Fundamentals of Chinese Language Processing

Chu-Ren Huang
Dept. of Chinese and Bilingual Studies
Hong Kong polytechnic University
Churen.huang@inet.polyu.edu.hk

Qin Lu
Department of Computing
Hong Kong Polytechnic University
csluqin@comp.polyu.edu.hk

1 Introduction

This tutorial gives an introduction to the fundamentals of Chinese language processing for text processing. Today, more and more Chinese information are available in electronic form and over the internet. Computer processing of Chinese text requires the understanding of both the language itself and the technology to handle them. This tutorial is targeted for both Chinese linguists who are interested in computational linguistics and computer scientists who are interested in research on processing Chinese.

2 Content Overview

This tutorial consists of two parts. The first part overviews the grammar of the Chinese language from a language processing perspective based on naturally occurring data. The second part overviews Chinese specific processing issues and corresponding computational technologies.

The grammar introduced is a descriptive grammar of general-purpose, present-day standard Mandarin Chinese, which is fast becoming an internationally spoken language. Real examples of actual language use will be illustrated based on a data driven and corpus based approach so that its links to computational linguistic approaches for computer processing are naturally bridged in. A number of important Chinese NLP resources are also presented. On the technology side, the tutorial mainly covers Chinese word segmentation and Part-of-Speech tagging. Word segmentation problem has to deal with some Chinese language unique problems such as unknown word detection and named entity recognition which are the emphasis of this tutorial.

3 Tutorial Outline

Part 1: Highlights of Chinese Grammar for NLP
- 1.1 Preliminaries: Orthography and writing conventions
- 1.2 Basic unit of processing: word or character?
  - a. Word-forms vs. character forms
  - b. Word-senses vs. character-senses
- 1.3 Part-of-Speech: important issues in defining word classes
- 1.4 Word formation: from affixation to compounding
- 1.5 Unique constructions and challenges
  - a. Classifier-noun agreement
  - b. Separable compounds (or ionization)
  - c. 'Verbless' Constructions
- 1.6. Chinese NLP resources

Part 2: Text Processing
- 2.1 Lexical processing
  - a. Segmentation
  - b. Disambiguation
  - c. Unknown word detection
  - d. Named Entity Recognition
- 2.2 Syntactic processing
  - a. Issues in PoS tagging
  - b. Hidden Markov Models
- 2.3 NLP Applications

References

Academia Sinica Balance Corpus of Mandarin Chinese. http://www.sinica.edu.tw/SinicaCorpus/

Chao, Y. R. 1968. A Grammar of Spoken Chinese. Berkeley: University of California Press.

Huang, C.-R., K.-j. Chen and B. K. T'sou. 1996. Readings in Chinese Natural Language Processing. *Journal of Chinese Linguistics Monograph Series No. 9.* Berkeley: POLA.

T'sou, B. K. 2004. Chinese Language Processing at the Dawn of the 21st Century. In C.-R. Huang and W. Lenders. Eds. Computational Linguistics and Beyond. Pp. 189-206. Taipei: AcademiaSinica.

Miao, S.Q., Wei, Z.H. 2007, Chinese Text Information Processing Principles and Applications (In Chinese). Tsinghua University Press.

Tutorial Abstracts of ACL-IJCNLP 2009, page 1,
Suntec, Singapore, 2 August 2009.