An Improved Method of Computing Chinese Sentence Similarity

Lu Wang[a], Zhongshi He[b]
College of computer, Chongqing University, Chongqing, China
[a]wonglou@163.com, [b]zshe@cqu.edu.cn

***Abstract*—The Chinese sentence similarity has been used widely in the field of Chinese information processing. There are many methods proposed to measure the similarity of Chinese sentences, but the accuracy of these methods is still lower. In order to improve this problem, this paper proposes an improved method to compute the similarity of Chinese sentences based on predecessors' research. The method proposed not only takes into account semantic meaning of sentences based on HowNet, CiLin and Maximum Weight Matching, but also considers the syntactic information based on tree kernel algorithm. Meanwhile, some special cases, such as antonym, double negation, are also considered. Experiments show that this improved method has relatively higher accuracy in Chinese sentence similarity computation.**

Keywords-Sentence Similarity; HowNet; CiLin; Tree Kernel; Maximum Weight Matching

I. INTRODUCTION

In the field of natural language processing, especially in the Chinese information processing, calculation of sentence similarity is a basis and a key research topic. Sentence similarity computation has a wide range of applications in the real world, its research directly determine a number of other research in related fields, such as in the document summarization system [1], in the example-based machine translation system [2], in the QA system [3] based on Frequently Asked Questions (FAQ) as well as the machine information retrieval, information filtering etc. The calculation of sentence similarity is one of the key technologies [4].

The current main methods of calculating sentence similarity from domestic and foreign scholars are as follows.

In literature [5], it proposes a sentence similarity computing method based on component. The method in literature [6] proposes that sentence similarity is composed of words semantic similarity and syntactic structure similarity by using the theory of HNC concept. Reference [7] presents an algorithm between very short texts of sentence length. It introduces a method that takes account of semantic information and word order information implied in the sentences. The paper [8] proposes a method which takes into account both semantic and syntax. Literature [9] presents a new calculation model based on multiple characteristics of the Chinese sentence similarity. Reference [10] raises a relationship vector model considering the collocation relationship among keywords and synonym keywords. Other related methods include morphological similarity [11], sentence-length similarity [12], edit distance similarity [13] and etc.

Although there are many methods of computing sentence similarity, they have three major drawbacks. First, the sentences are represented in a very high dimensional space with hundreds of dimensions, this always lead to poor performance. Second, once the similarity method is designed for an application domain, it cannot be adapted easily to other domains. Third, the accuracy of most of current methods is lower. To address these drawbacks, this paper proposes a method based on the combination of HowNet, CiLin and syntactic structure, which not only can obtain higher accuracy, but also can be used generally in applications requiring sentence similarity computation.

II. AN IMPROVED METHOD OF COMPUTING CHINESE SENTENCE SIMILARITY

A. The Computing of Word Similarity

CiLin, compiled by JiaJu, Mei, is a widely used semantic dictionary. Many scholars are based on CiLin to calculate the similarity between words. Such as JiuLe, Tian [14], however, for some words, especially for opposites, the similarity value calculated by this method, to some extent, is unreasonable. For example, considering the following words:

TABLE I. SIMILARITY VALUE OF REFERENCE [14]

Word1	Word2	Similarity value of reference[14]
beautiful	ugly	0.869
good	bad	1.0
ponder	ideology	1.0
trusted follower	trusted subordinate	0.1

In general, the words, such as “beautiful” and “ugly”, “good” and “bad” are opposite from the semantic point of view. Therefore the similarity of them should be very low. However, it is very high in the reference of [14]. In a similar way, “trusted follower” and “trusted subordinate” are synonymous, but the similarity of them is very low in the reference of [14].

HowNet is a commonsense knowledge base, which aims to revealing the relationship of concept and the relationship of the concept’s attributes. Among the methods of computing the similarity between words based on HowNet, one of the most representative methods is proposed by qun liu [15]. However, for some words, the similarity value calculated by this method, still have room to be improved. For example, considering the following words:

TABLE II. SIMILARITY VALUE OF REFERENCE [15]

Word1	Word2	Similarity value of reference[15]
apple	banana	1.0
good	bad	0.81
origin	source	0.042
do one's best	endeavor	0.044

In general, although "apple" and "banana" are fruits, but, after all, there are differences between them. So, the similarity value should not be 1. "good" and "bad" are relative to ever other, the similarity between them should be very low. In a similar way, "origin" and "source" , "do one's best" and "endeavor" are synonymous in general, however the similarity between them is very low in the re reference of [15]. Meanwhile, the above words were just collected randomly. Objectively, there must be a lot of words as the above situation.

Just as the examples given above, the meaning of words can be expressed from many aspects. Either CiLin or HowNet, only gives the semantic meaning at a certain level. In order to extract the semantic information from the most aspects, this paper proposes a method based on both CiLin and Hownet to computing the similarity between words. The proposed method as follows:

First, calculate the similarity based on CiLin, and denote it by SimCiLin;

Second, calculate the similarity based on Hownet, and denote it by SimHowNet;

Third, judge if these two words are opposites through searching an antonym list. If they are opposites, then denote this item by SimAnti, and set its value 0.001.

At last, calculate the arithmetic mean of the above three items.If they are not opposites, just calculate the arithmetic mean of SimCiLin and SimHownet.

While, the method based on CiLin is reference to JiuLe, Tian [14]. The method based on Hownet is reference to qun, liu [15]. The antonym list is collected from the Internet, which has a 4000 converses and needs to expand.

B. The Computing of Chinese Sentence Similarity

Through segmentation and part-of-speech tags, a sentence can be expressed as a set of words. For example, after segmentation, two sentences can be expressed as the following two sets:

$$S_1 = \{w_{1,1}, w_{1,2}, w_{1,3}, ...w_{1,n}\}$$
$$S_2 = \{w_{2,1}, w_{2,2}, w_{2,3}, ...w_{2,m}\}$$

Here, this paper uses maximum weight matching algorithm to calculate the similarity between the above two word sets.

First of all, construct a complete bipartite graph based on the above two sets. Where, set S1 and S2 serve as the disjoint sets of bipartite graph, the words in set S1, S2 serve as their vertexes respectively.

Second, calculate the similarity between every word in S1 and S2, and put the similarity value as a weight of their corresponding edges. The method of calculating word similarity is according to the method proposed in section A.

It may have a problem that the count of words in S1, S2 are different. Hence, in order to meet the condition of |S1|=|S2|, we should put some virtual vertexes into the fewer set (S1 or S2).

Here, assumes |S1|>|S2|, then a complete bipartite graph formed as the following figure:

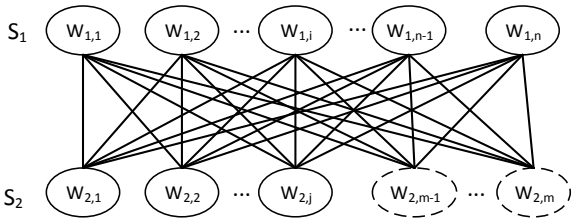

Figure 1. Bipartite graph of sentence S_1, S_2

Where, the vertexes $W_{2,m-1}$,……$W_{2,m}$ are the virtual nodes in set S_2, and then assign a definite small weight to the edge between real vertex and virtual vertex.

Third, apply Kuhn-Munkres algorithm [16] to calculate a maximum weight match from the above complete bipartite graph.

And note it as: M={e_{1x},e_{2y},……e_{ij}……},where e_{ij} is the matching edge between W_{1i} in S_1 and W_{2j} in S_2. The weight of e_{ij} is the similarity value of W_{1i} and W_{2j}.

Finally, let set M multiply the weight of its vertex and then calculate their average value, that is similarity of set S_1 and Set S_2 and is also the similarity of two sentences. The formula is as follows:

$$Sim(S_1, S_2) = w_1 W_{1S_2} + w_2 W_{2S_2} + ... + w_i W_{iS_2} + ... + w_m W_{mS_2} \quad (1)$$

Where, w_i is the weight of vertex (word) W_{1i}, W_{iS2} is the corresponding weight of match.

Considering Chinese sentences are composed of main components and secondary components. The main components are subject, predicate and object. And secondary components are attribute, adverbial and complement. In general, main components play a main role in sentence and secondary components play a second role. And as we all know, subject and object usually are noun or pronoun, and predicate usually are verb or adjective. Therefore, here, the method of giving weight to words is: giving a relative large weight to noun, verb, adjective and adverb, and a relative little weight to others.

C. Sentence Similarity Based on Syntactic Parsing Tree

Syntactic structure information is a very important feature of sentence. In order to comprehensively calculate the similarity between sentences, this paper also takes advantage of syntactic structure information of sentence. By the parsing tool, Stanford parser, it can easily parse a sentence to be a tree structure. Then, use a tree kernels algorithm to calculate the syntactic similarity of sentences.

For example, after be parsed, the syntactic parsing trees of Chinese sentences “I am a student of teacher” and “I am a student” are like figure 2:

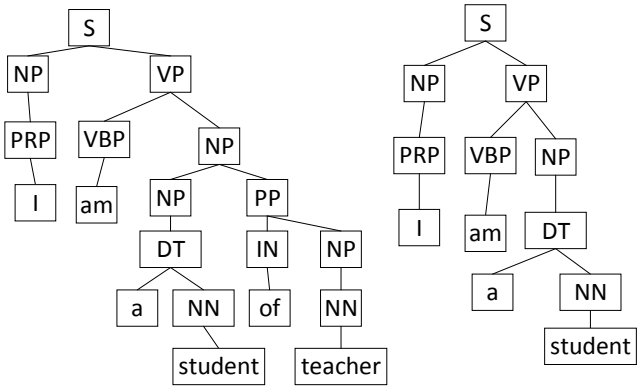

Figure 2. Parsing tree example

The most direct form of a sentence is tree structure. Through analyzing the structure of sentences, we find that tree kernel can accurately match the syntax of sentence.

Tree kernels have been widely used in many applications such as Natural Language Processing (NLP) problems, Support Vector Machines or Principal Component Analysis.

Tree kernels can be used to form representations which are sensitive to large sub-structures of trees or state sequences. It caught as much information as possible from the structure of tree to calculate the syntactic similarity by matching the same sub trees. Kernels match the syntax tree in a hierarchical way.

In this section, we use the method proposed by Collins to calculate the syntactic similarity [17]. Conceptually we begin by enumerating all tree fragments that occur in the training data 1 … … n. Note that this is done only implicitly, each tree is expressed by an n dimensional vector where the i^{th} component counts the number of occurrences of the i^{th} tree fragment. We define the function hi(T) to be the number of occurrences of the i^{th} tree fragment in tree T, so that T is now represented as $h(T)=(h_1(T), h_2(T),\ldots, h_n(T))$. Then we can get the method of syntactic similarity, the formula is as follows:

$$Sim(S_1,S_2)=Sim(T_1,T_2)=h(T_1)*h(T_2)=\sum_i^n h_i(T_1)*h_i(T_2)$$
$$=\sum_{n_1\in N_1}\sum_{n_2\in N_2}\sum_i I_i(n_1)I_i(n_2)=\sum_{n_1\in N_1}\sum_{n_2\in N_2} c(n_1,n_2) \quad (2)$$

Where n_1 and n_2 are the node number of set T_1, T_2. We define $c(n_1,n_2)=\sum_i I_i(n_1)I_i(n_2)$ and I(n) to be 1 if subtree I is at node n and 0 otherwise. Next, we note that $c(n_1,n_2)$ can be calculated in polynomial time, due to the following recursive definition:

If the productions at n_1 and n_2 are different. $c(n_1,n_2)=0$.

If the productions at n_1 and n_2 are the same, and n1 and n2 are pre-terminals, then $c(n_1,n_2)=1$.

Else if the productions at n_1 and n_2 are the same, and n_1 and n_2 are not pre-terminals,

$$c(n_1,n_2)=\prod_{j=1}^{nc(n_1)}(1+C(ch(n_1,j),C(ch(n_2,j))) \quad (3)$$

Where $nc(n_1)$ is the number of children of n_1 in the tree, because the productions at n_1/ n_2 are the same, we have $nc(n_1) = nc(n_2)$. The i’th child-node of n_1 is ch (n_1, i).

D. An Improved Method of Computing Sentence Similarity

As we know, the meaning of sentence can be expressed from many levels. Although the similarity of sentence structure can be calculated by using the syntactic information of sentence. However , it is not enough that just take syntactic information into account. For example, as follows:

(1) zhang san is my teacher.

(2) I am a student of zhang san.

Obviously, the meaning of the above two sentence to express is almost the same. But if just calculate the syntactic similarity of them, the similarity value will be very low, which is not consistent with common sense.

On the contrary, if apply the method mentioned in section B, the similarity value is almost 0.98, which is more close to common sense.

Here, have a special case that is when a sentence is double negative, which means a sentence contains two antonyms, for example:

(1) I am a good person.

(2) I am not a bad person.

In this case, it will be not ideal if use the method mentioned in section B. Therefore, in order to acquire the most meaning of sentence, this paper proposes an improved method that is combine the method in section B with the method in section C.

In this way, not only consider the semantic meaning, but also take the syntactic information into account. Meanwhile, the special case mentioned above is also considered.

Consequently, the results coincide more exactly with the common sense. The proposed method describes as follows:

(1) Calculate the sentence similarity with the method proposed in section B, and denote it with simSemantic;

(2) Calculate syntactic similarity of sentence with the method mentioned in section C, and denote it with simPTree;

(3) Judge if the two sentences contain two pairs of antonyms, if so, denote this term as simAnti, and set its value with 1. Finally, calculate the arithmetic average of the above three items.

(4) If not, just directly calculate the arithmetic average of simSemantic and simPTree.

The process of overall algorithm is as following figure:

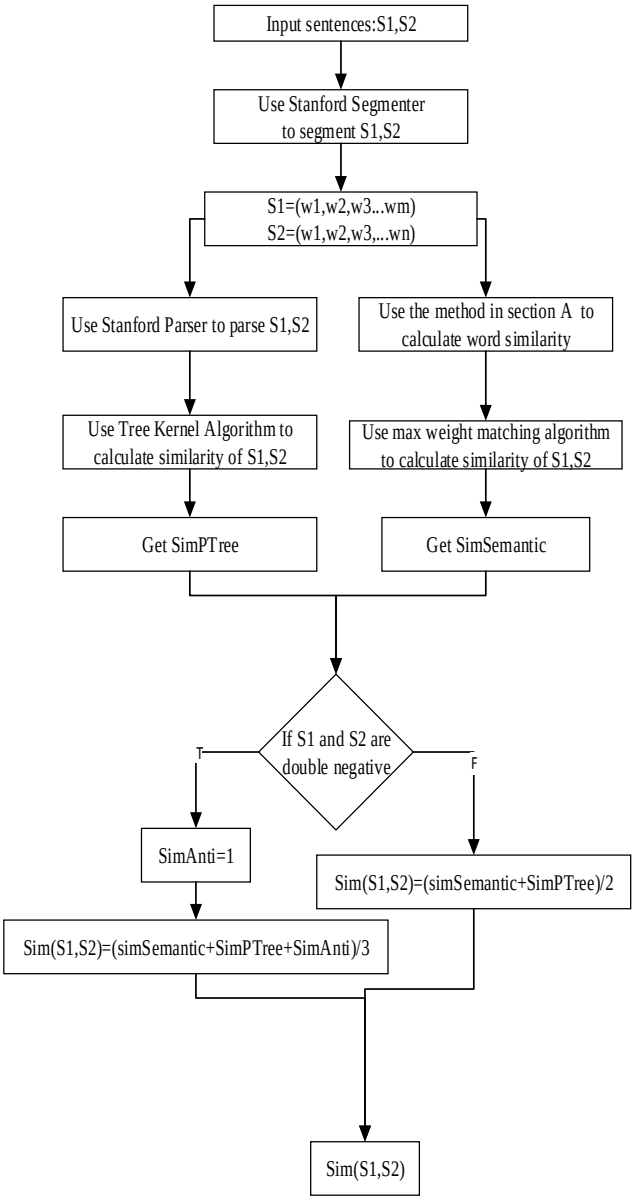

Figure 3. Process of proposed method

III. EXPERIMENTAL RESULTS

In order to verify the accuracy and generality, two test sets are used to evaluate the performance of the proposed method. The test set collected from the news of Sina.

The test set consists of 300 sentences, these sentences are divided into two parts, 250 noisy set and 50 standard set. The standard set is classified into 10 categories manually, and each category consists of about 4 to 6 sentences.

There are three ways to assess a similarity metric. First is a theory test, this assessment is a qualitative estimation, which is relatively rough. The second assessment compare with the results of human subjective judgments to see the extent of their match. The third way is to examine the application in a particular field. In this paper, the second method is used. We use limited experimental results to compare.

$$Accuracy\ Rate = \frac{sentence\ similarity\ of\ system}{sentence\ similarity\ of\ experts} \times 100\% \quad (4)$$

We compare the performance of our method to three other methods. The first one is the cosine method, which computes sentence similarity basing on Vector Space Model. The second one is semantic method, which is based on HowNet. The last one is the word overlap method, which computes sentence similarity basing on a number of words shared by two sentences.

The experimental result is as follows:

TABLE III. THE PERFORMANCE ON DATA SET

Method	Accuracy
cosine	52.67%
semantic	68.85%
word overlap	71.77%
this paper	74.34%

From the results above, it can be said the method proposed in this paper is better than the other methods, which just consider one feature. The reason why semantic method is relatively lower is that it does not take into account sentence syntactic information. The reason why the method proposed in this paper has a relatively high accuracy is that it not only considers the semantic meaning based on Hownet and Cilin, but also take into account sentence structure information and double negative sentence.

In a word, the experimental result shows that our method has a better perform in accuracy, but some further work will be needed to further improve the accuracy.

IV. CONCLUSION

This paper proposes an improved method of Chinese sentence similarity computation. The method, first, calculates semantic similarity of words based on HowNet, CiLin, Then introduces max weight matching algorithm to calculate sentence similarity. Meanwhile, uses tree kernel algorithm to calculate the sentence structure similarity. At last, combines the two algorithms to produce the final sentence similarity. Experiment shows that the method can obtain higher accuracy in Chinese sentence similarity computation.

Much further work will be needed. At first, we should improve the accuracy of similarity computing without reducing the efficiency of the system. Second, the weight of different features is assigned by ours experience, in some cases may not be optimal.

REFERENCES

[1] G. Erkan and D.R. Radev, “LexRank: Graph-Based Lexical Centrality As Salience in Text Summarization”, Artificial Intelligence Reasearch, Vol.22, 2004, pp.457-479.

[2] Y. Liu and C.Q. Zong, "Example-Based Chinese-English MT", Proc.2004 IEEE Int'l Conf Systems, Man, and Cybernetics, Vol.1-7, 2004, pp.6093-6096.

[3] P. Achananuparp, X.H. Hu, and X.J. Shen, "Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community", Proceedings of QAWeb 2008 Workshop, Beijing, China, 2008.

[4] Li Yue-lei,Shi Rui-feng.The Calculations of Chinese Sentence Semantic similarity[J]. Computer Scienc- e, 2006, 35(4A):1003-1010.

[5] Zheng Cheng,Xia Qingsong and Sun Changnian. Sentence Similarity Based on Constituent, in: Computer Technology and Development, Vol.22, No. 12(2012).

[6] Wu Zuoyan, Wang Yu. New Measure of Sentences Similarity Based on Hierarchical Network of Concepts Theory and Dependency Parsing, in: Computer Engineering and Applications, 50(3):97-102(2014).

[7] Yuhua Li, Zuhair Bandar and David McLean et al.A Method for Measuring Sentence Similarity and its Application to Conversational Agents. Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, p. 820-825, AAAI Press, Miami Beach, FL (2004).

[8] Shan Jianfang, Liu Zongtian and Zhou Wen. Sentence Similarity Measure Based on Events and Content Words, in: Academia SINICA Computing Centre, IEEE Xplore, January 21(2010).

[9] Li Chunmei, Xu Qingsheng. Research ofChinese Sentence Similarity Based on Multiple Characteristics, in: Computer Technology and Development,http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0922.055.html.

[10] Yin Yaoming, Zhang Dongzhan. Sentence Similarity Computing Based on Relation Vector Model, in: Computer Engineering and Applications, 50(2):198-203(2014).

[11] Tang Qi. Research of Sentence Similarity Computation Based on Semantic Analysis, North China Electric Power University (2009).

[12] Xue Huifang. Sentence Similarity Computing Theory and Application, Northwest University (2011).

[13] Vibhanshu Abhishek. Keyword Generation for Search Engine Advertising Using Semantic Similarity between Terms,WWW2007, May 8–12, Banff, Canada(2007).

[14] Jiu-le T, Wei Z. Words similarity algorithm based on Tongyici Cilin in semantic web adaptive learning system [J]. Journal of Jilin University (Information Science Edition), 2010, 28(6): 602-608.

[15] Liu Q, Li S. Word similarity computing based on How-net [J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76.

[16] Gong Q. Graph Theory and Network Optimization Algorithm [M]. Chongqing: Chongqing University Press, 2009: 86 — 95.

[17] Collins M, Duffy N, Convolution Kernels for Natural Language, Proceedings of the 42nd Annual Meeting on Association for ComPutational Linguisties Table of Contents, Spain, 2004:119-126.