Original Paper

ALBERT-Based Self-Ensemble Model With Semisupervised Learning and Data Augmentation for Clinical Semantic Textual Similarity Calculation: Algorithm Validation Study

Junyi Li, ME; Xuejie Zhang, PhD; Xiaobing Zhou, PhD

School of Information Science and Engineering, Yunnan University, Kunming, China

Corresponding Author:
Xiaobing Zhou, PhD
School of Information Science and Engineering
Yunnan University
East Outer Ring Road
Chenggong District, Kunming
Kunming, 650091
China
Phone: 86 87165031748
Email: zhouxb@ynu.edu.cn

Abstract

Background: In recent years, with increases in the amount of information available and the importance of information screening, increased attention has been paid to the calculation of textual semantic similarity. In the field of medicine, electronic medical records and medical research documents have become important data resources for clinical research. Medical textual semantic similarity calculation has become an urgent problem to be solved.

Objective: This research aims to solve 2 problems—(1) when the size of medical data sets is small, leading to insufficient learning with understanding of the models and (2) when information is lost in the process of long-distance propagation, causing the models to be unable to grasp key information.

Methods: This paper combines a text data augmentation method and a self-ensemble ALBERT model under semisupervised learning to perform clinical textual semantic similarity calculations.

Results: Compared with the methods in the 2019 National Natural Language Processing Clinical Challenges Open Health Natural Language Processing shared task Track on Clinical Semantic Textual Similarity, our method surpasses the best result by 2 percentage points and achieves a Pearson correlation coefficient of 0.92.

Conclusions: When the size of medical data set is small, data augmentation can increase the size of the data set and improved semisupervised learning can boost the learning efficiency of the model. Additionally, self-ensemble methods improve the model performance. Our method had excellent performance and has great potential to improve related medical problems.

(JMIR Med Inform 2021;9(1):e23086) doi: 10.2196/23086

KEYWORDS

data augmentation; semisupervised; self-ensemble; ALBERT; clinical semantic textual similarity; algorithm; semantic; model; data sets

Introduction

With the rapid development of computers and artificial intelligence, information availability has begun to show exponential growth. We are already in an era of information explosion. When faced with a large amount of information, time is wasted screening valid information. In addition, a large amount of information is stored in the form of text. Whether involving cluster storage or referring to related information, efficient information matching and screening is crucial. The importance of text information processing research has become very obvious. With major breakthroughs in the research of related algorithms in natural language processing and artificial intelligence, increasingly, research has been devoted to text information processing.

Textual similarity calculation [1] is a key technology for efficient information screening and matching in the field of text

processing. Previous work [2-8] has proposed some methods for textual similarity calculation, for example, traditional text similarity calculation methods [2], word similarity calculation [3], vector space model [4], and latent Dirichlet allocation model [5]. At present, with the development of deep learning and neural networks, methods based on neural networks have become popular, for example, word vector embedding method [6,7] and one-hot representation [8]. At the same time, these methods can also be clinically applied.

In the field of medicine, with the rapid increase in electronic medical data [9], electronic medical records and medical documents have become important data resources for medical clinical research. However, most of these data resources are stored unprocessed or in heterogeneous text formats. To understand the content of text data, it is necessary to integrate structured and heterogeneous clinical data resources, medical records, and scientific research documents. Similarity calculation can improve information retrieval performance for medical resources and effectively allow the integration of heterogeneous clinical data. The concept of semantic similarity evaluation is the key to understanding text data resources, which can effectively allow the processing, classification, and structured processing of those resources. For example, a semantic similarity method can be used to semantically analyze patient medical records to identify similar cases and find the best solution.

However, a large number of publicly available medical data sets are restricted because of privacy, and there are insufficient sources of medical data sets. The scarcity of data sets has led to the slow development of natural language processing (NLP) in the medical field. In recent years, more researchers have begun to pay attention to this issue. Therefore, competitions related to textual semantic similarity calculation have been produced, such as SemEval [10], to develop an automated method, and the 2019 National NLP Clinical Challenges (N2C2) Open Health Natural Language Processing (OHNLP) [11,12] shared task Track 1 on Clinical Semantic Textual Similarity (STS) [13], for systems based on semisupervised learning. An example of clinical STS is shown in Figure 1. The score indicates the similarity between the 2 sentences are and fall within an ordinal range, ranging from 0 to 5, where 0 means that the 2 sentences are completely different (ie, their meanings do not overlap) and 5 means that the 2 sentences have complete semantic equivalence.

Figure 1. An example from the Clinical STS.

Sentence 1:

nortriptyline [PAMELOR] 50 mg capsule 1 capsule by mouth every bedtime.

Sentence 2:

Tylenol Extra Strength 500 mg tablet 2 tablets by mouth every bedtime.

Score:

1

Teams that participated in the 2019 N2C2 OHNLP Clinical STS challenge demonstrated good results with methods such as multitask learning, XLNet, and ClinicalBERT methods. In the challenge, we used recursive neural networks and variants of these neural networks for experiments, such as long short-term memory neural networks [14], convolutional neural networks [15,16], capsule neural networks [17], and ordered long short-term memory neural networks. In addition, we combined some popular deep learning mechanisms, such as attention [18] and Siamese [19,20] networks. Through comparative experimental research, we obtained a Pearson correlation coefficient of 0.66 [21] in the official submission, which was not a satisfying result. Compared with other teams' methods, our model had 2 drawbacks. First, because the size of clinical data sets was small, there were not enough data to train the model, which led to insufficient learning and understanding of the model. Second, our model was based on a recurrent neural network. Due to the influence of the forget gate in the recurrent neural network, important information may be lost in the process of long-distance propagation, which prevents the model from extracting key information. As a result, the learning efficiency of the model decreased.

To address the abovementioned problems, this paper proposes a self-ensemble [22] ALBERT [23] model under semisupervised learning [24,25] with easy data augmentation (EDA) [26] to calculate the semantic similarity of clinical text.

Methods

Overview

In this section, we introduce 3 highlights of our method. Our method uses data augmentation and semisupervised learning to expand the scale of the data set from different levels. We pretrained ALBERT (based on self-ensemble methods) to strengthen the acquisition of key information and improve the performance of the model, and semisupervised learning and data augmentation methods were used to expand the number of data sets and increase the representation of data sets, which can prevent self-ensemble methods from overfitting.

Data Augmentation

By using external general domain data sets for semisupervised learning, we indirectly solved the problem of insufficient data. However, for medical data, semisupervised learning does not directly increase the amount of medical data. Therefore, we used an EDA method to directly increase the amount of medical data.

Generally, data augmentation is used in computer vision to flip, zoom, and add noise to a picture. These operations can increase small amounts of data, which can help train a more robust model; however, for text data, data augmentation is mainly used for operations such as replacing, adding, and deleting text. Previous work [27,28] has proposed some methods for data augmentation in NLP. For example, a study [27] translated sentences into French and then into English to generate new data. Other work has used data noising as smoothing [28]. However, these methods are highly time- and resource-consuming thus are not often used in practice.

In this paper, we use the form of EDA [26] shown in Table 1. Due to the irreplaceability of proper nouns in medical data, the selection range of the replacement operation has been optimized to keep proper nouns as much as possible. The size of medical data set increased from 1642 to 16,411 after EDA. We can intuitively see a substantial increase in the amount of medical data. We verified that this method increases the size of data set.

Table 1. Sentences generated using EDA.

Operation	Sentence 1	Sentence 2	Sentence 3
None[a]	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours as needed.	A lady is running her cute dog through an agility course.	A beautiful woman with a young girl pose with bear statues in front of a store.
Synonym replacement	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours as indeed.	A lady is running her cute dog through an legerity course.	A beautiful woman with a young girl pose with bear figurines in front of a store.
Random insertion	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by every mouth every 4 hours as needed.	A lady is running her cute dog through an amazing agility course.	A beautiful woman with a young girl pose with lovely bear statues in front of a store.
Random deletion	oxycodone [ROXICODONE] 5 mg tablet 0.5-1 tablets by mouth every 4 hours.	A lady is running her dog through an agility course.	A woman with a young girl pose with bear statues in front of a store.

[a]None indicates that this sentence did not undergo any operation.

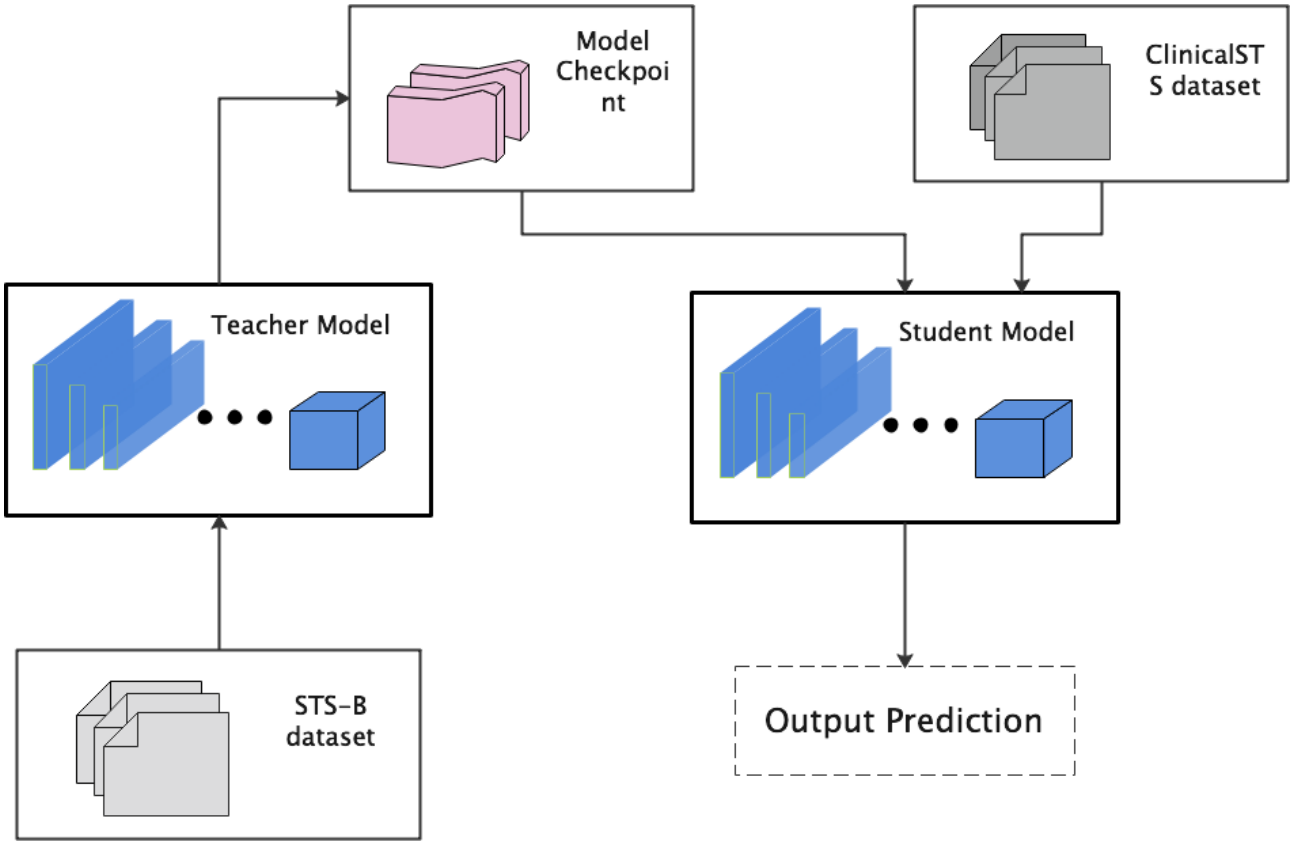
Semisupervised Learning

Because there was not a sufficient amount of medical data, the training of the model was not complete. To solve this problem, we used the semisupervised learning method in transfer learning.

The semisupervised [29] pretraining task in NLP is a form of transfer learning that aims to establish a wide range of semantic understanding to promote the performance improvement of training and testing tasks. It has been proven that semisupervised pretraining in transfer learning is very effective in benchmark NLP tasks, and the application prospects in medical NLP tasks are particularly broad. Nonspecific pretraining tasks are used for general medical domain tasks; however, commonly used and publicly available data sets are not specific to the medical domain and may not be well summarized. Therefore, the transfer of nonspecific pretraining tasks and the promotion of language models to medical domain tasks are very important for future model development.

To improve traditional semisupervised learning, we used the *teacher* and *student* idea in data distillation [30,31] to improve the design of semisupervised learning. Teacher–student refers to the same training process. The beginning of the student's training is the end of the teacher's training, which can deepen the learning of the model. We used the teacher–student approach to design semisupervised learning. The teacher part uses a data set from the common domain, using the STS-B data set from the General Language Understanding Evaluation standard of the general domain. The student part uses a clinical text data set. Our semisupervised learning method is shown in Figure 2.

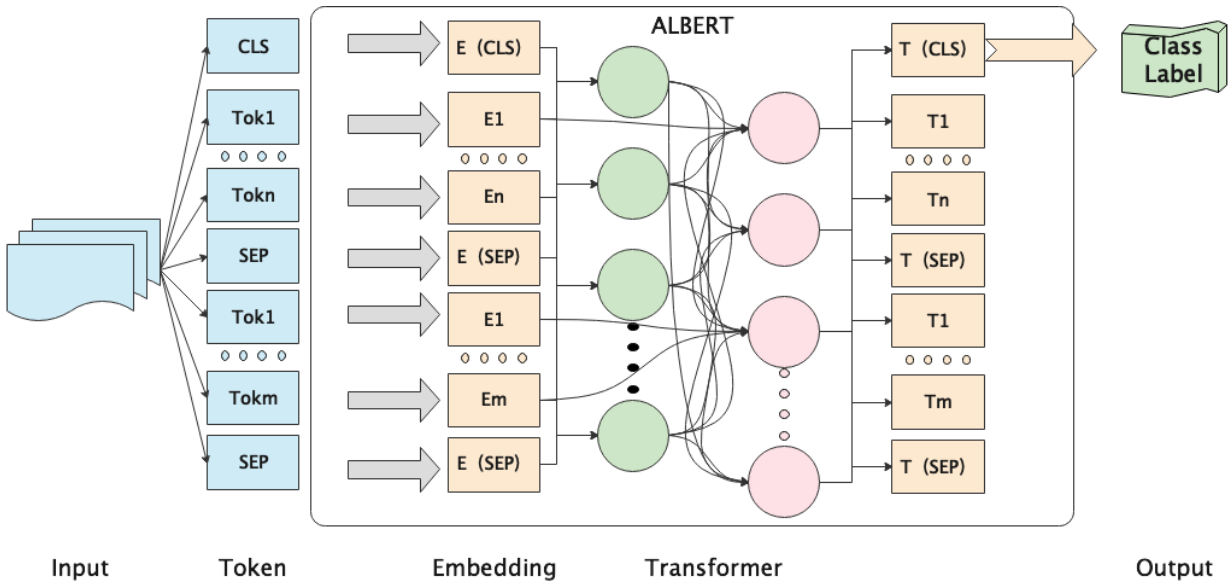
Figure 2. Semisupervised learning.



Self-Ensemble ALBERT Model

ALBERT has been applied to some tasks, such as natural language inference [32], sentiment analysis [33], causality analysis [34], and medical machine reading [35]. The self-attention structure is the core part of the transformer mechanism. The self-attention structure can directly calculate the similarity between words, which can intuitively solve the problem of long-distance information dependence. The combined self-attention structure transformer's semantic feature extraction ability is better than those of long short-term memory and convolutional neural networks, and it performs better under the combined action of decomposed embedding parameters and cross-layer shared parameters. Therefore, the pretrained self-attention structure, namely, the pretrained ALBERT model, was applied to our model. ALBERT is a variant of BERT that adds 2 methods of decomposing embedded parameters and sharing parameters across layers. It has 3 improvements. First, ALBERT decomposes embedding, which makes a large number of parameters sparse and reduces the number of dictionaries. Second, ALBERT adopts cross-layer parameter sharing, which reduces the parameter scale and improves the training speed. Third, ALBERT uses intersentence coherence, which makes the model unaffected by specific tasks. The architecture of the ALBERT model is shown in Figure 3.

Figure 3. Model architecture.



Following ALBERT, we first embedded the input data. Our embedding representation is constructed by the sum of token embedding, segment embedding, and location embedding. The input sequence is $S = [s_1, s_2, ..., s_n]$, where n is the number of words in the input. The tokens "[CLS]" and "[SEP]" were added at the beginning and end of each instance, respectively.

Then, we input the data into the ALBERT model, which is made up of n transformer stacks,

$$S_m = \text{Transformer}(S_{m-1}), \quad (1)$$

where S_m is the output of transformer stack m.

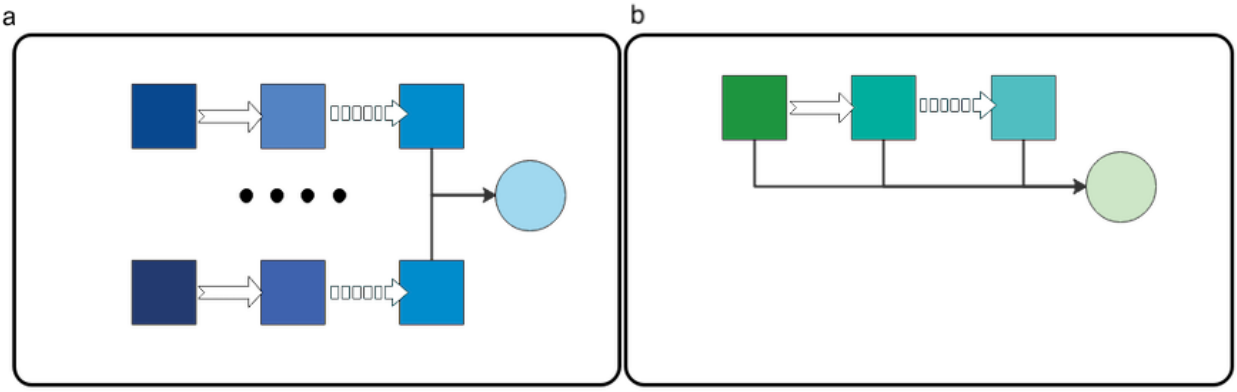
Since the results do not need to be normalized, we did not use an activation function.

To achieve the best performance, the ALBERT model was fine-tuned. ALBERT models are usually fine-tuned using stochastic gradient descent methods. In fact, fine-tuning the performance of ALBERT is usually sensitive to different random seeds and orders of the training data, especially if the last training sample is noisy. To alleviate this situation, an ensemble method was used to combine multiple fine-tuning models because it can reduce overfitting and improve model generalization. The ensemble ALBERT model usually has better performance than a single ALBERT model. However, training multiple ALBERT models simultaneously is time-consuming. It is often impossible to train multiple models with limited time and GPU resources. Therefore, we improved the model ensemble method to fine-tune the ALBERT model. Our model's ensemble method is called self-ensemble. The self-ensemble architecture is shown in Figure 4. The formula for self-ensemble is

$$\text{ALBERT}_{\text{vote}} = \text{Max}\left(\sum_{n=1}^{k} \text{ALBERT}(S_k)\right), \quad (2)$$

where ALBERT(S_k) represents the checkpoints of the model with k training steps.

Figure 4. (a) Traditional ensemble vs (b) self-ensemble architecture.



Data Sets

The Clinical STS shared task data set was collected from electronic health record in the Mayo Clinic clinical data warehouse. Since the Mayo Clinic has completed the system-wide electronic health record conversion of all care locations from General Electric to Epic, the Clinical STS shared

task data set will be extracted from the historical General Electric and Epic systems.

STS-B is a carefully selected English data set used in shared tasks between SemEval and SEM STS between 2012 and 2017. The data was divided into a training set, a development set, and a test set. The development set can be used to design new models and adjust hyperparameters. STS-B can be used to make comparable assessments in different research work and improve the tracking of the latest technology.

Table 2 shows the size of data set in the Clinical STS data set and the STS-B data set. The STS-B data set was used for the semisupervised learning training model. The STS-B data set comes from a data set collected by the general domain criterion General Language Understanding Evaluation. The Clinical STS data set was used to test the experimental results. The Clinical STS data set was provided by the competition organizer.

The STS-B data set provides paired text summaries, which are mainly from STS tasks in SemEval obtained over the years. The Clinical STS data set provides pairs of clinical text summaries, which are sentences extracted from clinical notes. This task assigns a numerical score to each pair of sentences to indicate their semantic similarity. Table 3 shows that the scores fall within an ordinal range, ranging from 0 to 5, where 0 means that the pair of sentences are completely different (ie, their meanings do not overlap) and 5 means that the pair of sentences have complete semantic equivalence.

Table 2. The size of data set.

Data set	Training	Validation	Test
STS-B	5749	1500	1379
Clinical STS	1642	N/A[a]	412

Table 3. Similarity scores with examples.

Score	Sentence 1	Sentence 2
0	The patient has missed 0 hours of work in the past seven days for issues not related to depression.	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.
1	nortriptyline [PAMELOR] 50 mg capsule 1 capsule by mouth every bedtime.	Tylenol Extra Strength 500 mg tablet 2 tablets by mouth every bedtime.
2	bupropion [WELLBUTRIN XL] 300 mg tablet sustained release 24 hour 1 tablet by mouth one time daily.	Flintstones Complete chewable tablet 1 tablet by mouth two times a day.
3	Given current medication regimen, the following parameters should be monitored by outpatient providers: None	Given current medication regimen, the following parameters should be monitored by outpatient providers: lithium level
4	The diagnosis and treatment plan were explained to the family/caregiver who expressed understanding of the information presented.	Explained diagnosis and treatment plan; patient expressed adequate understanding of the information presented today.
5	Learns best by: verbal instructions as procedure is being performed, reading, seeing, listening.	Learns best by: verbal instruction while procedure is performed, reading, seeing, listening.

Metric

We used the Pearson correlation coefficient as an evaluation criterion for the performance of the task. The Pearson correlation coefficient,

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-E(X))(Y-E(Y)))}{\sqrt{D(X)}\sqrt{D(Y)}} \quad (3)$$

where E is the mathematical expectation (or mean), D is the variance, and $\mathrm{Cov}(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$ is the covariance of random variables X and Y, is used to measure the degree of correlation between 2 variables.

Experimental Setting

In the experiments, we used Intel Xeon 2.2 GHz and Nvidia Tesla V100 32 GHz processors. Since we use semisupervised learning and self-ensemble techniques, our model will be stored by the checkpoint. The input dimensions of each of our data sets are the same. The optimal setting for the length of the input sequence is 64, and the optimal setting for the batch size was 32. The optimal setting for the checkpoint was 200. The optimal setting of the training step was 3598. In the experiments, we did not cross-train on the data set.

Results

Performance Comparison

Table 4 shows the top 5 performance results for the 2019 N2C2 OHNLP Track 1 Clinical STS, the value that we obtained during the challenge, and the value obtained by the method presented in this paper. Our current method achieves a good result—the Pearson correlation coefficient value exceeded the best result by 2 percentage points.

Table 4. Results on the test set for Clinical STS.

Methods	Pearson correlation coefficient
Multitask learning, ClinicalBERT	0.90
Multitask learning, BERT	0.89
BERT, XLNet	0.88
BERT	0.87
BERT, XLNet	0.87
Our previous method[a]	0.66
Our method in this paper	0.92

[a]Ordered short long-term memory and attention.

Data Augmentation

The EDA method uses text replacement and deletion operations, optimizes the selection range of replacement and deletion, and retains the medical proper nouns in the data set. Table 5 shows the effect of using EDA on the model performance. After EDA, the size of medical data set is expanded, and the model's performance was greatly improved.

Table 5. Comparison between the model with and without EDA.

Methods	Pearson correlation coefficient
Without EDA[a]	0.88
With EDA	0.92

[a]EDA: easy data augmentation.

Semisupervised Learning

The semisupervised learning method uses the general domain data set STS-B for training to solve the problem of insufficient medical data. Table 6 shows the effect of using semisupervised learning on the model performance. We can see that semisupervised learning can greatly improve the efficiency of the model.

Table 6. Comparison between the model with and without semisupervised learning.

Methods	Pearson correlation coefficient
Without semisupervised learning	0.87
With semisupervised learning	0.92

Self-Ensemble ALBERT

Table 7 shows the effect of using the self-ensemble method on the model performance. We can see that the efficiency of the model with self-ensemble is better than that of the ordinary ensemble model. Additionally, self-ensemble greatly shortens the training time of the model, reduces the calculation time of the algorithm, and improves the efficiency of the algorithm.

BERT and ALBERT are pretrained models with the same self-attention structure. As shown in Table 8, the performance of ALBERT is better than that of BERT on the Clinical STS data set.

Table 7. Comparison among the model without ensemble, the model with ensemble, and the model with self-ensemble.

Method	Pearson correlation coefficient
None	0.85
Ensemble[a]	0.89
Self-ensemble	0.92

[a]Ensemble represents an ensemble method through multiple ALBERT models.

Table 8. Comparison between the ALBERT and BERT models.

Methods	Runtime (minutes)	Convergence speed[a] (steps)	Pearson correlation coefficient
BERT	50	3300	0.86
ALBERT	32	2700	0.92

[a]Convergence speed is measured using the training steps.

Discussion

Overview

This paper makes the following contributions. First, we used the EDA text data augmentation method. This method increased the number of data through a series of operations and enriched the semantics of the data. Second, for the problem of insufficient medical data, we used a semisupervised learning method. This method relied on the use of external data to enrich the semantics. Third, to solve the problem of learning complex semantics and the loss of key semantic information, we used the self-ensemble ALBERT model for semantic similarity calculation of clinical text. This method not only improves the results of the semantic similarity calculation of clinical text but also, due to the improvement of the self-ensemble of our model, allows the algorithm to shorten its running time and improve its efficiency. With these techniques, our model obtained a Pearson correlation coefficient of 0.92.

In order to test the influence of the method on performance, we conducted ablation experiments on EDA, semisupervised learning, and self-ensemble. At the same time, in order to verify the performance of the model, we also performed ablation experiments on ALBERT.

Conclusions

Compared with other models and methods, combining an EDA and self-ensemble ALBERT model under semisupervised learning to perform clinical textual semantic similarity calculations can save a large amount of training time and allows more data to be trained at the same time. This brings great convenience for practical applications and scientific research.

In the future, we will study how to combine reinforcement learning to process natural language to further improve the performance of the model and handle the dilemma of bloated or erroneous in electronic health records caused by the increasing use of copy and paste.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61463050, Grant 61762091 and Grant 12061088, and the Science Foundation of Yunnan Education Department under Grant 2020Y0011.

Conflicts of Interest

None declared.

References

1. Karwatowski M, Russek P, Wielgosz M, Koryciak S, Wiatr K. Energy efficient calculations of text similarity measure on FPGA-accelerated computing platforms. Parallel Processing and Applied Mathematics 2016 Apr 2;9573:31-40 [FREE Full text] [doi: 10.1007/978-3-319-32149-3_4]
2. Quan X, Liu G, Lu Z, Ni X, Wenyin L. Short text similarity based on probabilistic topics. Knowl Inf Syst 2009 Sep 17;25(3):473-491. [doi: 10.1007/s10115-009-0250-y]
3. Song W, Feng M, Gu N. Question similarity calculation for FAQ answering. 2007 Presented at: Third International Conference on Semantics Knowledge and Grid (SKG); October 29-31; Shan Xi, China p. 298-301. [doi: 10.1109/skg.2007.247]
4. Li L, Zhu AH, Su T. An improved text similarity calculation algorithm based on vsm. AMR 2011 Apr;225-226:1105-1108. [doi: 10.4028/www.scientific.net/amr.225-226.1105]
5. Zhang L, Zhang L, Du B. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci Remote Sens Mag 2016 Jun;4(2):22-40. [doi: 10.1109/mgrs.2016.2540798]
6. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Oct Presented at: 19th Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25–29; Doha, Qatar p. 1532-1543. [doi: 10.3115/v1/d14-1162]
7. Kusner M, Sun Y, Kolkin N. From word embeddings to document distances. 2015 Presented at: International Conference on Machine Learning; July 6-11; Lille, France p. 957-966.
8. 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):1-7 [FREE Full text] [doi: 10.1186/s12911-020-1045-z] [Medline: 32349764]
9. Ritchie J, Welch B. Categorization of third-party apps in electronic health record app marketplaces: systematic search and analysis. JMIR Med Inform 2020 May 29;8(5):e16980 [FREE Full text] [doi: 10.2196/16980] [Medline: 32469324]

10. Cera D, Diabb M, Agirrec E, Lopez-Gazpioc I, Speciad L. SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).: Association for Computational Linguistics; 2017 Presented at: 11th International Workshop on Semantic Evaluation; August 3-4; Vancouver, Canada p. 1-14. [doi: 10.18653/v1/s17-2001]
11. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L. Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018 Aug Presented at: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; August 29-September 1; Washington DC, USA. [doi: 10.1145/3233547.3233672]
12. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/OHNLP track on clinical semantic textual similarity: overview. JMIR Med Inform 2020 Nov 27;8(11):e23375 [FREE Full text] [doi: 10.2196/23375] [Medline: 33245291]
13. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Nov;87:12-20 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]
14. Ma X, Tao Z, Wang Y, Yu H, Wang Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies 2015 May;54:187-197. [doi: 10.1016/j.trc.2015.03.014]
15. Shin H, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016 May;35(5):1285-1298 [FREE Full text] [doi: 10.1109/TMI.2016.2528162] [Medline: 26886976]
16. Wang B, Zhang X, Zhou X, Li J. A gated dilated convolution with attention model for clinical cloze-style reading comprehension. Int J Environ Res Public Health 2020 Feb 19;17(4) [FREE Full text] [doi: 10.3390/ijerph17041323] [Medline: 32092861]
17. Zhu Z, Peng G, Chen Y, Gao H. A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis. Neurocomputing 2019 Jan;323:62-75. [doi: 10.1016/j.neucom.2018.09.050]
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA, USA p. 5998-6008.
19. Bertinetto L, Valmadre J, Henriques JF, Vedaldi A, Torr PHS. Fully-convolutional siamese networks for object tracking. 2016 Nov Presented at: European Conference on Computer Vision; October 8-16; Amsterdam, Netherlands p. 850-865 URL: https://link.springer.com/chapter/10.1007/978-3-319-48881-3_56 [doi: 10.1007/978-3-319-48881-3_56]
20. Wu L, Wang Y, Gao J, Li X. Where-and-when to look: deep siamese attention networks for video-based person re-identification. IEEE Transactions on Multimedia 2019 Jun;21(6):1412-1424. [doi: 10.1109/tmm.2018.2877886]
21. Eisinga R, Grotenhuis MT, Pelzer B. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int J Public Health 2013 Aug;58(4):637-642. [doi: 10.1007/s00038-012-0416-3] [Medline: 23089674]
22. Jung H, Kim B, Lee I, Lee J, Kang J. Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method. BMC Med Imaging 2018 Dec 03;18(1):48 [FREE Full text] [doi: 10.1186/s12880-018-0286-0] [Medline: 30509191]
23. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. 2019 Presented at: International Conference on Learning Representations; April 26-30; Addis Ababa.
24. Huang G, Song S, Gupta JND, Wu C. Semi-supervised and unsupervised extreme learning machines. IEEE Trans Cybern 2014 Dec;44(12):2405-2417. [doi: 10.1109/TCYB.2014.2307349] [Medline: 25415946]
25. Enguehard J, O'Halloran P, Gholipour A. Semi-supervised learning with deep embedded clustering for image classification and segmentation. IEEE Access 2019;7:11093-11104. [doi: 10.1109/access.2019.2891970]
26. Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Nov Presented at: EMNLP-IJCNLP 2019; November 3-7; Hong Kong, China p. 6382-6388. [doi: 10.18653/v1/d19-1670]
27. Yu AW, Dohan D, Luong MT. Qanet: Combining local convolution with global self-attention for reading comprehension. 2018 Presented at: International Conference on Learning Representations; April 30-May 3; Vancouver, Canada.
28. Xie Z, Wang SI, Li J. Data noising as smoothing in neural network language models. 2017 Presented at: International Conference on Learning Representations; April 24-26; Toulon, France.
29. Hussain A, Cambria E. Semi-supervised learning for big social data analysis. Neurocomputing 2018 Jan;275:1662-1673. [doi: 10.1016/j.neucom.2017.10.010]
30. Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 Nov Presented at: IEEE Conference on Computer Vision and Pattern Recognition; July 21-26; Honolulu, HI, USA p. 4133-4141. [doi: 10.1109/cvpr.2017.754]

31. Pan Y, He F, Yu H. A novel enhanced collaborative autoencoder with knowledge distillation for top-N recommender systems. Neurocomputing 2019 Mar;332:137-148. [doi: 10.1016/j.neucom.2018.12.025]
32. Williams A, Nangia N, Bowman S R A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6; New Orleans, Louisiana p. 1112-1122.
33. Zampieri M, Nakov P, Rosenthal S. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation.: Association for Computational Linguistics; 2020 Presented at: The 28th International Conference on Computational Lingustics (COLING-2020); September 13-14; Barcelona (online) p. 1425-1447.
34. Yu HQ. Dynamic causality knowledge graph generation for supporting the Chatbot health care system. In: Proceedings of the Future Technologies Conference (FTC) 2020. 2020 Presented at: Future Technologies Conference (FTC) 2020; October; Vancouver, Canada p. 30-45. [doi: 10.1007/978-3-030-63092-8_3]
35. Li D, Hu B, Chen Q. Towards medical machine reading comprehension with structural knowledge and plain text. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).: Association for Computational Linguistics; 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); November; Online p. 1427-1438. [doi: 10.18653/v1/2020.emnlp-main.111]

Abbreviations

EDA: easy data augmentation
GLUE: General Language Understanding Evaluation
OHNLP: Open Health Natural Language Processing
N2C2: National NLP Clinical Challenges
NLP: natural language processing
STS: semantic textual similarity

Edited by Y Wang; submitted 31.07.20; peer-reviewed by S Liu, L Wang, D Mordaunt; comments to author 22.09.20; revised version received 22.11.20; accepted 15.12.20; published 22.01.21

Please cite as:
Li J, Zhang X, Zhou X
ALBERT-Based Self-Ensemble Model With Semisupervised Learning and Data Augmentation for Clinical Semantic Textual Similarity Calculation: Algorithm Validation Study
JMIR Med Inform 2021;9(1):e23086
URL: http://medinform.jmir.org/2021/1/e23086/
doi: 10.2196/23086
PMID: 33480858