Original Paper

Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models

Xi Yang, PhD; Xing He, MSc; Hansi Zhang, MSc; Yinghan Ma, BSc; Jiang Bian, PhD; Yonghui Wu, PhD

Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States

Corresponding Author:
Yonghui Wu, PhD
Department of Health Outcomes and Biomedical Informatics
University of Florida
2004 Mowry Road
Gainesville, FL, 32610
United States
Phone: 1 352 294 8436
Email: yonghui.wu@ufl.edu

Abstract

Background: Semantic textual similarity (STS) is one of the fundamental tasks in natural language processing (NLP). Many shared tasks and corpora for STS have been organized and curated in the general English domain; however, such resources are limited in the biomedical domain. In 2019, the National NLP Clinical Challenges (n2c2) challenge developed a comprehensive clinical STS dataset and organized a community effort to solicit state-of-the-art solutions for clinical STS.

Objective: This study presents our transformer-based clinical STS models developed during this challenge as well as new models we explored after the challenge. This project is part of the 2019 n2c2/Open Health NLP shared task on clinical STS.

Methods: In this study, we explored 3 transformer-based models for clinical STS: Bidirectional Encoder Representations from Transformers (BERT), XLNet, and Robustly optimized BERT approach (RoBERTa). We examined transformer models pretrained using both general English text and clinical text. We also explored using a general English STS dataset as a supplementary corpus in addition to the clinical training set developed in this challenge. Furthermore, we investigated various ensemble methods to combine different transformer models.

Results: Our best submission based on the XLNet model achieved the third-best performance (Pearson correlation of 0.8864) in this challenge. After the challenge, we further explored other transformer models and improved the performance to 0.9065 using a RoBERTa model, which outperformed the best-performing system developed in this challenge (Pearson correlation of 0.9010).

Conclusions: This study demonstrated the efficiency of utilizing transformer-based models to measure semantic similarity for clinical text. Our models can be applied to clinical applications such as clinical text deduplication and summarization.

(JMIR Med Inform 2020;8(11):e19735) doi: 10.2196/19735

KEYWORDS

clinical semantic textual similarity; deep learning; natural language processing; transformers

Introduction

Semantic textual similarity (STS) is a natural language processing (NLP) task to quantitatively assess the semantic similarity between two text snippets. STS is usually approached as a regression task where a real-value score is used to quantify the similarity between two text snippets. STS is a fundamental NLP task for many text-related applications, including text deduplication, paraphrasing detection, semantic searching, and question answering. In the general English domain, semantic evaluation (SemEval) STS shared tasks have been organized annually from 2012 to 2017 [1-6], and STS benchmark datasets were developed for evaluation [6]. Previous work on STS often used machine learning models [7-9] such as support vector machine [10], random forest [11], convolutional neural networks [12], and recurrent neural networks [13] and topic modeling techniques [8] such as latent semantic analysis [14] and latent Dirichlet allocation [15]. Recently, deep learning models based on transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT) [16], XLNet [17], and Robustly optimized BERT approach (RoBERTa) [18] have

demonstrated state-of-the-art performances on the STS benchmark dataset [19] and remarkably outperformed the previous models. More recently, the Text-to-Text Transfer Transformer model [20] and the StructBERT model [21] have further improved the performance on the STS benchmark. These studies demonstrated the efficiency of transformer-based models for STS tasks.

Rapid adoption of electronic health record (EHR) systems has made longitudinal health information of patients available electronically [22,23]. EHRs consist of structured, coded data and clinical narratives. The structured EHR data are typically stored as predefined medical codes (eg, International Classification of Diseases, 9th/10th Revision, codes for diagnoses) in relational databases. Various common data models were used to standardize EHR data to facilitate downstream research and clinical studies [24]. However, clinical narratives are often documented in a free-text format, which contains many types of detailed patient information, such as family history, adverse drug events, and medical imaging result interpretations, that are not well captured in the structured medical codes [25]. As free text, the clinical notes may contain a considerable amount of duplication, error, and incompleteness for various reasons (eg, copy-and-paste or using templates and inconsistent modifications) [26,27]. STS can be applied to assess the quality of the clinical notes and reduce redundancy to support downstream NLP tasks [28]. However, up until now, only a few studies [29-31] have explored STS in the clinical domain due to the limited data resources for developing and benchmarking clinical STS tasks. Recently, a team at the Mayo Clinic developed a clinical STS dataset, MedSTS [32], which consists of more than 1000 annotated sentence pairs extracted from clinical notes. Based on the MedSTS dataset, the 2018 BioCreative/Open Health NLP (OHNLP) challenge [33] was organized as the first shared task examining advanced NLP methods for STS in the clinical domain. In this challenge, two different teams explored various machine learning approaches, including several deep learning models [30,31]. Later, more teams competed in the 2019 National NLP Clinical Challenges (n2c2)/OHNLP STS challenge with a larger clinical STS dataset [34]. During this challenge, many new emerging NLP techniques, such as transformer-based models, were explored.

This study presents our machine learning models developed for the 2019 n2c2/OHNLP STS challenge. We explored state-of-the-art transformer-based models (BERT, XLNet, and RoBERTa) for clinical STS. We systematically examined transformer models pretrained using general English corpora and compared them with clinical transformer models pretrained using clinical corpora. We also proposed a representation fusion method to ensemble the transformer-based models. In this challenge, our clinical STS system based on the XLNet model achieved a Pearson correlation score of 0.8864, ranked as the third-best performance among all participants. After the challenge, we further explored a new transformer-based model, RoBERTa, which improved the performance to 0.9065 and outperformed the best performance (0.9010) reported in this challenge. This study demonstrated the efficiency of transformer-based models for STS in the clinical domain.

Methods

Dataset

The 2019 n2c2 organizers developed a corpus of 2054 sentence pairs derived from over 300 million deidentified clinical notes from the Mayo Clinic's EHR data warehouse. The sentence pairs were divided into a training set of 1642 sentence pairs for model development and a test set of 412 sentence pairs for evaluation. Similar to the annotation scheme in the general English domain, the challenge corpus was annotated by assigning a similarity score for each sentence pair as a number on a scale from 0.0 to 5.0, where 0.0 indicates that the semantics of the two sentences are entirely independent (ie, no overlap in their meanings), and 5.0 signifies that two sentences are semantically equivalent. Annotators used arbitrary similarity scores between 0.0 and 5.0, such as 2.5 or 3.5, to reflect different levels of equality. Table 1 presents the descriptive statistics of the datasets. The distribution of similarity scores is quite different between the training and test datasets. In the training set, the range with the most cases (509/1642, 31.0%) was (3.0, 4.0], whereas in the test set, most scores (238/412, 57.8%) were distributed in the range (0.0, 1.0]. In this study, we denoted this challenge dataset as STS-Clinic. In addition to the STS-Clinic, we also used a general English domain STS benchmark dataset from the SemEval 2017 [6] as an external source. We merged the original training and development datasets to create a unique dataset of 7249 annotated sentence pairs. We denoted this combined general English domain dataset as STS-General and used it as a complementary training set for model development in this study. Compared to the STS-Clinic, the similarity scores in STS-General were more evenly distributed in different ranges (Table 1).

Table 1. Descriptive statistics of the datasets.

Dataset	Sentence pairs, n	Annotation distribution, n (%)				
		[0.0, 1.0]	(1.0, 2.0]	(2.0, 3.0]	(3.0, 4.0]	(4.0, 5.0]
STS-Clinic[a] Training	1642	312 (19.0)	154 (9.4)	394 (24.0)	509 (31.0)	273 (16.6)
STS-Clinic Test	412	238 (57.8)	46 (11.2)	32 (7.8)	62 (15.0)	34 (8.3)
STS-General Training	7249	1492 (20.6)	1122 (15.5)	1413 (19.5)	1260 (17.4)	1962 (27.1)

[a]STS: semantic textual similarity.

Preprocessing of Sentence Pairs

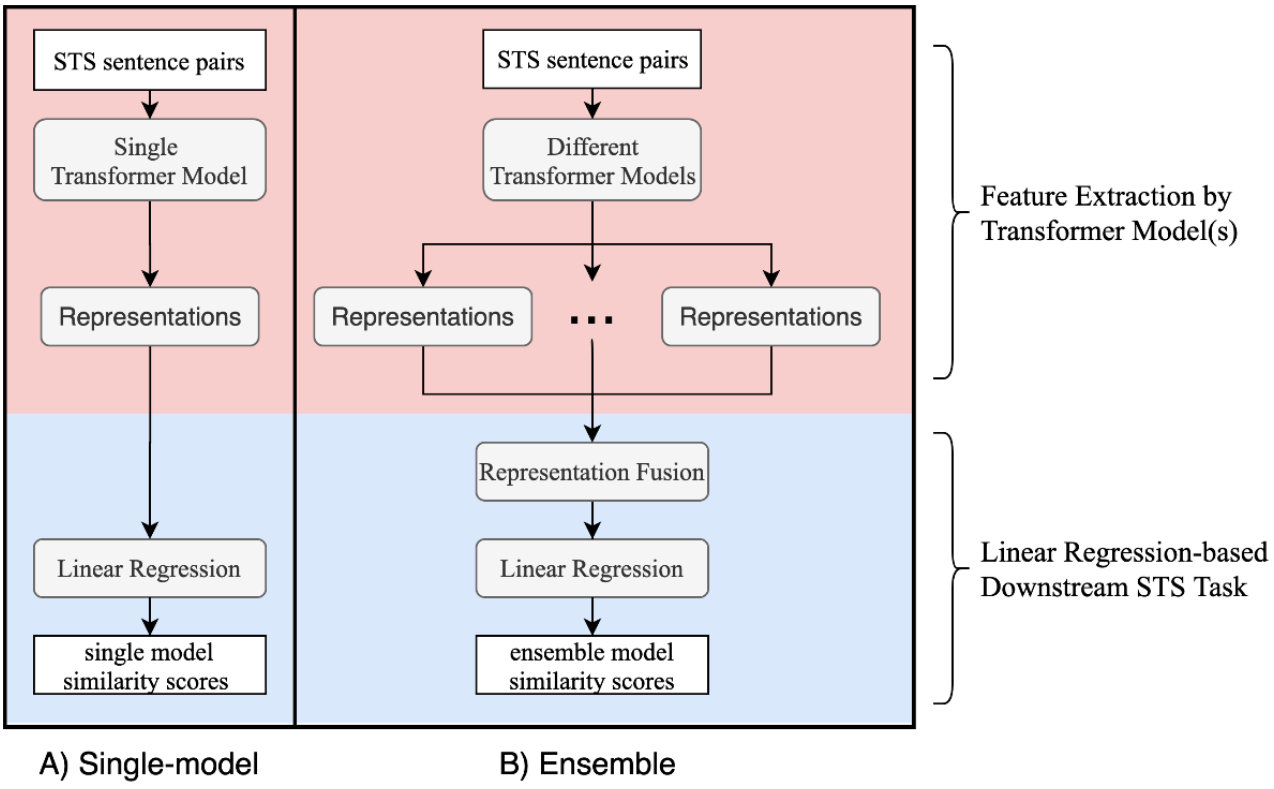
We developed a preprocessing pipeline to normalize each sentence pair, including (1) converting all words to lower case; (2) inserting white spaces to separate words from punctuation (eg, "[ab/cd]" → "[ab / cd]"; "abc,def" → "abc , def"); and (3) replacing two or more spaces or tabs ("\t") with a single space. We did not remove any stop-words from the sentences and kept the original formats of the numbers without any conversion. Since different transformer models adopted different tokenization strategies (eg, WordPiece for BERT, byte pair encoding for RoBERTa, and SentencePiece for XLNet), our preprocessing automatically picked the appropriate tokenizer according to the transformer model in use.

Transformer Model-Based STS System

In this study, we investigated three transformer models (BERT, XLNet, and RoBERTa) for clinical STS. BERT is a bidirectional transformer-based encoder model pretrained with a combination of masked language modeling (MLM) and next sentence prediction. RoBERTa has the same architecture as BERT but pretrained with a robust optimizing strategy. The RoBERTa pretraining procedure used dynamic MLM but removed the next sentence prediction task. XLNet is a transformer-based model pretrained with the bidirectional autoregressive language modeling method. Unlike the MLM used by BERT and RoBERTa, the autoregressive language model uses data permutation instead of data corruption and reconstruction. All three transformer models provided two different settings: a "BASE" setting and a "LARGE" setting. The main difference between the BASE model and the LARGE model is the number of layers. For example, the BERT-base model features 12 layers of transformer encoder layers, 768 hidden units in each layer, and 12 attention heads, while the BERT-large consists of 24 transformer blocks with a hidden size of 1024 and 16 attention heads. The total number of parameters for the BERT-large model is approximately 340 million, which is about 3 times more than the BERT-base model. In this study, we explored general transformers (pretrained using general English corpora) using both the BASE model and the LARGE model. We also examined clinical transformers pretrained using clinical notes from the MIMIC-III database. For clinical transformers, we adopted the BASE settings as we did not observe additional benefits from using the LARGE setting.

As shown in Figure 1, our STS system has two modules: (1) a transformer model–based feature learning module and (2) a regression-based similarity score learning module. In the feature learning module, transformer-based models were applied to learn distributed sentence-level representations from sentence pairs. In the similarity score learning module, we adopted a linear regression layer to calculate a similarity score between 0.0 and 5.0 according to the distributed representations derived from the transformers. We explored both single-model and ensemble solutions. Figure 1A shows the single-model solution where only one transformer-based model was used for feature representation learning. Figure 1B shows the ensemble solution where different transformer models were integrated. Ensemble learning is an efficient approach to aggregate different machine learning models to achieve better performance [35]. In this work, we tried different strategies to combine the distributed representations from two or three transformers as a new input layer for the similarity score learning module. We explored several methods to combine the distributed representations from different transformers, including (1) simple head-to-tail concatenation, (2) pooling, and (3) convolution.

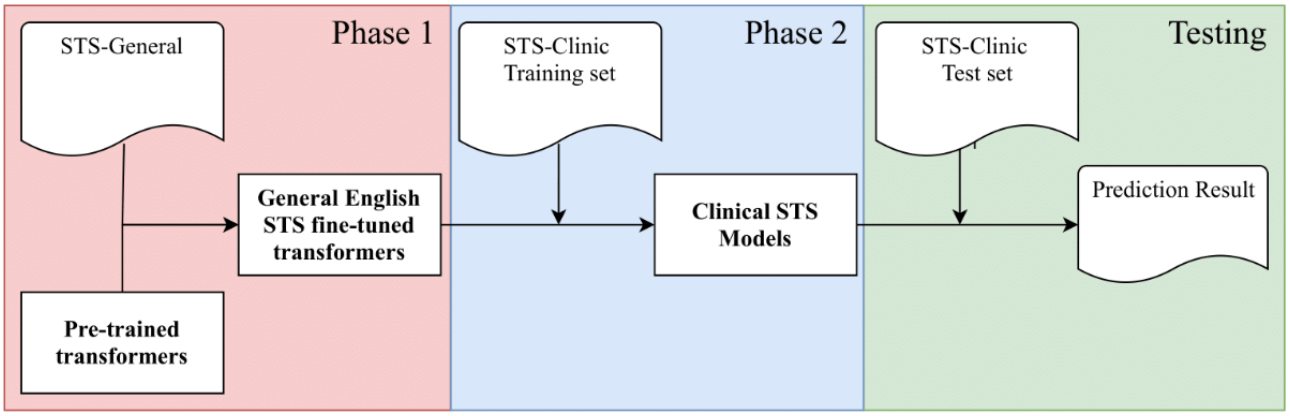
Figure 1. An overview of our single-model and ensemble solutions for clinical STS. STS: semantic textual similarity.



Training Strategy

As shown in Figure 2, we adopted a two-phase procedure to train our clinical STS models. In the first phase, an intermediate STS model was fine-tuned using the STS-General corpus. Subsequently, the intermediate model was further fine-tuned using the STS-Clinic corpus in phase 2. The fine-tuned model from the second phase was used for final testing. We used 5-fold cross-validation for hyperparameter optimization in both phase 1 and phase 2 training. We optimized the epoch number, batch size, and learning rate according to the cross-validation results.

Figure 2. The two-stage procedure for clinical STS model development. STS: semantic textual similarity.



Experiments and Evaluations

In this study, we implemented our STS system using the Transformers library developed by the HuggingFace team [36]. We also used the PyTorch-based general transformer models trained using general English corpora maintained by the HuggingFace team. The clinical transformer models were derived by further pretraining these general transformer models with clinical notes from the MIMIC-III database [37]. Table 2 shows the hyperparameters used for each transformer model. For evaluation, the results were calculated as the Pearson correlation scores using the official evaluation script provided by the 2019 n2c2/OHNLP challenge organizers. To report the *P* value for each Pearson correlation score, we adopted the SciPy package [38].

Table 2. Hyperparameters for transformer models.

Model	Number of epochs	Batch size	Learning rate[a]
BERT-base[b]	4	8	1.00E-05
BERT-mimic	3	8	1.00E-05
BERT-large	3	8	1.00E-05
XLNet-base	3	4	1.00E-05
XLNet-mimic	3	4	1.00E-05
XLNet-large	4	4	1.00E-05
RoBERTa-base[c]	3	4	1.00E-05
RoBERTa-mimic	3	4	1.00E-05
RoBERTa-large	3	4	1.00E-05
BERT-large + XLNet-large	4	8	1.00E-05
BERT-large + RoBERTa-large	3	4	1.00E-05
RoBERTa-large + XLNet-large	4	4	1.00E-05
BERT-large + XLNet-large + RoBERTa-large	3	2	1.00E-05

[a]The learning rate is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function [39].

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]RoBERTa: Robustly optimized BERT approach.

Results

Table 3 compares the performance of the different transformer models on the test dataset. The RoBERTa-large model achieved the best Pearson correlation of 0.9065 among all models, which outperformed the two models we developed and submitted during the challenge, including the XLNet-large (a Pearson correlation score of 0.8864) and the BERT-large models (a Pearson correlation score of 0.8549). For RoBERTa and XLNet, the models developed using the LARGE setting pretrained using general English corpora achieved better performances than their BASE settings (0.9065 vs 0.8778 for RoBERTa; 0.8864 vs 0.8470 for XLNet, respectively), whereas the BERT-base achieved a Pearson correlation score of 0.8615 that outperformed the BERT-large model's score of 0.8549. For all transformers, the models pretrained using general English corpora (in both LARGE settings and BASE settings) outperformed their corresponding clinical models pretrained using clinical notes from the MIMIC-III database. Among the ensemble models, the BERT-large + RoBERTa-large model achieved the best Pearson correlation score of 0.8914, which is remarkably lower than the best model, RoBERTa-large. We also observed that the performances of ensemble models were often in between the two individual models (eg, BERT-large + RoBERTa-large achieved 0.8914, which is between the BERT-large score of 0.8549 and RoBERTa-large score of 0.9065). The ensemble model of all three transformers achieved a Pearson correlation of 0.8452, which was even worse.

Table 3. Performances of the Pearson correlation on the test set.

Model	Pearson correlation on test set	*P* value
BERT-base[a]	0.8615	<.001
BERT-mimic	0.8521	<.001
BERT-large[b]	0.8549	<.001
XLNet-base	0.8470	<.001
XLNet-mimic	0.8286	<.001
XLNet-large[b,c]	0.8864	<.001
RoBERTa-base[d]	0.8778	<.001
RoBERTa-mimic	0.8705	<.001
RoBERTa-large	*0.9065*	<.001
BERT-large + XLNet-large[b]	0.8764	<.001
BERT-large + RoBERTa-large	0.8914	<.001
RoBERTa-large + XLNet-large	0.8854	<.001
BERT-large + XLNet-large + RoBERTa-large	0.8452	<.001

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]The challenge submissions.

[c]The best challenge submission (ranked 3rd).

[d]RoBERTa: Robustly optimized BERT approach.

Discussion

Principal Results

Clinical STS is a fundamental task in biomedical NLP. The 2019 n2c2/OHNLP shared task was organized to solicit state-of-the-art STS algorithms in the clinical domain. We participated in this challenge and developed a deep learning–based system using transformer-based models. Our best submission (XLNet-large) achieved the third-best performance (a Pearson correlation score of 0.8864) among the 33 teams. Based on our participation, we further explored RoBERTa models and improved the performance to 0.9065 (RoBERTa-large), demonstrating the efficiency of transformer models for clinical STS. We also further explored three different ensemble strategies to develop ensembled models using transformers. Our experimental results show that the ensemble methods did not outperform the unified individual models. Another interesting finding is that the transformers pretrained using the clinical notes from the MIMIC-III database did not outperform the general transformers pretrained using general English corpora on clinical STS. One possible reason might be that the clinical corpora we used for training are relatively small compared with the general English corpus. Further investigation examining these findings is warranted.

Experiment Findings

Although previous studies [40-44] have shown that pretraining transformer models with domain-specific corpora could enhance their performances in domain-related downstream tasks (such as clinical concept extraction), our results in this study indicated that this strategy might not be helpful for clinical STS. For all three types of transformers explored in this study, the models pretrained using general English text consistently obtained

higher scores than the corresponding models pretrained using clinical text. For example, the Pearson correlation score achieved by the RoBERTa-mimic was 0.8705; however, the RoBERTa-base yielded a higher performance of 0.8778. Tawfik et al [45] have similarly observed that the PubMed pretrained BioBERT did not outperform the corresponding general BERT model pretrained using English text on clinical STS.

In the clinical STS task, using STS-General (an STS corpus annotated in the general English domain) as an extra training set in addition to STS-Clinic could efficiently improve performances for transformer-based models. Taking the RoBERTa model as an example, the RoBERTa-large fine-tuned using only the clinical text (ie, STS-Clinic) achieved a Pearson correlation score of 0.8720; however, the same model fine-tuned with both the general English text (ie, STS-General) and clinical text (ie, STS-Clinic) achieved a score of 0.9065 (approximately 0.035 higher). We observed similar results for BERT and XLNet. Without Phase 1 (Figure 2), the BERT-large and XLNet-large models achieved Pearson correlation scores of 0.8413 and 0.8626, respectively, which are lower than the results we submitted (0.8549 and 0.8864) using two-phase training. We looked into the training datasets for possible reasons. Although the STS-General and STS-Clinic were extracted from different domains, there are common contents shared between them. First, the annotation guidelines between the two datasets were highly aligned. For both datasets, the annotation scale is from 0.0 to 5.0, and each score reflects the same similarity level. Since the two STS datasets were annotated by different annotators, subjective annotation bias might be introduced (eg, the judgement and agreement of semantic similarity among annotators might be different in the two datasets). However, our experiment results showed that training with both datasets improved the performance despite the potential annotation bias. Second, a considerable portion of STS-Clinic sentence pairs are common descriptions that do not require comprehensive clinical knowledge to interpret the semantics. Typical examples include sentences extracted from Consultation Note or Discharge Summary as follows:

> *Plan: the patient stated an understanding of the program, and agrees to continue independently with a home management program.*
>
> *Thank you for choosing the name M.D. care team for your health care needs!*

On the other hand, there are many sentences in the STS-General associated with healthcare. An example is exhibited below:

> *Although obesity can increase the risk of health problems, skeptics argue, so do smoking and high cholesterol.*

Tang et al [30] have demonstrated that combining representations derived from different models is an efficient strategy in clinical STS. We explored similar strategies to combine sentence-level distributed representations, including vector concatenation, average pooling, max pooling, and convolution. Surprisingly, our results showed that such ensemble strategies did not help transformer-based STS systems. For example, for the ensemble model derived from the BERT-large and the XLNet-large models (ie, BERT-large + XLNet-large), the achieved Pearson correlation scores for vector concatenation, average pooling, max pooling, and convolution were 0.8764, 0.8760, 0.8799, and 0.8803, respectively. All the results were approximately 0.01 lower than that for XLNet-large (0.8864). We also observed that ensemble models' performances were consistently in between the two individual models (0.8549 for BERT-large and 0.8864 for XLNet-large). Future studies should examine this finding.

To examine the statistical significance among different models' results, we used a 1-tailed parametric test based on the Fisher Z-transformation [46], adopted in the previous SemEval STS shared tasks [2-4]. Our best model (ie, RoBERTa-large) achieved a statistically significant higher performance than most of our other solutions (see Multimedia Appendix 1) but was not significantly better than the models XLNet-large (P=.07), BERT-large + RoBERTa-large (P=.13), and RoBERTa-large + XLNet-large (P=.06). The significance analysis indicated that these four models performed very similarly to each other.

Error Analysis

We compared the system prediction from our best model (ie, RoBERTa-large) with the gold standards and identified sentence pairs with the largest discrepancy in terms of the similarity score. Among the top 50 sentence pairs, 26 of them had labeled scores in the range of 0.0 to 1.0, and only 6 sentence pairs had gold standard STS scores over 3.0. We further split the testing results into two subsets using a threshold score of 2.5 on gold standards and calculated the mean and median of the differences between the gold standards and predictions. For the subgroup consisting of sentence pairs with gold standard scores over 2.5, the mean and median of difference were 0.46 and 0.37. For the other subset (difference≤2.5), the mean and median of difference were 0.69 and 0.66. Therefore, it was more challenging for the system to predict appropriate STS scores for sentence pairs with low similarity (gold standard score≤2.5) than for those with high similarity.

We also observed that sentence pairs with high similarity scores usually have a similar sentence structure where many words occur in both sentences. Therefore, we hypothesized that the STS models will assign higher scores to sentence pairs that share a large portion of their lexicons and similar syntax. To test our hypothesis, we adopted the BertViz package [47] to profile the attention pattern of the RoBERTa-large model (ie, our best STS model). BertViz can generate the attention pattern between two sentences by linking words via lines, where the line weights reflect the attention weights; higher line weights indicate higher attention weights between the two words. Table 4 and Figure 3 show an example for two sentence pairs on a similar topic from the training and test sets. In the first example from the training set, the attention pattern has three dominant attention weights (eg, "questions-questions") and the similarity score for this sentence pair is labeled as 5.0. However, the attention pattern for the sentence pair from the test set also has similar dominant attention weights (such as "questions-questions") but was labeled with a similarity score of 0.0.

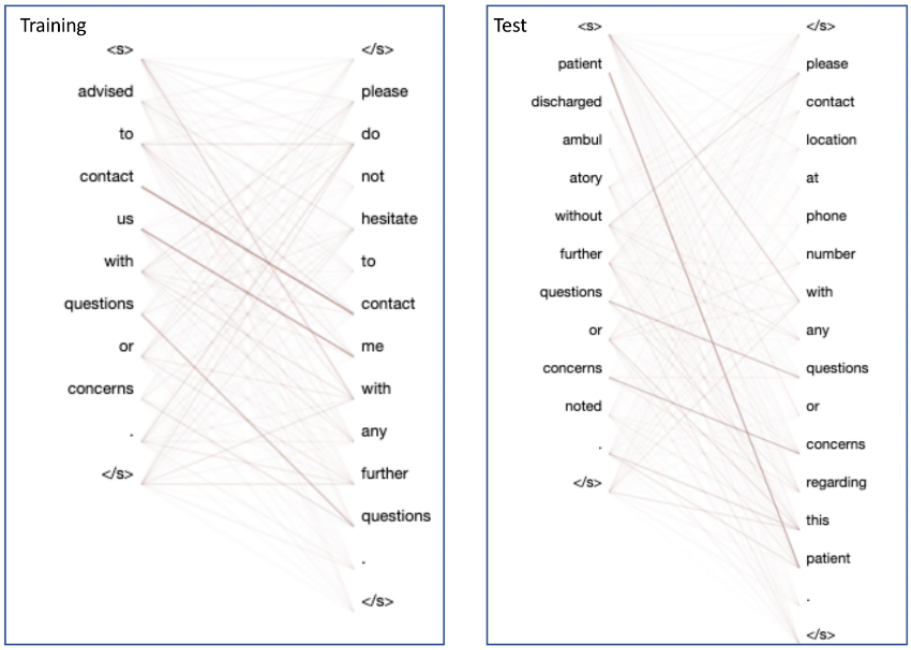
Table 4. Transformer model attention visualization on two examples from STS-Clinic.

Category	Sentence pair	Gold standard	Prediction
Training	• S1[a]: advised to contact us with questions or concerns. • S2: please do not hesitate to contact me with any further questions.	5	N/A[b]
Test	• S1: patient discharged ambulatory without further questions or concerns noted. • S2: please contact location at phone number with any questions or concerns regarding this patient.	0	2.5

[a]S: sentence.

[b]N/A: not applicable.

Figure 3. Transformer model attention visualization on two examples from STS-Clinic. STS: semantic textual similarity.



Limitations

This study has limitations. First, it is worth exploring methods to effectively integrate clinical resources with general English resources in transformer-based models. In this study, we explored an approach by pretraining transformer-based models with a clinical corpus (ie, MIMIC-III corpus). However, our results showed that this approach was not efficient. Therefore, new strategies to better integrate medical resources are needed. Second, our clinical STS systems performed better for sentence pairs with high similarity scores (ie, similarity score≥3 in gold standard) whereas, for the sentence pairs with low similarity scores (ie, similarity score<2 in gold standard), our systems still need to be improved. How to address this issue is one of our future focuses.

Conclusions

In this study, we demonstrated transformer-based models for measuring clinical STS and developed a system that can use various transformer algorithms. Our experiment results show that the RoBERTa model achieved the best performance compared to other transformer models. Our study demonstrated the efficiency of transformer-based models for assessing the semantic similarity for clinical text. Our models and system could be applied to various downstream clinical NLP applications. The source code, system, and pretrained models can be accessed on GitHub [48].

Acknowledgments

Research reported in this publication was supported by (1) the University of Florida Clinical and Translational Science Institute, which is supported in part by the National Institutes of Health (NIH) National Center for Advancing Translational Sciences under award number UL1TR001427; (2) the Patient-Centered Outcomes Research Institute under award number ME-2018C3-14754; (3) the Centers for Disease Control and Prevention under award number U18DP006512; (4) the NIH National Cancer Institute under award number R01CA246418; and (5) the NIH National Institute on Aging under award number R21AG061431-02S1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and Patient-Centered Outcomes Research Institute.

We would like to thank the n2c2 organizers for providing the annotated corpus and the guidance for this challenge. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the graphics processing units used for this research.

Authors' Contributions

XY, JB, and YW were responsible for the overall design, development, and evaluation of this study. YM, HZ, and XH were involved in conducting experiments and result analysis. XY, JB, and YW wrote and edited this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The 1-tailed parametric test results based on Fisher Z-transformation.
[XLSX File (Microsoft Excel File), 10 KB-Multimedia Appendix 1]

References

1. Agirre E, Cer D, Diab M, Gonzalez-Agirre A. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. : Association for Computational Linguistics; 2012 Presented at: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics; June 7-8, 2012; Montréal, Canada p. 385-393 URL: https://www.aclweb.org/anthology/S12-1051
2. Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. *SEM 2013 shared taskmantic Textual Similarity. : Association for Computational Linguistics; 2013 Presented at: *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics; June 13-14, 2013; Atlanta, USA p. 32-43 URL: https://www.aclweb.org/anthology/S13-1004
3. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. 2014 Presented at: The 8th International Workshop on Semantic Evaluation (SemEval 2014); Aug 23-24, 2014; Dublin, Ireland p. 81-91 URL: https://www.aclweb.org/anthology/S14-2010 [doi: 10.3115/v1/s14-2010]
4. 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 Presented at: The 9th International Workshop on Semantic Evaluation (SemEval 2015); June 4-5, 2015; Denver, USA p. 252-263 URL: https://www.aclweb.org/anthology/S15-2045/ [doi: 10.18653/v1/s15-2045]
5. Agirre E, Banea C, Cer D, Diab M, Gonzalez-Agirre A, Mihalcea R, et al. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. 2016 Presented at: The 10th International Workshop on Semantic Evaluation (SemEval-2016); June 16-17, 2016; San Diego, USA p. 497-511 URL: https://www.aclweb.org/anthology/S16-1081/ [doi: 10.18653/v1/S16-1081]
6. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. 2017 Presented at: The 11th International Workshop on Semantic Evaluation (SemEval-2017); Aug 3-4, 2017; Vancouver, Canada p. 1-14 URL: https://www.aclweb.org/anthology/S17-2001/ [doi: 10.18653/v1/s17-2001]
7. Farouk M. Measuring Sentences Similarity: A Survey. ArXiv 2019 Oct 06 [FREE Full text] [doi: 10.17485/ijst/2019/v12i25/143977]
8. Ramaprabha J, Das S, Mukerjee P. Survey on Sentence Similarity Evaluation using Deep Learning. In: J. Phys.: Conf. Ser. 2018 Apr 25 Presented at: National Conference on Mathematical Techniques and its Applications (NCMTA 18); Jan 5–6, 2018; Kattankulathur, India URL: https://iopscience.iop.org/article/10.1088/1742-6596/1000/1/012070 [doi: 10.1088/1742-6596/1000/1/012070]
9. Gomaa WH, Fahmy AA. A Survey of Text Similarity Approaches. IJCA 2013 Apr 18;68(13):13-18. [doi: 10.5120/11638-7118]
10. Béchara H, Costa H, Taslimipoor S, Gupta R, Orasan C, Corpas PG, et al. MiniExperts: An SVM Approach for Measuring Semantic Textual Similarity. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) Denver, Colorado: Association for Computational Linguistics; 2015 Presented at: The 9th International Workshop on Semantic Evaluation (SemEval 2015); Jun 4-5, 2015; Denver, USA p. 96-101 URL: https://www.aclweb.org/anthology/S15-2017/ [doi: 10.18653/v1/S15-2017]
11. Buscaldi D, Flores J, Ruiz I, Rodriguez I. SOPA: Random Forests Regression for the Semantic Textual Similarity task. 2015 Presented at: The 9th International Workshop on Semantic Evaluation (SemEval 2015); Jun 4-5, 2015; Denver, USA p. 132-137 URL: https://www.aclweb.org/anthology/S15-2024/ [doi: 10.18653/v1/s15-2024]
12. He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks. 2015 Presented at: The 2015 Conference on Empirical Methods in Natural Language Processing; Sept 19 – 21, 2015; Lisbon, Portugal p. 1576-1586 URL: https://www.aclweb.org/anthology/D15-1181/ [doi: 10.18653/v1/d15-1181]

13. Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. : AAAI Press; 2016 Presented at: The Thirtieth AAAI Conference on Artificial Intelligence; Feb 12–17, 2016; Phoenix, USA p. 2786-2792 URL: https://dl.acm.org/doi/10.5555/3016100.3016291 [doi: 10.5555/3016100.3016291]
14. Kashyap A, Han L, Yus R, Sleeman J, Satyapanich T, Gandhi S, et al. Robust semantic text similarity using LSA, machine learning, and linguistic resources. Lang Resources & Evaluation 2015 Oct 30;50(1):125-161. [doi: 10.1007/s10579-015-9319-2]
15. Niraula N, Banjade R, Ştefănescu D, Rus V. Experiments with Semantic Similarity Measures Based on LDA and LSA. 2013 Presented at: The First International Conference on Statistical Language and Speech Processing (SLSP 2013); July 29-31, 2013; Tarragona, Spain p. 188-199 URL: https://link.springer.com/chapter/10.1007/978-3-642-39593-2_17 [doi: 10.1007/978-3-642-39593-2_17]
16. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018 [FREE Full text]
17. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv 2019 [FREE Full text]
18. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019 [FREE Full text]
19. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv 2019 [FREE Full text]
20. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv 2019 [FREE Full text]
21. Wang W, Bi B, Yan M, Wu C, Bao Z, Xia J, et al. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. arXiv 2019 [FREE Full text]
22. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015 Mar 18;36:345-359. [doi: 10.1146/annurev-publhealth-031914-122747] [Medline: 25581157]
23. Obeid JS, Beskow LM, Rape M, Gouripeddi R, Black RA, Cimino JJ, et al. A survey of practices for the use of electronic health records to support research recruitment. J Clin Transl Sci 2017 Aug;1(4):246-252 [FREE Full text] [doi: 10.1017/cts.2017.301] [Medline: 29657859]
24. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
25. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
26. Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. AMIA Annu Symp Proc 2011;2011:1612-1620 [FREE Full text] [Medline: 22195227]
27. Wang MD, Khanna R, Najafi N. Characterizing the Source of Text in Electronic Health Record Progress Notes. JAMA Intern Med 2017 Aug 01;177(8):1212-1213 [FREE Full text] [doi: 10.1001/jamainternmed.2017.1548] [Medline: 28558106]
28. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Jt Summits Transl Sci Proc 2010 Mar 01;2010:1-5 [FREE Full text] [Medline: 21347133]
29. Soganciogl G, Öztürk H, Özgür A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics 2017 Jul 15;33(14):i49-i58 [FREE Full text] [doi: 10.1093/bioinformatics/btx238] [Medline: 28881973]
30. Xiong Y, Chen S, Qin H, Cao H, Shen Y, Wang X, et al. Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):72 [FREE Full text] [doi: 10.1186/s12911-020-1045-z] [Medline: 32349764]
31. Chen Q, Du J, Kim S, Wilbur WJ, Lu Z. Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):73 [FREE Full text] [doi: 10.1186/s12911-020-1044-0] [Medline: 32349758]
32. 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]
33. Rastegar-Mojarad M, Liu S, Wang Y, Afzal N, Wang L, Shen F, et al. BioCreative/OHNLP Challenge 2018. 2018 Presented at: The 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB); Aug 29 - Sept 1, 2018; Washington DC, USA p. 575 URL: https://doi.org/10.1145/3233547.3233672 [doi: 10.1145/3233547.3233672]
34. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. Overview of the 2019 n2c2/OHNLP Track on Clinical Semantic Textual Similarity. JMIR Medical Informatics 2020 [FREE Full text] [doi: 10.2196/23375]
35. Zhang C, Ma Y, editors. Ensemble Learning. In: Ensemble machine learning: methods and applications. New York, USA: Springer; 2012.
36. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv 2019 [FREE Full text]

37. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
38. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
39. Murphy K. Machine Learning: A Probabilistic Perspective. Cambridge, USA: MIT Press; 2012.
40. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, 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] [Medline: 31501885]
41. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv 2019 [FREE Full text]
42. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. arXiv 2019 [FREE Full text]
43. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. 2019 Presented at: The 18th BioNLP Workshop and Shared Task (BioBLP 2019); Aug 1, 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/w19-5006]
44. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 01;26(11):1297-1304. [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
45. Tawfik NS, Spruit MR. Evaluating sentence representations for biomedical text: Methods and experimental results. J Biomed Inform 2020 Apr;104:103396. [doi: 10.1016/j.jbi.2020.103396] [Medline: 32147441]
46. Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika 1915 May;10(4):507. [doi: 10.2307/2331838]
47. Vig J. A Multiscale Visualization of Attention in the Transformer Model. arXiv 2019 Jun 12 [FREE Full text] [doi: 10.18653/v1/p19-3007]
48. 2019 N2C2 Track-1 Clinical Semantic Textual Similarity. GitHub. URL: https://github.com/uf-hobi-informatics-lab/2019_N2C2_Track1_ClinicalSTS.git [accessed 2020-11-02]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
MIMIC-III: Medical Information Mart for Intensive Care
MLM: masked language modeling
n2c2: National Natural Language Processing Clinical Challenges
NIH: National Institutes of Health
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
RoBERTa: Robustly optimized BERT approach
SemEval: semantic evaluation
STS: semantic textual similarity

Edited by Y Wang; submitted 27.07.20; peer-reviewed by F Li, J Lei, A Mavragani; comments to author 06.10.20; revised version received 19.10.20; accepted 26.10.20; published 23.11.20

Please cite as:
Yang X, He X, Zhang H, Ma Y, Bian J, Wu Y
Measurement of Semantic Textual Similarity in Clinical Texts: Comparison of Transformer-Based Models
JMIR Med Inform 2020;8(11):e19735
URL: http://medinform.jmir.org/2020/11/e19735/
doi: 10.2196/19735
PMID: 33226350