# Extract Chinese Unknown Words from a Large-scale Corpus Using Morphological and Distributional Evidences

**Kaixu Zhang**† and **Ruining Wang**† and **Ping Xue**‡ and **Maosong Sun**†
†State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China
{kareyzhang,wangruining.student,sunmaosong}@gmail.com
‡The Boeing Company
ping.xue@boeing.com

## Abstract

The representative method of using morphological evidence for Chinese unknown word (UW) extraction is Chinese word segmentation (CWS) model, and the method of using distributional evidence for UW extraction is accessor variety (AV) criterion. However, neither of these methods has been verified on large-scale corpus. In this paper, we propose extensions to remedy the drawbacks of these two methods to handle large-scale corpus: (1) for CWS, we propose a generalized definition of word to improve the recall; and (2) for AV, we propose a restricted version to decrease noise. We carry out experiments on a Chinese Web corpus with approximate 200 billion Chinese characters. Experimental results show that our methods outperform the baselines, and the combination of the two evidences can further improve the performance. Moreover, our methods can also efficiently segment the corpus on the fly, which is especially valuable for processing large-scale corpus.

## 1 Introduction

A Chinese word is constructed with one or more Chinese characters. Chinese characters can ambiguously combine to form Chinese words, and there are no explicit delimiters in the text to indicate word boundaries. It is thus crucial for most Chinese natural language processing tasks to maintain a large word list. Given that Chinese language has several productive word creation mechanisms, identification and extraction of UW is an important task for or Chinese NLP tasks.

Chinese unknown word (UW) extraction aims to extract UWs from a given corpus and enrich the word list. Two types of information can be used to determine whether a string of characters in question is a Chinese UW or not, namely the characters that construct the string in question, and the neighbors that this string of characters appears with. The first type of information can be regarded as the morphological evidence, while the second can be viewed as the distributional evidence.

The representative method of using morphological evidence is Chinese word segmentation (CWS) model. CWS is to identify every word token in a given sentence. Using the CWS model, we can define the word-string ratio (WSR) to extract UWs. The representative method of using distributional evidence is the accessor variety (AV) criterion (Feng et al., 2004a).

WSR is directly derived from the CWS method based on character tagging (Xue, 2003). This CWS method is based on the morphological information of the strings in question and their context. Strings with high WSR are considered as words, for high WSR indicates that the corresponding string is segmented as a word by this CWS method with high probability. Though the performance of the CWS method is relatively high, it leaves a number of UWs unrecognized or incorrectly recognized due to erroneous segmentation.

The AV criterion (Feng et al., 2004a) is based on the distributional information. Strings that have various contexts can be considered as words. It is shown that this method works well even for UWs with frequency of about 10. But both words and non-words with high frequency tend to have high AV. This brings noise for UW extraction in a large-scale corpus.

However, neither of these methods has been verified on large-scale Web corpus. In fact, as we observe, they both show certain deficiencies when dealing with large-scale corpus.



*Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 837–845,
Chiang Mai, Thailand, November 8 – 13, 2011. 

The emergence of online documents that contain various UWs poses challenges to UW extraction and other Chinese NLP tasks, but also provides a rich resource to make the UW extraction more meaningful.

Taking the two methods above as baselines, we propose two new methods, namely, generalized word string ratio (GWSR), and restricted accessor variety (RAV), by extending the current methods respectively, in order to overcome the relevant problems for large-scale corpus.

For GWSR, we propose a sophisticated way to generalize the definition of word in the CWS model. This method can extract more UWs that cannot be correctly segmented as words. For RAV, as opposed to AV, we restricted the accessors to be a small set of word pairs $(\mathbf{w}_l, \mathbf{w}_r)$ such that $\mathbf{w}_l$ appears right before the string in question and, at the same time, $\mathbf{w}_r$ appears right after the string in question. RAV is especially suitable for a large-scale corpus in which UWs occur with relatively high frequency.

We carry out experiments on a Chinese Web corpus with approximate 200 billion Chinese characters. Experiment results show that our methods outperform the corresponding baselines, and the combination of our methods can further improve the performance. Some examples are shown in the experiments section.

We further investigate the effect of corpus size to UW extraction. The numbers of Chinese characters in corpora range from 20 million to 200 billion. Moreover, our methods can also efficiently segment the corpus on the fly, which is practical for processing large-scale corpus.

The contribution of this paper is twofold. First, we proposed two UW extraction methods which outperform the baselines based on morphological and distributional evidence. Second, our experiments were conducted on corpora with up to 200 billion Chinese characters and provided insights about the effect of corpus size on UW extraction.

## 2 Background

### 2.1 CWS as Character Tagging

CWS aims to segment Chinese sentences into words. A practical CWS model needs to handle UWs, which are also named as out-of-vocabulary (OOV) words. If a corpus is perfectly segmented, the UW extraction task is also accomplished.

Xue (2003) proposed a character sequence tagging framework for CWS. Comparing to other methods, it has better performance on dealing with the UWs (Ling et al., 2003; Peng et al., 2004). The sequence tagging framework is also used for named entity identification in English (McCallum and Li, 2003), which is related to Chinese UW extraction.

In this framework, the input is a raw Chinese sentence $\mathbf{s}$, denoted as a sequence of characters $c_i$:

$$\mathbf{s} = \overline{c_1 \dots c_n} \tag{1}$$

The output of the character sequence tagging is a sequence $\mathbf{t}$ of tags $t_i$ corresponding to the input characters:

$$\mathbf{t} = \overline{t_1 \dots t_n} \tag{2}$$

where $t_i \in \{\mathrm{B}, \mathrm{M}, \mathrm{E}, \mathrm{S}\}$. The tags B / M / E indicate that the corresponding character is at the beginning / middle / end position of a multi-character word. The tag S indicates that the corresponding character is a single character word. The segmentation result of this sentence can thus be determined by the tag sequence.

Given an input sentence $\mathbf{s}$, the output sequence of tags $\mathbf{t}$ is calculated as

$$\mathbf{t} = \arg\max_{\mathbf{t}'} W^T \Phi(\mathbf{s}, \mathbf{t}') \tag{3}$$

where $\Phi$ returns a feature vector of the pair $(\mathbf{s}, \mathbf{t}')$, and W is a vector of feature weights. The decoding is to find a $\mathbf{t}$ that maximizes the objective function.

Machine learning methods such as maximum entropy (Ng and Low, 2004), conditional random field model (Peng et al., 2004) and perceptron (Jiang et al., 2009) have been used for this framework.

The features in this framework are mainly composed by character unigrams, character bigrams and tag bigrams. In Chinese, a character is usually a morpheme. Therefore the CWS model based on the character tagging framework can be regarded as a UW extraction method using morphological information.

However, in contrast to CWS, UW extraction focuses on identifying substrings in a corpus that are potential words independent of the environments where they may occur. Though the performance of the CWS method is relatively high, the poor recall of UWs 'is still the Achilles heel of segmentation systems' (Emerson, 2005). The

CWS methods also fails to capture distributional information of the strings in question.

### 2.2 UW Extraction and the Accessor Variety Criterion

There are methods proposed for UW extraction based on morphological evidence, distributional evidence, or both (Chen and Ma, 2002; Ma and Chen, 2003; Feng et al., 2004a; Hong et al., 2009).

Some methods can be used for both UW extraction and CWS (Sun et al., 1998; Feng et al., 2004b; Jin and Tanaka-Ishii, 2006; Zhao and Kit, 2008). But for CWS, these methods are not comparable with the character tagging based CWS methods (Zhao and Kit, 2008), because the character tagging based CWS methods can better capture the morphological information.

We focus on a UW extraction method based on the distributional information, namely the accessor variety (AV) criterion (Feng et al., 2004a).

Assuming that a string is likely a meaningful unit if it occurs in different linguistic environments (Feng et al., 2004a), AV is defined as:

$$\mathrm{AV}(\mathbf{v}) = \min\{\mathrm{L_{av}}(\mathbf{v}), \mathrm{R_{av}}(\mathbf{v})\} \tag{4}$$

The $\mathrm{L_{av}}(\mathbf{v})$ is defined as the number of distinct Chinese characters that precede $\mathbf{v}$ plus the number of times that $\mathbf{v}$ appears at the beginning of a sentence. The $\mathrm{R_{av}}(\mathbf{v})$ is defined as the number of distinct Chinese characters that succeed $\mathbf{v}$ plus the number of times that $\mathbf{v}$ appears at the end of a sentence. The larger the $\mathrm{AV}(\mathbf{v})$ is, the more likely $\mathbf{v}$ is a word.

In order to fulfill this method, an extra dictionary is needed. And three ad hoc rules are used to discard strings which contain adhesive characters and cannot be words. The details can be found in (Feng et al., 2004a).

This method works well even for strings with low frequency because any distinct character is regarded as an accessor. However, when applying the method to a large-scale corpus in which strings in question are of high frequency, the noise increases considerably due to the lenient definition of the accessor.

## 3 Our Model

In this section, first we will introduce two UW extraction methods based on a character tagging based CWS model. Then we propose a UW extraction method called restricted accessor variety based on the distributional evidence. Finally, we discuss the combination of these methods.

### 3.1 Morphological Evidence

#### 3.1.1 Word-string Ratio

A character tagging based CWS model, which is based on the morphological evidence, can be directly used to extract UWs in a corpus. The Word-string ratio (WSR) provides a straightforward way to determine whether a string in question is a word or not. Strings with high WSR are regarded as UWs.

WSR is defined as the ratio of the frequency of $\mathbf{v}$ that is segmented as a word to the frequency $\mathbf{v}$ that occurs as a string in the corresponding corpus:

$$\mathrm{WSR}(\mathbf{v}) = \frac{\mathrm{WF}(\mathbf{v})}{\mathrm{SF}(\mathbf{v})} \tag{5}$$

where word frequency $\mathrm{WF}(\mathbf{v})$ is the number of times that string $\mathbf{v}$ is segmented by the CWS model as a word in the corpus, and string frequency $\mathrm{SF}(\mathbf{v})$ is the number of times that string $\mathbf{v}$ appears in the corpus.

Now we discuss how to define the words in the CWS model. Since the tag sequence $\mathbf{t}$ in the character tagging framework may contain conflicts (e.g., the tag sequence "B B" means that two immediately connected characters are both at the beginning of multi-character words, which is impossible), we use an alternative way to define the words in the output. This new definition is also a preparation for the definition of the generalized word that we will propose.

Recall the decoding process of the CWS model in the character tagging framework described in Equation 3. Given a sentence $\mathbf{s} = \overline{c_1 \dots c_n}$, we define $\mathrm{Conf}(m)$ as the confidence that there is a word boundary after the m-th character $c_m$ ($t_m \in \{\mathrm{E}, \mathrm{S}\}$):

$$\begin{aligned}\mathrm{Conf}(m) = &\max_{t_m \in \{\mathrm{E,S}\}} W^T \Phi(\mathbf{s}, \mathbf{t}) - \\ &\max_{t_m \in \{\mathrm{B,M}\}} W^T \Phi(\mathbf{s}, \mathbf{t})\end{aligned} \tag{6}$$

Obviously, $\mathrm{Conf}(m) > 0$ indicates that there is more likely a word boundary after the $m$-th character, while $\mathrm{Conf}(m) < 0$ indicates that there is less likely a word boundary after the $m$-th character.

If the string in question $\overline{c_i \cdots c_j}$ in $\mathbf{s}$ is a word, two confidences of the string boundaries $\mathrm{Conf}(i -$

1) and $\text{Conf}(j)$ should be positive and the confidences inside the string $\text{Conf}(k)$ should be negative for $k = i, \cdots, j-1$ . In other words, the string $\overline{c_i \cdots c_j}$ is regarded as a word if and only if:

$$\text{Conf}_{\text{o}}(\overline{c_i \cdots c_j}) > 0 > \text{Conf}_{\text{i}}(\overline{c_i \cdots c_j}) \quad (7)$$

where $\text{Conf}_{\text{o}}(\overline{c_i \cdots c_j})$ and $\text{Conf}_{\text{i}}(\overline{c_i \cdots c_j})$ are defined as:

$$\text{Conf}_{\text{o}}(\overline{c_i \cdots c_j}) = \min\{\text{Conf}(i-1), \text{Conf}(j)\} \quad (8)$$

$$\text{Conf}_{\text{i}}(\overline{c_i \cdots c_j}) = \max_{k=i,\cdots,j-1} \text{Conf}(k) \quad (9)$$

Roughly speaking, words defined according to tag sequence $\overline{t_i \cdots t_j}$ and words defined according to the definition above are identical. In a test set of 107 thousand words, there are only 6 sentences of which the results are not identical.

#### 3.1.2 Generalized Word-String Ratio

Although the CWS model can achieve relatively high performance. It fails to segment many instances of UWs correctly. This makes them hard to be extracted based on WSR.

In order to address this deficiency, we define a notion of generalized word, and we use the Generalized Word-String Ratio (GWSR) to extract UWs as a modified version of WSR. This idea is derived from Liu et al. (2008) and Zhang et al. (2010). For convenience, the term "word" in the rest part of this subsection always refers to a string that is segmented as a word by CWS model.

We define a cost function of string $\mathbf{v} = \overline{c_i \cdots c_j}$ based on the confidence function:

$$\text{Cost}(\mathbf{v}) = \max\{0 - \text{Conf}_{\text{o}}, \text{Conf}_{\text{i}} - 0\} \quad (10)$$

If a string $\mathbf{v}$ is segmented as a single word, the cost function returns a non-positive value; otherwise, it returns a positive value. The larger this value is, the less likely this string can be regarded as a word.

Now we can define GWSR of string $\mathbf{v}$:

$$\text{GWSR}_{\text{th\_LW}}(\mathbf{v}) = \frac{\text{GWF}_{\text{th\_LW}}(\mathbf{v})}{\text{SF}(\mathbf{v})} \quad (11)$$

where the generalized word frequency $\text{GWF}_{\text{th\_LW}}(\mathbf{v})$ is the number of times that string $\mathbf{v}$ appears with $\text{Conf}_{\text{o}}(\mathbf{v}) > \text{Conf}_{\text{i}}(\mathbf{v})$ and $\text{Cost}(\mathbf{v}) \leq$ th_LW in a certain sentence. Here th_LW is a threshold. The inequality $\text{Conf}_{\text{o}}(\mathbf{v}) > \text{Conf}_{\text{i}}(\mathbf{v})$ should be always satisfied. Otherwise it will bring in noise.

Note that $\text{WF}(\mathbf{v}) = \text{GWF}_0(\mathbf{v})$ means that when th_LW $= 0$ only words are regarded as generalized words.

The GWSR provides a way to allow UWs which are incorrectly segmented by the CWS model to be extracted. As a side effect, more noise may be brought in.

### 3.2 Distributional Evidence

The AV criterion is a method to extract UWs based on the distributional evidence. Here we propose a new version of the distribution-based criterion called restricted accessor variety (RAV) which is more suitable for the extraction from a large-scale corpus. We will describe this method and then discuss the difference between RAV and AV.

The RAV method can be divided into two steps. First, we identify the restricted contexts that words tend to appear in. Second, we count the number of distinct restricted contexts that the string in question appears in.

We define a restricted accessor pair as a pair of words that has the ability to match the majority of words. First, we define the matching between a pair of words $(\mathbf{w}_l, \mathbf{w}_r)$ and a string $\mathbf{v}$. We say that $(\mathbf{w}_l, \mathbf{w}_r)$ and $\mathbf{v}$ match if:

$$\frac{t(\mathbf{w}_l, \mathbf{v}, \mathbf{w}_r)}{f(\mathbf{v})} > \text{th\_RAV} \quad (12)$$

where th_RAV is a threshold. $f(\mathbf{v})$ is the number of times that $\mathbf{v}$ appears as a word or a sequence of words in the corpus segmented by our CWS model. $t(\mathbf{w}_l, \mathbf{v}, \mathbf{w}_r)$ is the number of times that the string $\mathbf{v}$ appears right after $\mathbf{w}_l$ and before $\mathbf{w}_r$ where $\mathbf{w}_l$ and $\mathbf{w}_r$ are also segmented as words.

Given a dictionary, we can find m word pairs which are most likely to match words in the dictionary. These pairs construct a set of restricted accessor pairs $\mathbf{R}$.

The RAV of a string v is defined as the number of restricted accessor pairs that match this string:

$$\text{RAV}(\mathbf{v}) = \sum_{(\mathbf{w}_l, \mathbf{w}_r) \in \mathbf{R}} 1_{[\text{th\_RAV}, \infty)} \left( \frac{t(\mathbf{w}_l, \mathbf{v}, \mathbf{w}_r)}{f(\mathbf{v})} \right) \quad (13)$$

where $1_{[\text{th\_RAV}, \infty)}$ is an indicator function to indicate whether $(\mathbf{w}_l, \mathbf{w}_r)$ and $\mathbf{v}$ match:

$$1_{[\text{th\_RAV}, \infty)}(x) = \begin{cases} 1 & x \in [\text{th\_RAV}, \infty) \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

The more restricted accessor pairs a string matches, the more likely it is a word.

Notice that the distribution-based method is usually used to measure the semantic distance between words. In our approach, by setting m (the number of restricted accessor pairs) to a relatively small number, the restricted accessor pairs can match any UW, no matter what meaning the word has or what word category it belongs to. Examples of the restricted accessor pairs will be shown in Section 4.3.

RAV is different from AV in at least four ways.

First, RAV is normalized. This prevents RAV from possible noise in the large-scale corpus. In such a corpus, a high frequency string tends to have more accessors which will bring noise to the AV criterion. In contrast, noise may be filtered out by a threshold in Eq. 13.

Second, RAV only considers restricted accessor pairs rather than any characters that precede or succeed the strings. This is also designed to further decrease the noise from a large-scale corpus.

Third, RAV does not need an ad hoc procedure to discard strings with adhesive characters, which prevents RAV from improperly discarding UWs. These adhesive characters in Chinese may also have the ability to be as a morpheme in a word. For example, "地" can be used as a function mark following an adverb, while it can also be a morpheme with the meaning "ground/territory" to form many UWs like "飞地" (enclave, literally "flying territory").

Last but not least, as RAV only concerns with a small number of restricted accessor pairs, RAV is more effective and efficient than AV in a large-scale corpus.

### 3.3 Combine Morphological and Distribu-tional Evidences

The morphological evidence and the distributional evidence represent the properties of different aspects of the "wordhood". The morphology-based method concerns with the possible character sequence that forms the string in question, and treats each occurrence of the string independently. The distribution-based method is not concerned with how a string is made up by characters, but with the context the string is in.

Evidence shows that these two methods are complementary; we expect to get a better performance by combining them. In this paper we only propose a simple way of linear combination.

## 4 Experiments

### 4.1 Dataset and Evaluation Method

A dictionary is needed to distinguish unknown words from known words. We used the same dictionary that Feng et al. (2004a) used. Totally 119,803 words in this downloaded dictionary are used as the known words.

SogouT corpus is an open and free large-scale Web corpus . This Web corpus was also used by Li and Sun (2009) in their semi-supervised CWS model. After certain process to remove non-text content such as the HTML tags, we obtained 119 million web pages consisting of 203 billion Chinese characters.

The whole corpus is denoted as LARGE. We sampled about one percent of these pages as a smaller corpus called MIDDLE, and further sampled about one percent of these pages in MIDDLE as SMALL. Corpora with different sizes are used to investigate how the size of the corpus influences the performances of different UW extraction methods.

It is difficult to evaluate the performance of UW extraction directly on such a large corpus. We used a partial evaluation method similar to the method used by Feng et al. (2004a). We sampled 2000 sentences from a balanced corpus (YUWEI corpus) consisting of news articles, academic articles, textbook articles, novels and other types of texts. Various UWs appear in these sentences. Since Chinese words commonly consist of 2, 3 or 4 characters (Chang and Su, 1997), only strings with length of 2, 3 or 4 in these sentences are considered as the UW candi-dates (strings in question). After filtering the strings that already appear in the dictionary, the remaining 111,536 strings are used as the test set.

In order to annotate these strings in the test set, we used the record of a Chinese input method software as an auxiliary data . We selected strings that are frequently inputted by users and manually annotated 1,630 of them as UWs. The annotation may have bias because the low frequency words tend to be ignored for the manually annotation. But the annotation is still independent of these UW extraction methods we use. Some of the UWs are 小诸葛 (nickname of a Chinese general Bai Chongxi), 二氯甲烷 (methylene chloride), 冬修 (to build in the winter by peasants in their slack

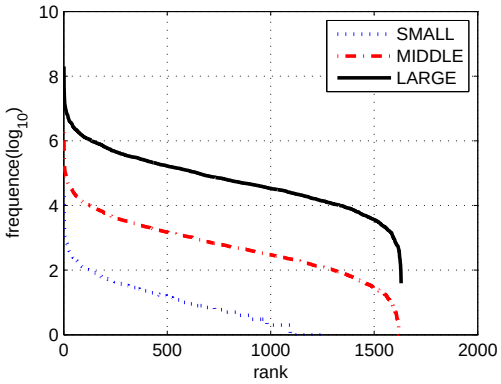


Figure 1: The frequencies of UWs in three corpora

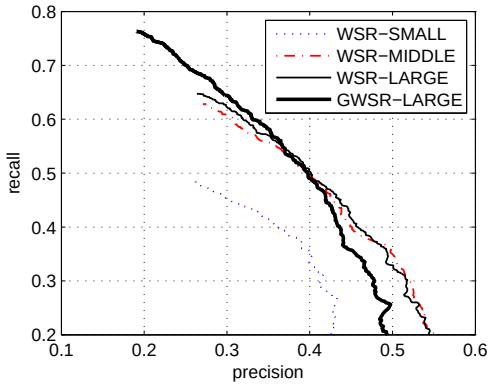


Figure 2: The precision-recall curves for WSR and GWSR on three corpora. The curves of GWSR on the SMALL and MIDDLE corpora are not showed due to the limitation of space.

season) and 官印 (official seal).

Figure 1 is an overview of the frequencies of the UWs in the corpora we used. We can observe the differences between these corpora. A number of UWs even do not appear in the SMALL corpus, while most of the UWs appear with a frequency higher than 10,000 in the LARGE corpus.

We use precision and recall as the evaluation measures:

$$\text{precision} = \frac{\text{\# of retrieved unknown words}}{\text{\# of retrieved words}} \quad (15)$$

$$\text{recall} = \frac{\text{\# of retrieved unknown words}}{\text{\# of annotated unknown words}} \quad (16)$$

The precision-recall curves of every method are drawn for the comparison. For each of these methods, a single threshold can be used to control the number of strings that are extracted. The precision-recall curves are drawn according to these thresholds.

Notice that in the evaluation, all the known words (words that are already in the dictionary) are not counted.

### 4.2 Morphological Evidence: WSR and GWSR

In this subsection we describe the implementation of our character tagging based CWS model, and the experiment results of the WSR and GWSR methods.

| Template |
|---|
| $c_{i-1}t_i$, $c_it_i$, $c_{i+1}t_i$ |
| $c_{i-2}c_{i-1}t_i$, $c_{i-1}c_it_i$, $c_ic_{i+1}t_i$, $c_{i+1}c_{i+2}t_i$ |
| $t_{i-1}t_i$ |

Table 1: The feature templates for the CWS model

The CWS model we use is based on the averaged perceptron (Collins, 2002). The features templates are listed in Table 1, which are similar to the templates used for a CRF-based model (Tseng et al., 2005).

The training set provided by Microsoft Research in SIGHAN bake-off 2005 (Emerson, 2005) is used to train our CWS model. The F-measure on the test set was 0.963. This is comparable with the reported best 0.964, which is from a CRF-based model (Tseng et al., 2005).

Additional techniques were used to speed up the decoding of our CWS model. A modified double-array trie (Aoe et al., 1992) data structure was implemented to store and retrieve the feature values. Fix-point numbers (integer) rather than floating point numbers are used for the calculation without losing accuracy. With some other minor improvements, the decoding speed of one process is up to 2 million characters per second. This makes it possible to segment the large-scale Web corpus on the fly.[1]

Since the WSR and GWSR for strings with low string frequency are not precise enough, we discard strings with low string frequency using the thresholds 0, 15 and 1,500 for the SMALL, MIDDLE and LARGE corpora, respectively. The threshold th_LW discussed in Section 3 for GWSR is set to 2.

Figure 2 shows the precision-recall curves for WSR and GWSR on three corpora.

On the SMALL corpus, the performance of

[1]The modified version is available at http://code.google.com/p/perminusminus/

WSR is poor, for the majority of UWs appear with a low frequency or even do not appear in this corpus. On the MIDDLE corpus, the performance is better than the one on the SMALL corpus. But on the LARGE corpus, the performance improvement is not observable comparing to the performance on the MIDDLE corpus. The GWSR behaves similarly to WSR when we enlarge the corpus. This phenomenon indicates that the methods based on the morphological evidence cannot benefit from a larger corpus if the frequencies of corresponding strings are high enough.

Consider this with Figure 1, we find that a frequency of about 100 is enough for WSR and GWSR to determine whether the corresponding string is a word or not.

Now we compare GWSR with WSR on the LARGE corpus in Figure 2. GWSR has a better performance in the left part of the precision-recall curve but a poorer performance in the right part of this curve. We can say that GWSR has a better recall but a poorer precision, which is consistent with our discussion in Section 3.1. Comparing to WSR, the GWSR can extract more UWs, while it brings in more noise as a side effect. The advantage of simply using GWSR instead of WSR is not obvious.

### 4.3 Distributional Evidence: AV and RAV

The words used to induce the restricted accessor pairs set are all the known words that appear in the 2,000 sentences we sampled from YUWEI corpus. We induce a restricted accessor pairs set of 50 word pairs, which are most likely to match these words. Some of the pairs are (#,#), (#, 和), (到, 的) and (没有,#). # is used as a special word to denote the beginning and the end of a sentence.

Among these 50 word pairs, only 3 pairs contain words with 2 characters. Other words are all single character words. Nearly all the words are function words. 和 (and), 到 (to), 的 (a particle) and 没有 ('no' or 'do not') which are in the pairs we showed are all frequently used function words.

We found that the word with the highest frequency in these pairs is "#", which is consistent with the claim by Li and Sun (2009) that punctuation marks are useful for CWS.

All these pairs have the ability to match a majority of words in different word categories or with different meanings. This result also benefits from the fact that Chinese words do not have inflection

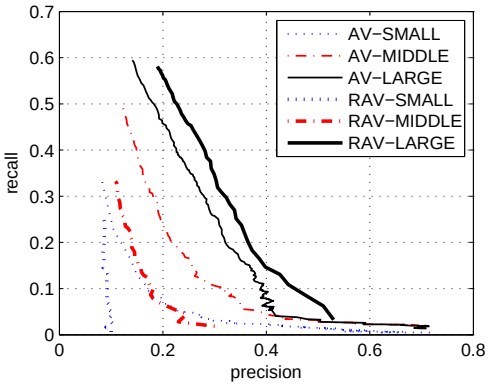


Figure 3: precision-recall curves for AV and RAV on three corpora

and agglutination. For example, an adjective like 幸福 (happy) can be used as a noun (happiness), a verb (be happy) or an adverb (happily) without changing the form.

We calculated AV and RAV for the strings in the test set on these three corpora. We found that RAV is more efficient than AV. According to the formulas, RAV only needs to assign a vector of 50 integers (we use 50 restricted accessor pairs in the experiments) for each string, while AV needs to assign a hash table for each string. Plus, the additional process for AV to discard strings with adhesive characters is also implemented according to the instruction in (Feng et al., 2004a).

Results are shown in Figure 3. Notice that the precision and recall for AV are lower than those reported by Feng et al. (2004a), for in our experiments the known words are not counted. Counting known words, the precision and recall are comparable with those reported by them.

We see that for both methods, larger size of the corpus improves the performances. We can even expect that more data can further enhance these methods.

AV and RAV behave differently when we enlarge the corpus. On both the SMALL and MIDDLE corpus, the AV method is better than the RAV method, whereas on the LARGE corpus, RAV outperforms AV. The reason is that the RAV method strongly depends on the size of corpus. Even 1000 occurrences may not be enough for a relatively accurate RAV of a string. This characteristic is also quite different from the WSR and LWSR methods.

Thus RAV is more suitable than AV when we have a large-scale corpus. Plus RAV does not need an ad hoc process to discard strings with adhesive

characters.

### 4.4 Combine Morphological and Distribu-tional Evidences

In this subsection we first show the differences of the errors made by the methods based on morphological and distributional evidences, respectively. Then we combine these two kinds of methods and show that the performance will be further improved.

(G)WSR and RAV are based on different evidences. The errors of these methods are thus quite different.

| Non-words | GWSR | RAV |
|---|---|---|
| 逍遥法 (part of "逍遥法外") | 0.879 | 6 |
| 脱但 (off but) | 0.817 | 2 |
| 一书里 (in a book of) | 0.671 | 10 |
| 一个女孩 (a girl) | 0 | 50 |
| 实验结果 (experiment result) | 0 | 49 |
| 严格把关 (to strictly check) | 0 | 43 |

Table 2: Some false positive examples for GWSR and RAV in the LARGE corpus

Table 2 shows some non-words that are incorrectly regarded as UWs by GWSR or RAV. Some non-words sun as 逍遥法 have high GWSR values, for they tend to be segmented as words by the CWS models. But they may have low RAV values for they are hard to be used as single syntactic units. Multi-word compounds such as 一个女孩 have high RAV for they have similar distribution as words. But they may have low GWSR because the CWS model tends to segment them into smaller parts with high confidence.

We linearly combine the scores of the morphological and distributional evidences. WSR and GWSR are used as the morphological evidence, respectively. For the AV method contains a filtering process, we cannot assign a value for every string. Only the RAV is used as the distributional evidence.

WSR and GWSR range from 0 to 1, while RAV ranges from 0 to 50. In our experiments, the weights for the morphological and distributional evidences are 50 and 0.8, respectively.

Two thick lines in Figure 4 show the performances of the combinations of the evidences. The precision of combing WSR and RAV (dashed thick line) is increased comparing to WSR. If we replace WSR by GWSR in the combined method

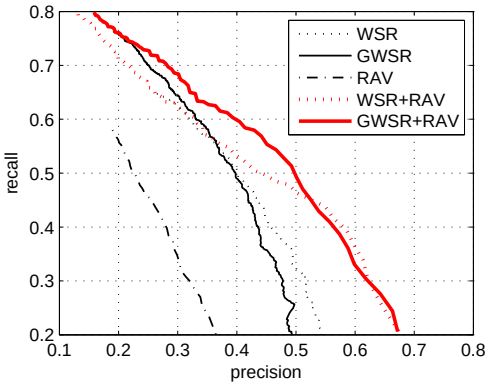


Figure 4: The combination of the morphological evidence and the distributional evidence on the LARGE corpus

(solid thick line), the recall is further increased without observably losing the precision. So the combination of GWSR and RAV outperforms the combination of WSR and RAV.

## 5 Conclusion

We discussed two UW extraction methods, namely morphology based and distribution based methods. The WSR based on morphological evidence has a relative high performance, while it does not benefit from the use of a large-scale corpus. The performance of the accessor variety (AV) based on distributional evidence improves gradually as we enlarge the corpus. We also proposed two extended methods. The method based on generalized word-string ratio (GWSR) has higher recall comparing to WSR. The restricted accessor variety (RAV) is specially designed for the large-scale web corpus in which the UWs are with high frequency.

Our methods outperformed the baselines, and the combination of the two methods can further improve the performance.

In the future, we will explore how to optimize the combination of GWSR and RAV to further improve UW extraction and the performance of the CWS models in general.

## Acknowledgments

This work is suooprted by the Tsinghua-Boeing Joint Research Project.

The author would like to thank Dr. Zhiyuan Liu for his helpful discussion, and Jianzhi Zeng for the proofreading of this paper.

## References


J. Aoe, K. Morimoto, and T. Sato. 1992. An efficient implementation of trie structures. *Software: Practice and Experience*, 22(9):695–721, September.

J. S Chang and K. Y Su. 1997. An unsupervised iterative method for chinese new lexicon extraction. *International Journal of Computational Linguistics & Chinese Language Processing*, 1(1):101 ‐ 157.

K. J. Chen and W. Y. Ma. 2002. Unknown word extraction for chinese documents. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

M. Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, page 1 ‐ 8.

T. Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 123–133. Jeju Island, Korea.

H. Feng, K. Chen, X. Deng, and W. Zheng. 2004a. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

H. Feng, K. Chen, C. Kit, and X. Deng. 2004b. Unsupervised segmentation of chinese corpus using accessor variety. *Natural Language Processing ‐ IJCNLP 2004*, pages 694–703.

C. M Hong, C. M Chen, and C. Y Chiu. 2009. Automatic extraction of new words based on google news corpora for supporting lexicon-based chinese word segmentation systems. *Expert Systems with Applications*, 36(2):3641 ‐ 3651.

W. Jiang, L. Huang, and Q. Liu. 2009. Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging ╟a case study. In *Proceedings of the 47th ACL*, page 522 ‐ 530, Suntec, Singapore, August. Association for Computational Linguistics.

Z. Jin and K. Tanaka-Ishii. 2006. Unsupervised segmentation of chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 428–435. Association for Computational Linguistics.

Z. Li and M. Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

G. O.H.C Ling, M. Asahara, and Y. Matsumoto. 2003. Chinese unknown word identification using character-based tagging and chunking. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, page 197 ‐ 200.

Y. Liu, B. Wang, F. Ding, and S. Xu. 2008. Information retrieval oriented word segmentation based on character associative strength ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1061–1069. Association for Computational Linguistics.

W. Y. Ma and K. J. Chen. 2003. A bottom-up merging algorithm for chinese unknown word extraction. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 31–38. Association for Computational Linguistics.

A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, page 188 ‐ 191.

H. T. Ng and J. K. Low. 2004. Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based. In *Proc of EMNLP*.

F. Peng, F. Feng, and A. McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th international conference on Computational Linguistics*, page 562. Association for Computational Linguistics.

M. Sun, D. Shen, and B. K Tsou. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1265–1271. Association for Computational Linguistics Morristown, NJ, USA.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171. Jeju Island, Korea.

N. Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

K. Zhang, M. Sun, and P. Xue. 2010. A local generative model for chinese word segmentation. *Information Retrieval Technology*, pages 420–431.

H. Zhao and C. Kit. 2008. An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. In *The Third International Joint Conference on Natural Language Processing (IJCNLP-2008), Hyderabad, India*.