中国語の未知語抽出における形態素解析とチャンキングの利用

ゴー.チュイリン　浅原 正幸　松本 裕治

中国語のテキストの解析には、わかち書きにより単語の境界を明らかにする必要があるが、未知語が多くなるにつれ、解析が困難となる。中国語は自由に文字を組み合わせて単語を作ることができるため、全ての未知語を辞書に登録するのは不可能である。本研究では、冗長形態素解析とSVMに基づくチャンキングの組み合わせにより正確に未知語を抽出する手法を提案する。実験の結果から、人名と組織名の抽出は高精度を得たが、一般的な未知語については、それらよりも精度が低かった。

キーワード：中国語、未知語、わかち書き、形態素解析、チャンキング

Chinese Unknown Word Identification Based on Morphological Analysis and Chunking

GOH Chooi Ling　Masayuki ASAHARA　Yuji MATSUMOTO

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 based on n-best morphological analysis and SVM-based chunking. Our target is to detect person names, organization names, and general unknown words. The experiments and results obtained have shown that this approach can yield quite high precision and recall for person name and organization name detection but moderate results for the other.

Keywords : Chinese, unknown words, word segmentation, morphological analysis, chunking

1 Introduction

Like many other Asian languages (Thai, Japanese, etc), written Chinese does not delimit words by, for instance, spaces (unlike English). And there is no clue to tell where the word boundaries are as there is only one single type of characters that is the hanzi (unlike Japanese) and one single form for this type of characters (unlike Arabic). 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. In other words, it is out of vocabulary. As for any other language, no dictionary as big as we may think, will be able to register all geographical names, person names, organization names etc. And in Chinese too, 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.

†奈良先端科学技術大学院大学 情報科学研究科
‡Graduate School of Information Science, Nara Institute of Science and Technology

We propose an "all-purpose" statistical 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. In a first step, we segment and assign the n-best POS tags to words in the text using a statistical morphological analyzer. In a second step, we break segmented words into characters, and assign each character its n-best POS tags. In a last step, we use a Support Vector Machine-based chunker to extract all the unknown words. The same method can be found in [Asahara and Matsumoto, 2003] for Japanese unknown word detection.

2 Proposed Method

We shall now describe the above 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 language share semantically heavily loaded characters, i.e. kanji for Japanese, hanzi for Chinese. Based on this assumption, a Japanese morphological analyzer may do well enough on Chinese for our purpose. Although we will not do a full morphological analysis for Chinese but we will just adopt *ChaSen* to do initial segmentation and POS tagging.

This morphological analyzer is based on Markov Models. Morphological analysis is defined as the determination of the POS tag sequence if a segmentation into a word sequence is given. The goal is to find the POS sequence and word sequence that maximize the following probability:

$$T = \arg\max_{T} P(T|W)$$

By using Bayes's rule, $P(T|W)$ can be decomposed as the product of the tag and word probablilities.

$$\arg\max_{T} P(T|W) = \arg\max_{T} P(W|T)P(T)$$

We make the following approximations that the tag probability is determined by the preceding tag only and that the word probability is determined by the tag of the word. The probabilities are estimated from the frequencies in a tagged corpus using Maximum Likelihood Estimation. With these parameters, the most probable tag and word sequences are determined using the Viterbi algorithm.

In practice, we use log likelihood as cost. Maximizing the probability is equivalent to minimizing the cost. N-best analysis means that we will produce the top n-best answers which fall within a certain cost. The n-best answers are picked for each character in the order of the accumulated cost from the beginning of the sentence. If the difference between the cost of the best answer and the n-best answer exceeds a predefined cost threshold, we abandon the n-best answer. The cost threshold is defined as the lowest probability in all events which occur in the training data.

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.

As the output of n-best analysis, each character receives a number of POS tags. This POS tag information 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*, to extract unknown words from the output of the n-best morphological analysis. Basically, SVM are binary classifiers that search for hyperplanes with the largest possible margin between positive and negative samples. *YamCha* extends binary classification to n-class classification because for NLP purposes, we would normally want to classify into several classes, as in the case for POS tags or base phrase chunking. Mainly two straightforward methods are used for this extension, the "one-vs-rest method" and the "pairwise method". In the "one-vs-rest method", n binary classifiers compare one class with the rest of the classes. In the "pairwise method", we use $\binom{n}{2}$ binary classifiers, between all pairs of classes. We simply chose "pairwise method" in this experiment because it is more efficient during the training. Details of the system can be found in [Kudoh and Matsumoto, 2001].

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.

3.1 Data Preparation

With regard to this corpus, two ways of preparing the dictionary were used. A first experiment was for the extraction of person names and organization names, and a second one was for general unknown word detection.

For person name and organization name extraction, all person names and organization names were deleted from the dictionary. There were 4,690 person names and 2,871 organization names in the corpus. These names were tagged with the chunk identification "BIO" as described above.

For general unknown word detection, all words that occurred only once in the corpus were deleted from the dictionary, and were thus treated as unknown words. 12,730 unknown words were created under this condition. As in the first experiment, these unknown words were marked with "BIO" tags.

3.2 Data Preprocessing

In this corpus, Chinese person names are segmented into two parts: family name and first name. For example, "邓/nr 小 平/nr" (Deng Xiaoping), "陈/nr 方/nr 安 生/nr" (Chen-Fang Ansheng)[1], etc. We have combined them together as one entity for chunking.

For organization names, the components are tagged separately but with brackets showing that they are organization names. For examples, "[中 国/ns 国 际/n 广 播/vn 电 台/n]nt" (China Radio International), "[全 国/n 人 大/j 常 委 会/j]nt" (National People's Congress), etc. We treat the whole phrase as a chunk for this task.

4 Results

We now present the results of our experiments. Recall, precision and F-measure are defined with the equations below, as is usual in such experiments.

$$recall = \frac{\#\ of\ correctly\ extracted\ words}{\#\ of\ total\ unknown\ words}$$
$$precision = \frac{\#\ of\ correctly\ extracted\ words}{\#\ of\ total\ regconized\ words}$$
$$F - measure = \frac{2 \times recall \times precision}{recall + precision}$$

4.1 Person Name Extraction

Table 2 shows the results of person name extraction. The accuracy for retrieving person names

[1]This is a name with two family names. It happened with some women after married by adding the family name of the husbands.

Position	Char.	POS(best)	POS(2nd)	POS(3rd)	Family Name	Chunk
i - 2	江	n-S	Ng-S	*	Y	B
i - 1	泽	Ag-S	Ng-S	*	N	I
i	民	Ng-S	a-B	j-S	N	I
i + 1	主	n-B	a-E	*	N	O
i + 2	席	n-E	q-S	Ng-S	Y	O

Figure 1: An illustration of chunking process 'President Jiang Zemin'

Table 2: Results for person name extraction

	Recall	Precision	F-measure
Best/F	83.37	86.06	***84.69***
Best/B	79.45	86.84	82.98
2nd/F	83.76	84.67	84.21
2nd/B	79.16	***87.37***	83.06
3rd/F	***84.05***	85.13	84.59
3rd/B	79.35	86.37	82.71
Best+FamN/F	**85.81**	87.52	86.66
Best+FamN/B	84.44	**89.25**	**86.78**

Best - use ony best POS tag, 2nd - use first 2 best POS tags, 3rd - use first 3 best POS tags, F - forward parsing, B - backward parsing, FamN - add family name as feature

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 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)[2], "弗 兰 克·卡 恩" (Frank Kahn)[3], "伯 瑞 恩" (Boraine)[4], etc.

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

Some examples of the extracted person names are shown below. The left side shows the best segmentation and POS tag from the n-best analysis.

李/Ng 鹏/Ng → 李 鹏 (Li Peng)
刘/nr 我/r 成/v → 刘 我 成 (Liu Wocheng)

[2]the former Prime Minister of Morocco
[3]Western Cape Attorney General of South Africa in 1998
[4]Truth Commission Deputy Chairman in 1998

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 a quite high 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").

4.2 Organization Name Extraction

Table 3: Results for organization name extraction

	Recall	Precision	F-measure
Best/F	54.66	70.85	61.71
Best/B	63.25	***79.36***	70.40
2nd/F	55.76	73.32	63.34
2nd/B	***64.17***	78.00	***70.41***
3rd/F	55.58	72.04	62.75
3rd/B	63.07	77.53	69.56

Table 3 shows the result for organization name extraction. Organization names are best extracted by using only 2-best answers with 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 high possibly to be an organization name. Some examples of the extracted organization names are shown below.

国际/n 原子能/n 机构/n → 国际原子能机构 (International Atomic Energy Agency)

香港/ns 中华/nz 电力/n 公司/n → 香港中华电力公司 (Hong Kong CLP Power)

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 judgment.

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.

Table 4: Results for unknown word extraction in general

	Recall	Precision	F-measure
Best/F	56.77	65.28	60.70
Best/B	***58.43***	63.82	***61.00***
2nd/F	56.27	***65.43***	60.46
2nd/B	57.93	64.05	60.83
3rd/F	55.88	64.66	59.95
3rd/B	57.23	62.70	59.84

In general, around 60% accuracy (F-measure) was achieved for unknown word detection. Adding new features does not yield better results as those shown in Table 4. This may be explained by the fact that more features may confuse the system in the right identification of the location of unknown words. Below are examples of extracted unknown words.

处/n 变/v 不/d 惊/unk → 处变不惊 (face a problem without panic)

绿/a 树/n 成/v 荫/unk → 绿树成荫 (green trees make shade)

通知/n 书/n → 通知书 (notice letter)

Surprisingly, common Chinese phrases which were not annotated as unknown words, were also extracted. It happens that they can be grouped together like idioms. For example, "天蓝海碧" (blue sky green sea), "星移物换" (stars move things change), etc. This shows that this approach may work as well for extracting idiomatic phrases in Chinese. Normally, if a word is unknown, then all characters in the word are segmented separately as monosyllabic characters. This is an important clue to detect the location of unknown words. More investigation in this direction is however required.

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. Table 5 shows the results and a comparison with the 1-best POS tag method.

Table 5: Comparison between word and character based chunking

	Recall	Precision	F-measure
PersN/Char/F	***83.37***	86.06	***84.69***
PersN/Char/B	79.45	***86.84***	82.98
PersN/Word/F	78.73	84.00	81.28
PersN/Word/B	73.58	82.78	77.88
OrgN/Char/F	54.66	70.85	61.71
OrgN/Char/B	***63.25***	79.36	***70.40***
OrgN/Word/F	55.21	***82.51***	66.16
OrgN/Word/B	57.77	82.29	67.88
UnkW/Char/F	56.77	***65.28***	60.70
UnkW/Char/B	***58.43***	63.82	***61.00***
UnkW/Word/F	51.72	63.38	56.96
UnkW/Word/B	51.64	63.44	56.93

It is proved that character based chunking yields better results than word based chunking in most of the cases. Except for organization name extraction, word based chunking gives better precision, but yet the recall is low compared with the other.

6 Comparison with Other Works

There are basically two methods to extract unknown words, statistical and rule based ap-

proaches. 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 unknown words as well. It is a better solution for rule based models as the maintenance of the rules is eased. 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 occurrences 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 its is 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 do n-best analysis and use chunking based on SVM. 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 Future Works

As shown in the results, n-best answers did not yield significantly better results. Therefore, finding alternative solutions for improvement is required.

Some of the suggestions are such as adding more character types like transliterated characters, second and third characters used for Chinese person name, different classification like "one-vs-rest method", different context window sizes, different chunking representation etc. We can also add in rule-based model to increase the precision.

8 Conclusion

We proposed an "all-purpose" method for Chinese unknown word detection. Our method is based on an n-best morphological analysis that generates n-best POS tags using Markov Models, followed by a chunking based on character features done using Support Vector Machines. Unfortunately, n-best morphological analysis did not give significantly better results than those obtained with the first best answer. We have 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

[Asahara and Matsumoto, 2003] Masayuki Asahara and Yuji Matsumoto. 2003. Unknown Word Indentification in Japanese Text Based on Morphological Analysis and Chunking. In *IPSJ SIG Notes Natural Language*, No. 2003-NL-154-8, pages 47–54.

[Chen and Ma, 2002] 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.

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

[Matsumoto et al., 2002] 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.

[Shen et al., 1997] 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).

[Zhang et al., 2002] 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.*