A Morpheme-based Part-of-Speech Tagger for Chinese

Guohong Fu
School of Computer Science and Technology
Heilongjiang University
Harbin 150080, P.R. China
ghfu@hotmail.com

Jonathan J. Webster
Department of Chinese, Translation and Linguistics
City University of Hong Kong
83 Tat Chee Avenue, Hong Kong, P.R. China
ctjjw@cityu.edu.hk

Abstract

This paper presents a morpheme-based part-of-speech tagger for Chinese. It consists of two main components, namely a morpheme segmenter to segment each word in a sentence into a sequence of morphemes, based on forward maximum matching, and a lexical tagger to label each morpheme with a proper tag indicating its position pattern in forming a word of a specific class, based on lexicalized hidden Markov models. This system have participated four closed tracks for POS tagging at the Fourth International Chinese Language Processing Bakeoff sponsored by the ACL-SIGHAN.

1 Introduction

Part-of-speech (POS) tagging aims to assign each word in a sentence with a proper tag indicating its POS category. While a number of successful POS tagging systems have been available for English and many other languages, it is still a challenge to develop a practical POS tagger for Chinese due to its language-specific issues. Firstly, Chinese words do not have a strict one-to-one correspondence between their POS categories and functions in a sentence. Secondly, an ambiguous Chinese word can act as different POS categories in different contexts without changing its form. Thirdly, there are many out-of-vocabulary (OOV) words in real Chinese text whose POS categories are not defined in the dictionary used. All these factors make it much more difficult to achieve a high-performance POS tagger for Chinese.

Recent studies in Chinese POS tagging focus on statistical or machine learning approaches with either characters or words as basic units for tagging (Ng and Low, 2004; Fu and Luke, 2006). Very little research has been devoted to resolving Chinese POS tagging problems based on morphemes. In our system, we prefer morphemes to characters or words as tagging units for three reasons. First, words are made of morphemes instead of characters (Wu and Tseng, 1995; Packard, 2000). Second, most morphemes are productive in word formation (Baayen, 1989; Sproat and Shih, 2002; Nishimoto, 2003), particularly in the formation of morphologically-derived words (MDWs) and proper nouns, which are the major source of OOV words in Chinese texts. Third, Packard (2000) indicates that Chinese do have morphology. Moreover, morphology proves to be a very informative cue for predicting POS categories of Chinese OOV words (Tseng *et al*, 2005). Therefore, we believe that a morpheme-based framework would be more effective than the character- or word-based ones in capturing both word-internal morphological features and word-external contextual information for Chinese POS disambiguation and unknown word guessing (UWG) as well.

Thus we present a morpheme-based POS tagger for Chinese in this paper. It consists of two main components, namely a morpheme segmentation component for segmenting each word in a sentence into a sequence of morphemes, based on the forward maximum matching (FMM) technique, and a lexical tagging component for labeling each segmented morpheme with a proper tag indicating its position pattern in forming a word of a specific type, based on lexicalized hidden Markov models (HMMs). Lack of a large morphological knowledge base is a major obstacle to Chinese morphological analysis (Tseng and Chen, 2002). To overcome this problem and to facilitate morpheme-based POS tagging as well, we have also developed a statistically-based technique for automatically extracting morphemes from POS-tagged corpora. We participated in four closed tracks for POS tagging at the Fourth International Chinese Language Processing Bakeoff sponsored by the ACL-SIGHAN and tested our system on different testing corpora. In this paper, we also made a summary of this work and give some brief analysis on the results.

The rest of this paper is organized as follows: Section 2 is a brief description of our system. Section 3 details the settings of our system for different testing tracks and presents the scored results of our system at this bakeoff. Finally, we give our conclusions in Section 4.

2 System Description

2.1 Chinese Morphemes

In brief, Chinese morphemes can be classified into free morphemes and bound morphemes. A free morpheme can stand by itself as a word (viz. a basic word), whereas a bound morpheme can show up if and only if being attached to other morphemes to form a word. Free morphemes can be subdivided into true free morphemes and pseudo free morphemes. A pseudo free morpheme such as 然而 ran2-er2 'however' can only stand alone, while a true free morpheme like 生产 SHENG-CHAN 'produce' can stand alone by itself as a word or occur as parts of other words. Chinese affixes include prefixes (e.g. 非 fei1 'non-', 伪 wei3 'pseudo'), infixes (e.g. 分之 fei1-zhi1) or suffixes (e.g. 性 xing4 '-ity', 主义 zhu3-yi4 '-ism'), in terms of their positions within a word.

2.2 Formulation

To perform morpheme-based Chinese POS tagging, we represent a POS-tagged word in a Chinese sentence as a sequence of lexical chunks with the aid of an extended IOB2 tag set (Fu and Luke 2005). A lexical chunk consists of a sequence of constituent morphemes associated with their corresponding lexical chunk tags. A lexical chunk tag follows the format T1-T2, indicating the POS category *T2* of a word and the position pattern *T1* of a constituent morpheme within the word. As shown in Table 1, four position patterns are involved in our system, namely *O* for a single morpheme as a word by itself, *I* for a morpheme inside a word, *B* for a morpheme at the beginning of a word and *E* for a morpheme at the end of a word.

Tag	Definition	Corresponding morpheme types
O	A morpheme as a word by itself	Free morphemes
I	A morpheme inside a word	Free morphemes and infixes
B	A word-initial morpheme	Free morphemes and prefixes
E	A word-final morpheme	Free morphemes and suffixes

Table 1. Extended IOB2 tag set

2.3 Affix Extraction

Due to the increasing involvement of affixation in Chinese word formation, affixes play a more and more important role in Chinese POS tagging. In morpheme extraction, affixes are very useful in determining whether a given word is derived by affixation. To extract affixes from corpora, we consider three statistics, i.e. morpheme-position frequency $Count\,(m,T1)$, morpheme-position probability $MPP(m,T1) = Count(m,T1)/Count(m)$ and morphological productivity. Following the proposal in (Baayen, 1989), the morphological productivity of a morpheme m with a position pattern $T1$, denoted as $MP(m,T1)$, can be defined as

$$MP(m,T1) = \frac{n1(m,T1)}{Count(m,T1)} \qquad (1)$$

where $n1(m,T1)$ is the number of word types that occur only once in the training corpus and at the same time, are formed by the morpheme m with the position pattern $T1$.

To estimate the above statistics for affix extraction, we only take into account the three position patterns B, I and E, for prefixes, infixes and suffixes, respectively. Thus we can extract affixes from training data with the following three conditions: $Count(m,T1) \geq TH_{MPF}$, $MPP(m,T1) \geq TH_{MPP}$ and $MP(m,T1) \geq TH_{MP}$, where TH_{MPF}, TH_{MPP} and TH_{MP} are three empirically-determined thresholds.

2.4 Morpheme Extraction

The goal of morpheme extraction is to identify MDWs and proper nouns in training corpora and prevent them from getting into the morpheme dictionary for POS tagging. In the present system, the following criteria are applied to determine whether a word in training data should enter the morpheme dictionary.

Completeness. With a view to the completeness of the morpheme dictionary, all characters in training data will be collected as morphemes.

Word length. In general, shorter morphemes are more productive than longer ones in word formation. As such, the length of a morpheme should not exceed four characters.

Word frequency. By this criterion, a word is selected as a morpheme if its frequency of occurrences in training data is higher than a given threshold.

MDWs. By this criterion, words formed by morphological patterns such as affixation, compounding, reduplication and abbreviation will be excluded from the morpheme dictionary.

Proper nouns. In some training corpora like the PKU corpus, some special tags are specified for proper nouns. In this case, they will be used to filter proper nouns during morpheme extraction.

2.5 Lexicalized HMM Tagger

As shown in Figure 1, our system works in three main steps as follows.

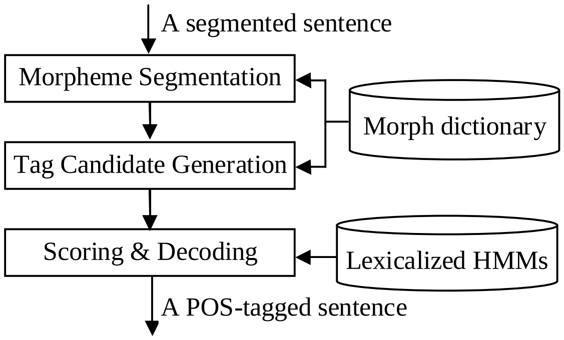

Figure 1. Overall architecture of our system

Morpheme segmentation. In this step, the FMM technique is employed to segment each word in a sentence to a sequence of morphemes associated with their position tags within the word.

Tag candidate generation. In this step, all possible POS candidates are generated for each word in the sentence by consulting the morpheme dictionary with its constitute morphemes and their related position patterns. All these candidates are stored in a lattice.

Scoring and Decoding. In this step, the lexicalized HMMs are first employed to score each candidate in the lattice and the Viterbi decoding algorithm is further used to search an optimal sequence of POS tags for the sentence. The details of lexicalized HMMs can be seen in (Lee *et al*, 200) and (Fu and Luke, 2005).

3 Evaluation Results

3.1 System Settings for Different Tracks

The POS tagging task at the fourth ACL-SIGHAN bakeoff consists of five closed tracks. We participated four of them, namely CKIP, CTB, NCC and PKU. Therefore our system is trained only using the relevant training corpora provided for the bakeoff. Furthermore, the morpheme dictionaries for these tracks are also extracted automatically from the relevant training data with the method presented in Sections 2.3 and 2.4. Table 2 illustrated the number of morphemes extracted from different training data.

Source	Training data (tokens/word types)	Number of morphemes
CKIP	721551 / 48045	30757
CTB	642246 / 42133	26330
NCC	535023 / 45108	28432
PKU	1116754 / 55178	30085

Table 2. Number of morphemes extracted from the training data for SIGHAN POS tagging bakeoff

3.2 Evaluation Results

Track	Total-A	IV-R	OOV-R	MT-R
CKIP-O	0.9124	0.9549	0.4756	0.8953
CTB-O	0.9234	0.9507	0.52	0.9051
NCC-O	0.9395	0.969	0.4086	0.9059
PKU-C	0.9266	0.9574	0.4386	0.9079

Table 3. Scores of our system for different tracks

Table 3 presents the scores of our system for different tracks. It should be noted that four measures are employed in the 4th ACL-SIGHAN bakeoff to score the performance of a POS tagging system, namely the overall accuracy (Total-A) and the recall with respect to in-vocabulary words (IV-R), OOV words (OOV-R) or multi-POS words (MT-R).

Although our system has achieved a promising performance, there is still much to be done to improve it. First, the quality of the morpheme dictionary is of particular importance to morpheme-based POS tagger. Although the present study proposed a statistical technique to extract morphemes from tagged corpora, further exploration is still needed on the optimization of this technique to acquire a more desirable morpheme dictionary for Chinese POS tagging. Second, morphological patterns prove to be informative cues for Chinese POS disambiguation and OOV word prediction. However, such a knowledge base is not publicly available for Chinese. As such, in the present study we only made use of certain surface morphological features, namely the position patterns of morphemes in word formation. Future research might usefully extend the present method to explore systematically more precise morphological features, including morpheme POS categories and morpho-syntactic rules for Chinese POS tagging.

4 Conclusion

In this paper we have presented a morpheme-based POS tagger for Chinese. We participated in four closed tracks at the fourth SIGHAN bakeoff. The scored results show that our system can achieve an overall accuracy of 0.9124-0.9395 for different corpora. However, the present system is still under development, especially in morphological knowledge acquisition. For future work, we hope to improve our system with a higher quality morpheme dictionary and more deep morphological knowledge such as morpheme POS categories and morpho-syntactic rules.

Acknowledgments

This study was supported in part by CityU Strategic Research Grant for fundable CERG (No. 7001879 & 7002037).

References

E. Nishimoto. 2003. Measuring and comparing the productivity of Mandarin Chinese suffixes. Computational Linguistics and Chinese Language Processing, 8(1): 49-76.

G. Fu and K.-K. Luke. 2005. Chinese named entity recognition using lexicalized HMMs. ACM SIGKDD Explorations Newsletter, 7(1): 19-25.

G. Fu and K.-K. Luke. 2006. Chinese POS disambiguation and unknown word guessing with lexicalized HMMs. International Journal of Technology and Human Interaction, 2(1): 39-50.

H. Tseng and K.-J. Chen. 2002. Design of Chinese morphological analyzer. In: Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing, 1-7.

H. Tseng, D. Jurafsky, and C. Manning. 2005. Morphological features help POS tagging of unknown words across language varieties. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

H.T. Ng and J.K. Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? Word-based or character-based?. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, 277-284.

J. Packard. 2000. Morphology of Chinese: A linguistic and cognitive approach. Cambridge University Press, Cambridge, UK.

R. Sproat and C. Shih. 2002. Corpus-based methods in Chinese morphology. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan.

R.H. Baayen. 1989. A corpus-based study of morphological productivity: Statistical analysis and psychological interpretation. Ph.D. thesis, Free University, Amsterdam.

S.-Z. Lee, T.-J. Tsujii, and H.-C. Rim. 2000. Lexicalized hidden Markov models for part-of-speech tagging. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Saarbruken, Germany, 481-487.

Z. Wu, G. Tseng. 1995. ACTS: An automatic Chinese text segmentation systems for full text retrieval. Journal of the American Society for Information Science, 46(2): 83-96.