Multi-Classifier for Chinese Unknown Word Detection

GOH Chooi Ling ASAHARA Masayuki MATSUMOTO Yuji
Graduate School of Information Science, Nara Institute of Science and Technology.
{ling-g,masayu-a,matsu}@is.aist-nara.ac.jp

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. The occurrences of unknown words have made the task more difficult, because they cannot be segmented correctly. 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 a new word. 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. Therefore, we will have to decide to what extent we want to detect the unknown words. Since we have to be consistent between the dictionary and the corpus used for our experiment, meaning the definition of words are the same, we choose to use the dictionary and corpus provided by Peking University. The dictionary contains 88,910 entries and the corpus have about 1 million words.

From our survey in this corpus, about 4.5% of the words are unknown. According to the part-of-speech tag (POS), 29% of the unknown words are numbers, 20% are time nouns, 17% are person names, and 34% for other types. In other words, almost 50% of the unknown words are made up from number types (numbers and time nouns), which is a trivial task for detection. As for person names, normally they consist of family names and given names, which somehow have similar patterns for recognition. In [Goh et al., 2003], a unified solution is proposed for all types of unknown words, but the results are not quite satisfactory. Therefore, we propose a method that will detect these unknown words type-by-type (person names, numbers, time nouns and others), by training one classifier for each type of unknown words, in order to get optimal results. Our experimental results show that the precision increased by 2% comparing with only using one classifier. Besides, the merit of this method is that we could get the type of unknown words for these three types, and left only others for POS tag guessing.

2 Proposed Method

2.1 Baseline Method

The basic of the method is the same as described in [Goh et al., 2003], which comprises of two statistical models. First, a Hidden Markov Model-based morphological analyzer [Matsumoto et al., 2002] is used to initially segment and POS tag the text. A post-processing will join continuous characters of type number and alphabet. Then, the output, (segmented words with POS tags), is converted into characters, and we assign each character with some features. Finally, a Support Vector Machine-based classifier [Kudo and Matsumoto, 2001] is used to detect the location of unknown words. The process is illustrated in Figure 1.

The features that we use for classification are as below. From the output of morphological analysis, each word will have a POS tag. 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 exam-

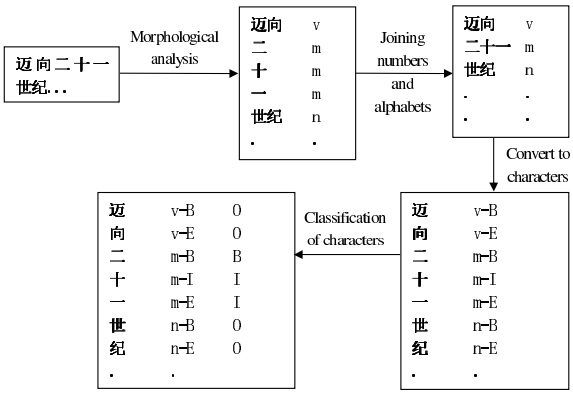

Figure 1: Unknown Word Detection Process - 'Looking forward to 21st century'

ple, 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.

Table 1: Position tags in a word

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

We also define character type as feature. Strictly saying, there is no character type in Chinese character, but we can group them according to its 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 and 160 transliteration characters. A character is assigned with one of these four types: SURNAME (a family name), FOREIGN (a transliteration character), BOTH (can be used as family name or 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.

As for the output of classification, we only need 3 tags to identify the location of unknown words, namely tag "B" (the beginning of an unknown word), tag "I" (inside an unknown word), or tag "O" (outside of an unknown word). Two characters at each side of the character are used as context window. We can either parse the sentence forwardly or backwardly. 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.

Loc.	Char.	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	O
i+2	妻	n-E	OTHER	O

Figure 2: An illustration of classification process - 'Zhou Xiulan couple'

SVM is known for binary classification, where only two classes involved. 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 will now define 3 ways of classification.

2.2 One-Classifier-One-Type Classification

If we regard all the unknown words as one single type of unknown 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 to which type they are referred to, as shown in Figure 3.

2.3 One-Classifier-Multi-Type Classification

As mentioned in previous section, about 67% of the unknown words are of types numbers,

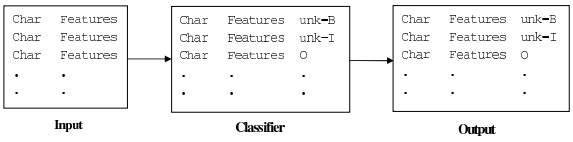

Figure 3: One-Classifier-One-Type Classification

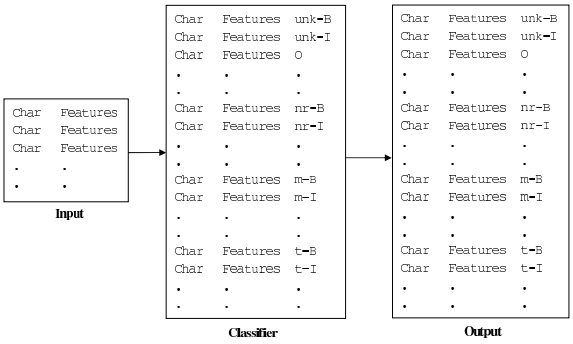

Figure 4: One-Classifier-Multi-Type Classification

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, instead of only 3 classes, we will 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.

2.4 Multi-Classifier-One-Type Classification

Our idea came from [Zhang et al., 2003], where a hierarchical model is used for different types of unknown word detection, such as person names, location names and organization names. If we use only one classifier for all types of unknown words, we will have to use the same features, same parameters for all of them. From our past experience, we realized that different types of unknown words need different features and parameter. For example, numbers are best detected by using only the POS+position tag as features, without the character type, and by using forward parsing.

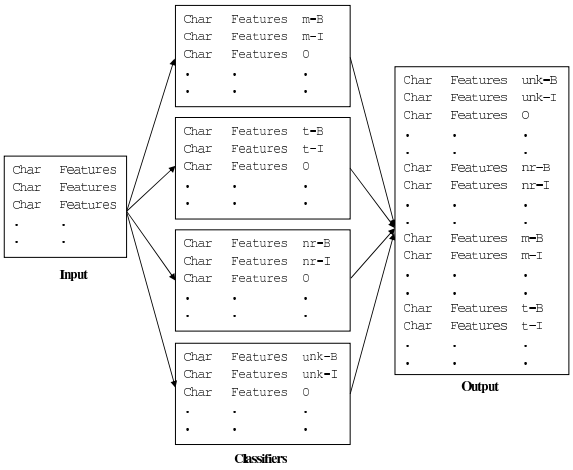

Figure 5: Multi-Classifier-One-Type Classification

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 can combine the outputs from each classifier to obtain the final output. This method is shown in Figure 5. We make no effort to combine the result, but just give priority to the type with higher precision. As a result, the sequence of priority is "time nouns > numbers > person names > others".

3 Experimental Results

We divide the corpus into a proportion of 80%/20% for training and testing respectively.

Table 2 shows the individual results produced by each of the classifiers from Multi-Classifier-One-Type approach. From this table, we realize that each type of unknown words needs different features and parsing direction. Therefore, our final result is composed by choosing the best result from each classiffier. Table 3 shows the experimental results for all the approaches.

The brackets show the results where the types of unknown words are not considered. It is because in One-Classifier-Multi-Type and Multi-Classifier-One-Type, there are possiblilities that a number is treated as time noun, or a person name is treated as general unknown word, and so on. Therefore, we evaluate our

Table 2: Individual F-measure of Multi-Classifier-One-Type

	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: Experimental 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	75.09 (75.94)	77.45 (78.38)	75.65 (76.63)	**77.66** (78.61)
	Multi-C-One-T	76.97 (77.56)			
Precision	One-C-One-T	(85.94)	(85.44)	(85.90)	(85.24)
	One-C-Multi-T	86.12 (87.09)	86.11 (87.15)	85.69 (86.80)	85.47 (86.51)
	Multi-C-One-T	**88.22 (88.91)**			
F-measure	One-C-One-T	(81.18)	(82.28)	(81.31)	(82.20)
	One-C-Multi-T	80.23 (81.14)	81.56 (82.53)	80.36 (81.40)	81.38 (82.37)
	Multi-C-One-T	**82.21 (82.85)**			

results in these two ways. For example, if a number is output as a time noun, it is considered correct with the first one (in brackets), but is considered wrong with the other one. We realize that Multi-Classifier-One-Type has done slightly better than others by F-measure. However, the recall is not good compared with One-Classifier-One-Type, the winning point here is the precision. We get quite satisfactory precision for time nouns (99.24%), numbers (98.29%) and person names (89.09%), and resonable for general unknown words (72.87%).

4 Conclusion

As a conclusion, Multi-Classifier-One-Type has improved slightly on the results by F-measure, as the precision obtained is higher. Although setting up more classifiers is more resource and time consuming, but the advantage of this method is that we could get the types of numbers, time nouns and person names straightaway, and left only the others for POS tag guessing, yet the accuracy is maintained.

References

[Goh et al., 2003] Goh, Chooi Ling, Asahara, Masayuki, and Matsumoto, Yuji. (2003). Chinese Unknown Word Identification Using Character-based Tagging and Chunking. In *Companion Volume to the Proceedings of ACL 2003, Interactive Poster/Demo Sessions*, pages 187–200.

[Kudo and Matsumoto, 2001] Kudo, Taku and Matsumoto, Yuji. (2001). Chunking with Support Vector Machines. In *Proceedings of NAACL 2001.*

[Matsumoto et al., 2002] Matsumoto, Yuji, Kitauchi, Akira, Yamashita, Tatsuo, Hirano, Yoshitaka, Matsuda, Hiroshi, Takaoka, Kazuma, and Asahara, Masayuki. (2002). *Morphological Analysis System ChaSen version 2.2.9 Manual.* Nara Institute of Science and Technology.

[Zhang et al., 2003] Zhang, Hua-Ping, Liu, Qun, Cheng, Xue-Qi, Cheng, Zhang, Hao, and Yu, Hong-Kui (2003). Chinese Lexical Analysis Using Hierarchical Hidden Markov Model In *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, pages 63–70.