Industrial Parsing of Software Manuals

Richard F. E. Sutcliffe, Heinz-Detlev Koch, and Annette McElligott (editors)
(University of Limerick, University of Heidelberg, and University of Limerick)

Amsterdam: Editions Rodopi
(Language and computers: Studies in
practical linguistics, edited by Jan Aarts
and Willem Meijs, volume 17), 1996,
xi+277 pp; paperbound, ISBN
90-420-0102-X, $28.00, Dfl 45.00

Reviewed by
John Carroll
University of Sussex

This book is a collection of articles written by research teams (eight from seven countries) that participated in the workshop Industrial Parsing of Software Manuals, held at the University of Limerick, Ireland, in 1995. However, unlike a typical proceedings volume, the book has a strong unifying theme: reporting the behavior and measuring the performance of a collection of parsing systems on a single text using the same evaluation criteria. The book also has a well-defined structure: all the articles have a standardized organization with the same section headings and tables. Both aspects are intended to facilitate direct comparison between the systems. Indeed, the book is targeted at the nonspecialist reader who wants a parser and knows the sort of result required from it, setting out to provide some initial answers to questions such as what the most appropriate algorithm would be, whether an existing parser could be adapted, whether its output could be converted to the form required, and what coverage and accuracy might be expected.

The text used for testing consisted of utterances extracted from Dynix, Lotus, and Trados software user manuals, such as: *For information, refer to "undoing one or more actions" in this chapter and Automatic Substitution of Interchangeable Elements.* Participants were asked to report the outcome of parsing the text, firstly with an unmodified system, then after relevant lexicon alterations, and finally, after grammar alterations as well. The articles report parser performance along four dimensions:

1. ability, in principle, to identify particular types of construction, ranging from recognition of verbs and nouns, through recognition of phrase boundaries, to attachment of prepositional phrases and analysis of co-ordination and gapping, amongst others;
2. coverage, expressed in terms of the percentage of utterances for which the parser is able to produce some analysis (whether correct or not);
3. efficiency, giving the time taken to analyze the utterances, specifying the type of machine used; and
4. accuracy, measuring the proportion of each type of construction (as in 1, above) that was identified correctly.

One unexpected outcome of the workshop was the great diversity in the types of output produced by the parsers (the appendices illustrate this by giving analyses

from all the systems for five of the utterances). This diversity made direct, quantitative evaluation of accuracy difficult, so the reported results are necessarily to some extent subjective. Thus, a second theme of the book—in the form of two chapters, one by Lin and one by Atwell, following the editorial introduction—concerns standardization of parse output and using this as an objective basis for evaluating parser accuracy.

Lin argues that the current widely used method of measuring parser accuracy—with respect to manually-annotated phrase boundaries in a test text (Grishman, Macleod, and Sterling 1992)—is flawed. He demonstrates that a high score for phrase boundary correctness does not guarantee that a reasonable semantic reading can be produced; conversely, many phrase boundary disagreements stem from systematic differences between parsing schemes that are well-justified within the context of their own theories. He then elaborates an earlier proposal of his (Lin 1995) for evaluation based on dependency structure annotation. Atwell, though, referring to the multiple layers of syntactic markup specified in the current European EAGLES (1996) guidelines, comments that in transforming constituency-based analyses into a dependency-based representation, certain kinds of grammatical information would be lost that might be important for further stages of processing, such as "logical" information (e.g. location of traces, or moved constituents). Atwell goes on to propose a common encoding format for parser output that would allow notational differences to be factored out, although it still does not form a basis for straightforward quantitative evaluation.

The rest of the book comprises eight chapters, one from each participating research team. The systems described (and the institution at which they were developed and from which the research team came, if the same) are these: ALICE (University of Manchester Institute of Science and Technolology), ENGCG (University of Helsinki), Sleator and Temperley's (1991) Link Parser, PRINCIPAR (University of Manitoba), a robust system constructed from the Alvey NL Tools (Briscoe et al. 1987), SEXTANT (Rank Xerox Research Centre, Grenoble), DESPAR (National University of Singapore), and TOSCA (University of Nijmegen). The approaches to parsing taken by these systems cover a wide spread, and include implementations of linguistically motivated phrase structure and principle-based theories, and systems based on categorial grammar, hand-crafted finite-state constraints, and extended hidden Markov models.

Surprisingly, given the obvious care with which the editors set up this enterprise so that the various systems could be compared, there is no concluding chapter discussing the strengths and weaknesses of the competing approaches and summarizing the results reported and lessons learned. Instead, the introductory chapter ends with a disappointing section less than a page long, offering a few generalized and anodyne remarks about the exercise as a whole. Except for the odd confused entry in the index, the book is in general well-produced, though the use of leading zeros in numbers in every table in the book (e.g., 00, 002.9, 098%) impairs readability and mars the otherwise good presentation.

Although the systems described are diverse, the book cannot be taken as a representative overview of the field of robust parsing, as one significant class of system is not represented: that of statistical constituency-based parsers trained on Treebanks (see, for example, Magerman [1995], Carroll and Briscoe [1996]; Charniak [1996]; Collins [1996]). This weakens the claim of the book to be a source of reliable answers for the nonspecialist about the state of the art. However, each chapter serves to summarize and exemplify a single approach, and as a whole I would recommend the book as an accessible and readable survey of a range of current parsing systems and techniques.

References

Briscoe, Ted, Claire Grover, Branimir Boguraev, and John Carroll. 1987. A formalism and environment for the development of a large grammar of English. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pages 703–708.

Carroll, John and Ted Briscoe. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. In *Proceedings of the 1st ACL SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 92–100.

Charniak, Eugene. 1996. Tree-bank grammars. Technical Report CS-96-02, Brown University, Department of Computer Science.

Collins, Michael. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

EAGLES, Text Corpora Working Group. 1996. *EAGLES preliminary recommendations for the syntactic annotation of corpora, EAG—TCWG—SASG1/P-B, March 1996.* http://www.ilc.pi.cnr.it/EAGLES96/home.html.

Grishman, Ralph, Catherine Macleod, and John Sterling. 1992. Evaluating parsing strategies using standardized parse files. In *Proceedings of the 3rd Association for Computational Linguistics Conference on Applied Natural Language Processing*, pages 156–161.

Lin, Dekang. 1995. A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1420–1425.

Magerman, David. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283.

Sleator, Daniel and Davy Temperley. 1991. Parsing English with a link grammar. Technical Report CMU-CS-91-196, Carnegie Mellon University, School of Computer Science.

John Carroll is a UK EPSRC Advanced Fellow at the University of Sussex, working on robust parsing of text and applications to real-world tasks. Carroll's address is: Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton BN1 9QH, UK; e-mail: john.carroll@cogs.susx.ac.uk