Book Reviews

The Oxford Handbook of Computational Linguistics

Ruslan Mitkov (editor)
(University of Wolverhampton)

Oxford: Oxford University Press, 2003,
xx+784 pp; hardbound, ISBN
0-19-823882-7, £76.00, $150.00

Reviewed by
Peter Jackson
Thomson Legal & Regulatory

This collection of invited papers covers a lot of ground in its nearly 800 pages, so any review of reasonable length will necessarily be selective. However, there are a number of features that make the book as a whole a comparatively easy and thoroughly rewarding read. Multiauthor compendia of this kind are often disjointed, with very little uniformity from chapter to chapter in terms of breadth, depth, and format. Such is not the case here. Breadth and depth of treatment are surprisingly consistent, with coherent formats that often include both a little history of the field and some thoughts about the future. The volume has a very logical structure in which the chapters flow and follow on from each other in an orderly fashion. There are also many cross-references between chapters, which allow the authors to build upon the foundation of one another's work and eliminate redundancies.

Specifically, the contents consist of 38 survey papers grouped into three parts: Fundamentals; Processes, Methods, and Resources; and Applications. Taken together, they provide both a comprehensive introduction to the field and a useful reference volume. In addition to the usual author and subject matter indices, there is a substantial glossary that students will find invaluable. Each chapter ends with a bibliography, together with tips for further reading and mention of other resources, such as conferences, workshops, and URLs.

Part I covers the full spectrum of linguistic levels of analysis from a largely theoretical point of view, including phonology, morphology, lexicography, syntax, semantics, discourse, and dialogue. The result is a layered approach to the subject matter that allows each new level to take the previous level for granted. However, the authors do not typically restrict themselves to linguistic theory. For example, Hanks's chapter on lexicography characterizes the deficiencies of both hand-built and corpus-based dictionaries, as well as discussing other practical problems, such as how to link meaning and use. The phonology and morphology chapters provide fine introductions to these topics, which tend to receive short shrift in many NLP and AI texts.

Part I ends with two chapters, one on formal grammars and one on complexity, which round out the computational aspect. This is an excellent pairing, with Martín-Vide's thorough treatment of regular and context-free languages leading into Carpenter's masterly survey of problem complexity and practical efficiency.

Part II is more task based, with a focus on such activities as text segmentation, part-of-speech tagging, parsing, word sense disambiguation, anaphora resolution, speech recognition, and text generation. (One might wonder why speech recognition is in

Part II rather than Part III, given that the chapter makes no reference to other areas that have appropriated these techniques and applied them elsewhere.) Some of these chapters make the obvious connections to topics in Part I, but others could have done more in this regard. However, there are many forward references to applications in Part III to which these techniques are pertinent.

The levels of treatment accorded to the topics in Part II are perhaps a little mixed, with some being less introductory than others. Thus Carroll's survey of parsing techniques provides an abstract overview for the researcher who already has a firm grasp of the computational issues and can appreciate the differences in parsing strategy and structural bookkeeping that he describes. Similarly, Karttunen's treatment of finite-state technology is more a summary of transducers for NLP than an introduction to the field. This is probably fine for a handbook of this kind, although it might limit the usefulness of these chapters for some students.

By contrast, Mikheev's chapter on segmentation and Voutilainen's chapter on POS tagging provide more background for the general reader and are comprehensible by nonexperts. Each of these chapters provides enough material to get a student started on the conception and planning phases of a segmentation or tagging project. Mitkov's exposition of anaphora resolution is particularly clear, being full of illuminating (and often entertaining) examples that highlight the distinctions between different kinds of anaphora, such as coreferring and non-coreferring. Kittredge's overview of sublanguages and controlled languages is also well organized and a model of clarity.

Middle chapters in Part II concentrate upon technologies to solve somewhat higher-level problems, such as natural language generation, speech recognition, and text-to-speech synthesis. Lamel and Gauvain's treatment of speech recognition is again an overview, rather than an introduction to the field, and is unlikely to be accessible to nonexperts. For example, the cepstral transformation and the Mel scale could be better motivated; neither is formally defined or linked to a glossary entry. Many students of NLP will not be familiar with these concepts and will not understand their importance in linear prediction and filterbank analyses using hidden Markov models.

These fairly specialized topics are then followed by useful chapters on subjects of interest to most computational linguists. Samuelsson's chapter on statistical methods does a very efficient job of imparting the basics of probability theory, hidden Markov models, and maximum-entropy models, together with a little dry humor. Mooney concentrates on the induction of symbolic representations of knowledge, such as rules and decision trees, in his chapter on machine learning. This focus avoids overlap with more statistical learning methods, such as naive Bayes, and allows room for covering case-based methods, such as nearest-neighbor algorithms.

Hirschman and Mani's chapter on evaluation represents a valiant attempt to cover, in a few pages, what remains a neglected topic in computational linguistics and natural language processing. It's a sad fact of life that gold standard–based approaches, such as those used in the Message Understanding and Text Retrieval Conferences, only take one so far in proving the effectiveness of research prototypes or final products. Measures such as precision and recall are useful yardsticks, but the real issue is, what value does the system deliver to an end user? More specifically, what does the system enable a knowledge worker to do that he or she could not do before? Academic researchers are typically not well placed to either pose or answer such questions, but any purveyor of natural language software must somehow address them. The section on evaluation of mature output components is the most relevant here.

McEnery provides an able introduction to corpus linguistics, albeit with a primary focus upon English, and briefly summarizes some of the advances that annotated corpora have enabled. Vossen's chapter on ontologies provides many useful pointers

to resources around the globe and makes an explicit attempt to outline areas of NLP in which such resources can and have been used. However, it is clear that the value of ontological approaches has yet to be fully demonstrated and that many of the tools are still in their infancy. Part II ends with a compact and readable overview of lexicalized tree-adjoining grammars by Joshi, which both motivates the formalism and illustrates its power.

Part III provides overviews of important areas such as machine translation, information retrieval, information extraction, question answering, and summarization. These chapters will be particularly attractive to practitioners in these fields, as they provide succinct and realistic overviews of what can and cannot be achieved by current technology. I confess to having read these chapters first. In fact, it might not be a bad strategy for some readers to dive straight into an application area in which they are particularly interested, and then read other chapters as needed, using the cross-references as a guide.

Machine translation is accorded two chapters, one that discusses the earlier, rule-based approaches and one that deals with more recent, empirical approaches based on parallel corpora. Both chapters give the general reader a good feel for the issues, the strengths and limitations of the various methods, and the kinds of tools that are currently available to assist translators. Somers's brief survey of statistical approaches to MT is particularly insightful on the topic of early successes and subsequent lack of improvement.

In the information retrieval chapter, Tzoukerman, Klavans, and Strzalkowski provide a frank assessment of how little impact natural language processing has had upon current search engine technology, beyond the application of tokenization and stemming rules. Whether attempting to apply WordNet to query expansion or seeking to disambiguate query terms, researchers have typically either failed to deliver improvements or failed to scale complex solutions to applications of commercial value. In attempting to elucidate why this is the case, the authors provide good coverage of the experimental literature and related research systems, such as CLARIT and DR-LINK. They conclude that NLP techniques to date have either been too weak to have a measurable impact or too expensive in terms of effort or computation to be cost-effective.

Grishman's information-extraction chapter provides a clear exposition of two problems: identifying proper names and recognizing events. Grishman provides an overview of the work done under the auspices of the Message Understanding Conferences in these areas, as well as an update on machine-learning approaches to the problem of building extraction patterns. Hearst's chapter on text data mining distinguishes this area from information retrieval and text categorization, linking the field to exploratory data analysis. Further chapters of interest include Hovy on summarization, André on multimodal and media systems, and Grefenstette and Segond on multilingual NLP.

In looking for gaps in the book as a whole, one cannot help noticing that the chapters on ontologies, word senses, and lexical knowledge acquisition (by Matsumoto) are among the few to touch upon semantic information processing. This is in marked contrast to many AI and NLP collections from the 1970s, in which articles on knowledge representation languages and text interpretation schemes abounded. Also absent are connectionist models of speech and language, which were perhaps more popular in the 1980s than they are today. These omissions may reflect a new realism in the field, in which the emphasis is now upon methods that are scalable, less knowledge intensive, and more amenable to empirical evaluation.

Overall, this is an impressive volume that demonstrates just how far the field has progressed in the last decade. During that time, we are fortunate to have seen many

advances in both the theory and practice of computational linguistics research, and one feels that these must be attended by improvements in natural language processing in the near future. When one combines the newer corpus-based approaches with continued advances in algorithms and representations in other areas, and then factors in annual increases in computing power and storage capability, one sees a recipe for further successes on hard problems like speech recognition, machine translation, and broad-coverage parsing.

Peter Jackson is vice-president of research and development at Thomson Legal & Regulatory, where he leads a group that specializes in the retrieval, mining, and classification of legal information. Over the last 20 years, he has published books and papers on expert systems, theorem proving, information extraction, and text categorization. His address is Thomson Legal & Regulatory, D1-N329, 610 Opperman Drive, St. Paul, MN 55123; e-mail: Peter.Jackson@Thomson.Com; URL: http://members.aol.com/jacksonpe/music1/home.htm.