# A Method for Automatic POS Guessing of Chinese Unknown Words

**Likun Qiu**
Department of Chinese Language and Literature, Peking University / No.5 Yiheyuan Road, Haidian District, Beijing, China 100871
NEC Laboratories, China
qiulk@pku.edu.cn

**Changjian Hu, Kai Zhao**
NEC Laboratories, China / 14F, Building.A, Innovation Plaza, No.1 Tsinghua Science Park, Zhongguancun East Road, Haidian District, Beijing, China 100084
{huchangjian, zhaokai}@research.nec.com.cn

## Abstract

This paper proposes a method for automatic POS (part-of-speech) guessing of Chinese unknown words. It contains two models. The first model uses a machine-learning method to predict the POS of unknown words based on their internal component features. The credibility of the results of the first model is then measured. For low-credibility words, the second model is used to revise the first model's results based on the global context information of those words. The experiments show that the first model achieves 93.40% precision for all words and 86.60% for disyllabic words, which is a significant improvement over the best results reported in previous studies, which were 89% precision for all words and 74% for disyllabic words. Further, the second model improves the results by 0.80% precision for all words and 1.30% for disyllabic words.

## 1 Introduction

Since written Chinese does not use blank spaces to denote word boundaries, Chinese word segmentation becomes an essential task for natural language processing, as in many other Asian languages (Thai, Japanese, Tibetan, etc.). It is difficult to build a complete dictionary comprising all words, for new words are constantly being created. As such, unknown words may greatly influence the effectiveness of text processing. Studies on unknown words include detection, POS guessing, sense classification, etc. Current methods for automatic unknown word detection have been relatively successful and widely used in many systems, yet automatic POS guessing for unknown words still remains a challenge for natural language processing research.

The task of POS guessing is quite different from traditional POS tagging. Traditional POS tagging involves assigning a single POS tag to a word token, provided that it is known what POS tag this word can take on in principle. This task requires a lexicon that lists possible POS tags for all words. However, unknown words are not in the lexicon, so the task of POS guessing of unknown words involves the guessing of a correct POS for an unknown word from the whole POS set of the current language. Obviously, traditional methods of POS tagging cannot effectively solve the problem of POS guessing of unknown words.

In previous work, two types of features have been used for the task of POS guessing of unknown Chinese words. One type is contextual feature, including local contextual features and global contextual features, and the other is internal component feature. Previous work has mainly used context information to guess the POS tags of unknown Chinese words, while a few designs looked at internal component features. Although there have been some attempts to combine the two types of features together, no reasonable explanation of the relationship between the two types of features has been given.

It is well known that the properties of a structure always depend on its internal component structure. As such, it is natural for us to wonder whether models based on internal component features alone can perform the POS guessing task for unknown Chinese words with both high pre-





*Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),* pages 705–712
Manchester, August 2008

cision and high recall. Here we present a model based on internal component features of unknown words using CRFs (conditional random fields). The results are very good, especially for multi-syllabic words (excluding disyllabic words).

While in the previous model the precision of POS guessing of disyllabic words is relatively high, there is still much room for further improvement. Considering that the usages of a word in real text can show the properties of a word, and that these features may be a useful complement to internal component features, we designed a scheme to effectively utilize the two types of features together. In this scheme, credibility scores for former guessing results are computed, and only those words with relatively lower credibility scores are revised by the model based on global context information. The model based on global context information simulates the behavior of linguists in judging the POS of a word. Though the recall of the latter model is low, it can revise some incorrect guessing results of the initial model.

According to Lu (2005), there are six main types of unknown Chinese words:

(1) Abbreviation (acronym): e.g., 中美 zhongmei (*China-U.S*).
(2) Proper names (person's name, place name, company name): e.g., 王安石 Wang Anshi (person's name), 槟城 Penang (an island in Malaysia; place name), 华为 Huawei (company name).
(3) Derived words (words with affixes): e.g., 总经理 zong-jingli (*general manager*), 现代化 xiandai-hua (*modernize*).
(4) Compounds: e.g., 获允 huoyun (*obtain permission*), 泥沙 nisha (*mud*), 突变 tubian (*sudden change*)
(5) Numeric compounds: e.g., 四千日圆 siqian riyuan (*four thousand Japanese yen*), ２００３年 2003nian (*year 2003*)
(6) Reduplicated words: e.g., 应不应该 yingbuyinggai (*should or should not*), 出出进进 chuchujinjin (*go in and go out*)

Proper names and numeric compounds are all nouns, so they don't have the problem of POS guessing. We will focus on abbreviation, derived words, compounds, and reduplicated words.

The remainder of this paper is organized as follows: in Section 2, we introduce some previous work on POS guessing of unknown Chinese words. Sections 3, 4, and 5 describe our proposed method, which includes two models and an additional process linking the two models together. In detail, Section 3 considers POS guessing for unknown Chinese words as a sequence-labeling problem and proposes a model based on internal component features to solve this task. In Section 4, we compute a credibility score for each guessing result based on the sequence type of the words' internal component structure. This links the model in Section 3 and that of Section 5 together. Section 5 describes a model based on global context information to revise the guessing results of the initial model that have relatively lower credibility scores. Section 6 shows the experiments and results of our methods and a comparison with previous work. Section 7 presents our conclusions.

## 2 Previous Work

Considering the features used during POS guessing, we have classified previous studies on POS guessing of unknown words into three types.

The first type use only contextual features, including local context and global context. For example, Nakagawa and Matsumoto (2006) proposed a probabilistic model to guess the POS tags of unknown words by considering all the occurrences of unknown words with the same lexical form in a document. The parameters were estimated using Gibbs sampling. They also attempted to apply the model to semi-supervised learning, and conducted experiments on multiple corpora. The highest precision in the Chinese corpus of their experiments was 67.85%.

The second type use only internal component features, such as that of Chen and Bai (1997) and Wu and Jiang (2000). Chen and Bai (1997) examined all unknown nouns, verbs, and adjectives and reported 69.13% precision using Dice metrics to measure the affix-category association strength and an affix-dependent entropy weighting scheme for determining the weightings between prefix-category and suffix-category associations. This approach is effective in processing derived words such as 现代化 xiandai-hua (*modernize*), but performs poorly when encountering compounds such as 保值 baozhi (*inflation-proof*). Wu and Jiang (2000) calculated *P(Cat,Pos,Len)* for each character, where *Cat* is the POS of a word containing the character, *Pos* is the position of the character in that word, and *Len* is the length of that word. They then calculated the POS probabilities for each unknown word as the joint probabilities of *P(Cat,Pos,Len)* for its com-

ponent characters. This approach was applied to unknown nouns, verbs, and adjectives of two to four characters in length. This approach exhibits lower recall for multi-syllabic words even if the training corpus is significantly large.

The third type attempt to combine internal component features and context information, such as that of Lu (2005) and Goh et al. (2006). Lu (2005) describes a hybrid model that combines a rule-based model with two statistical models for the task of POS guessing of unknown Chinese words. The rule-based model includes 35 manual rules concerning the type, length, and internal structure of unknown words, and the two statistical models utilize context information and the likelihood for a character to appear in a particular position of words of a particular length and POS category, one of which is Wu and Jiang's (2000) model. It achieves a precision of 89.00%, a significant improvement over the best result reported in previous studies, which was 69.00%. Goh et al. (2006) propose a method for guessing the part-of-speech tags of detected unknown words using contextual and internal component features with maximum entropy models. Both Lu (2005) and Goh et al. (2006) use only local context and not global context. As far as internal component features are concerned, Lu (2005) uses only the word category feature in his rule-based model while Goh et al. (2006) uses only the first character and last character features. From the above studies, we may find that methods based on internal component features are very promising, but this kind of features still needs much more attention. Moreover, none of them has proved that methods based on context information can improve the results of methods based on internal component features. They only attempted to utilize different types of features together and to give simultaneous results for both types of features.

Our method is among the third type of studies, but is different from the rest in the scheme of combining the two types of features together. In our method, internal component features play a more important role. We will prove that a model based on this type of features alone can perform very well. The other type of features acts as a useful supplement and can improve the results of some words in a certain degree. The two models are linked together by assigning a credibility score for each POS guessing result generated by the initial model. The results with a relatively lower credibility score are identified and put through reconsideration by a method based on global context information.

## 3 Model Based on Internal Component Features

In this model, we consider the task of POS guessing of unknown words as a problem of sequence labeling. The inspiration for this approach came from our observations of how humans understand a word. Usually an unknown word is regarded by human as a sequence of characters or morphemes that can be partitioned into several segments, where each segment is relatively coherent in meaning.

### 3.1 Conditional Random Fields

In contrast with other models for labeling sequences, such as HMM and MEMM, CRFs are good at avoiding the label bias problem. They condition on the entire observation sequence, thus avoiding the need for independent assumptions between observations and vastly expanding the set of features that can be incorporated into the model without violating its assumptions. One of the main advantages of a conditional model is its ability to explore a diverse range of features relevant to a specific task. As many studies have shown, CRFs are the best models for solving the sequence labeling problem (Lafferty et al., 2001; Vail et al., 2007). So we chose to use CRFs to solve the POS guessing problem.

### 3.2 The POS Guessing Model

While training, we use the words of a dictionary as training data. Those words will be considered as sentences and then segmented and assigned POS-tags by a standard word segmentation and POS tagging tool. By training with this data, we will obtain a POS guessing model for unknown words. While testing, we still consider an unknown word as a sentence and process it with the same tool.

In the dictionary, most words have one POS-tag while a few have more. The monosyllabic words were omitted from the training.

**Feature Analysis**

In our CRFs model, we employed three main types of features: the components of words, the lengths of those components, and the POS tags of those components.

Before the CRFs training, we analyzed the internal component structure of the dictionary words and assigned a proper POS-tag to each component. There are four analysis schemes that

are different from each other in two aspects. The first aspect is the type of the component, which may be a character or immediate constituent (IC). Here, "immediate constituent" means constituents that directly form a word. For example, 科学技术部 kexuejishubu (*department of science and technology*) has the following constituents: 科 ke, 学 xue, 技 ji, 术 shu, 部 bu, 科学 kexue, and 技术 jishu, in which only 科学 kexue, 技术 jishu and 部 bu are the immediate constituents of 科学技术部 kexuejishubu. The second aspect is regarding the consideration of the POS-tag of the component. The four analysis schemes are listed in Table 1.

| POS / Component | With | Without |
|---|---|---|
| Character | Scheme 1 | Scheme 2 |
| Immediate Constituent | Scheme 3 | Scheme 4 |

Table 1. Analysis Schemes

In Scheme 1 and Scheme 2, the tool segments a dictionary word until all components are characters, and in Scheme 1 only, each component is given a POS tag. In Scheme 3 and Scheme 4, the tool segments a dictionary word only once to get its immediate constituents, and in Scheme 3 only, each component is given a POS tag. For instance, 科学技术部 kexuejishubu (*department of science and technology*) will be segmented as "科/N 学/N 技/N 术/N 部/N" in Scheme 1 and "科学/N 技术/N 部/N" in Scheme 3.

**Feature Template Selection**

For each type of feature, we used the five templates in Figure 1. So in Schemes 1 and 3 there are 15 templates, and in Schemes 2 and 4 there are 10 templates.

For instance, when training, 科学技术部 will be transformed as 科学/N/N_B 技术/N/N_M 部/N/N_E in Scheme 3, in which N denotes that the POS of the word 科学技术部 is noun, while B, M and E denote the beginning, middle and end positions of the word.

U01:%x[-1,i]: the former component's *i*th feature
U02:%x[0,i]: the current component's *i*th feature
U03:%x[1,i]: the next component's *i*th feature
U04:%x[-1,i]/%x[0,i]: the former component's *i*th feature and the current component's *i*th feature
U05:%x[0,i]/%x[1,i]: the current component's *i*th feature and the next component's *i*th feature

Figure 1. List of Templates(i=1-3)

By using a dictionary with these feature templates for the training, we obtain a POS guessing model for unknown Chinese words. If an unknown word such as 用电 yongdian (*electricity used*) is tested, it would be analyzed as 用/V 电/N (*to use, electricity*) for the feature extraction and then it would be tagged as 用/V/N_B 电/N/N_E. That is, the word is assigned a POS of noun.

## 4 Credibility Computation

The initial model is based on the hypothesis that the syntactical properties of a word depend on its internal structure. But the internal structure of some words are ambiguous. For instance, both 用语 yongyu (*vocabulary* [literally, "used words"]) and 用劲 yongjing (*exert* [literally, "use strength"]) both have the sequence V1N1 (which is combined by a POS sequence of "VN" and a length sequence of "11"), yet the former one is a noun and the latter one is a verb.

There are some other sequences like V1N1. All the words (especially disyllabic words) fitting to these sequences bring difficulty for POS guessing model in Section 3. In this section, we attempt to identify those words by computing a credibility score for each type of sequence. The lower the score, less credible the result of the model.

In detail, Formula 1 is used to compute the credibility of a word that has a certain type of sequence, e.g., 用语 yongyu (*vocabulary* [literally, "used words"]) with the sequence "V1N1".

$$C_k = \frac{Count(S_k \mid P = P_j) - Count(S_k \mid P = P_{j+1})}{Count(S_k)} \qquad (1)$$

In Formula 1, $C_k$ denotes the credibility score of words that have the *k*th type of sequence $S_K$. $S_K$ denotes a sequence as $P_1L_1P_2L_2\ldots\ldots P_nL_n$, in which *n* means the quantity of components in the sequence $S_K$; $P_n$ *and* $L_n$ mean the POS and length of the *n*th component of any word that has the sequence $S_K$, respectively; $Count(S_K)$ denotes the quantity of words in the dictionary that have the sequence $S_K$; and $Count(S_K|P=P_j)$ and $Count(S_K|P=P_{j+1})$ denote the quantity of words in the dictionary that have the sequence $S_K$ and are tagged as POS $P_j$ and $P_{j+1}$, respectively, in which $P_j$ and $P_{j+1}$ are the two POSs that make the value of $Count(S_K|P=P_j)$ a maximum of two. Some sequences with lower credibility scores are listed in Table 2.

For instance, the sequence type of the word 用电 yongdian (*electricity used*) is V1N1, so its credibility score is 0.65. If the threshold is 0.8, this word will be considered a low-credibility word and put through reconsideration by the following model.

| Sequence | Credibility Score | Proportion |
|---|---|---|
| Vg1Ng1 | 0.7 | 0.50% |
| V1Ng1 | 0.67 | 1.90% |
| Vg1N1 | 0.67 | 0.55% |
| V1N1 | 0.65 | 2.99% |
| V1Vn2 | 0.57 | 0.04% |
| A1A1 | 0.55 | 0.22% |
| N1A1 | 0.48 | 0.15% |
| V1V2 | 0.44 | 0.05% |
| V1Vi2 | 0.25 | 0.08% |

Table 2. Examples of Sequences with Low Credibility Scores

## 5 Model Based on Global Contextual Features

Words with relatively lower credibility scores (given in Section 4) will be revised by a model based on global context. In this paper, we implement a model of voting by syntactical templates, which derived from research results of linguists.

This process requires a relatively large corpus that can provide enough context instances for each under-processed word. It is difficult for us to find a corpus that can provide enough instances of most unknown words, because many of such instances have only been used for a relatively short time. In this paper, we use search engine as the source of corpus, i.e., throw a word to a search engine and pick out instances from the returned snippets.

Linguists have summarized systematic rules for judging the POS of a Chinese word based on its global distribution in real text (Guo, 2002). For example, generally a verb or adjective can be modified by the word "不" (*not*) while a noun cannot. Based on this knowledge, we designed a set of syntactical templates listed in Table 3. The templates indicate whether a word can be used in such ways.

For every word, we build phrases based on these templates (see Table 3 for instances of 喜欢 xihuan (*like*)) and send the phrases to a search engine as queries. For each query, the search engine returns some snippets, which are generally in sentence form. Then each word gets three scores through a voting process in which the sentences act as "voters." The three scores, Score(N), Score(V), and Score(A), denote the likelihood score for the word to be a noun, a verb or an adjective, respectively. Each voter votes by following the criteria given in Figure 2. In Figure 2, Value(N), Value(V), and Value(A) are constant values that are used to balance the three scores.

| Templates | ~ | 不+~ | 开始+~ | 进行+~ | 予以+~ | 仍+~ | 所+~ | 很+~ | 很不+~ |
|---|---|---|---|---|---|---|---|---|---|
| Instance | 喜欢 | 不喜欢 | 开始喜欢 | 进行喜欢* | 予以喜欢* | 仍喜欢 | 所喜欢 | 很喜欢 | 很不喜欢 |

Table 3. Syntactical Templates with Instances of "喜欢"[1]

If the unknown word follows a transitive verb and is at the end of a sentence or subsentence, Score(N)+=Value(N);
If the unknown word follows a quantitative word and is at the end of a sentence or subsentence, Score(N)+=Value(N);
If the unknown word follows the word "不", "仍" or "所", Score(V)+=Value(V);
If the unknown word follows the word "开始", "进行" or "予以" and is at the end of a sentence or subsentence, or there is a following word that is not a verb, Score(V)+=Value(N);
If the unknown word follows the word "不", "很" or "很不", Score(A)+=Value(A).

Figure 2. Criteria for Voting

For each instance, Score(N), Score(V), and Score(A) will be added to the scores Value(N), Value(V), and Value(A), respectively.

Although these templates are effective, there are some exceptions brought by morphology analysis errors or other reasons, so we use an outstanding method to filter the exceptions. We

[1] Here "*" means the structure is invalid.

compute an outstanding value with Formula 2 to judge whether the voting result is acceptable.

$$O = \frac{Max(Score(POS)) - Max'(Score(POS))}{Max(Score(POS))} \quad (2)$$

In Formula 2, $O$ means the outstanding value of a voting result; $Max(Score(POS))$ means the maximum score among the three scores and $Max'(Score(POS))$ means the maximum score between the other scores. If $O$ is larger than a threshold, we assume the voting result to be acceptable and adopt the result to revise the POS guessing result of the initial model.

For instance, Score(N), Score(V), and Score(A) of the word 用语 yongyu (*vocabulary* [literally, "used words"]) are 50, 5 and 3 respectively. So O(用语)=(50-5)/50=0.9.

## 6 Experiments and Results

### 6.1 Data Preparation

The model based on CRFs is trained on the Modern Chinese Grammar Information Dictionary (Yu, 1998) and tested on the Contemporary Chinese Corpus of Peking University (Yu et al., 2002). The corpus is segmented and POS-tagged. Both the dictionary and corpus were constructed by the Institute of Computational Linguistics, Peking University. The corpus was built using the content of all the news articles of the People's Daily newspaper published in China from January to June 1998. We selected all verbs, nouns and adjectives from the dictionary, excluding monosyllabic words, as training data. The nouns, verbs and adjectives in the corpus but not in the dictionary were considered to be unknown words and used as testing data. The distribution of word length of the training and testing data is presented in Table 4.

| Word Length | Training | Testing |
|---|---|---|
| Disyllabic | 40,103 | 11,108 |
| Tri-syllabic | 12,167 | 12,901 |
| Four-character | 1,180 | 1,055 |
| Five-character | 0 | 279 |
| Total | 53,450 | 25,343 |

Table 4. Distribution of Word Length in Training and Testing Data

We used ICTCLAS 1.0 (Zhang, 2002) to do word segmentation and POS tagging, because ICTCLAS is known as one of the best tools for those functions. "CRF++, Yet Another CRF" toolkit (Kudo, 2005) was used as the implementation of CRFs model and www.Baidu.com as the search engine for our model based on contextual features.

### 6.2 Results of the Proposed Method

The results for the four schemes of our method based on internal component features are listed in Table 5[2]. From these results we may see that Scheme 1 is the best and Scheme 3 the second best, which means POS-tag of internal components is very useful feature in the POS guessing work. The comparison between Scheme 1 and Scheme 3 indicates that character-based scheme is good for processing tri-syllabic words and five-character words while IC-based scheme is good for processing disyllabic words and four-character words. Considering that most tri-syllabic words and five-character words are derivative words, while disyllabic words and four-character words are compounds, the results show that the character-based scheme is good for processing derivative words while the IC-based scheme is good for processing compounds. All the following improvements will be based on Scheme 1.

We assign the threshold of credibility score as 0.8, and then there are 2,234 words with a credibility score lower than the threshold. These words are then put through the revision process. In the revision model, we set the values of Value(N), Value(V), Value(A) and the outstanding threshold as 4, 1, 1, and 0.5, respectively, based on experience. All above thresholds are experimentally determined. Finally, 1,357 out of the 2,234 words pass the outstanding examination. Among them, 462 results were different from the former results and 302 of those were correctly revised, which resulted in the precision of disyllabic words reaching 87.90% (see Table 6). Moreover, other 895 words, which have the same result in the two models, reaches the precision of 91.2%. That means the credibility will be very high when the two models generate the same result.

Although we believe that the former method may have equal effectiveness to most man-made rules, there are still several rules that must be incorporated in order to simplify our machine-learning method. Here we incorporated two reduplication rules to process two types of reduplicated unknown words, respectively. The form of the first kind of words is "$V_1$不$V_2$, such as 应不应该 yingbuyinggai (*should or should*

[2] Precision, recall, F-measure are the same.

*not*) and the form of the second kind of words is "$V_1V_1V_2V_2$," such as 出出进进 chuchujinjin (*go in and go out*). If a four-character word is associated with one of the two forms and the first character is a verb, we revise its POS as a verb tag.

The two reduplication rules correctly revised 68 four-character words, which increased the precision of four-character words to 97.10%, a significant improvement over the previous best result, which was 92.89% (see Table 6).

| Word Length | Precision of Scheme 1 | Precision of Scheme 2 | Precision of Scheme 3 | Precision of Scheme 4 | Best Result |
|---|---|---|---|---|---|
| Disyllabic | 86.60% | 86.01 | 86.65% | 85.21% | 86.65% |
| Tri-syllabic | 99.22% | 99.17% | 98.65% | 97.48% | 99.22% |
| Four-character | 92.03% | 91.47% | 92.89% | 89.76% | 92.89% |
| Five-character | 100.00% | 98.20% | 98.92% | 98.92% | 100.00% |
| Total | 93.40% | 93.08% | 93.15% | 91.80% | 93.40% |

Table 5. Results for Four Schemes of The Model Based on Internal Component Features

| Word Length | Total number | Precision (Corresponding value of before) |
|---|---|---|
| Disyllabic | 11,108 | 87.90% (86.60%) |
| Tri-syllabic | 12,901 | 99.22% (99.22%) |
| Four-character | 1,055 | 97.10% (92.89%) |
| Five-character | 279 | 100% (100%) |
| Total | 25,343 | 94.20% (93.40%) |

Table 6. Results of Revision by Voting Model and Two Rules

### 6.3 Comparison with Previous Work

Wu & Jiang's (2000)[3] method is the most analogous with our method, yet they did not directly report the results in their paper. In this paper, we implement their model using the same data as our method. The results of Wu & Jiang's (2000) model are listed in Table 7. It shows that their model can guess the POS for disyllabic words with a relatively good F-measure (83.60%). However, the recall is not high for disyllabic (79.11%) and tri-syllabic (82.70%) words, and quite low for four-character (20.95%) and five-character (0%) words. Our model in Section 3 not only improves F-measure to 93.40%, but also improves recalls to 86.60%, 99.22%, 92.03% and 100% for multi-syllabic words in turn (Table 5, Scheme 1).

Lu (2005) proposed a hybrid model that achieved a precision of 89% for all words and 74% for disyllabic words. Compared with that method, the hybrid model in this paper improves the precision to 94.20% for all words and 87.90% for disyllabic words. Although the experiments were not taken on the same data, the figures reflect the difference of power between methods in a certain degree.

| Word Length | Total Number | Tagged Number | Precision | Recall | F-measure (Corresponding value of our method) |
|---|---|---|---|---|---|
| Disyllabic | 11,108 | 10,408 | 84.43% | 79.11% | 81.68% (87.90%) |
| Tri-syllabic | 12,901 | 11,091 | 96.20% | 82.70% | 88.94% (99.22%) |
| Four-character | 1,055 | 225 | 98.22% | 20.95% | 34.53% (97.10%) |
| Five-character | 279 | 0 | 0 | 0 | 0 (100%) |
| Total | 25,343 | 21,724 | 90.58% | 77.65% | 83.60% (94.20%) |

Table 7. Results of Wu & Jiang's (2000)

## 7 Conclusion and Future Work

The results of this experiment show that our model based on internal component features can achieve quite good results in POS guessing for unknown Chinese words, both in precision and recall. This proves that the internal component

[3] Lu (2005) implemented Wu & Jiang's (2000) model with a relatively small corpus as the training data. The precision of Wu & Jiang's method reported by the paper is 77.90% with a recall of 63.82%.

features of unknown words can be very useful in POS guessing. Moreover, the trained model based on internal component features is universal and robust. One evidence is that the model can identify POS correctly for most five-character words, even when there is no training data for that type of words.

Our results also show that the contextual features of unknown words can be an important complement to help improve POS guessing. Although models based on contextual features alone can't achieve the same precision and recall as models based on internal component features do, we may use contextual features as a complement in processing those words with ambiguous structure.

In contrast with Lu (2005), we don't use many manual rules. This does not mean that we believe those rules are useless in POS guessing. In fact, our initial model based on the CRFs model has learned the structure rules of Chinese words and can even give a credibility score for each rule. That is, most of the rules have been incorporated by the utilization of the CRFs model.

In the future, to improve the results, we attempt to manually revise the training data. Notice that the training data was formed by segmenting and tagging POS of each word in a dictionary using an existing tool like ICTCLAS. However, these tools usually generate quite a few errors on the words, because they are designed to handle sentence but not word. These errors were not revised in the experiment, which damaged the performance. Thus, by manually revising the training data, we hope to improve the results in a certain degree.

Although our experiments are mainly based on contemporary Chinese, we believe that this method will also be applicable to other Asian languages such as Japanese.

## References


Andy Wu and Zixin Jiang. 2000. Statistically-enhanced New Word Identification in a Rule-based Chinese System. In *Proceedings of the 2nd Chinese Language Processing Workshop*, pages 46–51.

Chao-Jan Chen, Ming-Hong Bai, and Keh-Jiann Chen. 1997. Category Guessing for Chinese Unknown Words. In *Proceedings of the Natural Language Processing Pacific Rim Symposium*, pages 35–40.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2006. Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing. In *Journal of Chinese Language and Computing* 16 (4):185-206

Douglas L. Vail, Manuela M. Veloso, and John D. Lafferty. 2007. Conditional Random Fields for Activity Recognition. In *Proceedings of 2007 International Joint Conference on Autonomous Agents and Multi-agent Systems*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of International Conference on Machine Learning*.

Kevin Zhang. ICTCLAS1.0. http://www.nlp.org.cn/project/project.php?proj_id=6.

Rui Guo. 2002. *Studies on Part-of-speech of Contemporary Chinese*. Commercial Press, Beijing, China.

Shiwen Yu. 1998. *Dictionary of Modern Chinese Grammar Information*. Tsinghua University Press. Beijing, China.

Shiwen Yu, Huiming Duan, Xuefeng Zhu, and Bing Sun. 2002. *The Basic Processing of Contemporary Chinese Corpus at Peking University*. Technical Report, Institute of Computational Linguistics, Peking University, Beijing, China.

T Nakagawa, Y Matsumoto. 2006. Guessing Parts-of-speech of Unknown Words Using Global Information. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of Association for Computational Linguistics*, pages 705–712.

Taku Kudo. 2005. *CRF++: Yet Another CRF toolkit*. http://chasen.org/~taku/software/CRF++.

Xiaofei Lu. 2005. Hybrid Methods for POS Guessing of Chinese Unknown Words. In *Proceedings of the 43th Annual Meeting of Association for Computational Linguistics Student Research Workshop*, pages 1–6.