Chinese Unknown Word Identification Using Character-based Tagging and Chunking

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

Abstract

Since written Chinese has no space to delimit words, segmenting Chinese texts becomes an essential task. During this task, the problem of unknown word occurs. It is impossible to register all words in a dictionary as new words can always be created by combining characters. We propose a unified solution to detect unknown words in Chinese texts. First, a morphological analysis is done to obtain initial segmentation and POS tags and then a chunker is used to detect unknown words.

1 Introduction

Like many other Asian languages (Thai, Japanese, etc), written Chinese does not delimit words by spaces and there is no clue to tell where the word boundaries are. Therefore, it is usually required to segment Chinese texts prior to further processing. Previous research has been done for segmentation, however, the results obtained are not quite satisfactory when unknown words occur in the texts. An unknown word is defined as a word that is not found in the dictionary. As for any other language, all possibilities of derivational morphology cannot be foreseen in the form of a dictionary with a fixed number of entries. Therefore, proper solutions are necessary for the detection of unknown words.

Along traditional methods, unknown word detection has been done using rules for guessing their location. This can ensure a high precision for the detection of unknown words, but unfortunately the recall is not quite satisfactory. It is mainly due to the Chinese language, as new patterns can always be created, that one can hardly efficiently maintain the rules by hand. Since the introduction of statistical techniques in NLP, research has been done on Chinese unknown word detection using such techniques, and the results showed that statistical based model could be a better solution. The only resource needed is a large corpus. Fortunately, to date, more and more Chinese tagged corpora have been created for research purpose.

We propose an "all-purpose" unknown word detection method which will extract person names, organization names and low frequency words in the corpus. We will treat low frequency words as general unknown words in our experiments. First, we segment and assign POS tags to words in the text using a morphological analyzer. Second, we break segmented words into characters, and assign each character its features. At last, we use a SVM-based chunker to extract the unknown words.

2 Proposed Method

We shall now describe the 3 steps successively.

2.1 Morphological Analysis

ChaSen is a widely used morphological analyzer for Japanese texts (Matsumoto et al., 2002). It achieves over 97% precision for newspaper articles. We assume that Chinese language has similar characteristics with Japanese language to a certain extent, as both languages share semantically heavily loaded characters, i.e. kanji for Japanese, hanzi for Chinese.

Based on this assumption, a model for Japanese may do well enough on Chinese. This morphological analyzer is based on Hidden Markov Models. The target is to find the word and POS sequence that maximize the probability. The details can be found in (Matsumoto et al., 2002).

2.2 Character Based Features

Character based features allow the chunker to detect unknown words more efficiently. It is especially the case when unknown words overlap known words. For example, *ChaSen* will segment the phrase "邓颖超生前..." (Deng Yingchao before death) into "邓/颖/超生/前/..." (Deng Ying before next life). If we use word based features, it is impossible to detect the unknown person name "邓颖超" because it will not break up the word "超生" (next life). Breaking words into characters enables the chunker to look at characters individually and to identify the unknown person name above.

The POS tag from the output of morphological analysis is subcategorized to include the position of the character in the word. The list of positions 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.

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

Character types can also be used as features for chunking. However, the only information at our disposal is the possibility for a character to be a family name. The set of characters used for transliteration may also be useful for retrieving transliterated names.

2.3 Chunking with Support Vector Machine

We use a Support Vector Machines-based chunker, *YamCha* (Kudo and Matsumoto, 2001), to extract unknown words from the output of the morphological analysis. The chunker uses a polynomial kernel of degree 2. Please refer to the paper cited for details.

Basically we would like to classify the characters into 3 categories, B (beginning of a chunk), I (inside a chunk) and O (outside a chunk). A chunk is considered as an unknown word in this case. We can either parse a sentence forwardly, from the beginning of a sentence, or backwardly, from the end of a sentence. There are always some relationships between the unknown words and the their contexts in the sentence. We will use two characters on each left and right side as the context window for chunking.

Figure 1 illustrates a snapshot of the chunking process. During forward parsing, to infer the unknown word tag "I" at position *i*, the chunker uses the features appearing in the solid box. Reverse is done in backward parsing.

3 Experiments

We conducted an open test experiment. A one-month news of year 1998 from *the People's Daily* was used as the corpus. It contains about 300,000 words (about 1,000,000 characters) with 39 POS tags. The corpus was divided into 2 parts randomly with a size ratio for training/testing of 4/1.

All person names and organization names were deleted from the dictionary for extraction. There were 4,690 person names and 2,871 organization names in the corpus. For general unknown word, all words that occurred only once in the corpus were deleted from the dictionary, and were treated as unknown words. 12,730 unknown words were created under this condition.

4 Results

We now present the results of our experiments in recall, precision and F-measure, as usual in such experiments.

4.1 Person Name Extraction

Table 2 shows the results of person name extraction. The accuracy for retrieving person names was quite satisfiable. We could also extract names overlapping with the next known word. For example, for the sequence "邓/Ng 颖/Ag 超生/v 前/f 使用/v 过/v

Position	Char.	POS(best)	Family Name	Chunk
i - 2	江	n-S	Y	B
i - 1	泽	Ag-S	N	I
i	民	Ng-S	N	I
i + 1	主	n-B	N	O
i + 2	席	n-E	Y	O

Figure 1: An illustration of chunking process ‘President Jiang Zemin’

的/u 物品/n” (The things that Deng Yingchao used before death), the system was able to correctly retrieve the name “邓颖超” although the last character is part of a known word “超生”. It could also identify transliterated foreign names such as “菲拉利” (Filali)[1], “弗兰克.卡恩” (Frank Kahn)[2], “伯瑞恩” (Boraine)[3], etc.

Table 2: Results for person name extraction

	Recall	Precision	F-measure
For	***83.37***	86.06	***84.69***
Back	79.45	***86.84***	82.98
+FamN/For	**85.81**	87.52	86.66
+FamN/Back	84.44	**89.25**	**86.78**

For - forward parsing, Back - backward parsing, +FamN - add family name as feature

Furthermore, it was proved that if we have the information that a character is a possible character for family name, it helps to increase the accuracy of the system, as the last two rows of Table 2 show.

Some person names that could not be extracted are such as in the sequence “老/a 张/q 仍/d 很/d 乐观/a” (Lao Zhang is still very positive). In this example, “老张仍” was extracted as a person name, however the right name is “老张” only. This is because the next character of the unknown ones is a monosyllabic word, thus there is higher possibility that it is joined with the unknown word as a chunk. Another example is “户/q 主张/v 宝/n 军/n” (The owner Zhang Baojun), where the family name “张” has been joined with the known word “主张” (suggest) before it. Therefore, the person name “张宝军” was not extracted (the correct segmentation should be “户主/n 张宝军/nr”).

[1]the former Prime Minister of Morocco

[2]Western Cape Attorney General of South Africa in 1998

[3]Truth Commission Deputy Chairman in 1998

4.2 Organization Name Extraction

Table 3 shows the result for organization name extraction. Organization names are best extracted by using backward parsing. This may be explained by the fact that, in Chinese, the last section of a word is usually the keyword showing that it is an organization name, such as, “公司” (company), “集团” (group), “机构” (organization), etc. By parsing the sentence backwardly, these keywords will be first looked at and will have higher possibility to be identified.

Table 3: Results for organization name extraction

	Recall	Precision	F-measure
For	54.66	70.85	61.71
Back	***63.25***	***79.36***	***70.40***

There are quite a number of organization names that could not be identified. For example, “襄樊市志达出租汽车公司” (Xiangfan City Zhida Car Rental Company), “上海庄妈妈净菜社服务有限公司” (Shanghai Zhuang Mother Jingcaishe Service Limited Company). This could be because the names are too long, and the 2 characters left and right context window is not enough for the system to make a correct judgement.

4.3 Unknown Words Extraction in General

As mentioned above, we deleted all words that occur only once from the dictionary to artificially create unknown words. Those “unknown words” included common nouns, verbs, numbers, etc. The results for this experiment are shown in Table 4.

In general, around 60% accuracy (F-measure) was achieved for unknown word detection, and backward parsing seems doing slightly better than forward parsing.

Table 4: Results for unknown word extraction in general

	Recall	Precision	F-measure
For	56.77	***65.28***	60.70
Back	***58.43***	63.82	***61.00***

5 Comparison with Word Based Chunking

As to ensure that character based chunking is better than word based chunking, we have carried out an experiment with word based chunking as well.

The results showed that character based chunking yields better results than word based chunking. The f-measure (<word based> vs <character based>) for person name extraction is (81.28 vs 84.69), for organization name is (67.88 vs 70.40), and for general unknown word is (56.96 vs 61.00) respectively.

6 Comparison with Other Works

There are basically two methods to extract unknown words, statistical and rule based approaches. In this section, we compare our results with previous reported work.

(Chen and Ma, 2002) present an approach that automatically generates morphological rules and statistical rules from a training corpus. They use a very large corpus to generate the rules, therefore the rules generated can represent patterns of unknwon words as well. While we use a different corpus for the experiment, it is difficult to perform a comparison. They report a precision of 89% and a recall of 68% for all unknown word types. This is better than our system which achieves only 65% for precision and 58% for recall.

In (Shen et al., 1997), local statistics information are used to identify the location of unknown words. They assume that the frequency of the occurences of an unknown word is normally high in a fixed cache size. They have also investigated on the relationship between the size of the cache and its performance. They report that the larger the cache, the higher the recall, but not the case for precision. They report a recall of 54.9%, less than the 58.43% we achieved.

(Zhang et al., 2002) suggest a method that is based on role tagging for unknown words recognition. Their method is also based on Markov Models. Our method is closest to the role tagging idea as this latter is also a sort of character based tagging. The extension in our method is that we first do morphological analysis and then use chunking based on SVM for unknown word extraction. In their paper, they report an F-measure of 79.30% in open test environment for person name extraction. Our method seems better with an F-measure of 86.78% for person name extraction (for both Chinese and foreign names).

7 Conclusion

We proposed an "all-purpose" method for Chinese unknown word detection. Our method is based on an morphological analysis that generates segmentations and POS tags using Markov Models, followed by a chunking based on character features using Support Vector Machines. We have also shown that character based features yields better results than word based features in the chunking process. Our experiments showed that the proposed method is able to detect person names and organization names quite accurately and is also quite satisfactory even for low frequency unknown words in the corpus.

References

Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown Word Extraction for Chinese Documents. In *COLING-2002: The 19th International Conference on Computational Linguistics Vol. 1*, pages 169–175.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with Support Vector Machines. In *Proceedings of NAACL 2001*.

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

Dayang Shen, Maosong Sun and Changning Huang. 1997. The application & implementation of local statistics in Chinese unknown word identification. In *COLIPS*, Vol. 8. (in Chinese).

Kevin Zhang (Hua-Ping Zhang), Qun Liu, Hao Zhang, and Xue-Qi Cheng. 2002. Automatic Regcognition of Chinese Unknown Words Based on Roles Tagging. In *Proceedings of 1st SIGHAN Workshop on Chinese Language Processing*.