International Journal of Hybrid Information Technology
Vol.8, No.4 (2015), pp.105-116
http://dx.doi.org/10.14257/ijhit.2015.8.4.13

Integrate Chinese Semantic Knowledge into Word Sense Disambiguation

Zhang Chun-Xiang[1, 2], Sun Lu-Rong[3], Gao Xue-Yao[3], Lu Zhi-Mao[4], and Yue Yong[5]

[1]*School of Software, Harbin University of Science and Technology, Harbin 150080, China*
[2]*College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China*
[3]*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*
[4]*School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China*
[5]*Department of Computer Science and Technology, University of Bedfordshire, Park Square, Luton LU1 3JU, United Kingdom*
z6c6x666@163.com

Abstract

Word sense disambiguation is important for many applications in natural language processing fields including machine translation, information retrieval and automatic summarization. In this paper, left word unit and right word unit are extracted for improving the quality of word sense disambiguation (WSD) starting from the target polysemous word. Their semantic knowledge is mined from Tongyici Cilin which is a Chinese semantic lexicon. A new method of word sense disambiguation is proposed with semantic information of left word unit and right word unit. The classifier of word sense disambiguation is built based on bayesian model. SemEval-2007: Task#5 is used as training corpus and test corpus. Experimental results show that the disambiguation classifier's precision is improved and demonstrate the effectiveness of the method.

Keywords: *Word sense disambiguation; Polysemous word; Semantic lexicon; Semantic information; Disambiguation classifier*

1. Introduction

The purpose of word sense disambiguation is to determine the correct sense of an ambiguous word in a specific context. It is an important research topic in natural language processing fields. The precision of word sense disambiguation has a great impact on machine translation, information retrieval, text analysis, automatic summarization and other related applications. WSD can be divided into supervised method, unsupervised method and semi-supervised method.

Wang uses semantic diffusion kernels to smooth BoW representation in WSD systems. These kernels model semantic similarities by means of a diffusion process on a graph which is defined by lexicon and co-occurrence information. Kernels are obtained based on a matrix exponentiation transformation on the given kernel matrix, and they apply higher order co-occurrences to get semantic similarities between terms [1]. Bordes gives a neural network architecture for embeding multi-relational graphs into a flexible continuous vector space and encoding semantics of these graphs in order to assign high probabilities to plausible components. It is applied to word sense disambiguation in a context of open-text semantic parsing. At the same time, it assigns a structured meaning representation to

ISSN: 1738-9968 IJHIT
Copyright ⓒ 2015 SERSC

a sentence [2]. Navigli gives a graph-based WSD algorithm in which few parameters are provided and sense-annotated data are not needed. At the same time, he uses this algorithm to identify several measures of graph connectivity best suited for WSD [3]. Yang proposes a novel model based on distance between words for WSD. It is built on graph-based WSD models and makes full use of distance information between words [4]. Fan selects features based on information gain for WSD. She mines location information in contexts of ambiguous words according to information gain. The purpose is to improve the efficiency of knowledge acquisition and the quality of word sense classifier [5]. Lu gives a supervised WSD method which formalizes senses of a polysemous word with interesting term weight based on vector space model. At the same time, k-nearest neighbor algorithm is used to deal with WSD [6]. Guo combines evidence from a monolingual WSD system with that from a multilingual WSD system. In this monolingual system, a graph-based in-degree algorithm is used. In this multilingual system, an all-words unsupervised approach is adopted [7]. Faralli presents a minimally-supervised framework for performing domain-driven word sense disambiguation. A bootstrapping method is used to get glossaries for several domains from webs. Then, these glosses are used as sense inventories for fully unsupervised domain-driven WSD [8]. Navigli gives a multilingual joint WSD approach and uses a large multilingual knowledge base BabelNet to perform the graph-based WSD across different languages. The wide-coverage multilingual lexical knowledge and robust graph-based algorithms are adopted. At the same time, several different methods are combined to solve WSD task [9]. Ponzetto presents a method to extend WordNet automatically with a large amount of semantic relations from Wikipedia. These high-quality semantic relations are provided for disambiguation algorithms which are short of knowledge. Experiments show that their performances outperform state-of-the-art supervised WSD systems [10]. Schwab applies three unsupervised stochastic algorithms to word sense disambiguation including genetic algorithms, simulated annealing algorithms and ant colony algorithms. At the same time, comparative experiments are conducted to evaluate these 3 algorithms' performances [11]. Huang gives a position-based algorithm in order to measure the context similarity, in which contextual words are assigned with positional weights. The correct sense of an ambiguous word is determined based on the context similarity between a new instance and pre-labeled instances. Senseval-2 English lexical samples are used as test corpus. Experimental results show that the proposed method achieves good performances [12]. Niu presents a new method to partition the mixed data including labeled data and unlabeled data. The principle is to maximize a stability criterion defined on classification results from an extended label propagation algorithm over all possible values of model order in mixed data. When the model order identification algorithm and the extended label propagation algorithm are combined as WSD classifier, its performance outperforms SVM [13]. Le uses unlabeled data for WSD within a semi-supervised learning framework. He solves 3 problems with the help of classifier combination strategies, including the imbalance of training data, the confidence of new labeled examples and the final classifier generation. Experiments show that the proposed solution improves the quality of supervised WSD methods [14]. Huang gives a novel algorithm of word sense disambiguation in which semi-supervised statistical learning methods are used. He uses small-scale labeled data to build an initial classifier with a certain accuracy rate and extends training data with a variety of thresholds. Experimental results show the proposed method has a higher performance [15]. Le gives a framework for weighted combination of WSD classifiers based on dempster-shafer theory of evidence and the ordered weighted averaging operators. He finds some features which provide complementary linguistic information for contexts, and combines these information sources based on dempster's rule of combination and owa operators [16].

 Copyright ⓒ 2015 SERSC

In this paper, Tongyici Cilin is applied to extract semantic codes of left word and right word around an ambiguous word. The extracted semantic codes are used as discriminative features and the bayesian model is applied to decide correct senses of polysemous words.

2. Extracting Discriminative Features for WSD

Discriminative features in texts can be expressed by word units in a certain language environment and reflect co-occurrences of linguistic information between word units. Linguistic information often includes word, part-of-speech, location, length, syntactic category and semantic category. Linguistic information usually lies on the surface layer of a sentence. After a sentence is segmented and every word is analyzed with semantic lexicon, discriminative features can be gotten. Now, precisions and performances of these analysis tools are good, so that we can get discriminative features with a certain disambiguation capabilities.

Semantic category of an ambiguous word is determined by its context. The context provides discriminative information for WSD. The ambiguous word is viewed as center and a word window is opened to extract the contextual information. When the size of the window is larger, it contains more discriminative information. However, it is difficult to get a large labeled corpus for WSD in reality. If the size of the window is too large, it will cause data sparseness in process of training WSD classifier. In this paper, semantic categories of left word unit and right word unit around an ambiguous word are only used as discriminative features to determine its correct sense. For Chinese sentence *C* including ambiguous word *w*, the algorithm of extracting its discriminative features is shown as follows:

(1)Use word segmentation tool to segment *C* and get Chinese words.

(2)Use part-of-speech tagging tool to mark Chinese words and their part-of-speech tags are obtained.

(3)Locate ambiguous word *w* in Chinese sentence *C*.

(4)Ambiguous word *w* is viewed as center. Its left and right word units are gotten.

(5)Look up Tongyici Cilin to determine semantic codes of left word and right word.

For Chinese sentence containing ambiguous word ‘wang4’, the process of extracting discriminative features is shown as follows:

Chinese sentence: zhan4 de1 di1 le1 ye3 bu4 xing2 , deng1 gao1 cai2 neng2 wang4 yuan3 .

Word segmentation: zhan4/ de1/ di1/ le1/ ye3/ bu4 xing2/ ,/ deng1 gao1/ cai2/ neng2/ wang4/ yuan3/ ./

Part-of-speech tagging: zhan4/v de1/u di1/a le1/u ye3/d bu4 xing2/a ,/w deng1 gao1/v cai2/d neng2/v wang4/v yuan3/a ./w

Semantic lexicon gives semantic categories of words and provides rich semantic knowledge for word sense disambiguation. Tongyici Cilin is a semantic category dictionary in Chinese and provides semantic codes for words. A semantic code has 3 layers. In Tongyici Cilin, there are 12 big categories, 94 small categories and 1428 subcategories. It describes a semantic classification architecture from top to bottom and from generality to specificity. All Chinese words are collected and organized according to their categories in Tongyici Cilin. In every word’s entry, semantic code is used to represent its semantic category. For Chinese word ‘wang4’, its semantic code is Dk15. It shows that Chinese word ‘wang4’ is located in big category D, small category k and subcategory 15. This semantic classification architecture can

Copyright ⓒ 2015 SERSC

be viewed as a tree. The root node' sons are all big categories. Sons of big categories are small categories. Sons of small categories are subcategories. In this paper, Tongyici Cilin is applied to determine semantic codes of left word, right word and an ambiguous word.

In Chinese, 'wang4' is an ambiguous word. It has five kinds of different senses in Tongyici Cilin. Table 1 lists its semantic codes and correspondent Chinese synonyms. The first semantic code is Da26. Its Chinese synonyms are 'ming2 sheng1', 'qing2 mian4' and 'mian4 zi1'. The second semantic code is Df08. Its Chinese synonyms are respectively 'zhi3 wang4', 'yuan4 wang4' and 'yi4 tu2'. The third semantic code is Kb01. Its Chinese synonyms are 'xiang4', 'bei4' and 'zai4'. The fourth semantic code is Fc04. Its Chinese synonym is 'kan4'. The fifth semantic code is Gb04. Its Chinese synonyms are respectively 'xi1 wang4', 'yao1 qiu2' and 'zhui1 qiu2'. The sixth semantic code is Ca23. Its Chinese synonym is 'ri4'. The seventh semantic code is Hi02. Its Chinese synonyms are respectively 'fang3 wen4', 'jin4 jian4' and 'tan4 wang4'.

Table 1. Seven Semantic Codes and Chinese Synonyms of Word 'wang4'

Semantic code	Chinese synonyms
Da26	'ming2 sheng1', 'qing2 mian4', 'mian4 zi1'
Df08	'zhi3 wang4', 'yuan4 wang4', 'yi4 tu2'
Kb01	'xiang4', 'bei4', 'zai4'
Fc04	'kan4'
Gb04	'xi1 wang4', 'yao1 qiu2', 'zhui1 qiu2'
Ca23	'ri4'
Hi02	'fang3 wen4', 'jin4 jian4', 'tan4 wang4'

The semantic tree of Chinese word 'wang4' is shown in Figure 1.

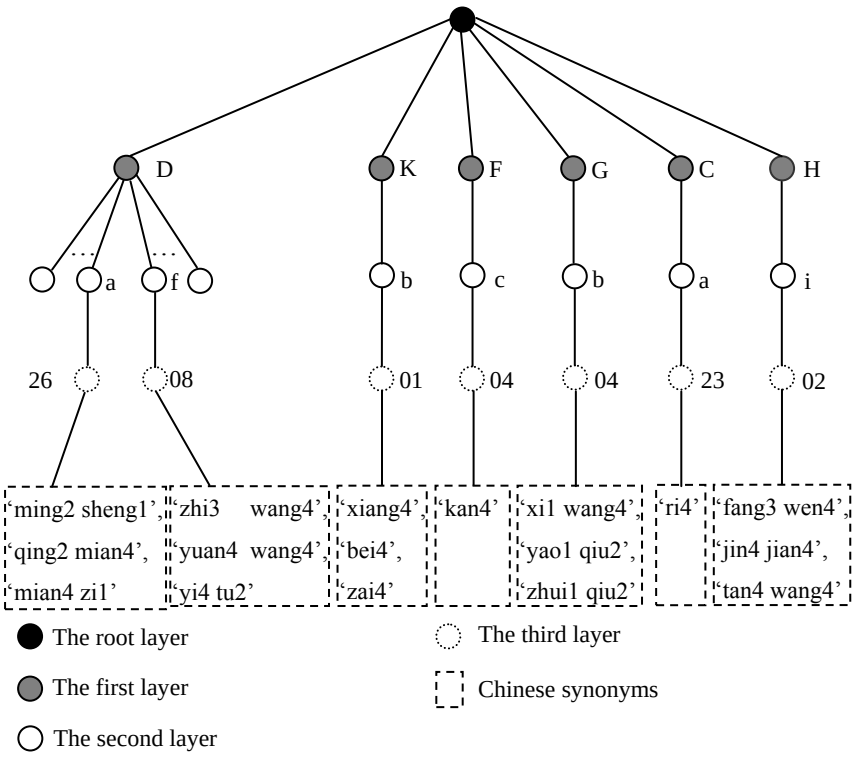

Figure 1. Semantic Tree of Chinese Word 'wang4'

 Copyright ⓒ 2015 SERSC

In the above example, ‘wang4’ is an ambiguous word. According to contexts, we can infer that its semantic code is Fc04 and its meaning is ‘kan4’ in the above example.

Left word of ‘wang4’ is Chinese word ‘neng2’. In Tongyici Cilin, ‘neng2’ is also an ambiguous word. It has 4 different meanings. The first semantic code is Ee17. Its synonyms are ‘neng2 gan4’ and ‘wu2 neng2’. The second semantic code is Dd14. Its synonyms are respectively ‘li4 qi4’, ‘li4 liang4’ and ‘neng2 liang4’. The third semantic code is De04. Its synonyms are ‘zhi4 hui4’, ‘cai2 neng2’, ‘neng2 li4’ and ‘gong1 fu1’. The fourth semantic code is Gc02. Its synonyms are respectively ‘neng2’, ‘neng2 gou4’ and ‘bu4 neng2’. It means that the discriminative context is also ambiguous and can not provide any guidance for word sense disambiguation. When the context is unambiguous, WSD classifier can decide correct senses of ambiguous words. Here, dice coefficient is used to determine semantic codes of left word and right word.

Dice coefficient is a measurement function of collection similarity. It is used for comparing the similarity of two samples. Dice coefficient is used to determine semantic code S_w of Chinese ambiguous word w in Tongyici Cilin, which is as shown in formula (1).

$$S_w = \arg\max_{s \in SenSet(w)} SenSim(w, w_s) \tag{1}$$

$$SenSim(w, w_s) = \frac{sim(w, w_s)}{length(w) + length(w_s)} \tag{2}$$

Here, $SenSet(w)$ is a set which contains semantic codes of ambiguous word w. For example, $SenSet$(‘neng2’)={Ee17, Dd14, De04, Gc02}. Here, w_s is synonymous with word w under semantic code S. For example, w_{Ee17}=‘neng2 gan4’_‘wu2 neng2’. The value of $sim(w, w_s)$ is the number of Chinese characters shared together by w and w_s. For example, sim(‘neng2’, ‘neng2 gan4’_‘wu2 neng2’)=3. There are three same Chinese characters ‘neng2’ in w and w_s. Here, $length(X)$ is the number of Chinese characters in string X. For example, $length$(‘neng2’)=1, $length$(‘neng2 gan4’_‘wu2 neng2’)=4. The similarity of semantic code Ee17 is calculated as shown in formula (3).

$$\begin{aligned} &SenSim('neng2','neng2\ gan4'\_'wu2\ neng2') \\ &= \frac{sim('neng2','neng2\ gan4'\_'wu2\ neng2')}{length('neng2') + length('neng2\ gan4'\_'wu2\ neng2')} = \frac{3}{1+4} = 0.6 \end{aligned} \tag{3}$$

For every semantic code of word ‘neng2’, the above method is used to calculate its similarity. The results are shown in Table 2.

Table 2. Compute Similarities of All Semantic Codes for Word ‘neng2’

Semantic code	Chinese synonyms	Similarity
Ee17	‘neng2 gan4’, ‘wu2 neng2’	0.6
Dd14	‘li4 qi4’, ‘li4 liang4’, ‘neng2 liang4’	0.28
De04	‘zhi4 hui4’, ‘cai2 neng2’, ‘neng2 li4’, ‘gong1 fu1’	0.33
Gc02	‘neng2’, ‘neng2 gou4’, ‘bu4 neng2’	0.67

According to equation (1), we can decide that the semantic code of Chinese word ‘neng2’ is Gc02.

Right word of ‘wang4’ is Chinese word ‘yuan3’. In Tongyici Cilin, ‘yuan3’ is also an ambiguous word. It has 2 different meanings. The first semantic code is Eb21. Its synonyms are ‘yuan3’ and ‘jin4’. The second semantic code is Ed32. Its synonyms are respectively ‘qin1 mi4’ and ‘shu1 yuan3’. Equation (2) is used to compute the similarity of semantic code Eb21. Here, $SenSet$(‘yuan3’)={Eb21, Ed32}. w_{Eb21}=‘yuan3’_‘jin4’.

Copyright ⓒ 2015 SERSC

sim(‘yuan3’, ‘yuan3’_‘jin4’)=2. *length*(‘yuan3’)=1. *length*(‘yuan3’_‘jin4’)=2. The similarity of semantic code Eb21 is calculated as shown in formula (4).

$$\begin{aligned} &SenSim('\text{yuan3}','\text{yuan3}'\_'\text{jin4}') \\ &= \frac{sim('\text{yuan3}','\text{yuan3}'\_'\text{jin4}')}{length('\text{yuan3}')+length('\text{yuan3}'\_'\text{jin4}')} = \frac{2}{1+2} = 0.67 \end{aligned} \tag{4}$$

For every semantic code of word ‘yuan3’, the above method is used to calculate its similarity. The results are shown in Table 3.

Table 3. Compute Similarities of All Semantic Codes for Word ‘yuan3’

Semantic code	Chinese synonyms	Similarity
Eb21	‘yuan3’, ‘jin4’	0.67
Ed32	‘qin1 mi4’, ‘shu1 yuan3’	0.4

According to equation (1), we can decide that the semantic code of Chinese word ‘yuan3’ is Eb21.

Discriminative features Gc02 and Eb21 are applied to determine the correct meaning of Chinese ambiguous word ‘wang4’.

3. Bayesian Classifier based on Semantic Knowledge

Bayesian model infers the current occurrence probability of an incident based on its past occurrence probability. Here, bayesian decision rule is applied to word sense disambiguation based on semantic codes of left and right words. For ambiguity word *w*, the process of determining its correct meaning is described in formula (5).

$$S = \arg\max_{i=1,2,\ldots,m} P(S_i \mid context) \tag{5}$$

Here, ambiguity word *w* has *m* meanings and their semantic codes include S_1, S_2, …, S_m. Word *w* is located in *context*. P(*X*) is the probability that *X* appears. P(S_i|*context*) is the probability that semantic code of *w* is S_i under *context*. The value of *i* is from 1 to *n*. When the value of P(S_i|*context*) is maximum, semantic code of word *w* is S_i. Bayesian rule guarantees that the error probability of decision is the smallest. Here, *context* is comprised of left word unit’s semantic code s_code_L and right word unit’s semantic code s_code_R. The process of determining *w*’s correct sense is shown in formula (6).

$$\begin{aligned} S &= \arg\max_{i=1,2,\ldots,m} P(S_i \mid s\_code_L, s\_code_R) \\ &= \arg\max_{i=1,2,\ldots,m} \frac{P(S_i, s\_code_L, s\_code_R)}{P(s\_code_L, s\_code_R)} \\ &\approx \arg\max_{i=1,2,\ldots,m} P(S_i, s\_code_L, s\_code_R) \\ &= \arg\max_{i=1,2,\ldots,m} P(s\_code_L, s\_code_R \mid S_i) P(S_i) \\ &\approx \arg\max_{i=1,2,\ldots,m} P(s\_code_L \mid S_i) P(s\_code_R \mid S_i) P(S_i) \end{aligned} \tag{6}$$

In Tongyici Cilin, a semantic code is divided into three layers. Semantic code s_code_L and s_code_R can all be denoted as $sc_1sc_2sc_3$. Capital English letters are used to

 Copyright ⓒ 2015 SERSC

express code sc_1 in the first layer. Its range is from A to L. Lowercase English letters are applied to represent code sc_2 in the second layer. Two digits are used to express code sc_3 in the third layer. For example, Fc04 is a semantic code of ambiguous word 'wang4'. Here, sc_1=F, sc_2=c and sc_3=04. Parameters $P(s\_code_L|S_i)$ and $P(s\_code_R|S_i)$ are estimated as shown in formula (7).

$$P(s\_code \mid S_i) = \frac{n_1}{2 * n_2} \tag{7}$$

In training corpus, Chinese sentences including ambiguous word *w* labeled with semantic code S_i are collected and its number is denoted as n_2. Here, n_1 is the number of *s_code* in Chinese sentences including ambiguous word *w* labeled with semantic code S_i. Parameters $P(S_i)$ is estimated as shown in formula (8).

$$P(S_i) = \frac{n_2}{n} \tag{8}$$

Here, *n* is the number of Chinese sentences including ambiguous word *w*. The training and test process of bayesian classifier is shown in Figure 2.

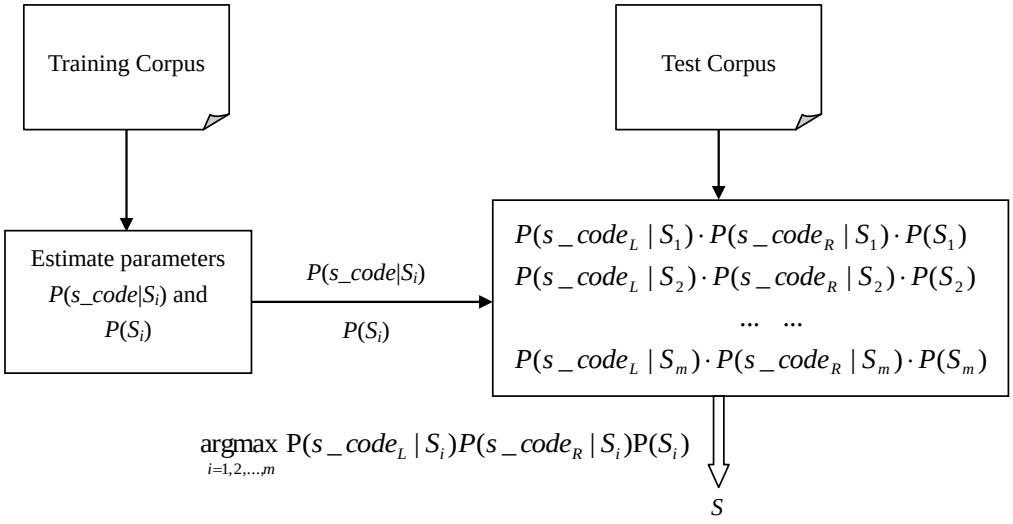

Figure 2. The Training and Test Process of Bayesian Classifier

4. Experiments

In order to measure the performance of the proposed method, SemEval-2007: #Task5 is used as training corpus and test corpus. Eight Chinese ambiguous words are selected for experiments including 'ben3', 'bu3', 'cheng2 li4', 'dui4 wu3', 'gan3', 'qi2 zhi3', 'tian1 di4' and 'chang2 cheng2'. Chinese sentences including these 8 ambiguous words are extracted from SemEval-2007: #Task5. These sentences are divided into two parts. One is training corpus and the other is test corpus.

In SemEval-2007: #Task5, there are three senses for ambiguous word 'ben3'. The first one is 'book' whose semantic code is Dk20 in TongYiCi CiLin. The second one is 'foundation' whose semantic code is Db08. The third one is 'capital' whose semantic code is Dj04. Ambiguous word 'bu3' has 3 different meanings. The first one is 'supply' whose semantic code is Ih05. The second one is 'repair' whose semantic code is Hj41. The third one is 'nourish' whose semantic code is Hj33. There are

Copyright ⓒ 2015 SERSC

three senses for ambiguous word ‘cheng2 li4’. The first one is ‘be founded’ whose semantic code is Hc05. The second one is ‘establish’. Although ‘establish’ is one sense of word ‘cheng2 li4’, no semantic code is used to label it in TongYiCi CiLin. So, we label it with -1. The third one is ‘be tenable’ whose semantic code is Ed13. Ambiguous word ‘dui4 wu3’ has 3 different meanings. The first one is ‘contingent’ whose semantic code is Aj07. The second one is ‘ranks’ whose semantic code is Di10. The third one is ‘troops’ whose semantic code is Di11. There are three senses for ambiguous word ‘gan3’. The first one is ‘rush for’ whose semantic code is Hj67. The second one is ‘drive’ whose semantic code is Hf01. The third one is ‘happen to’ whose semantic code is Hi07. Ambiguous word ‘qi2 zhi3’ has 3 different meanings. The first one is ‘stand’ whose semantic code is Dd11. The second one is ‘model’. But, no semantic code is used to label this sense. So, we label it with -1. The third one is ‘banner’ whose semantic code is Bp20. There are three senses for ambiguous word ‘tian1 di4’. The first one is ‘field of activity’ whose semantic code is Dd05. The second one is ‘heaven and earth’ whose semantic code is Bd01. The third one is ‘world’. But, no semantic code is used to label this sense. So, we label it with -1. Ambiguous word ‘chang2 cheng2’ has 3 different meanings. The first one is ‘the great wall’. But, no semantic code is used to label this sense. So, we label it with -1. The second one is ‘chang cheng’. But, no semantic code is used to label this sense. So, we label it with 1. The third one is ‘impregnable bulwark’ whose semantic code is Bn01. Their distributions are shown in Table 4.

Table 4. The Distribution of Training Corpus and Test Corpus

Ambiguous words	Semantic codes	Number of sentences in training corpus	Number of sentences in test corpus
‘ben3’	Dk20	30	10
	Db08	29	10
	Dj04	9	5
‘bu3’	Ih05	30	10
	Hj41	9	4
	Hj33	24	7
‘cheng2 li4’	Hc05	30	10
	-1	30	10
	Ed13	13	7
‘dui4 wu3’	Aj07	30	10
	Di10	24	9
	Di11	10	3
‘gan3’	Hj67	30	9
	Hf01	18	6
	Hi07	8	3
‘qi2 zhi3’	Dd11	30	10
	-1	9	4
	Bp20	11	4
‘tian1 di4’	Dd05	30	10
	Bd01	20	10
	-1	15	5
‘chang2 cheng2’	-1	28	10
	1	15	8
	Bn01	5	3

In order to measure the performance of the proposed method, two groups of experiments are conducted. In experiment 1, the word-based disambiguation method is used. Firstly, Chinese sentences are segmented into words. Secondly, left word and right word around ambiguous word are extracted as discriminative features. WSD classifier based on bayesian model is built in which left word and right word are used as features. The word-based disambiguation classifier is shown in formula (9). Parameter $P(w_L|S_i)$ is estimated as shown in formula (10). Here, n_3 is the number of w_L in

 Copyright ⓒ 2015 SERSC

Chinese sentences including ambiguous word w labeled with semantic code S_i. Parameter $P(w_R|S_i)$ is estimated as shown in formula (11). Here, n_4 is the number of w_R in Chinese sentences including ambiguous word w labeled with semantic code S_i.

$$\begin{aligned} S &= \underset{i=1,2,\ldots,m}{\operatorname{argmax}} P(S_i \mid w_L, w_R) \\ &= \underset{i=1,2,\ldots,m}{\operatorname{argmax}} \frac{P(S_i, w_L, w_R)}{P(w_L, w_R)} \\ &\approx \underset{i=1,2,\ldots,m}{\operatorname{argmax}} P(S_i, w_L, w_R) \\ &= \underset{i=1,2,\ldots,m}{\operatorname{argmax}} P(w_L, w_R \mid S_i) P(S_i) \\ &\approx \underset{i=1,2,\ldots,m}{\operatorname{argmax}} P(w_L \mid S_i) P(w_R \mid S_i) P(S_i) \end{aligned} \tag{9}$$

$$P(w_L \mid S_i) = \frac{n_3}{2 * n_2} \tag{10}$$

$$P(w_R \mid S_i) = \frac{n_4}{2 * n_2} \tag{11}$$

Training corpus for every ambiguous word is used to estimate parameters of WSD classifier and optimized it. Then the optimized classifier is applied to determine correct meanings of ambiguous words in test corpus. Accuracy rate is used to measure the classifier's performance. The classification result is shown in table 5.

In experiments 2, the semantics-based disambiguation method is used. Firstly, Chinese sentences are segmented into words. Secondly, get semantic codes of left word and right word around ambiguous word from Tongyici Cilin. Training corpus for every ambiguous word is used to estimate parameters $P(s\_code_R|S_i)$ and $P(S_i)$ as shown in formula (7) and formula (8). Then the optimized classifier is applied to determine correct meanings of ambiguous words in test corpus. The classification result is shown in Table 5.

Table 5. Accuracy Rate of Disambiguation on Test Corpus

Ambiguous words	Accuracy rate of word-based disambiguation(%)	Accuracy rate of semantics-based disambiguation(%)
'ben3'	68.0%	72.0%
'bu3'	40.0%	45.0%
'cheng2 li4'	59.3%	66.7%
'dui4 wu3'	36.4%	45.5%
'gan3'	27.8%	27.8%
'qi2 zhi3'	55.6%	55.6%
'tian1 di4'	50.0%	50.0%
'chang2 cheng2'	14.3%	23.8%

From Table 5, we can see that accuracy rate of experiment 2 is higher than or equal to that of experiment 1. The growth of accurate rate is 4% for word 'ben3'. For word 'bu3', its growth of accurate rate is 5%. The growth of accurate rate is 7.4% for word 'cheng2 li4'. For word 'dui4 wu3', its growth of accurate rate is 9.1%. For word 'gan3', 'qi2 zhi3' and 'tian1 di4', their accurate rates do not grow. The growth of accurate rate is 9.2% for word 'chang2 cheng2'. The reason is that left word and right word of ambiguous word are applied to determine its correct meaning in experiment 1. Maybe, there is data sparseness when parameters of WSD classifier are estimated. In experiment 2, semantic information of left word and right

Copyright ⓒ 2015 SERSC

word around ambiguous word is adopted to guide the disambiguation process. Semantic codes are discriminative features which have more generalization ability than words. More disambiguation information will be provided. It decreases the influence of data sparseness when parameters of WSD classifier are estimated. When files including parameters of classifiers are opened, we can find that lots of parameters are zero in classifier of experiment 1. But there are few parameters whose values are zero in experiment 2.

5. Conclusion

In this paper, semantic knowledge is introduced into the model of word sense disambiguation. In Chinese sentences, an ambiguous word is viewed as center. Its left and right words are extracted. Look up Tongyici Cilin to determine semantic codes of its left and right words. Their semantic codes are used as discriminative features. At the same time, bayesian model is applied to determine the correct meaning of an ambiguous word. The WSD classifier is optimized and tested. Experimental results show that its disambiguation performance is improved.

Acknowledgements

This work is supported by China Postdoctoral Science Foundation Funded Project under Grant Nos. 2014M560249 and Science and Technology Research Funds of Education Department in Heilongjiang Province under Grant Nos. 11541045.

References

[1] T. H. Wang, J. Y. Rao and Q. Hu, "Supervised word sense disambiguation using semantic diffusion kernel", Engineering Applications of Artificial Intelligence, vol. 27, (**2014**), pp. 167-174.
[2] A. Bordes, X. Glorot, J. Weston and Y. Bengio, "A semantic matching energy function for learning with multi-relational data: application to word-sense disambiguation", Machine Learning, vol. 94, no. 2, (**2014**), pp. 233-259.
[3] R. Navigli and M. Lapata, "An experimental study of graph connectivity for unsupervised word sense disambiguation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 4, (**2010**), pp. 678-692.
[4] Z. Z. Yang and H. Y. Huang, "Graph based word sense disambiguation method using distance between words", Journal of Software, vol. 23, no. 4, (**2012**), pp.776-785.
[5] D. M. Fan, Z. M. Lu, R. B. Zhang and S. S. Pan, "Chinese word sense disambiguation based on bayesian model improved by information gain", Journal of Electronics & Information Technology, vol. 30, no. 12, (**2008**), pp. 2926-2929.
[6] S. Lu, S. Bai, X. Huang and J. Zhang, "Supervised word sense disambiguation based on vector space model", Journal of Computer Research & Development, vol. 38, no. 6, (**2001**), pp. 662-667.
[7] W. W. Guo and M. Diab, "Combining orthogonal monolingual and multilingual sources of evidence for all words WSD", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Sweden, (**2010**), pp. 1542-1551.
[8] S. Faralli and R. Navigli, "A new minimally-supervised framework for domain word sense disambiguation", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Korea, (**2012**), pp. 1411-1422.
[9] R. Navigli and S. P. Ponzetto, "Joining forces pays off: multilingual joint word sense disambiguation", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Korea, (**2012**), pp. 1399-1410.
[10] S. P. Ponzetto and R. Navigli, "Knowledge-rich word sense disambiguation rivaling supervised systems", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Sweden, (**2010**), pp. 1522-1531.
[11] D. Schwab, J. Goulian and A. Tchechmedjiev, "Worst-case complexity and empirical evaluation of artificial intelligence methods for unsupervised word sense disambiguation", International Journal of Web Engineering and Technology, vol. 8, no. 2, (**2013**), pp. 124-153.
[12] S. L. Huang, X. L. Zheng, H. X. Kang and D. R. Chen, "Word sense disambiguation based on positional weighted context", Journal of Information Science, vol. 39, no. 2, (**2013**), pp. 225-237.
[13] Z. Y. Niu, D. H. Ji and C. L. Tan, "Learning model order from labeled and unlabeled data for partially supervised classification, with application to word sense disambiguation", Computer Speech and Language, vol. 21, no. 4, (**2007**), pp. 609-619.

 Copyright ⓒ 2015 SERSC

[14] A. C. Le, A. Shimazu, V. N. Huynh and L. M. Nguyen, "Semi-supervised learning integrated with classifier combination for word sense disambiguation", Computer Speech and Language, vol. 22, no. 4, (**2008**), pp. 330-345.
[15] Z. H. Huang, Y. D. Chen and X. D. Shi, "A novel word sense disambiguation algorithm based on semi-supervised statistical learning", International Journal of Applied Mathematics and Statistics, vol. 43, no. 13, (**2013**), pp. 452-458.
[16] C. A. Le, V. N. Huynh, A. Shimazu and Y. Nakamori, "Combining classifiers for word sense disambiguation based on dempster-shafer theory and owa operators", Data and Knowledge Engineering, vol. 63, no. 2, (**2007**), pp. 381-396.

Authors

Chun-Xiang Zhang, is Ph.D. and graduates from Ministry of Education-Microsoft Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, in Harbin Institute of Technology. He is also a professor in Harbin University of Science and Technology. His research interests are natural language processing, machine translation and machine learning. He has authored and coauthored more than fifty journal and conference papers in these areas.

Copyright ⓒ 2015 SERSC

 Copyright ⓒ 2015 SERSC