# QLUT at SemEval-2017 Task 1: Semantic Textual Similarity Based on Word Embeddings

**Fanqing Meng, Wenpeng Lu*, Yuteng Zhang, Jinyong Cheng, Yuehan Du, Shuwang Han**
Institute of Intelligent Information Processing, School of Information
QiLu University of Technology, Jinan, Shandong, China
mengfanqing678@163.com, lwp@qlu.edu.cn, zhangyuteng1029@163.com,
cjy@qlu.edu.cn, amaris_du@163.com, hanshuwang0909@163.com

## Abstract

This paper reports the details of our submissions in the task 1 of SemEval 2017. This task aims at assessing the semantic textual similarity of two sentences or texts. We submit three unsupervised systems based on word embeddings. The differences between these runs are the various preprocessing on evaluation data. The best performance of these systems on the evaluation of Pearson correlation is 0.6887. Unsurprisingly, results of our runs demonstrate that data preprocessing, such as tokenization, lemmatization, extraction of content words and removing stop words, is helpful and plays a significant role in improving the performance of models.

## 1 Introduction

Semantic Textual Similarity (STS) has been held in SemEval since 2012 (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015; Agirre et al., 2016), which is a basic task in natural language processing (NLP) field. It aims at computing the semantic similarity of two short texts or sentences, and the result will be evaluated on a gold standard set, which is made by several official annotators (Cer et al., 2017). In recent years, as an unsupervised method, word embedding (Mikolov et al., 2013a) becomes more and more popular in SemEval (Jimenez, 2016; Wu et al., 2016).

The paper describes the submission of our systems to STS 2017, which utilize word embedding method. Different from some teams who have used word embedding described above, what we pay attention to is the point of preprocessing evaluation data. With this consideration, we process the evaluation data with different method in order to verify whether it works or not.

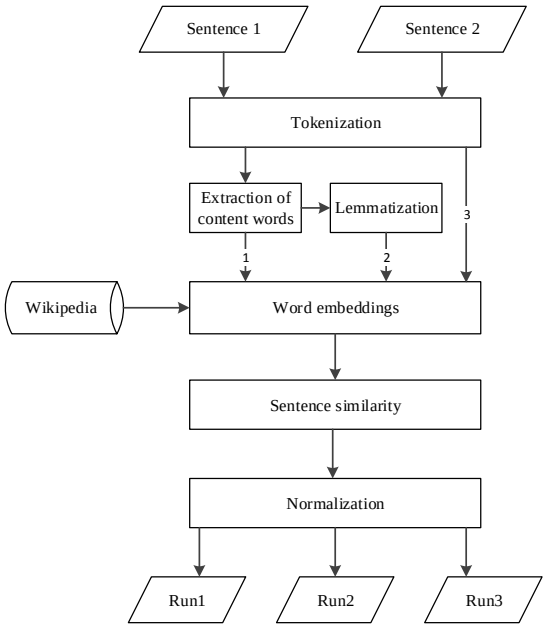


Figure 1: Framework of system.

The framework of our systems is showed in Figure 1. Its simple description is as follows:

**Tokenization:** This is to tokenize the two sentences of the system's input. Though the English sentence is tokenized naturally, the punctuations are not. For instance, the sentence "A person is on a baseball team." will be tokenized to "A person is on a baseball team .".

**Extraction of content words:** In this process, content words of the tokenized sentence will be extracted. For example, the tokenized sentence

*Corresponding author



*Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 150–153,
Vancouver, Canada, August 3 - 4, 2017. 

"A person is on a baseball team ." turns into "person is baseball team". In this paper, content words include nouns, verbs, adverbs or adjectives.

**Lemmatization:** It is known that words in English sentences have a variety of forms. This operation will lemmatize these words to their basic forms, for example, word "made" and "making" will be changed to "make". In addition, this process also convert the uppercase to lowercase, for instance, "Make" will be changed to "make".

**Word embeddings:** This process utilizes the word2vec toolkit[1] to train on the Wikipedia corpus, then the word embeddings can be obtained.

**Sentence similarity:** The similarity of two sentences is computed as the cosine of their sentence embeddings, which can be gotten easily (see 2.3).

**Normalization:** Due to the different range of the results of runs, similarity scores are normalized to meet the official standard.

## 2 System Overview

In STS 2017, we submit three system runs, all of which are unsupervised and utilize word embedding method after preprocessing.

### 2.1 Data Set

**Test Set:** The test set of the Track 5 (English monolingual pairs) consists of 250 sentence pairs. Each of these sentence pairs is in a line, split by tab.

**Gold Standard Set:** This set is the gold standard similarity score of 250 sentence pairs in the test set. The range of the score is from 0 to 5. More specially, 0 denotes that the two sentences are completely dissimilar; 1 means that the two sentences only have the same topic; 2 represents that the two sentences only have some details in common; 3 shows that the two sentences are approximately equivalent but they have some differences in the important details; 4 implies that the two sentences are roughly equivalent and some differences they have are not important; 5 indicates that the two sentences are completely equivalent.

### 2.2 Wikipedia Corpus

We use the unlabeled corpus, i.e., the English Wikipedia corpus, which have been processed by Rami Al-Rfou'[2]. The processed Wikipedia dumps have been tokenized in text format for all the languages which are considered in the evaluation. What we use in the system run is the English Wikipedia dump, after unzipped, a text file can be gotten and its size is 15.8 GB.

### 2.3 Method

In this competition, we use the word2vec toolkit on the Wikipedia corpus described above to train word embeddings. Before training word embeddings, we preprocess the text file in the corpus to transform its charset from Unicode to UTF-8, because UTF-8 is the default charset for us to run the word2vec toolkit. We set the training window size to 5 and default dimensions to 200, and choose the Skip-gram model. After trained on the corpus, the word2vec can generate a word embeddings file, in which each word in the corpus can be mapped to a word embedding of 200 dimensions. Each dimension of the word embedding is of floating point type double.

Mikolov has explained that the word embedding has semantic meaning (Mikolov et al., 2013a). Therefore, given two words, the semantic similarity of words can be easily obtained by the cosine of their word embeddings. Moreover, we can extend this to the semantic sentence similarity. Inspired by (Mikolov et al., 2013b; Wu et al., 2016), the sentence embedding of a sentence can be gained by accumulating the word embedding of all the words in it. Then by computing the cosine of two sentence embeddings, the semantic sentence similarity can be gotten as follows:

$$sim_{vec}(s_1, s_2) = \frac{\sum_{i=1}^{|s_1|} vec(w_i) \sum_{j=1}^{|s_2|} vec(w_j)}{\left|\sum_{i=1}^{|s_1|} vec(w_i)\right| \left|\sum_{j=1}^{|s_2|} vec(w_j)\right|}, \quad (1)$$

where $|s_1|$ and $|s_2|$ are the number of tokens, which sentence $s_1$ and $s_2$ include, respectively. Word $w_i$ represents the word, which belongs to $s_1$.

### 2.4 Runs

All of our runs utilize the same method described above, i.e., word emdeddings method. The only difference among them lies that each of these runs have different details in preprocessing the evaluation data. Here we clearly show their preprocessing operations in details.

**Run1:** We firstly use the Stanford CoreNLP toolkit[3] (Manning et al., 2014) to split each token for the sentence pairs in the evaluation data. Then

[1] https://code.google.com/p/word2vec/
[2] https://sites.google.com/site/rmyeid/projects/polyglot
[3] http://stanfordnlp.github.io/CoreNLP/

| Data set | Run1 | Run2 | Run3 | Run3' |
|---|---|---|---|---|
| Track 5 | 0.6155 | **0.6433** | 0.4924 | 0.5299 |

Table 1: Official evaluation results of our submitted runs on Track 5.

we tokenize all words with the help of the Stanford CoreNLP toolkit, then extract content words of the sentence pairs in the evaluation data.

**Run2:** As the operations of Run1, we tokenize the sentence pairs and extract content words for the sentence pairs in the evaluation data. Beyond that, we get the lemmas of these content words with the Stanford CoreNLP toolkit.

**Run3:** The only operation we do is to tokenize the sentence pairs of the evaluation data. Compared with Run1, all words are reserved in this run.

At last, in order to carry on the following evaluation, we normalize the output of these systems from [0, 1] to [0, 5].

The three runs are submitted to official evaluation, which are compared in Table 1.

In order to further consider the influence of stop words, we perform another group of experiences. Based on the runs in Table 1, we remove stop words which is from NLTK package. The corresponding results are shown in Table 2.

## 3 Evaluation

In the task, the official evaluation tool is based on Pearson correlation. A system run in each test set is evaluated by its Pearson correlation with the official provided gold standard set.

The results in Table 1 above shows that the system Run2 get the best performance of 0.6433. Compared with Run1, Run2 achieves a 2.78% improvement, which implies that to lemmatize content words can be helpful. The difference of 12.31% between Run1 and Run3 indicates that the extraction of content words can make a larger improvement for the similarity computation of the sentence pairs.

In order to further know the effect of lemmatization with Run3, we make the system Run3'. The only difference between them is that in the operation of preprocessing the data, Run3' makes the lemmatization of the sentence pairs in the data, on the contrary, Run3 do not do it. The contrast of Run3 and Run3' again confirms that lemmatization for computing the similarity of the sentence pairs can be effective.

| Data set | Run1- | Run2- | Run3- | Run3'- |
|---|---|---|---|---|
| Track 5 | 0.6473 | **0.6887** | 0.6341 | 0.6683 |

Table 2: Official evaluation results of our submitted runs after removing stop words on Track 5.

As is shown in Table 2, the relative performance of each run is similar with Table 1. Run2- get the best performance of 0.6887, which demonstrate the effectiveness of content words extraction and lemmatization. Each run in Table 2 achieves a better performance than that in Table 1, which demonstrates that it is necessary to remove stop words.

## 4 Conclusions and Future Work

The best Pearson correlation of our runs is 0.6887. Although our runs do not get the state-of-the-art performance, the result of these runs is acceptable. And it shows that word embeddings method is effective. Besides, in the competition, we can conclude that the appropriate preprocessing operation (such as tokenization, lemmatization, extraction of content words and removing stop words) for the data is helpful and necessary. In the future, with the help of word embeddings, we will explore some improved method to get a better performance.

## Acknowledgments

The work described in this paper is mainly supported by Natural Science Foundation of China under Grant 61502259 and 61202244, Natural Science Foundation of Shandong Province under Grant ZR2011FQ038. Thanks for the reviewers for their helpful suggestions.

## References

Eneko Agirre, Daniel Cer, Mona Diab and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics -- Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics, Montreal, Canada, pages 385--393. http://www.aclweb.org/anthology/S12-1051.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Association for Computational Linguistics, Atlanta, Georgia, pages 32--43. http://www.aclweb.org/anthology/S13-1004.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pages 81--91. http://www.aclweb.org/anthology/S14-2010.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria and Janyce Wiebe. 2015. 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)*. Association for Computational Linguistics, Denver, Colorado, pages 252--263. http://www.aclweb.org/anthology/S15-2045.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 497--511. http://www.aclweb.org/anthology/S16-1081.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio and Lucia Specia. 2017. 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, Vancouver, Canada, pages 1--14. http://www.aclweb.org/anthology/S17-2001.

Sergio Jimenez. 2016. Sergiojimenez at semeval-2016 task 1: Effectively combining paraphrase database, string matching, wordnet, and word embedding for semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 749--757. http://www.aclweb.org/anthology/S16-1116.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55--60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*. https://arxiv.org/abs/1310.4546.

Hao Wu, Heyan Huang and Wenpeng Lu. 2016. Bit at semeval-2016 task 1: Sentence similarity based on alignments and vector with the weight of information content. In Proceedings of the 10th International Workshop on Semantic Evaluation *(SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 686-690. http://www.aclweb.org/anthology/S16-1105.