

An Equivalent Pseudoword Solution to Chinese Word Sense Disambiguation

Zhimao Lu⁺ Haifeng Wang⁺⁺ Jianmin Yao⁺⁺⁺ Ting Liu⁺ Sheng Li⁺

⁺ Information Retrieval Laboratory, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, 150001, China

{lzm, tliu, lisheng}@ir-lab.org

⁺⁺ Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza, No. 1, East Chang An Ave., Beijing, 100738, China

wanghaifeng@rdc.toshiba.com.cn

⁺⁺⁺ School of Computer Science and Technology

Soochow University, Suzhou, 215006, China

jyao@suda.edu.cn

Abstract

This paper presents a new approach based on Equivalent Pseudowords (EPs) to tackle Word Sense Disambiguation (WSD) in Chinese language. EPs are particular artificial ambiguous words, which can be used to realize unsupervised WSD. A Bayesian classifier is implemented to test the efficacy of the EP solution on Senseval-3 Chinese test set. The performance is better than state-of-the-art results with an average F-measure of 0.80. The experiment verifies the value of EP for unsupervised WSD.

1 Introduction

Word sense disambiguation (WSD) has been a hot topic in natural language processing, which is to determine the sense of an ambiguous word in a specific context. It is an important technique for applications such as information retrieval, text mining, machine translation, text classification, automatic text summarization, and so on.

Statistical solutions to WSD acquire linguistic knowledge from the training corpus using machine learning technologies, and apply the knowledge to disambiguation. The first statistical model of WSD was built by Brown et al. (1991). Since then, most machine learning methods have been applied to WSD, including decision tree, Bayesian model, neural network, SVM, maxi-

mum entropy, genetic algorithms, and so on. For different learning methods, supervised methods usually achieve good performance at a cost of human tagging of training corpus. The precision improves with larger size of training corpus. Compared with supervised methods, unsupervised methods do not require tagged corpus, but the precision is usually lower than that of the supervised methods. Thus, knowledge acquisition is critical to WSD methods.

This paper proposes an unsupervised method based on equivalent pseudowords, which acquires WSD knowledge from raw corpus. This method first determines equivalent pseudowords for each ambiguous word, and then uses the equivalent pseudowords to replace the ambiguous word in the corpus. The advantage of this method is that it does not need parallel corpus or seed corpus for training. Thus, it can use a large-scale monolingual corpus for training to solve the data-sparseness problem. Experimental results show that our unsupervised method performs better than the supervised method.

The remainder of the paper is organized as follows. Section 2 summarizes the related work. Section 3 describes the conception of Equivalent Pseudoword. Section 4 describes EP-based Unsupervised WSD Method and the evaluation result. The last section concludes our approach.

2 Related Work

For supervised WSD methods, a knowledge acquisition bottleneck is to prepare the manually

tagged corpus. Unsupervised method is an alternative, which often involves automatic generation of tagged corpus, bilingual corpus alignment, etc. The value of unsupervised methods lies in the knowledge acquisition solutions they adopt.

2.1 Automatic Generation of Training Corpus

Automatic corpus tagging is a solution to WSD, which generates large-scale corpus from a small seed corpus. This is a weakly supervised learning or semi-supervised learning method. This reinforcement algorithm dates back to Gale et al. (1992a). Their investigation was based on a 6-word test set with 2 senses for each word.

Yarowsky (1994 and 1995), Mihalcea and Moldovan (2000), and Mihalcea (2002) have made further research to obtain large corpus of higher quality from an initial seed corpus. A semi-supervised method proposed by Niu et al. (2005) clustered untagged instances with tagged ones starting from a small seed corpus, which assumes that similar instances should have similar tags. Clustering was used instead of bootstrapping and was proved more efficient.

2.2 Method Based on Parallel Corpus

Parallel corpus is a solution to the bottleneck of knowledge acquisition. Ide et al. (2001 and 2002), Ng et al. (2003), and Diab (2003, 2004a, and 2004b) made research on the use of alignment for WSD.

Diab and Resnik (2002) investigated the feasibility of automatically annotating large amounts of data in parallel corpora using an unsupervised algorithm, making use of two languages simultaneously, only one of which has an available sense inventory. The results showed that word-level translation correspondences are a valuable source of information for sense disambiguation.

The method by Li and Li (2002) does not require parallel corpus. It avoids the alignment work and takes advantage of bilingual corpus.

In short, technology of automatic corpus tagging is based on the manually labeled corpus. That is to say, it still need human intervention and is not a completely unsupervised method. Large-scale parallel corpus; especially word-aligned corpus is highly unobtainable, which has limited the WSD methods based on parallel corpus.

3 Equivalent Pseudoword

This section describes how to obtain equivalent pseudowords without a seed corpus.

Monosemous words are unambiguous priori knowledge. According to our statistics, they account for 86%~89% of the instances in a dictionary and 50% of the items in running corpus, they are potential knowledge source for WSD.

A monosemous word is usually synonymous to some polysemous words. For example the words "信守, 严守, 恪守, 遵照, 遵从, 遵循, 遵守" has similar meaning as one of the senses of the ambiguous word "保守", while "康健, 强健, 健旺, 健壮, 壮健, 强壮, 精壮, 壮实, 敦实, 硬朗, 康泰, 健朗, 健硕" are the same for "健康". This is quite common in Chinese, which can be used as a knowledge source for WSD.

3.1 Definition of Equivalent Pseudoword

If the ambiguous words in the corpus are replaced with its synonymous monosemous word, then is it convenient to acquire knowledge from raw corpus? For example in table 1, the ambiguous word "把握" has three senses, whose synonymous monosemous words are listed on the right column. These synonyms contain some information for disambiguation task.

An artificial ambiguous word can be coined with the monosemous words in table 1. This process is similar to the use of general pseudowords (Gale et al., 1992b; Gaustad, 2001; Nakov and Hearst, 2003), but has some essential differences. This artificial ambiguous word need to simulate the function of the real ambiguous word, and to acquire semantic knowledge as the real ambiguous word does. Thus, we call it an *equivalent pseudoword* (EP) for its equivalence with the real ambiguous word. It's apparent that the equivalent pseudoword has provided a new way to unsupervised WSD.

把握(ba3 wo4)	S_1	信心/自信心
	S_2	握住/在握/把住/抓住/控制
	S_3	领会/理解/领悟/深谙/体会

Table 1. Synonymous Monosemous Words for the Ambiguous Word "把握"

The equivalence of the EP with the real ambiguous word is a kind of semantic synonym or similarity, which demands a maximum similarity between the two words. An ambiguous word has the same number of EPs as of senses. Each EP's sense maps to a sense of ambiguous word.

The semantic equivalence demands further equivalence at each sense level. Every corre-

sponding sense should have the maximum similarity, which is the strictest limit to the construction of an EP.

The starting point of unsupervised WSD based on EP is that EP can substitute the original word for knowledge acquisition in model training. Every instance of each morpheme of the EP can be viewed as an instance of the ambiguous word, thus the training set can be enlarged easily. EP is a solution to data sparseness for lack of human tagging in WSD.

3.2 Basic Assumption for EP-based WSD

It is based on the following assumptions that EPs can substitute the original ambiguous word for knowledge acquisition in WSD model training.

Assumption 1: Words of the same meaning play the same role in a language. The sense is an important attribute of a word. This plays as the basic assumption in this paper.

Assumption 2: Words of the same meaning occur in similar context. This assumption is widely used in semantic analysis and plays as a basis for much related research. For example, some researchers cluster the contexts of ambiguous words for WSD, which shows good performance (Schutze, 1998).

Because an EP has a higher similarity with the ambiguous word in syntax and semantics, it is a useful knowledge source for WSD.

3.3 Design and Construction of EPs

Because of the special characteristics of EPs, it's more difficult to construct an EP than a general pseudo word. To ensure the maximum similarity between the EP and the original ambiguous word, the following principles should be followed.

1) Every EP should map to one and only one original ambiguous word.

2) The morphemes of an EP should map one by one to those of the original ambiguous word.

3) The sense of the EP should be the same as the corresponding ambiguous word, or has the maximum similarity with the word.

4) The morpheme of a pseudoword stands for a sense, while the sense should consist of one or more morphemes.

5) The morpheme should be a monosemous word.

The fourth principle above is the biggest difference between the EP and a general pseudo word. The sense of an EP is composed of one or several morphemes. This is a remarkable feature

of the EP, which originates from its equivalent linguistic function with the original word. To construct the EP, it must be ensured that the sense of the EP maps to that of the original word. Usually, a candidate monosemous word for a morpheme stands for part of the linguistic function of the ambiguous word, thus we need to choose several morphemes to stand for one sense.

The relatedness of the senses refers to the similarity of the contexts of the original ambiguous word and its EP. The similarity between the words means that they serve as synonyms for each other. This principle demands that both semantic and pragmatic information should be taken into account in choosing a morpheme word.

3.4 Implementation of the EP-based Solution

An appropriate machine-readable dictionary is needed for construction of the EPs. A Chinese thesaurus is adopted and revised to meet this demand.

Extended Version of TongYiCiLin

To extend the TongYiCiLin (Cilin) to hold more words, several linguistic resources are adopted for manually adding new words. An extended version of the Cilin is achieved, which includes 77,343 items.

A hierarchy of three levels is organized in the extended Cilin for all items. Each node in the lowest level, called a minor class, contains several words of the same class. The words in one minor class are divided into several groups according to their sense similarity and relatedness, and each group is further divided into several lines, which can be viewed as the fifth level of the thesaurus. The 5-level hierarchy of the extended Cilin is shown in figure 1. The lower the level is, the more specific the sense is. The fifth level often contains a few words or only one word, which is called an atom word group, an atom class or an atom node. The words in the same atom node hold the smallest semantic distance.

From the root node to the leaf node, the sense is described more and more detailed, and the words in the same node are more and more related. Words in the same fifth level node have the same sense and linguistic function, which ensures that they can substitute for each other without leading to any change in the meaning of a sentence.

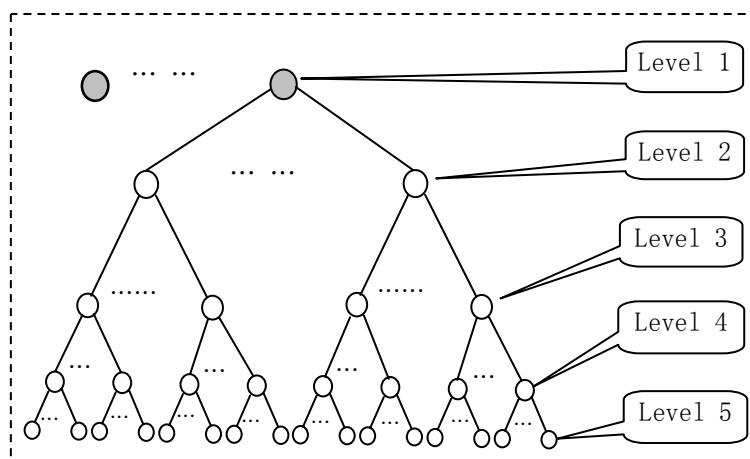


Figure 1. Organization of Cilin (extended)

The extended version of extended Cilin is freely downloadable from the Internet and has been used by over 20 organizations in the world¹.

Construction of EPs

According to the position of the ambiguous word, a proper word is selected as the morpheme of the EP. Almost every ambiguous word has its corresponding EP constructed in this way.

The first step is to decide the position of the ambiguous word starting from the leaf node of the tree structure. Words in the same leaf node are identical or similar in the linguistic function and word sense. Other words in the leaf node of the ambiguous word are called brother words of it. If there is a monosemous brother word, it can be taken as a candidate morpheme for the EP. If there does not exist such a brother word, trace to the fourth level. If there is still no monosemous brother word in the fourth level, trace to the third level. Because every node in the third level contains many words, candidate morpheme for the ambiguous can usually be found.

In most cases, candidate morphemes can be found at the fifth level. It is not often necessary to search to the fourth level, less to the third. According to our statistics, the extended Cilin contains about monosemous words for 93% of the ambiguous words in the fifth level, and 97% in the fourth level. There are only 112 ambiguous words left, which account for the other 3% and mainly are functional words. Some of the 3% words are rarely used, which cannot be found in even a large corpus. And words that lead to semantic misunderstanding are usually content words. In WSD research for English, only nouns, verbs, adjectives and adverbs are considered.

From this aspect, the extended version of Cilin meets our demand for the construction of EPs.

If many monosemous brother words are found in the fourth or third level, there are many candidate morphemes to choose from. A further selection is made based on calculation of sense similarity. More similar brother words are chosen.

Computing of EPs

Generally, several morpheme words are needed for better construction of an EP. We assume that every morpheme word stands for a specific sense and does not influence each other. It is more complex to construct an EP than a common pseudo word, and the formulation and statistical information are also different.

An EP is described as follows:

$$\begin{array}{l}
 W_{EP} \\
 \hline
 S_1 : W_{11}, W_{12}, W_{13}, \dots, W_{1k_1} \\
 S_2 : W_{21}, W_{22}, W_{23}, \dots, W_{2k_2} \\
 \vdots \\
 S_i : W_{i1}, W_{i2}, W_{i3}, \dots, W_{ik_i}
 \end{array}$$

Where W_{EP} is the EP word, S_i is a sense of the ambiguous word, and W_{ik} is a morpheme word of the EP.

The statistical information of the EP is calculated as follows:

1) $C(S_i)$ stands for the frequency of the S_i :

$$C(S_i) = \sum_k C(W_{ik})$$

2) $C(S_i, W_f)$ stands for the co-occurrence frequency of S_i and the contextual word W_f :

$$C(S_i, W_f) = \sum_k C(W_{ik}, W_f)$$

¹ It is located at <http://www.ir-lab.org/>.

Ambiguous word	citation (Qin and Wang, 2005)	Ours	Ambiguous word	citation (Qin and Wang, 2005)	Ours
把握(ba3 wo4)	0.56	0.87	没有(me1 you3)	0.75	0.68
包(bao1)	0.59	0.75	起来(qi3 lai2)	0.82	0.54
材料(cai2 liao4)	0.67	0.79	钱(qian2)	0.75	0.62
冲击(chong1 ji1)	0.62	0.69	日子(ri4 zi3)	0.75	0.68
穿(chuan1)	0.80	0.57	少(shao3)	0.69	0.56
地方(di4 fang1)	0.65	0.65	突出(tu1 chu1)	0.82	0.86
分子(fen1 zi3)	0.91	0.81	研究(yan2 jiu1)	0.69	0.63
运动(yun4 dong4)	0.61	0.82	活动(huo2 dong4)	0.79	0.88
老(lao3)	0.59	0.50	走(zou3)	0.72	0.60
路(lu4)	0.74	0.64	坐(zuo4)	0.90	0.73
Average	0.72	0.69	Note: Average of the 20 words		

Table 2. The F-measure for the Supervised WSD

4 EP-based Unsupervised WSD Method

EP is a solution to the semantic knowledge acquisition problem, and it does not limit the choice of statistical learning methods. All of the mathematical modeling methods can be applied to EP-based WSD methods. This section focuses on the application of the EP concept to WSD, and chooses Bayesian method for the classifier construction.

4.1 A Sense Classifier Based on the Bayesian Model

Because the model acquires knowledge from the EPs but not from the original ambiguous word, the method introduced here does not need human tagging of training corpus.

In the training stage for WSD, statistics of EPs and context words are obtained and stored in a database. Senseval-3 data set plus unsupervised learning method are adopted to investigate into the value of EP in WSD. To ensure the comparability of experiment results, a Bayesian classifier is used in the experiments.

Bayesian Classifier

Although the Bayesian classifier is simple, it is quite efficient, and it shows good performance on WSD.

The Bayesian classifier used in this paper is described in (1)

$$S(w_i) = \operatorname{argmax}_{S_k} \left[\log P(S_k) + \sum_{v_j \in c_i} \log P(v_j | S_k) \right] \quad (1)$$

Where w_i is the ambiguous word, $P(S_k)$ is the occurrence probability of the sense S_k , $P(v_j | S_k)$ is the conditional probability of the context word v_j , and c_i is the set of the context words.

To simplify the experiment process, the Naive Bayesian modeling is adopted for the sense classifier. Feature selection and ensemble classification are not applied, which is both to simplify the calculation and to prove the effect of EPs in WSD.

Experiment Setup and Results

The Senseval-3 Chinese ambiguous words are taken as the testing set, which includes 20 words, each with 2-8 senses. The data for the ambiguous words are divided into a training set and a testing set by a ratio of 2:1. There are 15-20 training instances for each sense of the words, and occurs by the same frequency in the training and test set.

Supervised WSD is first implemented using the Bayesian model on the Senseval-3 data set. With a context window of (-10, +10), the open test results are shown in table 2.

The F-measure in table 2 is defined in (2).

$$F = \frac{2 \times P \times R}{P + R} \quad (2)$$

Where P and R refer to the precision and recall of the sense tagging respectively, which are calculated as shown in (3) and (4)

$$P = \frac{C(\text{correct})}{C(\text{tagged})} \quad (3)$$

$$R = \frac{C(\text{correct})}{C(\text{all})} \quad (4)$$

Where $C(\text{tagged})$ is the number of tagged instances of senses, $C(\text{correct})$ is the number of correct tags, and $C(\text{all})$ is the number of tags in the gold standard set. Every sense of the ambiguous word has a P value, a R value and a F value. The F value in table 2 is a weighted average of all the senses.

In the EP-based unsupervised WSD experiment, a 100M corpus (People's Daily for year 1998) is used for the EP training instances. The Senseval-3 data is used for the test. In our experiments, a context window of (-10, +10) is taken. The detailed results are shown in table 3.

4.2 Experiment Analysis and Discussion

Experiment Evaluation Method

Two evaluation criteria are used in the experiments, which are the F-measure and precision. Precision is a usual criterion in WSD performance analysis. Only in recent years, the precision, recall, and F-measure are all taken to evaluate the WSD performance.

In this paper, we will only show the f-measure score because it is a combined score of precision and recall.

Result Analysis on Bayesian Supervised WSD Experiment

The experiment results in table 2 reveals that the results of supervised WSD and those of (Qin and Wang, 2005) are different. Although they are all based on the Bayesian model, Qin and Wang (2005) used an ensemble classifier. However, the difference of the average value is not remarkable.

As introduced above, in the supervised WSD experiment, the various senses of the instances are evenly distributed. The lower bound as Gale et al. (1992c) suggested should be very low and it is more difficult to disambiguate if there are more senses. The experiment verifies this reasoning, because the highest F-measure is less than 90%, and the lowest is less than 60%, averaging about 70%.

With the same number of senses and the same scale of training data, there is a big difference between the WSD results. This shows that other factors exist which influence the performance other than the number of senses and training data size. For example, the discriminability among the senses is an important factor. The WSD task becomes more difficult if the senses of the ambiguous word are more similar to each other.

Experiment Analysis of the EP-based WSD

The EP-based unsupervised method takes the same open test set as the supervised method. The unsupervised method shows a better performance, with the highest F-measure score at 100%, lowest at 59% and average at 80%. The results shows that EP is useful in unsupervised WSD.

Sequence Number	Ambiguous word	F-measure	Sequence Number	Ambiguous word	F-measure (%)
1	把握(ba3 wo4)	0.93	11	没有(me1 you3)	1.00
2	包(bao1)	0.74	12	起来(qi3 lai2)	0.59
3	料(cai2 liao4)	0.80	13	钱(qian2)	0.71
4	冲击(chong1 ji1)	0.85	14	日子(ri4 zi3)	0.62
5	穿(chuan1)	0.79	15	少(shao3)	0.82
6	地方(di4 fang1)	0.78	16	突出(tu1 chu1)	0.93
7	分子(fen1 zi3)	0.94	17	研究(yan2 jiu1)	0.71
8	运动(yun4 dong4)	0.94	18	活动(huo2 dong4)	0.89
9	老(lao3)	0.85	19	走(zou3)	0.68
10	路(lu4)	0.81	20	坐(zuo4)	0.67
Average	0.80	Note: Average of the 20 words			

Table 3. The Results for Unsupervised WSD based on EPs

From the results in table 2 and table 3, it can be seen that 16 among the 20 ambiguous words show better WSD performance in unsupervised SWD than in supervised WSD, while only 2 of them shows similar results and 2 performs worse. The average F-measure of the unsupervised method is higher by more than 10%. The reason lies in the following aspects:

1) Because there are several morpheme words for every sense of the word in construction of the EP, rich semantic information can be acquired in the training step and is an advantage for sense disambiguation.

2) Senseval-3 has provided a small-scale training set, with 15-20 training instances for each sense, which is not enough for the WSD modeling. The lack of training information leads to a low performance of the supervised methods.

3) With a large-scale training corpus, the unsupervised WSD method has got plenty of training instances for a high performance in disambiguation.

4) The discriminability of some ambiguous word may be low, but the corresponding EPs could be easier to disambiguate. For example, the ambiguous word "穿" has two senses which are difficult to distinguish from each other, but its EPs' senses of "越过/穿过/穿越" and "戳/捅/通/扎" can be easily disambiguated. It is the same for the word "冲击", whose EPs' senses are "撞击/磕碰/碰撞" and "损害/伤害". EP-based knowledge acquisition of these ambiguous words for WSD has helped a lot to achieve high performance.

5 Conclusion

As discussed above, the supervised WSD method shows a low performance because of its dependency on the size of the training data. This reveals its weakness in knowledge acquisition bottleneck. EP-based unsupervised method has overcome this weakness. It requires no manually tagged corpus to achieve a satisfactory performance on WSD. Experimental results show that EP-based method is a promising solution to the large-scale WSD task. In future work, we will examine the effectiveness of EP-based method in other WSD techniques.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-

Sense Disambiguation Using Statistical Methods. In *Proc. of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-1991)*, pages 264-270.

Mona Talat Diab. 2003. *Word Sense Disambiguation Within a Multilingual Framework*. PhD thesis, University of Maryland College Park.

Mona Diab. 2004a. Relieving the Data Acquisition Bottleneck in Word Sense Disambiguation. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pages 303-310.

Mona T. Diab. 2004b. An Unsupervised Approach for Bootstrapping Arabic Sense Tagging. In *Proc. of Arabic Script Based Languages Workshop at COLING 2004*, pages 43-50.

Mona Diab and Philip Resnik. 2002. An Unsupervised Method for Word Sense Tagging Using Parallel Corpora. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 255-262.

William Gale, Kenneth Church, and David Yarowsky. 1992a. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. In *Proc. of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, pages 101-112.

William Gale, Kenneth Church, and David Yarowsky. 1992b. Work on Statistical Methods for Word Sense Disambiguation. In *Proc. of AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, pages 54-60.

William Gale, Kenneth Ward Church, and David Yarowsky. 1992c. Estimating Upper and Lower Bounds on the Performance of Word Sense Disambiguation Programs. In *Proc. of the 30th Annual Meeting of the Association for Computational Linguistics (ACL-1992)*, pages 249-256.

Tanja Gaustad. 2001. Statistical Corpus-Based Word Sense Disambiguation: Pseudowords vs. Real Ambiguous Words. In *Proc. of the 39th ACL/EACL Student Research Workshop*, pages 61-66.

Nancy Ide, Tomaz Erjavec, and Dan Tufiş. 2001. Automatic Sense Tagging Using Parallel Corpora. In *Proc. of the Sixth Natural Language Processing Pacific Rim Symposium*, pages 83-89.

Nancy Ide, Tomaz Erjavec, and Dan Tufiş. 2002. Sense Discrimination with Parallel Corpora. In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 54-60.

Cong Li and Hang Li. 2002. Word Translation Disambiguation Using Bilingual Bootstrapping. In *Proc. of the 40th Annual Meeting of the Association*

- for *Computational Linguistics (ACL-2002)*, pages 343-351.
- Rada Mihalcea and Dan Moldovan. 2000. An Iterative Approach to Word Sense Disambiguation. In *Proc. of Florida Artificial Intelligence Research Society Conference (FLAIRS 2000)*, pages 219-223.
- Rada F. Mihalcea. 2002. Bootstrapping Large Sense Tagged Corpora. In *Proc. of the 3rd International Conference on Languages Resources and Evaluations (LREC 2002)*, pages 1407-1411.
- Preslav I. Nakov and Marti A. Hearst. 2003. Category-based Pseudowords. In *Companion Volume to the Proceedings of HLT-NAACL 2003, Short Papers*, pages 67-69.
- Hwee Tou. Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 455-462.
- Zheng-Yu Niu, Dong-Hong Ji, and Chew-Lim Tan. 2005. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. In *Proc. of the 43th Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 395-402.
- Ying Qin and Xiaojie Wang. 2005. A Track-based Method on Chinese WSD. In *Proc. of Joint Symposium of Computational Linguistics of China (JSCL-2005)*, pages 127-133.
- Hinrich. Schutze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1): 97-123.
- David Yarowsky. 1994. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-1994)*, pages 88-95.
- David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995)*, pages 189-196.