Book Review

Introduction to Chinese Natural Language Processing

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
(Chinese University of Hong Kong, Hong Kong Polytechnic University,
City University of Hong Kong, and San Diego State University)

Princeton, NJ: Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 4), 2010, x+148 pp; paperbound, ISBN 978-1-59829-932-8, $40.00; e-book, ISBN 978-1-59829-933-5, $30.00
or by subscription

Reviewed by
Min Zhang
Institute for Infocomm Research

This introductory book is a systematic and up-to-date overview of fundamental and focused knowledge of Chinese NLP for both practitioners and a general audience. The Chinese language has the largest number of native speakers in the world, with over one billion people speaking some style of Chinese. The globalization and the development of the Internet have significantly increased the participation of Chinese-speaking users in global business and social life, accounting for the highest growth rate in on-line population over the past decade.

Despite its increasing importance, the uniqueness of Chinese renders the computer processing of the language distinctive and challenging. With a clear awareness of many differences between Chinese and other languages in morphology, syntax, and semantics, the authors focus the subject matter on morphological analysis, which means they choose to pay close attention to the essential processing techniques that lay down the foundation of any advanced Chinese NLP system. For many readers who are puzzled by "why on earth we have to bother with Chinese NLP given the availability of English-language processing technologies," this book is suitable puzzle-solving material; for undergraduate and postgraduate students interested in the field, knowledge regarding the fundamentals of computer processing of Chinese language can be acquired; for a veteran already who knows everything in Chinese NLP, this is a useful and lightweight reference book. Overall, this book has a large potential audience of readers who are interested in Chinese NLP at all levels.

The book consists of eight chapters of main material, an appendix of linguistic resources, and a bibliography. Chapter 1 gives an introduction by raising the unique characteristics of Chinese. A couple of interesting examples are used to explain the problem of ambiguity caused by the lack of clear delimiters between words in Chinese text. Then it proceeds to illustrate how other types of difficulties are produced at morphological, syntactic, and semantic levels. From these discussions, Chapter 1 highlights and concludes that "The main difference in NLP between Chinese and other languages takes place in the first stage. This lays down the objective of the book, namely, to introduce the basic techniques in Chinese morphological analysis."

Chapter 2 describes the basic forms of morphological construction in Chinese written script, providing the necessary preparation for morphological analysis in the later chapters. It introduces the concepts of characters, morphemes, and words, as well as the typical morphological processes of word formation. In particular, the authors

provide sufficient details on the nature of the compounding that plays a predominant role in Chinese word formation and appears much more prevalent than that in English. A number of examples make the content easy to follow, especially for international learners who are not familiar with Chinese.

Given these primary linguistic characteristics, Chapter 3 particularizes the challenges that arise from a wide spectrum of morphological problems of Chinese. Compared to an alphabetic system like English, the difficulties of Chinese morphological processing not only result from its representation and writing conventions, such as a large character set, multiple co-existing variants of character sets and encoding standards, dialectal variations, special genres of punctuation, and so on, but also from its underlying linguistic heritage and distinctions. For example, the lack of formal morphological markers may render part-of-speech tagging troublesome due to few part-of-speech clues and multiple parts of speech for the same word; the indeterminacy may be worsened by homophony and homography. The authors emphasize the influences of different kinds of ambiguities and out-of-vocabulary (OOV) words and suggest that external knowledge and contextual information are important to ambiguity resolution.

Chapter 4 reviews Chinese word segmentation, a major challenge unique to Chinese and a few other Asian languages. A tree-based taxonomy is used to classify segmentation algorithms into character- and word-based approaches. Besides the rudimentary methods, emphasis is given to those methods using statistics and machine learning that take the advantage of large-scale annotated corpora to solve the segmentation ambiguity issue. Three main Chinese word segmentation standards are introduced, together with the performance measures and benchmark data widely used for evaluation.

As one of the major challenges in Chinese word segmentation, unknown word identification (UWI) is discussed separately in Chapter 5, probably because of its larger impact on accuracy and greater difficulty. Unknown words account for 60% of word segmentation errors. Proper nouns representing person, place, and organization names that are missing from the dictionary are common sources of unknown words. New names may appear on a daily basis, tending to intensify the OOV problem. This chapter describes various means for UWI from both generic and specific point of views. The general statistical methods based on co-occurrence statistics are depicted, followed by the techniques for identifying specific types of proper nouns based on their particular formation patterns. Two general methods worthy of review but missing are class-based language models (Fu and Luke 2004) and discriminative Markov models (Zhou 2005), both of which make significant improvements by leveraging various types of features.

Chapter 6 turns the reader's attention from word structure to meaning; it familiarizes them with the gist of the basic semantic concepts, relations, and resources, and helps build up knowledge for resolving deeper NLP problems such as word sense disambiguation, parsing, and language understanding. Three major contemporary Chinese thesauri, namely CILIN, HowNet, and CCD, are surveyed in comparison with the well-known English WordNet. For example, HowNet has three unique features compared to WordNet: (1) its concept definitions are based on sememes, which are specified in a structured mark-up language; (2) the semantic role relations can be connected across part-of-speech categories such as between nouns and verbs; and (3) the concepts are represented in both Chinese and English. Though structurally compatible, CCD differs from WordNet with some important extensions, such as more types of nouns, finer relations, and examples of collocations selected from real corpora with quantitative

descriptions. The three major Chinese linguistic resources at the lexical semantic level are valuable to Chinese NLP applications.

Chapters 7 and 8 concentrate on Chinese word collocation, which is a common lexical convention using the customary combination of words in expressions. Starting with the interesting example of *historical burden* as contrasted with *historical luggage* (*burden* and *luggage* can be expressed with the same Chinese word), Chapter 7 places emphasis on the concepts, distinguishing features, categorization, and data resources. Although widely used and easily comprehensible, collocation is defined in various ways and is therefore subject to controversy in the literature. After a review of English collocation, the authors accentuate the major differences between Chinese and English collocations, and then define Chinese collocation as a lexically restricted word combination with emphasis on semantic and syntactic considerations.

Chapter 8 elaborates on techniques for extracting Chinese collocations. The survey categorizes the existing methods into statistical, syntactic, semantic, and categorical approaches. Given a headword and a context window, a statistical approach identifies the word combinations from the context inclusive of the head word with significant lexical statistics as the collocations; a syntactic approach resorts to syntactic knowledge acquired from a parser to refine the candidates, assuming that collocation words are syntactically dependent; a semantic approach exploits semantic relations such as synonyms obtained from WordNet to overcome the inaccuracies caused by insufficient statistics; a categorical approach adopts a set of carefully tailored rules to identify collocations of different types, which may potentially achieve better results, but the method turns out to be less generalizable. This language-specific approach corresponds to the characteristics of Chinese collocations. The chapter would have been more helpful if it had included the work on collocativity measures based on limited modifiability with only shallow parsing (Wermter and Hahn 2004), statistical methods using accurate collocational information based on full parsing (Seretan and Wehrli 2006), and the more recent monolingual word alignment method (Liu et al. 2009).

The appendix provides a comprehensive and categorized list of linguistic resources. The references are generally complete and helpful to readers interested in further studies.

In summary, the book is clear and self-contained, discussing fundamental and critical issues in Chinese NLP. As the foundation of any NLP technology, morphological analysis (where Chinese morphology differs most apparently from other languages) is the first crucial procedure prior to syntactic and semantic processing. Although it would be more integrated to include syntactic and semantic analysis, I do believe the authors have made this choice wisely. Given the limited number of pages, it makes sense to focus on the essence with sufficient details to draw quick attention from many NLP readers who are sensitive to the most salient and distinctive subject matters. Instead of elaborating the Chinese part alone, the authors adopt a helpful writing style so that, whenever applicable, comparative discussions are made on the corresponding issues in other languages, especially English. This bilingual perspective sheds more light on what constitutes the noteworthy features in Chinese and the reason why customary techniques have to be or not be employed in Chinese NLP. The reader might have hoped for more polish and a better editorial quality in the book. For instance, a repeatedly referred-to concept OAS is not pre-defined (Section 4.4); not all the extracted co-words shown in Tables 8.1–8.6 are true collocations, which should have been boldface as stated in the text. In addition, one more chapter is needed to close the book by summarizing the book with some conclusions and discussions on further readings beyond the introduction at Chinese morphological level.

This book review was edited by Pierre Isabelle.

References

Fu, Guohong and Kang-Kwong Luke. 2004. Chinese unknown word identification using class-based LM. In *Proceedings of the First International Joint Conference on Natural Language Processing*, pages 704–713, Hainan Island.

Liu, Zhanyi, Haifeng Wang, Hua Wu, and Sheng Li. 2009. Collocation extraction using monolingual word alignment method. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 487–495, Singapore.

Seretan, Violeta and Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 953–960, Sydney.

Wermter, Joachim and Udo Hahn. 2004. Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 980–986, Geneva.

Zhou, GuoDong. 2005. A chunking strategy towards unknown word detection in Chinese word segmentation. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, pages 530–541, Jeju Island.

Min Zhang is a research scientist at Institute for Infocomm Research, Singapore. His research focuses on machine translation, information extraction, kernel-based statistical structure learning, and natural language processing. Min Zhang's address is 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore 138632; e-mail: mzhang@i2r.a-star.edu.sg.