A Simple Rule-based Approach to Organization Name Recognition in Chinese Text*

Wang Houfeng and Shi Wuguang

Institute of Computational Linguistics
School of Electronic Engineering and Computer Science
Peking University, Beijing, 100871, China
{wanghf, shiwuguang}@pku.edu.cn

Abstract. This paper presents a simple rule based approach to organization name recognition in Chinese text. Based on Chinese knowledge sources, our approach detects potential left and right boundaries in a text, and then determines whether a left-right boundary pair encloses an organization name by using a length constraint and non-organization name words/POS-tag constraints. Organization names with nested structure are also processed. This approach is easy to implement and the evaluation results are satisfactory.

1 Introduction

Named entity recognition (NER) is a fundamental component for many NLP applications, such as Information Extraction, Topic Detection and Tracking, Machine Translation and so forth. It is also a subtask of Message Understanding Conference. In recent years, this research has attracted much attention. Various approaches, have been proposed, e.g. maximum entropy model [1, 2], EM bootstrapping [3] and hidden Markov model [4]. While research work has mainly focused on NER in English, the study of NER in non-English has also made great advance [4, 5, 6]. However, organization name recognition in the Chinese text presents a special difficulty in NER, and no good performance in this field has been achieved yet.

In general, an integrated tool based on word segmentation and POS-tagging (Seg-Pos) is used to process many kinds of names, e.g. personal names, location names, date names and so on. Organization names, however, usually need to be processed independently. In [5] the author presented a role-based method by which all words are classified into roles, and as a result, the recognition of an organization name is simply reduced to role tagging. In [4], the author used both rule-based and statistics-based methods to identify organization names.

This paper focuses on multi-word organization name recognition. A one-word name, like 国务院 'State Council', is usually processed by a Seg-Pos tool.

* This work is funded by National Natural Science Foundation of Chinese (No. 60473138).