A Word Sense Disambiguation System Based on Bayesian Model

Chunxiang Zhang
School of Software, Harbin University of Science and Technology
Harbin, China
College of Information and Communication Engineering, Harbin Engineering University
Harbin, China
e-mail: z6c6x666@163.com

Shan He
School of Computer Science and Technology, Harbin University of Science and Technology
Harbin, China
e-mail: 1185402880@qq.com

Xueyao Gao
School of Computer Science and Technology, Harbin University of Science and Technology
Harbin, China
e-mail: xueyao_gao@163.com

***Abstract*—Research on word sense disambiguation (WSD) is of great importance in natural language processing fields. In this paper, a novel word sense disambiguation system is designed in which bayesian theory is applied to determine correct sense of an ambiguous word. Morphology knowledge in word unit is mined to guide WSD process. Neighboring morphology knowledge of an ambiguous word is used as feature for constructing WSD classifier. Word segmentation tool is integrated into this system and browser/server (B/S) framework is adopted. Experimental results show that the performance of WSD system is good.**

Keywords—word sense disambiguation; morphology knowledge; word segmentation tool; browser/server

I. INTRODUCTION

Word sense disambiguation is a technology that determines correct sense of an ambiguous word in a definite context. It has been an important search issue in natural language processing, and it has many applications containing text analysis, data mining, information retrieval, machine translation and so on.

Dhungana proposed a novel WordNet model in which ambiguous words and unambiguous words were organized. The purpose was to provide language knowledge for WSD process[1]. Nandanwar gave an unsupervised word sense disambiguation method in which HINDI WordNet was used to construct a sense graph and correct sense of a polysemous word was found from this sense graph[2]. Wessam gave a graph-based WSD method in which UMLS meta-thesaurus was taken as knowledge source. The purpose was to deal with the ambiguity in biomedical texts[3]. Akkaya proposed a word sense disambiguation method based on distributional semantic models, in which a compositional model was extended to exploit more contexts[4]. Liu applied a neural network to WSD in which neighboring words' knowledge of ambiguous word was mined as knowledge source with statistical methodology. Experiments showed that the proposed model had a good performance on word sense disambiguation[5]. Li presented a graph-based ranking method to extract features from limited language data, in which different features were adopted. Experiments showed that it had a highly effective performance[6]. Bachir applied a genetic algorithm to word sense disambiguation in modern standard arabic. A large corpus was used to train WSD model and a higher accuracy was gotten[7]. Kulkarni proposed a new WSD algorithm and lexical database WordNet was taken as disambiguation knowledge source, from which relationships between an ambiguous word and other words in a sentence were extracted[8]. Broda gave a WSD method that reduced human intervention with many text snippets, in which a classifier was constructed according to every cluster of a word[9]. Navigli proposed a structural semantic interconnection algorithm based on senses' structural specifications for each word. The best hypothesis was selected according to the related knowledge between sense specifications in process of word sense disambiguation[10].

In this paper, a WSD system is designed in which vocabularies besides an ambiguous word are used in disambiguation process. Then, a bayesian classifier is built according to disambiguation features. A word segmentation tool is employed to segment an input sentence into many word units. Semantic categories in 'Tongyici Cilin' are adopted and this system is implemented on a browser/server framework.

II. ARCHITECTURE OF WORD SENSE DISAMBIGUATION

In Chinese grammar, each word has its semantic category. An unambiguous word has a definite semantic category, but an ambiguous word has several different semantic categories. In a sentence, one semantic category can be chosen to describe its correct meanings. A semantic category of an ambiguous word

always occurs together with some neighboring vocabularies. Neighboring disambiguation knowledge of an ambiguous word can be mined from these vocabularies. Knowledge in neighboring word units of ambiguous words is extracted. The disambiguation knowledge and bayesian classification theory are utilized to guide WSD process. At the same time, word segmentation technology is adopted to divide the input sentence into word units. The purpose is to extract discriminative knowledge easily. The architecture of word sense disambiguation system proposed in this paper is shown in Figure 1.

As shown in Figure 1, there are 4 main models. They are respectively the model of word segmentation, the model of extracting discriminative features, the model of computing probability, and the model of WSD classifier. Here, bayesian model is utilized as WSD classifier. This is because that the classification performance is good and it is easy to train model parameters.

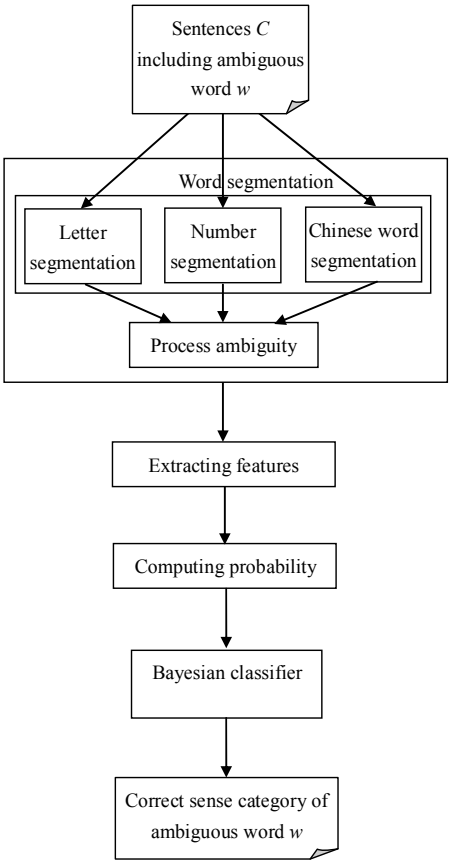

Fig. 1. The architecture of WSD system

In Figure 1, Chinese sentence *C* including ambiguous word *w* is firstly processed by word segmentation tool. In this system, Chinese word segmentation tool IKanalysis is adopted. In IKanalysis, there are mainly four models. They are respectively letter segmentation model, number segmentation model, the model of Chinese word segmentation, and the model of processing ambiguity. In letter segmentation model, English alphabet, arabic alphabet and mixed alphabet in Chinese sentence *C* are dealt with. In number segmentation model, Chinese numerals and quantifiers in sentence *C* are processed. In the model of Chinese word segmentation, Chinese vocabularies are dealt with. It is the core model in IKanalysis. In the model of processing ambiguity, ambiguous words are disambiguated briefly. Its task is to give every vocabulary a unique semantic category. Chinese sentence *C* is processed by these four components. In IKanalysis, a segmentation algorithm in most granular is adopted. After several iterations, the segmentation performance is better. When sentence *C* is input into IKanalysis, it will be dealt with respectively by three components including letter segmentation model, number segmentation model and the model of Chinese word segmentation. Each model has its independent dictionary and can implement a fast search by means of matching prefix character one by one. After sentence *C* is processed by these three models, a candidate set of word units is obtained. Then, these word units are processed by the model of processing ambiguity. Dictionary is an important component in IKanalysis. A data structure of dictionary tree is selected to increase the efficiency of segmentation. After IKanalysis is applied, there is one sense corresponding with every word in sentence *C*. Then, Chinese sentence *C* is processed by the module of extracting features. Here, two vocabularies around ambiguous word *w* are extracted as disambiguation features. If there is not left vocabulary or right vocabulary for ambiguous word *w*, disambiguation features will be set to null. For *w* in sentence *C*, there is a unique disambiguation feature.

The model of computing probability receives disambiguation features which are extracted by the model of extracting features. Chinese sentences containing ambiguous word *w* are collected to train parameters in bayesian classifier. The training process is implemented on a lot of sentences including an ambiguous word. So, the optimized parameters are gotten. In order to improve the classification efficiency, all parameters are stored in a file which will be loaded during initialization. Then, the optimized classifier is applied to determine correct sense of ambiguous word *w*. When sentence *C* is submitted into the system, it is segmented into different vocabularies. Then, discriminative features are extracted and bayesian classifier labels ambiguous word *w* with its correct semantic category automatically.

III. WSD Classifier Based on Bayesian Model

Bayesian classifier is adopted to disambiguate ambiguous words. It is a popular algorithm for the document classification problem. Word sense disambiguation system realizes an available classification function based on bayesian model here. The process of determining its correct sense is shown in Formula (1).

$$p(c \mid w) \approx \underset{c_i}{\operatorname{argmax}}\, p(w \mid c_i)\, p(c_i) \qquad (1)$$

where, c_i is the ith semantic category of word w in Chinese sense dictionary 'Tongyici Cilin'. $p(c_i|w)$ is the corresponding probability between ambiguous word w and sense category c_i. $p(c_i)$ denotes a prior probability that sense category c_i occurs in corpus. $p(w|c_i)$ is a condition probability that word w is labeled with sense category c_i in training corpus. Formula (1) indicates

that semantic category c_i whose probability $p(c_i|w)$ is maximum will be selected as correct sense for ambiguous word w.

From Chinese sentence C, we can get a unique feature which is composed of two vocabularies around ambiguous word w. w_l is viewed as the left vocabulary of ambiguous word w. w_r is defined as the right vocabulary of word w. So, the probability $p(w|c_i)$ can be calculated by $p(w_l,w_r|c_i)$. Based on the condition independence assumption, the computing process of $p(w_l,w_r|c_i)$ is shown in Formula (2).

$$p(w_l, w_r \mid c_i) \approx p(w_l \mid c_i) p(w_r \mid c_i) \qquad (2)$$

Human-annotated corpus in which every word is annotated with its semantic category is used to estimate parameters $p(w_l|c_i)$ and $p(w_r|c_i)$. Firstly, training corpus is divided according to different ambiguous words and Chinese sentences including the same word are extracted. For every word, training corpus is also separated by its different semantic categories.

Here, $p(w_l|c_i)$ is a condition probability where the left word is w_l and sense category of w is c_i. $p(w_r|c_i)$ is a condition probability where the right word is w_r and sense category of w is c_i. Under the condition where semantic category of word w is c_i, $p(w_l|c_i)$ and $p(w_r|c_i)$ are estimated. The number of Chinese sentences including ambiguous word w is n. Here, n_l is the number of sentences in which the left word of w is w_l and semantic category of word w is c_i. n_r is the number of sentences in which the right word of w is w_r and semantic category of word w is c_i. $p(w_l|c_i)$ is the proportion of n_l to n. $p(w_r|c_i)$ is the proportion of n_r to n.

IV. EXPERIMENTS

A WSD system based on bayesian classifier is given in this paper. In order to be operated conveniently by users, this system has a brief interface. The interface of this system is shown in Figure 2.

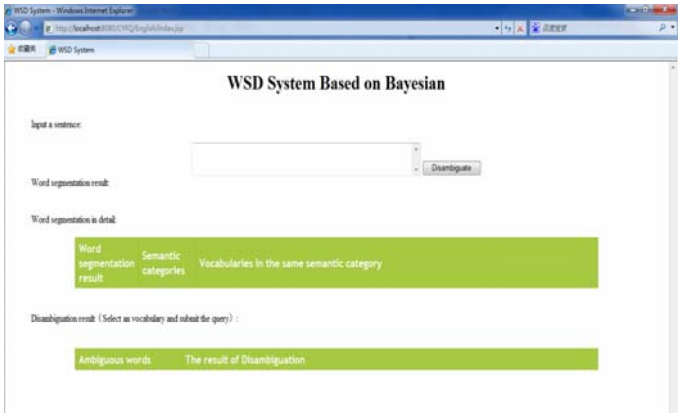

Fig. 2. The interface of WSD system

The interface has four components including the input unit, the unit of word segmentation result, the unit of word segmentation in detail, and disambiguation result unit. Users submit a sentence in the input unit. In the unit of word segmentation result, all words segmented by IKanalysis are shown. In the unit of word segmentation in detail, each word, its semantic categories and synonymy vocabularies are shown. For disambiguation result unit, there are two components. One shows correct semantic category of an ambiguous word and the other shows the disambiguation result in a tree structure.

The system is implemented on browser/server framework. A brief interface is designed on browser and an application program is completed on server. The system is disposed on tomcat server. Users input information in the interface. Then, the information is packaged and submitted into server. Firstly, the information is sent into a garbling filter in order to settle Chinese garbling phenomenon. Secondly, it is processed with a specific function which detects the information and puts it into word segmentation tool. Bayesian classifier receives word segmentation results and determines correct sense of an ambiguous word. This is the core function in word sense disambiguation system. The disambiguation results are formatted and set into a response function which sends the response message to the corresponding browser. Browser accepts the response message and extracts the disambiguation result. Thirdly, all disambiguation results are shown in the interface.

For this WSD system, its class diagram is shown in Figure 3.

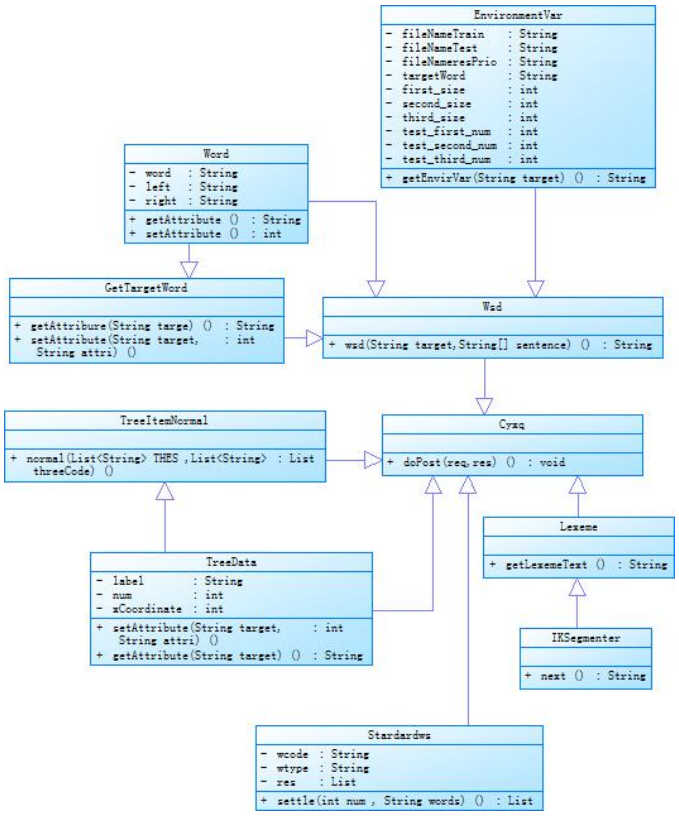

Fig. 3. The class diagram of WSD system

In Figure 3, Wsd class is the main class to train bayesian classifier. Word class and EnvironmentVar class are basic data structure for Wsd class. Word class stores basic information of ambiguous words and their disambiguation features. EnvironmentVar class stores global varies of Wsd class. GetTargetWord class is used to initial Word class and offers information for Wsd class. Cyxq class is a main class in WSD

system. It receives information input from interface and integrates all other classes. It sends the result of word segmentation and disambiguation to the interface. TreeData class is a basic data structure which gives semantic information of an ambiguous word. The purpose is to describe all semantic categories in a tree structure. IKSegmenter class and Lexeme class are two main classes in IKanalysis tool which receive a sentence and send the result of word segmentation to Cyxq class. TreeItemNormal class is utilized to initial TreeData class and offers information for Cyxq class. Stardardws class stores tree structure information of semantic categories.

This WSD system mainly contains two parts including browser program and server program. The former is used to interact with users. Html, jsp, ajax and javascript technologies are adopted to develop browser program. They are widely applicable technologies in browser program. The latter implements main logic function. IKanalysis is selected as a word segmentation tool in this system. Servlet technology is employed to develop program running on web server. It processes data submitted by user, and creates web contents dynamically.

Ambiguous word 'Tiao' is adopted to evaluate the performance of WSD system. It consists of three semantic categories 'CaiJi_SouLuo', 'Tiao_Tai_Bei' and 'Wa_Ti'. Sentences including word 'Tiao' are disambiguated by this WSD system and the accuracy is shown in Table 1.

TABLE I. THE DISAMBIGUATION ACCURACY

Number of instances	***Number of correct instances***	***Accuracy(%)***
14	7	50%

Table 1 describes that the accuracy of WSD system is 50%.

V. CONCLUSIONS

A WSD system is built in this paper. Bayesian theory and neighboring knowledge are adopted to construct the classifier. Sentences are segmented into several words by a word segmentation tool and those words are viewed as discriminative features. B/S framework is used to construct this WSD system. Experiments are conducted to test the performance of this system.

VI. ACKNOWLEDGMENTS

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

VII. REFERENCES

[1] Udaya Raj Dhungana, Subarna Shakya, "Word sense disambiguation using WSD specific WordNet of polysemy words", Proceedings of the 9th IEEE International Conference on Semantic Computing, (2015), pp. 148–52

[2] Lokesh Nandanwar, "Graph connectivity for unsupervised word sense disambiguation for HINDI language", Proceedings of the 2nd International Conference on Innovations in Information Embedded and Communication systems, (2015), pp. 1–4

[3] Gad El-Rab Wessam, "Unsupervised graph-based word sense disambiguation of biomedical documents", Proceedings of the 15th International Conference on e-Health Networking, Applications and Services, (2013) pp. 649–652

[4] Cem Akkaya, Janyce Wiebe, Rada Mihalcea, "Utilizing semantic composition in distributional semantic models for word sense discrimination and word sense disambiguation", Proceedings of the 6th IEEE International Conference on Semantic Computing, (2012), pp. 45–51

[5] Ting Liu, Zhimao Lu, Jun Lang, Sheng Li, "Chinese word sense disambiguation based on neural networks", Journal of Harbin Institute of Technology, vol. 12, no. 4, (2005), pp. 408–414

[6] Yeqing Li, Xiaoyu Qiu, "Word sense disambiguation based on feature ranking graph", Proceedings of the 29th International Conference on Advanced Information Networking and Applications Workshops, (2015), pp. 209–212

[7] Menai Mohamed El Bachir, "Word sense disambiguation using an evolutionary approach", Informatica, vol. 38, no. 2, (2014), pp. 155–169

[8] Manasi Kulkarni, Suneeta Sane, "An ontology clarification tool for word sense disambiguation", Proceedings of the 3rd International Conference on Electronics Computer Technology, (2011), pp. 292–296

[9] Bartosz Broda, Maciej Piasecki, "Semi-supervised word sense disambiguation based on weakly controlled sense induction", Proceedings of International Multiconference on Computer Science and Information Technology, (2009), pp. 17–24

[10] Roberto Navigli, Paola Velardi, "Structural semantic interconnections: a knowledge-based approach to word sense disambiguation", Institute of Electrical and Electronics Engineers Computer Society, vol. 27, no. 7, (2005), pp. 1075–1086