Journal of Chinese Language and Computing 15 (1): (1-12)

Training Multi-Classifiers for Chinese Unknown Word Detection

Chooi-Ling Goh, Masayuki Asahara and Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{ling-g,masayu-a,matsu}@is.naist.jp

Abstract

According to a survey in a corpus, majority of the unknown words in Chinese texts are numbers, time nouns and person names. Detection of numbers and time nouns are trivial tasks. These three types of unknown words may need different feature sets and parameters to achieve optimal results. For example, characters used in Chinese family names help in Chinese person name detection. Therefore, we propose a hierarchical model, with multiple classifiers (same model but different feature sets and parameters) for the detection. We create one classifier for each unknown word type: numbers, time nouns, person names and others. The experimental results show that our model can get higher precision (89%) comparing with only using one classifier (86%). Furthermore, this model enables us to detect the three unknown word types straightaway with high accuracy so that we can concentrate only on POS tag guessing for the other unknown types.

Keywords

Chinese, unknown words, word segmentation, POS tagging, machine-learning, chunking.

1. Introduction

Since written Chinese does not use blank spaces to indicate word boundaries, segmenting Chinese texts becomes an essential task for Chinese language processing. During the process, the occurrences of unknown words have made this task more difficult. They cannot be segmented correctly as they are not found in the dictionary. As for any other languages, even the largest dictionary we may think, will not be able to register all possible words such as proper names, numbers, and etc. This is particularly true in Chinese because almost any character can be used to form new words. As languages evolve, a dictionary will never be complete. Therefore, a proper solution to detect the unknown words is necessary.

The number of unknown words is very much depending on the size of the dictionary used. Certainly, the larger the dictionary, the less the unknown word occurrences in the texts. One can create a dictionary from a tagged corpus but that will not be a proper dictionary. Furthermore, if all words in the tagged corpus are used to create the dictionary, then there will be no unknown word in the texts. Therefore, it is important to define the meaning of unknown words properly. In (Goh et al. 2003), those words that occur only once in the corpus are treated as unknown words in their experiment. However, some people argue that this is not really true because even low frequency words are actually words in some dictionaries but those person names even with high frequency could not be found in a dictionary. A more natural way is by having a proper dictionary. We can consider those words that are not in a proper dictionary to be unknown words. In this case, some words in the corpus are not found in the dictionary and can be used as training data for unknown word detection (Chen and Bai 1997; Fu and Luke 2003). As far as we know, the definitions of words are different by institutions, such as Peking University Corpus, Penn Chinese Treebank and Taiwan Sinica Corpus. Therefore, the dictionary and the tagged corpus used must be consistent. We choose to use the dictionary and tagged corpus provided by Peking University[1]. The dictionary contains 88,910 entries and the corpus has about 1.1 million words.

From our survey in this corpus, about 4.5% of the words are unknown. According to the part-of-speech tags (POS), 29% of the unknown words are numbers (m), 20% are time nouns (t), 17% are person names (nr), and 34% for other types. That is to say, almost 50% of the unknown words are made up from number types (numbers and time nouns). The detection of number types is a trivial task although the production is high. As for Chinese person names, normally they consist of family names and given names, which somehow have similar patterns for recognition. And for foreign names, the characters used are limited to a set of characters which is used to spell the words by pronunciation in the foreign language. In (Goh et al. 2003), a unified solution is proposed for all types of unknown words, but the results are not quite satisfactory. Zhang et al. (2003) propose to detect the unknown words type by type, especially for Chinese person names, transliteration names, location names and organization names. Their purpose is more towards name entity extraction. Our purpose is mainly for word segmentation and we cover more general unknown word rather than named entity. We also propose to detect these unknown words type by type based on the frequency in the corpus (person names, numbers, time nouns and others). We train one classifier for each type of unknown words, in order to get optimal results. Our experimental results show that the precision increases by 3% comparing with only using one classifier, although we do not get any better by recall. However, the merit of this method is that we are able to get the type of unknown words for these three types straightaway, and left only others for POS tag guessing.

2. Previous Work

Some rule-based approaches have been proposed for unknown word detection. In (Chen and Bai 1997; Chen and Ma 2002), unknown word detection rules were generated automatically from a POS tagged corpus. These rules were used to determine whether a monosyllabic word is a word itself or is an unknown word morpheme after an initial segmentation. Then based on the detected unknown word morphemes, (Ma and Chen 2003)

[1] Institute of Computational Linguistics, Peking University, http://www.icl.pku.edu.cn/

proposed a bottom-up merging algorithm to extract the unknown words. They achieved an accuracy of 76% precision and 45% recall using Sinica corpus.

(Fu and Wang 1999; Fu and Luke 2003; Lai and Wu 1999; Shen et al. 1998; Zhang at al. 2002) proposed statistical models for unknown word detection. (Fu and Wang 1999; Fu and Luke 2003) began with an unsupervised method for identification, which is based on the character unigram on certain position to calculate word formation power of characters. They later proposed an integrated approach with word juncture model. They achieved 96.1 F-measure for overall segmentation and 81.2 for unknown word detection using Peking University corpus. Zhang et al. (2002) proposed the use of lexical role tags instead of fine-grained POS tags. For example, tag B as a family name, F as a prefix in a name, K as previous context before a name and etc. They defined different set of role tags for different types of unknown words such as person names, transliteration names, place names and others. Later, role tagging was carried out using Hidden Markov Models. They reported an F-measure of 79.3 for person name detection and 84.69 for transliteration name detection using the same Peking University corpus.

Research has also been done on hybrid models that combine rule-based and statistical based models. Nie et al. (1995) first used maximum matching algorithm to segment the text, and then used some heuristic rules for identification of words with fixed morphological patterns, and finally used some statistical models (character unigram as word formation power and statistical information on the occurrence rates of various character strings) to detect unknown words. They achieved 96% for overall segmentation including unknown words using Hua Xia Wen Zhai corpus. In (Zhou and Lua 1997), 4 steps were used for detection: (1) Word formation tagging by Hidden Markov Models and Viterbi algorithm. (2) N-gram grouping. (3) N-gram overlapping elimination. (4) Phrase elimination by heuristics rules. Their method yielded a precision of 92% and a recall of 70% using a corpus from Lianhe Zaobao, Singapore.

In general, the accuracy is better if one focuses only on certain types of unknown words such as person names, place names or transliteration names, with over 80%. However, for general unknown words, such as common nouns, verbs etc, the accuracy is ranging from 50% to 70%.

3. Proposed Method

3.1 Baseline Method

The basic idea of the method is described in (Goh et al. 2003), which comprises of two statistical models. First, a Hidden Markov Model-based (hereafter HMM) morphological analyzer is used to initially segment and POS tag the text. A post-processing joins continuous numbers and alphabetical characters, as to ease the detection process. Then, the output (segmented words with POS tags) is converted into characters, and each character is assigned with a position tag and POS information. Finally, a Support Vector Machine-based (hereafter SVM) chunker (Kudo and Matsumoto 2001) is used to decide whether the character is inside of an unknown word or not. The process is illustrated in Figure 1.

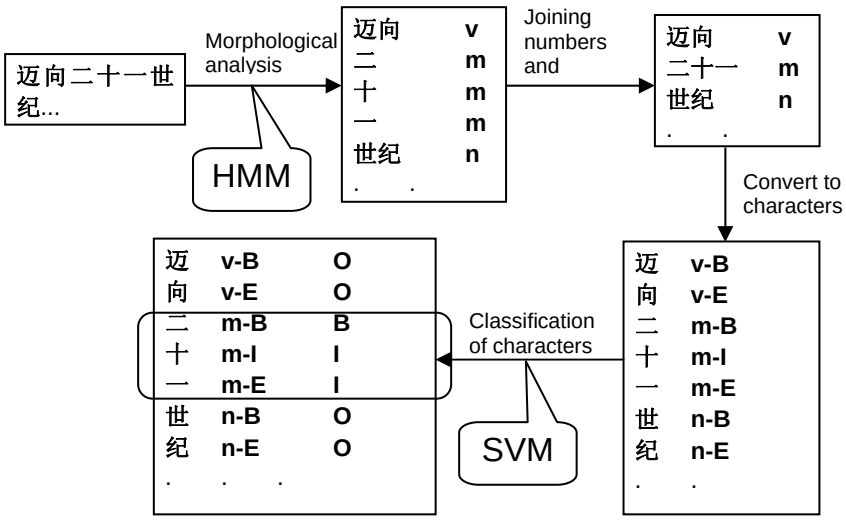

Figure 1 Unknown Word Detection Process – "Looking forward to 21st century"

The features that we use for classification by SVM are as below. Each word will have a POS tag from the output of morphological analysis. This POS tag is subcategorized to include the position of the character in the word. The list of position is shown in Table 1. For example, if a word contains three characters, then the first character is <POS>-B, the second is <POS>-I and the third is <POS>-E. A single character word is tagged as <POS>-S.

Tag	Description
S	one-character word
B	first character in a multi-character word
I	intermediate character in a multi-character word (for words longer than two characters)
E	last character in a multi-character word

Table 1 Position tags in a word

We also define character type as a feature. Strictly saying, there is no character type in Chinese language, but we can group them according to their usage, such as possible family names and transliteration characters (although they still can be used in other places). Currently we have collected 436 family names[2] and 160 transliteration characters[3]. A character is assigned with one of these four types: SURNAME (a family name), FOREIGN (a transliteration character), BOTH (can be used as both family name and transliteration character), or OTHER (not in any type). Finally, a character will have a POS tag with its position tag and a character type to be used as features during classification.

For the output of classification, we only need 3 basic tags to identify the location of unknown words, namely tag "B" (the beginning of an unknown word), tag "I" (inside of an

[2] Chinese family names are almost a fix set, where new family names are rarely created.
[3] Although the characters used for transliteration words are also limited, but they can be increased easily if there exist new pronunciations of new words.

unknown word), or tag "O" (outside of any unknown word). Two characters at both sides of the character are used as context window. Figure 2 shows an illustration of the classification process. The solid box shows the features used to determine the class of the character at location *i*. The characters tagged with "B" and "I" compose an unknown word "秀兰" (Xiulan), a person name.

Location	Character	POS + position tag	Char. type	Class
i-2	周	nr-S	SURNAME	O
i-1	秀	Vg-S	OTHER	B
i	兰	Ng-S	BOTH	I
i+1	夫	n-B	FOREIGN	?
i+2	妻	n-E	OTHER	?

Figure 2 An illustration of classification process – "Zhou Xiulan couple"

SVM is a binary classifier, where only two classes are considered. As we need more than two classes, we have chosen pairwise method to cater for multi-class binary classification. In each classifier, there are $\binom{n}{2}$ binary classifiers, where *n* is the number of classes. By using the method described above, we now define 3 approaches of classification. Note that we regard the $\binom{n}{2}$ binary classifiers as one multi-class classifier in the following section.

3.2 One-Classifier-One-Type Classification

In the first approach, if we regard all the unknown words as one single type of words, then we only need to classify the characters into 3 classes, namely unk-B, unk-I or O. The output will be the unknown words without knowing the types, as shown in Figure 3.

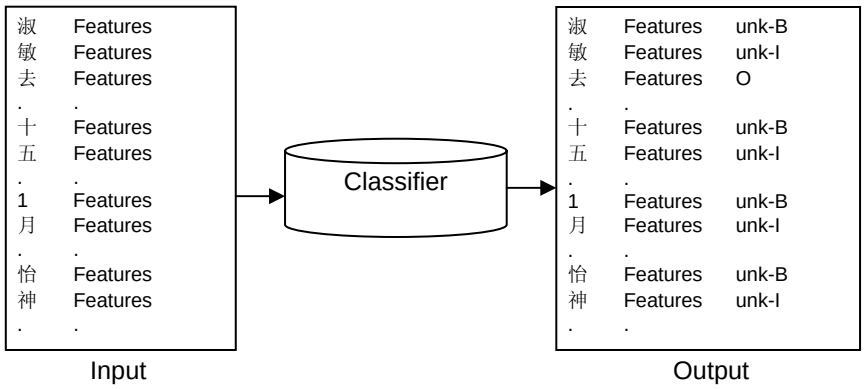

Figure 3 One-Classifier-One-Type

3.3 One-Classifier-Multi-Type Classification

From our survey in the corpus, about 67% of the unknown words are numbers, time nouns and person names. If we straightaway classify these three types during unknown word

detection process, then it will be grateful that we do not need to guess the category for these types anymore. Therefore, in the second approach, instead of only 3 classes, we define 9 classes for classification, namely nr-B, nr-I (for person names), m-B, m-I (for numbers), t-B, t-I (for time nouns), unk-B, unk-I (for others) and O. Figure 4 shows the classification process for this multi-type method.

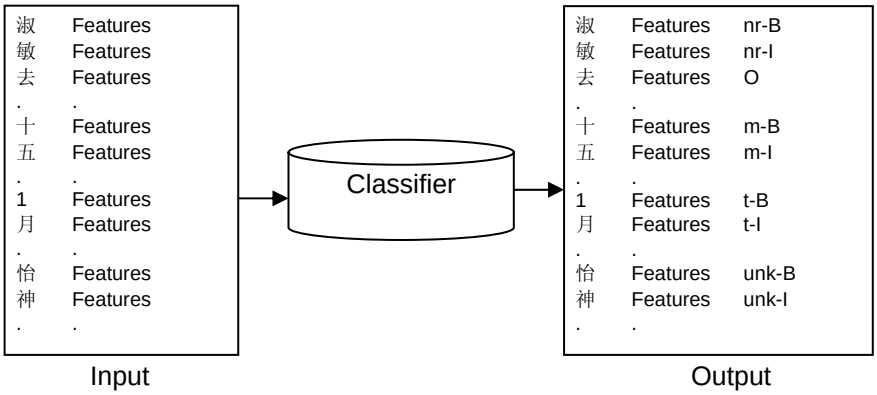

Figure 4 One-Classifier-Multi-Type

3.3 Multi-Classifier-Multi-Type Classification

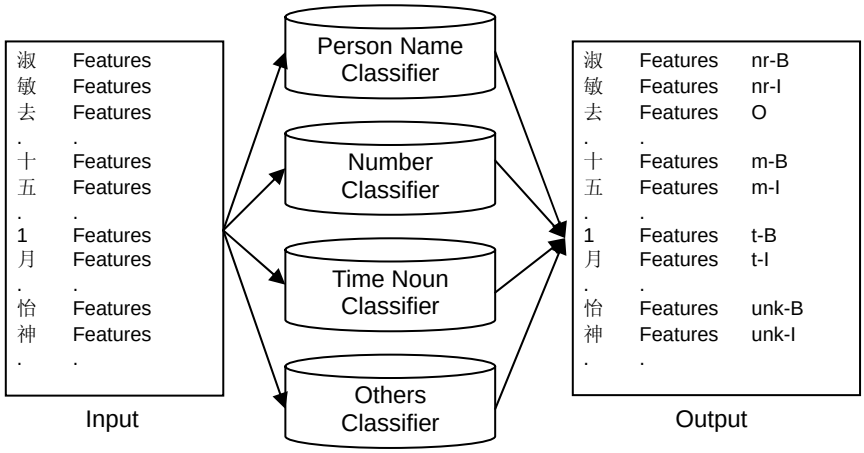

Figure 5 Multi-Classifier-Multi-Type

The third approach comes from the idea in (Zhang et al. 2003), where a hierarchical model was used for different types of unknown words. If we use only one classifier for all types of unknown words, we must use the same features, same parameters for all of them. From our past experiments, we realized that different types of unknown words need different feature sets and parameters. For example, numbers are best detected using only the POS+position tag as features, without the character type, and with forward parsing. Therefore, if we create one classifier for each type of unknown words, and use the best fitted features and parsing

direction, then, we may get optimal results for all of them. We then combine the outputs from each classifier to obtain the final output. This approach is shown in Figure 5. We make no effort to combine the result, but just give priority to the type with higher precision in case there are any conflicts where a character receives more than one tags. As a result, the sequence of priority is "time nouns > numbers > person names > others". In fact, there are not so many overlapping cases, more often with numbers and time nouns. Usually, time nouns are more preferred than numbers. We leave the more intelligent way to combine the outputs as future work.

4. Experimental Results

We use the corpus from Peking University (about 1.1 million words) for our experiment. We divide the corpus randomly into a proportion of 80%/20% for training and testing respectively. The dictionary contains 88,910 entries. Based on this dictionary, there are about 4.5% unknown words in the texts, which spread evenly between training and testing data. The detail distribution of the unknown words is shown in Table 2.

	# of words	# of unknown words	# of distinct unknown words	unknown word rate
Training data	911,551	40,733	17,027	4.47%
Testing data	209,896	10,033	5,201	4.78%
Total	1,121,447	50,766	20,424	4.53%

Table 2 Experimental Corpus

4.1 Effects of Character Types and Parsing Direction

	POS + position tag		POS + position tag & Char. Type	
	Forward	Backward	Forward	Backward
Person Name	82.43	84.18	84.25	**86.04**
Number	**97.06**	96.55	96.99	96.33
Time noun	95.84	97.30	95.79	**97.36**
Others	58.68	**61.97**	58.92	61.61

Table 3 Individual F-measure of Multi-Classifier-Multi-Type

Table 3 shows the individual results produced by each classifier in Multi-Classifier-Multi-Type approach. The first two columns show the results where character types are not used as the features, and the second two columns include character types as the features. Forward and Backward represent the parsing directions (read from the beginning of the sentence or reverse) during the SVM classification. This table shows that each type of unknown words needs different feature sets and parsing directions. Our final output is composed by choosing the best result from each classifier (as indicated in bold face).

4.2 Unknown Word Detection Results

		POS + position tag		POS + position tag & Char. Type	
		Forward	Backward	Forward	Backward
Recall	One-C-One-T	76.92	79.34	77.19	**79.38**
	One-C-Multi-T	75.94	78.38	76.63	78.61
	Multi-C-Multi-T	77.56			
Precision	One-C-One-T	85.94	85.44	85.90	85.24
	One-C-Multi-T	87.09	87.15	86.80	86.51
	Multi-C-Multi-T	**88.91**			
F-measure	One-C-One-T	81.18	82.28	81.31	82.20
	One-C-Multi-T	81.14	82.53	81.40	82.37
	Multi-C-Multi-T	**82.85**			

Table 4 Unknown Word Detection Results

Table 4 shows the overall unknown word detection results. We realize that the Multi-Classifier-Multi-Type approach has done slightly better than others by F-measure. Although the recall is worse compared with One-Classifier-One-Type, the improvement on the precision is significant (at 5% level).

In (Fu and Luke 2004), a class-based language model was introduced for Chinese unknown word identification. A hybrid model which composes of class-based word juncture models and class-based word formation patterns was proposed. The classes refer to the POS tags, which is similar to our method of dividing the unknown words into 4 types. Their method handles both internal word formation features and external contextual information which are important to identify the word boundaries. Since we are using the same corpus, namely the Peking University corpus, we have the same segmentation standard. However, their lexicon is smaller, only contains about 65,000 words (with 6.81% unknown words). They reported the accuracy of unknown word detection of 81.8, 80.8 and 82.5 for F-measure, recall and precision, respectively, and we have 82.85, 77.56 and 88.9. They have higher recall while we have better precision.

4.3 Results by Types of Unknown Words

In One-Classifier-Multi-Type and Multi-Classifier-Multi-Type approaches, there are possibilities that a number is detected as a time noun, or a person name is detected as other, and so on. Therefore, the overall accuracy drops a bit when we evaluate our results by types. Although we do not know the types by One-Classifier-One-Type approach, we just do the calculation by recall for comparison. We could not calculate the precision for One-Classifier-One-Type approach as the types of unknown word are not known. As shown in Table 5, the recall is better by One-Classifier-One-Type approach. However, we get high precision with Multi-Classifier-Multi-Type approach for time nouns (99.24%), numbers (98.29%) and person names (89.09%), and reasonable for others (72.87%).

In (Zhang et al. 2002), role tagging on characters was used for unknown word detection. They reported an F-measure of 79.30 for Chinese person name detection and 84.96 for transliteration name detection. We do not discriminate between Chinese and transliteration person names. We get 86.04 for both types, which is better then theirs. Fu and Luke (2004) get 86.4, slightly better than ours.

		Person Name	Number	Time Noun	Others	Overall
Recall	One-C-One-T	**(86.78)**	**(97.19)**	**(96.44)**	**(59.09)**	**(79.34)**
	One-C-Multi-T	80.25	96.48	95.70	56.26	77.45
	Multi-C-Multi-T	83.20	97.00	95.55	53.95	76.97
Precision	One-C-One-T	n.a.	n.a.	n.a.	n.a.	n.a.
	One-C-Multi-T	85.82	96.26	99.24	70.74	86.11
	Multi-C-Multi-T	**89.09**	**98.29**	**99.24**	**72.87**	**88.22**
F-measure	One-C-One-T	n.a.	n.a.	n.a.	n.a.	n.a.
	One-C-Multi-T	83.13	96.37	**97.44**	**62.67**	81.56
	Multi-C-Multi-T	**86.04**	**97.64**	97.36	62.00	**82.21**

Table 5 Results by Types of Unknown Words[4]

4.4 Results of Real Unknown Words

The training of our models requires a dictionary and a tagged corpus. Since the dictionary and the corpus are two different data sources, it also means that not all words in the training corpus are in the dictionary. Some people argued that although the unknown words are not in the dictionary, they probably have been seen in the training corpus. In this case, it is not a surprise that they can be detected correctly. Therefore, we also make an evaluation on those unknown words that occur only in the testing data but not in the training corpus. We refer to these unknown words as real unknown words. There are 4,427 (44%) real unknown words in the testing data. Table 6 shows the results for real unknown words. We get about 60% of recall with all approaches. The distribution of real unknown words is as below: person names (20%), numbers (13%), time nouns (1%) and others (66%). Originally the numbers and the time nouns have the highest unknown word distribution but they are not real. Most of them have been seen in the training data, therefore the detection is easier. The most difficult one is with the others type, which has the highest real unknown word distribution. We need more attention on this type in the future work.

		POS + position tag		POS + position tag & Char. Type	
		Forward	Backward	Forward	Backward
Recall	One-C-One-T	58.69	63.27	59.27	**63.43**
	One-C-Multi-T	57.40	61.44	58.69	61.73
	Multi-C-Multi-T	60.18			

Table 6 Results by Recall of Real Unknown Words

5. Error Analysis

We get quite satisfactory precision (88.91%) by using the proposed method. As there is no

[4] We show results of POS+position tag as features, with backward parsing for One-C-One-T and One-C-Multi-T as they have the best F-measures overall. On the contrary, the best result from each classifier is chosen to compose the final results for Multi-C-Multi-T.

single standard definition of words in Chinese, we could hardly say that the gold data is perfectly correct. Therefore, human judgment is necessary. Since there are not so many incorrectly detected words, we have gone through all the errors to examine what kind of mistakes has been made.

Surprisingly, there are quite a number of words in the error list which are said to be acceptable by human judgment. Out of 971 incorrect words, 380 words are acceptable. Appendix A shows some examples of these words. Some of these errors happen because of the non-standardization of segmentation. For example, "艺术史" (the history of art) is segmented as one word and "京剧/ 史/" (the history of Peking opera) is segmented as two words. There are also human errors like "史/ 泰龙/" where the name is segmented into a family name and a given name but our system has extracted it as one segment which is a correct one[5]. There are also some collocation phrases such as "大肚佛" (big stomach Buddha) and "百鸟朝凤" (hundred birds facing the phoenix), which to some people they can be considered as words too. If we consider these errors to be correct ones, then our method has achieved 93.24% precision. Again, we can conclude that our method can achieve high precision for unknown word detection.

6. Effects on Overall Segmentation

By replacing the new detected words with the original segmentation, we get the final segmentation. We get only 90.40 points F-measure using solely HMM. After unknown word detection by using Multi-Classifier-Multi-Type approach, we get 96.59 points, an improvement of 6.19 points.

In year 2003, a competition for Chinese word segmentation was carried out in SIGHAN[6] workshop to compare the accuracy of various methods (Sproat and Emerson 2003). It used to be difficult to compare the accuracy of various systems because the experiments had been done on different corpora. Therefore, this bakeoff intended to standardize the training and testing corpora, so that a fair evaluation could be made. The segmentation results of the open test for Peking University dataset are ranging from 88.6-95.9 of F-measure, and the recalls for unknown words are 50.3-79.9%. We did not retrain our model with their training materials, but just what we have on hand to run on the testing data. There are 1,253 (7.3%) unknown words in the test data based on our dictionary. We get an F-measure of 88.32 for segmentation by using only HMM and 95.11 after unknown word detection. The unknown word recall is 75.74% and precision is 89.19% according to our dictionary and the recall is 80.2% according to bakeoff dictionary.

7. Conclusions

As a conclusion, we have improved the unknown word detection result (F-measure) by detecting the unknown words type by type because the precision obtained is higher. Although setting up more classifiers consumes more resources and time, the advantage of this method is that we could get the types of numbers, time nouns and person names

[5] "史泰龙" (Stallone - an American actor) in fact is a transliteration foreign name but not a Chinese name although the first character can be a family name.

[6] A Special Interest Group of the Association of Computational Linguistics, http://www.sighan.org/.

straightaway. Only the others are left for POS tag guessing. As the precision is high, we are confident that the detected unknown words can be used to enlarge our lexicon. Finally, the detected unknown words also help to improve the accuracy of word segmentation.

References

[1] Chen, K.-J. and Bai, M.-H., 1997, Unknown Word Detection for Chinese by a Corpus-based Learning Method, In *Proceedings of ROCLING X*, pp. 159–174.

[2] Chen, K.-J. and Ma, W.-Y., 2002, Unknown Word Extraction for Chinese Documents, In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, vol. 1, pp. 169–175.

[3] Fu, G. and Luke, K.K., 2003, An Integrated Approach for Chinese Word Segmentation, In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC 17)*.

[4] Fu, G. and Luke, K.K., 2004, Chinese Unknown Word Identification Using Class-based LM, In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 262–269.

[5] Fu, G. and Wang, X., 1999, Unsupervised Chinese Word Segmentation and Unknown Word Identification, In *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS)*.

[6] Goh, C.-L., Asahara, M. and Matsumoto, Y., 2003, Chinese Unknown Word Identification Using Character-based Tagging and Chunking, In *Companion Volume to the Proceedings of ACL 2003, Interactive Poster/Demo Sessions*, pp. 197–200.

[7] Kudo, T. and Matsumoto, Y., 2001, Chunking with Support Vector Machines, In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 192–199.

[8] Lai, Y.-S. and Wu, C.-H., 1999, Unknown Word and Phrase Extraction Using a Phrase-Like-Unit- Based Likelihood Ratio, In *Proceeding of International Conference on Computer Processing of Oriental Languages (ICCPOL)*, pp. 5–9.

[9] Ma, W.-Y. and Chen, K.-J., 2003, A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction, In *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 31–38.

[10] Nie, J.-Y., Hannan, M.-L. and Jin, W., 1995, Unknown Word Detection and Segmentation of Chinese Using Statistical and Heuristic Knowledge, *Communications of COLIPS*, vol.5, pp. 47–57.

[11] Shen, D., Sun, M. and Huang, C., 1998, The Application & Implementation of Local Statistics in Chinese Unknown Word Identification, *Communications of COLIPS*, vol. 8.

[12] Sproat, R. and Emerson, T., 2003, The First International Chinese Word Segmentation Bakeoff, In *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 133–143.

[13] Zhang, H.-P., Liu, Q., Zhang, H., and Cheng, X.-Q., 2002, Automatic Recognition of Chinese Unknown Words Based on Roles Tagging, In *Proceedings of First SIGHAN Workshop on Chinese Language Processing*.

[14] Zhang, H.-P., Yu, H.-K., Xiong, D.-Y., and Liu, Q., 2003, HHMM-based Chinese

Lexical Analyzer ICTCLAS, In *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pp. 184–187.

[15] Zhou, G.-D. and Lua, K.-T., 1997, Detection of Unknown Chinese Words Using a Hybrid Approach, *Computer Processing of Oriental Language*, vol. 11, no. 1, pp. 63–75.

Appendix A Examples of Acceptable Detected Unknown Words

1. 对近代京剧史(Peking opera history)、论研究与批评，
2. 现代人(modern people)，何处是家乡？
3. 处事方式也非常深圳化(Shenzhen-ized)，
4. 田泳的四川话(Sichuan dialect)已经讲得很结巴，
5. 中国人是恋乡(loving home)的民族。
6. 这是“文化视角”的第二篇，上篇(previous article)话题是“
7. 我疼得勾着腰在地上蹦跳(jump up and down)，嘴里啊啊地叫着，
8. 老将军配有(possess with)专车，但他很少乘坐。
9. 湖北万名税官(tax officer)竞争上岗
10. 亚龙湾国家级旅游开发区、通什民族文化村(village of culture)、
11. 还趁着圩日在集市摆摊(set up a counter)咨询，踏进农户家现场
12. 啤酒店(beer house)里充满欢声笑语，弥漫着酒气芳香。
13. 总是让先来者登车，极少发生争抢、推挤(pushing)的现象。
14. 目前各华埠的舞龙(dragon dance)舞狮队正在抓紧彩排；
15. 幼儿园小朋友们举行了一年一度的『鸟婚(wedding of birds)』仪式。
16. 古人讲求心斋(purity of heart)，曾说一位工匠在雕刻时
17. 黑龙江呼兰河域，万历四十七年(year 47)被努尔哈赤所并，
18. 杂技团在北京木樨地搭起一座临时马戏棚(circus tent)演出，
19. 所谓练字就是练神(exercising the mind)，练神就是练心，
20. 每家都要把常年贴在灶头(kitchen place)上的灶公公像揭下来