Applying Conditional Random Fields to Chinese Shallow Parsing

Yongmei Tan, Tianshun Yao, Qing Chen and Jingbo Zhu

Natural Language Processing Lab
Northeastern University, Shenyang 110004
yongmeitan@hotmail.com, tsyao@mail.neu.edu.cn
chenqing@ics.neu.edu.cn, zhujingbo@mail.neu.edu.cn

Abstract. Chinese shallow parsing is a difficult, important and widely-studied sequence modeling problem. CRFs are new discriminative sequential models which may incorporate many rich features. This paper shows how conditional random fields (CRFs) can be efficiently applied to Chinese shallow parsing. We employ using CRFs and HMMs on a same data set. Our results confirm that CRFs improve the performance upon HMMs. Our approach yields the F1 score of 90.38% in Chinese shallow parsing with the UPenn Chinese Treebank. CRFs have shown to perform well for Chinese shallow parsing due to their ability to capture arbitrary, overlapping features of the input in a Markov model.

1 Introduction

Chinese shallow parsing is an important component of most text analysis systems in applications such as information extraction and summary generation. This problem has been widely studied and approached from different aspects. There are two main types of approaches to shallow parsing. One is base on rule-based methods; the other based on statistical methods. There is now a growing interest in applying machine-learning techniques to chunking, as they can avoid tedious manual work and are helpful in improving performance.

Much work has been done by researchers in this area. Li et al. used Maximum Entropy (ME) model to conduct Chinese chunk parsing [1], Zhang and Zhou used the inner structure and lexical information of base phrases to disambiguate border and phrase type [2]. Zhou et al. introduced the Chinese chunk parsing scheme and separated constituent recognition from full syntactic parsing, by using words boundary and constituent group information [3]. Zhao and Huang systematically defined Chinese base noun phrase from the linguistic point of view and presented a model for recognizing Chinese base noun phrases [4]. The model integrated Chinese base noun phrase structure templates and context features. These studies achieved promising results. However, comparing Chinese shallow parsing performance is difficult because those papers use different chunk definition and different data sets.

In this paper, we explore the practical issues in Chinese shallow parsing and present results on Chinese shallow parsing using Conditional Random Fields (CRFs).

CRFs [5] are models proposed recently that have the ability to combine rich domain knowledge, with finite-state decoding, sophisticated statistical methods, and discriminative, supervised training. In their most general form, they are arbitrary undirected graphical models trained to maximize the conditional probability of the desired outputs given the corresponding inputs. This method has been successfully applied in many NLP fields, such as POS tagging [5], noun phrase segmentation [6], Chinese word segmentation [7], named entity extraction [8] and Information Extraction [9][10].

In what follows, first, we briefly describe the general framework of Chinese shallow parsing and explore the practical issue in Chinese shallow parsing. Then, we describe CRFs, including how to conduct parameter estimation. Finally, we present experimental results and draw conclusions with possible future directions.

2 Chinese shallow parsing

Shallow parsing is the process of identifying syntactical phrases in natural language sentences. Several types of chunks – phrases that are derived from parse trees of Chinese sentences by flattening down the structure of the parse trees - provide an intermediate step to natural language understanding.

The pioneer work of Ramashaw and Marcus [11] has been proved to be an important inspiration source for shallow parsing. They formulate the task of NP-chunking as a tagging task where a large