Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/072L44

Calculating Similarity between Unknown Words Based on Combination Strategy

Fan Xing-Hua[1] *and Cao Rong-Li*[2]

Chinese Information Processing Lab, Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, 400065
Corresponding author: Xinghua Fan;

Received: 18 Oct. 2012, Revised: 5 Feb. 2013, Accepted: 8 Feb. 2013
Published online: 1 Jun. 2013

Abstract: This paper presents a method of calculating similarity between unknown words based on combination strategy. In the method, the best concept expression of an unknown word in corpus is obtained from the background of it, and then we construct context for the best concept expression. The connotation meaning of an unknown word is determined by the difference between the context of the best concept expression and its own context. The similarity between unknown words is calculated by utilizing semantic dictionary and the method of using the search engine to reconstruct corpus. Experimental results show that the precision, the recall and F1 in the method reaches 99.4%, 97.3% and 98.4% respectively, which are increased 21.4%, 3.2% and 13.1% respectively compared with the results got by using the method of similarity calculation based on HowNet known.

Keywords: Segmentation, Unknown Words, Similarity of Words, HowNet

1 Introduction

GBKfs As to the research that calculating similarity between unknown words which do not exist in semantic dictionary, the traditional solution utilizes the dictionary matching strategy [6] to segment the unknown word into a combination of registered words. The corresponding concept of the unknown word is represented with the combination of registered words, and then a combination of registered words is calculated. After studying, there we find are two aspects of problems about it: (1) if an unregistered word can be segmented into many combinations, how do we choose a combination for it, i.e. the mistaken segmentation problem? For example, should an unknown word “ (Ministry of State Security)” be segmented into a combination of “(quiet)” and “ (whole)” or a combination of “(safety)”and“(part)”? (2) can the combination of an unknown word express its concept, i.e. the abuse segmentation problem? For example, the concept of “ (Brownnose)” cannot be ex-pressed by the combination of “ (horse)”and “(fart)”. As to the unknown words above, the segmentation strategy cannot be applied to calculate their similarity. In order to solve the problems above, this paper represents a method of calculating similarity between unknown words based on combination strategy. The best concept expression of unknown words in corpus is obtained from the background of it, and then we construct context for the best concept expression. The connotation meaning of the unknown word is determined by the difference between the context of the best expression and its own context. Then the similarity between unknown words is calculated by utilizing semantic dictionary and the method of using the search engine to reconstruct corpus.

2 The Key Problems of Calculating Similarity between Unknown Words by Segmentation Strategy

When similarity between unknown words by segmentation strategy is calculated, the following two questions must be resolved.

Question1: which combination of registered words is the best concept expression of an unknown word?

Definition 1: (*CE*) candidate concept expression set of unknown words. For an unknown word W, if there are n

* Corresponding author e-mail: fanxh@cqupt.edu.cn

kinds combination called Assuming that time is continuous, and let x_1, x_2, x_3, ..., x_{n-1} and x_n by utilizing segmentation strategy, the candidate concept expression set of W has the form like $CE = \{x_1, x_2, x_3, \ldots, x_{n-1}, x_n\}$.

Definition 2: (*OEC*) best concept expression of unknown words. For an unknown word and its candidate concept expression set *CE*, if there exists a combination (or more) in *CE* which can express the concept expression of unknown word best in a given background corpus (or the whole corpus), we call it (or them) the best concept expression or *OEC* of an unknown word.

In this paper, the best concept expression is determined by calculating similarity between the background corpus of the unknown word and its candidate concept expression. The evaluation criterion can be also defined as follows:

$$P_i = \frac{m_i}{n} \tag{1}$$

In formula (1), P_i is the probability of the candidate concept expression x_i of *CE* as the best concept expression for the whole corpus; n is the number of background which contains unknown word; and mi is the number of background which takes x_i as its best concept express.

Question 2: how can we determine the best concept expression in corpus as the concept of an unknown word?

We first need to construct context for the best concept expression of an unknown word in the corpus when we judge the existence of the concept expression of an unknown word, and then we judge the similarity between the context of the unknown word and the context of the best concept expression in corpus. In the experiment, we set a threshold for these two contexts.

3 The Method of Calculating Similarity between Unknown Words

As a common knowledge library, HowNet [7] takes the concept of Chinese and English word as description object and the relationships between concepts and the relationship between attributes of concept as the basic content. HowNet is suitable for calculating similarity between Chinese words. Therefore, HowNet is the basic of the concrete calculating method.

3.1 Determination of the Best Candidate Concept Expression for Unknown Words

We segment the unknown words by utilizing the semantic dictionary HowNet and traverse all the segmentation combinations. According to the corresponding HowNet concept of a word, we construct the corresponding concept for each segmentation combination and put them into the expression set of candidate concepts.

In order to avoid the dilution of data, this paper takes the co-occurrence frequency score of primitive between the candidate concept expression and the context as the criterion.

Step1: Constructing co-occurrence frequency database of primitive

Co-occurrence frequency matrix [8] (YCDB) of primitive is a two-dimension table and each item corresponds to the co-occurrence frequency of a primitive pair (primitive 1, primitive 2). It is base of co-occurrence frequency score.

This paper adopts the method to construct co-occurrence frequency matrix of primitive and it is as follows:

i)Initializing the co-occurrence frequency matrix of primitive and setting each item to be 0;

ii)For the corpus, taking a sentence as a unit to calculate;

iii)For each sentence, pairing all the words of it;

iv)For each word pair, processing them as follows:

1): Finding concept description item for each word from HowNet;

2): Setting co-occurrence contribution of word to be 1, and distributing it evenly to all the concepts of it; Distributing co-occurrence contribution of a concept evenly to all the primitive of it;

3): Pairing all the primitives of word pair, and taking the product of all the co-occurrence frequency of each primitive pair as its co-occurrence frequency;

4): Summarizing the co-occurrence frequency of all primitive pairs to the co-occurrence frequency matrix.

Step2: Calculating the co-occurrence frequency score of primitive for candidate concept expression and context of unknown words

Essentially, similarity between candidate concept expression and context is the co-occurrence frequency score of primitive for candidate expression concept and context. So, we adopt the scoring standard as follows [4]:

$$score(S,C) = \frac{\sum_{\forall YE' \in C'} score(YS, YE')}{|C'|} \tag{2}$$

Here, C is the context where the unknown word appears, and S is the concept of candidate concept expression. The YS is a primitive set of S, YE' is concept set of each word in C, and C' is concept set of all words in C.

As for determination of the best concept expression of unknown word, namely, the solving of formula (1), we define the best concept expression (*OEC*) in a given context as candidate concept expression, which gets the superlative score in the context, and the concrete form is as follows:

$$OEC = \max_{\forall S \in EC} score(S,C) \tag{3}$$

After calculating *OEC* of the unknown word in each context of corpus, that is, the solving value of formula (1),

we can determine the best concept expression of unknown words.

3.2 Judging the Existence of Concept Expression for Unknown Words

3.2.1 Constructing the Context for the Best Concept Expression of Unknown Words

Step1: Getting each registered word included by the best concept express of an unknown word and searching their background corpus to get the con-texts;

Step2: Combining all the contexts of registered words obtained in step1 according to a same proportion for the best concept expression of an unknown word.

3.2.2 Judging the Existence of Concept Expression of Unknown Words

By mean of calculating the average score between the context of the best concept expression and the context of it in corpus, we can judge whether the concept expression of unknown word exists or not. If the answer is positive, the two con-texts are similar. Therefore, in order to measure their similarity, we give an estimation criterion as follows:

$$sim = \frac{1}{score(YS, GlobalYS)} (\sum_{i=0}^{K} score(S, C_i') - score(S, L)) \quad (4)$$

S in formula (4) denotes the concept of the candidate concept expression for unknown words, $GlobalYS$ is a primitive set which contains all the primitives of more than 1700, and k is the number of segmented contexts, which is decided by experiment; C_i' is an extended primitive set for the primitives of all the words in the i segmented context. $Score(S, L)$ in formula (4) represents the average score, which is defined as the average value of the scores of all the contexts. We take the best concept expression as best concept expression. The calculating formula is as follows:

$$score(S, L) = \sum_{i=0}^{K} score(S, C_i) \quad (5)$$

Here, $score(S, L)$ in formula (5) represents the average score of unknown word; and $score(S, C_i)$ means the scores of all contexts, which takes the best concept expression of the i unknown word in the corpus as the best concept expression.

3.3 Calculation of Similarity between Unknown Words

3.3.1 Construction of HowNet Concept for Unknown Words

Though the HowNet concept of segmenting unknown words can be combined by the HowNet concepts of resisted words, how can we construct the HowNet concept for a segmentable unknown word? Chinese is characteristic of gravity backward-shift [6], we suppose that unknown word is segmented into a combination of W_1 and W_2, and then we obtain the Cartesian Product of HowNet concepts of W_2, W_1, which is the set of HowNet concepts of the unknown word.

3.3.2 Similarity Computation Based on Constructed Corpus

Because the amount of information in the context of a low-frequency word is too little in a static corpus, similarity computed based on it is invalid. A viable way is to use search engine to retrieve related texts of words to construct a dynamic corpus, contexts extracted from the dynamic corpus can provide enough information to guarantee a good result in word similarity computed based on it.

i) Construct corpus;

For two Chinese words W_1 and W_2, take word W_1, word W_2 and word pair (W_1, W_2) as query term to be searched by BAIDU respectively. C_1, C_2 and C_3 are corresponding searching results (snippets, a brief window of text extracted around the query term). The corpus for word W_1 and W_2 is constructed by respectively selecting the top n_1, n_2 and n_3 snippets from C_1, C_2 and C_3. n_1, n_2 and n_3 should meet a proper proportion, which is confirmed by experiment.

ii) Context vector based on a constructed corpus

Because of the differences of similarity calculation between method based on constructed corpora and that based on conventional corpus, a dynamic context features selection to confirm the feature vector is presented in the paper. Its detailed method as follows:

Suppose that corpus C is constructed for a given word pair W_1 and W_2. A set of context features is:

$\{LEFT, RIGHT\} = \{\{WL_1, WL_2, \ldots, WL_n\}, \{WR_1, WR_2, \ldots, WR_m\}\}$

$LEFT$ is a set of words which appears in the left position of W_1 or W_2 in a corpus within a certain distance, RIGHT is a set of words which appears in the right position of W_1 or W_2 in a corpus within a certain distance. Usually, the distance is determined by experiment, in this paper is 6 words.

Context vector for word W_1 is:

$f(W_1) = \{I(W_1, WL_1), (W_1, WL_2), \ldots, (W_1, WL_n), (WR_1, W_1), (WR_2, W_1), \ldots, (WR_m, W_1)\}$

Context vector for word W_2 is:

$f(W_2) = \{I(W_2, WL_1), (W_2, WL_2), \ldots, (W_2, WL_n), (WR_1, W_2), (WR_2, W_2), \ldots, (WR_m, W_2)\}$

Here, $I(x,y)$ is the mutual information between x and y, representing the degree of correlation between them.

$$I(x,y) = \frac{P(xy)}{P(x)P(y)} \quad (6)$$

Here, $P(x)$ and $P(y)$ respectively represents probabilities of x and y in the constructed corpus C, $P(xy)$ represents the probability of co-occurrence of x and y within a certain distance.

In this paper a popular correlation coefficient in similarity computation is used [9].

3.3.3 Calculation of Similarity between Unknown Words

The essence of constructing the HowNet concept for unknown words is changing the unknown words into registered words. Therefore, we get similarity between unknown words by calculating similarity of registered words for HowNet. Part of experiments in this paper adopted the method used in reference [1] to calculate similarity.

In this paper, we utilize the method of uniting background corpus and dictionary and the method of using the search engine to reconstruct corpus to calculate similarity between unknown words. The framework of the method is as follows:

i) If the unknown words a, b can be segmented into combinations of registered words and its concept can be expressed by that of registered words in the combinations, then we can get the similarity of unknown words by calculating that between the registered words;

ii) Otherwise, respectively, take the word a, word b, pair a and b as the query terms, and then construct a context corpus for the words a and b utilizing their search results on Internet.Finally we can adopt a method based on word context to calculate similarity;

iii) Integrate the results from different sources to make them comparable in value.

4 Experiments

4.1 Experimental Data

The dataset used here is composed of four corpuses, which are from People's Daily in January, March and May and 52800 Netizen comments about topics in Sina and Sohu website randomly. The size of each corpus is about 8M. For the low-frequency words, whose frequency is less than 11, we find out all the unknown words which can be combined with registered words of HowNet. We pair all the words of the corpuses and then extract 0.005 percent of the un-synonyms pairs and 50 per-cent of the synonyms randomly by screening out the synonyms pairs and non-synonyms pairs artificially. We process the four corpuses with the same way to get the final experimental dataset.

4.2 Experiments of Segmenting the Words by Utilizing Co-occurrence Frequency Score of Primitive

Method FS: the dictionary segmentation method based on the forward maximum matching method of HowNet.

Method RS: the dictionary segmentation method based on the reverse maximum matching method of HowNet [2].

Method CF: the segmentation method proposed in this paper.

We use the following criterion to evaluate the segmentation result:

$$P = \frac{\textit{the number of words segmented correctly by utilizing this segmentation strategy}}{\textit{the total of test sets}} \times 100\%$$

The threshold of sim in the experiment is 0.00064, and the value of k is 5. If we can not construct k segmenting contexts for an unknown word; k is the maximum of the number of segmenting contexts. The results are shown in Table 1.

Table 1 The Comparison of Performance of Three Segmentation method

DatasetMethod	$FS(P)$	$RS(P)$	$CF(P)$
People'Daily in January 1998	0.9727	0.9557	0.9850
People'Daily in March 1998	0.9740	0.9692	0.9864
People'Daily in May 1998	0.9731	0.9568	0.9855
Netizen comment	0.9691	0.9621	0.9831
Average Performance	0.9722	0.9610	0.9850

As is indicated from the Table 1, the segmentation method based on background corpus is superior to the method based on dictionary matching, and it is consistent with theoretical analysis. According to the analysis, the information of unknown words is enhanced by introducing the background corpus.

4.3 Experiments of Similarity Calculation in HowNet Context by Segmentation Strategy

*Method*1(M1): calculating similarity by utilizing dictionary matching segmentation strategy [6].

*Method*2(M2): calculating similarity proposed in this paper.

We translate the evaluation of calculating similarity into question of classification for two categories, which is

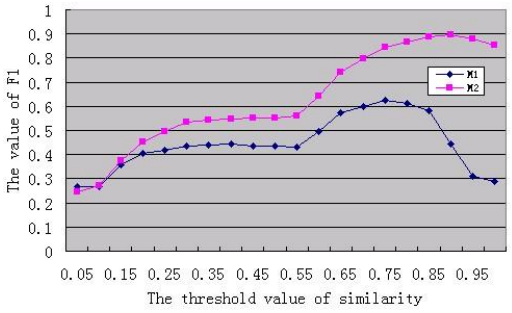

Fig. 1 the Experimental Performance Comparison of the Two Methods

related to classifying the word pairs into the synonyms word pair and the non-synonyms word pair. Given a threshold φ , if the similarity of a word pair is equal to or greater than φ , we classify it into synonyms category, otherwise, we put it into the non-synonyms category. We always use the following performance index to evaluate the transformed classification.

$$P = \frac{word\ pairs\ classified\ into\ category\ correctly}{word\ pairs\ classified\ into\ category\ of\ testing\ set} \times 100\%$$

$$R = \frac{word\ pairs\ classified\ into\ category\ correctly}{word\ pairs\ for\ a\ category\ of\ testing\ set} \times 100\%$$

$$F1 = \frac{2 \times P \times R}{P+R}$$

In this experiment, we use the parameters in reference [1] to calculate the similarity of HowNet unknown words: a=1.60; b_1= 0.5; b_2 =0.20; b_3=0.17; b_4 =0.13; g =0.2; d =0.2; the parameter k and the threshold sim in this experiment are same as those of experiment 4.2, and the experimental performance comparison of the two methods is shown in Figure 1.

The detailed data of comparison for the two methods is shown in Table 2 when the performance reaches to maximum.

Table 2 The Comparison of the Best Performance of Each Dataset

dataset	$M_1(sim = 0.75)$			$M_2(sim = 0.9)$		
	P	R	F_1	P	R	F_1
People'Daily in January	0.72 41	0.55 05	0.62 55	0.95 83	0.86 14	0.90 73
People'Daily in March	0.65 56	0.59 30	0.62 27	0.96 61	0.85 92	0.90 96
People'Daily in May	0.46 96	0.62 79	0.53 73	0.85 05	0.86 05	0.85 55
Netizen comments	0.89 81	0.56 73	0.69 53	1.0	0.88 24	0.90 45
Average performance	0.68 69	0.58 47	0.62 02	0.94 37	0.86 59	0.89 42

The figures of Table 2 indicate that the performance of calculating the similarity of unknown words with this method is improved by 25% in the long-text and 27% in the short-text comparing method proposed in paper [6]. Though the method proposed in this paper is based on the background corpus, it does not rely on the type of dataset. Therefore, the method of calculating similarity between unknown words by utilizing the background corpus and dictionary also has strong adaptability for all kinds of dataset.

4.4 Experiments of Evaluation of Method of Calculating Similarity Based on Combination Strategy of background corpus and dictionary

The evaluation of this experiment is realized by comparing the following two method of similarity calculation:

*Method*1: The method of similarity calculation based on HowNet known. In this method, the word pairs which can't be found in the HowNet will be set to error directly.

*Method*2: The combinational method of similarity calculation based on HowNet known and HowNet segment. In this method, the word pairs which can't be found in the HowNet at the same time will be set to error directly.

*Method*3: The method of similarity calculation based on strategy of combination which is proposed in this paper.

The performance of the two method of calculating similarity changes with synonymous threshold as shown in Figure 2, analyzing the figure can draw the following conclusions:

i) If the comparison of *method*1 and *method*3 can draw that, the performance of method of words'similarity calculation based on strategy of combination which is proposed in this paper relative to the method proposed in literature [5] have great increasing , suggest that the method proposed in this paper does full use of advantages of efficiency and accuracy using the method of similarity calculation based on dictionary, and combines advantage of method of calculating similarity based on statistics of wide coverage. The advantage is the method has wide coverage. Prove that the method of calculating similarity based on strategy of combination is rational and valid.

ii) Take three method of calculating similarity above when the performance reach maximum as an example.The detail data is shown in Table 3. As is shown in the table, we can get that the increase of *method*3 is not greatly relative to *method*2, the reason is that the word-pairs based on constructing corpus are less than 0.4% of total word-pairs, while the advantage of the *method*3 relative to *method*2 is that *method*3 can solve similarity calculation for words in this part. So *method*3 is a method of calculating similarity which is more comprehensive and more reasonable, but the performance of *method*2 depends on the proportion of the words based on constructing corpus in data set, higher proportion, greater impact to its performance.

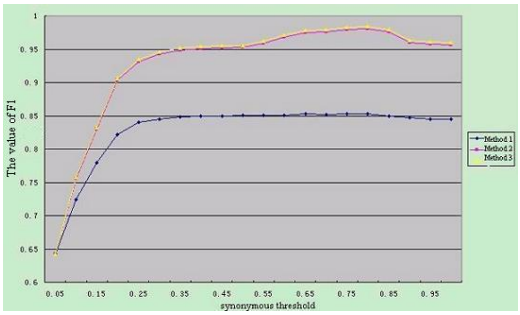

Fig. 2 Figure of the performance of the two method of calculating similarity changes with synonymous threshold

Table 3 Performance analysis of two method when the performance of calculating reach maximum

Method	Precision	Recall	F1
1	0.780	0.941	0.853
2	0.990	0.970	0.980
3	0.994	0.973	0.984

5 Conclusions

This paper presents a method of calculating similarity between unknown words based on combination strategy. This method effectively solves the mistaken segmentation problem and the abuse segmentation problem in the similarity calculation between unknown words based on the dictionary matching strategy. Meanwhile, the similarity calculation performance between unknown words is greatly improved. The experiment results show that the method in this paper achieves the desired effect.

Acknowledgement

This research was supported by the National Major Basic Research Development Project (2013CB329606).

References

[1] Sujian Li, et a1. Semantic computation in cheese question answering system [J]. Journal of Computer Science and Technology2002,17(6):933-939.
[2] Endong Gou, YanWei. The Similarity of English Words Based on Semantic Net. [J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(3): 43-48.
[3] Zhiling Zhang. Similarity of Words Based on Corpus li-brary, [J].Computer Application, 2006, 26(3): 638-644.
[4] ZhaoYan, Xiaolong Wang, Bingquan Liu. Semantic problem-solving strategy based on vector space model and maximum entropy model, [J], High Technology Let-ters. 2005, 15 (1): 1-6.
[5] LiuQun, Sujian Li. Calculating semantic similarity based on HowNet [C]. The 3rd Chinese lexical semantics proceedings. 2002,7(2):59-76.
[6] XiaTian. Research on semantics similarity of Chinese word, [J].Computer Engineer, 2007, 33(6): 192-194.
[7] Zhendong Dong, Dongqiang. HowNet[Z]. 2002-12. httpwww.keenage.corn.
[8] Erhong Yang, Guoqing Zhang, Yongkui Zhang. Semantics Disambiguation method of Chinese based on co-occurrence frequency of primitive, [J]. 2001, 38(7):833-838.
[9] GUO Qing-lin, LIN Yan-mei,TANG Qi.Similarity Computing of documents based on VSM,[J]. 2008, 25(11):3256-3258.

Fan Xing-Hua, born in 1972, Ph.D., and Professor. His research interests include artificial intelligence, natural language processing, information retrieval, network content security, and uncertain reasoning and fault diagnosis to complex system.

Cao Rong-Li is a graduate student of Chongqing University of Posts and Telecommunications, her's main research interests are in natural language processing and network content security.