the medications in the 2 text

snippets are completely different, but the STS score of the example predicted by our model is 2.5 as some other words are the same in the 2 text snippets. Just because our model cannot extract medical information comprehensively, there are lots of errors of the second type. For further improvement, we need a comprehensive information extraction module to extract polarity information and medications with usage and dosage attributes besides the current 9 kinds of clinical entities. A possible way is to integrate the existing tools specifically for polarity information extraction (such as SenticNet [47]) or medication extraction (such as MedEx [48]) into our model. We also find that the scores of mispredictions are close to 2.5, which may be caused by the different STS score distributions of the training and test sets. As shown in [Figure 3](#), the STS scores of most sentence pairs in the training set concentrate in [2.5, 3.5], whereas those in the test set concentrate in [0.5, 1.5]. The difference is remarkable. It is reasonable to obtain the STS scores of mispredictions around the average score of the training set.

Table 4. Examples of errors on the test set.

Number	Example
1	<ul style="list-style-type: none"> • Sentence 1: respiratory: positive for coughing up mucus (phlegm), dyspnea and wheezing. • Sentence 2: negative for coughing up blood and dry cough. • Gold-standard: 1.0 • Predicted: 2.5
2	<ul style="list-style-type: none"> • Sentence 1: ibuprofen [motrin] 800 mg tablet 1 tablet by mouth four time a day as needed. • Sentence 2: lisinopril 10 mg tablet 1 tablet by mouth one time daily. • Gold-standard: 1.0 • Predict: 2.4

Figure 3. Similarity interval distribution in the training and test data sets.



Effect of Entity-Level Representation

Although the results in Table 3 show that any one of the 2 entity-level representations enhances the BERT-based model, some limitations also exist. In the case of entity I, we only consider type semantic information, but no entity semantic information. In the case of entity II, only about 20% (220/1080) of clinical entities recognized by cTAKES [42] can be mapped to Mesh via dictionary look-up. There are 2 directions for improvement: (1) introduce entity semantic information into entity I, and (2) improve entity mapping performance in entity II and find a larger KG instead of Mesh.

Conclusions

In this paper, we propose an enhanced BERT-based model for ClinicalSTS by introducing a character-level representation and an entity-level representation. Experiments on the 2019 n2c2/OHNL track on ClinicalSTS in 2019 indicate that both the character-level representation and the entity-level representation can enhance the BERT-based ClinicalSTS model, and our enhanced BERT-based model achieves competitive performance with other state-of-the-art models. In addition, domain-specific pretrained BERT models are better than general pretrained BERT models.

Acknowledgments

This paper is supported in part by grants: National Natural Science Foundations of China (U1813215, 61876052, and 61573118), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), National Natural Science Foundations of Guangdong, China (2019A1515011158), Guangdong Province Covid-19 Pandemic Control Research Fund (2020KZDZX1222), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20180306172232154 and JCYJ20170307150528934), and Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

Conflicts of Interest

None declared.

References

- Evans R. Electronic Health Records: Then, Now, and in the Future. *Yearb Med Inform* 2016 May 20;Suppl 1:S48-S61 [FREE Full text] [doi: [10.15265/IYS-2016-s006](https://doi.org/10.15265/IYS-2016-s006)] [Medline: [27199197](https://pubmed.ncbi.nlm.nih.gov/27199197/)]
- Markel A. Copy and paste of electronic health records: a modern medical illness. *Am J Med* 2010 May;123(5):e9. [doi: [10.1016/j.amjmed.2009.10.012](https://doi.org/10.1016/j.amjmed.2009.10.012)] [Medline: [20399309](https://pubmed.ncbi.nlm.nih.gov/20399309/)]
- Kettl PA. A Piece of My Mind. *JAMA* 1992 Feb 12;267(6):798. [doi: [10.1001/jama.1992.03480060040014](https://doi.org/10.1001/jama.1992.03480060040014)]

4. Wu H, Toti G, Morley K, Ibrahim Z, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J Am Med Inform Assoc* 2018 May 01;25(5):530-537 [FREE Full text] [doi: [10.1093/jamia/ocx160](https://doi.org/10.1093/jamia/ocx160)] [Medline: [29361077](https://pubmed.ncbi.nlm.nih.gov/29361077/)]
5. Hanauer DA, Wu DT, Yang L, Mei Q, Murkowski-Steffy KB, Vydyswaran VV, et al. Development and empirical user-centered evaluation of semantically-based query recommendation for an electronic health record search engine. *J Biomed Inform* 2017 Mar;67:1-10 [FREE Full text] [doi: [10.1016/j.jbi.2017.01.013](https://doi.org/10.1016/j.jbi.2017.01.013)] [Medline: [28131722](https://pubmed.ncbi.nlm.nih.gov/28131722/)]
6. Plaza L, Díaz A. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs. In: Hopfe CJ, Rezgui Y, Métails E, Preece A, Li H, editors. *Natural Language Processing and Information Systems. NLDB 2010. Lecture Notes in Computer Science*, vol 6177. Berlin, Germany: Springer; 2010:293-303.
7. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino J, et al. AskHERMES: An online question answering system for complex clinical questions. *J Biomed Inform* 2011 Apr;44(2):277-288 [FREE Full text] [doi: [10.1016/j.jbi.2011.01.004](https://doi.org/10.1016/j.jbi.2011.01.004)] [Medline: [21256977](https://pubmed.ncbi.nlm.nih.gov/21256977/)]
8. Demner-Fushman D, Lin J. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics*. 2006 Jul Presented at: 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics Association for Computational Linguistics; 2006; Sydney, Australia p. 841-848 URL: <https://dl.acm.org/doi/10.3115/1220175.1220281> [doi: [10.3115/1220175.1220281](https://doi.org/10.3115/1220175.1220281)]
9. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J Am Med Inform Assoc* 2019 Nov 01;26(11):1163-1171 [FREE Full text] [doi: [10.1093/jamia/ocz163](https://doi.org/10.1093/jamia/ocz163)] [Medline: [31562516](https://pubmed.ncbi.nlm.nih.gov/31562516/)]
10. Agirre E, Cer D, Diab M, Gonzalez-Agirre A. Semeval-2012 task 6: A pilot on semantic textual similarity. In: *SemEval '12: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. New York, NY: ACM; 2012 Jun 7 Presented at: SemEval '12; 2012; Montréal, Canada p. 385-393 URL: <https://dl.acm.org/doi/10.5555/2387636.2387697> [doi: [10.18653/v1/s17-2001](https://doi.org/10.18653/v1/s17-2001)]
11. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. SEM 2013 shared task: Semantic Textual Similarity. 2013 Jun 13 Presented at: *Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*; 2013; Atlanta, GA p. 32-43. [doi: [10.18653/v1/s17-2001](https://doi.org/10.18653/v1/s17-2001)]
12. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. 2014 Aug 23 Presented at: *International Conference on Computational Linguistics (COLING)*; August 23-24, 2014; Dublin, Ireland p. 81-91. [doi: [10.3115/v1/s14-2010](https://doi.org/10.3115/v1/s14-2010)]
13. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. 2015 Jun 4 Presented at: *Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*; June 4-5, 2015; Denver, CO p. 252-263. [doi: [10.18653/v1/S15-2045](https://doi.org/10.18653/v1/S15-2045)]
14. Agirre E, Banea C, Cer D, Diab M, González-Agirre A, Mihalcea R, et al. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. 2016 Jun 2 Presented at: *Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*; June 16-17, 2016; San Diego, CA p. 497-511. [doi: [10.18653/v1/s16-1081](https://doi.org/10.18653/v1/s16-1081)]
15. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. *Lang Resources & Evaluation* 2018 Oct 24;54(1):57-72. [doi: [10.1007/s10579-018-9431-1](https://doi.org/10.1007/s10579-018-9431-1)]
16. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L, Shen F, et al. Overview of the BioCreative/OHNLN challenge 2018 Task 2: Clinical Semantic Textual Similarity. In: *Proceedings of the BioCreative/OHNLN Challenge 2018*. 2018 Presented at: *BioCreative/OHNLN Challenge 2018*; December, 2018; Washington, DC. [doi: [10.1145/3233547.3233672](https://doi.org/10.1145/3233547.3233672)]
17. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLN Track on Clinical Semantic Textual Similarity: Overview. *JMIR Med Inform* 2020 Nov 27;8(11):e23375 [FREE Full text] [doi: [10.2196/23375](https://doi.org/10.2196/23375)] [Medline: [33245291](https://pubmed.ncbi.nlm.nih.gov/33245291/)]
18. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019 Presented at: *Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*; 2019; Minneapolis, MN p. 4171-4186. [doi: [10.18653/v1/n19-1423](https://doi.org/10.18653/v1/n19-1423)]
19. Lipscomb C. Medical Subject Headings (MeSH). *Bull Med Libr Assoc* 2000 Jul;88(3):265-266 [FREE Full text] [Medline: [10928714](https://pubmed.ncbi.nlm.nih.gov/10928714/)]
20. Mueller J, Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press; 2016 Presented at: *Thirtieth AAAI Conference on Artificial Intelligence*; February 12-17, 2016; Phoenix, AZ p. 2786-2792.
21. Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks. 2016 Aug 11 Presented at: *5th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2020*; August 11, 2016; Berlin, Germany p. 148-157. [doi: [10.18653/v1/w16-1617](https://doi.org/10.18653/v1/w16-1617)]

22. Hu B, Lu Z, Li H, Chen Q. Convolutional Neural Network Architectures for Matching Natural Language Sentences. 2014 Dec 8 Presented at: Neural Information Processing Systems (NeurIPS); December 8-13, 2014; Montreal, Quebec, Canada p. 2042-2050 URL: <http://papers.nips.cc/paper/5550-convolutional-neural-network-architectures-for-matching-natural-language-sentences.pdf>
23. Wang K, Yang B, Xu G, He X. Medical Question Retrieval Based on Siamese Neural Network Transfer Learning Method. In: Database Systems for Advanced Applications. Cham, Switzerland: Springer International Publishing; Apr 24, 2019:49-64.
24. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark; 2017 Sep Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics; Stroudsburg, PA. [doi: [10.18653/v1/d17-1070](https://doi.org/10.18653/v1/d17-1070)]
25. He H, Gimpel K, Lin J. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015. Stroudsburg, PA: Association for Computational Linguistics; 2015 Sep 17 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP 2015); September 17-21, 2015; Lisbon, Portugal p. 1576-1586. [doi: [10.18653/v1/d15-1181](https://doi.org/10.18653/v1/d15-1181)]
26. Kim S, Kang I, Kwak N. Semantic Sentence Matching with Densely-Connected Recurrent and Co-Attentive Information. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul 17 Presented at: AAAI Conference on Artificial Intelligence; January 27 to February 1, 2019; Honolulu, HI p. 6586-6593. [doi: [10.1609/aaai.v33i01.33016586](https://doi.org/10.1609/aaai.v33i01.33016586)]
27. Yin W, Schütze H, Xiang B, Zhou B. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. TACL 2016 Dec;4:259-272. [doi: [10.1162/tacl_a_00097](https://doi.org/10.1162/tacl_a_00097)]
28. Wang Z, Hamza W, Florian R. Bilateral Multi-Perspective Matching for Natural Language Sentences. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI). 2017 Aug 19 Presented at: Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI); August 19-25, 2017; Melbourne, Australia p. 4144-4150. [doi: [10.24963/ijcai.2017/579](https://doi.org/10.24963/ijcai.2017/579)]
29. He H, Lin J. Pairwise Word Interaction Modeling with Deep Neural Networks for Semantic Similarity Measurement. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics; 2016 Jun 12 Presented at: Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL); June 12-17, 2016; San Diego CA p. 937-948. [doi: [10.18653/v1/n16-1108](https://doi.org/10.18653/v1/n16-1108)]
30. Tan C, Wei F, Wang W, Lv W, Zhou M. Multiway Attention Networks for Modeling Sentence Pairs. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. 2018 Jul 13 Presented at: Twenty-Seventh International Joint Conference on Artificial Intelligence; July 13-19, 2018; Stockholm, Sweden p. 4411-4417. [doi: [10.24963/ijcai.2018/613](https://doi.org/10.24963/ijcai.2018/613)]
31. Gong Y, Luo H, Zhang J. Natural Language Inference over Interaction Space. 2018 Apr 30 Presented at: 6th International Conference on Learning Representations, ICLR 2018; May 3, 2018; Vancouver, BC, Canada URL: <https://openreview.net/forum?id=r1dHXnH6->
32. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019 Dec 8 Presented at: Neural Information Processing Systems (NeurIPS), 2019; December 8-14, 2019; Vancouver, BC, Canada p. 5754-5764 URL: <http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf>
33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019 Jul 26. URL: <https://arxiv.org/abs/1907.11692> [accessed 2019-07-26]
34. Liu Z, Chen Y, Tang B, Wang X, Chen Q, Li H, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. J Biomed Inform 2015 Dec;58 Suppl:S47-S52 [FREE Full text] [doi: [10.1016/j.jbi.2015.06.009](https://doi.org/10.1016/j.jbi.2015.06.009)] [Medline: [26122526](https://pubmed.ncbi.nlm.nih.gov/26122526/)]
35. Xiong Y, Shen Y, Huang Y, Chen S, Tang B, Wang X, et al. A Deep Learning-Based System for PharmaCoNER. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019. Stroudsburg, PA: Association for Computational Linguistics; 2019 Dec 4 Presented at: 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019; November 4, 2019; Hong Kong, China p. 33-37. [doi: [10.18653/v1/d19-5706](https://doi.org/10.18653/v1/d19-5706)]
36. Dong C, Zhang J, Zong C, Hattori M, Di H. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. Cham, Switzerland: Springer International Publishing; 2016 Dec 2 Presented at: Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016; December 2-6, 2016; Kunming, China p. 239-250. [doi: [10.1007/978-3-319-50496-4_20](https://doi.org/10.1007/978-3-319-50496-4_20)]
37. Niu J, Yang Y, Zhang S, Sun Z, Zhang W. Multi-task Character-Level Attentional Networks for Medical Concept Normalization. Neural Process Lett 2018 Jun 18;49(3):1239-1256. [doi: [10.1007/s11063-018-9873-x](https://doi.org/10.1007/s11063-018-9873-x)]
38. Clark K, Manning C. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016. Stroudsburg, PA: The Association for Computer Linguistics; 2016 Aug 7 Presented at: 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016; August 7-12, 2016; Berlin, Germany. [doi: [10.18653/v1/p16-1061](https://doi.org/10.18653/v1/p16-1061)]

39. Wu T, Wang Y, Wang Y, Zhao E, Yuan Y, Yang Z. Representation Learning of EHR Data via Graph-Based Medical Entity Embedding. arXiv. 2019 Oct 7. URL: <https://arxiv.org/abs/1910.02574> [accessed 2019-10-07]
40. Yang J, Liang S, Zhang Y. Design Challenges and Misconceptions in Neural Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018. 2018 Aug 20 Presented at: 27th International Conference on Computational Linguistics, COLING 2018; August 20-26, 2018; Santa Fe, NM p. 3879-3889 URL: <https://www.aclweb.org/anthology/C18-1327/>
41. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):67 [FREE Full text] [doi: [10.1186/s12911-017-0468-7](https://doi.org/10.1186/s12911-017-0468-7)] [Medline: [28699566](https://pubmed.ncbi.nlm.nih.gov/28699566/)]
42. Apache cTAKESTM - clinical Text Analysis Knowledge Extraction System. URL: <https://ctakes.apache.org/> [accessed 2020-03-22]
43. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating Embeddings for Modeling Multi-relational Data. In: Advances in Neural Information Processing Systems. 2013 Dec 5 Presented at: 27th Annual Conference on Neural Information Processing Systems 2013; December 5-8, 2013; Lake Tahoe, NV p. 2787-2795 URL: <http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf>
44. Google Research: BERT. 2020. URL: <https://github.com/google-research/bert> [accessed 2020-08-06]
45. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019. 2019 Aug 1 Presented at: 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019; August 1, 2019; Florence, Italy p. 58-65. [doi: [10.18653/v1/w19-5006](https://doi.org/10.18653/v1/w19-5006)]
46. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
47. Malheiros Y. senticnet: Access SenticNet data using Python Internet. URL: <https://github.com/yurimalheiros/senticnetapi> [accessed 2020-12-16]
48. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association 2010 Jan 01;17(1):19-24. [doi: [10.1197/jamia.m3378](https://doi.org/10.1197/jamia.m3378)]

Abbreviations

- BERT:** bidirectional encoder representation from transformers
- ClinicalSTS:** clinical semantic textual similarity
- CNN:** convolutional neural network
- EHR:** electronic health record
- KG:** knowledge graph
- MLP:** multilayer perceptron
- NLP:** natural language processing
- OHNLP:** Open Health Natural Language Processing
- OOV:** out of vocabulary
- PCC:** Pearson correlation coefficient
- SemEval:** Semantic Evaluation
- STS:** semantic textual similarity

Edited by Y Wang; submitted 10.08.20; peer-reviewed by X Yang, M Manzanara, M Memon; comments to author 22.09.20; revised version received 10.11.20; accepted 16.11.20; published 29.12.20

Please cite as:

Xiong Y, Chen S, Chen Q, Yan J, Tang B

Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study

JMIR Med Inform 2020;8(12):e23357

URL: <http://medinform.jmir.org/2020/12/e23357/>

doi: [10.2196/23357](https://doi.org/10.2196/23357)

PMID: [33372664](https://pubmed.ncbi.nlm.nih.gov/33372664/)

©Ying Xiong, Shuai Chen, Qingcai Chen, Jun Yan, Buzhou Tang. Originally published in JMIR Medical Informatics (<http://medinform.jmir.org>), 29.12.2020. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <http://medinform.jmir.org/>, as well as this copyright and license information must be included.