Hybrid Methods for POS Guessing of Chinese Unknown Words

Xiaofei Lu
Department of Linguistics
The Ohio State University
Columbus, OH 43210, USA
xflu@ling.osu.edu

Abstract

This paper describes a hybrid model that combines a rule-based model with two statistical models for the task of POS guessing of Chinese unknown words. The rule-based model is sensitive to the type, length, and internal structure of unknown words, and the two statistical models utilize contextual information and the likelihood for a character to appear in a particular position of words of a particular length and POS category. By combining models that use different sources of information, the hybrid model achieves a precision of 89%, a significant improvement over the best result reported in previous studies, which was 69%.

1 Introduction

Unknown words constitute a major source of difficulty for Chinese part-of-speech (POS) tagging, yet relatively little work has been done on POS guessing of Chinese unknown words. The few existing studies all attempted to develop a unified statistical model to compute the probability of a word having a particular POS category for all Chinese unknown words (Chen et al., 1997; Wu and Jiang, 2000; Goh, 2003). This approach tends to miss one or more pieces of information contributed by the type, length, internal structure, or context of individual unknown words, and fails to combine the strengths of different models. The rule-based approach was rejected with the claim that rules are bound to overgenerate (Wu and Jiang, 2000).

In this paper, we present a hybrid model that combines the strengths of a rule-based model with those of two statistical models for this task. The three models make use of different sources of information. The rule-based model is sensitive to the type, length, and internal structure of unknown words, with overgeneration controlled by additional constraints. The two statistical models make use of contextual information and the likelihood for a character to appear in a particular position of words of a particular length and POS category respectively. The hybrid model achieves a precision of 89%, a significant improvement over the best result reported in previous studies, which was 69%.

2 Chinese Unknown Words

The definition of what constitutes a word is problematic for Chinese, as Chinese does not have word delimiters and the boundary between compounds and phrases or collocations is fuzzy. Consequently, different NLP tasks adopt different segmentation schemes (Sproat, 2002). With respect to any Chinese corpus or NLP system, therefore, unknown words can be defined as character strings that are not in the lexicon but should be identified as segmentation units based on the segmentation scheme. Chen and Bai (1998) categorized Chinese unknown words into the following five types: 1) acronyms, i.e., shortened forms of long names, e.g., *běi-dà* for *běijīng-dàxué* 'Beijing University'; 2) proper names, including person, place, and organization names, e.g., *Máo-Zédōng*; 3) derived words, which are created through affixation, e.g., *xiàndài-huà* 'modernize'; 4) compounds, which are created through compounding, e.g., *zhǐ-lǎohǔ* 'paper tiger'; and 5) nu-

Proceedings of the ACL Student Research Workshop, pages 1–6,
Ann Arbor, Michigan, June 2005.

meric type compounds, including numbers, dates, time, etc., e.g., *liǎng-diǎn* 'two o'clock'. Other types of unknown words exist, such as loan words and reduplicated words. A monosyllabic or disyllabic Chinese word can reduplicate in various patterns, e.g., *zǒu-zǒu* 'take a walk' and *piào-piào-liàng-liàng* 'very pretty' are formed by reduplicating *zǒu* 'walk' and *piào-liàng* 'pretty' respectively.

The identification of acronyms, proper names, and numeric type compounds is a separate task that has received substantial attention. Once a character string is identified as one of these, its POS category also becomes known. We will therefore focus on reduplicated and derived words and compounds only. We will consider unknown words of the categories of noun, verb, and adjective, as most unknown words fall under these categories (Chen and Bai, 1998). Finally, monosyllabic words will not be considered as they are well covered by the lexicon.

3 Previous Approaches

Previous studies all attempted to develop a unified statistical model for this task. Chen et al. (1997) examined all unknown nouns[1], verbs, and adjectives and reported a 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 blind to the type, length, and context of unknown words. 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 the *P(Cat,Pos,Len)* of its component characters. This approach was applied to unknown nouns, verbs, and adjectives that are two to four characters long[2]. They did not report results on unknown word tagging, but reported that the new word identification and tagging mechanism increased parser coverage. We will show that this approach suffers reduced recall for multisyllabic words if the training corpus is small. Goh (2003) reported a precision of 59.58% on all unknown words using Support Vector Machines.

Several reasons were suggested for rejecting the rule-based approach. First, Chen et al. (1997) claimed that it does not work because the syntactic and semantic information for each character or morpheme is unavailable. This claim does not fully hold, as the POS information about the component words or morphemes of many unknown words is available in the training lexicon. Second, Wu and Jiang (2000) argued that assigning POS to Chinese unknown words on the basis of the internal structure of those words will "result in massive overgeneration" (p. 48). We will show that overgeneration can be controlled by additional constraints.

4 Proposed Approach

We propose a hybrid model that combines the strengths of different models to arrive at better results for this task. The models we will consider are a rule-based model, the trigram model, and the statistical model developed by Wu and Jiang (2000). Combination of the three models will be based on the evaluation of their individual performances on the training data.

4.1 The Rule-Based Model

The motivations for developing a set of rules for this task are twofold. First, the rule-based approach was dismissed without testing in previous studies. However, hybrid models that combine rule-based and statistical models outperform purely statistical models in many NLP tasks. Second, the rule-based model can incorporate information about the length, type, and internal structure of unknown words at the same time.

Rule development involves knowledge of Chinese morphology and generalizations of the training data. Disyllabic words are harder to generalize than longer words, probably because their monosyllabic component morphemes are more fluid than the longer component morphemes of longer words. It is interesting to see if reduction in the degree of fluidity of its components makes a word more predictable. We therefore develop a separate set of rules for words that are two, three, four, and five

[1] Including proper names and time nouns, which we excluded for the reason discussed in section 2.

[2] Excluding derived words and proper names.

Chars	T1	T2	T3	T4	Total
2	1	2	1	2	6
3	2	6	2	5	15
4	2	2	0	8	12
5+	0	1	0	1	2
Total	5	11	3	16	35

Table 1: Rule distribution

or more characters long. The rules developed fall into the following four types: 1) reduplication rules (T1), which tag reduplicated unknown words based on knowledge about the reduplication process; 2) derivation rules (T2), which tag derived unknown words based on knowledge about the affixation process; 3) compounding rules (T3), which tag unknown compounds based on the POS information of their component words; and 4) rules based on generalizations about the training data (T4). Rules may come with additional constraints to avoid overgeneration. The number of rules in each set is listed in Table 1. The complete set of rules are developed over a period of two weeks.

As will be shown below, the order in which the rules in each set are applied is crucial for dealing with ambiguous cases. To illustrate how rules work, we discuss the complete set of rules for disyllabic words here[3]. These are given in Figure 1, where A and B refer to the component morpheme of an unknown AB. As rules for disyllabic words tend to overgenerate and as we prefer precision over recall for the rule-based model, most rules in this set are accompanied with additional constraints.

In the first reduplication rule, the order of the three cases is crucial in that if A can be both a verb and a noun, AA is almost always a verb. The second rule tags a disyllabic unknown word formed by attaching the diminutive suffix *er* to a monosyllabic root as a noun. This may appear a hasty generalization, but examination of the data shows that *er* rarely attaches to monosyllabic verbs except for the few well-known cases. In the third rule, a categorizing suffix is one that attaches to other words to form a noun that refers to a category of people or objects, e.g., *jiā* '-ist'. The constraint "A is not a verb morpheme" excludes cases where B is polysemous and does not function as a categorizing suffix

[3]Multisyllabic words can have various internal structures, e.g., a disyllabic noun can have a N-N, Adj-N, or V-N structure.

```
if A equals B
    if A is a verb morpheme, AB is a verb
    else if A is a noun morpheme, AB is a noun
    else if A is an adjective morpheme, AB is a stative
        adjective/adverb
else if B equals er, AB is a noun
else if B is a categorizing suffix AND A is not a verb
    morpheme, AB is a noun
else if A and B are both noun morphemes but not verb
    morphemes, AB is a noun
else if A occurs verb-initially only AND B is not a noun
    morpheme AND B does not occur noun-finally only,
    AB is a verb
else if B occurs noun-finally only AND A is not a verb
    morpheme AND A does not occur verb-initially only,
    AB is a noun
```

Figure 1: Rules for disyllabic words

but a noun morpheme. Thus, this rule tags *bèng-yè* 'water-pump industry' as a noun, but not *lí-yè* leave-job 'resign'. The fourth rule tags words such as *shā-xiāng* 'sand-box' as nouns, but the constraints prevent verbs such as *sōng-kòu* 'loosen-button' from being tagged as nouns. *Sōng* can be both a noun and a verb, but it is used as a verb in this word. The last two rules make use of two lists of characters extracted from the list of disyllabic words in the training data, i.e., those that have only appeared in the verb-initial and noun-final positions respectively. This is done because in Chinese, disyllabic compound verbs tend to be head-initial, whereas disyllabic compound nouns tend to be head-final. The fifth rule tags words such as *dīng-yǎo* 'sting-bite' as verbs, and the additional constraints prevent nouns such as *fú-xiàng* 'lying-elephant' from being tagged as verbs. The last rule tags words such as *xuě-bèi* 'snow-quilt' as nouns, but not *zhāi-shāo* pick-tip 'pick the tips'.

One derivation rule for trisyllabic words has a special status. Following the tagging guidelines of our training corpus, it tags a word ABC as verb/deverbal noun (v/vn) if C is the suffix *huà* '-ize'. Disambiguation is left to the statistical models.

4.2 The Trigram Model

The trigram model is used because it captures the information about the POS context of unknown words and returns a tag for each unknown word. We assume that the unknown POS depends on the previous two POS tags, and calculate the trigram probability $P(t_3|t_1, t_2)$, where t_3 stands for the unknown

POS, and t_1 and t_2 stand for the two previous POS tags. The POS tags for known words are taken from the tagged training corpus. Following Brants (2000), we first calculate the maximum likelihood probabilities $\hat{P}$ for unigrams, bigrams, and trigrams as in (1-3). To handle the sparse-data problem, we use the smoothing paradigm that Brants reported as delivering the best result for the TnT tagger, i.e., the context-independent variant of linear interpolation of unigrams, bigrams, and trigrams. A trigram probability is then calculated as in (4).

$$\hat{P}(t_3) = f(t_3)/N \quad (1)$$

$$\hat{P}(t_3|t_2) = f(t_2, t_3)/f(t_2) \quad (2)$$

$$\hat{P}(t_3|t_1, t_2) = f(t_1, t_2, t_3)/f(t_1, t_2) \quad (3)$$

$$P(t_3|t_1, t_2) = \lambda_1\hat{P}(t_3) + \lambda_2\hat{P}(t_3|t_2) + \lambda_3\hat{P}(t_3|t_1, t_2) \quad (4)$$

As in Brants (2000), $\lambda_1 + \lambda_2 + \lambda_3 = 1$, and the values of λ_1, λ_2, and λ_3 are estimated by deleted interpolation, following Brants' algorithm for calculating the weights for context-independent linear interpolation when the n-gram frequencies are known.

4.3 Wu and Jiang's (2000) Statistical Model

There are several reasons for integrating another statistical model in the model. The rule-based model is expected to yield high precision, as over-generation is minimized, but it is bound to suffer low recall for disyllabic words. The trigram model covers all unknown words, but its precision needs to be boosted. Wu and Jiang's (2000) model provides a good complement for the two, because it achieves a higher recall than the rule-based model and a higher precision than the trigram model for disyllabic words. As our training corpus is relatively small, this model will suffer a low recall for longer words, but those are handled effectively by the rule-based model. In principle, other statistical models can also be used, but Wu and Jiang's model appears more appealing because of its relative simplicity and higher or comparable precision. It is used to handle disyllabic and trisyllabic unknown words only, as recall drops significantly for longer words.

4.4 Combining Models

To determine the best way to combine the three models, their individual performances are evaluated in the training data first to identify their strengths. Based on that evaluation, we come up with the algorithm in Figure 2. For each unknown word, if the trigram model returns exactly one POS tag, that tag is prioritized, because in the training data, such tags turn out to be always correct. Otherwise, the guess returned by the rule-based model is prioritized, followed by Wu and Jiang's model. If neither of them returns a guess, the guess returned by the trigram model is accepted. This order of priority is based on the precision of the individual models in the training data. If the rule-based model returns the "v/vn" guess, we first check which of the two tags ranks higher in the list of guesses returned by Wu and Jiang's model. If that list is empty, we then check which of them ranks higher in the list of guesses returned by the trigram model.

```
for each unknown word
    if the trigram model returns one single guess, take it
    else if the rule-based model returns a non-v/vn tag, take it
    else if the rule-based model returns a v/vn tag
        if W&J's model returns a list of guesses
            eliminate non-v/vn tags on that list and return the
            rest of it
        else eliminate non-v/vn tags on the list returned by the
            trigram model and return the rest of it
    else if W&J's model returns a list of guesses, take it
    else return the list of guesses returned by the trigram
        model
```

Figure 2: Algorithm for combining models

5 Results

5.1 Experiment Setup

The different models are trained and tested on a portion of the Contemporary Chinese Corpus of Peking University (Yu et al., 2002), which is segmented and POS tagged. This corpus uses a tagset consisting of 40 tags. We consider unknown words that are 1) two or more characters long, 2) formed through reduplication, derivation, or compounding, and 3) in one of the eight categories listed in Table 2. The corpus consists of all the news articles from *People's Daily* in January, 1998. It has a total of 1,121,016 tokens, including 947,959 word tokens and 173,057 punctuation marks. 90% of the data are used for training, and the other 10% are reserved for testing. We downloaded a reference lexicon[4] containing 119,791

[4]From http://www.mandarintools.com/segmenter.html.

entries. A word is considered unknown if it is in the wordlist extracted from the training or test data but is not in the reference lexicon. Given this definition, we first train and evaluate the individual models on the training data and then evaluate the final combined model on the test data. The distribution of unknown words is summarized in Table 3.

Tag	Description
a	Adjective
ad	Deadjectval adverb
an	Deadjectival noun
n	Noun
v	Verb
vn	Deverbal noun
vd	Deverbal adjective
z	Stative adjective and adverb

Table 2: Categories of considered unknown words

Chars	Training Data		Test Data	
	Types	Tokens	Types	Tokens
2	2611	4789	387	464
3	3818	7378	520	764
4	490	1229	74	125
5+	188	698	20	56
Total	7107	14094	1001	1509

Table 3: Unknown word distribution in the data

5.2 Results for the Individual Models

The results for the rule-based model are listed in Table 4. Recall (R) is defined as the number of correctly tagged unknown words divided by the total number of unknown words. Precision (P) is defined as the number of correctly tagged unknown words divided by the number of tagged unknown words. The small number of words tagged "v/vn" are excluded in the count of tagged unknown words for calculating precision, as this tag is not a final guess but is returned to reduce the search space for the statistical models. F-measure (F) is computed as $2 * RP/(R + P)$. The rule-based model achieves very high precision, but recall for disyllabic words is low.

The results for the trigram model are listed in Table 5. Candidates are restricted to the eight POS categories listed in Table 2 for this model. Precision for the best guess in both datasets is about 62%.

The results for Wu and Jiang's model are listed in Table 6. Recall for disyllabic words is much higher than that of the rule-based model. Precision for disyllabic words reaches mid 70%, higher than that of the trigram model. Precision for trisyllabic words is very high, but recall is low.

Chars	Data	R	P	F
2	Training	24.05	96.94	38.54
	Test	27.66	96.89	43.03
3	Training	93.50	99.83	96.56
	Test	93.72	99.86	96.69
4	Training	98.70	99.02	98.86
	Test	99.20	99.20	99.20
5+	Training	99.86	100	99.93
	Test	100	100	100
Total	Training	70.60	99.40	82.56
	Test	69.72	99.34	81.94

Table 4: Results for the rule-based model

Guesses	1-Best	2-Best	3-Best
Training	62.01	93.63	96.21
Test	62.96	92.64	94.30

Table 5: Results for the trigram model

Chars	Data	R	P	F
2	Training	65.19	75.57	67.00
	Test	63.82	77.92	70.17
3	Training	59.50	98.41	74.16
	Test	55.63	99.07	71.25

Table 6: Results for Wu and Jiang's (2000) model

5.3 Results for the Combined Model

To evaluate the combined model, we first define the upper bound of the precision for the model as the number of unknown words tagged correctly by at least one of the three models divided by the total number of unknown words. The upper bound is 91.10% for the training data and 91.39% for the test data. Table 7 reports the results for the combined model. The overall precision of the model reaches 89.32% in the training data and 89.00% in the test data, close to the upper bounds.

6 Discussion and Conclusion

The results indicate that the three models have different strengths and weaknesses. Using rules that do not overgenerate and that are sensitive to the type, length, and internal structure of unknown words,

Chars	Training	Test
2	73.27	74.47
3	97.15	97.25
4	98.78	99.20
5+	100	100
Total	89.32	89.00

Table 7: Results for the combined model

the rule-based model achieves high precision for all words and high recall for longer words, but recall for disyllabic words is low. The trigram model makes use of the contextual information of unknown words and solves the recall problem, but its precision is relatively low. Wu and Jiang's (2000) model complements the other two, as it achieves a higher recall than the rule-based model and a higher precision than the trigram model for disyllabic words. The combined model outperforms each individual model by effectively combining their strengths.

The results challenge the reasons given in previous studies for rejecting the rule-based model. Overgeneration is a problem only if one attempts to write rules to cover the complete set of unknown words. It can be controlled if one prefers precision over recall. To this end, the internal structure of the unknown words provides very useful information. Results for the rule-based model also suggest that as unknown words become longer and the fluidity of their component words/morphemes reduces, they become more predictable and generalizable by rules.

The results achieved in this study prove a significant improvement over those reported in previous studies. To our knowledge, the best result on this task was reported by Chen et al. (1997), which was 69.13%. However, they considered fourteen POS categories, whereas we examined only eight. This difference is brought about by the different tagsets used in the different corpora and the decision to include or exclude proper names and numeric type compounds. To make the results more comparable, we replicated their model, and the results we found were consistent with what they reported, i.e., 69.12% for our training data and 68.79% for our test data, as opposed to our 89.32% and 89% respectively.

Several avenues can be taken for future research. First, it will be useful to identify a statistical model that achieves higher precision for disyllabic words, as this seems to be the bottleneck. It will also be relevant to apply advanced statistical models that can incorporate various useful information to this task, e.g., the maximum entropy model (Ratnaparkhi, 1996). Second, for better evaluation, it would be helpful to use a larger corpus and evaluate the individual models on a held-out dataset, to compare our model with other models on more comparable datasets, and to test the model on other logographic languages. Third, some grammatical constraints may be used for the detection and correction of tagging errors in a post-processing step. Finally, as part of a bigger project on Chinese unknown word resolution, we would like to see how well the general methodology used and the specifics acquired in this task can benefit the identification and sense-tagging of unknown words.

References

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pages 224–231.

Keh-Jiann Chen and Ming-Hong Bai. 1998. Unknown word detection for Chinese by a corpus-based learning method. *International Journal of Computational Linguistics and Chinese Language Processing*, 3(1):27–44.

Chao-Jan Chen, Ming-Hong Bai, and Keh-Jiann Chen. 1997. Category guessing for Chinese unknown words. In *Proceedings of NLPRS*, pages 35–40.

Chooi-Ling Goh. 2003. Chinese unknown word identification by combining statistical models. Master's thesis, Nara Institute of Science and Technology, Japan.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of EMNLP*, pages 133–142.

Richard Sproat. 2002. Corpus-based methods in Chinese morphology. Tutorial at the 19th COLING.

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.

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.