Chinese Segmentation with a Word-Based Perceptron Algorithm

Yue Zhang and **Stephen Clark**
Oxford University Computing Laboratory
Wolfson Building, Parks Road
Oxford OX1 3QD, UK
{yue.zhang,stephen.clark}@comlab.ox.ac.uk

Abstract

Standard approaches to Chinese word segmentation treat the problem as a tagging task, assigning labels to the characters in the sequence indicating whether the character marks a word boundary. Discriminatively trained models based on local character features are used to make the tagging decisions, with Viterbi decoding finding the highest scoring segmentation. In this paper we propose an alternative, word-based segmentor, which uses features based on complete words and word sequences. The generalized perceptron algorithm is used for discriminative training, and we use a beam-search decoder. Closed tests on the first and second SIGHAN bakeoffs show that our system is competitive with the best in the literature, achieving the highest reported F-scores for a number of corpora.

1 Introduction

Words are the basic units to process for most NLP tasks. The problem of Chinese word segmentation (CWS) is to find these basic units for a given sentence, which is written as a continuous sequence of characters. It is the initial step for most Chinese processing applications.

Chinese character sequences are ambiguous, often requiring knowledge from a variety of sources for disambiguation. Out-of-vocabulary (OOV) words are a major source of ambiguity. For example, a difficult case occurs when an OOV word consists of characters which have themselves been seen as words; here an automatic segmentor may split the OOV word into individual single-character words. Typical examples of unseen words include Chinese names, translated foreign names and idioms.

The segmentation of known words can also be ambiguous. For example, "

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 840–847,
Prague, Czech Republic, June 2007.