International Journal of Security and Its Applications
Vol.9, No.11 (2015), pp.209-216
http://dx.doi.org/10.14257/ijsia.2015.9.11.20

Chinese Word Sense Disambiguation Based on Beam Search

Zhang Chun-Xiang[1], He Shan[2], Gao Xue-Yao[21*] and Lu Zhi-Mao[3]

[1]School of Software, Harbin University of Science and Technology,
Harbin 150080, China
[2]School of Computer Science and Technology,
Harbin University of Science and Technology,
Harbin 150080, China
[3]School of Computer Science and Technology,
Dalian University of Technology,
Dalian 116024, China
{gao, xue-yao} gaoxueyao@hrbust.edu.cn

Abstract

Research on word sense disambiguation (WSD) is of great importance in natural language processing. In this paper, a new method based on beam search algorithm for Chinese WSD is proposed. By mining potential knowledge between phrase and semantic category in a sentence, this approach can construct its semantic network. It searches an optimal semantic category sequence from a Chinese sentence's semantic network with beam search algorithm, so that correct meanings of ambiguous words can be found from the optimal sequence. Experiments show that a better WSD performance is gotten.

Keywords*: word sense disambiguation; beam search algorithm; semantic network; optimal semantic category sequence*

1. Introduction

Yu proposes a new method of rule extraction for word sense disambiguation in which word features are used. Experimental results show that the proposed method is easier and more efficient than structural partial ordered attribute diagram methods[1]. Ondrej mines additional knowledge from lexicons and parallel texts, and designs a system for verb WSD. The disambiguation performance is improved[2]. Zhong gives a new method to evaluate sense distributions for short queries. At the same time, word senses is integrated into language model in information retrieval and synonym relations are considered[3]. Fakhrahmad develops a new WSD system based on a data mining algorithm and an expert system. In this expert system, word co-occurrences and association rules generated by data mining algorithm are utilized[4]. Chasin studies WSD methods with knowledge sources in specific domains. At the same time, their performance enhancements are investigated[5]. Nguyen considers word sense disambiguation as traveling salesman problem. Ant colony algorithm is used to select correct sense for an ambiguous word through maximizing total semantic relevance of its contexts[6]. Klapaftis analyses evaluation results in Semeval 2010 word sense induction task, and gives a new evaluation point for sense induction[7]. Zhao integrates thesaurus and machine learning algorithms into WSD, and provides a new semi-supervised WSD method in which contexts are weighed[8].

*Gao Xue-Yao[2] is the corresponding author.

ISSN: 1738-9976 IJSIA

Moro gives a united graph-based method for entity linking and word sense disambiguation in which identification of candidate meanings and a densest sub-graph heuristic are employed[9]. Johansson presents a new WSD method in which sense representations are embed into continuous vector space to represent words[10]. Fauceglia proposes a new method to determine event types, in which syntactic contexts of a verb are applied to select its correct sense[11]. Ponzetto extends WordNet with semantic relations from Wikipedia. Its semantic knowledge in high quality is applied to WSD[12]. Taghipour studies the effect of words' continuous space representation on two WSD methods. At the same time, their performances on senseval lexical samples and a domain-specific task are evaluated[13]. Li builds a model with conditional probabilities of sense paraphrases to select correct sense for an ambiguous word in a given context[14]. Faralli gives a novel minimally-supervised framework for domain-driven WSD, in which a bootstrapping method is used to collect domain-specific dictionaries from web[15].

Based on phrase semantics knowledge, a Chinese WSD model is given in this paper. A Chinese sentence's semantic network is constructed based on phrase semantics knowledge. Beam search algorithm is used to find an optimal semantic category sequence for this Chinese sentence, which determines correct meanings of ambiguous words.

2. Translation Model

A phrase is made up of words which are correlated in syntax and semantics. Some phrases are ambiguous because they contain ambiguous vocabularies. Tongyici Cilin gives the relevance between words and semantics. There are many different semantic categories for a vocabulary in Tongyici Cilin. So, one phrase has multiple semantic categories. A sentence is segmented into several phrases and there are multiple segmentation methods. In the same way, one sentence can also have different semantic category sequences, where only one can describe this sentence's correct meaning. If phrases' semantic categories are viewed as nodes and their adjacent relationships are regarded as edges, a sentence's semantic network is constructed. When semantic network is traveled, this sentence's semantic category sequence is obtained. In this paper, correlations between phrases and semantics are adopted to determine correct meanings of ambiguous words.

Given Chinese sentence C=w_1, w_2, …, w_n, the process of determining its optimal semantic category sequence from semantic network is shown in Formula (1).

$$S_{best} = \underset{S_i}{argmax} \mathrm{P}\left(S_i / C\right) \tag{1}$$

There are several semantic category sequences for sentence C in semantic network and they are described as S_1, S_2, …, S_m. Its optimal one is denoted as S_{best}. Sentence C is viewed as phrase sequence $ph^i{}_1$, …, $ph^i{}_j$, …, $ph^i{}_n$. S_i is regarded as semantic category sequence $s^i{}_1$, …, $s^i{}_j$, …, $s^i{}_n$. $s^i{}_j$ is semantic category of phrase $ph^i{}_j$. Formula (1) shows that our method's task is to find an optimal semantic category sequence S_{best} for sentence C from S_1, S_2, …, S_m.

Formula (1) derives from noisy channel model in statistical machine translation. The input's true meaning is seen as semantic category sequence S_{best} which is changed through noisy channel, and a sequence of phrases which compose Chinese sentence C is obtained on its other side. The task of WSD is to restore semantic category sequence S_{best} according to sentence C. In this way, S_{best} is an input in noisy channel model and sentence C is its output. Chinese vocabularies' ambiguity is an influence factor of noisy channel model.

There are several phrase segmentation methods for sentence C. C is segmented into multiple phrases $ph^i{}_1$, …, $ph^i{}_j$, …, $ph^i{}_n$ in the ith segmentation method. Semantic category of $ph^i{}_j$ is $s^i{}_j$ (j=1, 2, …, n). All phrases' semantic categories compose semantic category sequence S_i. In the case of conditional independence assumption, the probability that a sentence occurs with its semantic category sequence is viewed as a product of multiple occurrence probabilities between phrases and their corresponding semantic categories. So, probability $P(S_i/C)$ can be calculated according to Formula (2).

$$\mathrm{P}(\mathrm{S_i/C}) \approx \prod_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{P_c}(s_j^i \,/\, ph_j^i) \tag{2}$$

Here, $P(S_i/C)$ denotes the probability between sentence C and semantic category sequence S_i. $P_c(s^i_j/ph^i_j)$ is occurrence probability between phrase ph^i_j and semantic category s^i_j. Probability $P_c(s^i_j/ph^i_j)$ is computed as shown in formula (3).

$$\mathrm{P_c}(\mathrm{s^i_j/ph^i_j}) = \frac{count\left(\mathrm{ph^i_j, s^i_j}\right)}{count\left(\mathrm{ph^i_j}\right)} \tag{3}$$

Here, $(ph^i_j,\ s^i_j)$ denotes a pair between phrase and semantic category. Human-annotated corpus in which every word is annotated with its semantic category is used to estimate parameters in formula (3). $count(ph^i_j,\ s^i_j)$ is the number of sentences that contain phrase ph^i_j whose semantic category is s^i_j in human-annotated corpus. $count(ph^i_j)$ is the number of sentences that contain phrase ph^i_j in human-annotated corpus.

3. Beam Search for WSD

Sentence C can be divided by several segmentation methods. In each segmentation method, C is segmented into several phrases. Each phrase has its corresponding semantic category. If phrases' semantic categories are viewed as nodes and their adjacent relationships are regarded as edges, a sentence's semantic network is constructed. In semantic network, each path from initial node to terminal node denotes a semantic category sequence for sentence C. The first phrase's semantic category is located in initial node. The last phrase's semantic category is located in terminal node. Beam search algorithm is applied to search an optimal semantic category sequence for sentence C from its semantic network.

For Chinese sentence 'KeXue LiShi Shang Bei ShiShi TuiFan De LiLun HenDuo', its semantic network is built as shown in Figure 1.

KeXue	LiShi	Shang	Bei	ShiShi	TuiFan	De	LiLun	HenDuo
Ed12	Da07	Ca04	Kb05	Da21	Ha01	Ed01	Hi41	Eb01
Dk03	Dd01	Cb03	Bp28	Da01	Hc14	Kd01	Dk02	
	Da07 Ca04		Je13		Hi22	Bo29		
Dk03 Da07				Da01 Hi22			Dk02 Eb01	
			Kb05 Da01 Hi22					

Figure 1. Semantic Network of Sentence *C*

For 'KeXue LiShi', there are two segmentation methods. In the first method, it is divided into two phrases 'KeXue' and 'LiShi'. So, two pairs between phrase and semantic category including 'KeXue-Dk03' and 'LiShi-Da07' can be obtained. In the second method, it is viewed as a phrase 'KeXue LiShi'. So, a pair between phrase and semantic category 'KeXue LiShi-Dk03 Da07' is gotten.

In beam search algorithm, heuristic strategies are used. It finds an optimal semantic category sequence for sentence C from its semantic network, which maximizes the probability of formula (1). It chooses a node randomly as a basic one from network, and searches from basic node to other ones. When this algorithm finds a new node, this node is marked as a visited one. Semantic categories in all visited nodes are combined and compose a disambiguation hypothesis. Each disambiguation hypothesis represents current disambiguation result in process of searching. When all phrases in sentence are covered by a disambiguation hypothesis, this algorithm gets a disambiguation path for sentence *C*. The disambiguation algorithm is shown as follows.

1. Disambiguation hypothesis is initially empty.
   do

   ① Choose an unvisited node randomly from semantic network and its semantic category is set into current disambiguation hypothesis.
   ② Its corresponding phrase is marked as a disambiguated one and this node is marked as a visited one.
   Until all phrases are disambiguated.

2. Several disambiguation paths for sentence *C* are obtained from semantic network. Every path is corresponded with a semantic category sequence.

3. Probability between every semantic category sequence and sentence *C* will be calculated by Formula (1), and the semantic category sequence with maximum probability is considered as an optimal one.
   For the sentence in Figure 1, disambiguation hypothesis is empty at the beginning. Semantic category ‘Dk03’ is selected randomly and set into a disambiguation hypothesis. Then its corresponding phrase ‘KeXue’ is marked as a disambiguated one. Semantic category ‘Da07’ is selected and set into current disambiguation hypothesis. New hypothesis ‘Dk03 Da07’ is obtained. Phrase ‘KeXue LiShi’ is marked as disambiguated. Semantic category ‘Ca04’ can also be appended to current disambiguation hypothesis. New hypothesis ‘Dk03 Ca04’ is gotten. Then, phrase ‘KeXue Shang’ is marked as a disambiguated one. Hypothesis ‘Dk03 Da07’ can be expanded by semantic category ‘Ca04’. New hypothesis ‘Dk03 Da07 Ca04’ is obtained. At the same time, phrase ‘KeXue LiShi Shang’ is marked as disambiguated. The searching process is shown in Figure 2.

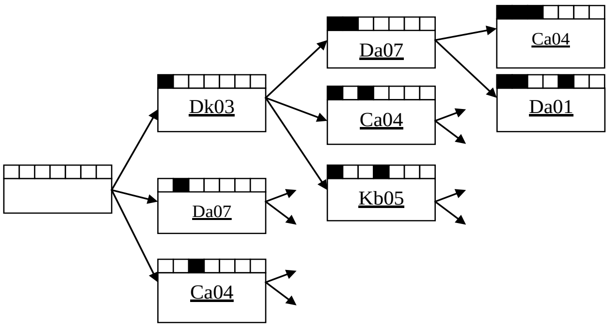

Figure 2. The Process of Beam Search Algorithm

Beam search algorithm can find multiple semantic category sequences for a sentence. The corresponding probability between sentence and semantic category sequence is calculated by Formula (1). The semantic category sequence with the maximum probability will be selected as an optimal one for the input sentence.

Stack architecture is employed to store disambiguation hypotheses in beam search algorithm. The stack is organized according to the number of words in disambiguated phrases. Each disambiguation hypothesis in the first stack contains only one word. There are only two words in the second stack. Each disambiguation hypothesis in the *n*th stack includes only n words, where all words in sentence *C* are disambiguated. The number of disambiguation hypothesis will increase gradually in stack. When the number of disambiguation hypothesis in a stack reaches a threshold value or this stack is full, it need be pruned.

4. Experiments

In order to evaluate the proposed method’s performance, human-annotated corpus is used as training set, which is developed by Ministry of Education-Microsoft Key Laboratory of Natural Language Processing and Speech in Harbin Institute of Technology. 10 thousand sentences are collected and each sentence is segmented into words. Every word is annotated manually with semantic category. The probability between phrase and semantic category is estimated on this training set. SemEval-2007 #Task5 is adopted as test set and six ambiguous words are selected. They are ‘DuiWu’, ‘Bu’, ‘Gan’, ‘ZhongYi’, ‘Cai’ and ‘Wang’. The distribution of these words is shown in Table 1.

Table 1. The Distribution of Test Corpus

Ambiguous words	The number of sentences
DuiWu	22
Bu	20
Gan	18
ZhongYi	16
Cai	19
Wang	13

There are three semantics for word ‘DuiWu’ including ‘contingent’, ‘ranks’ and ‘troops’. Their semantic categories are respectively ‘Di10’, ‘Aj07’ and ‘Di11’. Word ‘Bu’ has three semantics including ‘supply’, ‘repair’ and ‘nourish’. Their semantic categories are separately ‘Ih05’, ‘Hj41’ and ‘Hj33’. There are three semantics for word ‘Gan’ including ‘rush_for’, ‘drive’ and ‘happen_to’. Their semantic categories are respectively ‘Hj67’, ‘Hf01’ and ‘Ga12’. Word ‘ZhongYi’ has two semantics including ‘traditional_Chinese_medical_science’ and ‘practitioner_of_Chinese_medicine’. Their semantic categories are separately ‘Dk03’ and ‘Ae15’. There are two semantics for word ‘Cai’ including ‘vegetable’ and ‘dish’. Their semantic categories are respectively ‘Bh06’ and ‘Br06’. Word ‘Wang’ has two semantics including ‘gaze’ and ‘hope’. Their semantic categories are separately ‘Fc04’ and ‘Gb06’.

In order to evaluate this method’s performance, two experiments are designed. In the first experiment, a continuous disambiguation space for ambiguous vocabularies is built. There is significant knowledge in two words around an ambiguous word, whose morphologies are used as disambiguation features. A bayesian model is taken to construct a classifier for WSD. In the second experiment, formula (1) is adopted to build WSD classifier, and beam search algorithm is employed to search an optimal semantic category sequence from semantic network. The accuracy rate of disambiguation is illustrated in Table 2.

Table 2 shows that experiment 1 has a higher accuracy than experiment 2. For word ‘ZhongYi’, the improvement of accuracy rate is highest. For word ‘DuiWu’, its improvement is lowest.

Table 2. The Accuracy Rate of Disambiguation

Ambiguous words	Experiment 1	Experiment 2
DuiWu	36.3%	40.9%
Bu	40%	50%
Gan	27.8%	44.4%
ZhongYi	37.5%	75%
Cai	33.3%	47.2%
Wang	69.2%	76.1%

There are two advantages in the proposed method. On the one hand, the probability between phrase and semantic category is employed, so that semantic knowledge in phrases can be used to guide WSD process. This can decrease the degree of ambiguity in an ambiguous word's context. On the other hand, a semantic network is constructed for a sentence, and a beam search algorithm is used to search an optimal semantic category sequence. So, the knowledge in phrase can be mined fully.

Conclusion

In this paper, knowledge between phrase and semantic category is used to build disambiguation model. In order to utilize semantic knowledge fully, a semantic network is constructed for Chinese sentence. Beam search algorithm is employed to find an optimal semantic category sequence. Experimental results show that accuracy rate of WSD is improved by this method.

Acknowledgements

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

References

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

[2] D. Ondrej, F. Eva, H. Jan and P. Martin, "Using parallel texts and lexicons for verbal word sense disambiguation", Proceedings of the Third International Conference on Dependency Linguistics, **(2015)**, pp. 82-90.

[3] Z. Zhong and H. T. Ng, "Word sense disambiguation improves information retrieval", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, **(2012)**, pp. 273-282.

[4] S. M. Fakhrahmad, M. H. Sadreddini and M. Z. Jahromi, "A proposed expert system for word sense disambiguation: deductive ambiguity resolution based on data mining and forward chaining", Expert Systems, vol. 32, no. 2, **(2015)**, pp. 178-191.

[5] R. Chasin, A. Rumshisky, O. Uzuner and P. Szolovits, "Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods", Journal of the American Medical Informatics Association, vol. 21, no. 5, **(2014)**, pp. 842-849.

[6] 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.

[7] I. P. Klapaftis and S. Manandhar, "Evaluating word sense induction and disambiguation methods", Language Resources and Evaluation, vol. 47, no. 3, **(2013)**, pp. 579-605.

[8] G. Z. Zhao and W. L. Zuo, "Semi-supervised word sense disambiguation via context weighting", Modern Technologies in Materials, Mechanics and Intelligent Systems, vol. 1049, **(2014)**, pp. 1327-1338.

[9] A. Moro, A. Raganato and R. Navigli, "Entity linking meets word sense disambiguation: a unified approach", Transactions of the Association for Computational Linguistics, **(2014)**, pp. 231-244.

[10] R. Johansson, "Combining relational and distributional knowledge for word sense disambiguation", Proceedings of the 20th Nordic Conference of Computational Linguistics, **(2015)**, pp. 69-78.

[11] N. Fauceglia, Y. C. Lin, X. Z. Ma and E. Hovy, "Word sense disambiguation via PropStore and OntoNotes for event mention detection", Proceedings of the 3rd Workshop on Events at the 2015

Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies, **(2015)**, pp. 11-15.

[12] 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.

[13] K. Taghipour and H. T. Ng, "Semi-supervised word sense disambiguation using word embeddings in general and specific domains", The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics, **(2015)**, pp. 314-323.

[14] L. L. Li, B. Roth and C. Sporleder, "Topic models for word sense disambiguation and token-based idiom detection", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, **(2010)**, pp. 1138-1147.

[15] 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, **(2012)**, pp. 1411-1422.

Authors

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. 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.