Unknown Word Extraction for Chinese Documents

Keh-Jiann Chen
Institute of Information science,
Academia Sinica
kchen@iis.sinica.edu.tw

Wei-Yun Ma
Institute of Information science,
Academia Sinica
ma@iis.sinica.edu.tw

Abstract

There is no blank to mark word boundaries in Chinese text. As a result, identifying words is difficult, because of segmentation ambiguities and occurrences of unknown words. Most previous works focus their attention only on the resolution of ambiguous segmentation. The problem of unknown word identification is considered more difficult and needs further investigation. Conventionally unknown words were extracted by statistical methods for statistical methods are simple and efficient. However the statistical methods without using linguistic knowledge suffer the drawbacks of low precision and low recall. Because character strings with statistical significance might be phrases or partial phrases instead of words and low frequency new words are hardly identifiable by statistic methods. In addition to statistical information, we try to use as much information as possible, such as morphology, syntax, semantics, and world knowledge. The identification system fully utilizes the context and content information of unknown words in the steps of detection process, extraction process, and verification process. A practical unknown word extraction system was implemented which online identifies new words, including low frequency new words, with high precision and high recall rates.

1 Introduction

One of the most prominent problems in computer processing of Chinese language is identification of the word sequences of input sentences. There is no blank to mark word boundaries in Chinese text. As a result, identifying words is difficult, because of segmentation ambiguities and occurrences of unknown words (i.e. out-of-vocabulary words).

Most papers dealing with the problem of word segmentation focus their attention only on the resolution of ambiguous segmentation. The problem of unknown word identification is considered more difficult and needs to be further investigated. According to an inspection on the Sinica corpus (Chen etc., 1996), a 5 million word Chinese corpus with word segmented, it shows that 3.51% of words are not listed in the CKIP lexicon, a Chinese lexicon with more than 80,000 entries. The result is compatible with Sampson's finding of the dictionary coverage on the LOB corpus (Sampson, 1989).

Identifying Chinese unknown words from a document is difficult; since

1. There is no blank to mark word boundaries;
2. Almost all Chinese characters and words are morphemes;
3. Morphemes are syntactic ambiguous and semantic ambiguous;
4. Words with same morpho-syntactic structure might have different syntactic categories;
5. No simple rules can enumerate all types of unknown words;
6. Online identification from a short text is even harder, since low frequency unknown words are not identifiable by naive statistical methods.

It is difficult to know where unknown words occur in a text since all Chinese characters can either be a morpheme or a word and there are no blank to mark word boundaries. Therefore without (or even with) syntactic or semantic checking, it is difficult to tell whether a character in a particular context is a part of an unknown word or whether it stands alone as a word. Compound words and proper names are two major types of unknown words. It is not possible

to list all of the proper names and compounds neither in a lexicon nor enumeration by morphological rules. Conventionally unknown words were extracted by statistical methods for statistical methods are simple and efficient. However the statistical methods without using linguistic knowledge suffer the drawbacks of low precision and low recall. Because character strings with statistical significance might be phrases or partial phrases instead of words and low frequency new words are hardly identifiable by statistic methods.

Conventional common statistical features in problem of unknown word extraction are mutual information (Church 90), entropy (Tung 94), association strength (Smadja 93, Wang 95) and dice coefficients (Salton 83, Smadja 96) etc. Chang etc. (Chang etc. 97) iteratively apply the joint character association metric which is derived by integrating above statistical features. Their performance is recall rate:81%, precision rate: 72% in disyllabic unknown word, recall rate:88%, precision rate: 39% in trisyllabic unknown word, and recall rate:94%, precision rate: 56% in four-syllabic unknown word.

Chang etc. (1994) used statistical methods to identify personal names in Chinese text which achieved a recall rate of 80% and a precision rate of 90%. Chen & Lee (1994) used morphological rules and contextual information to identify the names of organizations. Since organizational names are much more irregular than personal names in Chinese, they achieved a recall rate of 54.50% and a precision rate of 61.79%. Lin etc. (1993) made a preliminary study of the problem of unknown word identification. They used 17 morphological rules to recognize regular compounds and a statistical model to deal with irregular unknown words, such as proper names etc.. With this unknown word resolution procedure, an error reduction rate of 78.34% was obtained for the word segmentation process. Since there is no standard reference data, the claimed accuracy rates of different papers vary due to different segmentation standards. In this paper we use the Sinica corpus as a standard reference data. As mentioned before, the Sinica corpus is a word-segmented corpus based on the Chinese word segmentation standard for information processing proposed by ROCLING (Huang et al, 1997). Therefore it contains both known words and unknown words, which are properly segmented. The corpus was utilized for the purposes of training and testing.

From the above discussion, it is known that identification of unknown words is difficult and need to adopt different methods in identifying different types of unknown words. The objective of this research is to find methods to extract unknown words from a document and identify their syntactic and semantic categories. Although both processing are interrelated, for limiting scope of this paper, we will focus our discussion on the extraction process only and leave the topics of syntactic and semantic category predictions to other papers.

2 Steps to Identify Unknown Words

In addition to statistical information, we try to use as much information as possible, such as morphology, syntax, semantics, and world knowledge, to identify unknown words. The identification system fully utilizes the context and content information of unknown words in each three steps of processes, i.e. detection process, extraction process, and verification process. The detection process detects the occurrences of unknown words for better focusing, so that on the next step extraction process, it needs only focus on the places where unknown were detected. In addition, it also helps in identifying low frequency unknown words, which hardly can be identified by conventional statistical extraction methods. The extraction process extracts unknown words by applying morphological rules and statistical rules to match for different types of unknown words. As usual, tradeoff would occur between recall and precision. Enriching the extraction rules might increase recall rates, but it also increases the ambiguous and false extractions and thus lowers the precision. The final verification process comes to rescue. It resolves ambiguous and false extractions based on the morphological validity, syntactic validity, and statistical validity.

3 Unknown Word Detection

Conventionally a word segmentation process

identifies the words in input text by matching lexical entries and resolving the ambiguous matching (Chen & Liu, 1992, Sproat et al, 1996). Hence after segmentation process the unknown words in the text would be incorrectly segmented into pieces of single character word or shorter words. If all occurrences of monosyllabic words are considered as morphemes of unknown words, the recall rate of the detection will be about 99%, but the precision is as low as 13.4% (Chen & Bai, 1998). Hence the complementary problem of unknown word detection is the problem of monosyllabic known-word detection, i.e. to remove the monosyllabic known-words as the candidates of unknown morphemes. A corpus-based learning method is proposed to derive a set of syntactic discriminators for monosyllabic words and monosyllabic morphemes (Chen & Bai, 1998).

The following types of rule patterns were generated from the training corpus. Each rule contains a key token within curly brackets and its contextual tokens without brackets. For some rules there may be no contextual dependencies. The function of each rule means that in a sentence, if a character and its context match the key token and the contextual tokens of the rule respectively, this character is a proper word (i.e. not a morpheme of an unknown word). For instance, the rule "{Dfa} Vh" says that a character with syntactic category Dfa is a proper word, if it follows a word of syntactic category Vh.

Rule type	Example
char	{的}
word char	不 {願}
char word	{全} 世界
category	{T}
{category} category	{Dfa} Vh
category {category}	Na {Vcl}
char category	{就} VH
category char	Na {上}
category category char	Na Dfa {高}
char category category	{極} Vh T

Table1. Rule types and Examples

Rules of the 10 different types of patterns above were generated automatically by extracting each instance of monosyllabic words in the training corpus. Every generated rule pattern was checked for applicability and accuracy. At the initial stage, 1455633 rules were found. After eliminating the low applicability rules, i.e. frequency less than 3, there are 215817 rules remained. At next stage, the rules with accuracy greater than 98% are selected for better recall rate. However the selected rules may subsume each other. Shorter rule patterns are usually more general than the longer rules. A further screening process is applied to remove the redundant rules. The final rule sets contain 45839 rules and were used to detect unknown words in the experiment. It achieves the detection rate of 96% and the precision rates of 60%. Where detection rate 96% means that for 96% of unknown words in the testing data, at least one of its morpheme was detected as part of unknown word. However the boundaries of unknown words are still not known. For more detail discussion, see (Chen & Bai 1998). For convenience, hereafter we use (?) to mark detected morphemes of unknown words and () to mark the words which are not detected as morphemes of unknown words.

4 Unknown Word Extraction

At detection stages, the contextual rules were applied to detect fragments of unknown words, i.e. monosyllabic morphemes. The extraction rules will be triggered by the detected morphemes only. The extraction rules are context, content, and statistically constrained. Rule-design targets for high recall rate and try to maintain high precision at the mean time. Since it is hard to derive a set of morphological rules, which exactly cover all types of unknown words. Our approach is that if morphological structures of certain types of unknown words are well established, their fine-grain morphological rules will be designed. Otherwise statistical rules are designed without differentiate their extracted word types. Redundancy is allowed to achieve better coverage. Both morphological rules and statistical rules use context, content and statistical information in their extraction. The only difference is that the design of morphological rules is based on the morphological structures, i.e. content information. The context and statistical information is for

verification. On the other hand the design of statistical rules is based on the statistic information, and the context and content information is for verification.

4.1 Morphological rules

Since there are too many different types of unknown words, we cannot go through the detail extraction processes for each different type. It will be exemplified by the personal name extraction to illustrate the idea of using different clues in the extraction process. First of all the content information is used, each different type of unknown words has its own morphological structure. For instance, a typical Chinese personal name starts with a last name and followed by a given name. The set of last names is about one hundred. Most of them are common characters. Given names are usually one or two characters and seldom with bad meaning. Based on the above structure information of Chinese personal names, the name extraction rules are designed as shown in Table 2. Context information is used for verification and determining the boundary of the extracted word. For instance, in the last rule of Table 2, it uses context information and statistical information to resolve ambiguity of the word boundary. It is illustrated by the following examples.

1) after detection : 張(?) 明(?) 正() 要() 殺() 人()。
extractnion : 張明正 要 殺 人。
Ming-Zheng Zhang want kill somebody.
or 張明 正 要 殺 人。
Ming Zhang just want kill somebody.

In the examples 1), there are two possible candidates of personal names, 張明 and 張明正. By context information, the bi-gram (NAME, 正) is less freguent than (NAME, 要) in the corpus, so without considering statistical constraints, it would suggest that 張明正 is a correct extraction instead of 張明. However, the locality of the keywords is very important clue for identification, since the keywords of a text are usually unknown words and they are very frequently reoccurred in the text. Conventional new word extraction techniques are very much relied upon the statistical information. We will discuss this topic in more details in the next section. The statistical information is used here for verification. For instance, if an another sentence which is like 張(?) 明(?) 來() 了() occurs in a same document, it suggests 張明 is the correct extraction, since the statistical constraint $prob_{document}(正 \mid 張明) < 1$ rejects 張明正.

Rule type	Constraints & Procedure
$ms_i(?)\ ms_{i+1}(?)\ ms_{i+2}(?)$	$combine(i, i+1, i+2)$
$ms_i()\ \ ms_{i+1}(?)\ ms_{i+2}(?)$	$combine(i, i+1, i+2)$
$ms_i(?)\ ms_{i+1}()\ \ ms_{i+2}(?)$	$combine(i, i+1, i+2)$
$ds_i()\ \ ms_{i+1}(?)$	$combine(i, i+1)$
$ms_i(?)\ \ ds_{i+1}()$	$combine(i, i+1)$
$ms_i(?)\ ms_{i+1}(?)\ ps_{i+2}()$	$combine(i, i+1)$
$ms_i(?)\ ms_{i+1}(?)\ ms_{i+2}()$	as follows:

$$
\begin{aligned}
&if\ \ prob_{document}(ms_{i+2} \mid ms_i ms_{i+1}) < 1 \\
&\quad combine(i, i+1)\ \ as\ \ a\ \ disyllabic\ \ name \\
&elsif\ \ freq_{coupus}(NAME, ms_{i+2}, word_{i+3}) \geq 1 \\
&\quad combine(i, i+1)\ \ as\ \ a\ \ disyllabic\ \ name \\
&elsif\ \ freq_{coupus}(NAME, word_{i+3}) \geq freq_{coupus}(NAME, ms_{i+2}) \\
&\quad combine(i, i+1, i+3)\ \ as\ \ a\ \ trisyllabic\ \ name \\
&\quad else\ \ combine(i, i+1)\ \ as\ \ a\ \ disyllabic\ \ name
\end{aligned}
$$

Notes: ms denotes monosyllable. ds denotes disyllable. ps denotes polysyllable which consists of more than one syllable. word denotes a word which could consist of any number of syllable. ms_i must belong to Common Chinese Last Name Set, such as 陳, 王…etc.

Table 2. Rule types of Chinese personal name

4.2 Statistical Rules

It is well known that keywords often reoccur in a document (Church, 2000) and very possible the keywords are also unknown words. Therefore statistical extraction methods utilize the locality of unknown words. The idea is that if two consecutive morphemes are highly associated then combine them to form a new word. Mutual information-like statistics are very often adopted in measuring association strength between two morphemes (Church & Merser, 1993, Sproat et al, 1996). However such kind of statistic does not work well when the sample size is very limited. Therefore we propose to use

reoccurrence frequency and fan-out numbers to characterize words and their boundaries (Chien, 1999). 12 statistical rules are derived to extract unknown words. Each rule is triggered by detected morphemes and executed in iteration. The boundaries of unknown words might extend during each iteration until no rule could be applied. Following are two typical examples of statistical rules.

Rule id	Pattern	Statistical constraint
R1	Lm(?) Rm()	S1
R2	Lm(?) Rm(?)	S2

S1: $P(Lm \mid Rm) \geq 0.8$ and $P(Rm \mid Lm) \geq 0.8$ and $Freq(LmRm) \geq 2$

S2: $((P(Lm \mid Rm) \geq 0.8$ or $P(Rm \mid Lm) \geq 0.8)$ *and* $Freq(LmRm) \geq 2)$ or $(P(Lm \mid Rm) \geq 0.8$ and $P(Rm \mid Lm) \geq 0.8)$

Table 3. Two examples of statistical rules

The rule R1 says that Lm and Rm will be combined, if both conditional probability P(Lm|Rm)>=0.8 and P(Rm|Lm)>=0.8 hold and the string LmRm occurred more than once in the processed document. Conditional probabilities constrain the fan-out number on each side of morpheme, i.e. the preceding morpheme of Rm should almost be limited to Lm only and vice versa. The threshold value 0.8 is adjusted according to the experimental results, which means at least four out of five times the preceding morpheme of Rm is Lm and vice versa. However the statistical constraints are much loose when the right morpheme Rm is also a detected morpheme, as exemplified in R2. You may notice that it also accepts the unknown words occurred only once in the document.

Conventional statistical extraction methods are simple and efficient. However if without supporting linguistic evidences the precision of extraction is still not satisfactory, since a high frequency character string might be a phrase or a partial phrase instead of a word. In addition to statistical constraint, our proposed statistical method requires that a candidate string must contain detected morphemes. In other words, the statistical rules are triggered by detected morphemes only. Furthermore the morphological structure of extracted unknown word must be valid. A validation process will be carried out at the different stages for all extracted unknown words.

5 Verification

To verify a correct extraction depends on the following information.

1. Structure validity: the morphological structure of a word should be valid.
2. Syntactic validity: the syntactic context of an identified new word should be valid.
3. Local consistency: the identified unknown words should satisfy the local statistical constraints, i.e. no inconsistent extension on the morphological structures. For instance, a new word was identified by the pattern rules, but if it violates the statistical constraints, as exemplified in 1), will be rejected.

Each extracted candidate will be evaluated according to the validity of above three criteria. For the candidates extracted by the statistical rules, their structure validity and syntactic validity are checked after extraction. On the other hand, for the unknown words extracted according to the morphological rules, their structure validity and syntactic validity are checked at extraction stage and their local statistical consistency is checked after extraction. To verify the structure validity and syntactic validity of the unknown words extracted by statistical methods, their syntactic categories are predicted first, since statistical rules do not classify unknown word types. The prediction method is adopted from (Chen, Bai & Chen, 1997). They use the association strength between morpheme and syntactic category to predict the category of a word. The accuracy rate is about 80%. Once the syntactic category of an unknown word is known its contextual bi-gram will be checked. If the bi-grams of (*preceding word/category, unknown word category*) and (*unknown word category, following word/category*) are syntactically valid, i.e. the bi-gram patterns are commonly occurred in the corpus, the extracted word is considered to be a valid word. Otherwise this candidate will be rejected.

5.1 Final Selection

It is possible that the extracted candidates

conflict each other. For instance, in the following example, both candidates are valid. 班乃特, Bennet" is extracted by name rules and "律師班, lawyer-class" is extracted by suffix rules.

name ==> 安然 公司 律師 班乃特 說 ,
An-jan company lawyer Bennett said,
suffix ==> 安然 公司 律師班 乃 特 說 ,
An-jan company lawyer-class is special said,

The extracted new words will form a word lattice. The selection process finds the most probable word sequence among word lattice as the final result. In the current implementation, we used a very simple heuristics of maximizing the total weights of words to pick the most probable word sequence. The weight of a word w is defined to be freq(w)*length(w), where freq(w) is the occurrence frequency of w in the document and length is the number of characters in w. For the above example, "班乃特, Bennett" occurred 5 times and "律師班, lawyer-class" occurred twice only in the document. Therefore the final result is

安然 公司 律師 班乃特 說 ,
An-jan company lawyer Bennett said ,
"Bennett, the lawyer of An-jan company, said…"

6 Experimental Results

In the current implementation, the morphological rules include the rules for Chinese personal names, foreign transliteration names, and compound nouns. In addition to the morphological rules, twelve constrained statistical rules were implemented to patch the under coverage of the morphological rules. Although the current implementation is not complete, morphological rules of many other types of unknown words were not included, such as rules for compound verbs. The experiment results still show that the proposed methods work well and the morphological rules and the statistical rules complement each other in the extraction and verification.

The Sinica balanced corpus version 3.0 contains 5 million segmented words tagged with pos, which provides the major training and testing data. The training data contains 8268 documents with 4.6 million words. We use it to train the detection rules and morphological rules. We randomly pick 100 documents from rest of the corpus, which contains 17585 words and average 11.6 unknown word types per document as the testing data. A word is considered as an unknown word, if either it is not in the CKIP lexicon or it is not identified by the word segmentation program as foreign word (for instance English) or a number. The CKIP lexicon contains about 80000 entries.

The precision and recall rates are provided. The target of our approach is to extract unknown words from a document, so we define "correct extractions" as unknown word types correctly identified in the document. The precision and recall rate formulas are as follows:

NC_i = number of correct extractions in document i
NE_i = number of extracted unknown words in document i
NT_i = number of total unknown words in document i

$$\text{Precision rate} = \frac{\sum_{i=1}^{i=100} NC_i}{\sum_{i=1}^{i=100} NE_i} \qquad \text{Recall rate} = \frac{\sum_{i=1}^{i=100} NC_i}{\sum_{i=1}^{i=100} NT_i}$$

	Match#	Extract#	Precision	Recall
Morphological rules	541	590	92%	47%
Statistical rules	455	583	78%	39%
Total system	791	890	89%	68%

Table 4. Experimental results

The recall rate of the current system is not very high, because not all of the morphological rules were implemented and some of the word tokens in the testing data are arguable.

7 Conclusions and Future Works

Unknown word extraction is a very hard task. In addition to statistical information, it requires supporting knowledge of morphological, syntactic, semantic, word type specific and common sense. One important trend is to look harder for sources of knowledge and managing knowledge that can support unknown word identification. A word segmented and tagged corpus is essential for the success of the whole research. The corpus

provides the major training and testing data. It also supports plenty of unknown words and their contextual data to derive extraction rules. In this work we are managing to use the structure information, the context environment, and statistical consistency of the unknown words and to increase the recall and precision of the extraction process. The syntactic and semantic classifications for unknown words are executed in parallel with the extraction process. Both classification processes are very hard and need further researches.

8 References

Chang J. S.,S.D. Chen, S. J. Ker, Y. Chen, & J. Liu,1994 "A Multiple-Corpus Approach to Recognition of Proper Names in Chinese Texts", *Computer Processing of Chinese and Oriental Languages*, Vol. 8, No. 1, 75-85.

Chen, H.H., & J.C. Lee, 1994,"The Identification of Organization Names in Chinese Texts", *Communication of COLIPS*, Vol.4 No. 2, 131-142.

Chen, K.J. & S.H. Liu, 1992,"Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th Coling*, pp. 101-107.

Chen, K.J., C.R. Huang, L. P. Chang & H.L. Hsu, 1996,"SINICA CORPUS: Design Methodology for Balanced Corpora," *Proceedings of PACLIC 11th Conference*, pp.167-176.

Chen, K. J., M. H. Bai, K. J. Chen, 1997, "Category Guessing for Chinese Unknown Words." *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, pp. 35-40. NLPRS '97 Thailand.

Chen, K.J. & Ming-Hong Bai, 1998, "Unknown Word Detection for Chinese by a Corpus-based Learning Method," international Journal of Computational linguistics and Chinese Language Processing, Vol.3, #1, pp.27-44.

Chen, K.J., Chao-Jan Chen. 1998. "A Corpus Based Study on Computational Morphology for Mandarin Chinese（語料庫為本的中文複合詞構詞律模型研究）." *Quantitative and Computational Studies on the Chinese Language*. Benjamin K. T'sou, Tom B.Y. Lai, Samuel W. K. Chan, William S-Y. Wang, ed. HK: City Univ. of Hong Kong. pp.283-306.

Chiang, T. H., M. Y. Lin, & K. Y. Su, 1992," Statistical Models for Word Segmentation and Unknown Word Resolution," Proceedings of ROCLING V, pp. 121-146.

Chien, Lee-feng, 1999," PAT-tree-based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval," Information Processing and Management, Vol. 35, pp. 501-521.

Church, K. W., & R. L. Mercer, 1993, "Introduction to the Special Issue on Computational Linguistics Using Large Corpora." *Computational Linguistics, Vol. 19, #1*, pp. 1-24

Church, Kenneth W., 2000," Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p*p", Proceedings of Coling 2000, pp.180-186.

Huang, C. R. Et al.,1995,"The Introduction of Sinica Corpus," *Proceedings of ROCLING VIII*, pp. 81-89.

Huang, C.R., K.J. Chen, & Li-Li Chang, 1997, "Segmentation Standard for Chinese Natural Language Processing," International Journal of Computational Linguistics and Chinese Language Processing, Accepted.

Lin, M. Y., T. H. Chiang, & K. Y. Su, 1993," A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation," Proceedings of ROCLING VI, pp. 119-137.

Mo, R.P., Y.J. Yang, K.J. Chen, and C.R. Huang, 1993 "Determinative-Measure Compounds in Mandarin Chinese: Formation Rules and Parser Implementation", *Readings in Chinese Natural Language Processing, Journal of Chinese Linguistics Monograph Series Number 9*.

Sampson, Geoffrey, 1989,"How Fully Does a Machine-usable Dictionary Cover English Text?", *Literary and Linguistic Computing* Vol. 4, pp.29-35.

Sproat, R., C. Shih, W. Gale, & N. Chang,1996, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22(3),377-404.

Sun, M. S., C.N. Huang, H.Y. Gao, & Jie Fang, 1994, "Identifying Chinese Names in Unrestricted Texts", *Communication of COLIPS*, Vol.4 No. 2, 113-122.

Chang, Jing-Shin and Keh-Yih Su, 1997a. "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", to appear in International Journal of Computational Linguistics & Chinese Language Processing, 1997.