Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 235096, 6 pages
http://dx.doi.org/10.1155/2015/235096

Research Article

Word Sense Disambiguation for Chinese Based on Semantics Calculation

Yuntong Liu and Hua Sun

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

Correspondence should be addressed to Yuntong Liu; liuyt_liuyt@126.com

Received 29 October 2014; Revised 23 December 2014; Accepted 3 January 2015

Academic Editor: Chih-Cheng Hung

In order to use semantics more effectively in natural language processing, a word sense disambiguation method for Chinese based on semantics calculation was proposed. The word sense disambiguation for a Chinese clause could be achieved by solving the semantic model of the natural language; each step of the word sense disambiguation process was discussed in detail; and the computational complexity of the word sense disambiguation process was analyzed. Finally, some experiments were finished to verify the effectiveness of the method.

1. Introduction

Currently, semantics is becoming more and more important in natural language processing. Scholars had made great progress in WSD research by analyzing the semantic relations.

Based on the semantic relevancy calculated according to HowNet, a WSD method was discussed [1]. A WSD algorithm which could disambiguate the word sense of the polysemy by the semantic relatedness in WordNet was proposed [2]. A two-stage WSD method was researched according to the semantic information on the Wiki [3]. Using the distance between words in the graph based model, a graph based WSD method was studied [4]. The Chinese sentence could be disambiguated based on HowNet in a question answering system [5]. A WSD algorithm based on the semantic relevancy in HowNet was researched [6]. A k-pruning algorithm for semantic relevancy calculating model of natural language was studied [7]. WSD can be achieved by solving a model based on WordNet [8]. According to the semantic tree in WordNet, word sense could be disambiguated [9, 10].

Although the research had made considerable achievements, the WSD results were still not accurate enough in practice. In order to solve the problem more effectively and accurately, a WSD algorithm for the Chinese sentences was proposed. By the method, a Chinese clause could be disambiguated by analyzing the semantic relevancy. In the end, we verified the effectiveness of the method through some experiments.

2. The Basic Theory

2.1. The Semantic Relevancy Calculation Model. Suppose that each word W_i (except for the predicate words) in a sentence (C_S) semantically describes another word W_{Gi}; the semantic relevancy between W_i and W_{Gi} could be represented by the correlation function $\mathrm{Rel}(W_i, W_{Gi})$.

Suppose there are m kinds of parsing process for the sentence C_S; in the ith parsing process f_{Ai}, V are the predicate words, S are the subject words, and O are the object words. The semantic relevancy of the sentence for f_{Ai} can be expressed by formula (1), as shown in Figure 1:

$$f_{Ai} = K_{SVO} * (\mathrm{Rel}(S, V) + \mathrm{Rel}(O, V)) + \sum_{i=1}^{n} * \mathrm{Rel}(W_i, W_{Gi}). \quad (1)$$

In formula (1), n is the number of words in C_S (not including S, V, and O), K_{SVO} is the weight coefficient, generally, K_{SVO} should be proportional to the length of the sentence, and $K_{SVO} > 1$.

The Basic Principle of Model Solution. The most reasonable parsing process would be the parsing process which had the max semantic relevancy in all the m kinds of parsing process.

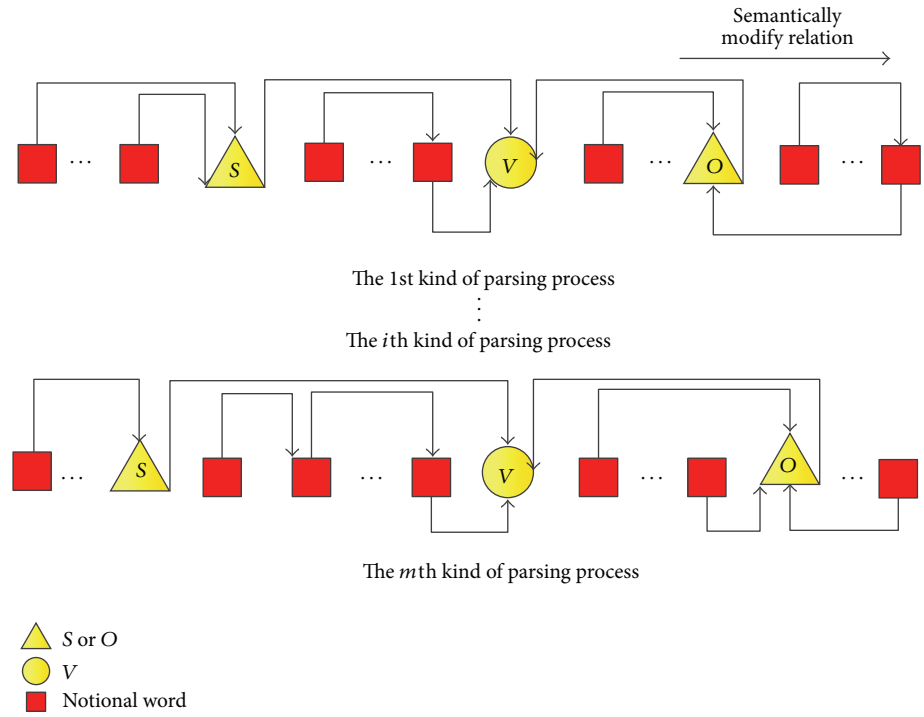

FIGURE 1: The m kinds of parsing process for a sentence.

In the calculation process, the grammatically partial words should be neglected.

2.2. The Basic Method to Solve the Model. According to the semantic structure, all the sentences in Chinese could be divided into two kinds:

(i) the simple sentences: the sentences without subordinate sentences,

(ii) the complex sentences: the sentences with subordinate sentences.

In the process to solve the model, a simple sentence might be selected, and resolute it to a word, and repeat the resolution process until the sentence becomes a simple sentence. And, in the resolution process, WSD could be finished.

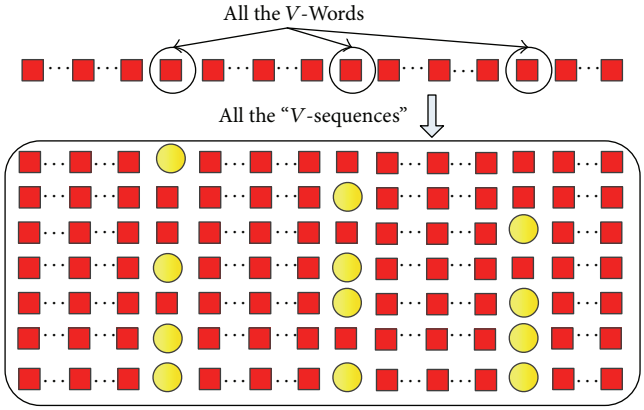

FIGURE 2: All the "V-sequences" for a sentence with 3 "V-words."

3. The Word Sense Disambiguation Process

Most words in a Chinese sentence are polysemies; the WSD process could be solved by the following steps.

3.1. Get All the "V-Sequences" for a Sentence. If a word (W) in a sentence is polysemy and one of the senses may be a verb or an adjective, the word W could be classified as "V-Word." "V-Word" is a word that may be the predicate word (V) in the sentence.

Select all the "V-Words"; the other words remain unchanged; we can arrange all the possible "V-sequences" for the sentence. When a "V-Word" is arranged, no matter how many senses of the "V-Word" were, the word would be treated as two kinds: {V, a common vocabulary}. In mathematical theory, a sentence with n "V-Words" could be arranged in $2^n - 1$ kinds of "V-sequence." As an example, Figure 2 shows all the "V-sequences" for a sentence with 3 "V-Words."

3.2. Get All the Simple Sentences for a "V-Sequence". Generally a simple clause contains only one "V-Word," so it is easy to get all the simple sentences by the exhaustive method. As an example, Figure 3 shows all the simple sentences for a "V-Word (V_i)" in a V-sequence.

In Figure 3, there might be $m * n$ kinds of the simple sentences for V_i at most.

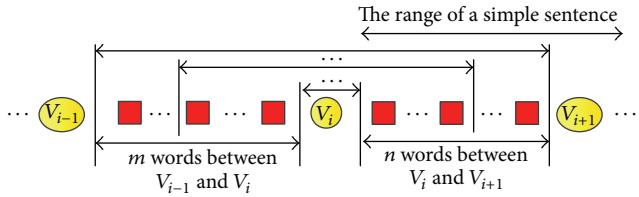

Figure 3: All the simple sentences for V_i.

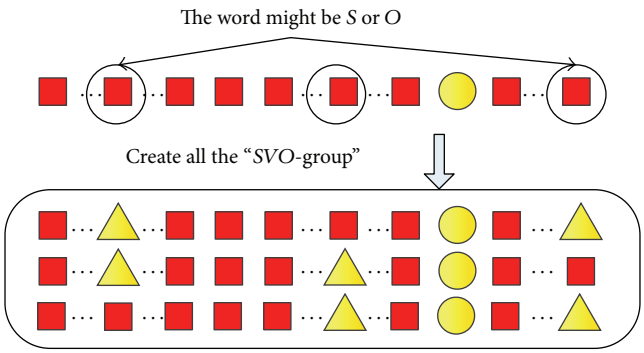

Figure 4: All the "SVO-group" for a simple sentence.

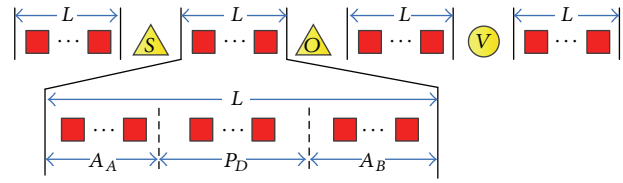

Figure 5: Dividing a simple sentence into segments.

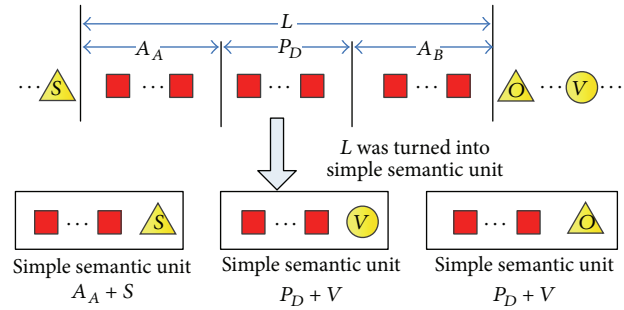

Figure 6: Turning L into simple semantic units.

3.3. Get All the "SVO-Group" for a Simple Sentence. Get all the words which might be the subject words (S) or the object words (O) by calculating the semantic relevancy; if the value of Rel(W, V) is greater than the threshold, the word might be S or O. It is easy to get all the "SVO-group" for a simple sentence, as shown in Figure 4.

3.4. Dividing a Simple Sentence into Segments. Generally, a sentence (C_S) could be divided into several segments as in Figure 5.

In Figure 5, L is the segment between S, V, and O, A_B is prepositive attributive, A_A is postpositive attributive, and P_D is adverbial.

3.5. Turning the Segments into Some Simple Semantic Units. A segment L between S, V, and O could be turned into several simple semantic units in semantic logic as in Figure 6.

For any simple semantic units, there would be the following semantic features:

(i) for any word W_i, W_{Gi} is in the same simple semantic unit (W_i semantically describes another word W_{Gi});

(ii) in the semantic analyzing process, a simple semantic unit could be treated as a whole, and its internal grammatical structure had no effect on the other analyzing process.

3.6. WSD for Simple Semantic Units. Most words are polysemies in natural language, so a semantic description relation graph (SDRG) for a simple semantic unit could be created for WSD. In the SDRG, all the senses of a polysemy are created as a "Generalized Vertex" and each sense is created as a vertex in the "Generalized Vertex" set.

As shown in Figure 7, W_i^j is the jth sense of W_i and W_p^q is the qth sense of W_p, and W_p^q had been semantically described by W_i^j; then a directed edge between W_i^j and W_p^q should be created, and, in words list, a directed edge should be created at the same time.

In Figure 7, there were the following key features.

(i) Except for the final "Generalized Vertex," there is only one goal for any "Generalized Vertex" to describe, and the outdegree of any "Generalized Vertex" is 1.

(ii) In each "Generalized Vertex," all the edges of a spanning tree must connect to the same vertex.

(iii) A SDRG, for each parsing method, must be a spanning tree of the complete graph of all the "Generalized Vertices."

So, the best SDRG of the best parsing method for the simple semantic units must be the maximum spanning tree (MST) of the complete semantic description relation graph of all the "Generalized Vertices." The specific details had been discussed in references [11].

3.7. Get the Best Simple Clause Resolution Sequence. According to formula (1), we could calculate the semantic relevancies of each simple clause and sum up all the values. There would be many different resolution sequences, so we should search and calculate each resolution sequence by exhaustive method during the calculation process. The resolution sequences with the best semantic relevancies are the best parsing method in semantics.

After Step 1 to Step 7, the semantic model could be solved and each polysemous word could be disambiguated.

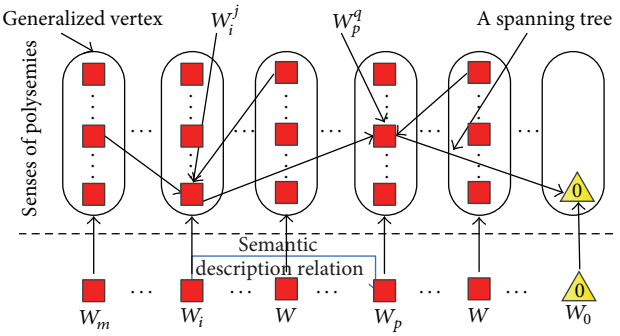

Figure 7: The SDRG of a semantic unit.

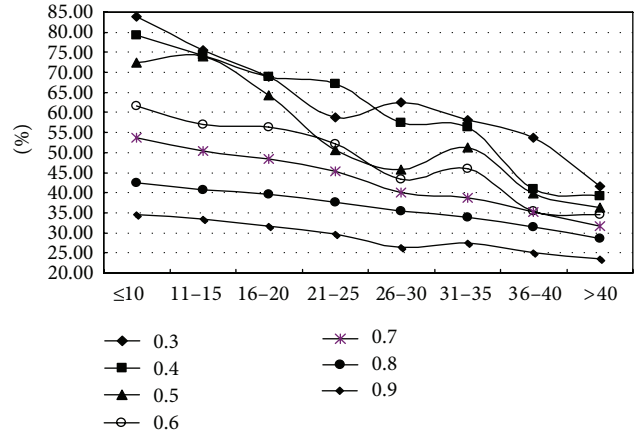

Figure 8: The correct rates for different conditions.

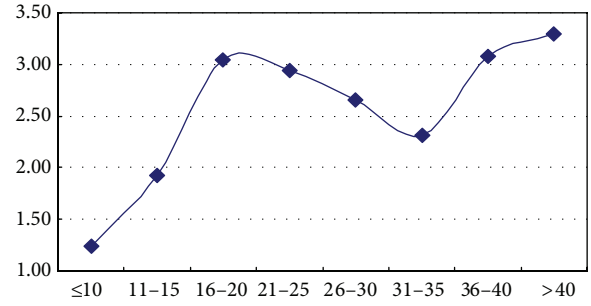

Figure 9: $(0.05 * m^5 * t^2)$/the average time.

4. The Computational Complexity

The key difficult problem is the computational complexity because the exhaustive methods are used in each step. Is the method too complex for calculating?

Suppose a sentence contained m words and n "V-Words," each word is t senses averagely; the time complexity for each step is analyzed as follows (Figure 9).

Step 1. Consider $O(2^n)$; each "V-Word" contains only 2 kinds of grammatical function.

Step 2. Consider $O((m/(n+1))^2)$; because the average length between two "V-Words" is $m/(n+1)$, the average length of a simple clause would be $(m/(n+1))^2$.

Step 3. Consider $O(k)$; k is a constant number. In theory, the time complexity is $O((m/(n+1))^2/4) * O(3^t)$; however, there would be less loss of accuracy if top-k method was to be adopted.

Step 4. Consider $O(4)$; a simple sentence might be divided into 4 segments (L) except for SVO.

Step 5. Consider $O(3)$; a segment L might be turned into 3 simple semantic units.

Step 6. Consider $O(((m/(n)) * t)^2/2)$; a segment L might be turned into 3 simple semantic units; the average length of semantic units is between m/n and $m/(3n)$; an approximate algorithm for the problem was discussed in [11].

Step 7. Consider $O(n!)$; there would be n simple clauses, so the maximum kind of simple clause resolution sequence is $n!$.

The Time Complexity. Consider the following:

$$O(2^n) * O\left(\left(\frac{m}{n+1}\right)^2\right) * O(k) * O(4) * O(3) * O(n!) * O\left(\frac{(m/n) * t)^2}{2}\right) = O\left(2^n * n! * m^4 * \frac{t^2}{n^4}\right). \quad (2)$$

Averagely, n is less than 5, t is less than 5, and only the value of m would be great, so the time complexity would not high enough for calculation in practice.

5. Experimental Results and Analysis

In the experiments, 200 Chinese sentences were selected and the HowNet was used as the lexical semantics library when calculating the semantic relevancy between two words (Windows XP; CPU: Xeon E5-2403, 2 GHz; memory: 8 G).

From the experimental results (Table 1), we can see the following.

(i) The correct rates decrease with the length of the clause.

(ii) The computational complexity increases with the length of the clause.

(iii) The time of solving the semantic model is in the same order of $m^5 * t^2$; this means that the computational complexity is $O(m^5 * t^2)$ in practice.

Using the same 200 Chinese sentences and the method in [1], we made some comparative experiments; the results are shown in Table 2.

In theory and practice, the correct rates would decrease with the length of the clause (Figure 8), but the author did not treat the different length of the clause in [1].

Table 1: Experimental results.

Sentences length	Sentences number	Number of meaning items	The average time (ms)	The correct rates for different threshold						
				0.3	0.4	0.5	0.6	0.7	0.8	0.9
≤10	11	3.4	15215	83.9%	79.3%	72.4%	61.6%	53.7%	42.4%	34.5%
11–15	26	4.3	178433	75.6%	74.1%	74.0%	57.0%	50.3%	40.8%	33.4%
16–20	42	4.1	521731	68.9%	69.0%	64.2%	56.3%	48.5%	39.6%	31.6%
21–25	51	3.5	1338696	58.7%	67.2%	50.7%	52.2%	45.3%	37.7%	29.7%
26–30	36	3.1	3115185	62.5%	57.5%	45.7%	43.4%	40.0%	35.4%	26.3%
31–35	18	2.8	6639568	58.1%	56.4%	51.2%	46.0%	38.8%	33.9%	27.5%
36–40	10	2.4	7427105	53.8%	40.9%	39.8%	35.4%	35.3%	31.4%	25.0%
>40	6	2.3	11814337	41.7%	39.1%	36.4%	34.5%	31.7%	28.6%	23.5%

Table 2: The correct rates of the comparative experiments.

Sentences length	Sentences number	Threshold = 0.8		Threshold = 0.5	
		Our method	The comparative method	Our method	The comparative method
≤10	11	42.4%	37.2%	72.4%	62.0%
11–15	26	40.8%	34.1%	74.0%	64.6%
16–20	42	39.6%	35.4%	64.2%	58.5%
21–25	51	37.7%	32.8%	50.7%	46.1%
26–30	36	35.4%	35.3%	45.7%	52.0%
31–35	18	33.9%	23.4%	51.2%	47.7%
36–40	10	31.4%	19.7%	39.8%	37.1%
>40	6	28.6%	16.8%	36.4%	21.7%

6. Summaries

In this paper, a word sense disambiguation method for Chinese based on semantics calculation was researched and WSD could be achieved by solving the semantic relevancy calculation model, and the relations between accuracy and the time complexity were explored by experiments. However, the experimental data was not enough and the accuracy was not high enough. These problems will be explored in the future research.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the Project of the National Characteristic Specialty of Computer Science and Technology (Grant no. 2009TS11576) and the Science and Technology Research Key Project of Education Department of Henan Province (Grant no. 13B520894).

References

[1] G. Z. Wang and X. F. Wang, "Word sense disambiguating method based on HowNet semantic relevancy computation," *Journal of Anhui University of Technology (Natural Science)*, vol. 25, no. 1, pp. 71–75, 2008.

[2] K. K. Deepesh, C. Jyotirmayee, and C. Alok, "Improvement in WSD by introducing enhancements in English WordNet structure," *International Journal on Computer Science and Engineering*, vol. 4, no. 7, pp. 1366–1370, 2012.

[3] C. Li, A. Sun, and A. Datta, "TSDW: two-stage word sense disambiguation using Wikipedia," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 6, pp. 1203–1223, 2013.

[4] Z.-Z. Yang and H.-Y. Huang, "Graph based word sense disambiguation method using distance between words," *Journal of Software*, vol. 23, no. 4, pp. 776–785, 2012.

[5] K. Jia, "Query expansion based on word sense disambiguation in Chinese question answering system," *Journal of Computational Information Systems*, vol. 6, no. 1, pp. 181–187, 2010.

[6] S. Q. Tian, T. Qiaoyan, and T. cheng, "Disambiguating method for computing relevancy based on hownet semantic knowledge," *Journal of the China Society for Scientific and Technical Information*, vol. 28, no. 5, pp. 706–771, 2009.

[7] Y. T. Liu, "K-pruning algorithm for semantic relevancy calculating model of natural language," *Journal of Theoretical and Applied Information Technology*, vol. 48, no. 3, pp. 231–235, 2013.

[8] K. Fragos, "Modeling wordnet glosses to perform word sense disambiguation," *International Journal on Artificial Intelligence Tools*, vol. 22, no. 2, Article ID 1350003, pp. 1345–1352, 2013.

[9] A. Minca and S. Diaconescu, "An approach to knowledge-based Word Sense Disambiguation using semantic trees built on a WordNet lexicon network," in *Proceeding of the 6th International Conference on Speech Technology and Human-Computer Dialogue (SpeD '11)*, pp. 1–6, May 2011.

[10] D. S. Suvitha and R. Janarthanan, "Enriched semantic information processing using WordNet based on semantic relation

network," in *Proceedings of the International Conference on Computing, Electronics and Electrical Technologies (ICCEET '12)*, pp. 846–851, Kumaracoil, India, March 2012.

[11] Y. Liu and H. Sun, "Word sense disambiguation for the simple semantic units based on dynamic programming," *Computer Engineering and Design*, vol. 4, no. 35, pp. 2939–2943, 2014.