Handbook of Natural Language Processing (second edition)

Nitin Indurkhya and Fred J. Damerau (editors)
(University of New South Wales; IBM Thomas J. Watson Research Center)

Boca Raton, FL: CRC Press, 2010, xxxiii+678 pp; hardbound, ISBN 978-1-4200-8592-1, $99.95

Reviewed by
Jochen L. Leidner
Thomson Reuters Corp. and Linguit Ltd.

The *Handbook of Natural Language Processing* is a revised edition of an earlier handbook (Dale, Moisl, and Somers 2000). This second edition was prepared by Nitin Indurkhya, a researcher at the University of New South Wales, and the late text processing pioneer Fred J. Damerau of the IBM T. J. Watson Research Center (d. 27 January 2009), whose 1964 paper introduced a version of what is now known as the Damerau-Levenshtein distance, a metric of the similarity between two strings and a dynamic programming algorithm to compute it efficiently (Damerau 1964). Damerau also invented automatic hyphenation (Damerau 1970) and worked on early question-answering systems. Indurkhya, who is also affiliated with a consulting company, Data-Miner Pty Ltd., maintains a companion wiki for the book.[1]

The book has three parts, totaling 26 chapters. The first part, Classical Approaches, essentially covers techniques that were known prior to the statistical revolution, that is, before natural language processing people in the mainstream embraced techniques that speech engineers were already using successfully for awhile. The second part, Empirical and Statistical Approaches, covers state-of-the-art data-driven models.[2]

Part three, Applications, shows some techniques closer to applications. If you are talking to a computational linguist, information extraction is seen as an application, but if you are talking to business people, they will see it as a general technology area, from which many application products and services can be built. The handbook "aims to cater to the needs of NLP practitioners and language-engineering professionals in academia as well as in industry. . . . The prototypical reader is interested in the practical aspects of building NLP systems and may also be interested in working with languages other than English" (p. xxii). Hence it would have been nice to introduce some descriptions of actual products that generate revenue (even if this meant that this particular part of the handbook would become outdated more quickly) in order to demonstrate how the NLP parts are embedded in non-NLP technology, and how these products are embedded in the businesses that use them. For example, the application chapter "Information Retrieval" does not describe how the topics in other parts were applied in Web search engines or enterprise search products, as one might have expected. Rather, it basically is another technical chapter—and its probabilistic IR material could just as well have been presented in part two (statistical techniques).

1 `http://cgi.cse.unsw.edu.au/~handbookofnlp`.

2 As someone with a deep interest in methodology, this reviewer cannot resist pointing out that "empirical" is not an intrinsic property of any particular algorithm or technical approach, but a methodology, and as such not necessarily tied to statistical algorithms only. So the dichotomy of "classical" vs. "empirical and statistical" conflates technical approach (statistics vs. rule-based) and epistemological approach (empirical vs. rationalist).

© 2011 Association for Computational Linguistics

Each chapter describes one sub-field of natural language processing, and starts with its own chapter table of contents, which gives the reader a good idea of the structure of the chapter ahead. However, from a reference-work point of view, it would have been beneficial to have a consistent set of sub-headings for each chapter. For example, it seems arbitrary that the chapter on question answering has a section on evaluation, but the chapter on statistical parsing does not.

We can ask of a book, as of any other commercial product, "What is its competition?" First, an obvious class of competitors is textbooks such as Jurafsky and Martin (2008) and, to a lesser extent, Manning and Schütze (1999) (which covers only statistical techniques), at least one of which (and probably both) are already on the professional reader's bookshelf. A second class of direct competitor is other handbooks in the field, such as Mitkov (2003). Third, a more indirect form of competition is posed by the open access ACL Anthology and the ACM Digital Library (which is not open access, but many professionals and academics are members) in combination with Web search engines. We shall address these in turn here.

First, Jurafsky and Martin's book (2008) is a strong competitor to the handbook. For example, its description of the Viterbi algorithm is much closer to an actual implementation than the handbook's mathematical formulation. Second, the other NLP handbook, edited by Mitkov (2003), although now over half a decade old, has stood the test of time well, and it has a similar coverage (and page count). But Indurkhya and Damerau naturally offer coverage of much more recent literature. Third, the option of retrieving the original papers free of charge, perhaps using a downloadable *ACM Computing Surveys* article as a starting point, should be mentioned in a decade where the death of the tangible book is predicted by many. But not all topic areas in the handbook are covered by up-to-date surveys elsewhere. Due to the handbook's survey-like style—as a tendency, it prefers extensive citations of external papers to giving actual descriptions of the methods—using the handbook requires Web access in tandem as well.

Because the handbook stays on the theoretical side, this reviewer felt compelled to give examples of the kind of questions industry practitioners are confronted with: *Which NP chunkers for Spanish are available under a LGPL license? How do I index a text with named entity markup in Lucene? What's the state of the art in company name recognition for Brazilian Portuguese in terms of F_1-score? How can I get GATE's HashGazetteer to recognize multi-token names?* It appears that even after this book, there is still a market for a desk-side companion book addressing these and other practical questions.

A short review cannot possibly do justice to a 678-page volume covering a whole discipline, but it may be prudent to point out a few specific highlights and shortcomings. The information extraction chapter by Hobbs and Riloff contains an excellent discussion of the limits of state-of-the-art systems (Section 21.5, "How good is information extraction?"). Its Fig. 21.3 on page 526 shows that published academic information extraction systems have hit a quality ceiling of $F_1 = 60\%$ (on the MUC datasets 1991–1998). The real highlight of the book (with respect to its own goals as set out in the preface) is a part of the chapter on biomedical NLP by Cohen. Sections 25.5 ff. ("Getting up to speed"), which point newcomers to the key papers of the field, key tools used by most practitioners (MetaMap, MMTx, ABNER, OSCAR3, KeX, LingPipe), and the main corpora, datasets, and thesauri (PubMed, MEDLINE, GENIA, Entrez Gene/LocusLink, UMLS), come closest to addressing the professional audience aimed at by the handbook. The acknowledgments point at the likely reason for this chapter's high practical utility: It evolved from a tutorial given by the author.

The handbook is remarkably free of errors; the authors and editors did a careful job, and were supported by a team of technical reviewers of international acclaim.

Consequently, this reviewer only encountered minor typos such as "Finite-sate" (p. 57). It is also very carefully typeset, and printed on high-quality paper. But the sparse index does not contain entries for precision, recall, or *F*-score. The NLP expert might not need these, but if so then again he or she may not need many parts of this book; yet in an interdisciplinary field such as the processing of language by machine, it is good to err on the side of caution, as researchers and practitioners have a multitude of backgrounds. The understanding of evaluation metrics is also of paramount importance if the book is intended to be useful to managers (managers and business people are often surprised by the fact that NLP/IR systems do not return perfect results).

If you are a researcher or graduate student who wants an up-to-date and succinct 25-page starting point to get into a new sub-area, this handbook is for you. For the technical professional who is looking for all he or she needs to build a prototype system, speech and language textbooks are perhaps more useful because they contain more algorithm descriptions and examples. Overall, this handbook is a readable and high-quality set of up-to-date surveys of all major fields of natural language processing for anyone who does not already own another broad treatise of the field.

References

Dale, Robert, Herbert Moisl, and Harold Somers. 2000. *Handbook of Natural Language Processing* (first edition). Marcel Dekker, New York.

Damerau, Frederick J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Damerau, Frederick J. 1970. Automatic Hyphenation Scheme. United States Patent 3,537,076.

Jurafsky, Daniel and James H. Martin. 2008. *Speech and Language Processing* (second edition). Prentice Hall, Upper Saddle Hill, NJ.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

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

Jochen L. Leidner is a Senior Research Scientist with the Thomson Reuters Corporation and a co-founder and director of Edinburgh-based Linguit Limited. He holds a Ph.D. in Informatics from the University of Edinburgh and Master's degrees in Computer Speech, Text, and Internet Technologies from the University of Cambridge, and in Computational Linguistics, English, and Computer Science from the University of Erlangen-Nuremberg. His research areas include information extraction, spatial resolution/location-aware systems, question answering, mobile search, NLP/IR methodology, and software architecture for NLP/IR systems. Leidner's address is: Thomson Reuters Corporation, Research and Development, 610 Opperman Drive, St. Paul, MN 55123 USA; e-mail: `leidner@acm.org`.