International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.263-270
http://dx.doi.org/10.14257/ijdta.2015.8.6.24

Chinese Word Sense Disambiguation Based on Hidden Markov Model

Zhang Chun-Xiang[1, 2], Sun Yan-Chen[3], Gao Xue-Yao[3*] and Lu Zhi-Mao[4]

[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

**Corresponding author E-mail: gaoxueyao@hrbust.edu.cn (*Gao Xue-Yao*)*

Abstract

Word sense disambiguation (WSD) is important for natural language processing. It plays important roles in information retrieval, machine translation, text categorization and topic tracking. In this paper, the transition among senses of words is considered. For an ambiguous word, its semantic codes and its left word's semantic codes are taken as disambiguation features. At the same time, a new method based on hidden Markov model (HMM) is proposed for Chinese word sense disambiguation. Chinese Tongyici Cilin is used to determine semantic codes of words. HMM is optimized in training corpus. The WSD classifiers based on HMM is tested. Experimental results show that the accuracy of word sense disambiguation is improved.

Keywords: *word sense disambiguation; semantic code; hidden Markov model; Tongyici Cilin*

1. Introduction

In natural language, words are always ambiguous. The task of word sense disambiguation is to ascertain the specific meaning of a word according to its context. WSD plays an important role in natural language processing fields. In Chinese sentence 'yi wei lao zhong yi gei ta kan bing', word 'zhong yi' means a practitioner of Chinese medicine. In Chinese sentence 'zhong yi he xi yi xiang jie he de cheng guo', word 'zhong yi' means traditional Chinese medical science.

Liu proposes a novel algorithm for simple semantic units in order to use semantic knowledge more quickly and effectively. This algorithm is based on dynamic programming method [1]. Simonini gives an automatic method to build a generic sense inventory which is used as a reference for WSD. The community detection algorithm is applied to extract insight from big data in order to construct the inventory [2]. Nguyen regards WSD as a traveling salesman problem. Its purpose is to maximize general semantic relatedness of context. Then, he solves this problem by ant colony optimization algorithm [3]. Akkaya utilizes a lot of labeled data for subjectivity word sense disambiguation. The data is semi-automatically produced with cluster and label strategy. He describes an iterative constrained clustering algorithm to improve the clustering purity [4]. Agirre proposes an algorithm based on random walks for word sense disambiguation. The algorithm is better than other graph-based methods in precision when it runs on a graph which is built from

ISSN: 2005-4270 IJDTA

WordNet and extended WordNet [5]. Dhillon uses feature relevance prior to select discriminative features for word sense disambiguation. Then, he uses the transfer of knowledge from similar words to learn the prior over features, from which a higher accurate model is gotten [6]. De optimizes PageRank algorithm for word sense disambiguation. Its accuracy is kept and the processing time is decreased [7]. Broda proposes a method based on clustering text snippets. His purpose is to reduce human intervention for word sense disambiguation [8]. Lefever uses a language-independent framework to abstract senses from word alignments on a parallel corpus, in order to determine correct senses of ambiguous words [9]. But, the monolingual sense inventory is not utilized in his method. Ponzetto extends semantic relations of Wikipedia to WordNet and builds semantic corresponding relationships between them [10]. Experiments show that unsupervised algorithm is close to supervised one in precision when high-quality semantic relations are provided. Yu proposes a novel method of rule extraction by attribute features for word sense disambiguation [11]. Huang gives a new semi-supervised algorithm to obtain high-quality labeled data for WSD, in which he builds an initial classifier with a certain accuracy rate in a few of labeled data [12]. Wu discusses different strategies of opening contextual windows for word sense disambiguation in order to construct an optimized bayes classifier. Its purpose is to find out effective rules to select contexts including more discriminative information [13]. Le uses unlabeled data to determine senses of ambiguous words within a semi-supervised learning framework. He applies combination strategies to solve three piecemeal problems occurred in a general bootstrapping algorithm [14]. Abdalgader proposes a novel unsupervised similarity-based algorithm for word sense disambiguation. The algorithm is operated by calculating semantic similarities between glosses of target word and a contextual vector [15]. Preotiuc-Pietro studies feature selection methods for unsupervised word sense disambiguation and presents a new method based on web n-gram features [16].

In this article, we view semantic categories of target word and left word as disambiguation features. Then, hidden Markov model is used to construct disambiguation classifier. At the same time, the process of disambiguation is taken as the decoding problem of hidden Markov model. Human annotated corpus is used to train model parameters. Then, the optimized classifier is tested.

2. Word Sense Disambiguation Based on HMM

Hidden Markov model is a probability model about time series. It describes an unobserved state sequence which hides in a Markov chain. State and observation are corresponded with each other. HMM can generate a stochastic observation sequence. In the process of word sense disambiguation, a sentence can be expressed as a Markov chain. A word in a sentence could be viewed as an observation. Semantics of words will be viewed as hidden states. The left word of ambiguous word can affect its semantic category. HMM is a strong sequential model. HMM is suitable for solving disambiguation problem. We take semantic categories of left words as disambiguation features. HMM is used for building disambiguation model λ=(S, W, A, B, π). Here, a sequence of semantic categories is determined by parameter π and parameter A. The sequence of observations is determined by parameter B.

S denotes a set which contains semantic categories of all words in a sentence. The number of semantic categories in set S is N.

W stands for a set containing all words in a sentence. The number of words in set W is M.

$A=[a_{ij}]_{N\times N}$ denotes transition probability matrix among semantic categories. Here, i=1, 2, …, N, j=1, 2, …, N. Element a_{ij} stands for transition probability from semantic category s_i to s_j.

$B=[b_j(k)]_{N\times M}$ denotes transformation probability matrix between words and semantic categories. This matrix is also called as confusion matrix. Here, k=1, 2, …, M, j=1, 2, …, N. Element $b_j(k)$ denotes a probability that word w_k is produced under semantic category s_j. It is also a probability that semantic category of w_k is s_j in training corpus.

$\pi=(\pi_1, \pi_2, \ldots, \pi_N)$ stands for a vector of start probabilities. Here, i=1, 2, …, N. π_i denotes a probability that semantic category s_i occurs in training corpus.

There are three problems in hidden Markov model. They are evaluating problem, learning problem and forecasting problem. The parameter training of model $\lambda=(S, W, A, B, \pi)$ is evaluating problem and learning problem. Word sense disambiguation is forecasting problem which is also called as decoding problem. The disambiguation process based on HMM is to find semantic sequence $S_T=s_1, s_2, \ldots, s_{\mathrm{T}}$. Its purpose is to maximize probability $\mathrm{P}(S_T|W_T, \lambda)$ with knowing model $\lambda=(S, W, A, B, \pi)$ and word sequence $W_T=w_1, w_2, \ldots, w_{\mathrm{T}}$. The length of W_T is T. In the process of word sense disambiguation, we could see word sequence $W_T=w_1, w_2, \ldots, w_{\mathrm{T}}$. Then, the semantic sequence behind W_T is deduced.

In HMM, state sequence with maximum probability is solved by Viterbi algorithm. This algorithm is based on dynamic programming method. According to word sequence W_T, the corresponding semantic category sequence with maximum probability S_T can be solved by Viterbi algorithm. In this paper, a Chinese sentence containing ambiguous words is segmented into words. The left word unit of ambiguous word is extracted. According to Tongyici Cilin, semantic categories of words are gotten. Semantic categories are used for the disambiguation process. Word units are viewed as nodes in a Markov chain. At t=1, the first word unit in Chinese sentence is viewed as the first node. At $t=T$, ambiguous word unit is regarded as the last node.

Using Viterbi algorithm for word sense disambiguation, we define a variable $\delta_t(u)$ which is shown in formula (1).

$$\delta_t(u)=\max_{s_1,s_2,\cdots,s_{t-1}} P(s_t=u, s_{t-1},\cdots,s_1,w_t,\cdots,w_1\,|\,\lambda) \tag{1}$$

The meaning of $\delta_t(u)$ is a maximum probability path that semantic category of word w_t is u. The recursive process is shown in formula (2).

$$\delta_{t+1}(u)=\max_{v\in\{s_1,\ldots,s_N\}}[\delta_t(v)a_{ji}]b_i(w_{t+1}) \tag{2}$$

Here, $u=s_1, s_2, \ldots, s_N$, t=1, 2, …, T-1.

According to formula (3), s_T is gotten as semantic category of word w_t.

$$S_T=\max_{u\in\{s_1,\ldots,s_N\}}\delta_T(u) \tag{3}$$

Backtrack to find optimal semantic sequence $S_q=s_1, s_2, \ldots, s_{\mathrm{T}}$ for $W_T=w_1, w_2, \ldots, w_{\mathrm{T}}$.

If the quantity of left word units is big, there will be a problem of data sparseness when we train model parameters. In this paper, we take semantic categories of left words as disambiguation features.

3. Train Model Parameters

In Tongyici Cilin, semantic categories of words are given. Semantic knowledge is the foundation of word sense disambiguation. The semantic classification structure in Tongyici Cilin is a tree. Semantic category in Tongyici Cilin contains three layers. For example, a semantic category of word ‘cai’ is Br06. Its Chinese synonym is ‘cai yao’. Code B denotes big category. Code r stands for middle category. Code 06 denotes small category. There are no semantic codes for punctuations, names and symbols in Tongyici Cilin. So, their semantic codes are all set to -1.

The training and classifying process of WSD model is shown in Figure 1. Training corpus is disposed for extracting disambiguation features. Then, transition probability matrix *A* and transformation probability matrix *B* are estimated on training corpus. Chinese sentence containing an ambiguous word is segmented into words. Semantic categories of ambiguous word and its left word are gotten. According to disambiguation features, Viterbi algorithm is used to determine the sense of ambiguous word.

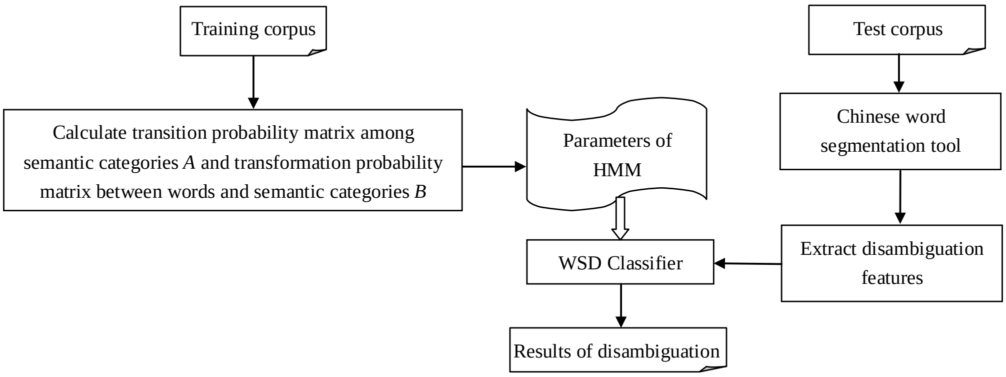
Training corpus
Test corpus
Calculate transition probability matrix among semantic categories *A* and transformation probability matrix between words and semantic categories *B*
Parameters of HMM
Chinese word segmentation tool
Extract disambiguation features
WSD Classifier
Results of disambiguation

Figure 1. The Training and Classifying Process of WSD Model

Training corpus is annotated by language engineers, in which every Chinese word is labeled with semantic category. Its form is shown in Figure 2. Here, semantic code comes from Tongyici Cilin. For example, ‘mei qing mu xiu/Eb30 de/Kd01 kong/-1 ling/-1 zuo/Fb03 zai/Kb01 yi pang/Cb06 mei/Ka18 kai kou/Hi12 ./-1’.

……	Word i-1	Semantic code of word i-1	Word i	Semantic code of word i	……

Figure 2. The Form of Training Corpus

Test corpus is composed of Chinese sentences including ambiguous words. For instance, ‘ci cai zhu yao sheng zhang yu hai ba er qian wu bai mi shang xia de cang shan ma long feng.’. It is processed by Chinese word segmentation tool. The result is ‘ci/ cai/ zhu yao/ sheng zhang/ yu/ hai ba/ er qian wu bai/ mi/ shang xia/ de/ cang shan/ ma long feng/ ./’.

The ambiguous word is ‘cai’ and its left word is ‘ci’. In Tongyici Cilin, semantic categories of word ‘ci’ are Ed61 and Cb30. Semantic categories of word ‘cai’ are Br06, Bh06 and Bh09.

Because training corpus is sparse, three-layer semantic code will make transition probability matrix *A* too sparse. In Tongyici Cilin, words are classified and coded by semantic distances. For code Cb30, its Chinese synonyms are ‘zhe li’ and ‘na li’. Chinese synonyms of code Cb05 are ‘nei’ and ‘wai’. For code Cb01, its Chinese synonyms are ‘fang xiang’ and ‘wei zhi’. These semantic categories are denoted by C big category and b middle category. They all stand for ‘direction’. When transition probability a_{ij} is calculated, two-layer semantic code is only considered. Transition probability a_{ij} is computed in formula (4). Here, $Num(s_i, s_j)$ is the number of sentences including the sequence of semantic category s_i, s_j. $Num(s_i)$ is the number of sentences containing semantic category s_i.

$$a_{ij} = Num(s_i, s_j) / Num(s_i) \tag{4}$$

In training corpus, transition probability matrix *A* is calculated. The results are shown as follows.

$$\begin{array}{ccccc} & Ed & Cb & Br & Bh \\ Ed & 0.02 & 0.02 & 0.01 & 0.01 \\ Cb & 0.02 & 0.04 & 0.01 & 0.01 \\ Br & 0.01 & 0.01 & 0.02 & 0.01 \\ Bh & 0.01 & 0.05 & 0.01 & 0.03 \end{array}$$

Here, $b_j(k)$ denotes the probability that semantic category of word w_k is s_j. Its calculation process is shown in formula (5). Here, $Num(s_j, w_k)$ is the number of sentences including word w_k whose semantic category is s_j. $Num(w_k)$ is the number of sentences containing word w_k in training corpus.

$$b_j(k) = Num(s_j, w_k) / Num(w_k) \qquad (5)$$

According to formula (5), transformation probability matrix *B* is shown as follows.

$$\begin{array}{ccc} & ci & cai \\ Ed & 0.50 & 0.00 \\ Cb & 0.50 & 0.00 \\ Br & 0.00 & 0.55 \\ Bh & 0.00 & 0.45 \end{array}$$

The probability that Ed is taken as semantic code of 'ci' is 0.50. The probability that Cb is chosen as semantic code of 'ci' is 0.50. The probability that Br is taken as semantic code of 'cai' is 0.55. The probability that Bh is chosen as semantic code of 'cai' is 0.45.

Start probability vector is π=(0.50, 0.50, 0.00, 0.00). Word sequence is W_T=ci, cai and the value of T is 2.

Viterbi algorithm is applied to determine the sequence of semantic categories S_T=Ed, Bh. The path of semantic categories could be selected from a grid graph of Viterbi. The graph is shown in Figure 3. Here, the dotted line stands for an optimal and selected path of semantic categories.

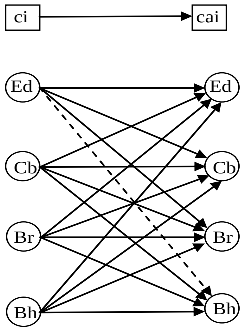

Figure 3. The Grid Graph of Viterbi

Here, semantic code of word 'ci' is Ed. Its meanings are 'zhe ge', 'na ge', 'mou ge', 'ge ge', 'qi ta' and 'he' in Tongyici Cilin. The semantic code of word 'cai' is Bh. Its meaning is 'shu cai'.

4. Experiment

We use SemEval-2007 #Task5 as test corpus to measure the performance of the proposed method. Ten frequently-used ambiguous words are chosen for test. These words are 'bu', 'chen li', 'dui wu', 'ri zi' and 'shi'. The distribution of test corpus is shown in Table 1.

Table 1. The Distribution of Test Corpus

Ambiguous words	The number of sentences
bu	20
cheng li	27
dui wu	22
ri zi	32
shi	16

Two experiments are designed and conducted. In experiment 1, morphology is taken as disambiguation feature. Bayes model is used for word sense disambiguation. Training corpus in SemEval-2007 #Task5 is applied to optimize bayes model. In experiment 2, semantic code is chosen as disambiguation feature. Viterbi algorithm is used for word sense disambiguation. HMM is trained by HIT human annotated corpus. The scale of training corpus is 10000 sentences. Experimental results are shown in Table 2.

Table 2. Accuracy Rates of Disambiguation in Two Experiments

	Experiment 1	Experiment 2
bu	40.0%	50.0%
cheng li	59.3%	74.1%
dui wu	36.4%	45.5%
ri zi	46.9%	62.5%
shi	62.5%	62.5%

From Table 2, we can see that accuracy rate of word 'bu' is increased by 10%. For word 'cheng li', 14.8 percent improvement of accuracy rate is obtained. Accuracy rate of word 'dui wu' is increased by 9.1%. For word 'ri zi', 15.6 percent improvement of accuracy rate is obtained. Accuracy rate of word 'shi' is unchanged. Compared with experiment 1, accuracy rate of disambiguation in experiment 2 is improved. There are two reasons. Firstly, the disambiguation classifier has more language generalization ability in which semantic codes are viewed as disambiguation features. Secondly, HMM is adopted. When we determine sense of target word, semantic category of its left word is considered. So the sense selection is more accurate.

5. Conclusion

In this paper, semantic code of left word is taken as disambiguation feature. Then hidden Markov model is used for building WSD classifier. HMM is trained with HIT human annotated corpus and Tongyici Cilin. Semantic category of target word in test corpus is determined by the optimized HMM. Comparative experiments show that its disambiguation performance is improved.

Acknowledgement

This work is supported by China Postdoctoral Science Foundation Funded Project (2014M560249) and Natural Science Foundation of Heilongjiang Province of China (F2015041).

References

[1] Y. T. Liu, J. Xiong and J. L. Cui, "A word sense disambiguation algorithm for the simple semantic units based on semantic relevancy", Journal of Computational Information Systems, vol. 10, no. 4, (**2014**), pp. 1555-1563.

[2] G. Simonini and F. Guerra, "Using big data to support automatic word sense disambiguation", Proceedings of the 2014 International Conference on High Performance Computing and Simulation, **(2014)**, pp. 233-259.

[3] K. H. Nguyen and C. Y. Ock, "Word sense disambiguation as a traveling salesman problem", Artificial Intelligence Review, vol. 40, no. 4, **(2013)**, pp. 405-427.

[4] C. Akkaya, J. Wiebe and R. Mihalcea, "Iterative constrained clustering for subjectivity word sense disambiguation", Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, **(2014)**, pp. 269-278.

[5] E. Agirre, O. L. D. Lacalle and A. Soroa, "Random walks for knowledge-based word sense disambiguation", Computational Linguistics, vol. 40, no. 1, **(2014)**, pp. 57-84.

[6] P. S. Dhillon and L. H. Ungar, "Transfer learning, feature selection and word sense disambguation", Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language, **(2009)**, pp. 257-260.

[7] C. D. De, R. Basili, M. Luciani, F. Mesiano and R. Rossi, "Robust and efficient page rank for word sense disambiguation", Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, **(2010)**, pp. 24-32.

[8] B. Broda and M. Piasecki, "Semi-supervised word sense disambiguation based on weakly controlled sense induction", Proceedings of the International Multi-conference on Computer Science and Information Technology, **(2009)**, pp. 17-24.

[9] E. Lefever, V. Hoste and C. M. De, "Para-sense or how to use parallel corpora for word sense disambiguation", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, **(2011)**, pp. 317-322.

[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, **(2010)**, pp. 1522-1531.

[11] J. P. Yu, C. Li, W. X. Hong, S. X. Li and D. M. Mei, "A new approach of rules extraction for word sense disambiguation by features of attributes", Applied Soft Computing Journal, vol. 27, **(2015)**, pp. 411-419.

[12] 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.

[13] C. B. Wu and Q. Zhang, "Word sense disambiguation based on bayesian classifier with tailored context window", Journal of Computational Information Systems. vol. 8, no. 12, **(2012)**, pp. 5195-5202.

[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] K. Abdalgader and A. Skabar, "Unsupervised similarity-based word sense disambiguation using context vectors and sentential word importance", ACM Transactions on Speech and Language Processing, vol. 9, no. 1, **(2012)**, pp. 1-21.

[16] D. Preotiuc-Pietro and F. Hristea, "Unsupervised word sense disambiguation with N-gram features", Artificial Intelligence Review, vol. 41, no. 2, **(2014)**, pp. 241-260

Author

Chun-Xiang Zhang, he 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. At the same time, he is a post doctor in College of Information and Communication Engineering, in Harbin Engineering University. His research interests are natural language processing and machine learning.