International Journal of Database Theory and Application
Vol.7, No.6 (2014), pp.203-210
http://dx.doi.org/10.14257/ijdta.2014.7.6.18

Disambiguate Chinese Word Sense Based on Linguistics Knowledge

Chun-Xiang Zhang[1], Long Deng[2], Xue-Yao Gao[2] and Zhi-Mao Lu[3]

[1]School of Software, Harbin University of Science and Technology, Harbin 150080, China
[2]College 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

z6c6x666@163.com

Abstract

Word sense disambiguation (WSD) is important to many application problems in natural language processing fields, such as machine translation, parsing analysis and information retrieval. In this paper, we propose a new method to determine correct sense categories of Chinese words based on linguistics knowledge. The left word string and the right word string around the ambiguous word are respectively analyzed. Their syntactic structures are obtained for determining its intended sense. Syntactic category and part of speech are extracted as disambiguation features. A naive bayesian model is used as the classifier. Experimental results showed that the accuracy rate of classification arrives at 64%. The performance of disambiguation is improved.

Keywords: *natural language processing; sense categories; ambiguous word; bayesian model*

1. Introduction

Polysemy is an inherent property of natural language and the correct sense is dependent on its contexts. The task of word sense disambiguation is to automatically choose the intended sense of an ambiguous word in a given context [1]. According to the disambiguation knowledge, disambiguation methods can be divided into two categories. One is the statistical method and the other is the knowledge-based one. Dhillon proposes a novel approach to incorporate a feature relevance prior for each word. He transfers the knowledge from similar words to learn this prior over features, from which the higher accuracy model is gotten [2]. Pedersen presents a corpus-based method for word sense disambiguation that builds an ensemble of naive bayesian classifiers, each of which is based on lexical features that represent co-occurring words in varying size windows of contexts [3]. Guo presents a system that combines evidence from a monolingual WSD system together with that from a multilingual WSD system to yield state of the art performance on standard all-words data sets [4]. The monolingual system is based on a modification of the graph-based algorithm. Izquierdo explores on the performance of abstraction and sense grouping [5]. Experiments demonstrate that base concepts are able to group word senses into an adequate medium level of abstraction to perform the supervised class-based disambiguation and the semantic classes provide rich information about polysemous words. Ponzetto presents a method to automatically extend WordNet with large amounts of semantic relations from an

ISSN: 2005-4270 IJDTA

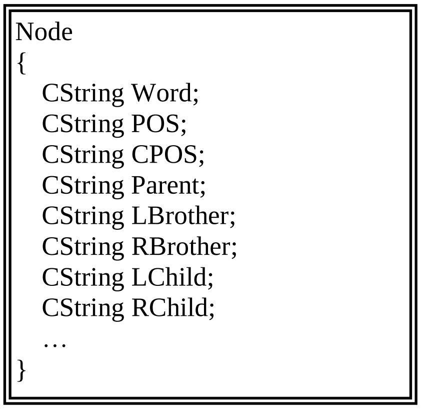
encyclopedia Wikipedia [6]. Experimental results show that knowledge-lean disambiguation algorithms compete with state-of-the-art supervised WSD systems in a coarse-grained all-words set when it is provided with a vast amount of high-quality semantic relations. Zhong presents a supervised English all-words WSD system IMS [7]. The flexible framework of IMS allows users to integrate different preprocessing tools, additional features and different classifiers. Experiments demonstrate that IMS achieves state-of-the-art results. Cruys presents a unified model for the induction of word senses from text and the subsequent disambiguation of particular word instances with the automatically extracted sense inventory [8]. The induction step and the disambiguation step are based on the following principle: words and contexts are mapped to a limited number of topical dimensions in a latent semantic word space. Khapra conducts many experiments on supervised approaches and unsupervised ones[9]. Experimental results show that if there is any sense marked corpora, a small amount of annotation in any other domain can deliver the goods as if exhaustive sense marking is available in that domain. Tanigaki proposes a novel smoothing model with a combinatorial optimization scheme for word sense disambiguation from untagged corpora[10]. He introduces a smoothing method in context-sense space to cope with data sparsity from a large variety of linguistic context and sense.

In this paper, we analyze phrasal structures of contexts around an ambiguous word and put forward a new disambiguation model in which linguistics knowledge is fused. Nodes related closely to the ambiguous word are found in a parsing tree. At the same time, the syntactic information and the part-of-speech information of these nodes are extracted as disambiguation features. The bayesian classification decision method is adopted, and the performance of word sense disambiguation is improved.

2. Extract Disambiguation Features from Parsing Tree

A Chinese sentence which contains the ambiguity word is segmented into words firstly. Secondly the sentence is split into phrase structures and each phrase is marked with its syntactic category. Thirdly, its parsing tree is established according to the hierarchical relationships. Then, the parsing tree is traveled. The syntactic information and part of speech information in its father node, its left brother node and its right brother nodes are acquired as disambiguation features.

In order to obtain disambiguation features conveniently from the parsing tree, every Chinese word is stored in data structure Node. Data structure Node is shown in Figure 1.



```
Node
{
   CString Word;
   CString POS;
   CString CPOS;
   CString Parent;
   CString LBrother;
   CString RBrother;
   CString LChild;
   CString RChild;
   …
}
```

Figure 1. Data Structure of Node

Node contains the following properties. Word stands for the Chinese word. POS stores its part of speech. Phrase is the syntactic category of the phrase. CPOS stands for part of speech in its core node. Parent points to its parent node. LBrother points to its left-brother node. RBrother points to its right-brother node. LChild points to the most left child node. RChild points to the most right child node. For Chinese sentence CS and its parsing tree TS, the process of extracting disambiguation features for ambiguous word *w* is shown in Figure 2.

```
Node* Find_WNode(Node *pNode, CString w)
{
   if(pNode==NULL)
      return NULL;
   if(!pNode->IsLeafNode()){
      if(pNode->Word==w)
         return pNode;
   }
   else{
      if(pNode->RBrother!=NULL)
         Find_WNode(pNode->RBrother, w);
   }
}
CString Feature_Extraction(Node *TS, CString w)
{
   Node *p=Find_WNode(TS, w);
   CString f="";
   if(p!=NULL){
      if(!p->LBrother->IsLeafNode())
         f="LS="+p->LBrother->POS;
      else
         f="LP="+p->LBrother->Phrase+"LCS="+p->LBrother->CPOS;
      if(!p->RBrother->IsLeafNode())
         f=f+"RS="+p->RBrother->POS;
      else
         f=f+"RP="+p->RBrother->Phrase+"RCS="+p->RBrother->CPOS;
      f=f+"PP="+p->Parent->Phrase+"PCS="+p->Parent->CPOS;
   }
   return f;
}
```

Figure 2. The Algorithm of Extracting Disambiguation Features

For Chinese sentence CS which contains ambiguous word 'shuo', the process of acquiring disambiguation features is shown as follows:

Chinese sentence CS: Ta/ yi/ shuo/ dao/ zhe/ jian/ shi er/, jiu/ kai shi/ ku/ ./

Syntactic analysis results of CS: S[SS[Ta/r VP[yi/d VO[shuo/vg VO[dao/vg BNP[BMP[zhe/r jian/m] shi er/ng]]]]] ,/wo VP[jiu/d VP[kai shi/vg ku/vg]] ./wj]

Here, '[' and ']' are paired with each other. They are utilized together to describe syntactic structures of a Chinese sentence. The label before the '[' stands for syntactic category of a phrase, such as S, SS, VO, BNP, BMP and VP. The label after the '/' is part of speech of a Chinese word such as r, vg, ng, d, m, wo and wj. Each pair '[' and ']' describes a phrase structure, and also shows a hierarchical structure of a sentence. From outside to inside, each pair of '[' and ']' extends downward as an independence node to form a tree, namely the parsing tree. The parsing tree of Chinese sentence CS is shown in Figure 3.

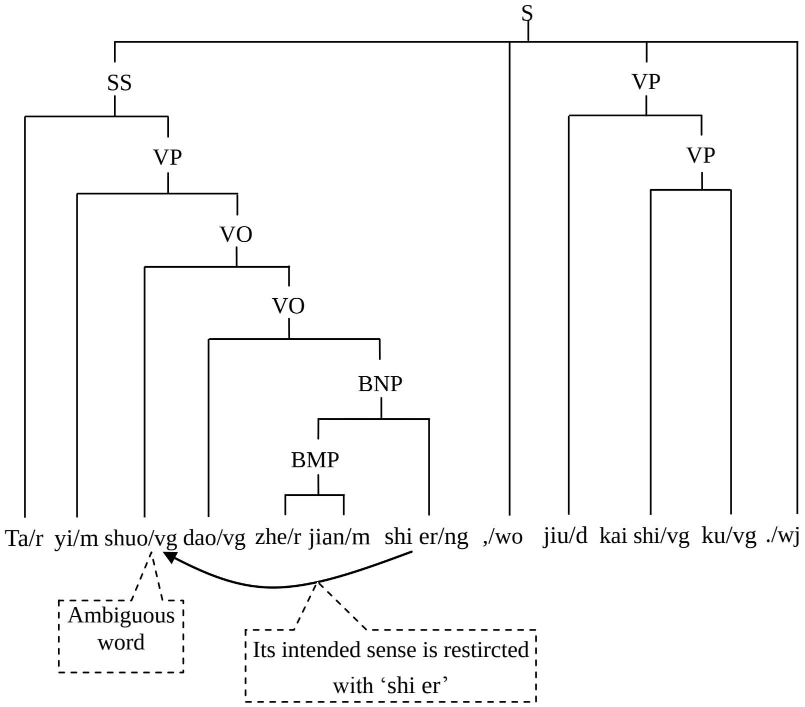


Figure 3. Parsing Tree of Chinese Sentence

In Chinese, 'shuo' is an ambiguous word. It contains four meanings in TongYiCi CiLin which is a Chinese semantic lexicon. The first sense category is Hi12, and its Chinese synonyms are 'speak' and 'chat'. The second sense category is Hg12, and its Chinese synonyms are 'explain', 'note', 'quote' and 'consult'. The third sense category is Hi16, and its Chinese synonyms are 'introduce' and 'recommend'. The fourth sense category is Hi34, its Chinese synonyms are 'encourage' and 'advise'.

From Figure 3, we can see that the sense category of Chinese word 'shuo' should be Hi12. If we open a linear window, the left and right adjacent unit of 'shuo' are respectively 'yi/m' and 'dao/vg'. If we use part of speech to determine the sense of 'shuo', disambiguation features are m and vg. From the grammatical perspective, the meaning of 'shuo' is influenced directly by word unit 'shi er/ng'. The word unit 'shi er/ng' is far from the 'shuo/vg' and its correct meaning can not be determined in the linear window. The parsing tree provides a tree window for extracting disambiguation information and we can obtain effective disambiguation features from the tree window. Using the proposed feature extraction method, we can acquire syntactic category VO of its right brother node. The core node of its right brother is word unit 'shi er/ng', and the core part of speech of its right brother node is ng. The syntactic category of its parent node is VP, and part of speech of its core node is ng. In parsing tree, the left brother node of 'shuo' is empty. We extract part of speech of the left adjacent word unit 'yi/m'. The disambiguation features obtained from the tree window are m, VO, ng, VP and ng. We can see that disambiguation features extracted from the tree window are more reasonable for determining the correct sense of ambiguous word 'shuo'.

3. Disambiguation classifier for determining sense

A naive bayesian classifier is a simple probabilistic model based on applying bayes theorem with strong independence assumptions. Bayesian classifier is appropriate to single point classification problems. Single point classification refers that the symbol corresponding condition in the state has nothing to do with other symbols corresponding condition in the state. It is a statistical system that tries to quantify the tradeoff between various decisions with probabilities and costs. In simple terms, a naive bayesian classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For some types of probability models, naive bayesian classifiers can be trained very efficiently in a supervised learning setting. In process of estimating parameters of naive bayesian models, the method of maximum likelihood is used. An advantage of the bayesian classifier is that it only requires a small amount of training data to estimate parameters and it is necessary for classification.

In the given context, the probability that the correct sense appears should be maximal. Here, ambiguous word *w* has n sense categories. They are s_1, s_2, …, s_n. The context that *w* appears is denoted as *w_context*. In *w_context*, there are three kinds of disambiguation features. They are respectively disambiguation features F_L in its left brother node, disambiguation features F_R in its right brother node and disambiguation features F_P in its father node. The discrimination process of ambiguous word *w* is shown in formula (1).

$$\begin{aligned} S &= \arg\max_{i=1,2,\ldots,n} P(s_i \mid w\_context) \\ &= \arg\max_{i=1,2,\ldots,n} \frac{P(s_i, w\_context)}{P(w\_context)} \\ &\approx \arg\max_{i=1,2,\ldots,n} P(s_i, w\_context) \\ &= \arg\max_{i=1,2,\ldots,n} P(w\_context \mid s_i) \cdot P(s_i) \\ &= \arg\max_{i=1,2,\ldots,n} P(F_L, F_R, F_P \mid s_i) \cdot P(s_i) \\ &\approx \arg\max_{i=1,2,\ldots,n} P(F_L \mid s_i) \cdot P(F_R \mid s_i), \cdot P(F_P \mid s_i) \cdot P(s_i) \end{aligned} \tag{1}$$

If the left brother is a leaf node, F_L is its part of speech. If the left brother is not a leaf node, F_L is its syntactic category and core part of speech. If the right brother is a leaf node, F_R is its part of speech. If the right brother is not a leaf node, F_R is its syntactic category and core part of speech. F_P includes syntactic category and core part of speech. The process of solving the bayesian model follows the decision at a given error rate minimum principle.

4. Experiment

In order to measure the performance of the proposed method, 105 Chinese sentences including ambiguous words are collected in this paper. Firstly, Chinese word segmentation tool is used to segment every Chinese sentence into a sequence of words. Two human annotators check the segmentation results manually. Secondly, Chinese part-of-speech tagging tool is applied to mark each word's part of speech. Two human annotators check the tagging results manually. Thirdly, the syntactic analysis tool is used to build phrase structures of Chinese sentences. Two human annotators check every parsing tree manually.

These two human annotators label semantic catagories of ambiguous words by hand according to Chinese semantic lexicon TongYiCi CiLin. These sentences are divided into two parts. One is the training data set, and the other is the test data set. The training data consists of 80 sentences and there are 25 sentences in test data.

In order to compare the performance of the proposed method, two experiments are conducted. In experiment 1, the linear window is utilized for the extraction of disambiguation features and the bayesian classifier is used for determining correct senses of ambiguous words. In experiment 2, the tree window is applied to extract disambiguation features and formula (1) is used for determining correct senses of ambiguous words.

The training data is utilized to train the classifiers respectively in these two groups of experiments. Then the optimized classifiers are respectively used to classify the test data. Experimental results are shown in Table 1.

Table 1. Accuracy Rates of Disambiguation in Two Experiments

	Accuracy rate
Experiment 1	60%
Experiment 2	64%

From table 1, we can see that the accuracy rate of experiment 2 is higher than that of experiment 1 and it reaches 64%. The reason is that syntactic information is used to guide the disambiguation process in experiment 2 and the extracted disambiguation features have a better performance.

5. Conclusions

In this paper, linguistics knowledge is introduced into the model of word sense disambiguation. Chinese sentences are segmented into words, and every word is tagged with part of speech and parsing trees are built for Chinese sentences. We extract disambiguation features from the parsing tree and a new method of determining correct senses of ambiguous words based on the naive bayesian classifier is given. The corpus tagged with sense categories is used to train the classifier and the optimized classifier is applied to determine its correct sense. Comparative experiments show that the performance of word sense disambiguation is improved.

Acknowledgements

This work is supported by Science and Technology Research Funds of Education Department in Helongjiang Province under Grant Nos. 12531106.

References

[1] P. F. Brown, “Word sense disambiguation using statistical methods”, Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, **(1991)**; California.

[2] P. S. Dhillon, “Transfer learning, feature selection and word sense disambiguation”, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, **(2009)**, Singapore.

[3] T. Pedersen, “A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation”, Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, **(2000)**; Seattle.

[4] 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, **(2010)**; Sweden.

[5] R. Izquierdo and A. Suárez, “An empirical study on class-based word sense disambiguation”, Proceedings of the 12th Conference of the European Chapter of the ACL, **(2009)**; Greece.

[6] 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)**; Sweden.

[7] Z. Zhong and H. T. Ng, "It makes sense: a wide-coverage word sense disambiguation system for free text", Proceedings of the ACL System Demonstrations, **(2010)**; Sweden.

[8] T. V. D. Cruys and M. Apidianaki, "Latent semantic word sense induction and disambiguation", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, **(2011)**; Oregon.

[9] M. M. Khapra and A. Kulkarni, "All words domain adapted WSD: finding a middle ground between supervision and unsupervision", Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, **(2010)**; Sweden.

[10] K. Tanigaki and M. Shiba, "Density maximization in context-sense metric space for all-words WSD", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, **(2013)**; Sofia.