Word Semantic Similarity Calculation Based on Domain Knowledge and HowNet

Xinyuan Feng[1]*, Jianguo Wei[1], Wenhuan Lu[*,2], Jianwu Dang[3]
[1]Department of Computer Science and Technology
[2]Department of Computer Software
[3]Japan Advanced Institute of Science and Technology, Tianjin University, Tianjin, China
*Corresponding author, e-mail: xinyuanfeng@tju.edu.cn

Abstract

Word semantic similarity is the foundation of semantic processing, and is a key issue in many applications. This paper argues that word semantic similarity should associate with domain knowledge, which traditional methods did not take into account. In order to adopt domain knowledge into semantic similarity measurement, this paper proposed a sensitive words sets approach. For this purpose, we also propose a new approach for sememe similarity calculation. This method distinguishes three different positional relationships between two sememes, and the results have shown that our method overperformed than other methods based on a Chinese knowledge base 'HowNet'.

Keywords*: information processing, HowNet, words semantic similarity, sensitive words sets, domain knowledge*

1. Introduction

Semantic similarity measurement of words is broadly used in many applications, such as information retrieval, information extraction, word sense disambiguation, text classification and machine translation, etc [1-4]. For English word sets, WordNet is the most popular knowledge base, which has been widely applied in many studies. For Chinese, there is another knowledge base named HowNet.

For calculating word semantic similarity in Chinese, the one based on HowNet is one of the key computing frameworks and it has a certain degree of application in the industry [5, 6]. The HowNet is a knowledge system created by Dong et al. spanning over ten years [7]. The HowNet has a complex internal structure and is rich in vocabulary semantic knowledge and knowledge of the world. Since Liu, Li [8] proposed the method which uses the hyponymy of the sememe to compute the similarity of two sememes and then to compute the similarity of words, Li [7], Jiang [10] and a number of scholars in this area have done a lot of further research.

However, the conventional methods of word similarity calculation based on HowNet are not considered the difference of the same words in diverse domain knowledge. The same words in different domain knowledge may have different similarities.The results of traditional calculation methods of two words are the same, in any case and domain. This is not consistent with the fact that the words are closely related to the domain knowledge in natural language. In this paper, we proposed a sensitive words set (SWS) approach to take domain knowledge into account for computing the word semantic similarity. The proposed semantic similarity calculation framework was evaluated by HowNet knowledge base.

The remainder of the paper is structured as follows: in section 2 we discuss the method of word similarity calculation. We discuss the methods of concept similarity computing in section3. Following that section 4 introduces the sememe similarity calculation methods. Section 5 gives the experimental results and data analysis, and we give the conclusion in section 6.

2. Word Semantic Similarity Calculation

In natural language environment, the semantic similarity of the same pair of words may have a large difference in different areas of knowledge. Therefore, each word in HowNet may

Received August 18, 2013; Revised September 16, 2013; Accepted October 15, 2013

have more than one concept. During computing of word semantic similarity, we should select the right concept according to the different areas of knowledge to calculate the word semantic similarity. The role of the sensitive words set is to realize this purpose.

But not all words have more than one concept. If the words have nothing to do with the domain knowledge, we call them isolated words. Referring to many words similarity methods based on HowNet, we defined the isolated two words similarity calculation as the maximum similarity between all of its concepts. For two words W_1, W_2, if W_1 has n concepts: $(C_{11}, C_{12}, ..., C_{1n})$ and W_2 has m concepts: $(C_{21}, C_{22}, ..., C_{2m})$. Thus, the formula of isolated word similarity is:

$$\begin{aligned} &Sim(W_1, W_2) = \max Sim(C_{1i}, C_{2j}) \\ &\qquad \text{wher e } i = 1,2,...,n;\ j = 1,2,...,m \quad \text{□□□} \end{aligned} \tag{1}$$

2.1. Sensitive Words Set

The sensitive words set is defined as: a set of keywords to reflect a particular field of knowledge. In different domain knowledge, there are some words which usually used very frequently and have a lot of semantic relevance. These words can reflect the semantic environment. Thus, we can use the sensitive words set to distinguish words in different semantic environments.

2.2. How to Select the Right Concepts

The definition of the semantic similarity between concept and word is the maximum value of concept C and concepts set $(C_1, C_2, ..., C_n)$ which belongs to the word W. The formula is as follow:

$$\begin{aligned} &Sim(\mathrm{C}, \mathrm{W}) = Max(Sim(\mathrm{C}, C_i)) \\ &\qquad \text{wher e } i = 1,2,...,n. \end{aligned} \tag{2}$$

For the words W_1 and W_2, in the case of given sensitive words set, we use each concept in the concepts set $(C_{11}, C_{12}, C_{13}, ...)$ of the word W_1 to compare with the words W_j in given sensitive words set. In the processing, we record the values of $Sim(C_{1i}, W_j)$ and when it is greater than the threshold value $\oint$, we mark the number as S_i. Then we select the concept C_{1x} whose S_i is the greatest. We select the concept C_{2y} of word W_2 using the same method. Thus, word semantic similarity computing becomes concepts semantic similarity calculating. The formula is:

$$Sim(W_1, W_2) = Sim(C_{1x}, C_{2y}) \tag{3}$$

3. Concepts Semantic Similarity Calculation

3.1. The Semantic Similarity of Functional Words' Concepts

In this paper, for the semantic similarity of functional words' concepts, we use the method which was proposed of Liu and Li [8]. The semantic similarity between functional word and notional word is defined as 0. The semantic similarity among functional words is calculated by their syntactic sememes or relational sememes.

3.2. The Semantic Similarity of Notional Words' Concepts

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations. In HowNet, notional words are made up of four parts, they are: the first basic sememe, other basic sememes, relational sememes and relational symbol sememes. So, the semantic similarity of concept consists of the four parts, too. The most intuitive method of concept semantic similarity calculation is to set a weight value for each part and to do the addition of them. The advantage of this method is simple but it has obvious defects. This paper adopts the idea of Li [9] to calculate the concept semantic similarity. Our formula is:

$$Sim(C_1,C_2)=\beta_1 Sim_1(S_1,S_2)+\beta_2 Sim_2(S_1,S_2)+\beta_3 Sim_3(S_1,S_2). \quad (4)$$

$Sim_1(S_1,S_2)$, the similarity of two concepts' all the basic sememes; $Sim_2(S_1,S_2)$, the similarity of relational sememes; $Sim_2(S_1,S_2)$, the similarity of relational symbol sememes. β_1,β_2,β_3 are all weight coefficients.

What worth noting is that while relational sememes or relational symbol sememes are empty, we set the similarity of them as the similarity of basic sememes rather than set the similarity of them as 1. This method is more suitable to conditions.

4. Sememe Semantic Similarity Calculation

4.1. Related Works

At present, many studies have been focusing on word semantic similarity calculation. Among these, Lin [11], Agirre and Rigau [12] have given their reasonable semantic similarity calculation methods. Liu [8], Li [9] and some other researchers have also proposed their own method to calculate the sememe semantic similarity of Chinese characters.

The sememes in HowNet are the basic unit to describe concepts, and they are stored in a tree structure. The similarity of sememes is described by their positions in the tree structure. In this paper, we proposed an improved method based on the distance between sememes.

Among the methods based on distance between sememes, the formula of Liu [8] is:

$$Sim(S_1,S_2)=\frac{\alpha}{d+\alpha}. \quad (5)$$

□□

α is an adjustable parameter and represents the path length when the similarity is 0.5. d means the path length of two sememes S_1,S_2 in the sememe level system.

Li [9] put forward their formula by introducing the hierarchy depth of sememes as follows:

$$Sim(S_1,S_2)=\frac{\alpha\times\min(h_1,h_2)}{\alpha\times\min(h_1,h_2)+d}. \quad (6)$$

□□

α is an adjustable parameter. h_1,h_2 mean the depth of the two sememes in the sememe level system. d means the path length of two sememes S_1,S_2.

4.2. An Improved Method of Sememe Semantic Similarity Calculation

The above calculation methods describe the positional relationship between sememes by the distance between sememes. But, they do not distinguish the following clear differences between positions of sememes.

In Figure 1(1), S_1 and S_2 are in one branch of the tree. In Figure 1(2), S_1 and S_2 have the same path length as they are in Figure 1(1), but they are in different branches. Thus, their depth is much closer than it in Figure 1(1). Obviously, the similarities of two structures above are different. In this paper, we distinguish them, thus, the similarity is more accordance with the structure of sememes.

From the organizational structure of sememes, we divided the positional relationship of two sememes into three categories. The first type is that a sememe is an ancestor of the other sememe, such as Figure 1(1). The second type is that two sememes have the same ancestor sememe, for instance, Figure 1(2). The third type is that they do not have the same ancestor sememe. For the third kind, we set the similarity as a small constant while the similarity of the second type has relation to their public ancestor sememe. Thus, we defined the three types of sememe semantic similarities as follows:

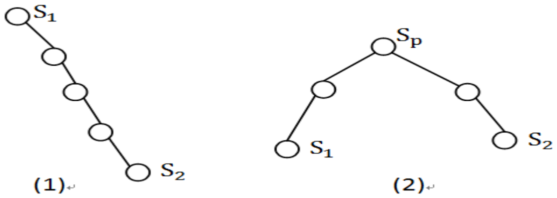

Figure 1. The position structure of sememes

The first class:

$$Sim(S_1,S_2)=\frac{\alpha\times h}{\alpha\times h+Dist(S_1,S_2)}. \quad (7)$$

□□

□

The second class:

$$Sim(S_1,S_2)=kSim(S_1,S_p)+(1-k)Sim(S_2,S_p)$$

$$where,k=\frac{x_1}{x_1+x_2}then,1-k=\frac{x_2}{x_1+x_2}. \quad (8)$$

□□□

The third class:

$$Sim(S_1,S_2)=\theta \quad (9)$$

□

□ □□□

Where, α is an adjustable parameter; S_1 and S_2 are two sememes; S_p is the first common ancestor sememe of S_1 and S_2 ; x_1 represents the shortest path length of two sememes S_1 and S_p ; x_2 represents the shortest path length of two sememes S_2 and S_p ; h represents the depth of the first common ancestor sememe; $Dist(S_1,S_2)$, represents the shortest path length of two sememes S_1 and S_2 , that is semantic distance. θ is a small constant.

In particular, we define the similarity of sememe and null value as a small constant δ and the similarity of concrete word and sememe as a small constant γ. According to Liu [8], for the similarity of concrete words, if the two words are same, the similarity is 1, otherwise, the similarity is 0.

4.3. Comparison of Three Calculation Methods

In this section, we compare the calculated results of the three sememe similarity calculation methods. First, we give the structure of sememes in HowNet. The semantic distance of two sememes is their shortest path in the Figure 2. Results are shown in Table 1.

In Figure 2, you can see that the sememes' first ancestor sememe in Group1 and Group2 respectively is "Things" and "Material" and their semantic distance is 2 and 4. The difference of sememes in the same group is that they have different position relationships.

Table 1. Sememe Similarity Calculation Results of Three Methods

Group	Word 1	Word 2	Liu, Li	Li, Li	Our
1	Internet	organization	0.444	0.615	0.500
	Things	Idea	0.444	0.615	0.333
2	Inanimate object	Tree	0.285	0.545	0.396
	Creature	Earth	0.285	0.545	0.403
	Naturals object	Plant	0.285	0.545	0.428

Comparing the calculated results in Table 1, the method of Liu does not consider the effect of the depth of sememes. The similarities of different two sememes are same as long as their semantic distance is equal. By considering the depth of sememes, the method of Li, Li can distinguish the semantic distance in different depth. But, for the sememes of same semantic distance and depth, all the methods do not distinguish the three different positional relationships. Our method distinguishes that and the results show this clearly and we can see that the more balance the sememes tree is, the greater the similarity of sememes. From the above analysis: our method can be more accurate for computing the similarity between sememes.

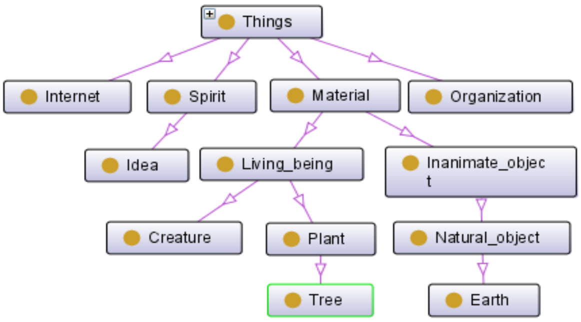

Figure 2. The Structure of Sememes in HowNet

5. Experimental Results and Data Analysis

The following experimental data gives the maximum similarity of two words and results with three sensitive words sets. The three sets are: “Chemistry”, “Biology”, “Track-and- Field Sports” and are respectively denoted as Field1, Field2, Field3.

The parameter values are selected as follows: β_1=0.7, β_2=0.15, β_3=0.15, α=0.5, the similarity between empty set and non-empty set is 0.1. The similarity between threshold of concept and words in sensitive words set $\vartheta$is 0.4.

Table 2. Similarity of Same Words in Different Fields

Word 1	Word 2	Field 1	Field 2	Field 3
People	Head	0.111	0.080	0.517
People	Material	0.245	0.191	0.787
Virus	Electricity	0.382	0.323	0.479
Raise	Metabolism	0.555	0.295	0.395
Raise	Scrabble	0.708	0.541	0.576

From the experimental results, we can see that: as we introduced the sensitive words set, the same words can have different similarity in different field. This is because the same word selects different concept according to different field. Thus, the same words’ similarities in different field are difference. What concept the words are select just as follows:

Table 3. What Concept the Words are Select in Different Fields

Word	Field 1	Field 2	Field 3
People	Null	People |He	Attribute |Manner
Head	Parts	Parts	Attribute |Style
Material	Information	Material|umbrella name	Attribute |Quality
Virus	Software	Microorganism	Null
Electricity	Electricity	Letter	Null
Raise	Null	Optimization	Promote
Metabolism	Null	Metabolism |Cure	Exchange
Scrabble	Null	Strip	Move

From Table 3, we can see that: the same word selects different concepts in different fields, such as: “Material” selects “Information” in Field1, “Material \|Umbrella name” in Field2 and “Attribute \|Quality” in Field3. Thus, when we calculate the word similarity with domain knowledge, we can select different concepts with the sensitive words sets. Therefore, in the case of selecting the appropriate sensitive words set, we can get the similarity that is more consistent with the domain knowledge.

Meanwhile, we draw the conclusions from the processing of selecting concepts in words: if two words have multiple different semantic concepts, they prefer to select the corresponding concepts, thus it can avoid the minimum similarity and we select the pair of concepts depending on the field of knowledge rather than the maximum similarity. We achieve that the same words in different field of knowledge has different similarity, because of the selected pair of concepts depending on field.

6. Conclusion

In this paper, we proposed the concept of sensitive words set. By this concept, it shows that the semantic similarity of the same words pair may different according to different domain knowledge. The experimental results illustrate that one can get a more identical word semantic similarity by considering the domain knowledge. The results show that the proposed SWS approach can improve disambiguation performance.

In order to realize the SWS-based semantic similarity calculation, we propose a new sememe similarity calculation through indepth analysis of the organization of sememes. This method distinguishes the three different positional relationships of two sememes and gives three kinds of corresponding calculation methods. The experimental results have shown that: comparing with traditional methods, our method is better for sememes which have the same semantic distance but different positions.

The results denote that our proposed SWS-based word semantic similarity calculation framework can take domain knowledge into account reasonably for semantic distant measurement. More experiments will be conducted for evaluating this framework in the further work.

Acknowledgements

This work is sponsored by the national natural science foundation of china under contract No. 61175016. And it is supported by 973 Program (Contract No. 2013CB329305). This study was also supported by the Ministry of Science and Technology of China with contract No. 2011BAH16B 1-02.

References

[1] Y Li, ZA Bandar, D Mclean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering.* 2003; 15: 871-882.

[2] L Han, T Finin, P Mcnamee, A Joshi, Y Yesha. Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy. *IEEE Transactions on Knowledge and Data Engineering.* 2013; 25: 1307-1322.

[3] R Mihalcea, C Corley, C Strapparava. *Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity.* Proceedings of the 21st National Conference on Artificial Intelligence. Boston. 2006; 775-780.

[4] Pei-Ying Zhang. A How Net-Based Semantic Relatedness Kernel for Text Classification. *Journal of Telkomnika.* 2013; 11(4): 1909-1915.

[5] Naji Hasan, Shu Gao, Malek, Al-Gabri, Zi-long Jiang. An optimal semantic network-based approach for web service composition with qos. *Journal of Telkomnika.* 2013; 11(8): 4505-4511.

[6] T Xia, Z Fan, L Liu. A Chinese Natural Language Interface Based on ALICE. *Transaction of Beijing Institute of Technology.* 2004; 24(10): 885-889.

[7] D Dong, Q Dong. HowNet, http: //www.keenage.com,1999.

[8] Q Liu, S Li. The Word Similarity Computing Based on How-net. *Computational Linguistics and Chinese Language Processing*. 2002; 7(2): 59-76.

[9] F Li, F Li. A New Approach Measuring Se-mantic Similarity in Hownet 2000. *Chinese Information Processing.* 2007; 21(3): 99-105.

[10] M Jiang, S Xiao, H Wang. An Improved Word Similarity Computing Method Based on HowNet. *Chinese Information Processing.* 2008; 22(5): 84-89.

[11] D Lin. *An Information-Theoretic Definition of Similarity.* Proceedings of the 15th International Conference on Machine Learning. 1998; 296-304.

[12] E Agirre, G Rigau. *Word Sense Disambiguation Using Conceptual Density*. Proceedings of the 16th International Conference on Computational Linguistics.1996; 258-264.