Chapter 4.12
Chinese POS Disambiguation and Unknown Word Guessing with Lexicalized HMMs

Guohong Fu
The University of Hong Kong, Hong Kong

Kang-Kwong Luke
The University of Hong Kong, Hong Kong

ABSTRACT

This article presents a lexicalized HMM-based approach to Chinese part-of-speech (POS) disambiguation and unknown word guessing (UWG). In order to explore word-internal morphological features for Chinese POS tagging, four types of pattern tags are defined to indicate the way lexicon words are used in a segmented sentence. Such patterns are combined further with POS tags. Thus, Chinese POS disambiguation and UWG can be unified as a single task of assigning each known word to input a proper hybrid tag. Furthermore, a uniformly lexicalized HMM-based tagger also is developed to perform this task, which can incorporate both internal word-formation patterns and surrounding contextual information for Chinese POS tagging under the framework of HMMs. Experiments on the Peking University Corpus indicate that the tagging precision can be improved with efficiency by the proposed approach.

INTRODUCTION

While a number of successful part-of-speech (POS) tagging systems have been reported for English and many other languages over the past years, it is still a challenge to develop a practical Chinese POS tagger due to the language-specific issues in Chinese POS tagging. First, there is not a strict one-to-one correspondence for a Chinese word between its POS and its function in a sentence. Second, an ambiguous Chinese word can act as different POS categories in different contexts without changing its form. Third, there are many unknown words in real Chinese text whose POS categories are not defined in the dictionary used. Furthermore, there are almost no explicit morphological features in Chinese words, such as inflexion, derivation, and capitalization in English, that can be used directly for POS disambiguation and unknown word guessing (UWG). All these factors make it much more difficult to achieve a high-performance POS tagger for Chinese.

Recent study on POS tagging has focused on machine-learning approaches, including hidden Markov models (HMMs) (Brants, 2000; Weichedel et al., 1993), transformation-based error-driven learning (TBL) (Brill, 1995), maximum entropy model (Ratnaparkhi, 1996), support vector machines (SVMs) (Nakagawa et al., 2001), and loglinear models (Fu & Wang, 2002). Machine-learning approaches have the advantage of robustness. However, it is difficult for most machine-learning techniques to keep a balance between their capacity and their computational cost (Nakawa et al., 2002). Pla and Molina (2004) have proved that HMM-based taggers have the highest training and tagging speed. However, they usually achieve relatively low tagging precision, because standard HMMs only take into account contextual POS information for tagging. On the contrary, some learning models such as ME and SVMs are capable of handling much richer lexical information for POS tagging. However, they usually require higher computational cost, which inevitably will result in reduction of efficiency in training and tagging. This will be a serious problem in processing a large amount of data or in some online applications. In order to tackle this problem, some recent work, such as Lee, et al. (2000) and Pla and Molina (2004) suggested the use of lexicalization techniques to enhance the standard HMMs. Their experiments have shown that the lexicalized models can improve POS tagging precision without increasing much computational cost in training and processing.

POS disambiguation and unknown word guessing are two key issues in developing a Chinese POS tagger for practical applications. On the one hand, Chinese is highly ambiguous with respect to part of speech. Consequently, the first task of a POS tagger is how to find a proper POS for each ambiguous word in a sentence. On the other hand, there are many unknown or out-of-vocabulary (OOV) words in real Chinese texts whose POS categories are not defined in advance in the system dictionary. So, a practical tagger should be capable of predicting or guessing with accuracy the POS categories for these unknown words in an open-ended text.

Following the line of Lee et al. (2000) and Pla and Molina (2004), we propose in this article a unified approach to Chinese POS disambiguation and unknown word guessing. In order to explore word-internal morphological features for Chinese POS tagging, we introduce four types of pattern tags that indicate the way of a lexicon word to present itself in a real segmented sentence. Further, we define a hybrid tag set by merging these patterns with POS tags, with which Chinese POS disambiguation and unknown word guessing can be unified as a single process of assigning each known word (KW) in input a proper hybrid tag. Moreover, a statistical tagger is developed based on the uniformly lexicalized HMMs. In this way, three kinds of features — contextual tag information, surrounding lexical information, and word-internal word-formation patterns — can be incorporated for Chinese POS tagging under an efficient HMM-based framework. As a consequence, the tagger's performance can be improved with efficiency in training and tagging.

The rest of this article is organized as follows: The second section presents a novel formulation for Chinese POS tagging. Next a uniform lexicalization technique is introduced to enhance the standard HMMs for POS tagging. The fourth section gives a brief description of the tagging algorithm. Finally, we report our experimental results on the Peking University corpus in the fifth section and give our concluding remarks on this work in the final section.