Book Review

Handbook for Language Engineers

Ali Farghaly (editor)
(SYSTRAN Software Corporation)

Stanford, CA: CSLI Publications (CSLI lecture notes, number 164) (distributed by the University of Chicago Press), 2003, xi+442 pp; hardbound, ISBN 1-57586-395-2, $62.50; paperbound, ISBN 1-57586-396-0, $25.00, £17.50

Reviewed by
Ruslan Mitkov
University of Wolverhampton

The reader learns from this book's introduction (written by Ali Farghaly) that the objective of the book is to "equip linguists embarking on NLP assignments." The introduction also explains why language engineers are needed and summarizes the contents of each chapter.

"Domain Analysis and Representation" by Farghaly and Bruce Hedin offers a good discussion of the importance of domain analysis in natural language processing (NLP). It provides a helpful background on the notion of sublanguages and correctly notes that distinctions between the different domains can be blurred, as domains often overlap. The chapter would have been more convincing if there were further examples of NLP applications that benefit from narrowing down the domain of their operation. The chapter also discusses the analysis of domain into topics. Section 2.4.4 covers statistical approaches to classification, but no references or further reading pointers are given.

"The Language of the Internet" by Naomi Baron gives an easy-to-read and useful chronological overview of the developments on the Internet. It provides information about a number of technologies that are used on the Internet, comments on the changing styles of natural language use, and briefly overviews Web markup and programming languages and the Semantic Web. At the same time, I found this chapter too general. It would have been useful if the chapter had covered specific new NLP applications that are relevant to the Internet, such as question answering, and had covered in greater detail low-quality machine translation for e-mail and chat or text categorization of Web pages.

"Grammar Writing, Testing and Evaluation" by Miriam Butt and Tracy Holloway King provides a very good and accessible historical and linguistic account of grammars and parsing in NLP. It covers deep and shallow parsing and associated techniques as well as testing and evaluation of grammars and parsers. Morphological analyzers and part-of-speech taggers are briefly outlined too. A good practical point is the section on documentation of grammar writing.

"Ontologies" by Natalya Noy is a concise and plainly written introduction to the topic of ontologies and their development. It clarifies key terms in the ontology development jargon and includes an overview of major ontologies, ontology libraries, and

ontology development tools. All this information would be appreciated by anyone new to the field wishing to find the basic instruction material on the topic. Unfortunately, the discussion does not explicitly deal with how an ontology can be used in an NLP task, although ontologies are now frequently employed in a wide variety of tasks (e.g., information extraction, coreference resolution, word sense disambiguation). The use of ontologies for machine translation is mentioned but not really explained.

In "Text Mining, Corpus Building and Testing," Karine Megerdoomian covers important topics for a linguist embarking on NLP research. I found this chapter quite informative and useful, especially with regard to the use of corpora in computational linguistics. However, the problem is that too many topics have been included, and as a result, many of these topics are not properly covered. To start with, the title of the chapter is a bit misleading; the only discussion about text mining is Section 6.2.4, in a rather unusual context: between Section 6.2.3 "Corpus Analysis," and Section 6.3, "Tokenization." This two-page section does not sufficiently cover (or even mention) the whole variety of problem areas in text mining and practices for addressing them. Also, some of the important work on text mining is not mentioned, and a reference to Hearst (1999) at least would have been welcome. My view is that it would have been better to focus the chapter on corpora and corpus annotation and not to cover low-level text-processing tasks that could have been described in another chapter; or the reader could have been referred to textbooks or handbooks such as Manning and Schütze (2000) or Dale, Moisl, and Somers (2000). On the other hand, the use of XML in corpus annotation could have been described in greater detail. It is perfectly clear that not everything can be covered in a limited space, but there is no reference to ELDA or TRACTOR, even though the issue of languages with fewer resources is discussed. The "Linguist" list is suggested as a source of information about corpora, but not the "Corpora" list, which in my view is a better choice. Concordances, collocations, and frequency lists are discussed, but there is no reference to Wordsmith. Finally, in Section 6.5.4, GATE is called a ''graphic interface tool," which seems to me an oversimplification.

"Statistical Natural Language Processing" by Chris Callison-Burch and Miles Osborne contains a concise and informative introduction to the topic of machine learning and statistical approaches to NLP. It presents a good overview of the general procedure for building and testing a statistical model of a linguistic phenomenon. The topics discussed are a bit unbalanced: while it is quite detailed on basic notions (training vs. testing, baseline, evaluation measures, etc.), more advanced topics (maximum entropy, kernel methods, ensemble approaches) are discussed too briefly. In the end, the chapter contains a useful overview of application of statistical methods to various NLP tasks, in which they are contrasted with rule-based approaches, allowing the reader to form a clear idea of their strengths and weaknesses.

In "Knowledge Representation for Language Engineering," Matthew Stone focuses on "uses of knowledge representation that are more directly practical." The chapter is competently written, but I wonder if for linguists it would not have been more appropriate to discuss in greater detail traditional forms of linguistic knowledge employed in NLP systems, such as morphological, lexical, syntactic, semantic, and discourse information, rather than focusing on real-world and commonsense knowledge, given that the latter is rarely incorporated in practical NLP systems.

"Speech Recognition and Understanding" by Jan Amtrup is an accessible and useful introduction. To start with, different scenarios of speech applications are well illustrated. Then text-to-speech systems and speech synthesis are discussed. Following the section on speech recognition and continuing into the section on language modeling, various statistical approaches and techniques are introduced that might, however,

be too difficult for linguists to understand. Important issues such as prosody, dialogue systems, speech-to-speech translation, and speech resources are also covered. Given the range of tasks and applications presented, perhaps it would have been more accurate to title this chapter "Speech Processing."

The closing chapter, "Language Engineering and the Knowledge Economy," by Farghaly, argues that a shift is taking place in the industrialized countries from a product-based to a knowledge-based economy. Even though the author mentions that language engineering has a role to play in processing and assimilating massive amounts of language data, I found the presentation of little relevance to computational linguistics proper.

In my opinion, the coverage of topics is insufficient for a "handbook of language engineering." In addition, I would have expected a more practical orientation. It is understandable that every volume has space constraints, but practical applications in which linguists can and do play a role, such as text summarization, term extraction (and terminology processing in general), machine translation, and computer-assisted language learning, would have been welcome.

With regard to the editor's and publisher's roles, among the good points of the book is the fact that the reader will benefit from good cross-chapter referencing. Another good point is that the paperback version of the book is cheap and affordable ($25). Among the weak points of the book is the fact is that copyediting is far from perfect; for instance, the style of references is not identical across the chapters, and I also came across a number of spelling mistakes.

To conclude: the book could be of use to linguists embarking on research or projects in NLP. At the same time, some of the weaknesses outlined above will limit its overall benefit.

References

Dale, Robert, Hermann Moisl, and Harold Somers. 2000. *Handbook of Natural Language Processing*. Dekker.

Hearst, Marti. 1999. Untangling text data mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, College Park, MD, 3–10.

Manning, Christopher and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Ruslan Mitkov is Professor of Computational Linguistics and Language Engineering at the University of Wolverhampton. His publications and interests cover areas such as anaphora resolution, machine translation, text summarization, centering, term extraction, and generation of multiple-choice tests. He is author of the monograph *Anaphora Resolution* (Longman, 2002), editor of the *Oxford Handbook of Computational Linguistics* (Oxford University Press, 2003), and editor-in-chief of the book series *Natural Language Processing* (John Benjamins). Mitkov's address is School of Humanities, Languages and Social Sciences, University of Wolverhampton, Stafford St, Wolverhampton, WV1 1SB, United Kingdom; e-mail: R.Mitkov@wlv.ac.uk.