

RESEARCH

Open Access



# Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity

Ying Xiong<sup>1</sup>, Shuai Chen<sup>1</sup>, Haoming Qin<sup>1</sup>, He Cao<sup>1</sup>, Yedan Shen<sup>1</sup>, Xiaolong Wang<sup>1</sup>, Qingcai Chen<sup>1,2</sup>, Jun Yan<sup>3</sup> and Buzhou Tang<sup>1,2\*</sup>

From BioCreative/OHNLN Challenge 2018  
Washington, D.C., USA. 29 August-01 September 2018

## Abstract

**Background:** Semantic textual similarity (STS) is a fundamental natural language processing (NLP) task which can be widely used in many NLP applications such as Question Answer (QA), Information Retrieval (IR), etc. It is a typical regression problem, and almost all STS systems either use distributed representation or one-hot representation to model sentence pairs.

**Methods:** In this paper, we proposed a novel framework based on a gated network to fuse distributed representation and one-hot representation of sentence pairs. Some current state-of-the-art distributed representation methods, including Convolutional Neural Network (CNN), Bi-directional Long Short Term Memory networks (Bi-LSTM) and Bidirectional Encoder Representations from Transformers (BERT), were used in our framework, and a system based on this framework was developed for a shared task regarding clinical STS organized by BioCreative/OHNLN in 2018.

**Results:** Compared with the systems only using distributed representation or one-hot representation, our method achieved much higher Pearson correlation. Among all distributed representations, BERT performed best. The highest Person correlation of our system was 0.8541, higher than the best official one of the BioCreative/OHNLN clinical STS shared task in 2018 (0.8328) by 0.0213.

**Conclusions:** Distributed representation and one-hot representation are complementary to each other and can be fused by gated network.

**Keywords:** Clinical semantic textual similarity, Gated network, Distributed representation, One-hot representation

\* Correspondence: [tangbuzhou@gmail.com](mailto:tangbuzhou@gmail.com)

<sup>1</sup>Department of Computer Science, Harbin Institute of Technology, Shenzhen, Guangdong, China

<sup>2</sup>Peng Cheng Laboratory, Shenzhen, Guangdong, China

Full list of author information is available at the end of the article



## Background

Electronic Health Records (EHRs) that record patients' complete information, including family history, general situation, chief complaint, examination, lab test, diagnosis, assessment, and plan, etc., have been widely used to help medical experts to improve processes of care on patient outcomes. The key to secondary use of EHRs lies in high quality. However, the quality of EHRs has met challenges such as frequent use of copy-and-paste, templates, and smart phrases which lead to bloated or erroneous clinical notes [1]. A study of 23,630 clinical notes written by 460 clinicians showed that 46% of the text in the clinical records copied other clinical records, 36% was imported from templates, and only 18% was manually entered [2]. To aggregate data from diverse sources and minimize data redundancy, BioCreative/OHNLNLP organized a shared task to evaluate the semantic similarity between text snippets (also called sentences in this paper) of clinical texts in 2018. In this shared task, the similarity between two clinical text snippets ranged from 0 to 5, where 0 means that the two clinical text snippets are not semantically similar at all, and 5 indicates that the two clinical text snippets are entirely equal. In the past few years, SemEval workshop has launched STS shared task in the general domain many times [3–8]. In the clinical area, BioCreative/OHNLNLP first organized an STS shared task in 2018.

As many NLP applications such as QA, IR, etc. usually used STS as a core component, large quantities of researchers have contributed to STS and achieved great success. STS is a typical regression problem, and how to model sentence pairs is the key to solutions. There are two main types of representations to model sentence pairs: one-hot representation and distributed representation. The one-hot representation that depends on manually-crafted features suffers from sparsity. The distributed representation that learns dense real-value vector from unlabeled data automatically by neural networks have shown great potentialities. Most studies focus on one type of representations. In this paper, we proposed a novel framework to fuse the two types of representations using a gated network. In the case of distribution representations, we compared some current state-of-the-art neural networks such as CNN, Bi-LSTM, and BERT. To evaluate our method, we conducted experiments on the clinical STS corpus of BioCreative/OHNLNLP 2018 by comparing our method with the methods that only using one type of representation and the official best method on the clinical STS shared task. Experimental results showed that: 1) the proposed method achieved much higher Pearson correlation than the methods only using one type of representations. 2) BERT performed better than other distributed representations. 3) Our method achieved the highest Pearson

correlation of 0.8541, higher than the best official one of the clinical STS shared task (0.8328) by 0.0213.

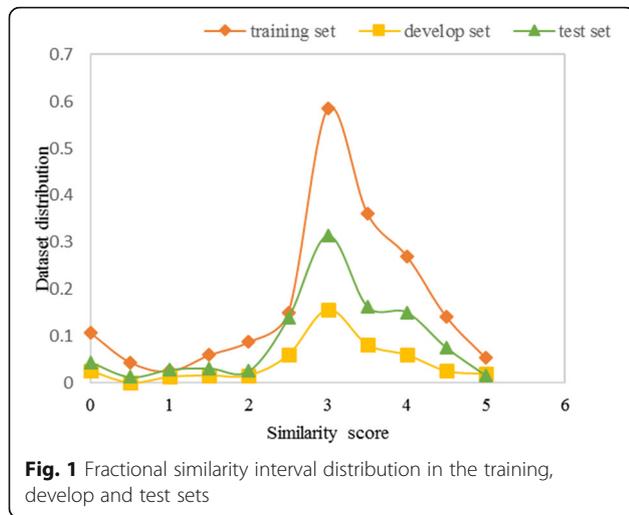
## Related work

There are two main types of sentence representation: (1) sparse one-hot representation based on manually extracted features. (2) densely distributed representation learnt from large labeled data. Within a long period, there have been a large number of feature extraction methods proposed to represent sentence by one-hot vector. Goma et al. [9] summarized several types of features and various similarity computation methods: string-based similarity computation methods such as N-gram [10–12], corpus-based similarity methods [13–16] and knowledge-based similarity computation methods [17–19]. In recent years, neural networks have become mainstream methods for sentence representation and STS. Bromley et al. [20] firstly presented a Siamese architecture to encode sentence pairs. Based on previous work, Mueller et al. [21] used Siamese recurrent architecture learning sentence representation. Tang et al. [22] used deep belief network to learn sentence representation. He et al. [23] proposed a novel pairwise word interaction method to measure the sentence semantic similarity. Gong et al. [24] further hierarchically extracted semantic features from interaction space. Tai et al. [25] used tree-structured LSTM to improve the sentence representation. Subramanian et al. [26] used transfer learning to learn sentence representation. In recent years, neural language models have been also utilized for sentence representation, such as ELMo [27] and GPT [28]. Some researchers extracted features at different granularities and combined them with distributed representations, such as He et al. [29] and Wang et al. [30]. Ji et al. [31] combined the features with distributed representation, our work was similar to Ji's work, but we used a novel gate to choose how to combine one-hot representation and distributed representation.

## Methods

### Task definition

Formally, the clinical STS task is to determine the similarity of a pair of given sentences, denoted by  $sim(s_1, s_2)$ , where  $s_1$  is a sentence of length  $m$  and  $s_2$  is a sentence of length  $n$ . We used  $s_{ij}$  to denote the  $j$ -th word of  $s_i$ . In this study, the similarity of a sentence pair ranged from 0 to 5, where 0 represents the two sentences are not semantically similar, and 5 represents the two sentences are semantically equal. Besides, we used  $D$  and  $O$  to describe a sentence's distributed representation and one-hot representation respectively.



**Fig. 1** Fractional similarity interval distribution in the training, develop and test sets

**Dataset**

The BioCreative/OHNLP organizer manually annotated 750 sentence pairs with semantic similarity ranging from 0 to 5 for system development and 318 sentence pairs for system test. We further divided the 750 sentence pairs into a training set and a develop set using stratified sampling to guarantee that the develop set is a representative of the overall dataset. Figure 1 shows the fractional similarity interval distribution in the training, development and test sets, and Table 1 lists some annotated examples.

**Table 1** Annotated examples

Score	Example
0	$s_1$ : discuss necessity member healthcare team male female participate procedure $s_2$ : report represent interpretation original data trace store electronic record esophageal laboratory
1	$s_1$ : mother blood type o + hepatitis b negative hiv negative found gb positive $s_2$ : patient undergone genetic test found brca1 2 negative well bart negative
2	$s_1$ : patient discharge home ambulate without assistance discharge instruction give patient $s_2$ : patient left without see ambulate without assistance family drive accompany husband wife
3	$s_1$ : negative cardiovascular review system historian denies chest pain dyspnea exertion $s_2$ : negative cardiovascular review system historian denies chest pain diaphoresis syncope palpitation
4	$s_1$ : patient education ready learn apparent learn barrier identify learn preference include listen $s_2$ : assistance somali interpreter ready learn apparent learn barrier identify learn preference include listen
5	$s_1$ : nurse visit ten minute half spent counsel point test $s_2$ : nurse visit ten minute half spent consultation point test

**Data processing**

We preprocessed each sentence as follows: 1) used NLTK tool (<http://www.nltk.org/>) for tokenization and lemmatization; 2) converted Arabic numerals into English numbers. For example, the sentence “Indication, Site, and Additional Prescription Instructions: Apply 1 patch every 24 hours; leave on for up to 12 hours within a 24 hour period” became “indication site additional prescription instruction apply one patch every twenty four hour leave twelve hour within twenty four hour period” after preprocessing.

**Distributed representation and one-hot representation fusion**

Figure 2 shows an overview architecture of our distributed representation and one-hot representation fusion system based on a gated network for the clinical STS task of BioCreative/OHNLP 2018 (i.e., task2). The system consists of three components: (1) sentence pair representation – distributed representation and one-hot representation; (2) representation fusion with gated network; (3) neural network to compute sentence similarity. We described some of them in the following sections in detail.

**Distributed representation**

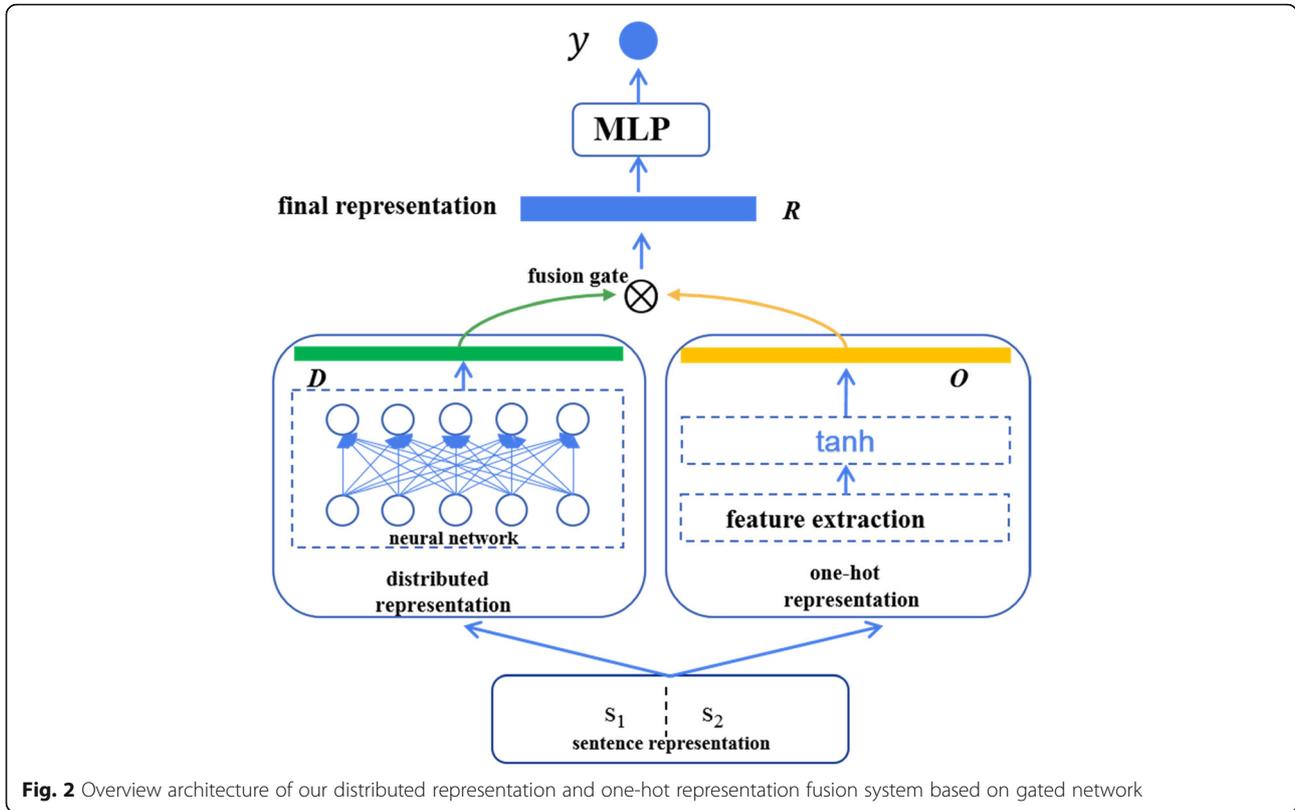
In this study, we investigated three types of distributed representations: Siamese CNN [32], Siamese RNN [21] and BERT [33], where Siamese CNN and Siamese RNN are two popular neural networks used to represent sentence pair, while BERT is a new language representation method proposed recently.

- (1) **Siamese CNN** is composed of two CNNs, each of which represents a sentence, and the two CNNs share weights. The representation of sentence pair  $(s_1, s_2)$  is obtained as follows:

$$\begin{aligned}
 CNN(\cdot) &= avg\_pool(convolution(\cdot)) \\
 D_{cnn} &= [CNN(s_1), CNN(s_2)],
 \end{aligned}
 \tag{1}$$

where *avg\_pool* is the average pooling operation, *convolution* is the convolution operation,  $s_1$  and  $s_2$  are the two input sentences.

- (2) **Siamese RNN**, similar to Siamese CNN, is composed of two RNNs that represent each one sentence respectively and share weights. In our study, we adopted Bi-directional Long Short Term Memory (Bi-LSTM) networks as an implementation of RNN, where each word  $i$  at  $s_1$  and  $s_2$  is represented as:



**Fig. 2** Overview architecture of our distributed representation and one-hot representation fusion system based on gated network

$$\begin{aligned} \overrightarrow{h}_i^{s_1} &= \overrightarrow{LSTM}(\overrightarrow{h}_{i-1}^{s_1}, s_{1i}) & i = 1, \dots, m \\ \overleftarrow{h}_i^{s_1} &= \overleftarrow{LSTM}(\overleftarrow{h}_{i+1}^{s_1}, s_{1i}) & i = m, \dots, 1, \end{aligned} \quad (2)$$

$$\begin{aligned} \overrightarrow{h}_i^{s_2} &= \overrightarrow{LSTM}(\overrightarrow{h}_{i-1}^{s_2}, s_{2i}) & i = 1, \dots, n \\ \overleftarrow{h}_i^{s_2} &= \overleftarrow{LSTM}(\overleftarrow{h}_{i+1}^{s_2}, s_{2i}) & i = n, \dots, 1, \end{aligned} \quad (3)$$

where  $\overrightarrow{LSTM}$  and  $\overleftarrow{LSTM}$  are the forward and backward LSTMs.

The sentence pair  $(s_1, s_2)$  is described as:

$$D_{lstm} = \left[ \overrightarrow{h}_m^{s_1}, \overleftarrow{h}_1^{s_1}, \overrightarrow{h}_n^{s_2}, \overleftarrow{h}_1^{s_2} \right] \quad (4)$$

(3) **BERT (Bidirectional Encoder Representations from Transformers)** is a language representation method to obtain deep bidirectional representations of sentences by jointly conditioning on both left and right context in all layers from free text unsupervised. In our study,

the representation of a sentence pair  $(s_1, s_2)$  was denoted by

$$D_{bert} = BERT([s_1, s_2]) \quad (5)$$

We trained a new BERT model on MIMIC III starting from the pre-trained model released by Google (<https://github.com/google-research/bert>).

#### One-hot representation

We followed Tian's work [34] to extract the following two types of features: (1) Sentence-level features: IDF (inverse document frequency) [35] and sentence length; (2) Sentence pair-level features:  $N$ -gram overlaps defined in eq. (6), and distances or similarities between the two input sentences calculated by cosine, Manhattan, Euclidean, Chebyshev, polynomial kernel, RBF kernel, Laplacian kernel and sigmoid kernel after each sentence is represented by the average vector of all words' embeddings (<https://github.com/mmihaltz/word2vec-GoogleNews-vectors>).

$$NGO(s_1, s_2) = 2 \left( \frac{|Ngram(s_1) \cap Ngram(s_2)|}{|Ngram(s_1)| + |Ngram(s_2)|} \right) \tag{6}$$

where  $Ngram(s_i)$  ( $i = 1,2$ ) is a  $N$ -gram set extracted from  $s_i$ . In our study, unigrams, bigrams and trigrams were considered.

### Fusion gate

Inspired by the gated network mechanism in variants of RNN such as LSTM and GRU (Gated Recurrent Unit), we introduced a gate to leverage distributed representation and one-hot representation. Before fusion, we adopted the tanh function as an activation function to convert the two types of representation into the same space. So that the final representation of sentence pair  $(s_1, s_2)$   $R$  can be obtained in the following way:

$$D_{norm} = \tanh(W_d \cdot D + b_d) \tag{7}$$

$$O_{norm} = \tanh(W_o \cdot O + b_o) \tag{8}$$

$$f = \sigma(W_f \cdot [D_{norm}, O_{norm}] + b_f) \tag{9}$$

$$R = f * D_{norm} + (1-f) * O_{norm} \tag{10}$$

Where  $W_d, W_o, W_f$  are weights matrices;  $b_d, b_o, b_f$  are bias vectors;  $\sigma$  is the sigmoid activation function;  $f$  is leverage coefficient between the distributed representation and the one-hot representation.

### Experiments

We started from the baseline systems that only used one type of representations (distributed representation or one-hot representation), then concatenated the two types of representations, and finally fused the two types of representations with a

gated network. All systems were evaluated on the clinical STS corpus of the BioCreative/OHNLP challenge in 2018, and Pearson correlation was used to measure the performance of the systems.

### Results

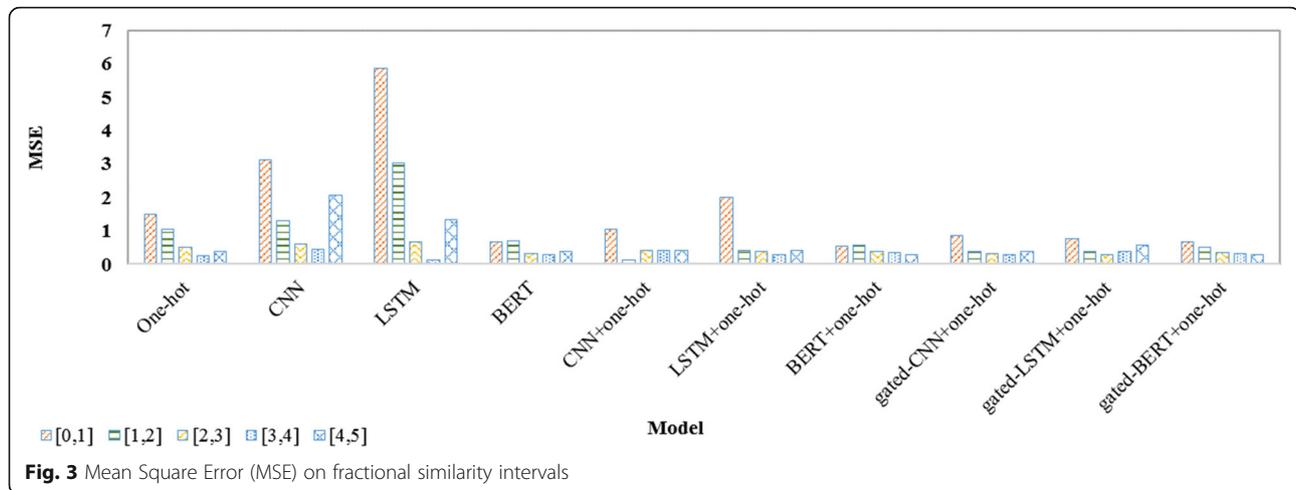
As shown in Table 2, the baseline system only using one-hot representation achieved much higher Pearson correlation than the baseline system just applying CNN or LSTM, but lower Pearson correlation than the baseline system only using BERT. The highest Pearson correlation of the baseline systems was 0.8461. When concatenating each distributed representation with the one-hot representation, we received higher Pearson correlation, indicating that the two types of representations are mutually complementary. For example, when we concatenated BERT with one-hot representation, we obtained a Pearson correlation of 0.8525, higher than the baseline system only using BERT by 0.0064 and the baseline system solely using one-hot representation by 0.0586. Instead of concatenating any distributed representation with one-hot representation, fusing them brought more significant improvement in Pearson correlation. The Pearson correlation difference between the systems that using concatenation strategy and fusion strategy ranged from 0.0016 to 0.0359. Among three distributed representations, our system achieved highest Pearson correlation of 0.8541 when using BERT for fusion, higher than the best official one of the BioCreative/OHNLP clinical STS shared task (0.8328) by 0.0213.

### Discussion

In this study, we investigated three state-of-the-art distributed representation methods, that is, CNN, Bi-

**Table 2** Performance of systems on the clinical STS corpus of the BioCreative/OHNLP shared task in 2018

Method	Score Interval					Overall
	[0,1]	[1, 2]	[2, 3]	[3, 4]	[4, 5]	
Baseline						
One-hot	0.5567	0.2311	0.0998	0.2409	0.1167	0.7939
CNN	0.3960	-0.0850	-0.0090	0.0370	-0.0654	0.4444
LSTM	0.3920	-0.2945	0.2088	-0.0538	-0.0303	0.4275
BERT	0.7613	0.1206	0.2635	0.2530	0.1210	0.8461
Concatenation						
CNN + one-hot	0.5406	0.6917	0.1352	0.2539	0.0744	0.8083
LSTM+one-hot	0.5850	0.3415	0.2269	0.2173	0.2155	0.8030
BERT+one-hot	0.6684	0.3038	0.2309	0.2425	0.2203	0.8525
Fusion (gated network)						
CNN + one-hot	0.6973	0.2324	0.1675	0.2336	0.0864	0.8442
LSTM+one-hot	0.6253	0.3583	0.1869	0.2550	0.1018	0.8379
BERT+one-hot	0.6872	0.1605	0.3238	0.2822	0.1666	<b>0.8541</b>



LSTM, and BERT, and proposed a novel framework based on a gated network to fuse distributed representation and one-hot representation of sentence pairs. Among the systems only using any one distributed representation or one-hot representation, the system using BERT achieved highest Pearson correlation, but the system using one-hot representation produced much higher Pearson correlation than the method using CNN or Bi-LSTM. Both concatenation and fusion of distributed representation and one-hot representation brought improvement, and the fusion with gated network performed better.

The reason why the system using CNN or Bi-LSTM performed much worse than that using BERT or one-hot representation lies in the following two aspects: 1) the word embeddings used in CNN or Bi-LSTM were trained on a much smaller corpus than BERT; 2) one-hot representation had an advantage over CNN and Bi-LSTM on sentence pairs not very similar when the embeddings were trained on a small corpus. For example, it was easy to determine that “it be appropriate to retain the patient at the present level of care since the patient be make progress but have not yet achieve the goal articulate in the individualize treatment plan” and “the patient demonstrates the ability to fire the ta g and fh1 of the operative extremity” are not semantically similar (i.e., similarity of 0) when we applied the system using one-hot representation as there was no  $N$ -gram overlapped by the two sentences, but a little semantically similar (e.g., similarity of 2.02 when using CNN) when we applied the system using CNN or Bi-LSTM. The improvement because of concatenation or fusion of distributed representation and one-hot representation mainly came from sentence pairs of high similarity. As an illustration, we compared mean square error (MSE) on fractional similarity intervals as shown in Fig. 3.

For further improvement, there are three possible directions as follows: 1) fuse multiple distributed representations

with one-hot representation as different distributed representations may be complementary; 2) increase more data for word embedding training and model training; 3) introduce domain knowledge into our framework. All of them will be investigated in the future.

## Conclusion

In this paper, we proposed a novel framework to fuse distributed representation and one-hot representation using a gated network for clinical STS. Experiments on a benchmark dataset showed that the two types of representations were complementary and gated network was a good way for representation fusion.

## Abbreviations

BERT: Bidirectional encoder representations from transformers; Bi-LSTM: Bidirectional long short term memory networks; CNN: Convolutional neural network; EHR: Electronic health records; IDF: Inverse document frequency; IR: Information retrieval; NLP: Natural language processing; QA : Question answer; STS: Semantic textual similarity

## Acknowledgements

Not applicable.

## About this supplement

This article has been published as part of BMC Medical Informatics and Decision Making Volume 20 Supplement 1, 2020: Selected Articles from the BioCreative/OHNLN Challenge 2018 – Part 2. The full contents of the supplement are available online at <https://bmcmmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-1>.

## Authors' contributions

The work presented here was carried out in collaboration between all authors. YX, SC, HQ and BT designed deep learning methods and experiments. HC and YS designed featured-based methods and experiments. YX and BT contributed to the writing of the manuscript. XW, QC and JY provided guidance and reviewed the manuscript critically. All authors have approved the final manuscript.

## Funding

This work is supported in part by grants: NSFCs (National Natural Science Foundations of China) (U1813215, 61876052 and 61573118), Special Foundation for Technology Research Program of Guangdong Province (20158010131010), Strategic Emerging Industry Development Special Funds

of Shenzhen (JCYJ20170307150528934 and JCYJ20180306172232154), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

#### Availability of data and materials

Our annotated corpus was supplied by BioCreative/OHNL organization on clinical semantic textual similarity shared task.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Department of Computer Science, Harbin Institute of Technology, Shenzhen, Guangdong, China. <sup>2</sup>Peng Cheng Laboratory, Shenzhen, Guangdong, China. <sup>3</sup>Yidu Cloud (Beijing) Technology Co., Ltd, Beijing, China.

Published: 30 April 2020

#### References

- Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. In: Proceedings of American Medical Informatics Association Annual Symposium. AMIA; 2011. p. 1612.
- Wang MD, Khanna R, Najafi N. Characterizing the source of text in electronic health record progress notes. *JAMA Intern Med.* 2017;177:1212–3.
- Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, et al. Semeval-2014 task 10: multilingual semantic textual similarity. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014); 2014. p. 81–91.
- Agirre E, Cer D, Diab M, Gonzalez-Agirre A, Guo W. \* SEM 2013 shared task: semantic textual similarity. In: Second joint conference on lexical and computational semantics (\* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity; 2013. p. 32–43.
- Agirre E, Diab M, Cer D, et al. Semeval-2012 task 6: A pilot on semantic textual similarity. In: Proceedings of the 6th International Workshop on Semantic Evaluation. (SemEval 2012); 2012. p. 385–93.
- 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. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015); 2015. p. 252–63.
- 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. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016); 2016. p. 497–511.
- Cera D, Diab M, Agirrec E, Lopez-Gazpio I, Speciad L, Donostia BC. SemEval-2017 task 1: semantic textual similarity multilingual and cross-lingual focused evaluation; 2017.
- Gomaa WH, Fahmy AA. A survey of text similarity approaches. *Int J Comput Appl.* 2013;68:13–8.
- Barrón-Cedeno A, Rosso P, Agirre E, Labaka G. Plagiarism detection across distant language pairs. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010). Beijing: Coling 2010 Organizing Committee; 2010. p. 37–45.
- Wan S, Dras M, Dale R, Paris C. Using dependency-based features to take the ‘para-farce’ out of paraphrase. In: Proceedings of the Australasian language technology workshop 2006; 2006. p. 131–8.
- Madnani N, Tetreault J, Chodorow M. Re-examining machine translation metrics for paraphrase identification. In: Proceedings of the 2012 conference of the north American chapter of the Association for Computational Linguistics: human language technologies. USA: Association for Computational Linguistics; 2012. p. 182–90.
- Landauer TK, Dumais ST. A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol Rev.* 1997;104:211.
- Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav Res Methods Instrum Comput.* 1996;28:203–8.
- Gabrilovich E, Markovitch S. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc.; 2007. p. 1606–11.
- Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th research on computational linguistics international conference; 1997. p. 19–33.
- Lin D. Book Reviews: WordNet: An Electronic Lexical Database. *Computational Linguistics.* 1999;25. <https://www.aclweb.org/anthology/J99-2008>.
- Fernando S, Stevenson M. A semantic similarity approach to paraphrase detection. In: Proceedings of the 11th annual research colloquium of the UK special interest Group for Computational Linguistics; 2008. p. 45–52.
- Mihalcea R, Corley C, Strapparava C. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. In: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1. AAAI Press; 2006. p. 775–80.
- Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R. Signature verification using a “siamese” time delay neural network. In: Advances in neural information processing systems; 1994. p. 737–44.
- Mueller J, Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press; 2016. p. 786–92.
- Tang D, Qin B, Liu T, Li Z. Learning sentence representation for emotion classification on microblogs. In: Proceedings of Natural Language Processing and Chinese Computing. Springer; 2013. p. 212–23.
- 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; 2016. p. 937–48.
- Gong Y, Luo H, Zhang J. Natural language inference over interaction space. *ArXiv Prepr ArXiv170904348*; 2017.
- Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 1: long papers); 2015. p. 1556–66.
- Subramanian S, Trischler A, Bengio Y, Pal CJ. Learning general purpose distributed sentence representations via large scale multi-task learning. *ArXiv Prepr ArXiv180400079*; 2018.
- Peters M, Neumann M, Lyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 conference of the north American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long papers); 2018. p. 2227–37.
- Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018. [https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf).
- 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; 2015. p. 1576–86.
- Wang Z, Hamza W, Florian R. Bilateral Multi-Perspective Matching for Natural Language Sentences. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press; 2017. p. 4144–50.
- Ji Y, Eisenstein J. Discriminative improvements to distributional sentence similarity. In: Proceedings of the 2013 conference on empirical methods in natural language processing; 2013. p. 891–6.
- Yin W, Schütze H, Xiang B, Zhou B. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *Trans Assoc Comput Linguist.* 2016;4:259–72.
- Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*; 2018. p. abs/1810.04805. <http://arxiv.org/abs/1810.04805>.
- Tian J, Zhou Z, Lan M, Wu Y. ECNU at SemEval-2017 task 1: leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017); 2017. p. 191–7.
- Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag.* 1988;24:513–23.

#### Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.