Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 473-482
© The Association for Computational Linguistics and Chinese Language Processing

An Unsupervised Approach to Chinese Word Sense Disambiguation Based on Hownet[1]

Hao Chen*, Tingting He*, Donghong Ji+ and Changqin Quan*

Abstract

The research on word sense disambiguation (WSD) has great theoretical and practical significance in many fields of natural language processing (NLP). This paper presents an unsupervised approach to Chinese word sense disambiguation based on Hownet (an electronic Chinese lexical resource). In our approach, contexts that include ambiguous words are converted into vectors by means of a second-order context method, and these context vectors are then clustered by the k-means clustering algorithm. Lastly, the ambiguous words can be disambiguated after a similarity calculation process is completed. Our experiments involved extraction of terms, and an 82.62% average accuracy rate was achieved.

Keywords: Word Sense Disambiguation, Hownet,Second-Order Context, K-Means Clustering

1. Introduction

Ambiguous words are widely used in various kinds of natural languages and are distributed widely. Word sense disambiguation (WSD) has long been studied as an important problem in natural language processing (NLP), and the research on WSD has great theoretical and practical significance in many areas of NLP.

Much work has been done on WSD. In [Lu *et al*. 2002], the author presented an unsupervised approach to WSD based on a vector space model. This method defines the

[1] The study was supported by the National Natural Science Foundation of China (NSFC), (GrantNo10071028), the National Language Application Project of the Tenth five-year plan of China (GrantNo ZDI105-43B), the Ministry of Education of China, and a Research Project for Science and Technology (Grant No ZDI105-43B).

* Department of Computer Science, Huazhong Normal University, Wuhan, China 430079
E-mail: chenhaoccnu@163.com; tthe@mail.ccnu.edu.cn; quanchqin@163.com
The author for correspondence is Tingting He.
+ Institute for Infocomm Research, Heng Mui Keng Terrace 21, Singapore 119613
E-mail: dhji@i2r.a-star.edu.sg

[Received April 14; Revised June 15, 2005; Accepted August 15, 2005]

matrix of a word and calculates the importance of the context to depict the word, so that the precision position of the word in vector space can be located. However, a lot of information provided by the word sequence in the context is ignored by *tf.idf.* method. Li Juan-zi used *Cilin*(Chinese dictionary) to get semantic vectors, using a machine-readable method [Li 1999]. There are three drawbacks when using *Cilin* as a lexical knowledge base: (1) Semantic vectors are hard to determine accurately because of the coarse classification of word senses. (2) There are not enough words; *Cilin* only contains sixty thousand. (3) *Cilin* is based on the semantic frame of a hierarchical tree, with which it is hard to express the semantic relativity, especially the domain relativity between words. Yet such information plays rather important role in the WSD processing.

This paper presents an unsupervised approach to WSD based on Hownet. Compared with the above works, the main characteristics of ours are as follows:

(1) The synonymous set provided by Hownet can express word senses more concretely.

(2) Contexts that include ambiguous words are converted into vectors by means of second-order contexts, so that more information about word senses in contexts can be provided.

(3) Features are selected based on the extraction of terms.

Section 2 describes our method. In section 2.1, we introduce some related information about Hownet. Section 2.2 describes the method for establishing second-order contexts and the method for clustering these context vectors by means of the k-means algorithm. Section 2.3 describes the method for calculating the similarity between context vectors. In section 3, we describe experiments that we conducted using our method. We compare our method with other similar methods and find that we were able to obtain better results. In section 4, we conclude our work and offer some suggestions for further improvement.

2. An Unsupervised Approach to WSD Based on Hownet

With the rich lexical resources of Hownet, we firstly convert contexts that include ambiguous words into context vectors; then, the definitions (DEFs) of ambiguous words in Hownet can be determined by calculating the similarity between these context vectors. The whole process is completed automatically, so a sense-labeled corpus is not needed.

2.1 Hownet and WSD

Hownet [Dong and Dong 2000], which describes concepts in Chinese and English, is a knowledge database that determines the relationships between concepts or the attributions of concepts. In our method, we focus on the knowledge dictionary.

2.1.1 The Record Mode of the Knowledge Dictionary

The knowledge dictionary is the fundamental document in Hownet. In this document, a record contains a concept of a word and its description. Each record is made up of 4 parts .In each part, the part on the left side of "=" contains the name of the data and the part on the right side contains the value of the data. Their alignment is as follows:

W_C= Chinese Word;

W_E= English Word;

E _X= Example of the word;

G_C= Part of speech in Chinese;

DEF= Definition.

2.1.2 An Example of Ambiguous Word in Hownet:

One sense of "沽(Gu)":

No =032339;

W_C=沽(Gu);

G_C=v;

W_E= sell;

DEF=sell, commercial;

The other sense of "沽(Gu)":

No=032340;

W_C=沽(Gu);

G_C=v;

W_E= buy;

DEF=buy, commercial.

2.1.3 Word Senses of Ambiguous Words

Several different W_C in Hownet correspond to the same DEF; in other words, each word sense has a synonymous set. Table 1 gives examples that show the numbers of synonyms of

ambiguous words in Hownet.

Table 1. The numbers of synonyms of ambiguous words in Hownet

Ambiguous Word	DEF	The number of synonyms	Total number
健康	A value|属性值,kind|类型,ordinary|普,desired|良	1	47
	A value|属性值,physique|体格,strong|强,desired|良	38	
	attribute|属性,physique|体格,&Animal Human|动物	8	
造就	result|结果,#succeed|成功,desired|良	29	53
	cultivate|培养	24	

Perhaps the words in a set of synonyms are also ambiguous words. We can solve this problem using the following method: for each word in the set of synonyms, we collect its contexts and cluster these contexts by means of the k-means algorithm (section 2.2). Each word has several categories. We select one category with the smallest distance from the others, and the contexts in that category are the contexts that we want. Finally, we synthesize the contexts as the context set of the ambiguous word.

2.2 K-Means Clustering Based on Second-Order Contexts

The k-means clustering approach presented by [MacQueen 1967], is an important method for data mining and knowledge discovery, and it has the characteristics of simplicity and fast convergence.

When the k-means clustering algorithm is used, the first problem is translating the content of the document into a form that can be processed mathematically. Here, we convert contexts that include ambiguous words into vectors by means of second-order contexts. Our method is as follows:

(1) We Extract terms from the contexts that include ambiguous words by the method that offered in [Pantel and Lin 2001].

(2) A parameter called “m” represents the number of calculated occurrences. We then select terms with high “m” values as features. This step is based on our assumption that there is at least one feature in each context. If the context contains several features, we calculate the sum of the vectors of the features.

In the following, we will take 健康（health）as an example. We select the features whose numbers of occurrences are above 8 from 445 texts. Table 2 presents the co-occurrence frequency of two features in 445 texts; and Table 3 presents the vectors of 5 contexts in Table 2.

Table 2. The co-occurrence frequency of two features

feature/m $context_i$ |feature	持续 /11	快速 /11	发展 /13	老年人 /9	生活方式 /9	精神 /8	心理 /10	教育 /16	身体 /8	教师 /9	事业 /10
C1 持续	\	7	7	1	1	1	2	1	1	0	5
C1 快速	7	\	7	0	2	1	2	0	1	0	5
C1 发展	7	7	\	1	3	1	0	2	3	2	6
C2 老年人	1	0	1	\	5	3	2	1	4	5	0
C2 生活方式	1	2	3	5	\	1	1	2	3	3	0
C3 精神	1	1	1	3	1	\	2	0	3	2	1
C3 心理	2	2	0	2	1	2	\	4	2	4	2
C3 教育	0	0	2	1	2	0	4	\	0	6	5
C4 身体	1	1	3	4	3	3	2	0	\	3	4
C4 教师	0	0	2	5	3	2	4	6	3	\	6
C5 事业	5	5	6	0	0	1	2	5	4	6	\

Table 3. Vectors of contexts

Feature $Context_i$	持续	快速	发展	老年人	生活方式	精神	心理	教育	身体	教师	事业
C_1	14	14	14	2	6	3	4	3	5	2	16
C_2	2	2	4	5	5	4	3	3	7	8	0
C_3	3	3	3	6	4	2	6	4	5	12	8
C_4	1	1	5	9	6	5	6	6	3	3	10
C_5	5	5	6	0	0	1	2	5	4	6	0

Then we perform clustering using the Novel K-means algorithm [Li *et al.* 2002]. K-means adopts the iterative way to renew. In each iteration, the value of the objective function decreases. When the smallest value of the objective function is related, the best clustering result can be obtained. During k-means clustering, determining k is always a key problem .If the ambiguous word has n word senses in Hownet, we can define k as being equal to n. But if the contexts we extract from the corpus do not contain n word senses, we can using the following method to get k: for each k, we cluster separately from 2 to n. Take k＝3 as an

example, to each category of the three, we can get the highest score between the category and the word sense in Hownet. Then we can take the average score of the 3 highest scores as the final score. Lastly, we can determine k when its corresponding score is the highest. In the example discussed in section 2.2, we find 5 contexts in which “health” occurs .We determine that k＝3 when the highest average score is obtained, in other words, we can get the best result by 3 categories .The results are as follows: （C_1），（$C_2 \cdot C_4 \cdot C_5$），（C_3）.

2.3 A Way to Calculate the Similarity between Two Vectors

The similarity between two context vectors can be calculated using the ‘cosine’ formula, which is show below:

$$sim(Q, D2) = \frac{\sum_{j=1}^{t} wqj \times wd_{ij}}{\sqrt{\sum_{j=1}^{t} \left(wqj\right)^2 \times \sum_{j=1}^{t} \left(wd_{ij}\right)^2}} \quad . \tag{1}$$

The formula makes good use of the vector space model, which can formalize the content of the context into a spot and convert it into a vector. This is because documents can be defined in real number space, models can be identified, many algorithms in other realms can be applied, and the calculability and maneuverability of documents in NLP has greatly improved.

3. Experiment and Results

3.1 Experiment Based on Hownet

The steps in the algorithm are as follows:

(1) Select some contexts containing ambiguous words (W) from a corpus;

(2) Cluster these contexts into several categories using the k-means clustering algorithm;

(3) Find the word sense (w_i) of W (w_1, w_2, w_n) in Hownet and compose the synonym set for each word sense (w_i). We then extract some contexts containing synonyms from the corpus and finally, convert the contexts into second-order contexts. Each word sense is replaced by a vector, which is constructed by means of second-order contexts based on its synonyms set.

(4) Calculate the similarity between the categories in (2) for word sense; the similarity of the word sense with the largest value is the correct one.

Experimental data are shown in Table 4. Concerning the experiment, several points must be noted:

(1) The training data in our experiment came from a modern Chinese corpus.

(2) The extraction of contexts takes the sentence as a unit, and a context that lacks features is removed.

(3) Hownet, 2000 version, by Dong Zheng-Dong was adopted in our experiment.

(4) Our experiment does not have the difference between open test and close test because of the unsupervised approach.

Table 4. Experimental data and results (Hownet)

Ambiguous words	Number of contexts		Number of experimental contexts		Accuracy (%)	
	Word sense	Number of synonyms	All contexts	Correct contexts		Average (%)
材料 (cai 'liao)	S1	3	65	51	78.46	82.62
	S2	61				
	S3	7				
改(gai)	S1	33	72	60	83.33	
	S2	44				
代表 (dai biao)	S1	20	84	68	80.95	
	S2	47				
	S3	0				
	S4	6				
健康 (jian'kang)	S1	8	50	40	80.00	
	S2	38				
	S3	1				
保持 (bao'chi)	S1	27	66	56	84.48	
	S2	51				
打气 (da qi)	S1	24	46	40	86.95	
	S2	13				
挂彩 (gua'cai)	S1	21	55	47	85.45	
	S2	22				
表现 (biao'xian)	S1	2	75	61	81.33	
	S2	37				
	S3	24				

3.2 Discussion

We have compared our results with results reported in the literature [Lu *et al*. 2002]; [Li 1999]. The results in [Lu *et al*. 2002] are listed in Table 5 (excerpted from reference [Lu *et al*. 2002] page 1086) along with the word senses of ambiguous words from the <Modern Chinese Lexicon>, 1996 edition. Figure1 shows the accuracy of 5 ambiguous words; the results [Li 1999] are listed in Table 6 (excerpted from [Li 1999] page 76), and the word sense of ambiguous words came from Cilin, a Chinese thesaurus.

Table 5. Results of experiments in reported in [Lu et al. 2002]

Polysemous words	Number of senses	Accuracy (%)
材料（cai'liao）	4	72.83
改(gai)	3	78.71
表现(biao'xian)	3	90.39
发表(fa'biao)	2	89.11
健康(jian'kang)	2	75.07

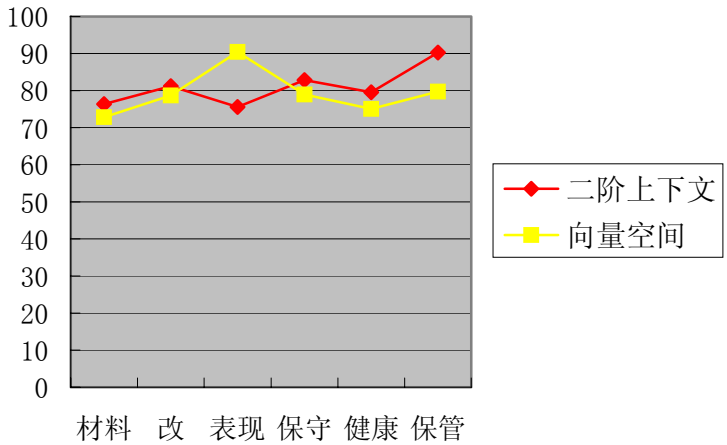

Figure1. Accuracy of 5 words based on Hownet

Table 6. Results of experiments reported in [Li 1999]

Polysemous words①	Sense ID②	Accuracy③ (%)
材料(cai'liao)	Dk17/ba06/al03	81.7
改(gai)	ih02/hg18/hj66	70.6
表现(biao'xian)	Jd06/di20/hj59	68.9
发表(fa'biao)	Hc11/hi14/jd03	73.4
健康(jian'kang)	ed43/eb37	70.1

①Ambiguous words, ②type of ambiguity, ③accuracy.

Notice: we did not select "Publish" as an ambiguous word in our experiment because several word senses of "Publish" in Hownet are very close to each other.

Compared with the results reported in [Lu *et al*. 2002], we obtained good results for some of the ambiguous words. What is more, for the selected 5 words, the average accuracy improved. Compared with the results reported in [Li 1999], the accuracy achieved was greatly improved.

4. Conclusion

Cognitive linguists believe that the similarity of contexts has a great influence on the process of differentiating diverse semantemes [Miller and Charles 1991]. This paper adopted this assumption that the similarity of contexts determines the similarity of semantemes of the words. The experiments proved that our approach is feasible. However, there are still some shortcomings:

(1) The selection of features. The features that we select must provide as much information for word sense disambiguation as possible and, at the same time, should contain as little noise as possible. This is important for clustering.

(2) The accuracy of clustering. The clustering of k-means clustering has some drawbacks in that it is sensitive to the center of the cluster and it is difficult to get a superior overall solution of overall situation. Thus the accuracy we achieved was influenced by the accuracy of clustering.

(3) Expanding the scope of ambiguous words. It is difficulty to evaluate the overall results of our approach, so many people are concerned about the issue. The applicability of our approach should be further tested in large-scale experiments and with other languages.

References

Dong, Z. D., and Q. Dong, "Hownet," http://keenage.com, 2000.

Li, F., B. Xue, and Y. L. Huang, "A Novel K-means Clustering Based on Initial Central Superior," *Computer Science*, 29(7), 2002, pp.94-96.

Li, J. Z., "The Research on Chinese Word Sense Disambiguation," Ph.D.Thesis, Beijing: Tsinghua University, 1999.

Lu, S., S. Bai, and X. Huang, "An Unsupervised Approach to Word Sense Disambiguation Based on Sense-Words in Vector Space Model," *Journal of Software*, 13(6), 2002, pp.1082-1089.

MacQueen, J., "Some methods for classification and analysis of multivariate observations," In *proceedings of 5th Symp. Math. Statist, Prob*, 1967, Berkeley, pp.218-297.

Miller, G. A., and W. Charles, "Contextual Correlates of Semantic Similarity " *Language and Cognitive Processes*, 6(1), 1991, pp.1-28.

Pantel, P., and D.K. Lin, “A Statistical Corpus-Based Term Extractor,”In *proceedings of AI*, 2001, Canadian, pp.36-46.