Chinese Word Sense Disambiguation based on Context Expansion

Zhizhuo Yang [1,2] *Heyan Huang* [1,2]
(1) Department of Computer Science, Beijing Institute of Technology, No.5 Yard, Zhong Guan Cun South Street Haidian District, Beijing, 100081,ChinaNo.5 Yard, Zhong Guan Cun South Street Haidian District, Beijing, 100081,China
(2) Beijing Engineering Applications Research Center of High Volume Language Information Processing and Cloud Computing, Beijing Institute of Technology
10907029@bit.edu.cn, hhy@bit.edu.cn

ABSTRACT

Word Sense Disambiguation (WSD) is one of the key issues in natural language processing. Currently, supervised WSD methods are effective ways to solve the ambiguity problem. However, due to lacking of large-scale training data, they cannot achieve satisfactory results. In this paper, we suppose synonyms for context words that can provide more knowledge for WSD task, and present two different WSD methods based on context expansion. The first method regards Synonyms as topic contextual feature to train Bayesian model. The second method treats context words made up of synonyms as pseudo training data, and then derives the meaning of ambiguous words using the knowledge from both training and pseudo training data. Experimental results show that the second method can significantly improve traditional WSD accuracy by 2.21%. Furthermore, it also outperforms the best system in SemEval-2007.

KEYWORDS: Data sparseness, Context expansion, Bayesian model, Synonym, Parameter estimation

Proceedings of COLING 2012: Posters, pages 1401–1408,
COLING 2012, Mumbai, December 2012.

1. Introduction

Word Sense Disambiguation (WSD), the task of identifying the intended meaning (sense) of words in a given context is one of the most important problem in natural language processing. Various approaches have been proposed to deal with the WSD problem. Hwee found that the supervised machine learning methods are the most successful approach to WSD when contextual features have been used to distinguish ambiguous words in these methods (Hwee and Bin Wang, 2003). However, word occurrences in the context are too diverse to capture the correct pattern, which means that the dimension of contextual words will be very large when all words are used in robust WSD system. It has been proved that expanding context window size around the target ambiguous word can help to enhance the WSD performance. However, expanding window size unboundedly will bring not only useful information but also some noise which may deteriorate the WSD performance. Can we find another way to expand context words without bringing too much noise?

In this paper, we propose to conduct WSD based on context expansion, which acquires WSD knowledge from synonymy dictionary. The assumption of our approach is that contextual words around ambiguous word can be substituted by synonymy, and the new context represented by synonymy expresses the same meaning, thus the sense of the ambiguous word in new context remains unchanged. Therefore, the new context can provide more knowledge for us to improving WSD performance. Under this assumption, we propose two methods to integrate the contribution of synonymy into supervised WSD model. The first method directly considers synonymy as contextual feature, and exploit synonymy feature to train supervised WSD model. The second method treats the new context represented by synonymy as pseudo training data. In the method, the pseudo and authentic training data are both utilized to train supervised model. Consequently, the sense of ambiguous word is not only determined by authentic training data, but also pseudo training data. Experiments are carried out on dataset and the results confirm the effectiveness of our approach. The synonym for context word can significantly improve the performance of WSD.

The rest of paper is organized as follows: Section 2 briefly introduces the related work. The proposed method is described in detail in Section 3, and experimental results are presented in Section 4. Lastly we conclude this paper in Section 5.

2. Related Work

Generally speaking, Word Sense Disambiguation methods are either knowledge-based or corpus-based. In addition, the latter can further be further divided into two kinds: unsupervised ones and supervised ones. In this paper we focus on supervised WSD method.

More recently, WSD approaches based on pseudo word have gained much attention in the NLP community (Yarowsky, 1995; Leacock et al, 1998; Mihalcea and Moldovan, 1999; Agirre and Marinez, 2004; Brody and Lapata, 2008; Lu et al, 2006). The pseudo words can simulate the function of the real ambiguous words. In most cases, synonyms are used as pseudo words to acquire semantic knowledge as the real ambiguous word does. Specifically, These approaches exploits a sense inventory such as WordNet or corpus to collect pseudo words for ambiguous words, and use pseudo words to automatically create sense label data which can subsequently serve to train any supervised classifier. Such approaches are often regarded as weakly supervised learning or semi-supervised learning methods. Inspired by these approaches, we use synonymy of context to provide more knowledge for WSD task. Different from previous approach, we generate training data from another perspective. Instead of utilizing synonymy of ambiguous word to acquire instance form corpus, context words around ambiguous word are extended by synonymy to produce training data in our method 2. Moreover, the method 2 and previous pseudo words based approach can be applied simultaneously in WSD task.

3. Proposed Approach

3.1 The Bayesian Classifier

Naïve Bayesian model have been widely used in most classification task, and was first used in WSD by Gale et al. The classifier works under the assumption that all the feature variables are conditionally independent given the classes. For word sense disambiguation, the context in which an ambiguous word occurs is represented by a vector of feature variables $F = \{f_1, f_2, ..., f_n\}$. The sense of ambiguous word is represented by variables $S = \{s_1, s_2, ..., s_n\}$. Finding the right sense of the ambiguous word equal to choosing the sense s' that maximizes the conditional probability as follow:

$$s' = \arg\max_{s_i \in S} \prod_{f_j \in F} P(f_j \mid s_i) P(s_i) \tag{1}$$

The probability of sense $P(s_i)$ and the conditional probability of feature f_j with observation of sense $P(f_j \mid s_i)$ are computed via Maximum-Likelihood Estimation:

$$P(s_i) = \frac{C(s_i)}{\sum_{s_i \in S} C(s_i)} \tag{2}$$

$$P(f_j \mid s_i) = \frac{C(f_j, s_i)}{C(s_i)} \tag{3}$$

where $C(s_i)$ is the number of sense s_i that appears in training corpus. $C(f_j, s_i)$ is the number of occurrences of feature f_j in context with sense s_i in the training corpus. We use “add one” data smooth strategies to avoid data sparse problem when estimating the conditional probabilities of the model.

3.2 WSD Methods based on Context Expansion

Synonyms are different words with almost identical or similar meanings. In this paper, we extend context around ambiguous word into a larger dimension using synonyms, which could provide more knowledge and clues for WSD task. In the previous study of WSD, the most widely used assumption is that words of the same meaning usually play the same role in a language. The assumption can be further extended as words of the same meaning often occur in similar context. Base on the above assumptions, we propose basic assumption in this study, synonyms of context express similar meaning to that of original context, thus the sense of ambiguous word appear in the two similar contexts remains unchanged. For example, in Chinese sentence "可以使消费者清楚地知道自己的钱花在何处" (makes consumers to clearly know where your money goes), the target ambiguous word is "使", it has two meanings as a verb in HowNet (Dong, 2000) which are "make" and "use". We can easily infer the meaning of ambiguous word as "make" based on the context. After word segmentation, the context around ambiguous word can be expanded into synonyms set as figure 1. Since the context nearby ambiguous word has the largest impact to the sense of ambiguous word, only three contextual word "可以，消费者，清楚地" are listed and expanded with synonyms in the figure. We simply expand each contextual word with only four synonyms in the figure, actually more synonyms could be added into the synonyms set. Given ambiguous word and synonyms set for each contextual word, some reliable training data could be generated. For example, "可使顾主明白地",“足以使购买者明晰地" and "得以使买主清晰地", etc. The ambiguous word "使" express the same meaning "make" in all of these training examples. It is obvious that synonyms provide additional knowledge for training model, and the knowledge can be exploited to improve the WSD performance.

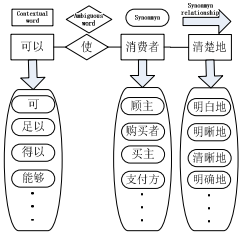

FIGURE 1 – WSD Method base on Context Expansion

The first method we proposed is that treating these expanded synonyms as topic feature. Then, we use these features together with other features to train the classifier. The method is quite straightforward. If contextual words near ambiguous word appear once in training data, synonyms of these contextual words are supposed to appear once in the corpus at the same time. For example, in the previous example, if the testing instance contain contextual word "可" or "顾主", it is likely that the sense of

ambiguous word “使” could be inferred as “make” by Bayesian classifier. But it should be noted that the method has its own shortcomings. The authentic training data is labeled by human while the training data which consists of synonyms is generated automatically by machine. Thus the latter training data contain some noise compared to former training data, and should not play the same role while deducing the sense of the ambiguous word.

In order to overcome the disadvantage of method 1, we proposed method 2. The data which consist of synonyms are regarded as pseudo training data in method 2. The pseudo and authentic training data are both utilized to train the classifier. Instead of using formula (3) to compute the conditional probability of feature with sense, we apply follow formula to compute $P(f_j \mid s_i)$:

$$P(f_j \mid s_i) = \frac{C_a(f_j, s_i)}{C_a(s_i)} + \lambda \frac{C_p(f_j, s_i)}{C_p(s_i)} \tag{4}$$

Here, $C_a(s_i)$ and $C_p(s_i)$ are the number of sense s_i that appears in authentic and pseudo training corpus respectively. $C_a(f_j, s_i)$ and $C_p(f_j, s_i)$ are the number of occurrences of feature f_j with sense s_i in authentic and pseudo training corpus respectively. Parameter λ adjusts the influence of two different kinds of training data. We can set λ to a larger value to let the pseudo training data play a stronger role, and vice versa. In the model, pseudo training data always play a lesser role to determine the sense of ambiguous word. Furthermore, we can set different value to λ for different kinds of ambiguous word.

We encounter two problems when expanding the contextual word with synonyms. The first problem is that not all the synonyms are suitable for generating training data. For example, contextual word “清楚” has synonyms such as “清晰”,“明晰”,“历历” and “不可磨灭” in dictionary. It is obvious that “历历” and “不可磨灭” should not be added into the expanded synonyms set, since the collocations of those synonyms with ambiguous word are rarely occur in large-scale corpus. In addition, the contextual words are not monotonous in most cases, and we do not know which sense of the ambiguous word should be expanded by synonyms. For example , Chinese word “可以” has three meanings in dictionary. They are “不错”,“认可” and “可” respectively. Which sense should be expanded by synonyms in order to generate appropriate training data? To solve the above problems, we exploit word collocation relationship to restrict expansion of synonyms, i.e., only synonyms co-occurrence with ambiguous word that exceeded a certain number are used to train classifier. This strategy can not only filter out uncommonly used collocations, but also solve the problem of noise caused by ambiguity of contextual word. The collocation parameter threshold $threshold\_cooc$ will be adjusted in the experiment.

4. Experiment

4.1 Experimental Setup

Synonyms dictionary: The extended TongYiCiCiLin[1] which was developed by HIT-SCIR is applied to look up synonyms. The items in Cilin are organized as hierarchy of five levels. From the root level to the leaf level, the lower the level is, the more specific the sense is. Since the words in fifth level have similar sense and linguistic function, they can be substituted for each other without changing the meaning of the sentence.

Collocation relationship: In the experiment, Sogou Chinese collocation relation[2]was used to filter out uncommonly used collocations. The collocation corpus involves more than 20 million collocation relations and more than 15000 high-frequency words, which was extracted from over 100 million internet pages on web in October 2006.

Training and testing data: In SemEval-2007, the 4th international workshop on semantic evaluations under conference of ACL-2007 (Jin et al, 2007), we used task#5 multilingual Chinese English lexical sample to test our methods. Macro-average precision (Liu et al., 2007) was used to evaluate word sense disambiguation performance.

Since we aim to evaluate discriminating power of synonymy feature, in the experiment, only some basic features such as topic words, collocations, and words assigned with their positions were used. We compare two baseline methods with our methods, the two baseline methods are as follows:

(1) Original: WSD method based on traditional Bayesian Classifier.

(2) SRCP_WSD (Xing Y, 2007): The system participated in semeval-2007 and won the first place in multilingual Chinese English lexical sample task.($p_{mar} = 74.9\%$)

Our methods:

(1) Method_1: The first method we proposed. This method was based on traditional Bayesian classifiers, which use synonym feature and basic features to train model.

(2) Method_2: The second method we proposed. This method was also based on Bayesian classifiers, which use Basic features to train model. But this method computed the conditional probability using formula (4) .

4.2 Evaluation Results

Because not all words in the sentence are useful for WSD, the contextual words are restricted by syntactic filters, i.e., only the words with a certain part of speech are added.

(1) In order to compare the performances of various methods, table 1 gives the average precision of four methods. It can be seen that method_1 and method_2 obtain improvement over original method, which shows that the methods we propose are

[1] It is located at http://ir.hit.edu.cn/.

2 http://www.sogou.com/labs/dl/r.html

effective. Moreover, method_2 also outperforms the best system participated in SemEval-2007.

	Original	Method_1	SRCP_WSD	Method_2
Average precision (p_{mar})	0.7336	0.7447	0.7490	**0.7557**
Improving performance (%)	0	1.11	1.54	2.21

TABLE 1 –Experimental result of 4 methods

(3) In order to investigate how the threshold of co-occurrences number influence the performance, experiment on different *threshold _cooc* was conducted, and the results are shown in Figure 2. The figure shows the curves for two methods when *threshold _cooc* ranges from 5 to 35. We can see that the performance of the two methods first increases and then decreases with the increase of *threshold _cooc* . The trend demonstrates that extremely small or large co-occurrences number will deteriorate the results. Because a small number means that too many synonyms co-occur with ambiguous word and the number of synonyms exceeds the number that are used to train the classifier. These synonyms introduce noisy knowledge. On the other hand, a large number means very few synonyms are used to train the classifier and this cannot provide sufficient knowledge. The best performance was achieved when set *threshold _cooc* to 25.

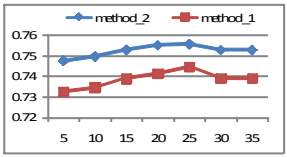

FIGURE 2 – Comparison result of different *threshold _cooc*

(4) In order to investigate how the λ parameter in formula (4) influences the performance, we conduct experiment with different value of λ as shown in figure 3. In this experiment, we set different λ to different values for ambiguous nouns and verbs contained in testing instance. The red curve represent verb while blue curve represent noun. We can see from the figure, the best experimental results were achieved when λ is set to (0, 1), and the optimal value of λ for noun and verb were set to 0.8 and 0.5 respectively. Because the collocation relationship between noun ambiguous and contextual word would be different with that relationship between verb ambiguous and contextual word, the synonym of contextual word should have larger impact on ambiguous nouns and smaller impact on ambiguous verbs.

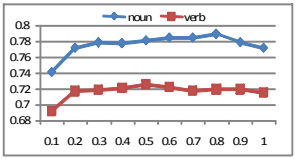

FIGURE 3 – Comparison result of different λ_noun and λ_verb using method_2

Conclusion and perspectives

In this paper, we proposed two novel methods for supervised word sense disambiguation by leveraging synonym for context around ambiguous word. The experimental results on dataset demonstrate the effectiveness of our methods. In current study, we search synonym by extended TongYiCiCiLin. In future work, we will retrieve more synonyms from HowNet and large-scale corpus to expand context nearby ambiguous word, attempting to further improve the performance of WSD.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No.61132009, the Primary Funds of National Defense.

References

Hwee Tou Ng, Bin Wang, Yee Seng Chan. (2003). Exploiting Parallel Texts for Word Sense Disambiguation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. pages 455--462.

David Yarowsky. (1995). Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA, June.

C. Leacock, M.Chodorow, and G.A. Miller. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147-166.

R. Mihalcea and I. Moldovan. (1999). A Method for Word Sense Disambiguation of Unrestricted Text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 152-158.

E. Agirre and D. Mart'inez. (2004). Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 25-32.

Samuel Brody and Mirella Lapata. (2008). Good Neighbors Make Good Senses: Exploiting Distributional Similarity for Unsupervised WSD. In *Proceedings of COLING*, pages 65-72.

Lu, Zhimao, Wang Haifeng and Yao Jianmin. (2006). An Equivalent Pseudoword Solution to Chinese Word Sense Disambiguation. In *Proceedings of the 44th annual meeting of the Association for Computational Linguistics*, pages 457-464.

Dong ZD, Dong Q. Hownet. 2000. http://keenage.com.

Peng Jin, YunfangWu, and Shiwen Yu. (2007). SemEval-2007 task 05: Multilingual Chinese-English lexical sample task. In: Eneko Agirre, ed. *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Prague, Czech Republic: Association for Computational Linguistics. pages 19-23.

Xing Y. (2007). SRCB-WSD : Supervised Chinese Word Sense Disambiguation with Key Features. In : *Proceedings of the 4th in Workshop on Semantic Evaluations*. pages 300-303.