Chinese Word Sense Disambiguation with PageRank and HowNet

Jinghua Wang
Beijing University of Posts and Telecommunications
Beijing, China
wjh_smile@163.com

Jianyi Liu
Beijing University of Posts and Telecommunications
Beijing, China
jianyilui@sohu.com

Ping Zhang
Shenyang Normal University
Shenyang, China
pinney58@163.com

Abstract

Word sense disambiguation is a basic problem in natural language processing. This paper proposed an unsupervised word sense disambiguation method based PageRank and HowNet. In the method, a free text is firstly represented as a sememe graph with sememes as vertices and relatedness of sememes as weighted edges based on HowNet. Then UW-PageRank is applied on the sememe graph to score the importance of sememes. Score of each definition of one word can be computed from the score of sememes it contains. Finally, the highest scored definition is assigned to the word. This approach is tested on SENSEVAL-3 and the experimental results prove practical and effective.

1 Introduction

Word sense disambiguation, whose purpose is to identify the correct sense of a word in context, is one of the most important problems in natural language processing. There are two different approaches: knowledge-based and corpus-based (Montoyo, 2005). Knowledge-based method disambiguates words by matching context with information from a prescribed knowledge source, such as WordNet and HowNet. Corpus-based methods are also divided into two kinds: unsupervised and supervised (Lu Z, 2007). Unsupervised methods cluster words into some sets which indicate the same meaning, but they can not give an exact explanation. Supervised machine-learning method learns from annotated sense examples. Though corpus-based approach usually has better performance, the mount of words it can disambiguate essentially relies on the size of training corpus, while knowledge-based approach has the advantage of providing larger coverage. Knowledge-based methods for word sense disambiguation are usually applicable to all words in the text, while corpus-based techniques usually target only few selected word for which large corpora are made available (Mihalcea, 2004).

This paper presents an unsupervised word sense disambiguation algorithm based on HowNet. Words' definition in HowNet is composed of some sememes which are the smallest, unambiguous sense unit. First, a free text is represented as a sememe graph, in which sememes are defined as vertices and relatedness of sememes are defined as weighted edges. Then UW-PageRank is applied on this graph to score the importance of sememes. Score of each definition of one word can be deduced from the score of sememes it contains. Finally, the highest scored definition is assigned to the word. This algorithm needs no corpus, and is able to disambiguate all the words in the text at one time. The experiment result shows that our algorithm is effective and practical.

2 HowNet

HowNet (Dong, Z. D, 2000) is not only a machine readable dictionary, but also a knowledge base which organizes words or concepts as they represent in the object world. It has been widely used in word sense disambiguation and pruning, text categorization, text clustering, text retrieval, machine translation, etc (Dong, Z. D, 2007).

2.1 The content and structure of HowNet

HowNet is an online common-sense knowledge based unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and English equivalents. There are over 16000 word records in the dictionary. This is an example

No.=017625	No.=017630
W_C=打	W_C=打
G_C=V	G_C=V
E_C=打鼓	E_C=打酱油
W_E=hit	W_E=buy
G_E=V	G_E=V
DEF=beat|打	DEF=buy|买，commercial|商

This is two of the concepts of word "打": "No." is the entry number of the concept in the dictionary; "G_C" is the part of speech of this concept in Chinese, and "G_E" is that in English; "E_C" is the example of the concept; "W_E" is the concept in English; "DEF" is the definition.

Definitions of words are composed of a series of sememes (usually more than one，like DEF No.017630 contains "buy|买" and "commercial|商")，like "beat|打" which is the smallest unambiguous unit of concept. First sememe of the definition like "buy|买" of DEF No.017630 is the main attribution of the definition. Sememes have been classified into 8 categories, such as attribute, entity, event role and feature, event, quantity value, quantity, secondary feature and syntax. Sememes in one category form a tree structure with hypernymy / hyponymy relation. Sememes construct concepts, e.g. definition, so the word sense disambiguation task of assigning definition to word can be done through the computation of sememes.

2.2 The similarity of sememes

The tree structure of sememes makes it possible to judge the relatedness of them with a precision mathematical method. Rada (Rada, R, 1989) defined the conceptual distance between any two concepts as the shortest path through a semantic network. The shortest path is the one which includes the fewest number of intermediate concepts. With Rada's algorithm, the more similar two concepts are, the smaller their shortest path is, and so we use the reciprocal of the length of shortest path as the similarity. Leacock and Chodorow (Leacock, C, 1998) define it as follows:

$$sim_{lch}(c_1, c_2) = \max[-\log(length(c_1, c_2)/(2D))$$

where *length(c1, c2)* is the shortest path length between the two concepts and D is the maximum depth of the taxonomy.

Wu and Palmer (Wu, Z., 1994) define another formula to measure the similarity

$$sim_{wup}(c_1, c_2) = \frac{2 \cdot depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

depth is the distance from the concept node to the root of the hierarchy. *lcs(c1,c2)* is the most specific concept that two concepts have in common, that is the lowest common subsumer.

3 PageRank on Sememe Graph

PageRank is an algorithm of deciding the importance of vertices in a graph. Sememes from HowNet can be viewed as an undirected weighted graph, which defines sememes as vertices, relations of sememes as edges and the relatedness of connected sememes as the weights of edges. Because PageRank formula is defined for directed graph, a modified PageRank formula is applied to use on the undirected weighted graph from HowNet.

3.1 PageRank

PageRank (Page, L., 1998) which is widely used by search engines for ranking web pages based on the importance of the pages on the web is an algorithm essentially for deciding the importance of vertices within a graph. The main idea is that: in a directed graph, when one vertex links to another one, it is casting a vote for that other vertex. The more votes one vertex gets, the more important this vertex is. PageRank also takes account the voter: the more important the voter is, the more important the vote itself is. In one word, the score associated with a vertex is determined based on the votes that are cast for it, and the score of the vertex casting these votes. So this is the definition:

Let G=(V,E) be a directed graph with the set of vertices V and set of edges E, when E is a subset of V×V. For a given vertex V_i, let $In(V_i)$ be the set of vertices that point to it, and let $Out(V_i)$ be the set of edges going out of vertex V_i. The PageRank score of vertex V_i is

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{S(V_j)}{\left|Out(V_j)\right|}$$

d is a damping factor that can be set between 0 and 1,and usually set at 0.85 which is the value we use in this paper (Mihalcea, R., 2004).

PageRank starts from arbitrary values assigned to each vertex in the graph, and ends when the convergence below a given threshold is achieved. Experiments proved that it usually stops computing within 30 iterations (Mihalcea, R., 2004).

PageRank can be also applied on undirected graph, in which case the out-degree of a vertex is equal to the in-degree of the vertex.

3.2 PageRank on sememe graph

Sememes from HowNet can be organized in a graph, in which sememes are defined as vertices, and similarity of connected sememes are defined as weight of edges. The graph can be constructed as an undirected weighted graph.

We applied PageRank on the graph with a modified formula

$$S(V_i) = (1-d) + d * \sum_{j \in C(V_i)} \frac{weight(E_{ij}) \cdot S(V_j)}{\left|D(V_j)\right|}$$

C(Vi)is the set of edges connecting with V_j, weight(E_{ij})is the weight of edge E_{ij} connecting vertex V_i and V_j, and $D(V_j)$ is the degree of V_j. This formula is named UW-PageRank. In sememe graph, we define sememes as vertices, relations of sememes as edges and the relatedness of connected sememes as the weights of edges. UW-PageRank is applied on this graph to measure the importance of the sememes. The higher score one sememe gets, the more important it is.

4 Word sense disambiguation based on PageRank

To disambiguate words in the text, firstly the text is converted to an undirected weighted sememe graph based on HowNet. The sememes which are from all the definitions for all the words in the text form the vertices of the graph and they are connected by edges whose weight is the similarity of the two sememes. Then, we use UW-PageRank to measure the importance of the vertex in the graph, so all the sememes are scored. So each definition of one word can be scored based on the score of the sememes it contains. Finally, the highest scored definition is assigned to the word as its meaning.

4.1 Text representation as a graph

To use PageRank algorithm to do disambiguation, a graph which represents the text and interconnects the words with meaningful relations should be built first. All the words in the text should be POS tagged first, and then find all the definitions pertaining to the word in HowNet with its POS. Different sememes from these definitions form the vertices of the graph. Edges are added between the vertices whose weights are the similarity of the sememes. The similarity can be measured by the algorithm in Section 2.2. As mentioned in Section 2.1, all the sememes in HowNet are divided into eight categories, and in each category, sememes are connected in a tree structure. So based on the algorithms in Section 2.2, each two sememes in one category, i.e. in one tree, have a similarity more than 0, but if they are in different category, they will have a similarity equal to 0. As a result, a text will be represented in a sememe graph that is composed of several small separate fully connected graphs.

Assumed that a text containing "word1 word2 word3" is to be represented in a graph. The definition (DEF) and sememes for each word are listed in Table 1.

Table 1. "*Word1 Word2Word3*"

Word	Definition	Sememes
Word1	DEF11	S1,S5
	DEF12	S2
	DEF13	S8
Word2	DEF21	S6
	DEF22	S7,S9
Word3	DEF31	S3
	DEF32	S4

Sememes are linked together with the weight of relatedness. For example, S1 and S2 are connected with an edge weighted 0.3.The relation of word, DEF and sememes is represented in Figure1, and sememe graph is in Figure 2.

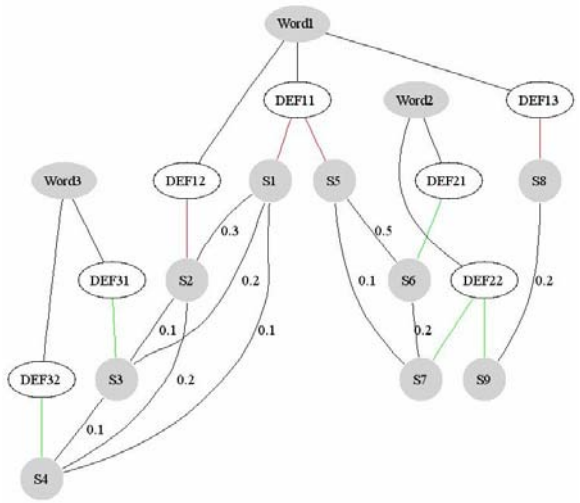

Figure 1. Word-DEF-Sememe Relation

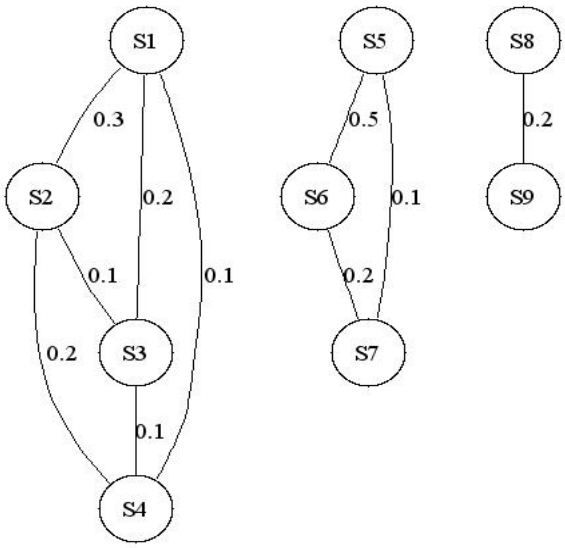

Figure 2. Sememe Graph

4.2 Word sense disambiguation based on PageRank

Text has been represented in a sememe graph with sememes as vertices and similarity of sememes as the weight of the edges. Then, UW-PageRank is used to measure the importance of the vertex, i.e. sememes in the graph. The score of all the vertices in Figure 1 is in Table 2.

Table 2. Score of Sememes

Vertex	S1	S2	S3	S4	S5
UW-PageRank Score	0.179	0.175	0.170	0.165	0.202
Vertex	S6	S7	S8	S9	
UW-PageRank Score	0.208	0.176	0.181	0.181	

Each definition of the words is scored based on the score of the sememes it contains.

$$Sense(Word) = \arg\max_{1 \le i \le m}(Score(DEF_i))$$

$DEF_i \in Word$, DEF_i is the i sense of the word.

We use two methods to score the definition:

Mean method

HowNet uses sememes to construct definitions, so the score of the definition can be measured through an average score of all the sememes it contains. And we chose the definition of the highest score as the result.

$$Score(DEF) = \frac{1}{n}\sum_{1 \le i \le n} Score(S_i)$$

$S_i \in DEF$, S_i is the i sememe of DEF.

First Sememe method

First sememe of one DEF is defined as the most important meaning of the DEF. So we use another method to assign one DEF to one word taking first sememe into consideration. For all the DEF of one word, if one first sememe of one DEF gets the highest score, the DEF is assigned to the word.

$$Score(DEF) = Score(FirstSememe)$$

If several DEFs have the same first sememe or have the same score, we sort all the other sememes are from high score to low score, then comparison is made among this sememes from the beginning to the end until one of the sememes has the highest score among them, and finally the DEF containing this sememe is assigned to the word.

The performance of the two methods will be tested and compared in Section5.

With the "Means" (M) and "First Sememe" (FS) methods, text in Section 4.1 gets the result in Table 3.

Table3. Result of "*Word1 Word2 Word3*"

Word	Definition	Score (M)	Result(M)	Result(FS)
Word1	DEF11	0.191	DEF11	DEF13
	DEF12	0.175		
	DEF13	0.181		
Word2	DEF21	0.208	DEF21	DEF21
	DEF22	0.179		
Word3	DEF31	0.170	DEF31	DEF31
	DEF32	0.165		

Table 4. Experimental Result

Word	Baseline	R+M	L +M	W+M	R+FS	L +FS	W+FS	Li
把握	0.25	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	**0.53**	0.32
材料	0.33	**0.6**	0.5	**0.6**	0.4	0.4	0.3	0.74
老	0.1	**0.42**	**0.42**	0.46	0.35	0.35	0.38	0.26
没有	0.25	0.73	**0.75**	0.56	0.67	**0.75**	0.56	0.39
突出	0.17	0.5	0.57	0.64	0.43	0.5	0.64	**0.67**
研究	0.33	**0.47**	0.27	0.13	**0.47**	0.27	0.13	0.27
Average Precision	0.24	**0.54**	0.51	0.49	0.48	0.47	0.42	0.44

5 Experiment and evaluation

We chose 96 instances of six words from SENSEVAl-3 Chinese corpus as the test corpus. Words are POS tagged. We use precision as the measure of performance and random tagging as the baseline. We crossly use Rada's (R), Leacock & Chodorowp's (L), and Wu and Palmer's (W) methods to measure the similarity of sememes with mean method (M) and first sememe (FS) scoring the DEF. The precision of the combination algorithm is listed in Table 4.

Li (Li W., 2005) used naive bayes classifier with features extracted from People's Daily News to do word sense disambiguation on SENSEVAL-3. The precision is listed in line "Li" of table as a comparison.

The average precision of our algorithm is around two times higher than the baseline, and 5 of the 6 combination algorithm gets better performance than Li. And for 5/6 word case, our algorithm gets the best performance. Among the three methods of measure the similarity of sememes, Rada's method gets the best performance. And between the two methods of scoring definition, "Mean method" works better, which indicates that although the first sememe is the most important meaning of one definition, the other sememes are also very important, and the importance of other sememes also should be taken into consideration while scoring the definition. Of all the combination of algorithms, "Rada + Mean" gets the best performance, which takes a reasonable way to measure the similarity of two sememes and comprehensively scores the definition based on the importance of its sememes in the sememe graph from the whole text.

6 Related works

Many works in Chinese word sense disambiguation with HowNet. Chen Hao (Chen Hao, 2005) brought up a k-means cluster method base on HowNet, which firstly convert contexts that include ambiguous words into context vectors; then, the definitions of ambiguous words in Hownet can be determined by calculating the similarity between these context vectors. To do disambiguation, Yan Rong (Yan Rong, 2006) first extracted some most relative words from the text based on the co-occurrence, then calculate the similarity between each definition of ambiguous word and its relative words, and finally find the most similar definition as its meaning. The similarity of definitions is measured by the weighted mean of the similarity of sememes, and the similarity of sememes is measured by a modified Rada's formula. Gong YongEn (Gong YongEn, 2006) used a similar method with Yan, and more over, he took recurrence of sememes into consideration. Compare with those methods, our method has a more precious sememes' similarity measure method, and make full use of the structure of its tree structure by representing text in graph and use UW-PageRank to judge sememes' importance in the whole text, that is the most obvious difference from them. Mihalceal (Mihalceal, 2004) first provide the semantic graph method to do word sense disambiguation, but her work is totally on English with WordNet, which is definitely different in meaning representation from HowNet. WordNet uses synsets to group similar concepts together and differentiate them, while HowNet use a close set of sememes to construct concept definitions. In Mihalceal's method, the

vertexes of graph are synsets, and in ours are sememes. And after measure the importance of sememes, an additional strategy is used to judge the score of definition based on the sememes.

7 Conclusion

An unsupervised method is applied to word sense disambiguation based on HowNet. First, a free text is represented as a sememe graph with sememes as vertices and relatedness of sememes as weighted edges. Then UW-PageRank is applied on this graph to score the importance of sememes. Score of each definition of one word can be deduced from the score of sememes it contains. Finally, the highest scored definition is assigned to the word. Our algorithm is tested on SENSEVAL-3 and the experimental results prove our algorithm to be practical and effective.

Acknowledgment

This study is supported by Beijing Natural Science Foundation of (4073037) and Ministry of Education Doctor Foundation (20060013007).

References

Chen hao, He Tingting, Ji Donghong, Quan Changqing, 2005. *An Unsupervised Approach to Chinese Word Sense Disambiguation Based on Hownet*, Computational Linguistics and Chinese Language Processing, Vol. 10, No. 4, pp. 473-482

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

Dong Zhendong, Dong Qiang, Hao Changling, 2007. *Theoretical Findings of HowNet*, Journal of Chinese Information Processing, Vol. 21, No. 4, P3-9

Gong Y., Yuan C., Wu G., 2006. *Word Sense Disambiguation Algorithm Based on Semantic Information*, Application Research of Computers, 41-43.

Leacock, C., Chodorow, M., 1998.*Combing local context and WordNet Similarity for word sense identification*, in: C.Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, 305-332

Li W., Lu Q., Li W., 2005. *Integrating Collocation Features in Chinese Word Sense Disambiguation*, Integrating Collocation Features in Chinese Word Sense Disambiguation. In Proceedings of the Fourth Sighan Workshop on Chinese Language Processing, 87-94.

Lu Z., Liu T., Li, S., 2007. *Chinese word sense disambiguation based on extension theory*, Journal of Harbin Institute of Technology, Vol.38 No.12, 2026-2035

Mihalcea, R., Tarau, P., Figa, E., 2004. *PageRank on Semantic Networks, with application to Word Sense Disambiguation*, in Proceedings of The 20st International Conference on Computational Linguistics

Montoyo, A., Suarez, A., Rigau, G. and Palomar, M. 2005. Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods, Volume 23, Journal of Machine learning research , 299-330.

Page, L., Brin, S., Motwani, R., and wingorad, T., 1998. *The pagerank citation ranking: Bringing order to the web Technical report*, Stanford Digital Library Technologies Project.

Rada, R., Mili,E,.Bicknell, Blettner, M., 1989. *Development and application of a metric on semantic nets*, IEEE Transactions on Systems, Man and Cybernetics 19(1) 17-30

Wu, Z., Plamer, M., 1994. *Verb semantics and lexical selection*, in 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, 133-138

Yan R., Zhang L., 2006. *New Chinese Word Sense Disambiguation Method*, Computer Technology and Development, Vol. 16 No.3, 22-25