International Journal of Knowledge and Language Processing
Volume **2**, Number 2, April **2011**

www.ijklp.org
AKLP International ©2011 ISSN 2191-2734
pp. **44-55**

Word Sense Disambiguation Based on Expanding Training Set Automatically

Pengyuan Liu
Applied Linguistics Research Institute, Beijing Language and Culture University
Beijing, 100083, China
Pengyuan_liu@acm.org

Received March 2011; revised April 2011

ABSTRACT. *Based on our supposition one sense per n-gram, three experiments of word sense disambiguation had been carried out in this paper. Each of them is corresponding to a different kind of sense granularity The first experiment is on event detection in the SemEval-2010 Task, the PKU_HIT system has been built for the three sub-tasks including target verb WSD. The second experiment is on Multilingual Chinese-English Lexical Sample task in SemEval-2007, a prototype naive Bayes system has been built. And finally the last experiment is on Infrequent Sense Identification for Mandarin Text to Speech Systems in SemEval-2010, the PengYuan@PKU system has been built. All these three systems expanding training set automatically based on one sense pre N-gram supposition. The experiment results show this method is simple but effective, especially in WSD of the coarse sense granularity.*

Keywords: word sense disambiguation; one sense per N-gram; expanding training set

1. **Introduction.** Word Sense Disambiguation (WSD) has been described as the task which selects the appropriate meaning (sense) to a given word in a given context where this meaning is distinguishable from other senses potentially attributable to that word. WSD is an important problem in NLP and an essential preprocessing step for many applications including machine translation, question answering and information extraction. WSD is a difficult task while state-of-the-art systems are still often not good enough for real-world applications despite the fact that it has been the focus of much research over the years. One major factor that makes WSD difficult is a relative lack of manually annotated corpora, which hampers the performance of supervised systems.

In order to achieve high performance, supervised approaches require large training sets where instances are hand-annotated with the most appropriate word senses. Producing this kind of knowledge is extremely costly: at a throughput of one sense annotation per minute[1] and tagging one thousand examples per word, dozens of person-years would be required for enabling a supervised classifier to disambiguate all the words in the English lexicon with high accuracy.

This paper shows a strategy of expanding annotated examples from some hand-annotated examples automatically. On one hand, the hand-annotated examples can be utilized, on the other hands the expanding examples may enabling a higher performance supervised

classifier. To our knowledge, the methods of auto acquiring sense-labeled instances include using parallel corpora like Gale et al. [2] and Ng et al.[3], extracting by monosemous relative of WordNet like Leacock et al. [4], Mihalcea and Moldovan [5], Agirre and Mart ńez [6], Mart ńez et al.[7] and PengYuan et al. [8]. The method proposed by Mihalcea and Moldovan [9] is also an effective way.

Following the celebrated supposition One Sense Per Collocation (OSPC) of Yarowsky [10], Our previous work [11] shows that with high probability, a polysemous word has One Sense Per N-gram (OSPN), and therefore local sources have enough information to determine the sense. We tested this empirical hypothesis by experimenting on Chinese Word Sense Tagging Corpus (STC), and discovered that it holds with over 85.9% agreement for both nouns and verbs.

Based on OSPN, we designed three WSD systems on three semantic evaluation tasks. All these three systems expanding training set automatically from origin training set of three tasks individually. The first system is the PKU_HIT system on event detection in the SemEval-2010 Task #11. We participate in the evaluation. The second system is a prototype naive Bayes system on Multilingual Chinese-English Lexical Sample task in SemEval-2007 #5. The last system is PengYuan@PKU system on Infrequent Sense Identification for Mandarin Text to Speech Systems in the SemEval-2010 task #15. This system is also our participating system.

This paper is organized as follows: Section 2 introduces the supposition, some definitions and describes the experiment on Chinese Word Sense Tagging Corpus. Section 3 presents the three application WSD systems. Section 4 makes some discussions. Finally, the conclusion and future work are presented in Section 5.

2. **One Sense Per Ngram.** In checking the error annotates of China Daily, where some words have already been sense-tagged by an auto-tagging system, some modifications were made for the convenience of this research. For each tagging word, we applied different ways to sort the sentences and discover that they have almost the same label when their N-gram is nearly the same by accident. We study the word label in STC immediately and the results are listed in Table 1.

TABLE 1: A typical 3-gram sense-label distribution for the polysemous word 成 cheng2.

N-gram	Freq. as !0-2	Freq. as !0-3
便成了	12	0
已经成了	6	0
已成定局	12	0
也成了	24	0
说成是	0	12
制成的	0	9
炼成的	0	57
办成了	0	5
不成问题	10	0
几乎成了	6	0

Table 1 shows the example of the sense label of a Chinese ambiguous word 成 cheng2(!0-2/!0-3) for the 3-gram pattern. "!0-2/!0-3" is the main sense (become/succeed) label of 成 which has been listed in the Chinese Semantic Dictionary (CSD) [12]. The underlined Chinese words in the N-gram column are function words (we follow the [10]'s definition of function words and content words). It does not show all 3-gram examples of 成 since the list will be too long. The very polarized result in Table 1 shows that N-gram can determine sense.

2.1. Definition and Formalization. For a target word in text, w_0, we wish to assign a sense label, s_i, from a fixed set of candidates, $S = (s_1, s_2...,s_{|S|})$. Assume that our target word w_0 occurs in a given window size sequence of context tokens: $C^0_{(i,j)} = (c_{-i}, c_{-i+1}, ... c_{-1}, w_0, c_1 ... c_{j-1}, c_j, i, j>0)$, which can be called context patterns[13]. For certain words and their particular sequence, we called them the N-gram of the word directly.

To any N-gram of a word w_0, if we only consider direct syntactic relationships such as verb/object, subject/verb, and adjective/noun pairs, or only choose one content words, it is just like the definition of Collocation between w_0 and c_m in [10]. The statistical properties between n-gram and senses of target word also reveal the statistical properties between collocation and senses if we take all n-grams for n>1 into consideration. Collocation can be just considered as one special kind of N-gram. Any conclusion of collocation could be thought of as a particular case of N-gram.

Many problems in NLP can be viewed as assigning labels to a particular word in text, given the word's context. Here, the definition of word sense is just a predefined label set of possible choices which can be chosen in the decision process. It could be translations in a language such as English, or the entries in a dictionary.

We define the sense of an N-gram as the sense (label) which the target word of the N-gram is tagged with. We define the entropy of a N-gram as the mean entropy of the distribution Pr(Sense|N-gram). Note that all of the N-gram entries in Table 1 have zero as one of the frequency counts. Although these zero counts had a contrary example in a larger corpus, we still have no plans of computing entropy to smooth as [10] caused the parallel comparison to counteract its influence largely.

2.2. Experiment on STC. STC is an ongoing project[5]of building a sense-tagged corpus [12]. Up to the present, STC has completed semantic annotation of more than three months of People's Daily. The set of semantic labels used are from CSD. We choose the sense-tagged 1, 2 and 3 months of People's Daily 2000 as our evaluation corpus. Table 2 is the overview of the evaluation corpus.

TABLE 2: Overview of the STC.

Class	Types	Tokens
Tagged verbs	461	144607
Tagged nouns	509	24545
All words	76102	3379761

From the tagged polymoneous words, we choose 247 verbs and 106 nouns which all have more than 20 instances in STC. We enumerate all N-grams (considered only 2-gram and 3-gram) of these words whose N-gram frequency counts are more than twice, list all

[5] http://icl.pku.edu.cn/ icl_groups/corpus/dwldform1.asp

their labels counts, compute the entropy of every N-gram. For contrast, we also take POS into account. The results are showed in Table 3.

In Table 3, (*i*,*j*) in *N-gram* column refers to the N-gram window as showed in section 2.1. In *class* column, *BL* refers to the results of general N-gram which do not consider POS of words, as our baseline, *W* refers to the target word, *F* refers to function word, and *C* refers to the content words. *FW*, *CW*, *WC* and etc. indicate the combination of N-gram words pattern. For instance, *FCW* refers to the N-gram like (f_{-2}, c_{-1}, w_0), f_{-2} is a function word, c_{-1}, is a content word and w_0 is the target word. Agreement (Agr.) means the underlying probability distributions of sense conditional on N-gram. For example, for the 2-gram pattern *WC*, the value of 0.931 indicates that on average, given a specific 2-gram we will expect to see the same 93.1% the time. This mean distribution is also reflected in the entropy (Ent.) column.

TABLE 3: Experimental results of OSPC on STC

N-gram	**Class**		**Agr.**		**Ent.**	
			verb	**noun**	**verb**	**noun**
(-1,0)		BL	0.902	0.966	0.171	0.062
	(F)	FW	0.886	0.935	0.200	0.115
	(C)	CW	0.908	0.980	0.162	0.036
(0,1)		BL	0.924	0.959	0.135	0.070
	(F)	WF	0.905	0.936	0.173	0.113
	(C)	WC	0.931	0.968	0.121	0.053
(-2,0)		BL	0.942	0.984	0.094	0.027
	(F)	FFW	0.921	0.980	0.129	0.032
		CFW	0.925	0.981	0.121	0.031
		FCW	0.946	0.981	0.089	0.032
	(C)	CCW	0.948	0.991	0.084	0.017
(-1,1)		BL	0.954	0.987	0.076	0.023
	(F)	FWF	0.934	0.992	0.109	0.016
		FWC	0.961	0.977	0.063	0.040
		CWF	0.949	0.988	0.084	0.020
	(C)	CWC	0.959	0.991	0.068	0.016
(0,2)		BL	0.952	0.987	0.079	0.021
	(F)	WFF	0.939	0.990	0.101	0.017
		WCF	0.945	0.980	0.092	0.029
		WFC	0.947	0.989	0.086	0.017
	(C)	WCC	0.965	0.987	0.058	0.020
ALL	(F)		0.917	0.967	0.142	0.059
ALL	(C)		0.942	0.983	0.099	0.028
ALL	BL		0.935	0.977	0.111	0.041

For the N-grams studied, it appears that the hypothesis of OSPN holds with high probability for disambiguation of nouns and verbs. The results in the Agr. Column of Table 3 quantify the validity of this claim. Accuracy varies from 88.6% to 99.2% for different patterns of N-gram and part of speech, with a mean of 95.6%.

3. T**hree application WSD systems.** Based on OSPN, we designed three WSD systems on

three semantic evaluation tasks. All these three systems expanding training set automatically from origin training set of three tasks individually.

3.1. The PKU_HIT System on SemEval-2010 task #11[14]**.** The objective of the task is to detect and analyze basic event contents in Chinese news sentences, similar to the frame semantic structure extraction task in SemEval-2007. However, this task is a more complex as it involves three interrelated subtasks: (1) target verb word sense disambiguation (WSD), (2) sentence semantic role labeling (SRL) and (3) event detection (ED). This paper will introduce the WSD module here, one can read [15] for the detail of PKU_HIT System.

For the WSD module, we consider the subtask as a general WSD problem. First of all, we automatically extract many instances from an untagged Chinese corpus using OSPN. Then we train a Naïve Bayesian (NB) classifier based on both the extracted instances and the official training data. We then use the NB classifier to predict situation the description formula and natural explanation of each target verb in testing data.

We suppose that one sense per 3-gram that we consider one sense per 3-gram (*w*-1*w*verb*w*1) and we can extract instances with this pattern. For all the 27 multiple-sense target verbs in the official training data, we found their 3-gram (*w*-1*w*verb*w*1) and extracted the instances with the same 3-gram from a Chinese monolingual corpus – the 2001 People's Daily (about 116M bytes). We consider the same 3-gram instances should have the same label. Then an additional sense labeled training corpus is built automatically in expectation of having 95.4% precision at most. And this corpus has 2145 instances in total (official training data have 4608 instances).

We build four systems to investigate the effect of our instances expansion using the Naïve Bayesian classifier. System configuration is shown in Table 4. In column 1, BL means baseline, X means instance expansion, 3 and 15 means the window size. In column 2, w_i is the *ith* word relative to the target word, $w_{i-1}w_i$ is the 2-gram of words, w_j/j is the word with position information ($j \in [-3,+3]$). In the last column, 'O' means using only the original training data and 'O+A' means using both the original and additional training data. Syntactic feature and parameter optimizing are not used in this module.

TABLE 4: The system configuration

WSD Systems	**Features**	**Window Size**	**Training data**
BL_3		±3	O
X_3	$w_{i,}$	±3	O+A
BL_15	$w_{i-1}w_{i,}w_j/j$	±15	O
X_15		±15	O+A

Table 5 shows the official result of the WSD system. BL_3 with window size three using the original training corpus achieves the best result in our submission. It indicates the local features are more effective in our systems. There are two possible reasons why the performances of the X system with instance expansion are lower than the BL system. First, the additional instances extracted based on 3-gram provide a few local features but many topical features. But, local features are more effective for our systems as mentioned above. The local feature related information that the classifier gets from the additional instances is not sufficient. Second, the granularity of the WSD module is too small to be distinguished by 3-grams. As a result, the additional corpus built upon 3-gram has more exceptional

instances (noises), and therefore it impairs the performance of X_3 and X_15. Taking the verb ‘ 属于’ (belong to) as an example, it has two senses in the task, but both senses have the same natural explanation: ‘归一 某方面或为某方所有’ (part of or belong to), which is always considered as the sense in general SRL. The difference between the two senses is in their situation description formulas: ‘partof (x,y)+NULL’ vs. ‘belongto (x,y)+NULL’.

TABLE 5: Official results of the WSD systems

Systems	**Micro-A (%)**	**Macro-A (%)**	**Rank**
BL_3	81.30	83.81	3/7
X_3	79.82	82.58	4/7
BL_15	79.23	82.18	5/7
X_15	77.74	81.42	6/7

3.2. The NB System on SemEval-2007 task #5. The Multilingual Chinese-English Lexical Sample task (MCELS) [16] includes 40 Chinese ambiguous words: 19 nouns and 21 verbs are selected for evaluation. Each sense of a word is provided at least 15 instances and at most 40 instances, in which around 2/3 of the instances are used as the training data and 1/3 as the test data. Table 1 presents the number of words under each part of speech (POS), the average number of senses for each POS and the number of instances in the training and test sets, respectively.

Like Section 3.1, we start from an initial labeled set, use the labeled instances to extract their 3-gram of the target words, then the sense of the 3-gram is the label of the corresponding instance. More instances contain the same 3-gram can be extracted from a large corpus or web. For every sense of every target word, we can get many corresponding N-grams and many instances. We still use Naïve Bayes classifier here.

TABLE 6: Experiment results of WSD. S-EX, S-MCELS and S-COMB refer to the system which train with the instances extracted, train instances of MCELS and both instances respectively. The “+” column is the promotion of performance. The MFS line means the result of most frequent sense provided by the MCELS. The last line SRCB-WSD is the best system on SemEval-2007 task #5.

Window	**S-EX**		**S-MCELS**		**S-ALL**		+
	Micro	Macro	Micro	Macro	Micro	Macro	
-2,2	0.612	0.641	0.691	0.729	0.693	0.735	+0.006
-3,3	0.614	**0.655**	0.695	**0.734**	0.698	0.742	+0.008
-4,4	0.579	0.613	0.691	0.724	0.689	0.731	+0.007
-5,5	0.585	0.625	0.690	0.719	0.697	**0.746**	**+0.027**
-7,7	0.594	0.631	0.672	0.700	0.670	0.702	+0.002
-9,9	0.587	0.624	0.680	0.707	0.666	0.702	*-0.005*
-12,12	0.582	0.626	0.667	0.706	0.664	0.710	+0.004
-15,15	0.568	0.616	0.651	0.691	0.659	0.708	+0.017
MFS	0.405	0.462	0.405	0.462	0.405	0.462	N/A
SRCB-WSD	0.717	0.749	0.717	0.749	0.717	**0.749**	N/A

The training set of MCELS is our initial labeled set. We use all the 3-gram (-1,1) and 8832 instances are extracted from the 2001 year People's Daily (about 116M bytes). Features are selected with the words, POS and the 2-grams in the context window which target words are centered. Table 4 shows the results that all the three systems including *S-EX* which only use the extracted instance are prior to the *MFS*. Combining all the instances (*S-ALL*) made a slight improvement (2.7%) using only the instances in MCELS training samples (*S- MCELS*). The result indicates that the instances extracted by OSPN can help disambiguation.

The promotion of performance is not much, for the very unbalance sense distribution between the extracted instances and MCELS training samples. We suppose that the performance would improve a lot if instances are extracted from web and the sense distribution could be controlled to match that of MCELS training samples. For infrequent sense we can identified like Peng [17]. For Predominant Word Senses, we can estimate sense distribution automatically like Diana [18].

3.3. The PengYuan@PKU System on SemEval-2010 task #15. This task required systems to disambiguating the homograph word, a word that has the same POS (part of speech) but different pronunciation. In this case, we still considered it as a WSD (word sense disambiguation) problem, but it is a little different from WSD. In this task, two or more senses of the same word may correspond to one pronunciation. That is, the sense granularity is coarser than traditional WSD.

The challenge of this task is the much skewed distribution in real text: the most frequent pronunciation accounts for usually over 80%. In fact, in the training data provided by the organizer, we found that the sense distribution of some words is distinctly unbalanced. For each of these words, there are fewer than ten instances of one sense whereas the dominant sense instances are hundreds or more. At the same time, according to the task description on the task 15 of SemEval-2010 (http://semeval2.fbk.eu/semeval2.php?location=tasks), the test dataset of this task is intentionally divided into the infrequent pronunciation instances and the frequent ones by half and half. Apparently, if we use traditional methods and only the provided training dataset to train whatever classifier, it is very likely that we will get a disambiguation result that all (at least the overwhelming number) the test instances of these words would be labeled with the most frequent pronunciation (sense) tag. Then our system is meaningless for the target of the task is focused on the performance of identifying the infrequent sense.

The core system is a supervised system based on the ensembles of Naïve Bayesian classifiers. The complemented training data is extracted from an untagged Chinese corpus – People's Daily of the year 2001 automatically.

TABLE 7: Features and their weights used in one Naïve Bayesian classifier

Features	Description	weights
$w_{-i}\ldots w_i$	Content words appearing within the window of $\pm i$ words on each side of the target word	1
$w_j/j,j\in[-3,3]$	Word forms and their position information of the words at fixed positions from the target word.	3
$w_{k-1}w_k,k\in(-i,i]$	word bigrams appearing within the window of $\pm i$	1 when i>3, else 3
$P_{k-1}P_k,k\in(-i,i]$	POS bigrams appearing within the window of $\pm i$	1

The features and their weights of context used in one single Naïve Bayesian classifier are described in Table 7. The ensemble strategy of our system is like [19]. The windows of context have seven different sizes (*i*): 3, 5, 7, 9, 11, 13 and 15 words. The first step in the ensemble approach is to train a separate Naïve Bayesian classifier for each of the seven window sizes.

Each of the seven member classifiers votes for the most probable sense given the particular context represented by that classifier; the ensemble disambiguates by assigning the sense that receives the majority of the votes.

TABLE 8: The overview of the training data before and after the extracting stage

N-gram		**Increasing Instances Number**	
3-gram	(-1,1)	246	1026(9135)
	(-2,0)	229	
	(0,2)	551	
2-gram	(-1,0)	1123	2967(9135)
	(0,1)	1844	

TABLE 9: The sense distributions of the training data before and after the extracting stage

Target Words	**Sense Distribution**					
	Before (O)		**After**			
			(O+E3)		**(O+E2)**	
背	128	51	128	66	128	262[6]
车	503	83	503	83	503	194
澄清	168	13	168	16	168	23
冲	175	10	175	27	175	88
当	487	42	487	63	487	267
合计	134	44	134	44	134	49
见长	125	11	125	11	125	12
看	2020	8	2020	12	2020	25
落	300	3	300	6	300	32
没	268	3	268	4	268	45
上	1625	41	1625	346	1625	1625
系	144	13	144	15	144	33
兄弟	136	8	136	9	136	16
应	1666	253	1666	847	1666	1567
攒	142	17	142	17	142	17
转	438	76	438	136	438	414

For all the 16 multiple-sense target words in the training data of task 15, we found the N-gram of infrequence sense instances and extracted[7] the instances with the same N-gram from People's Daily of the year 2001(about 116M bytes). We extracted as many as

[6] We intentionally control the sense distribution of word ("背") and change it from approximately 2.5:1 to 1:2 so as to investigate the influence.

[7] In order to guarantee the extracted instances are not duplicated in the training data or in the test data in case, our system filters the repeated instances automatically if they are already in the original training or test dataset.

possible until the total number of them is equal to the dominant sense instance number. We appointed the same N-gram instances the same sense tag and merge it into the original training corpus. Table 2 and 3 show the overview and the sense distribution of the training data before and after the extracting stage. Number 9135 in brackets of Table 8 is the instance number of original training corpus. O, O+E3, O+E2 in Table 9 mean original training data, original training data plus extracted 3-gram instances and original training data plus extracted 2-gram instances respectively. Limited to the scale of the corpus, the unbalance sense distribution of some words does not improve much.

TABLE 10: Official results 1 of PengYuan@PKU

System ID	Micro Accuracy	Macro Accuracy	Rank
_3.001	0.974	0.952	1/9
_3.1	0.965	0.942	2/9
_2.001	0.965	0.941	3/9
_2.1	0.965	0.942	2/9
Baseline	0.924	0.895	

TABLE 11: Official results 2 of PengYuan@PKU

Words	Precision				
	_3.001	_3.1	_2.001	_2.1	baseline
背	**0.844**	0.789	0.789	0.789	0.711
车	**0.976**	0.962	0.969	0.962	0.863
澄清	**0.901**	**0.901**	**0.901**	**0.901**	0.901
冲	0.978	**0.989**	0.978	**0.989**	0.957
当	**0.925**	0.853	0.864	0.853	0.925
合计	**0.956**	0.944	**0.956**	0.944	0.700
见长	**0.971**	0.956	0.956	0.956	0.956
看	**0.998**	0.997	0.997	0.997	0.996
落	**0.987**	0.974	0.974	0.974	0.987
没	0.956	0.963	**0.971**	0.963	0.956
上	**0.983**	0.975	0.969	0.975	0.978
系	0.924	**0.949**	0.937	**0.949**	0.886
兄弟	**0.986**	**0.986**	**0.986**	**0.986**	0.959
应	0.986	**0.989**	**0.989**	**0.989**	0.869
攒	0.875	**0.900**	0.875	**0.900**	0.838
转	**0.981**	0.946	0.953	0.946	0.844

Macro Accuracy is the average disambiguation precision of each target word. Micro Accuracy is the disambiguation precision of total instances of all words. For task 15 whose instance distribution of the target words is very unbalanced in the test dataset, Macro Accuracy maybe a better evaluation indicator. Our systems achieved from 1st to 4th position (ranked by Macro Accuracy) out of all nine systems that participated in this task. Our best system is PengYuan@PKU_3.001 which uses original training data plus extracted 3-gram

instances as our training data, $P(S)$ is tuned to 0.5 and smoothness variable λ is equal to 0.001.

From the official result in Table 10 and Table 11 we can see, for this task, our classifier and strategy of extracting infrequency instances is effective. Basically, for each target word, the performances of our systems are superior to the baseline.

From Table 11, we also see the performances of our systems are influenced by different λ and different instance extracting patterns. Comparatively smaller probability λ of nonoccurrence features is better. Using the Extracting 3-gram instances is better than that of using 2-gram. (By using the 3-gram method of extracting instances, we obtain a better result than that of 2-gram.)

Our original idea for the system is two-folds. On one hand, we consider the relieving of data sparseness through more instances extracted by 2-gram pattern can achieve a better performance than that of 3-gram pattern, though the instances extracted through 2-gram pattern induce more noise. On the other hand, we assume that the performance would be better if we had given a larger probability of nonoccurrence features, for this strategy favors more infrequent sense instances. However the unbalance of sense distribution in the real test data as is shown in Table 9 went beyond our expectation. It is very hard for us to evaluate our system from the viewpoint of smoothness and instance sense distribution.

4. Discussion. Based on OSPN, this paper designed three supervised WSD systems as in section 3. And each of them is corresponding with a kind WSD of sense granularity as the Table 12 shows.

TABLE 12: The Sense Granularity among the 3 Tasks

System	PKU_HIT	NB	PengYuan@PKU
Sense Granularity	finer than common	common	Coarser than common

The result of PKU_HIT shows that our strategy is almost failed. It introduces more and more noise of incorrect instances when expanding sense-tagged examples. The fine granularity sense of a word needs more context than only Ngram context. The MCELS task is a common WSD task, our system promote the performance of the baseline system. Due to the influence of distribution of the sense-instances, it does not show a surprising improvement. The PengYuan@PKU system is a relative completed system and it disambiguates the sense of words coarser than that of common WSD task. The system controls the distribution of instances which extracted automatically and the noise of the new instances is much lower than those of PKU_HIT and NB. Finally this system performs well and gets the-state-of-art result.

5. **Conclusions and Future Work.** Based on OSPN and extracting sense-tagged instances automatically, this paper presents three WSD systems which each of them are corresponding with a kind WSD of sense granularity. The experiments on SemEval2010 task #11 and task #15 and SemEval2007 task #5 shows that that strategy is effective. The performance is influenced by many factors such as the distribution of different sense instances, the sense granularity and the smoothness variable. This strategy is better when it faces the coarser sense granularity than that of common WSD.

In future study, we will do followings so as to improve the performance further:

(1) Search the internet to get more instances.
(2) Adding some rules to reduce the noise.
(3) Getting more Ngram by self-training and bootstrapping.
(4) Combining some semantic lexicon.

Acknowledgment. This work is supported by the National Natural Science Foundation of China under Grant No.60903063.

REFERENCES

[1] Philip Edmonds, Designing a task for SENSEVAL-2. Technical report, University of Brighton, U.K. 2000.

[2] William A. Gale, Kenneth W. Church and David Yarowsky, A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(2):415-539,1992.

[3] Hwee Tou Ng, Bin Wang, Yee Seng Chan, Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. *Proceedings of the 41st ACL*, 455-462, Sappora, Japan, 2003

[4] Claudia Leacock, Martin Chodorow and George A. Miller, Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147~166, 1998.

[5] Rada Mihalcea and Dan I. Moldovan, An automatic method for generating sense tagged corpora. *Proceedings of AAAI-99*, Orlando, FL, 461–466,1999.

[6] Eneko Agirre and David Mart‘inez, Unsupervised WSD based on automatically retrieved examples: The importance of bias. *Proceedings of the International Conference on Empirical Methods in Natural Language Processing, EMNLP*, 25~32, 2004.

[7] David Mart‘inez, Eneko Agirre and Xinglong Wang, Word relatives in context for word sense disambiguation. *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, 42~50, 2006.

[8] Liu Peng-yuan Zhao Tie-jun Yang Mu-yun Li Zhuang, Unsupervised Translation Disambiguation Based on Equivalent PseudoTranslation Model. *Journal of Electronics & Information Technology*. 30(7):1690-1695, 2008.

[9] Rada Mihalcea and Dan .I. Moldovan, An iterative approach to word sense disambiguation. *Proceedings of FLAIRS-2000*, pp 219–223, Orlando, FL, May, 2000.

[10] Yarowsky David, “One sense per collocation,” In Proceedings of the ARPA Workshop on Human Language Technology, pp. 266-271, 1993.

[11] PengYuan Liu, Shui Liu ShiQi Li and ShiWen Yu, One Sense Per N-gram. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops NLPOE, pp. 195-198, 2010.

[12] Wu, Y., P. Jin, Y. Zhang and S. Yu, “A Chinese corpus with word sense annotation,” Proceedings of ICCPOL2006, pp. 414-21, 2006.

[13] Bergsma S., D. Lin and R. Goebel, “Distributional identification of nonreferential pronouns,” In Proceedings of the 45th Annual Meeting of the Association forComputational Linguistics (ACL), pp. 10-18, 2008.

[14] Qiang Zhou, SemEval-2010 task 11: Event detection in Chinese News Sentences. *Proceedings of SemEval-2010, 2010.*

[15] **S**hiqi Li, Pengyuan Liu, Tiejun Zhao, Qin Lu and Hanjing Li, PKU_HIT: An Event Detection System Based on Instances Expansion and Rich Syntactic Features. Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, Uppsala, Sweden, 15-16 July, pp. 304–307,2010.

[16] Peng Jin, Yunfang Wu and Shiwen Yu, "SemEval-2007 task 05: multilingual Chinese-English lexical sample." Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pp.19-23, 2007.

[17] Peng Jin, One Class SVMs for infrequent sense identification. International Journal of Knowledge and Language Processing. 1(1), pp.1-18, 2010.

[18] Diana McCarthy, Rob Koeling, Julie Weeds and John Carroll, Finding predominant word senses in untagged text, Proceedings of ACL-2004, 2004.

[19] Ted. Pedersen, A Simple Approach to Building Ensembles of Naïve Bayesian Classifiers for Word Sense Disambiguation. *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA, May, pp. 63-69, 2000.