What You Need to Know about Chinese for Chinese Language Processing

Chu-Ren Huang
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
churen.huang@inet.polyu.edu.hk

1 Introduction

The synergy between language sciences and language technology has been an elusive one for the computational linguistics community, especially when dealing with a language other than English. The reasons are two-fold: the lack of an accessible comprehensive and robust account of a specific language so as to allow strategic linking between a processing task to linguistic devices, and the lack of successful computational studies taking advantage of such links. With a fast growing number of available online resources, as well as a rapidly increasing number of members of the CL community who are interested in and/or working on Chinese language processing, the time is ripe to take a serious look at how knowledge of Chinese can help Chinese language processing.

The tutorial will be organized according to the structure of linguistic knowledge of Chinese, starting from the basic building block to the use of Chinese in context. The first part deals with characters as the basic linguistic unit of Chinese in terms of phonology, orthography, and basic concepts. An ontological view of how the Chinese writing system organizes meaningful content as well as how this onomasiological decision affects Chinese text processing will also be discussed. The second part deals with words and presents basic issues involving the definition and identification of words in Chinese, especially given the lack of conventional marks of word boundaries. The third part deals with parts of speech and focuses on definition of a few grammatical categories specific to Chinese, as well as distributional properties of Chinese PoS and tagging systems. The fourth part deals with sentence and structure, focusing on how to identify grammatical relations in Chinese as well as a few Chinese-specific constructions. The fifth part deals with how meanings are represented and expressed, especially how different linguistic devices (from lexical choice to information structure) are used to convey different information. Lastly, the sixth part deals with the ranges of different varieties of Chinese in the world and the computational approaches to detect and differentiate these varieties. In each topic, an empirical foundation of linguistics facts are clearly explicated with a robust generalization, and the linguistic generalization is then accounted for in terms of its function in the knowledge representation system. Lastly this knowledge representation role is then exploited in terms of the aims of specific language technology tasks. In terms of references, in addition to language resources and various relevant papers, the tutorial will make reference to Huang and Shi's (2016) reference grammar for a linguistic description of Chinese.

2 Resources

- Huang, Chu-Ren. 2009. Tagged Chinese Gigaword Version 2.0. Philadelphia: Lexical Data Consortium. University of Pennsylvania. ISBN 1-58563-516-2
- Sinica Corpus: Academia Sinica Balanced Corpus for Mandarin Chinese. http://www.sinica.edu.tw/SinicaCorpus
- Sinica BOW: Academia Sinica Bilingual Ontological Wordnet http://BOW.sinica.edu.tw
- Sinica TreeBank
  http://TreeBank.sinica.edu.tw/
- Chinese Wordnet 2005.
  http://cwn.ling.sinica.edu.tw
- Hantology 2006.
  http://hantology.ling.sinica.edu.tw

3 Outline

The tutorial will have six components according to the nature of linguistic knowledge of Chinese: 1)

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 23–24,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

characters, 2) words, 3) Parts of Speech, 4) Sentence and Structure, 5) Meaning: Representation and Expressive, and 6) Variations and Changes. Under each knowledge component, there will be 3 to 5 focus areas. In addition, relevant resources and language technology applications will be introduced together with the linguistic description or at the end of the lecture sections (for those language processing applications involving more than one linguistic issue.) Overall, two lecture sections of 80 minutes each will be given, each containing 5 topical groups (each topical group covers 2-3 focus areas described above). It is estimated that each topic group will take about 15 minutes to cover. Although the 15 minutes will not be enough for explication of finer details, participants will be able to access and acquire additional details from a comprehensive list references.

The three hour teaching plan is given below.

00:00-01:20 Characters, Words, and Parts-of-Speech

- -Component structure of Chinese characters: encoding and ontological issues
- -Writing system and processing of Chinese texts: myths and facts
- -Definition and identification of words in Chinese: with special foci on segmentation, and compounds
- -PoS and tagging in Chinese, with special foci on de, adjectives (or verbs), prepositions, and classifiers
- -Related issues and examples in Chinese Language processing

01:20-01:40: Coffee Break

01:40-03:00 Sentence, Meaning, and Variations

- -Aspectual and eventive systems of Chinese
- -Identification of grammatical relations: ba/bei, topic/argument, separable compounds and oblique arguments
- -Semantic relations and semantic selection
- -World Chineses: variations and changes and how to identify them
- -Related issues and examples in Chinese Language processing

4 Instructor

Chu-Ren Huang is currently a Chair Professor at the Hong Kong Polytechnic University. He is a Fellow of the Hong Kong Academy of the Humanities, a permanent member of the International Committee on Computational Linguistics, and President of the Asian Association of Lexicography. He currently serves as Chief Editor of the Journal Lingua Sinica, as well as Cambridge University Press? Studies in Natural Language Processing. He is an associate editor of both Journal of Chinese Linguistics, and Lexicography. He has served advisory and/or organizing roles for conferences including ALR, ASIALEX, CLSW, CogALex, COLING, IsCLL, LAW, OntoLex, PACLIC, ROCLING, and SIGHAN. Chinese language resources constructed under his direction include the CKIP lexicon and ICG, Sinica, Sinica Treebank, Sinica BOW, Chinese WordSketch, Tagged Chinese Gigaword Corpus, Hantology, Chinese WordNet, and Emotion Annotated Corpus. He is the co-author of a Chinese Reference Grammar (Huang and Shi 2016), and a book on Chinese Language Processing (Lu, Xue and Huang in preparation).