Sentence Similarity Research Based on Chinese FrameNet and Semantic Dependency Parsing

*Cheng Li,Yong Wu,*and *Binjun Wang*
College of information technology and cyber security,People's Public Security University of China,Beijing,China.

Abstract: According to the problems existing in the similarity comparison of Chinese sentences,this paper proposed a sentence similarity computing method which combined with advantages of Chinese frameNet method and semantic dependency parsing method.This method is based on the framework of semantic.And firstly,the method analyzed and calculated the similarity of two frameworks;Further,it analyzed semantic dependency relationship existing in the core frame elements from the two aspects of the meaning and the overall dependency relation;Finally,it put forward the fusion sentence similarity calculation formula.Experimental results show that compared with the method based on space vector model and based on HowNet and based on Frame Semantic Parsing, this method has higher accuracy in the similarity judgment of Chinese sentences.

Keywords: Chinese frameNet;semantic dependency parsing ;sentence similarity

1 Introduction

Sentence similarity research has shown great potential in the fields of machine translation[1],depth information retrieval[2],automatic question answering[3] and so on.

This paper will start from the Chinese FrameNet and identify target words of specific sentence firstly;next, calculate the similarity value of the two sentence framework name;and then calculate the semantic core frame elements in the dependency relation similarity;finally,measure the similarity of two sentences.

2 Related Work

At present,the research of Chinese sentence similarity is mainly divided into two aspects:One is the vector space model based on the word,and the other is the semantic sentence similarity method based on the word and syntax structure.Obviously,it is ideal to study text similarity based on syntax and semantics,But the research purely based semantic is difficult.Therefore,some experts have proposed semantic similarity research methods based on grammatical structure.At present,the more mature methods are the Chinese FrameNet method and the Semantic Dependency Parsing method.

Chinese FrameNet(CFN) is developed by Shanxi University,with Fillmore's frame semantics theory[4]and California University Berkeley FrameNet[5] as a reference method to construct a frame of semantic based on Chinese Corpus,and it includes frames,word elements,frame relationships,examples and chapters[6].

"The beggar pursues a dog.",as shown in figure 1.It is the result of using CFN semantic role automatic labeling tool.

Frame	Target word	Beggar	chased	a	dog.
following	chase	Thm	Tgt	Thm_c	

Figure 1. Framework annotation example

The target word(Tgt) in the example is "chase" and the frame name is "following"."Beggar" is the Theme of this frame and it is usually a living entity.And the Cotheme represents another moving object.

Semantic Dependency Parsing(SDP)[7] is the analysis of semantic associations between linguistic components within a sentence and semantic associations

are represented by dependency structures.Figure 2 shows the semantic dependency parsing of the example "he hears the explosion".

Figure 2. Semantic dependency analysis

The main dependency relation exist in above example is Affection(Aft) and Content(Cont),that is "hears→he" and "hears→explosion".

3 Sentence similarity computation based on Chinese FrameNet and Semantic Dependency Parsing

Through the above analysis,the Chinese FrameNet method is difficult to compare the similarities between different elements.For example,the example sentence shown in Figure 1 and "a dog chases a beggar.",their frames are "following",but the meaning of the sentences is completely different;The SDP method lacks a macro understanding of the whole sentence.For example,the example sentence shown in Figure 2 and "I smell dynamite.".the main semantic dependencies in the examples are Aft and Cont,but their frames are different,and the semantics of the two sentences are completely different.

Therefore,we propose a method of sentence similarity based on Chinese FrameNet and Semantic Dependency Parsing.

3.1 Similarity between frames

The CFN method first determines the target word of the sentence,and then determines the frame to which it is inspired based on the target word.In practical applications,we find that,the Chinese FrameNet and semantic computing research team of Shanxi University has many similar frames in the library of 323 frames accumulated over the years.If the target words of the two sentence can arouse similar frames,the two sentences may also have very high similarity.

Li Feng[8] proposed the semantic similarity algorithm of words,such as formula (1).

$$SimF(W_1, W_2)=\frac{a*\min(depth_{W_1},depth_{W_2})}{a*\min(depth_{W_1},depth_{W_2})+distance(W_1, W_2)} \quad (1)$$

The similarity between frames can be calculated in accordance with formula (1).Because CFN constructs the frame semantic resources and carries on the frame classification, and divides the same type of word element into a frame.So,in addition to some higher similarity frames,the similarity between most of the frame names are less than 0.4,and the similarity between more than half of frames is less than 0.1.

3.2 Similarity of sentence core frame elements based on Semantic dependency

If the two sentences have the same frames or a high degree of similarity is calculated in accordance with the 2.1,the semantic dependency relationships in the core frame elements within the sentence need to be further determined.

The similarity judgment of intra sentential semantic dependency is divided into two parts:single semantic dependency computation based on the meaning of words and the whole semantic dependency relation computation based on VSM.

3.2.1 Single semantic dependency computation based on the meaning of words

Semantic dependency parsing can be analyzed and determined by language technology platform of Harbin Institute of Technology[9].The results of the analysis are shown in figure 3,which are the example sentence shown in figure 1 and "a dog chases a beggar.".

```
<sent id="0" cont="乞丐追逐着一只狗。">
    <word id="0" cont="乞丐" pos="n" semparent="1" semrelate="Agt" />
    <word id="1" cont="追逐" pos="v" semparent="-1" semrelate="Root" />
    <word id="2" cont="着" pos="u" semparent="1" semrelate="mTime" />
    <word id="3" cont="一" pos="m" semparent="4" semrelate="Quan" />
    <word id="4" cont="只" pos="q" semparent="5" semrelate="Qp" />
    <word id="5" cont="狗" pos="n" semparent="1" semrelate="Pat" />
    <word id="6" cont="。" pos="wp" semparent="1" semrelate="mPunc" />
</sent>
<sent id="1" cont="狗追赶着乞丐。">
    <word id="0" cont="狗" pos="n" semparent="1" semrelate="Agt" />
    <word id="1" cont="追赶" pos="v" semparent="-1" semrelate="Root" />
    <word id="2" cont="着" pos="u" semparent="1" semrelate="mTime" />
    <word id="3" cont="乞丐" pos="n" semparent="1" semrelate="Datv" />
    <word id="4" cont="。" pos="wp" semparent="1" semrelate="mPunc" />
```

Figure 3. Semantic dependency analysis results of two example sentences

The above two sentences belong to the "following" frame,and the target words are "pursue/v" and "chase /v". The core frame elements are "beggars" and "dogs".According to the analysis of semantic

dependency relation, two sentences' Agt are "beggars" and "dog";Pat in figure 1 is "dog" and Datv in "a dog chases a beggar."is "beggar".

In the two sentences,the similarity of the specific words in the same or similar semantic dependency relation can be expressed by formula (1) and represented by $Sim(R_{1i},R_{2i})$.Among them,"1" and "2" represent the first sentence and the second sentence respectively,and "i" represents the corresponding or similar semantic dependency in the two sentences.

3.2.2 The whole semantic dependency relation computation based on VSM

The number of times a semantic dependency "r" appears in the sentence is represented by the weight "W",then all the semantic dependencies in sentences S_1 and S_2 are represented by vectors such as R_1 and R_2,as shown in formula (2),and a semantic dependency represents a dimension in vector space model whose value is the weight of the dependency "W".

According to the theory and method of the vector space model cosine method[10],it can be similar to the similarity of the semantic dependencies of two sentences,such as formula (3).

$$R_1 = R_1(w_{11}, w_{12}, \ldots, w_{1n})$$
$$R_2 = R_2(w_{21}, w_{22}, \ldots, w_{2n}) \quad (2)$$

$$SimR(R_1, R_2) = \frac{R_1 \bullet R_2}{|R_1| * |R_2|} \quad (3)$$

3.3 Sentence similarity analysis

If taking into account the structure of the sentence, the time and cost will be enormous.so this paper only consider the semanti`c frame and core frame elements in the dependency relation between sentences,and semantic frame and core frame elements of the dependency relation is called the effective collocation of, it not only can eliminate fees for core frame elements interference calculation results,also can reduce the complexity of computing.Based on the analysis of 3.1 and 3.2,the similarity computation of Chinese sentence combining Chinese FrameNet network and Semantic Dependency Parsing is presented in this paper, such as formula (4).

$$SIM(S_1, S_2)=\alpha * SimF(F_1, F_2) + \beta * \frac{\sum_{i=1}^{n} SimF(R_{1i}, R_{2i})}{n}] + \gamma * SimR(R_1, R_2) \quad (4)$$

Among them,$SIM(S_1,S_2)$ represents the similarity of sentences S_1 and S_2;F_1 and F_2 represent the frame of two sentences,and $Sim(F_1,F_2)$ represents the similarity between the two sentence frame;"n" represents the amount of corresponding or similar semantic dependencies pairs existing in the two sentence;α,β and γ are empirical parameters,they represent the contribution of sentence frame,the single semantic dependency relationship based on word sense and the semantic dependency relationship based on VSM.Their values are determined by a large number of experiments,and $0\leq\alpha\leq1,0\leq\beta\leq1,0\leq\gamma\leq1,\alpha+\beta+\gamma=1$.

Considering the semantic dependency relationship between the frames and the core frame elements and the contribution of the semantic similarity relation to the sentence similarity,and after a large number of value comparison experiments,$\alpha=0.3,\beta=0.3,\gamma=0.4$.This not only shows the semantic expression of the frame to the whole sentence,but stresses the importance of semantic dependency in the sentence.

4 Experiment and result analysis

The final test corpus contains 790 Chinese sentences,of which 90 are experimental standard set,700 are experimental test sets,and the test set includes matching set of 270 sentences and noise set of 430 sentences.Each Chinese sentence in the standard set has 3 similar statements in the matching set.

This paper adopts the detection method proposed in document [11].Extracting a sentence from the standard set each time,and calculating its similarity to each sentence in the test set,then ranking the resulting similarity from large to small.If the 2~3 sentences with the highest similarity are standard sentences with similar sentences,then the result is correct.Correct rate calculation as shown in formula(5).

$$r=\frac{\text{The correct number of sentences for the test results}}{\text{Number of sentences to be measured}} \quad (5)$$

We Use VSM method,HowNet calculation method proposed in document [12],FrameNet semantic analysis method proposed in document [13] and the method

proposed in this paper test Chinese sentences in the test corpus.The test example is shown in Table 1,and the test result is shown in table 2.

Table 1. Test case

Statement instance	Similarity value			
	1	2	3	This paper
爱夸张事实的孩子往往喜欢喜剧。	1.00	1.00	1.00	1.00
喜剧很受小学生欢迎。	0.38	0.73	0.49	0.56
孩子们爱看喜剧故事。	0.58	0.70	0.80	0.75
儿童喜欢喜剧。	0.39	0.68	0.50	1.00
孩子们有丰富的想象力。	0.43	0.35	0.05	0.19
母亲十分疼爱她的两个儿子。	0.46	0.20	0.21	0.53
丈夫总是不厌其烦地听着。	0.00	0.08	0.00	0.05
毛主席在杨家沟过生日。	0.00	0.05	0.01	0.01
凡是一般商店没有的，我们这里都有。	0.00	0.06	0.03	0.04
戏院的右边是书店。	0.05	0.01	0.02	0.03

(1: VSM;2:HowNet;3:FrameNet)

Table 2. Test results

Test Method	**The Number Of Test Sentences**	**Correct Sentences**	**Correct Rate**
VSM	90	37	41.11%
HowNet	90	66	73.33%
FrameNet	90	71	78.89%
This paper	90	76	84.44%

Can be seen from the experimental results,the method used in this paper is better than the first two methods,which analyses the dependency relation between the context and the framework inspired by the core frame elements, and don't analyze the word order and non core elements.By analyzing VSM method and HowNet method,we can indicate that the calculation result is not accurate while the sentence structure is complex and the verb is diverse.This makes it difficult for the algorithm to judge the keywords of sentences.After comparing the frame semantic dependency method and the other two methods in calculating correct sentences,we can indicate that when the sentence in the presence of Prod and Comp or two sentence words similarity is high,the method in this paper has higher accuracy.According to the basic understanding of Chinese,this method is more close to human understanding of Chinese.

5 Conclusion

This paper presents a calculation method of sentence similarity,which combines Chinese FrameNet and Semantic Dependency Parsing.This method across the surface grammatical structure and is more accurate to dig out the deep semantic relations of the sentences.This method only analyzes the core frame elements of sentences and ignores the analysis of non core framework elements when analyzing sentence components by Semantic dependency,and this can avoid interference from the unimportant components of the sentence on the results of the calculations.Next, we will further optimize the similarity calculation method,and propose an improved algorithm for

the similarity computation of long difficult sentences,and further analyze how to deal with the sentences which contain formula symbols.

Reference

[1]Yang L, Shouxun Q L. Fuzzy Matching in Machine Translation Evaluation[J]. Journal of Chinese Information Processing, 2005.

[2]Mai Z Y, Jin P, Zeng S. The Computation of Chinese Word Similarity Based on Large Scale Corpus[J]. Journal of Zhongyuan University of Technology, 2010.

[3]YANG S C, CHEN J J.Research on Sentence Similarity Computing in Chinese Automatic Answering[J]. Journal of the China Society for Scientific & Technical Information, 2008, 27(1):35-41.

[4]FILLMORE C J. Frame Semantics[M]// Linguistics in the Morning Calm. 1982:111-138.

[5]BAKER C F, Fillmore C J, Lowe J B. The Berkeley FrameNet Project[C]// Meeting of the Association for Computational Linguistics and, International Conference on Computational Linguistics. Association for Computational Linguistics, 1998:86-90.

[6]LIU K Y. Research on Chinese FrameNet Construction and Application technologies[J]. Journal of Chinese Information Processing, 2011, 25(6):46-52.

[7]GUO J, CHE W X, LIU T. Chinese Semantic Dependency Parsing[J]. Intelligent Computer & Applications, 2011, 21(6):53-56.

[8]LI F.An New Approach Measuring Semantic Similarity in Hownet 2000[J]. Journal of Chinese Information Processing, 2007, 21(3):99-105.

[9]LIU T, CHE W,LI Z. Language Technology Platform[J]. Journal of Chinese Information Processing, 2011, 2(6):13-16.

[10]LI Y C,TIAN Z, YOU J. A new character feature vector similarity function[J]. Computer Engineering & Science, 2013, 35(5):93-99.

[11]LI B, LIU T,QING B, et al. Chinese Sentence Similarity Computing Based on Semantic Dependency Relationship Analysis[J].Application Research of Computers, 2003.

[12]CHENG C P,WU Z G. A Method of Sentence Similarity Computing Based on Hownet[J].Computer Engineering & Science,2012, 34(2):172-175.

[13]Ru L, Wang Z, Li S, et al. Chinese Sentence Similarity Computing Based on Frame Semantic Parsing[J]. Journal of Computer Research & Development, 2013, 50(8):1728-1736.