Journal of Theoretical and Applied Information Technology
10th February 2013. Vol. 48 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SENTENCE SEMANTIC SIMILARITY CALCULATING METHOD BASED ON SEGMENTED SEMANTIC COMPARISON

YUNTONG LIU, YANJUN LIANG

School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, Henan China

ABSTRACT

In order to calculate sentence semantic similarity more accurately, a sentence semantic similarity calculating method based on segmented semantic comparison was proposed. Sentences would be divided into the trunk and the other segments by some grammar rules, and each segment might be divided into several shorter segments. When calculating the sentence semantic similarity between two sentences, the trunk and the other segments were set different weights, and the grammatical and semantic structure of the sentences would be analyzed, and the reasonable grammatical orders for segments in the two sentences would be chosen. With the method, the more reasonable and accurate sentence semantic similarity between two sentences could be calculated. Finally an experiment was provided to verify the effectiveness of the method.

Keywords: *Sentence Semantic Similarity; Semantic Similarity Calculating; Trunk of Sentence; Segments of Sentence;*

1. INTRODUCTION

In many technical fields, sentence semantic similarity calculating is the crucial technology, such as automatic question answering system, information extraction, and knowledge acquisition.

Sentence semantic similarity calculating had been deeply researched by many scholars, and they had proposed many algorithms. The semantic similarity of two sentences could be calculated using information from a structured lexical database and from corpus statistics[1]. The semantic similarity between sentences could be computed based on the semantic distances in WordNet[2][3]. The semantic similarity of two sentences could be got using information from corpus statistics based on WordNet[4]. An approach to building conversational agent by calculating semantic sentence similarity was researched[5]. A text similarity using corpus-based word similarity and string similarity was proposed in document [6]. A new sentence similarity based extractive technique for automatic text summarization was researched in document [7]. These algorithms can be divided into two categories: (1) the methods by statistical analysis according to the semantics similarity between vocabularies, however the results were not accurate enough because the grammatical structure the whole sentence were neglected; (2) the methods by analyzing the grammatical structure of a sentence according to some corpus, although the methods were effective in analyzing for one sentence, but the results were not precise enough because the two sentences might be very different in grammatical structure.

In order to get more accurate results, a sentence semantic similarity calculating method based on segmented semantic comparison was proposed. According to the general grammar rules which we had designed, the sentence could be divided into two components: the trunk and the other segments, and they were set different weights in calculating process; by the method, the more accurate similarity between two sentences could be calculated. Finally an experiment was provided to verify the effectiveness of the method.

2. THE GRAMMAR AND SEGMENTED METHOD FOR A SENTENCE

As a prerequisite, a grammar should be design to descript the sentence. By the grammar a sentence could be divided segments. The grammar had been designed according to Case Grammar[8], the details are as follows:

2.1 The Segmented Rules For The Trunk Segments

Suppose a sentence(C_S) is composed of the subject(S), the predicate(V), the object (O) and the other segments, and the appearing sequence in the sentence are S, V, O. The rules for the sentence can be described in the grammar rules (Figure 1, formula 1):

$$C_S \rightarrow LSLVLOL \qquad (1)$$

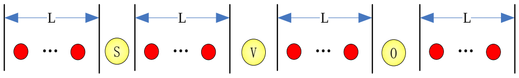

Figure 1: The Rules For The Trunk Segments

L is the parts between S, V, O.

2.2 The Segmented Rules For L

The part L is composed of two segments:

I. The attributive part (A_T);

II. The adverbial part (A_D).

The rules for L can be described in the grammar rules (Figure 2, formula 2):

$$L \rightarrow A_T A_D \,|\, A_D A_T \,|\, \epsilon \qquad (2)$$

L →ϵ means L might be null.

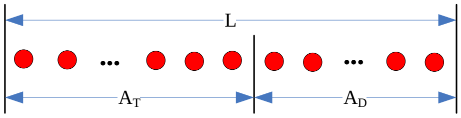

Figure 2. The Segmented Rules For L

2.3 The Segmented Rules For A_T And A_D

Suppose the attributive part (A_T) is composed $\{P_1, P_2 \ldots P_n\}$, P_i is an attribute or state to describe S or O;

Suppose adverbial part (A_D) is composed $\{P_1, P_2 \ldots P_n\}$, P_i is an case of the sentence to describe the reason, the result, etc., of the predicate(V);

The rules for A_T and A_D can be described in the grammar rules (formula 3):

$$\begin{aligned} A_D &\rightarrow A_D P \,|\, P \,|\, \epsilon \\ A_T &\rightarrow A_T P \,|\, P \,|\, \epsilon \end{aligned} \qquad (3)$$

Generally, the conjunctions and the prepositions in a sentence might be the semantic boundaries, so they were selected as grammatically-partial word (Figure 3):

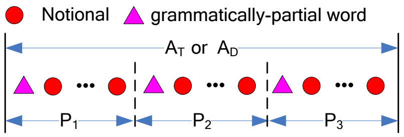

Figure 3. The Segmented Rules For A_T And A_D

2.4 The Processing Method For Different Grammatical Order

In this article, we using the grammatical order of the sentence by formula (1) (CS → LSLVLOL) as an example to describe the sentence s; in practice, a sentence might have different grammatical orders; all the possible grammatical orders were shown in Table 1:

Table 1: All the possible grammatical orders and their rules

Grammatical Orders	SVO	OVS	VSO	VOS	SOV	SOV	SV	OV	VS	VO
CS→	LSLVLOL	LOLVLSL	LVLSLOL	LVLOLSL	LSLOLVL	LSLOLVL	LSLVL	LOLVL	LVLSL	LVLOL

As shown in Table 1, there are 10 kinds of grammatical orders, and the part between S, V, O for any grammatical order are all L, so, the same method can be using to calculate the sentence semantic similarity.

3. CALCULATING SENTENCE SEMANTIC SIMILARITY

Calculating sentence semantic similarity had three stages:

I. Calculating the semantic similarity of the trunk;

II. Calculating the semantic similarity of L;

Ⅲ. Calculating the semantic similarity of the whole sentence.

Suppose there are two sentences C_{SA} and C_{SB}; according to the formula 1, set $C_{SA}=L_{A1}S_A L_{A2} V_A L_{A3} O_A L_{A4}$, $C_{SB}=L_{B1}S_B L_{B2} V_B L_{B3} O_B L_{B4}$.The semantic similarity between C_{SA} and C_{SB} could be calculated as follows:

3.1 Calculating the semantic similarity of the trunk

3.1.1 The semantic similarity between two words

Suppose there are two words W_i and W_j, and there is a lexical semantics library organized as a tree, such as Wordnet. The semantic similarity between W_i and W_j, could be calculated by formula 4:

$$\mathrm{Sim}(W_i,W_j)=\begin{cases}1-\mathrm{dis}(W_i,W_j)/D & W_i\infty W_j\\ 0 & !(W_i\infty W_j)\end{cases} \quad (4)$$

$W_i \infty W_j$ means W_i is the ancestor of W_j in the lexical semantics library (using wordnet), dis(W_i, W_j) is the distance between W_i and W_j in the tree, D is the maximum depth of the tree for the lexical semantics library.

3.1.2 Selecting the words of the trunk and segmenting the sentence

For each verb V_i(If no verb, then selecting the adjectives) in C_{SA} and each verb V_j in C_{SA}, the semantic similarity value Sim(V_i,V_j) could be calculated by formula 4; Selecting the V_i and V_j with the maximum Sim(V_i,V_j), set $V_A = V_i$ as the predicate of C_{SA} ,and set $V_B = V_j$ as the predicate of C_{SB}.

Using a similar method, the subject S and the object O of C_{SA} and C_{SB} could be selected.

When S,V,O was selected, all the segments of the C_{SA} and C_{SB} were determined. The semantic similarity for whole sentences could be calculated in the next step.

3.1.3 The formula for calculating the semantic similarity of the trunk

When we calculating the semantic similarity for whole sentences, it is obvious that the importance of the trunk is far more than the ordinary segments L, so they should be calculated by different methods; the semantic similarity of the trunk could be calculated as formula 5:

$$F_1(C_{SA},C_{SB})=(\mathrm{Sim}(S_A,S_B)+\mathrm{Sim}(V_A,V_B)+\mathrm{Sim}(O_A,O_B))/3 \quad (5)$$

3.2 Calculating the semantic similarity of the segment L

3.2.1 The jaccard similarity calculation formula

When calculating the semantic similarity of the segment L, the jaccard similarity calculation formula had been adopted, which was discussed in detail in document [9] [10].

The Jaccard similarity for two collections S and T could be calculated by formula 6:

$$\mathrm{Sim}(S,T) = |S\cap T| \;/\; |S\cup T| \quad (6)$$

$|S\cap T|$ is the number of the common elements of S and T, $|S\cup T|$ is the total number of the elements of S and T.

3.2.2 Selecting the corresponding segment for L

It had been discussed in 2.3 that the two sentences might be different grammatical orders in many cases. So for each the segment L_{Ai} in C_{SA}, the corresponding segment L_{Bj} should be selected by semantic logic.

The selection principle is: according to the semantic logic order of L_{Ai} in C_{SA} divided by S_A,V_A,O_A, selecting L_{Bj} in C_{SA} which is in the same semantic logic order.

For an instance shown in Figure 4, the corresponding four segments are as below:

$L_{A1} \leftrightarrow L_{B1}$

$L_{A2} \leftrightarrow L_{B3}$

$L_{A3} \leftrightarrow L_{B2}$

$L_{A4} \leftrightarrow L_{B4}$.

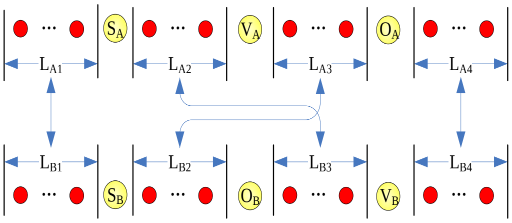

Figure 4. Selecting The Corresponding Segment

3.2.3 Calculating the semantic similarity of L

Suppose there are two segments L_A and L_B, according to 2.3, L_A and L_B could be divided into shorter segments, we can set $L_A=\{P_{A1}, P_{A2}, \ldots P_{An}\}$,$L_B=\{P_{B1}, P_{B2}, \ldots P_{Bm}\}$; For P_{Ai} and P_{Bj}, the jaccard similarity could be calculated by formula 6.

The jaccard similarity requires there must be common elements in P_{Ai} and P_{Bj}, the determination condition is as follows (formula 7):

$$\begin{aligned}&P_{Ai}\in\{P_{A1},\ P_{A2},\ \ldots P_{An}\}\\&P_{Bj}\in\{P_{B1},\ P_{B2},\ \ldots P_{Bm}\}\\&P_{Ai}=P_{Bj}\end{aligned} \quad (7)$$

However, in the calculating process, there might not be elements meeting the conditions $P_{Ai}=P_{Bj}$, because which means P_{Ai} and P_{Bj} must be the same words.

Actually, it need not to meet such stringent conditions in calculating process, we could modify the formula 7 as formula 8:

$$\begin{aligned}&P_{Ai}\in\{P_{A1},\ P_{A2},\ \ldots P_{An}\}\\&P_{Bj}\in\{P_{B1},\ P_{B2},\ \ldots P_{Bm}\}\\&\exists W_p \exists W_q (W_p\in P_{Ai})\cap(W_q\in P_{Bj})\cap\\&\quad(\mathrm{Sim}(W_p,W_q)>0)\end{aligned} \quad (8)$$

W_p is a word in P_{Ai}, W_q is a word in P_{Bj}.

The modification means: the requirement of $P_{Ai}=P_{Bj}$ would be reduced to that there is a pair of words having similar semantics in P_{Ai} and P_{Bj}.

3.2.4 Calculating the semantic similarity of the whole sentence.

When the semantic similarity of the trunk and all the segments had been calculated, the semantic similarity of the whole sentence could be calculated by formula 9:

$$F_2(C_{SA},C_{SB})=\alpha*F_1(C_{SA},C_{SB})+\beta*\sum_{i=1}^{n}Sim(L_{Ai},L_{Bi}) \quad (9)$$

α , β are the weight coefficient, $\alpha+\beta=1$.

4. EXPERIMENTAL RESULTS AND ANALYSIS

When we calculated the semantic similarity between two words we used Wordnet as the lexical semantics library, we selected 50 pairs of sentences as raw data for experiment; for each pair of sentences, set one sentence as C_{SA}, the other sentence as C_{SB}; and a threshold need be selected to determining right and wrong.

The value of α and β would have a great influence on the results, and they should be determined by experiments. So we set 11 groups values of α and β in the experiment.

In addition to this, we also made comparative experiments for the following two situations:

Situation 1: for each segment L between S,V,O, L was not divided into smaller segments;

Situation 2: for each segment L between S,V,O, L was divided into smaller segments;

The experimental results were shown in table 2, and the relations between α and the correct rates were shown in Figure 5.

Table 2: The Results For The 11 Groups Values Of α And β In The Experiment

Threshold	Situation	Correct Totol=50	α =0	α =0.1	α =0.2	α =0.3	α =0.4	α =0.5	α =0.6	α =0.7	α =0.8	α =0.9	α =1
			β =1	β =0.9	β =0.8	β =0.7	β =0.6	β =0.5	β =0.4	β =0.3	β =0.2	β =0.1	β =0
0.7	not divided	number	10	16	20	23	27	31	34	34	28	23	18
		rate	20%	32%	40%	46%	54%	62%	68%	68%	56%	46%	36%
	divided	number	11	19	25	30	33	36	43	44	41	34	26
		rate	22%	38%	50%	60%	66%	72%	86%	88%	82%	68%	52%
0.5	not divided	number	10	16	20	23	27	31	34	34	28	23	18
		rate	20%	32%	40%	46%	54%	62%	68%	68%	56%	46%	36%
	divided	number	11	19	25	30	33	36	43	44	41	34	26
		rate	22%	38%	50%	60%	66%	72%	86%	88%	82%	68%	52%
0..3	not divided	number	10	16	20	23	27	31	34	34	28	23	18
		rate	20%	32%	40%	46%	54%	62%	68%	68%	56%	46%	36%
	divided	number	11	19	25	30	33	36	43	44	41	34	26
		rate	22%	38%	50%	60%	66%	72%	86%	88%	82%	68%	52%

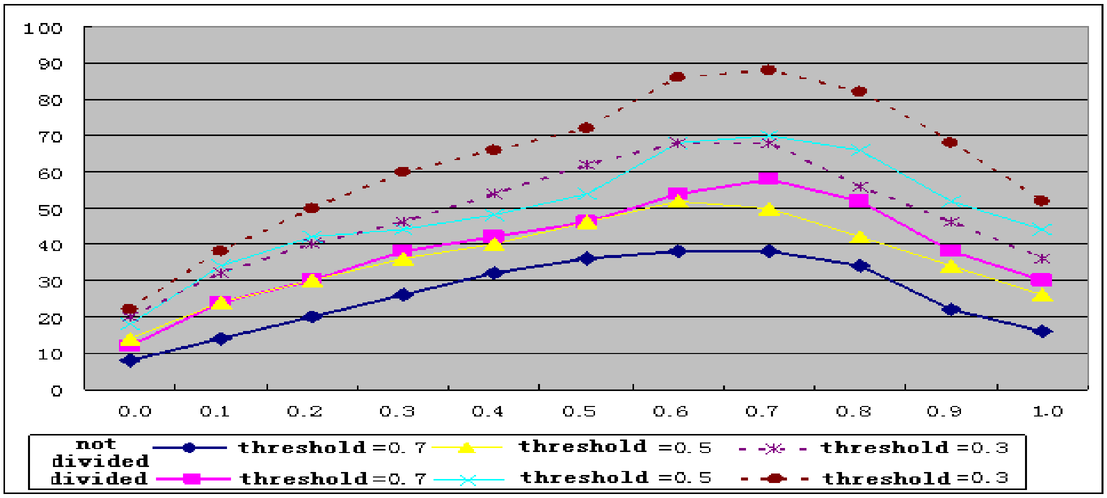

Figure 5. The Relations Between α And The Correct Rates

It can be seen from the experimental results:

I. The smaller threshold was chosen, the more similar sentences could be got; which is easy to understand according to the logic;

II. The trunk of the sentence is more important than any segments when calculating the semantic similarity. If the weight coefficient α of the trunk is between 0.6 and 0.8, the correct rate would be max.

III. For each segment L between S, V, O, if L was divided into shorter segments, the more accurate results could be achieved.

5. CONCLUSION

When calculating the sentence semantic similarity, if the sentence was divided into the trunk segment and the other segments and calculating the semantic similarity respectively, the more accurate results could be achieved and the calculating process would more fit to the semantic logic in the sentence. However, the relationship between the sentence length and results was not discussed, which would be the content of the far research.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. U1204402)

REFERENCES:

[1] Y. Li, D. Mclean, Z. Bandar, "A.Sentence similarity based on semantic nets and corpus statistics", *Knowledge and Data Engineering*, Vol. 18, No. 8, 2006, pp. 1138-1150.

[2] J. Julio, E. Marina, "Using Sentence Semantic Similarity Based on WordNet in Recognizing Textual Entailment" Advances in Artificial Intelligence- IBERAMIA 2010,Lecture Notes in Computer Science, Vol. 6433, 2010, pp. 366-375.

[3] D. Tufiş, D. Ştefãnescu, "Experiments with a differential semantics annotation for WordNet 3.0",*Decision Support Systems*, Vol. 53, No. 4, 2012, pp. 695-703.

[4] P. Selvi, N. Gopalan, "Sentence Similarity Computation Based on Wordnet and Corpus Statistics", Proceedings of the International Conference on Conference on Computational Intelligence and Multimedia Applications, November, 2007, pp.9-14.

[5] O. Karen, "An approach to conversational agent design using semantic sentence similarity", *Applied Intelligence*, Vol. 37, No. 4, 2012, pp. 558-568.

[6] A. Islam, D. Semantic, "Text similarity using corpus- based word similarity and string similarity", *ACM Transations on Knowledge Discovery Data*, Vol. 2, No. 2, 2008, pp.1-25.

[7] M. Ramiz, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", *Expert Systems with Applications*, Vol. 36, No. 4, 2009, pp. 7764-7772.

[8] C. Kehjiann, H. ChuRen, "Information-based Case Grammar", COLING '90 Proceedings of the 13th conference on Computational linguistics, Vol. 2, 1990, pp. 54-59.

[9] G. Ivchenko, S. Honov, "On the jaccard similarity test", *Journal of Mathematical Sciences.March*, Vol. 88, No. 6,1998, pp. 789-794.

[10] R. Rousseau, "Jaccard similarity leads to the Marczewski-Steinhaus topology for information retrieval", *Information Processing & Management*, Vol. 34, No. 1, 1998, pp. 87–94.