Mitkov, Ruslan, ed. (2003) The Oxford Handbook of Computational Linguistics, Oxford University Press.

Announced at http://linguistlist.org/issues/14/14-567.html

Roland Stuckardt, Johann Wolfgang Goethe University Frankfurt am Main, Germany

DESCRIPTION OF THE BOOK

The edited collection under review, the Oxford Handbook of Computational Linguistics (OHCL), belongs to the Oxford Handbooks series, which, according to the publisher, aims at providing „*an authoritative and state-of-the-art survey of current thinking and research*“ in particular subject areas. According to the editor, the OHCL addresses university researchers, teachers, and students in the fields of Computational Linguistics, Computer Science and Linguistics, as well as professionals such as industrial researchers, executives, software engineers, and translators.

The book consists of thirty-eight chapters authored by fifty experts from all over the world, and a preface by Martin Kay with a brief description of the history of the discipline of Computational Linguistics (CL). Each chapter covers a particular topic of CL. Chapter lengths vary between eleven and twenty-nine pages, with the majority of chapters comprising between fifteen and twenty pages. Access to the individual chapters is facilitated by abstracts. Each chapter includes a local list of bibliographic references, hints at further reading, and pointers (in particular, URLs) to relevant resources (software, corpora etc.).

The OHCL is divided into three parts, which are intended to reflect a natural progression from theory to practice.

Part I, “*Fundamentals*”, considers the issues typically covered by an introductory text on general linguistics, corresponding to the various levels of linguistic abstraction: 1. Phonology, 2. Morphology, 3. Lexicography, 4. Syntax, 5. Semantics, 6. Discourse, and 7. Pragmatics and Dialogue. However, the topics are discussed from a computational perspective to develop an understanding of the specific problems to be solved before a respective software technology can be implemented. Two further chapters, 8. Formal Grammars and Languages, and 9. Complexity, provide relevant background material of mathematical linguistics and theoretical computer science, covering, in particular, the fundamentals of automata theory, formal languages, and computational complexity.

Part II, “*Processes, Methods, and Resources*”, takes the computation-oriented discussion of the linguistic and mathematical fundamentals in part I as the point of departure. The following detailed description of the basic stages of text and speech processing and the employed methods, resources, and formalisms goes one step ahead towards automatic natural language processing (NLP). Eight chapters deal with specific problems of processing written and spoken language: 10. Text Segmentation, 11. Part-of-Speech Tagging, 12. Parsing, 13. Word-Sense Disambiguation 14. Anaphora Resolution, 15. Natural Language Generation, 16. Speech Recognition, and 17. Text-to-Speech Synthesis. Another nine chapters describe the most important general methods, resources, and formalisms employed in these processing stages: 18. Finite-State Technology, 19. Statistical Methods, 20. Machine Learning, 21. Lexical Knowledge Acquisition, 22. Evaluation, 23. Sublanguages and Controlled Languages, 24. Corpus Linguistics, 25. Ontologies, and 26 Tree-Adjoining Grammars.

Part III, "*Applications*", focuses on the application of NLP technology for solving real world problems, proceeding from the description of the NLP base technology that is provided in part II. Part III comprises twelve chapters: 27. Machine Translation: General Overview, 28. Machine Translation: Latest Developments, 29. Information Retrieval, 30. Information Extraction, 31. Question Answering, 32. Text Summarization, 33. Term Extraction and Automatic Indexing, 34. Text Data Mining, 35. Natural Language Interaction, 36. Natural Language in Multimodal and Multimedia Systems, 37. Natural Language Processing in Computer-Assisted Language Learning, and 38. Multilingual On-Line Natural Language Processing.

In addition, the OHCL provides a list of commonly used acronyms and a glossary with brief definitions of about sixhundred key terms of CL and NLP. There are two general indexes: one by subject, and one by author/person.

CRITICAL EVALUATION

It would be beyond the scope of this review to provide a detailed evaluation of each of the thirty-eight chapters. Thus, the OHCL will be commented on at a general level. To some specific subjects it will be looked at in detail.

The above description clearly shows that the OHCL covers all main topics of CL. It deals with the whole range of text, speech, and dialogue processing, and discusses issues of text and speech analysis as well as generation. The OHCL is unique in that it bridges between the linguistic fundamentals (as provided in part I), the respective software base technology (as described in part II), and possible applications (as discussed in part III). As such it complements the recent textbook by Jurafsky and Martin (2000), which focuses on the algorithmic, mathematical, and engineering aspects of NLP. Another positive aspect is the inclusion of brief and quite comprehensive surveys of the most central methods, resources, and formalisms, such as machine learning, statistics, finite-state technology, and corpora.

The general organization of the handbook is excellent. In many cases, multi-authored volumes consist of collections of more or less loosely connected articles. Regarding the OHCL, editorial efforts concerning the overall coherence have been highly successful, resulting in self-contained chapters that provide a large number of useful cross-references that foster the connectivity of the material. Coherence and general accessibility are enhanced by the uniform style of presentation, which enables the reader to quickly access the relevant resources and further readings without skimming through the whole body of text.

Content quality and actuality of the individual chapters are generally high, fostered by the selection of authors that are world-leading experts with extensive research experience on the respective topics. However, the assignment of multiple authors also entails a certain variance in the formal (organizational and presentational) quality of the individual contributions and, more importantly, in the accessibility for different types of readers. It goes without saying that tight page limits always involve a trade-off between coverage on one and verbosity and readability on the other hand. Regarding the OHCL, with respect to both accessibility and formal quality, the difference between the individual articles is considerable, which may only partly be attributed to the varying degree of complexity of the issues covered.

Beginning with the positive side of the gamut, the majority of chapters is of high organizational and presentational quality, and accessible, with reasonable efforts, for

advanced students, scientists, and professionals with moderate previous knowledge of CL, NLP, or Linguistics. Among the many excellently written chapters are: 2. Morphology, 3. Lexicography, 7. Pragmatics and Dialogue, 10. Text Segmentation, 15. Natural Language Generation, 25. Ontologies, 27. Machine Translation: General Overview, 28. Machine Translation: Latest Developments, 29. Information Retrieval, 34. Text Data Mining, 35. Natural Language Interaction, 37. NLP in Computer-Assisted Language Learning.

However, there is also room for further improvement. In part I, "*Fundamentals*", for instance, the strong direction of chapter 2 (Morphology) towards computational issues is in line with the computational perspective to be assumed. Some other articles, however, are to a lesser extent, such as the overall excellent chapter 5 on computational semantics, which could be further enhanced by an assessment of the general computational feasibility of the construction of compositional-semantic descriptions, their expected coverage, and their potential contribution to robust NLP. More importantly, the presentation and organization of the material covered by chapters 8 (Formal Grammars and Languages) and 9 (Complexity) should be revised. While these articles are useful references for researchers with previous knowledge, their presentation may be in parts too dense for an audience unfamiliar with these rather mathematical issues. The author of the last-mentioned chapter himself admits: "*For a true understanding of complexity, it is best to read a serious algorithms book […].*" (page 196). Since the topics covered in these two chapters are highly related, it may be reasonable to integrate them into a single chapter of forty to fifty pages. To enhance the coherence of this rather formal subject matter with the rest of the book, the important discussion of its implications for practical issues of NLP may be expanded; the respective material, which is currently scattered about various subsections (e.g., 9.2.6 and 9.3.8), should be put into a dedicated section.

Regarding part II, "*Processes, Methods, and Resources*", parts of the generally well-written chapter 16 on speech recognition may be considered as quite difficult to access. Section 16.2 (Acoustic Parameterization and Modeling) refers to advanced technical notions and presupposes a quite high amount of previous knowledge. On the other hand, it misses out some fundamentals, e.g. the notion of formants, which play a central role in the analysis of waveforms, and which are not included in the glossary either (cf. Jurafsky and Martin (2000), which provides a more extensive discussion of this topic). However, the other sections of this chapter are readily accessible without expert knowledge; in particular, the discussion of the performance of state-of-the-art speech recognition technology is excellent.

Regarding part III, "*Applications"*, chapter 31 on question answering might be, to a certain extent, enhanced. While this article is logically structured and well organized, some parts of the presentation may be perceived as too dense, particularly section 31.7 on answer extraction. The details on training a perceptron, in particular the formula for the computation of a relative comparison score and the empirically determined weights and threshold values, are of little value to a reader interested in the fundamentals of answer extraction. Instead, more room should be given to the discussion of the principal ideas of answer extraction methods, perhaps focusing on the two most promising approaches and referring the reader to the literature for further technical details.

There is an additional topic further editions of the OHCL should include: discourse parsing. Fostered in large parts by the seminal work of Marcu (2000), this subject area has made rapid progress in the last few years. A chapter on algorithmic approaches to discourse parsing would neatly fit in the book and supplement the current edition's single article on discourse in part II (14. Anaphora Resolution). As such, it would provide a natural link between chapter 6

on the linguistic fundamentals of discourse and important applications discussed in part III, in particular text summarization (chapter 32, which explicitly refers to Marcu's work). Moreover, part III might be enhanced if chapters on text categorization and internet search engines were included.

There are some further minor content-related issues. Chapter 14 (Anaphora Resolution) should include a brief discussion of the intricacies of anaphor and ellipsis resolution in dialogue; this important issue is referred to in chapter 35 on natural language interaction. Chapter 20 on machine learning should include approaches of unsupervised learning, which receive increasing attention in NLP since supervised learning requires annotated corpora the gathering of which is, in general, expensive. Chapter 22 on the central topic of evaluation should be allotted more pages. Some practical examples of formal, corpus-based evaluation tasks would be helpful. Furthermore, the comparatively dense section 22.4 on the evaluation of interactive systems should include more details. Chapter 30 on information extraction should provide more pointers to concrete systems, including an overview of software technology available for different languages. Finally, chapter 35 on natural language interaction should include a discussion on the emerging dialogue modeling standard VoiceXML.

Two further marginal organizational issues: the table of contents should provide at least the section headings in addition to the chapter headings. Chapter 2 employs five different levels of section embedding; this is inadequate with an article of only 23 pages.

CONCLUDING REMARKS

In accordance with the intentions set out by the editor and the publisher, the OHCL provides a comprehensive high-quality survey of the theoretical fundamentals of CL and state-of-the-art NLP, as it covers the base technology, the underlying methods, and a wide range of application scenarios. Most chapters provide self-contained surveys of specific CL topics that are adequate reading to an audience with moderate previous knowledge, ranging from advanced students of CL, Linguistics, and Computer Science to scientists and industrial researchers. While it neither substitutes introductory textbooks nor monographs on the issues under discussion, it is an excellent reference book that provides a wealth of information and enables the experienced reader to quickly enter into new subject areas of CL and NLP. Thanks to its unique structure, it neatly complements books that focus on the technological aspects of NLP such as the text of Jurafsky and Martin (2000). Whereas these authors provide more thorough descriptions and illustrations of algorithms, the particular strengths of the OHCL are the comprehensive computation-oriented discussion of the fundamental linguistic issues and the broad coverage of NLP methods and resources. It thus extensively accounts for the theoretical and methodological backgrounds of CL and NLP.

A final note on pricing: the relatively high price of GBP 95.- (US$ 150.-) quoted by the publisher may put off many potential readers. The publisher should consider issuing a moderately priced student's edition to make the OHCL affordable to the wide audience it definitely deserves.

REFERENCES

Daniel Jurafsky and James H. Martin (2000). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, Upper Saddle River, NJ.

Daniel Marcu (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, Cambridge, MA.

ABOUT THE REVIEWER

Roland Stuckardt works as a postdoctoral researcher and consultant in the fields of Computational Linguistics and Natural Language Engineering. He received his Ph.D. in 2000 from the University of Frankfurt am Main. His current research interests include anaphor resolution, information extraction, text summarization, question answering, and spoken dialogue systems. (web: http://www.stuckardt.de/)