The Effect of the Number of Features to Supervised Chinese Word Sense Disambiguation

Pengyuan Liu
Applied Linguistics Research Institute, Beijing Language and Culture University, Beijing, China
Email: liupengyuan@pku.edu.cn

***Abstract*—Although feature selection is very important during either supervised or unsupervised word sense disambiguation processing, there is no systematic study on investigating the relationship between the number of features and the performance as we know yet. This paper investigates the effect of the number of features to supervised Chinese word sense disambiguation through thousands of experiments on Semeval 2007 Multilingual Chinese-English Lexical Sample task dataset. It shows that local basic feature provides adequate information to do disambiguation and the influence of data sparseness is not as important on the performance as we think before from the number of features point of view.**

***Index Terms*—Chinese word sense disambiguation, number of features, window size, Local feature, Data sparseness**

I. INTRODUCTION

Word Sense Disambiguation (WSD) has been described as a 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, semantic role labeling and information extraction. At the same time, WSD is a difficult task despite the fact that it has been the focus of much research over the years. To our knowledge, to any language, there is no system which can disambiguate all the multi-sense words good enough for any real-world applications. One major factor the researchers often discussed is that what making WSD difficult is a relative lack of manually annotated corpora which hampers the performance of supervised systems. On the other hand, on the series of semantic evaluation task [1-4], the performance of the state-of-the-arts supervised WSD system which trained by the golden training set cannot reach 80% yet.

Researchers have used many kinds of the learning algorithms including Naïve Bayes Model, Maximum Entropy Model, Decision List, Decision Tree, k-Nearest Neighbor, Support Vector Machines and so on. Although there are also many studies on the comparison of different models [5-10], no founded evidence proved which model is the best or why this model is the best.

If how to acquiring the tagging corpus is not our matter for now, besides the learning algorithms, another main decision needs to be made in the design of a supervised system is the set of features to be used. Some researchers support Kaplan who studied the window size in which features are chosen and think the accuracy of sense resolution does not improve when more than four words around the target word are considered [11]. While some researchers [12, 13] think a large context window provides domain information which increases the accuracy of some words. Some researchers [13- 15] focus on feature selection. They do not use all the features in the context windows and try to select the most useful features for different words and got better results than the baselines which do not use feature selection.

In Chinese WSD, [15] also studies the feature selection and selects different features for every words and gets the state of the arts performance.

It shows that the entire context feature (1-gram, which is very important kind of the feature, rank 3rd, the 1st is the target word itself, the 2nd is the POS-tag of the target word) after feature selection for every target word is limited to 6 words (window size = 3) around the target word except one context 1-gram feature in [13]. In [15], we found that all the features it used finally is limited to the window size 5 and most of the features are limited to the window size 2. It also investigates the effect on result within different window sizes; it shows that using the features before feature selection for every word, the window size which getting the best performance is 1.

Although feature is very important in either supervised or unsupervised word sense disambiguation, there is no systematic study on investigating the relationship between the number of features and the performance as we know yet. This paper investigates the effect of the number of features to supervised Chinese word sense disambiguation through thousands of experiments on Semeval 2007 Multilingual Chinese-English Lexical Sample task dataset which is the same dataset with [15] by a Naive Bayes classifier. The experiment shows that local basic feature provides adequate information to do disambiguation and larger window size enlarges the probability of key local feature while sometimes the loss outweighs the gain if the window size is too large.

This paper is organized as follows: next section, it shows the contradiction results between [13] and [15] which is also the reason we do the experiments to study

doi:10.4304/jcp.8.2.313-318

TABLE I.
THE BEST RESULTS AFTER THE AUTOMATIC FEATURE SELECTION ALGORITHM IN [13]

Word.pos	Features	recall
authority.n	CW CP COL=1 VB NB	91.3%
facility.n	CP COL=1 VB PRB	98.3%
fatigue.n	CP B=5,3 NB	90.7%
holidy.n	CP	96.8%
lady.n	CW HNP	94.3%
mouth.n	CP COL=1 VB NB PB	93.3%
spade.n	CP CF=1 COL=2	97.0%
yew.n	CF=1	100%
solemn.a	CP COL=1 DT	96%
vital.a	CW CP NB	94.7%
collaborate.v	CW CP CF=1	90.0%
face.v	CP	100.0%
wonder.v	CP PA	90.0%

besides the arguments between local and topical feature. Section 3 introduces the Naive Bayes Classifier and the settings of the experiment. Section 4 describes the experiment which includes the dataset and evaluation of the task. It discusses the relationship between the performance and the number of features in section 5. Finally, some discussion and the conclusion and future work is presented in Section 6.

II. THE REMAIN FEATURES AFTER FEATURE SELECTION

Feature selection is an important process for every machine learning model[19]. Here we are only interested in the result on which features are selected and what is the final performance based on these features.

In [13], it presents an automatic feature selection algorithm and tests it on the Senseval-2 English lexical sample task. We show some of the best results of it in Table.1 (The recall is larger than 90%, coarse-grained). The meanings of the symbols of the features column in the Table.1 can be found in [13]. Table.1 shows that with the best performance, all the features selected by automatic feature selection algorithm are within window size 2 except one word - *fatigue.n*.

On the other hand, in [15], which is also a supervised approach to WSD, in the table of "Selected feature templates based on individual word", it shows that we'd best need the features within the context window 5.

There is an obvious divergence between their results. Is it because Chinese word sense needs more context words than English? Then how many features does Chinese word sense disambiguation need? If we want to know these we should know the relationship between the numbers of features and the performance first, due to the data sparseness the window sizes do not corresponding to features precisely.

TABLE II.
BASIC FEATURES USED IN NAÏVE BAYESIAN CLASSIFIER

Features	Description
W	Content words appearing within the window size ($\pm$i words on each side of the target word)
Wp	Word forms and their position information of the words at fixed positions from the target word.
Wn	word bigrams(W2) or trigrams(W3) appearing within the window size
P	POS of content words appearing within the window size
Pp	POS of word forms and their position information of the words at fixed positions from the target word.
Pn	POS bigrams(P2) or trigrams(P3) appearing within the window size

III. NAIVE BAYES CLASSIFIER AND SETTINGS

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

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

In (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:

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

In (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 basic features of the context used in Naïve Bayesian classifier are described in Table 2. We choose the features in common use here. The window sizes are set from ± 1 to ± 10.

IV. EXPERIMENTS AND RESULTS

A. Dataset

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 3 presents the number of words under each part of speech

TABLEIII.
SUMMARY OF THE SENSE INVENTORY AND NUMBER OF TRAINING AND TEST SETS

	#Average Senses	#Training Instances	#Training Instances
19 nouns	2.58	1019	364
21 verbs	3.57	1667	571

TABLE IV.
THE PMIR OF 8 BASIC TYPES OF FEATURE, WE ONLY SHOWS THE PMIR RESULTS FOR THE BEHAVIOR OF THE PMAR RESULTS IS BASICALLY THE SAME

WS	W	Wp	W2	W3	P	Pp	P2	P3
1	**0.661**	0.595	0.660	0.413	**0.570**	0.609	0.615	0.413
2	0.658	**0.618**	**0.668**	**0.533**	0.535	**0.627**	**0.624**	**0.596**
3	0.645	0.601	0.656	0.521	0.508	0.609	0.578	0.585
4	0.643	0.595	0.649	0.508	0.510	0.589	0.559	0.561
5	0.632	0.570	0.643	0.459	0.513	0.569	0.542	0.565
6	0.628	0.566	0.642	0.424	0.513	0.563	0.540	0.543
7	0.622	0.551	0.641	0.413	0.504	0.563	0.527	0.526
8	0.604	0.538	0.636	0.407	0.502	0.572	0.524	0.513
9	0.609	0.540	0.639	0.411	0.510	0.565	0.519	0.513
10	0.612	0.541	0.646	0.404	0.502	0.569	0.520	0.506

TABLE V.
THE PMIR OF 5 COMBINATION TYPES OF FEATURE, WE ONLY SHOWS THE PMIR RESULTS FOR THE BEHAVIOR OF THE PMAR RESULTS IS BASICLLY THE SAME

WS	PPpP2	PPpP3	WWpW2	WWpW3	WWpW2P2
1	0.618	0.604	0.660	0.670	0.710
2	**0.630**	**0.624**	0.682	0.675	0.716
3	0.609	0.616	**0.703**	**0.694**	**0.725**
4	0.612	0.604	0.698	0.683	0.717
5	0.591	0.602	0.702	0.686	0.716
6	0.589	0.593	0.688	0.670	0.701
7	0.574	0.584	0.689	0.658	0.713
8	0.593	0.580	0.666	0.652	0.698
9	0.579	0.574	0.682	0.649	0.692
10	0.576	0.560	0.677	0.651	0.689

(POS), the average number of senses for each POS and the number of instances in the training and test sets, respectively.

Two kinds of precisions are evaluated. One is micro-average:

$$P_{mir} = \sum_{i=1}^{N} m_i / \sum_{i=1}^{N} n_i \quad (3)$$

The other is macro-average:

$$P_{mar} = \sum_{i=1}^{N} p_i / N, p_i = m_i / n_i \quad (4)$$

N is the number of all target word-types. m_i is the number of correctly labeled test instances to one specific target word type and n_i is the number of all test instances for this word type.

B. Experiments and Results

In order to investigate the effect of different types of feature on the performance, we use 8 basic types of feature to do the experiments at first. The results are showed in table 4. In table 4, WS is the window size. We choose the window size (WS) from 1 to 10 because the average number of words per sentence of test data is about 20.

Table 4 shows that the features of all the best results (bold numbers) within window size 2. It supports the conclusion of some researchers in [11]. To have a close look, we compare all the 8 basic types of feature in Fig. 1. It shows that after the best micro-average peak at window sizes 2, their performance deteriorate as the window size get lager except P type of feature which always decreases along with the window size getting larger.

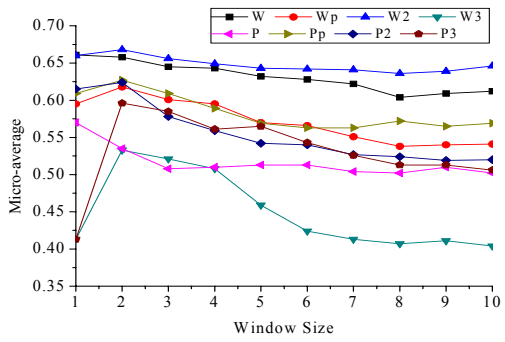

Figure 1. Pmir of 8 Basic Types of Feature Result decreases with the windows sizes after window size 2.

V. THE RELATIONSHIP BETWEEN PERFORMANCE AND THE NUMBER OF FEATURES

Due to the influence by the position of the target word and the data sparseness, window sizes are difficult to show the relationship between the features of test instances which appeared in the training feature set and the performance while the features relate with the final performance directly. In order to investigate the relationship between performance and the number of features, we define ANFS (Average Number of Features per Sense) for every experiment in (5).

$$ANFS_k = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{fn_j}{N \times sn_i} \quad (5)$$

In (5), k is the window size; M is the number of all test words; N is the number of the instances of one specific target word; fn_j is the number of features of one specific instance which appears in the training feature set and sn_i is the number of senses of one specific target word.

Fig. 2 shows the relationship between the value of ANFS and the rank of Pmir which ranks the Pmir results of the ten window-sizes by descending order. All the lines ended with the relative more features except P3. Although the maximum value of ANFS is about 15, the value of ANFS of the best result is under 3. It shows that about 80% features is noise or useless in all the features of window size 10.

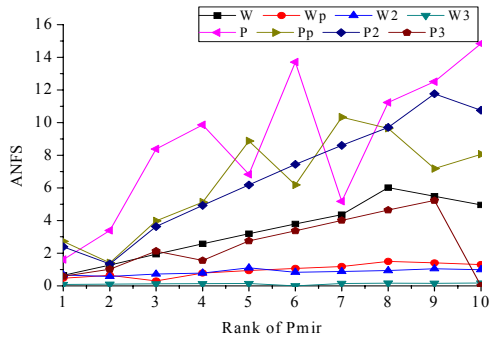

Figure 2. The Relationship of 8 Basic Type of Feature between ANFS and the Rank of Pmir which Ranks the Pmir Results of the Ten Window-sizes by Descending Order.

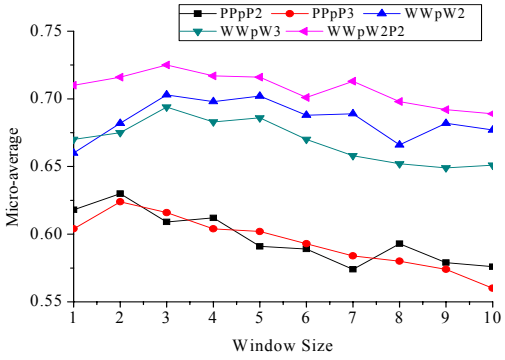

Figure 3. Pmir of 4 Word-related and POS-related Types Combination of Feature and the Best Feature Combination Result decreases with the windows sizes after window size 2 or 3.

We combine all the window sizes with the features and do thousands of experiments. We compare the results which only use word-related type feature (WWpW2 and WWpW3, which means using W, Wp and W2 types of feature and using W, Wp and W3 types of feature individually), only use POS-related type feature(PPpP2 and PPpP3, which means using P, Pp and P2 types of feature and using P, Pp and P3 types of feature individually) and the best combination of feature (Pmir = 0.725, which use W, Wp, W2 and P2 types of feature, outperforms the best system SRCB-WSD which Pmir = 0.717 in Semeval 2007 task #5) in table 5 and Fig. 3.

In Fig. 3 we find that the peak of Pmir happens at window size 2 or 3 and all the Pmir decrease along with the window size getting larger after the peak window size. The behavior between Pmir and all the types of combination of feature in Fig. 1 and Fig. 3 is by and large similar.

Table 6 shows the value of ANFS of 5 types of feature combination which Ranks it by Pmir descending order. In table 6, the bold number is the maximum value of ANFS, the italics number is the minimum value of ANFS. It shows that the Pmir performance is related with the value of ANFS, and high value of ANFS usually means a bad Pmir performance.

TABLE VI.
5 TYPE OF FEATURE COMBINATIONS BETWEEN ANFS AND THE RANK OF PMIR WHICH RANKS THE PMIR RESULTS OF THE TEN WINDOW-SIZES

Rank	PPpP2	PPpP3	WWpW2	WWpW3	WWpW2P2
1	8.55	6.71	*3.86*	3.24	7.49
2	*4.37*	10.17	5.67	4.97	9.72
3	16.96	*3.01*	4.79	4.12	5.22
4	12.85	13.51	7.31	2.24	11.86
5	31.17	16.70	6.52	*1.23*	15.92
6	20.82	19.79	2.82	5.77	*3.15*
7	24.56	22.67	8.84	6.52	13.96
8	34.21	25.39	**9.55**	7.28	17.82
9	**37.04**	28.00	8.12	**8.62**	19.60
10	27.98	**30.42**	1.80	7.96	**21.31**

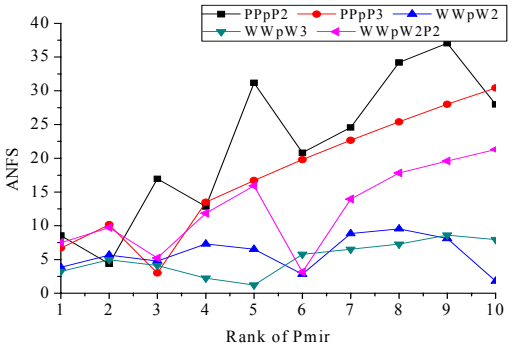

Figure 4. The Relationship of 5 Type of Feature Combinations between ANFS and the Rank of Pmir which Ranks the Pmir Results of the Ten Window-sizes by Descending Order.

Fig. 4 also shows the relationship between the value of ANFS and the rank of Pmir like Fig. 2. If we compare Fig. 2 and Fig. 4, we will find they show quite similar pattern. All the lines ended with the relative more features except WWpW3. Although the maximum value of ANFS is about 37, the value of ANFS of the best result is under 8. It shows that about 78% features is noise or useless in all the features of window size 10.

To our knowledge till now the state of the arts performance is 74.76%(Pmir) [15] on this task. So we could choose 75% as the boundary between good performance and bad performance.

Fig. 5 shows the comparison of ANFS between the good and the bad performance. It only shows the best combination (WWpW2P2) result. We can see the unbalance of features after window size 7. It shows that the naive Bayes classifier can keep making right choice as far as possible though there is more and more noise or useless features along with the window size getting larger. The classifier will be influenced heavily after a window size (e.g. window size 7) or the number of features threshold (e.g. half noise features).

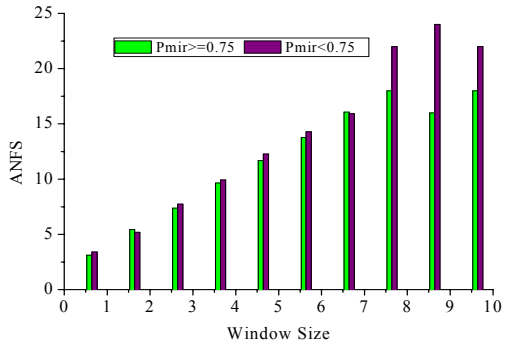

Figure 5. The Comparison of ANFS between the Pmir>=0.75 and Pmir <0.75 Performance.

Like Fig. 5, Fig. 6 shows the comparison of ANFS

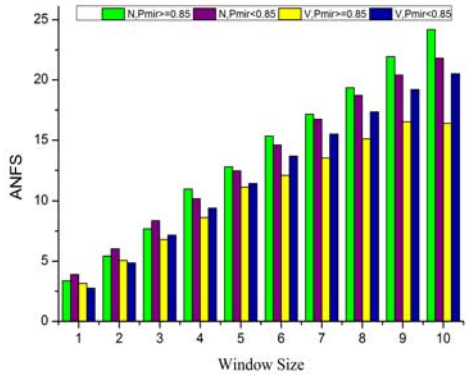

Figure 6. The Comparison of ANFS between the Pmir>=0.85 and Pmir <0.85 Performance in which N is Noun, V is verb.

which the Pmir performance is better than 0.85 or not (if we think Pmir=0.85 is a good enough performance to the real task like machine translation). It only shows the best combination (WWpW2P2) result. It shows that the ANFS is different or adverse between the Nouns and the Verbs. This result is fit our knowledge that Nouns need more topical information than Verbs.

VI. Discussion and Conclusion

In this paper, we use naive Bayes classifier to investigate the relationship between performance and feature in supervised Chinese word sense disambiguation on the Semeval 2007 task #5. The experiment shows that local basic feature provides quite adequate information to do disambiguation.

We think usually that the data sparseness is one of the problems which influence the performance a lot. From the result of experiments above we could see that the influence of data sparseness does not make the test instance lacking features too much to disambiguate overall. If we think about [17, 18] in which they show that local context has enough ability to discriminate different senses, and then the experiments above could be explained and concluded as follows:

Local features can catch enough features to disambiguate senses though it may be helpful if we enlarge window size for more features. Due to the data sparseness, classifier cannot do discrimination rightly when some key local features (like collocation) do not appear in training data set though there are some even many features do appear in the training data set. Larger window size enlarges the probability of key local features while sometimes the loss outweighs the gain if the window size is too large.

In future, we will go deep into investigating the relationship of feature-performance both of Chinese and English. We will do some comparisons between English and Chinese in this regard.

Acknowledgment

This work was supported by the project of National Natural Science Foundation of China (No.60903063) and the Fundamental Research Funds for the Central Universities. The earlier version of this paper has been presented in the Special Session on Natural Language Processing and Intelligent Computation of the 7th International Conference on Computational Intelligence and Security. The author also gratefully acknowledges the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] Edmonds, Philip and Adam Kilgarriff, editors. Special Issue on Evaluating Word Sense Disambiguation Systems. Natural Language Engineering. Cambridge University Press, vol. 8(4), 2002.

[2] Rada Mihalcea and Phil Edmonds, editors. Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, Association for Computational Linguistics Conference (ACL2004). 2004:44~48.

[3] Eneko Agirre, Llúis M`arquez and Richard Wicentowski, editors. SemEval 2007: Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics Conference. Prague, Czech Republic. June, 2007:23~24

[4] Katrin Erk and Carlo Strapparava, editors. SemEval 2010: Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics Conference. Uppsala, Sweden. June, 2010

[5] R. J. Mooney. Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, 1996:82~91

[6] H. T. Ng. Exemplar-Base Word Sense Disambiguation: Some Recent Improvements. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, 1997

[7] A. Fujii, K. Inui, T. Tokunaga, and H. Tanaka. Selective Sampling for Example-Based Word Sense Disambiguation. Computational Linguistics, 1998, 24(4):573~597

[8] T. Pedersen and R. Bruce. Knowledge Lean Word–Sense Disambiguation. In Proceedings of the Conference of the American Association for Artificial Intelligence, AAAI, 1998:800~805

[9] G. Escudero, L. M`arquez, and G. Rigau. A Comparison between Supervised Learning Algorithms for Word Sense

Disambiguation. In Proceedings of the Computational Natural Language Learning Workshop, CoNLL, 2000:31~36

[10] Y. K. Lee and H. T. Ng. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, 2002:41~48

[11] N. Ide and J. Veronis. Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics, 1998,24(1):1-40.

[12] Abolfazl K. Lamjiri, Osama El Demerdash, Leila Kosseim. Simple features for statistical Word Sense Disambiguation. Proceedings of Senseval-3: 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, 2004: 133-136.

[13] R. Mihalcea. 2002. Instance based learning with automatic feature selection applied to word sense disambiguation. In Proceedings of COLING'02, Taiwan.

[14] P. S. Dhillon and L. H. Ungar. Transfer learning, feature selection and word sense disambiguation. In Annual Meeting of the Association of Computational Linguistics, (ACL), August 2009:257-260

[15] He Jing-Zhou, Wang Hou-Feng. Chinese Word Sense Disambiguation Based on Maximum Entropy Model with Feature Selection. Journal of Software. 2010, 21(6):1287-1295.

[16] Peng Jin, Yunfang Wu and Shiwen Yu, "SemEval-2007 Task 05:Multilingual Chinese-English Lexical Sample," Proceedings of theFourth International Workshop on Semantic Evaluations (SemEval-2007),2007,pp.19-23.

[17] Yarowsky, D., "One sense per collocation," In Proceedings of the ARPA Workshop on Human Language Technology, 1993, pp. 266-271.

[18] Pengyuan Liu, Shui Liu and Shiqi Li, "One sense per N-gram," In Proc. of the NLPOE workshop of 2010 IEEE/WIC/ACM International Conference on WI-IAT, 2010:195-198.

[19] Tian Xia and Yanmei Chai. An improvement to TF-IDF:Term Distribution based Term Weight Algorithm. Journal of Computer. 2011, 6(3):413-420