PengYuan@PKU: Extracting Infrequent Sense Instance with the Same N-gram Pattern for the SemEval-2010 Task 15

Peng-Yuan Liu[1] Shui Liu[2] Shi-Wen Yu[1] Tie-Jun Zhao[2]
[1]Institute of Computational Linguistics, Peking University, Beijing, China
[2]Department of Computer Science, Harbin Institute of Technology, Harbin, China
{liupengyuan,yusw}@pku.edu.cn,{tjzhao,liushui}@mtlab.hit.edu.cn

Abstract

This paper describes our infrequent sense identification system participating in the SemEval-2010 task 15 on Infrequent Sense Identification for Mandarin Text to Speech Systems. The core system is a supervised system based on the ensembles of Naïve Bayesian classifiers. In order to solve the problem of unbalanced sense distribution, we intentionally extract only instances of infrequent sense with the same N-gram pattern as the complemental training data from an untagged Chinese corpus – People's Daily of the year 2001. At the same time, we adjusted the prior probability to adapt to the distribution of the test data and tuned the smoothness coefficient to take the data sparseness into account. Official result shows that, our system ranked the first with the best Macro Accuracy 0.952. We briefly describe this system, its configuration options and the features used for this task and present some discussion of the results.

1 Introduction

We participated in the SemEval-2010 task 15 on Infrequent Sense Identification for Mandarin Text to Speech Systems. 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 are 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 an 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.

In order to solve the problem of the unbalanced sense distribution in the training data and the fairly balanced sense distribution in the test data, we designed our PengYuan@PKU system, which attempts to extract infrequent sense instances only and adjust the prior probability so as to counteract the problem as far as possible. The core system is a supervised system based on the ensembles of Naïve Bayesian classifiers. The complemental training data is extracted from an untagged Chinese corpus – People's Daily of the year 2001 automatically. Besides the motivation of investigating the function of our method of compensating infrequent sense instances, we are also interested in the role where the smoothness plays when it encounters with such a data sparseness here.

In section 2, we will describe our system that includes the core classifier, its configuration options and features. In section 3, we will show the official results of this task and present some analyses and discussions. Section 4 is related

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 371–374,
Uppsala, Sweden, 15-16 July 2010.

works. The conclusion and future work are in section 5.

2 System Description

2.1 Naïve Bayesian Classifier and Features

For a naïve Bayesian classifier, the joint probability of observing a certain combination of context features with a particular sense is expressed as:

$$p(F_1, F_2, ..., F_n, S) = p(S)\prod_{i=1}^{n} p(F_i \mid S) \qquad (1)$$

In equation (1), (F_1, F_2,…, F_n) is feature variables, S is classification variable and $p(S)$ is the prior probability of classification variable. Any parameter that has a value of zero indicates that the associated word never occurs with the specified sense value. These zero values are smoothed by additive smoothing method as expressed below:

$$P(F_i \mid S_k) = \frac{C(F_i, S_k) + l}{C(S_k) + N} \quad , \quad \lambda \in (0,1) \qquad (2)$$

In equation (2), λ is the smoothness variable. $C(S_k)$ is the times of instances with S_k label. $C(F_i,S_k)$ is the concurrences times of F_i and S_k. N is the times of total words in the corpus.

The features and their weights of context used in one single Naïve Bayesian classifier are described in Table 1.

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

Table 1: Features and their weights used in one Naïve Bayesian classifier

2.2 Ensembles the Naïve Bayesian Classifiers

The ensemble strategy of our system is like Pederson (2000). 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.

2.3 Infrequent Sense Instances Acquisition

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 2: The overview 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[1]
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

Table 3: The sense distributions of the training data before and after the extracting stage

Our system uses a special heuristic rule to extract the sense labeled infrequent sense instances automatically. The heuristic rule assumes that *one sense per N-gram* which we testified initially through investigating a Chinese sense-tagged corpus STC (Wu et al., 2006). Our assumption is inspired by the celebrated *one sense per collocation* supposition (Yarowsky, 1993). STC is an ongoing project of building a sense-tagged

[1] 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.

corpus which contained the sense-tagged 1, 2 and 3 months of People's Daily of the year 2000. According to our investigation, to any target multi-sense word, given a specific N-gram (N>1) including the target word, we will expect to see the same label that range from 88.6% to 99.2% of the time on average. So, based on the training data, we can extract instance with the same N -gram pattern from the untagged Chinese corpus and we assume if the N-gram is the same then the sense-label is the same.

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[2] the instances with the same N-gram from People's Daily of the year 2001(about 116M bytes). We extracted as many as 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 2 is the instance number of original training corpus. O, O+E3, O+E2 in Table 3 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.

2.4 Other Configuration Options

Systems	Training Data	$p(S)$	λ
_3.001	O+E3	0.5	0.001
_3.1	O+E3	0.5	0.1
_2.001	O+E2	0.5	0.001
_2.1	O+E2	0.5	0.1

Table 4: The system configuration

To formula (1), we tune the prior probability of classification variable $p(S)$ as a constant to match the sense distribution of test data. Considering the data sparseness as there may have been in the test stage, to formula (2), we set 2 kinds of λ to investigate the effect of smoothness.

In total, we develop four systems based on various configuration options. They are showed in Table 4.

[2] 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.

3 Results and Discussions

3.1 Official Results

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 5: Official results 1 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

Table 6: Official results 2 of PengYuan@PKU

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 λ is equal to 0.001.

3.2 Discussions

From the official result in Table 5 and Table 6 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 6, 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 5 went beyond our expectation. It is very hard for us to evaluate our system from the viewpoint of smoothness and instance sense distribution.

4 Related Work

To our knowledge, the methods of auto-acquiring sense-labeled instances include using parallel corpora like Gale et al. (1992) and Ng et al. (2003), extracting by monosemous relative of WordNet like Leacock et al. (1998), Mihalcea and Moldovan (1999), Agirre and Martínez (2004), Martínez et al. (2006) and PengYuan et al. (2008). The method proposed by Mihalcea and Moldovan (2000) is also an effective way.

5 Conclusion and Future Work

We participated in the SemEval-2010 task 15 on Infrequent Sense Identification for Mandarin Text to Speech Systems. Official results show our system which extract infrequent sense instances is effective.

For the future studies, we will focus on how to identify the infrequent sense instances effectively based on the plan to change the proposition between dominant sense and infrequent sense step by step.

Acknowledgments

This work was supported by the project of National Natural Science Foundation of China (No.60903063) and China Postdoctoral Science Foundation funded project (No.20090450007).

References

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

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)*, 2006:42~50

David Yarowsky. 1993. One sense per collocation. *Proceedings of the ARPA Workshop on Human Language Technology*.

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*, 2004:25~32

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

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

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

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

Ted. Pedersen. 2000. 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*, pages 63-69, Seattle, WA, May.

Yunfang Wu, Peng Jin, Yangsen Zhang, and Shiwen Yu. 2006. A Chinese corpus with word sense annotation. *Proceedings of ICCPOL-2006*.

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