Natural Language Processing: The PLNLP Approach

Karen Jensen, George E. Heidorn, and Stephen D. Richardson (editors)
(Microsoft Corporation)

Boston: Kluwer Academic Publishers
(The Kluwer International Series in
Engineering and Computer Science;
Natural Language Processing and
Machine Translation, edited by Jaime
Carbonell), 1992, xv + 324 pp.
Hardbound, ISBN 0-7923-9279-5, $80.00

Reviewed by
Paul S. Jacobs
GE Research and Development Center

This is the first and final collection of a noteworthy set of research papers, which is both historical and timely. PLNLP, the Programming Language for Natural Language Processing, which emerged at IBM's Thomas J. Watson Research Center in Yorktown Heights about 15 years ago, is ostensibly the glue that holds the volume's 22 chapters together. But it is as much the corporate umbrella of IBM and its massive, international presence that makes the book a whole, and to a large extent the evolution of the research described here parallels the changes in the parent company. Just as there will never be another monolithic mainframe monopoly, there may never be another corporate environment in which so many researchers can cultivate their ideas for so long with relatively little disruption or focus on profit.

While many of the papers are excellent in their own right, this historical perspective makes them more interesting, and certainly justifies taking another look at some of them. Almost all of the book has appeared in print elsewhere at one time or another (starting in 1982), although some in sources that are not widely available. The book includes revised or shortened versions of some of the papers, along with a lucid and well-written introduction, and one new unpublished paper (Chapter 2).

PLNLP (often pronounced "Penelope" or "Plenelope," the editors point out) is the basis for some sizable applied natural language projects, the best known of which is the Critique text-processing system, a grammar and style checker that IBM took into development. Within PLNLP, the IBM group developed an extensive English grammar called PEG (PLNLP English Grammar) and a text-critiquing system called EPISTLE, which later became Critique. By the time Critique (along with the book's editors) moved from Yorktown to an IBM development lab in Bethesda, a variety of simpler, relatively inexpensive text-critiquing competitors had come to market through small companies. The editors now work for the Research Division of Microsoft.

The book is balanced with respect to theory and practice, and covers a variety of topics, including syntax and its application to text-critiquing, using on-line dictionaries, machine translation, semantics, and sense disambiguation. The new research contribution, "Towards Transductive Linguistics" (Chapter 2) by Alexis Manaster Ramer, gives a theoretical perspective. While it is interesting and readable, it is a bit awkward, appearing to retrofit various theories, including Chomsky's, to the IBM work and contrasting the transductive model embodied in Critique with the traditional generative model of syntax. Presumably this analysis is meant to help tie the rest of the papers together. After all, in grammar checking or machine translation, a system must do

more than simply decide whether a sentence is grammatically correct: it must produce analyses that help to correct the input or to provide an accurate translation. This could easily have been explained without a tutorial on transducers, complete with Turing-machine state table.

The papers on PEG and Critique, about a third of the book, provide interesting re-reading. These are followed by several papers on parsing and disambiguating dictionary definitions and using on-line dictionaries. The IBM groups pioneered the use of on-line resources in NLP, represented by papers by Klavans, Chodorow, and Wacholder; Binot and Jensen; and Ravin. Two newer papers, by Montemagni and Vanderwende and by Vanderwende alone, show that the work has continued; the Vanderwende article on noun sequences is extracted from a forthcoming Ph.D. thesis.

The machine translation work appears in several different places in the book, apparently representing the editors' attempt to organize the work into traditional categories like syntax and logical form. This is one place where things get confusing. Several MT systems are described, covering languages from Chinese to Norwegian. While the paper by Diana Santos gives some perspective on the use of PEG in MT and the motivation for transfer-based approaches, MT is too complicated and idiosyncratic to make the relationships between the systems apparent. A bit more editorial commentary might have helped here.

Two papers on sense discrimination, one by Braden-Harder and one by Tsutsumi, are an excellent foil for the syntactically oriented research. Neither has been published in any well-circulated form. Braden-Harder's paper, "Sense Disambiguation Using Online Dictionaries," is a synopsis of her 1991 thesis that reads well and gives valuable insights into the problem of sense interpretation and the role of dictionaries. This paper also stands out (along with Braden-Harder's other contribution with Stephen Richardson on Critique) as one that actually tries to evaluate the accuracy of the system. In a book full of large but often incomplete systems, the reporting of results, even sketchy ones, is refreshing.

The book wraps up with two additional papers on semantics. The final chapter, "The Paragraph as a Semantic Unit" by Zadrozny and Jensen, appeared in *Computational Linguistics* in a longer form (called "Semantics of Paragraphs") in 1991. While it's more ethereal and somewhat detached from the other papers, spanning linguistics and philosophy, and presenting examples that have never really been automatically analyzed, it provides nice food for thought about the future of research in the field. Just as the preoccupation with syntax embodied in the early years of PLNLP may be a thing of the past, the emphasis on sentences as the principal unit of communication is losing out to larger and longer structures.

In summary, this is an interesting collection that is likely to stay fresh for some time. The PLNLP era, like the mainframe era, may be ending, but the lessons of the research should stay with us for some time. Although the collection may not age like a fine wine, it is good reading and will be worth consulting periodically.

Paul S. Jacobs is a Computer Scientist with the Information Technology Laboratory at the GE Research and Development Center. Since September 1985, he has led research in natural language processing, particularly in text interpretation. He is on the editorial board of *Computational Linguistics*. Jacobs's address is GE Research and Development Center, 1 River Road, Schenectady, NY 12301; e-mail: psjacobs@crd.ge.com.