# Morphological features help POS tagging of unknown words across language varieties

**Huihsin Tseng**
Dept. of Linguistics
University of Colorado
Boulder, CO 80302
tseng@colorado.edu

**Daniel Jurafsky**
Dept. of Linguistics
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

**Christopher Manning**
Dept. of Computer Science
Stanford University
Stanford, CA 94305
manning@stanford.edu

## Abstract

Part-of-speech tagging, like any supervised statistical NLP task, is more difficult when test sets are very different from training sets, for example when tagging across genres or language varieties. We examined the problem of POS tagging of different varieties of Mandarin Chinese (PRC-Mainland, PRC-Hong Kong, and Taiwan). An analytic study first showed that unknown words were a major source of difficulty in cross-variety tagging. Unknown words in English tend to be proper nouns. By contrast, we found that Mandarin unknown words were mostly common nouns and verbs. We showed these results are caused by the high frequency of morphological compounding in Mandarin; in this sense Mandarin is more like German than English. Based on this analysis, we propose a variety of new morphological unknown-word features for POS tagging, extending earlier work by others on unknown-word tagging in English and German. Our features were implemented in a maximum entropy Markov model. Our system achieves state-of-the-art performance in Mandarin tagging, including improving unknown-word tagging performance on unseen varieties in Chinese Treebank 5.0 from 61% to 80% correct.

## 1 Introduction

Part-of-speech tagging is an important enabling task for natural language processing, and state-of-the-art taggers perform quite well, when training and test data are drawn from the same corpus. Part-of-speech tagging is more difficult, however, when a test set is drawn from a corpus that includes significantly different varieties of the language. One factor that may play a role in this cross-variety difficulty is the presence of test-set words that were unseen in cross-variety training sets.

We chose Mandarin Chinese to study this question of cross-variety and unknown-word POS tagging. Mandarin is both a spoken and a written language; as a written language, it is the official written language of the PRC (Mainland and Hong Kong), and Taiwan. Thus regardless of which dialect people speak at home, they write in Mandarin. But the varieties of Mandarin written in the PRC (Mainland and Hong Kong) and Taiwan differ in orthography, lexicon, and even grammar about as much as the British, American, and Australian varieties of English (or more in some cases). The corpus we use, Chinese Treebank 5.0 (Palmer et al., 2005), contains data from the three language varieties as well as different genres within the varieties. It thus provides a good data set for the impact of language variation on tagging performance.

Previous work on POS tagging of unknown words has proposed a number of features based on prefixes and suffixes and spelling cues like capitalization (Toutanova et al. 2003, Brants 2000, Ratnaparkhi 1996). For example, these systems followed Samuelsson (1993) in using n-grams of letter sequences ending and starting each word as unknown word features. But these features have mainly been tested on inflectional languages like English and German, whose derivational and inflectional affixes tend to be a strong indicator of word classes; Brants (2000), for example, showed that an English word ending in the suffix *-able* was very likely to be an adjective. Chinese, by contrast, has more than 4000 frequent affix characters. The amount of training data for each affix is thus quite sparse and (as we will show later) Chinese affixes are quite ambiguous in their part-of-speech identity. Furthermore, it is possible that n-gram features may not be suited to Chinese at all, since Chinese words are much shorter than English (averaging 2.4 characters per word compared with 7.7 for English, for unknown words in CTB 5.0 and Wall Street Journal (Marcus et el., 1993)).

In order to deal with these difficulties, we first performed an analytic study with the goal of understanding the morphological properties of unknown words in Chinese. Based on this analysis, we then propose new morphological features for addressing the unknown word problem. We also showed how these features could make use of a non-CTB corpus that had been labeled with very different POS tags, by converting those tags into features.

The remainder of the paper is organized as follows. The next section is concerned with a corpus analysis of cross language variety differences and introduces Chinese morphology. In Section 3, we evaluate a number of lexical, sequence, and linguistic features. Section 4 reviews related work and summarizes our contribution.

## 2 Data

Chinese Treebank 5.0 (CTB) contains 500K words of newspaper and magazine articles annotated with segmentation, part-of-speech, and syntactic constituency information. It includes data from three major media sources, XH[1] from PRC, HKSAR[2] from Hong Kong, and SM[3] from Taiwan. In terms of genre, both XH and HKSAR focus on politics and economic issues, and SM more on topics such as culture, health, education and travel. All of the files in CTB are encoded using Guo Biao (GB) and use simplified characters.

We did some cleanup of character encoding errors in CTB before running our experiments. Taiwan and Hong Kong still use the traditional forms of characters, while PRC-Mainland has adopted simplified forms of many characters, which also collapse some distinctions between characters. Additionally a different character set encoding is standardly used. The articles in HKSAR and SM originally used traditional characters and Big 5 encoding, but prior to inclusion in the CTB corpus they had been converted into simplified characters and GB. Some errors seem to have crept into this conversion process, accidentally leaving traditional characters such as 後 instead of simplified 后 (after), 於 for 于 (for), 甚麼 and 什麼 and 什么 (what), all of which we fixed. We also normalized half width numbers, alphabets, and punctuation to full width. Finally we removed the -NONE- traces left over from CTB parse trees.

## 3 Corpus analysis

We begin with an analytic study of potential problems for POS tagging on cross language variety data.

### 3.1 More unknown words across varieties?

We first test our hypothesis that a test set from a different language variety will contain more unknown words. Table 1 has the number of words in our devset that were unseen in the XH-only training set (we describe our training/dev/test split more fully in the next section). The devset contains equal amounts of data from all three varieties (XH, HKSAR, and SM). As table 1 shows, in data taken from the same source as the training data (XH), 4.63% of the words were unseen in training, compared to the much larger numbers of unknown words in the cross-variety datasets (14.3% and 16.7%). Some of this difference is probably due to genre as well, especially for the outlier-genre SM set.

Table 1 Percent of words in devset that were unseen in an XH-only training set. See Table 4 for more details.

| *Data Set* | *Lang Variety* | *Source* | *Genre* | *% unk* |
|---|---|---|---|---|
| XH | Mainland Mandarin | Xinhua | News | 4.6 |
| HKSAR | Hong Kong Mandarin | HKSAR | News | 14.2 |
| SM | Taiwan Mandarin | Sinorama | Magazine | 16.7 |
| **Devset** | Mix | Mix | Mix | 12.0 |

### 3.2 What are the unknown words?

In this section, we analyze the part-of-speech characteristics of the unknown words in our devset.

Table 2 Word class distribution of unknown words in devset, XH, HKSAR, SM. Devset represents the conjunction of the three varieties. CC, DT, LC, P, PN, PU, and SP are considered as closed classes by CTB.

| *Word class* | *Devset* | *XH* | *HKSAR* | *SM* |
|---|---|---|---|---|
| AD (adverb) | 74 | 2 | 23 | 49 |
| CC (coordinating conj.) | 7 | - | - | 7 |
| CD (cardinal number) | 151 | 108 | 23 | 20 |
| DT (determiner) | 10 | - | 6 | 4 |
| FW (foreign words) | 2 | 2 | - | - |
| JJ (other noun modifier) | 79 | 14 | 38 | 27 |
| LC (localizer/postposit) | 1 | - | 1 | - |
| M (measure word) | 12 | 2 | 4 | 6 |
| NN (common noun) | 1128 | 131 | 520 | 477 |
| NR (proper noun) | 400 | 92 | 156 | 152 |
| NT (temporal noun) | 53 | 3 | 38 | 12 |
| OD (ordinal number) | 4 | - | 4 | - |
| P (preposition) | 16 | 1 | 8 | 7 |
| PN (pronoun) | 10 | - | 3 | 7 |
| PU (punctuation) | 361 | - | 110 | 251 |
| SP(sentence final particle) | 1 | - | - | 1 |
| VA(predicative adjective) | 43 | 1 | 19 | 23 |
| VV (other verbs) | 497 | 25 | 215 | 257 |
| **Total** | 2849 | 381 | 1168 | 1300 |

Table 2 shows that the majority of Chinese unknown words are common nouns (NN) and verbs (VV). This holds both within and across different varieties. Beyond the content words, we find that 10.96% and 21.31% of unknown words are function words in HKSAR and SM data. Such unknown function words include the determiner *gewei* ("everybody"), the conjunction *huoshi* ("or"), the preposition *liantong* ("with"), the pronoun *nali* ("where"), and symbols used as quotes "「" and "」" (punctuation). XH does contain words with similar function (*huozhe*

[1] Xinhua Agency
[2] Information Services Department of Hong Kong Special Administrative Region
[3] Sinorama magazine

"or", *yu* "with", *dajia* "everybody", quotation marks "『"and "』"). Our result thus suggests that each Mandarin variety may have characteristic function words.

### 3.3 Cross language comparison

A key goal of our work is to understand the way that unknown words differ across languages. We thus compare Chinese, German, and English. Following Brants (2000), we extracted 10% of the data from the Penn Treebank Wall Street Journal (WSJ[4]) and NEGRA[5] (Brants et al., 1999) as observation samples to compare to the rest of the corpora.

In these observation samples, we found that Chinese words are more ambiguous in POS than English and German; 29.9% of tokens in CTB have more than one POS tag, while only 19.8% and 22.9% of tokens are ambiguous in English and German, respectively.

Table 3 shows that 40.6% of unknown words are proper nouns[6] in English, while both Chinese and German have less than 15% of unknown words as proper nouns. Unlike English, 60% of the unknown words in Chinese and German are verbs and common nouns. In the next section we investigate the cause of this similarity between Chinese and German unknown word distribution.

Table 3 Comparison of unknown words in English, German and Mandarin. The English and German data are extracted from WSJ and NEGRA. Chinese data is our CTB devset.

| *Language* | *English%* | *German%* | *Chinese%* |
|---|---|---|---|
| Proper nouns | 40.6 | 12.2 | 14.0 |
| Other nouns | 24.0 | 53.0 | 41.5 |
| Verbs | 6.8 | 11.4 | 19.0 |
| **ALL** | 100.0 | 100.0 | 100.0 |

## 4 Morphological analysis

In order to understand the causes of the similarity of Chinese and German, and to help suggest possible features, we turn here to an introduction to Chinese morphology and its implications for part-of-speech tagging.

[4] WSJ unknown words are those in WSJ 19-21 but unseen in WSJ 0-18; these are the devset and training set from Toutanova et al. (2003).

[5] The unknown words of NEGRA are words in a 10% randomly extracted set that were unseen in the rest of the corpus.

[6] We treat NNP (proper noun) and NNPS(proper noun plural) as proper nouns, NN(noun) and NNS(noun plural) as other nouns, and V* as verbs in WSJ. We treat NE (Eigennamen) as proper nouns, NN (Normales Nomen) as other nouns, and V* as verbs in NEGRA. We treat NR as proper nouns, NN and NT as other nouns, and V* as verbs in CTB.

### 4.1 Chinese morphology

Chinese words are typically formed by four morphological processes: affixation, compounding, idiomization, and reduplication, as shown in Table 4.

In affixation, a bound morpheme is added to other morphemes, forming a larger unit. Chinese has a small number of prefixes and infixes[7] and numerous suffixes (Chao 1968, Li and Thompson 1981). Chinese prefixes include items such as *gui* ("noble") in *guixing* ("your name"), *bu* ("not") in *budaode* ("immoral"), and *lao* ("senior") in *laohu* ("tiger") and *laoshu* ("mouse"). There are a number of Chinese suffixes, including *zhe* ("marks a person who is an agent of an action") in *zuozhe* ("author"), *shi* ("master") in *laoshi* ("teacher"), *ran* (-ly) in *huran* ("suddenly"), and *xin* (-ity or –ness) in *kenengxin* ("possibility").

Compound words are composed of multiple stem morphemes. Chao (1968) describes a few of the different compounding rules in Mandarin, such as coordinate compound, subject predicate compound, noun noun compound, adj noun compound and so on. Two examples of coordinate compounds are *anpai* ARRANGE-ARRANGE ("to arrange, arrangement") and *xuexi* STUDY-STUDY ("to study").

Table 4 Chinese morphological rules and examples

| | *Examples* |
|---|---|
| Prefix | *lao* ("senior") in *laohu* ( "tiger") |
| Suffix | *shi* ("master") in *laoshi* ("teacher") |
| Compounding | *xuexi* ("to study", "study") |
| Idiomization | *wanshiruyi* ("everything is fine") |
| Reduplication | *changchang* ("taste a bit") |

Compounding is extremely common in both Chinese and German. The phrase "income tax" is treated as an NP in English, but it is a word in German, *Einkommensteuer*, and in Chinese, *suodesui*. We suggest that it is this rich use of compounding that causes the wide variety of unknown common nouns and verbs in Chinese and German. However, there are still differences in their compound rules. German compounds can compose with a large number of elements, but Chinese compounds normally consist of two bases. Most German compounds are nouns, but Chinese has both noun and verb compounds.

Two final types of Chinese morphological processes that we will not focus on are idiomization (in which a whole phrase such as *wanshiruyi* ("everything is fine") functions as a word, and reduplication, in which a morpheme or word is repeated to form a new word such as the formation of *changchang* ("taste a

[7] Chinese only has two infixes, which are *de* and *bu* (not). We do not discuss infixes in the paper, because they are handled phrasally rather than lexically in CTB.

bit"), from *chang* "taste". (Chao 1968, Li and Thompson 1981).

## 4.2 Difficulty

The morphological characteristics of Chinese create various problems for part-of-speech tagging. First, Chinese suffixes are short and sparse. Because of the prevalence of compounding and the fact that the morphemes are short (1 character long), there are more than 4000 affixes. This means that the identity of an affix is often a sparsely-seen feature for predicting POS. Second, Chinese affixes are poor cues to POS because they are ambiguous; for example 63% of Chinese suffix tokens in CTB have more than one possible tag, while only 31% of English suffix tokens in WSJ have more than one tag. Most English suffixes are derivational and inflectional suffixes like *-able*, *-s* and *-ed*. Such functional suffixes are used to indicate word classes or syntactic function. Chinese, however, has no inflectional suffixes and only a few derivational suffixes and so suffixes may not be as good a cue for word classes. Finally, since Chinese has no derivational morpheme for nominalization, it is difficult to distinguish a nominalization and a verb.

These points suggest that morpheme identity, which is the major feature used in previous research on unknown words in English and German, will be insufficient in Chinese. This suggests the need for more sophisticated features, which we will introduce below.

# 5 Experiments

We evaluate our tagger under several experimental conditions: after showing the effects of data cleanup we show basic results based on features found to be useful by previous research. Next, we introduce additional morphology-based unknown word features, and finally, we experiment with training data of variable sizes and different language varieties.

## 5.1 Data sets

To study the significance of training on different varieties of data, we created three training sets: training set I contains data only from one variety, training set II contains data from 3 varieties, and is similar in total size to training set I. Training set III also contains data from 3 varieties and has twice much data as training set I. To facilitate comparison of performance both between and within Mandarin varieties, both the devset and the test set we created are composed of three varieties of data. The XH test data we selected was identical to the test set used in previous parsing research by Bikel and Chiang (2000). For the remaining data, we included HKSAR and SM data that is similar in size to the XH test set. Table 5 details characteristics of the data sets.

Table 5 Data set splits used. The unknown word tokens are with respect to Training I.

| *Data set* | *Sect'ns* | *Token* | *Unknown* |
|---|---|---|---|
| Training I | 26-270, 600-931 | 213986 | - |
| Training II | 600-931, 500-527, 1001-1039 | 204701 | - |
| Training III | 001-270, 301-527, 590-593, 600-1039, 1043-1151 | 485321 | - |
| *Devset* | | 23839 | 2849 |
| XH | 001-025 | 7844 | 381 |
| HKSAR | 500-527 | 8202 | 1168 |
| SM | 590-593, 1001-1002 | 7793 | 1300 |
| *Test set* | | 23522 | 2957 |
| XH | 271-300 | 8008 | 358 |
| HKSAR | 528-554 | 7153 | 1020 |
| SM | 594-596, 1040-1042 | 8361 | 1579 |

## 5.2 The model

Our model builds on research into loglinear models by Ng and Low (2004), Toutanova et al., (2003) and Ratnaparkhi (1996). The first research uses independent maximum entropy classifiers, with a sequence model imposing categorical valid tag sequence constraints. The latter two use maximum entropy Markov models (MEMM) (McCallum et al., 2000), that use log-linear models to obtain the probabilities of a state transition given an observation and the previous state, as illustrated in Figure 1 (a).

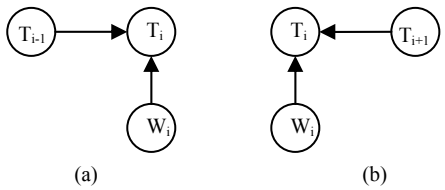


Figure 1 Graphical representation of transition probability calculation used in maximum entropy Markov models. (a) The previous state and the current word are used to calculate the transition probabilities for the next state transition. (b) Same as (a), but when model is run right to left.

Using left-to-right transition probabilities, as in Figure 1 (a), the equation for the MEMM can be formally stated as the following, where by $d_i$ represents the set of features the transition probabilities are conditioned on:

$$P(t, w) = \prod_i P(t_i \mid d_i)$$

Maximum entropy is used to calculate the probability $P(t_i \mid d_i)$ using the equation below. Here, $f_j(t_i, d_i)$ represents a feature derived from the available contextual information (e.g. current word, previous tag, next word, etc.)

$$P(t_i \mid d_i) = \frac{\exp(\sum_j \lambda_j f_j(t_i, d_i))}{\sum_{t' \in T} \exp(\sum_j \lambda_j f_j(t', d_i))}$$

We also used Gaussian prior to prevent overfitting. This technique allows us to utilize a large number of lexical and MEMM state sequence based features and also provides an intuitive framework for the use of morphological features generated from unknown word models.

## 5.3 Data cleanup

Before investigating the effect of our new features, we show the effects of data cleanup. Table 6 illustrates the .46 (absolute) performance gain obtained by cleaning character encoding errors and normalizing half width to full width.

We also clustered punctuation symbols, since training set I has too many (36) variety of punctuations, compared to 9 in WSJ. We clustered punctuations, for example grouping "《" and "〈" together. This mapping renders an overall improvement of .08%. All models in the following sections are then trained on font-normalized and punctuation clustered data.

Table 6 Improvement of tagging accuracy after data cleanup. The features used by all of the models are the identity of the two previous words, the current word and the two following word. No features based on the sequence of tags were used.

| *Models* | *Token A*[8] *%* | *Δ Token A%* | *Unk A %* |
|---|---|---|---|
| $2R_w+2L_w$ | 87.11 | - | 47.03 |
| +Cleanup | 87.57 | 0.46 | 48.54 |
| +PU | 87.65 | 0.08 | 49.26 |

## 5.4 Sequence features

We examined several tag sequence features from both left and right side of the current word. We use the term *lexical features* to refer to features derived from the identity of a word, and *tag sequence features* refer to features derived from the tags of surrounding words.

These features have been shown to be useful in previous research on English (Toutanova et al, 2003, Brants 2000, Thede and Harper 1999)

The models[9] in Table 7 list the different tag sequence features used; they also use the same lexical features from the model $2R_w+2L_w$ shown in Table 6. The table shows that Model $L_t+LL_t$ conditioning on the previous tag and the conjunction of the two previous tags yields 88.27%. As such, using the sequence features$<t_{i-1}, t_{i-1}t_{i-2}>$ achieves the current best result.

So far, there are no features specifically tailored toward unknown words in the model.

Table 7 Tagging accuracy of different sequence feature sets.

| *Models* | *Feature sets* | *Token A %* | *Unk A %* |
|---|---|---|---|
| $R_t+RR_t+2R_w+2L_w$ | $<t_i,t_{i+1}>,<t_i,t_{i+1},t_{i+2}>$ + lexical features | 88.10 | 50.11 |
| $L_t+LL_t+2R_w+2L_w$ | $<t_i,t_{i-1}>,<t_i,t_{i-1},t_{i-2}>$ +lexical features | 88.27 | 51.16 |

## 5.5 Unknown word model

Starting with Model $L_t+LL_t$ from the last section, we introduce 8 features to improve the performance of the tagger on unknown words. In the sections that follow, the model using affixation in conjunction with the basic lexical features described above is considered to be our baseline.

We considered words that occur less than 7 times in the training set I as rare; if $W_i$ is rare, an unknown word feature is used in place of a feature based on the actual word's identity. During evaluation, unknown word features are used for all words that occurred zero to 7 times in the training data. In addition, when tagging such rare and unknown words, we restrict the set of possible tags to just those tags that were associated with one or more rare words in the training data.

### 5.5.1 Affixation

Our affixation feature is motivated by similar features seen in inflectional language models. (Ng and Low 2004, Toutanova et al, 2003, Brants 2000, Ratnaparkhi 1996, Samuelsson 1993). Since Chinese also has affixation, it is reasonable to incorporate this feature into our model. For this feature, we use character *n*-gram prefixes and suffixes for *n* up to 4.[10] An example is:

资料袋 INFORMATION-BAG "folder"
$W_i$=资料袋 "a folder"
$F_{AFFIX}$={(prefix1,资), (prefix2,资料), (prefix3,资料袋), (suffix1,袋), (suffix2,料袋), (suffix3,资料袋)}

### 5.5.2 CTBMorph (CTBM)

While affix information can be very informative, we showed earlier that affixes in Chinese are sparse, short, and ambiguous. Thus as our first new feature we used a POS-vector of the set of tags a given affix could have. We used the training set to build a morpheme/POS dictionary with the possible tags for each

[8] We abbreviate accuracy as "A".

[9] Except where otherwise stated, during training, a count cutoff of 3 is applied to all features found in the training set. If a feature occurs fewer than 3 times, it is simply removed from the training data. All models are trained on training set I and evaluated on the devset.

[10] Despite the short average word length, we found that affixes up to size 4 worked better than affixes only up to size 2, perhaps mainly because they help with long proper nouns and temporal expressions.

morpheme. Thus for each prefix and suffix that occurs with each CTB tag in the training set I, we associate a set of binary features corresponding to each CTB tag. In the example below the prefix 资 occurred in both NN and VV words, but not AD or AS.

Prefix1=资, suffix1=袋
$F_{CTBM\text{-}pre}$= {(AD,0),(AS,0),…(NN,1),…(VV,1)}
$F_{CTBM\text{-}suf}$= {(AD,0),(AS,0),…(NN,1),…(VV,0)}

This model smoothes affix identity and the quantity of active CTBMorph features for a given affix expresses the degree of ambiguity associated with that affix.

Figure 2 Pseudo-code for CTBMorph

```
GenCTBMorphFeatureSet (Word W)
  FeatureSet f;
  for each t in CTB tag set:
    for each single-character prefix or suffix k of W
          if t.affixList contain k f.appendPair(t, 1)
      else f.appendPair(t, 0)
```

### 5.5.3 ASBC

One way to deal with robustness is to add more varied training data. For example the Academic Sinica Balanced Corpus[11] contains POS-tagged data from a different variety (Taiwanese Mandarin). But the tags in this corpus are not easily converted to the CTB tags. This problem of labeled data from very different tagsets can happen more generally. We introduce two alternative methods for making use of such a corpus.

#### 5.5.3.1 ASBCMorph (ASBCM)

The ASBCMorph feature set is generated in an identical manner to the CTBMorph feature set, except that rather than generating the morpheme table using CTB, another corpus is used. The morpheme table is generated from the Academic Sinica Balanced Corpus, ASBC (Huang and Chen 1995), a 5 M word balanced corpus written in Taiwanese Mandarin. As the CTB annotation guide[12] states, the mapping between the tag sets used in the two corpora is non-trivial. As such, the ASBC data can not be directly used to augment the training set. However, using our ASBCMorph feature, we are still able to derive some benefit out of such an alternative corpus.

#### 5.5.3.2 ASBCWord (ASBCW)

The ASBCWord feature is identical to the ASBCMorph feature, except that instead of using a table of tags that occur with each affix, we use a table of tags that a word occurs with in the ASBC data. Thus, a rare word in the CTB training/test set is augmented with features that correspond to all of the tags that the given word occurred with in the ASBC corpus, i.e. in this case, the pos tag of the identical word in ASBC, 资料袋.

$W_i$=资料袋
$F_{ASBCWord}$={(A,0),(Caa,0),(Cab,0)…(V_2,0)}

### 5.5.4 Verb affix

This feature set contains only two feature values, based on whether a list of verb affixes contains the prefix or suffix of an unknown word. We use the verb affix list created by the Chinese Knowledge Information Processing Group[13] at Academia Sinica. It contains 735 frequent verb prefixes and 282 frequent verb suffixes. For example,

Prefix1=资, suffix1=袋
$F_{verb}$={(verb prefix, 1), (verb suffix, 0)}

### 5.5.5 Radicals

Radicals are the basic building blocks of Chinese characters. There are over 214 radicals, and all Chinese characters contain one or more of them. Sometimes radicals reflect the meaning of a character. For example, the characters 猴 (monkey), 猪 (pig) 猫 (kitty cat) all contain the radical 犭 that roughly means "something that is an animal". For our radical based feature, we use the radical map from the Unihan database.[14] The radicals associated with the characters in the prefix and suffix of unknown words were incorporated into the model as features, for example:

Prefix1=资, suffix1=袋
$F_{RADICAL}$={(radical prefix, 贝), (radical suffix,衣)}

### 5.5.6 Named Entity Morpheme (NEM)

There is a convention that the suffix of a named entity points out the essential meaning of the named entity. For example, the suffix *bao* (newspaper) appears in Chinese translation of "WSJ", *huaerjierebao.* The suffix *he* (river) is used to identify rivers, for example in *huanghe* (yellow river).
To take advantage of this fact, we made 3 tables of named entity characters from the Chinese English Named Entity Lists (CENEL) (Huang 2002). These lists consist of a table of Chinese first name characters, a table of Chinese last name characters, and a

[11] The ASBC was originally encoded in traditional Big5 character, and we converted it to simplified GB.
[12] http://www.cis.upenn.edu/~chinese/posguide.3rd.ch.pdf
[13] http://turing.iis.sinica.edu.tw/affix/
[14] Unihan database is downloadable from their website: http://www.unicode.org/charts/unihan.html.

Table 8 Devset performance of the cumulatively rare word models, starting with the baseline. The second and third columns show the change in token accuracies and unknown word accuracies from the baseline for each feature introduced cumulatively. The fourth column shows the improvement from each feature set. The six columns on the right side of the table shows the error rate for the 5 most frequent tagsets of unknown words and the rest of unknown words.

| | | | | Error analysis: error rate % of unknown words in each POS | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Feature (add one in) | Token | Unk A% | Δ Unk A% | NN | VV | NR | PU | CD | Others |
| $L_t$+$LL_t$ | 88.27 | 51.16 | - | 16.67 | 57.14 | 68.25 | 100.00 | 16.56 | 60.86 |
| +Suffix | 89.70 | 60.74 | 9.58 | 12.50 | 41.65 | 44.75 | 100.00 | 5.30 | 37.25 |
| +Prefix → **baseline** | 90.03 | 63.66 | 2.92 | 10.55 | 36.62 | 40.00 | 100.00 | 3.97 | 34.76 |
| +CTBM | 91.48 | 76.13 | 12.47 | 13.74 | 31.99 | 36.00 | 1.99 | 0.00 | 20.58 |
| +ASBCM | 91.69 | 77.36 | 1.23 | 14.01 | 28.37 | 33.75 | 1.99 | 0.66 | 19.57 |
| +ASBCW | 91.85 | 78.84 | 1.48 | 13.30 | 23.54 | 33.50 | 1.42 | 0.00 | 17.93 |
| +Verb affix | 91.82 | 79.05 | 0.21 | 12.59 | 24.14 | 32.75 | 0.85 | 0.00 | 17.76 |
| +Radical | 91.85 | 79.09 | 0.04 | 11.88 | 24.75 | 33.50 | 0.85 | 0.00 | 18.78 |
| +NEM | 91.91 | 79.61 | 0.53 | 12.23 | 23.54 | 31.00 | 0.85 | 0.00 | 18.39 |
| +Length→**best** | 91.97 | 79.86 | 0.25 | 12.15 | 22.94 | 30.25 | 0.85 | 0.00 | 18.21 |

table of named entity suffixes such as organization, place, and company names in CENEL. Our named entity feature set contains 3 features, each corresponding to one of the three tables just described. To generate these features, first, we check if the prefix of an unknown is in the Chinese last name table. Second, we check if the suffix is in the Chinese first name table. Third, we check if the suffix of an unknown word is in the table of named entity suffixes. In Chinese last names are written in front of a first name, and the whole name is considered as a word, for example:

Prefix1=资, suffix1=袋
$F_{NEM}$={(last name, 0), (first name, 0), (NE suffix, 1)}

### 5.5.7 Length of a word

The length of a word can be a useful feature, because the majority of words in CTB have less than 3 characters. Words that have more than 3 characters are normally proper nouns, numbers, and idioms. Therefore, we incorporate this feature into the system. For example:

$W_i$=资料袋, $F_{length}$={(length , 3)}

### 5.5.8 Evaluation

Table 8 shows our results using the standard maximum entropy forward feature selection algorithm; at each iteration we add the feature family that most significantly improves the log likelihood of the training data given the model. We seed the feature space search with the features in Model $L_t$+$LL_t$. From this model, adding suffix information gives a 9.58% (absolute) gain on unknown word tagging. Subsequently adding in prefix makes unknown word accuracy go up to 63.66%. Our first result is that Chinese affixes are indeed informative for unknown words. On the right side of Table 8, we can see that this performance gain is derived from better tagging of common nouns, verbs, proper nouns, numbers and others. Because earlier work in many languages including Chinese uses these simple prefix and suffix features (Brants 2000, Ng and Low 2004) we consider this performance (63.66% on unknown words) as our baseline.

Adding in the feature set CTBM gives another 12.47% (absolute) improvement on unknown words. With this feature, punctuation shows the largest tagging improvement. The CTBM feature helps to identify punctuation since all other characters have been seen in different morpheme table made from the training set. That is, for a given word the lack of CTBM features cues that the word is a punctuation mark. Also, while this feature set generally helps all tagsets, it hurts a bit on nouns.

Adding in the ASBC feature sets yields another 1.23% and 1.48% (absolute) gains on unknown words. These two feature sets generally improve performance on all tagsets. Including the verb affix feature helps with common nouns and proper nouns, but hurts the performance on verbs. Overall, it yields 0.21% gain on unknown words. Finally, adding the radical feature helps the most on nouns, while subsequently adding in the name entity morphemes help to reduce the error on proper nouns by 2.50%. Finally, adding in feature length renders a 0.25% gain on unknown words. Commutatively, applying the feature sets results in an overall accuracy of 91.97% and an unknown word accuracy of 79.86%.

## 5.6 Experiments with the training sets of variable sizes and varieties

In this section, we compare our best model with the baseline model using different corpora size and language varieties in the training set. All the evaluations are reported on the test set, which has roughly equal amounts of data from XH, HKSAR, and SM.

The left column of Table 9 shows that when we train a model only on a single language variety and test on a mixed variety data, our unknown word accuracy is 79.50%, which is 18.48% (absolute) better than the baseline. The middle column shows when the training set is composed of different varieties and hence looks like the test set, performance of both the baseline and our best model improves.

Table 9 Comparison of the baseline and our best model. Using different training sets to evaluate on the test set. (McNemar's Test $p < .001$)

| | *Training I* | | *Training II* | | *Training III* | |
|---|---|---|---|---|---|---|
| | Token | Unk | Token | Unk | Token | Unk |
| Base-line | 89.17 | 61.02 | 92.54 | 74.78 | 93.51 | 81.11 |
| ***Best*** | *91.34* | *79.50* | *93.00* | *81.62* | *93.74* | *86.33* |

The right column shows the effect of doubling the training set size, using mixed varieties. As expected, using more data benefits both models.

These results show that having training data from different varieties is better than having data from one source. But crucially, our morphological-based features improve the tagging performance on unknown words even when the training set includes some data that resembles the test set.

How good are our best numbers, i.e. 93.7% on POS tagging in CTB 5.0? Unfortunately, there are no clean direct comparisons in the literature. The closest result in the literature is Xue et al. (2002), who retrain the Ratnaparkhi (1996) tagger and reach accuracies of 93% using CTB-I. However CTB-I contains only XH data and furthermore the data split is no longer known for this experiment (Xue p.c.) so a comparison is not informative. However, our performance on tagging when trained on Training I and tested on just the XH part of the test set is 94.44%, which might be a more relevant comparison to Xue et al. (2002).

## 6 Conclusion

Previous research in part-of-speech tagging has resulted in taggers that perform well when the training set and test set are both drawn from the same corpus. Unfortunately, for many potential real world applications, such an arrangement is just not possible.

Our results show that using sophisticated morphological features can help solve this robustness problem. These features would presumably also be applicable to other languages and NLP tasks that could benefit from the use of morphological information

Besides these tagging results, our research provides valuable analytic results on understanding the nature of unknown words cross-linguistically. Our results that unknown words in Chinese are not proper nouns like in English, but rather common nouns and verbs, suggest a similarity to German. We suggest this is because both German and Chinese, despite their huge differences in genetic, area, and other typological characteristics, tend to form unknown words through a similar word formation rule, compounding.

## 7 Acknowledgement

Thanks to Kristina Toutanova and Galen Andrew for their generous help and to the anonymous reviewers. This work was partially funded by ARDA AQUAINT and by NSF award IIS-0325646.

## 8 References

Bikel, Daniel and David Chiang. 2000. Two statistical parsing models applied to the Chinese Treebank. In *CLP* 2.

Brants, Thorsten. 2000. TnT: a statistical part-of-speech tagger. In *ANLP* 6.

Brants, Thorsten Wojciech Skut, Hans Uszkoreit. 1999. Syntactic Annotation of a German Newspaper Corpus In: *Anne Abeillé: ATALA sur le Corpus Annotés pour la Syntaxe Treebanks*.

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Huang, Chu-ren. and Keh-Jiann Chen. 1995. Academic Sinica Balanced Corpus. Technical Report 95-02/98-04. Academic Sinica.

Huang, Shudong. 2002. Chinese <-> English Name Entity Lists Version 1.0 beta. Catalog number: LDC2003E01.

Li, Charles and Sandra A Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

McCallum, Andrew, Dayne Freitag, Fernando Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *ICML* 17.

Marcus, Mitchel, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics*, 19.

Ng, Hwee Tou and Jin Kiat Low. 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? In *EMNLP* 9.

Martha Palmer, Fu-Dong Chiou, Nianwen Xue, Tsan-Kuang Lee. 2005. Chinese Treebank 5.0. Catalog number: LDC2005T01.

Ratnaparkhi, Adwait. 1996. A maximum entropy model forpart-of-speech tagging. In *EMNLP* 1.

Thede, Scott and Mary P. Harper. 1999. Second-order hidden Markov model for part-of-speech tagging. In *ACL 37*.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAACL* 2003.

Samuelsson, Christer. 1993. Morphological tagging based entirely on bayesian inference. In *NCCL* 9.

Xue, Nianwen, Fu-dong Chiou and Martha Palmer. 2002. Building a large-scale annotated Chinese corpus. In *COLING*.