Current Issues in Parsing Technology

Masaru Tomita (editor)
(Carnegie Mellon University)

Boston: Kluwer Academic Publishers
(The Kluwer International Series in
Engineering and Computer Science:
Natural Language Processing and
Machine Translation), 1991,
xix+296 pp.
Hardbound, ISBN 0-7923-9131-4, $75.00,
£44.75, Dfl 160.00

Reviewed by
Robert J. Kuhns
Science Applications International Corporation

Current Issues in Parsing Technology, edited by Masaru Tomita, contains 18 revised versions from the 45 papers that were presented at the International Workshop on Parsing Technologies (IWPT-89) in August 1989. (Other papers from the conference relating to generalized LR parsing can be found in another book edited by Tomita, aptly titled *Generalized LR Parsing*.) Seventeen of the contributions in the book report on implementations or formal results (e.g., theorems and proofs) associated with parsing issues, while the first chapter is an introduction by the editor that sets the context and objective for the conference and the book. This introductory chapter also briefly summarizes each of the other papers. Before discussing the common goal and commenting on whether the reports satisfy that objective, a short sketch of each contribution is provided.

1. A Brief Overview of the Contributions

Only the briefest and most general description is possible here. The interested reader is referred to the book for particular findings and results of the works.

"The computational implementation of principle-based parsers" by Sandiway Fong and Robert C. Berwick (Chapter 2) describes a principle-based (Government-Binding) parser (Principle-Ordering Parser) and examines the effects of principle ordering with respect to efficiency considerations.

"Parsing with Lexicalized Tree Adjoining Grammar" by Yves Schabes and Aravind K. Joshi (Chapter 3) compares parsing performance with a lexicalized grammar (of which Lexicalized Tree Adjoining Grammars are instances) and several parsing strategies.

"Parsing with Discontinuous Phrase Structure Grammar" by Harry Bunt (Chapter 4) investigates the parsing of discontinuous structures using a new formalism called Discontinuous Phrase-Structure Grammars that are extensions of Generalized or Head Phrase Structure Grammars.

"Parsing with Categorical Grammar in predictive normal form" by Kent Wittenburg and Robert E. Wall (Chapter 5) is a collection of formal results concerning variants of predictive normal form of Categorical Grammars. It is shown that variants eliminate spurious ambiguity from the parsing problem.

"PREMO: Parsing by conspicuous lexical consumption" by Brian M. Slator and Yorick Wilks (Chapter 6) describes a knowledge-based preference semantics parser (PREMO) that extracts and incorporates semantic information from a dictionary in the parsing process.

"Parsing, word associations, and typical predicate-argument relations" by Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle (Chapter 7) reports on ways that collocational constraints can have a positive impact on syntactic processing without resorting to semantic interpretation.

"Parsing spoken language using Combinatory Grammars" by Mark Steedman (Chapter 8) examines relationships between intonational and syntactic structure within a Combinatory Grammar framework.

"A dependency-based parser for topic and focus" by Eva Hajičová (Chapter 9) discusses methods for parsing written "free word order" language. By employing suprasegmental features, the techniques extend to spoken utterances.

"A probabilistic parsing method for sentence disambiguation" by T. Fujisaki, F. Jelinek, J. Cocke, E. Black, and T. Nishino (Chapter 10) reports on a probabilistic modeling procedure with linguistic expertise that is capable of handling a number of language processing problems, notably, ambiguity. Results are provided for both English and Japanese.

"Towards a uniform formal framework for parsing" by Bernard Lang (Chapter 11) presents formalisms for syntactic structures and parsing in order to compare computational aspects of each. Syntactic frameworks are considered special cases of Horn clauses, whereas parsing is expressed as a pushdown automaton.

"A method for disjunctive constraint satisfaction" by John T. Maxwell III and Ronald M. Kaplan (Chapter 12) examines the problem of processing disjunctive specifications and provides an algorithm that uses locality conditions of language resulting in better average-time performance.

"Polynomial parsing of extensions of context-free grammars" by K. Vijay-Shanker and David J. Weir (Chapter 13) presents a general scheme for polynomial-time recognition for languages generated by formalisms that are extensions of context-free grammars; viz., Linear Indexed Grammar, Combinatory Categorial Grammar, and Tree Adjoining Grammar.

"Overview of parallel parsing strategies" by Anton Nijholt (Chapter 14) surveys parallel parsing strategies. In addition to a discussion of multiple serial parser configurations, this paper examines parallel versions of CYK and Earley's algorithms.

"Chart parsing for loosely coupled parallel systems" by Henry S. Thompson (Chapter 15) examines loosely coupled parallel systems in which the processors do not share memory. The computation/communication impact of this design for parallel parsing is explored.

"Parsing with connectionist networks" by Ajay N. Jain and Alex H. Waibel (Chapter 16) reports on a connectionist system that acquires a set of grammar rules based on a sample corpus of sentences. Speech processing is also discussed with respect to these connectionist techniques.

"A broad-coverage natural language analysis system" by Karen Jensen (Chapter 17) presents a robust, data-driven parser. It describes a system capable of performing syntactic and semantic analyses and paragraph modeling.

"Parsing 2-dimensional language" by Masaru Tomita (Chapter 18) introduces a pair of two-dimensional parsing algorithms based on Earley's and Generalized LR algorithms.

2. Discussion

We now return to the first chapter ("Why parsing technologies?"), where Tomita provides the objective of the conference and an insightful analysis of several issues governing processing "real" texts. (A related discussion can be found in Tomita [1988].) In introducing the purpose of the book, Tomita notes that while there have been many advances in linguistics and in natural language processing, there have been very few practical systems that utilize theories of language. In short, there is a gap between theory and applications.

To support the view that a gap exists, Tomita distinguishes between the linguistically "interesting" (e.g., ambiguity and movement) and the linguistically "uninteresting" constructions that are readily found in everyday language (e.g., date and time expressions and idioms). Since scientific investigations, including the study of language, rely on abstractions and idealizations of phenomena, gaps between models and the observables are necessary byproducts of doing science. Linguistic theory attempts to explain the structure of language with respect to sets of assumptions or principles, and data that are idiosyncratic or without principled underpinnings are eliminated from examination by the abstraction process. Each theory, therefore, defines what is or is not worthy of investigation, i.e., the phenomena that can or cannot be described and explained by the theory, respectively. Since "real" language is a combination of the linguistically rich and impoverished, a system that is solely grounded on a particular theory will certainly lack wide coverage and robustness. The system must then have some mechanism to handle those aspects of language left unaccounted for by the theory. To bridge this gap between theory and application, developers, as Tomita indicates, have designed and implemented various techniques, including "efficient parsing algorithms, software engineering for linguistic knowledge, implementation of linguistic theories/formalisms, and stochastic/probabilistic approaches" (p. 1), and the purpose of the results in this book is "to make contributions to fill the gap between theories and practical systems" (*ibid.*). "This research area, which I [Tomita] call *Parsing Technologies,* is essential to bring down novel linguistic theories into practical applications" (p. 3). The reports certainly do make significant contributions to parsing technologies, in that the systems and techniques can handle or address a wide set of complex language issues. There is also a diversity of approaches to language processing ranging from those that are theory-oriented (e.g., Fong and Berwick) to those that are primarily data-driven (e.g., Jensen). Works on parallel and connectionist implementations (e.g., Jain and Waibel, Nijholt, and Thompson) are represented as are probabilistic approaches (e.g., Church et al. and Fujisaki et al.). Thus, there is both depth and breadth to the works described.

The research also permits new questions to be asked concerning syntactic issues. For instance, Fong and Berwick have developed a system that makes possible the investigation of computational efficiency with respect to orderings of abstract principles. Another effort reported by Lang provides unifying frameworks for syntactic formalisms and parsing strategies, thereby allowing various syntactic phenomena to be compared in the same terms and divergent parsing approaches to be examined from a common perspective. Other papers presenting formal results (e.g., Maxwell and Kaplan, Vijay-Shanker and Weir, and Wittenburg and Wall) also provide novel insights. In short, no contribution lacks technical depth and value. Let us now consider the question of whether the goal of the book is satisfied by the various contributions.

The gap between theory and practical systems is wide, and to assess works that purport to link the two areas we need a clearer notion of what constitutes a "practical system." For the purposes of this review and to remain consistent with Tomita's view, a practical system is one that has the capability of processing "real text," i.e., unrestricted text. (Systems that are commercial successes or acceptable to a narrow class of trained users do not qualify as practical ones if their input is only of a limited or restricted nature.) In this context, the contributions fall short of the desired goal. The systems are either weighted toward theory with little or no regard to application or developed for broad use but without a theoretical basis. In other words, the efforts can be categorized as falling to one or the other side of the gap, but none can be said to have succeeded in uniting the two sides. If the papers contained in this volume are indicative of research in parsing technology, then the vast amount of effort is clearly oriented to theoretical issues rather than practical applications.

A system that would satisfy the seemingly incompatible goals of being theory-based and yet capable of processing real-world texts that contain linguistically "uninteresting" structures could be designed in several ways. For instance, such a system could link modules of a theory with parsing actions. (Barton [1984] is an early example of a discussion of the interaction of parser designs and linguistic theory.) These theory-based actions would be coupled or co-routined with a set of construction-specific rules to handle the peripheral ("uninteresting") aspects of language. A control module would invoke the different components as appropriate. The effect of this dual approach to parsing is that a combination of "interesting" and "uninteresting" phenomena can be processed, and as a result, the gap is bridged. (Kuhns [1990] provides a particular theory-based application and Reyle and Rohrer [1988] is a collection of linguistic-oriented systems primarily addressing research issues from a variety of frameworks.)

Despite what I believe is a shortcoming in that contributions fail to satisfy the desired goal of bridging theory and application, the book, nevertheless, is an excellent source for current parsing technology. Each paper is technically rich and presents some aspect of parsing in a clear, detailed discussion. The book could provide supplementary reading for advanced courses on parsing issues as well as reference material for researchers and developers working on formal syntax or parsing systems.

References

Barton, G. Edward, Jr. (1984). "Toward a principle-based parser." AI memo 788, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Kuhns, Robert J. (1990). "Automatic indexing and Government-Binding theory." *Proceedings, 13th International Conference on Computational Linguistics* (COLING-90). Helsinki, Finland, Vol. 3, 397–399.

Reyle, Uwe, and Rohrer, Christian (eds.) (1988). *Natural Language Parsing and Linguistic Theories.* (Studies in Linguistics and Philosophy 35). Dordrecht: D. Reidel.

Tomita, Masaru (1988). "'Linguistic' sentences and 'real' sentences." *Proceedings, 12th International Conference on Computational Linguistics (COLING-88).* Budapest, Hungary, 453.

Robert J. Kuhns is Chief Scientist at Science Applications International Corporation. His main research area is large-scale text processing applications involving theory-based parsing systems. His address is: SAIC, 3 Cambridge Center, Suite 201, Cambridge, MA 02142.