Sentence Similarity Calculating Method Based on Word2Vec and Clustering

Wan-li SONG¹,²,*

¹Information Engineering School, Nanjing Xiaozhuang University, Nanjing, 211171, P.R. China
²Key Laboratory of Intelligent Information Processing, Nanjing Xiaozhuang University, Nanjing, 211171, P.R. China
*Corresponding author

Keywords: Sentence similarity, Clustering, Word2Vec.

Abstract. With the rapid development and all-round popularization of Internet, more and more data is stored in the form of text in the network platform. The massive data makes the redundancy of text information. It is very important to use text similarity technology to remove duplicate data. Therefore, how to effectively improve the accuracy and precision of text similarity calculation is an urgent problem. In this paper, we propose an improved method to calculate sentence semantic similarity. This method uses word2vec model to get the semantic information of the text, uses k-means algorithm to cluster the above results, then uses word2vec model for retraining, and finally gets the sentence similarity. Experimental results indicate that the performance of our algorithm is better improved compared with the traditional word2vec algorithm.

Introduction

The demand of information automatic processing technology for people is becoming more and more urgent [1]. The accuracy of sentence similarity calculation is the core issue of natural language automatic processing technology [2]. Sentence similarity calculation in information filtering technology, online automatic question answering system, automatic summarization technology, automatic machine translation technology, clustering and document classification technology [3]. These sentence similarity calculation results will directly affect the accuracy of information processing. In practical application, there is a very complex relationship between natural language words, which needs to be expressed in a way, semantic similarity is one of them. Due to the rich Chinese semantics, the accuracy of Chinese sentence similarity calculation results largely depends on the accuracy of word semantic similarity calculation [4]. Sentence similarity calculation is a very challenging research topic. Finding a calculation method that conforms to the characteristics of Chinese itself needs to overcome many difficulties [5].

In terms of the current research situation of text similarity, there are two kinds of methods to calculate it: one is based on statistics to calculate the similarity on sentence representation; the other is based on word sentence vector distance [6]. The method based on statistics is mainly used in scenes with a certain length (generally more than paragraphs) and achieves better results; while the method based on distance is mainly used in similarity calculation of short length texts (such as words, chunks, sentences) [7]. At present, the research of text similarity is mainly based on the research of word similarity. After a lot of research on word similarity, scholars at home and abroad have put forward many mature calculation methods. The calculation methods proposed by foreign scholars include: the method based on dictionary annotation, the method based on constituent character, the method based on word net, and the method based on direction. The calculation methods of the volume space model, search engine-based methods and other dictionary calculation methods and the other part of the calculation methods based on a large number of corpus training statistics [8]. At present, Chinese sentence similarity calculation methods are based on semantic calculation, using semantic dependency method, word shape and word order matching method, skeleton dependency tree-based
method, pattern-based method [9], editing distance-based method and so on. There are some defects in each of the above methods. For example, the calculation method based on the key semantic dictionary is more accurate in describing sentence semantics, but it is greatly limited by the dictionary itself. It requires a large number of dictionaries with vocabulary to be accurate, and it needs to be updated at all times. The calculation method based on editing distance is relatively complex and not widely used. The application scenarios will be greatly limited.

The traditional process of calculating text similarity based on word2vec model is as follows: first, use word segmentation tools to preprocess the data set, such as data cleaning, word segmentation, removing stop words, and then calculate the text similarity based on the training of word vector based on word2vec model. However, many factors will affect the evaluation of text similarity. How to optimize the model to improve the accuracy of text similarity calculation is the focus of this paper.

Our Approach

The method of retraining model based on K-means clustering algorithm is proposed: on the basis of word vector training of word2vec model, K-means clustering algorithm is used to classify the corpus, and then word2vec model is used to retrain word vector according to the classification results, and then the similarity of the text is calculated.

The evaluation of semantic similarity of word2vec model is based on context information, so it is necessary to cluster corpus to obtain context information. In this paper, we optimize the word2vec model, and propose an improved model which combines the word2vec and K-means algorithm (this paper calls it word2vec-kmeans model). The main work of the improved algorithm is as follows: on the basis of word 2vec model for word vector training, K-means algorithm is used to classify the corpus after training. After the classification results are obtained, word 2vec model is used for word vector retraining. Finally, similarity calculation is carried out on the basis of word vector set after retraining. Figure 1 shows the flow diagram of calculating text similarity with word2vec-kmeans model.

The main steps of text similarity evaluation based on word2vec-kmeans model are as follows.
(1) corpus preprocessing.
(2) word2vec trains word vectors.
(3) clustering of corpuses by K-means algorithm.
(4) training the clustering results twice to get the word vector results after training.
(5) calculate the similarity based on the result of retraining.

**Experiment Results and Analysis**

It generally requires large-scale corpus for using word2vec model training. These corpora need to be segmented accurately to show the training effect. In this paper, we use a collection of all the news data of Sogou laboratory, about 1 G, to train the word vector in this large corpus, which can obviously highlight the remarkable effect of word2vec model. The word vector tool used in this paper is the python version of gensim framework. In gensim, word2vec related APIs are all in the package'gensim. models. word2vec". After downloading the news data set of Sogou Laboratory, the data cleaning, word segmentation, removal of stop words, loading user-defined dictionaries and other preprocessing operations are carried out for the corpus.

Among all the steps we have mentioned in section 2, in our experiment, we will discuss how to determine the value of K of K-means algorithm in the step 3. Here we use the elbow method to determine the value of K. The key codes are shown in code segment 1 and the sum of squares error values corresponding to different K values are shown in Fig. 2.

**Code segment 1: to determine the K value of K-means**

01 # Use SSE to calculate the loss
02 SSE = [] # Store the sum of the squares of each result.
03 for k in range(1, 20):
04     estimator = KMeans(n_clusters=k) # Construct cluster
05     estimator.fit(word_vectors)
06     SSE.append(estimator.inertia_)
07     X = range(1, 20)
08     plt.xlabel('k')
09     plt.ylabel('SSE')
10     plt.plot(X, SSE, 'o-')
11     plt.show()

**Figure 2. K-means clustering SSE loss elbow diagram.**

As shown in Figure 2, when k is greater than 10, the slope is relatively small, and the curve behind is relatively gentle. Therefore, for this data set, k = 10 is the optimal clustering value.

In this section, through the comparative experiment between the optimized model and word2vec model, we calculate the similarity of 10 groups of sentences respectively, and analyze the calculation
results. All the 10 groups of sentences are taken from the corpus used in this paper (the corpus has been described in detail in the above chapters), the data is the real data, and table 1 is the comparison chart of the similarity results of 10 groups of sentences.

![Figure 3. Comparison of experimental results between the original model and the optimized model.](image)

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
<th>Word2vec model</th>
<th>Word2Vec-Kmeans model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy lifestyle for all</td>
<td>Action for a healthy lifestyle for all</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Social public carbon sink peer individual donation plan: one donates one tree to strive for green citizenship</td>
<td>Low carbon economy calls for green life</td>
<td>0.15</td>
<td>0.38</td>
</tr>
<tr>
<td>Collecting donations for the compatriots in Yushu earthquake stricken area, Qinghai Province</td>
<td>Siyuans focus public welfare fund helps children</td>
<td>0.41</td>
<td>0.67</td>
</tr>
<tr>
<td>Nine teachers and students from Zhejiang Province support teaching in Sichuan Province</td>
<td>Nine teachers in Zhejiang teach in Sichuan</td>
<td>0.89</td>
<td>0.96</td>
</tr>
<tr>
<td>Go to school in the depth of Liangshan</td>
<td>Children in school deep in Liangshan</td>
<td>0.69</td>
<td>0.56</td>
</tr>
<tr>
<td>Siyuans focus public welfare fund and Qumei furniture</td>
<td>Siyuans focus public welfare fund and Qumei furniture</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Healthy blood pressure activity</td>
<td>2009 healthy blood pressure theme activity kicked off</td>
<td>0.66</td>
<td>0.88</td>
</tr>
<tr>
<td>When data is represented in computer memory, the physical address is the same as the logical address and is continuous.</td>
<td>When data is represented in computer memory, the physical address is the same as the logical address and is a continuous structure</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>Direct selection sorting is adopted, with the total number of comparisons.</td>
<td>Use direct selection sorting, time complexity.</td>
<td>0.77</td>
<td>0.67</td>
</tr>
<tr>
<td>If the number of nodes of a complete binary tree is 1025, the depth of the binary tree can be deduced.</td>
<td>A binary tree with 18 nodes has the lowest height.</td>
<td>0.92</td>
<td>0.95</td>
</tr>
</tbody>
</table>

According to the comparison experiment data in Fig. 3 (the ordinate is the similarity calculation result, and the abscissa is the serial number of the comparison document), because the second group, the third group, and the seventh group are semantically similar, and the words in the sentence have the same category of words, the similarity calculation result is greatly improved compared with the word...
model. Because the first group, the fourth group, the sixth group, the eighth group and the tenth group are basically similar in semantics and structure, so the similarity rate is very high. Although the difference between the two models is basically small, the calculation results of the optimized model still have a certain degree of improvement compared with the original model. The fifth group and the ninth group have the same words, but the sentence semantic expression is basically different, the original model similarity calculation results are high. Therefore, from the data in the figure, it can be seen that the accuracy of the optimized word2vec-kmeans model is 1% to 20% higher than that of the original model. Especially in the sentence similarity evaluation with the same category and expression semantics, the effect is more obvious.

Summary
Through the analysis of Chinese sentence components, the similarity of different part of speech words in sentences is calculated respectively, and different types of corpus are trained separately by word2vec algorithm combined with K-means clustering algorithm to improve the accuracy of the model. The result shows that the method of sentence similarity calculation based on grouping training and distinguishing part of speech calculation is better than the conventional method. Because of the complexity of Chinese sentences and the different styles of the authors, the second training algorithm proposed in this paper needs to consider more characteristics of topic sentences in order to have a better calculation effect.

Acknowledgement
This work was partially supported by the following research grants: (1) No.2016NXY46 from the Research Foundation of Nanjing Xiaozhuang University and (2) No.19AIP07 from the Research Foundation of the Key Laboratory of Intelligent Information Processing of Nanjing Xiaozhuang University.

References