Generalized LR Parsing

Masaru Tomita (editor)
(Carnegie Mellon University)

Boston: Kluwer Academic Publishers,
1991, xii + 166 pp.
Hardbound, ISBN 0-7923-9201-9, $59.95,
£39.75, Dfl 135.00

Reviewed by
Giorgio Satta
University of Pennsylvania

Tomita's algorithm, also called the Generalized LR (GLR) parser, is a method for natural language parsing that extends standard parsing techniques for LR(k) grammars to cases of nondeterminism (see Tomita 1986). This algorithm has became quite popular within the computational linguistics community. At the first International Workshop on Parsing Technologies, held in Pittsburgh in 1989, several papers were presented reporting experimental work and theoretical results based upon Tomita's algorithm. These papers have been revised and collected in the book *Generalized LR Parsing* for the purpose of providing a reference for researchers having some interest in this algorithm specifically. I now discuss the contents of each contribution and conclude with some general remarks.

The opening chapter by Tomita and Ng introduces the Graph-Structured Stack (GSS), which is the basic idea in Tomita's algorithm for the simulation of nondeterminism. The authors then present a slightly modified version of the algorithm in (Tomita 1986) by discussing a sample computation on an ambiguous English sentence.

The next three chapters deal with performance issues, although from different perspectives. The contribution by Shann entitled "Experiments with GLR and chart parsing" reports on some experiments of average time evaluation of different parsing strategies implemented within a chart framework and an implementation of Tomita's parsing algorithm. The results can be summarized as follows: Tomita's algorithm performs best in cases of high ambiguity sentences, but shows the same performance as the left-corner method with top-down filtering in the case of a restricted domain corpus with low ambiguity sentences. Comparisons are made on the basis of net running time and the number of edges constructed by the different methods, with a shift operation in Tomita's algorithm counting as an edge construction. Both of those measures, however, allow comparisons only indirectly related to average-case time complexity. The first measure depends on how different methods are implemented. The latter is very weakly related to time, because an edge construction depends on a number of tests that can grow with the grammar and the input string length, and also since a reduction operation associated with a nonterminal shift cannot be considered an elementary operation as well. The reader should therefore be careful in the interpretation of the results, since these kinds of experiments require a uniform framework to be carried out (see for example Billot and Lang 1989). The reader should also be warned that the bottom-up and the bidirectional methods defined in this chapter involve polynomial computations of unbounded degree, but methods are found in the literature that perform much better, still using general context-free grammars.

The contribution by Johnson, entitled "The computational complexity of GLR parsing," discusses two issues regarding the worst-case computational complexity of

Tomita's algorithm. Johnson points out that a crude representation of a packed parse forest can lead the algorithm to use an amount of space exponential in the grammar size. This has to do with the fact that, for a given production, there are exponentially many different ways of matching immediate constituent boundaries within a long enough string. A similar point is made by Billot and Lang (1989), who give efficient solutions to this problem (based upon bilinear covers for the input grammar). The second point in this chapter is a demonstration that there exist context-free grammars G whose collection of LR(0) items is exponentially larger than $|G|$. Johnson exhibits input strings such that all these items are exploited by the algorithm, forcing exponential running time. Although these worst cases may not be relevant for natural language processing applications, it is interesting to note that nondeterministic LR automata are found in the literature that use a set of states always proportional to $|G|$—for example the LL/LR automaton proposed by Leermakers (1989); see also Schabes (1991) for computational complexity issues.

The contribution by Kipps entitled "GLR parsing in time $O(n^3)$" is also based on the above observation about the number of different matchings of immediate constituent boundaries. As a consequence, Kipps shows that in the worst case, an exponential amount of time with respect to the grammar length may be required by reduction operations in the original version of Tomita's algorithm. The author redesigns the *reduce* procedure of the algorithm using a tabular technique, thereby solving the problem in an efficient way. As a minor note, in the discussion of Earley's algorithm, Kipps attributes an overall amount of time $O(|w|^2)$ to the *predictor* operation, w being the input string. However, there are known implementations of such an operation that take $O(|w|)$ time (see for example Graham, Harrison, and Ruzzo 1980).

The contribution entitled "GLR parsing for ε-grammars" by Farshi focuses on a subclass of (noncyclic) context-free grammars with null productions that cannot be handled by the original version of Tomita's algorithm. This class has the following property: the number of null constituents preceding a symbol in a sentence of the language cannot in general be inferred using bounded lookahead. Farshi discusses a modification of Tomita's algorithm that repairs this shortcoming. The solution consists of extending the GSS to include cycles corresponding to possible parses of empty constituents. The author also discusses how to handle cyclic grammars, thereby extending the coverage of Tomita's algorithm to general-form context-free grammars. Note incidentally that, in discussing the importance of null productions, the author claims that a general context-free grammar can be exponentially more succinct than a context-free grammar in ε-free form; this is not true, and the size of the two grammars are related by a linear equation (see Sippu and Soisalon-Soininen 1988).

The contribution by Tanaka and Numazaki, entitled "Parallel GLR parsing based on logic programming," reports on an implementation of Tomita's algorithm within the framework of a concurrent logic programming language that is briefly introduced to the reader. The basic idea is to represent each entry in the LR table as a sequential process, and to run in parallel the processes corresponding to multiple actions within the same entry. Apart from the fact that the English of this chapter is rather clumsy, I find the presentation unconvincing for the following reasons: first, important details in the implementation of nondeterminism are missing. The reader is told that the operation of process splitting due to action conflicts involves a stack-copying operation and the GSS is replaced by a tree-structured stack, but the authors do not discuss the computational consequences of this. Furthermore, no comparison at all is offered to the reader with the existing literature on parallel context-free grammar parsing. The only work cited in the bibliography that describes a parallel parsing method is never mentioned in the chapter.

The two following chapters deal with the issue of stochastic parsing. The contribution entitled "GLR parsing with scoring" by Su, Wang, Su, and Chang, emphasizes the importance of scoring in best-solution–oriented parsing (in the context of a machine translation system, in this particular case) and focuses on syntactic scoring functions. The authors propose a general definition for the score of a derivation and adapt it to Tomita's algorithm by means of a decomposition in which each term corresponds to a transition between two shift actions. They also present a general truncation algorithm and provide an interesting discussion of its expected behavior as well as experimental results. One is left wondering how a GSS can be used in this framework: the authors do not deal with the problem of combining equal states having different scores in order to achieve further reduction of the search space.

The contribution by Wright and Wrigley, entitled "GLR parsing with probability," presents a theory for the construction of different kinds of stochastic LR tables from a stochastic context-free grammar. This allows the computation at parsing time of probabilistic distributions for next words, given the prefix of the input sentence analyzed so far. On the basis of the proposed framework, the authors discuss an application in uncertain input parsing. This chapter assumes considerable confidence with stochastic context-free grammars on the part of the reader. As a technical note, I observe that the authors compute the probability p_{BB} of all possible left-recursive derivations $B \stackrel{*}{\Rightarrow} Bw$, B a nonterminal and w a terminal string, as $\sum_w \Pr(B \stackrel{*}{\Rightarrow} Bw)$. But events $B \stackrel{*}{\Rightarrow} Bw$ are not mutually disjointed, and this summation is no longer a probability (such a quantity corresponds to quantity $Q_L(B \Rightarrow B)$ studied in Jelinek and Lafferty 1991). Probabilities p_{BB} for every nonterminal B can be correctly computed by solving a system of linear equations.

The last three chapters discuss applications of Tomita's algorithm to cases of corrupted input. The contribution by Malone and Felshin, entitled "GLR parsing for erroneous input," describes a system developed for use by language learners. The system is based on Tomita's algorithm, and is able to parse in the presence of ill-formed input of various kinds. Errors are grouped in different categories, and techniques to handle them are discussed along with the use of a scoring method. Among other things, the authors propose a standard lattice representation for the input sentence in order to deal with competing hypotheses deriving from typographical errors. This representation is well suited to the GSS, but any tabular parsing technique could have been used as well. In fact, I find the relationship of this paper to the rest of the book to be rather marginal: none of the proposed techniques depends at all upon the fact that the system is based on a nondeterministic shift-reduce automaton, and very little is said about LR parsing.

The contribution by Saito and Tomita in the next chapter, entitled "GLR parsing for noisy input," describes an application of Tomita's algorithm to a problem of error-correcting parsing within an automatic speech understanding system. A speech recognition device that produces a noisy phoneme sequence is coupled with Tomita's algorithm in an attempt to recover and parse the original sentence. The authors present a running example and briefly discuss the adopted scoring method and the pruning strategy. The reader is told that equal states having different scores are combined applying a Viterbi-like technique, but this important point is not discussed in any detail. Also, the proposal that is advanced is not related to the relevant literature on error-correcting parsing.

The closing chapter, by Kita, Kawabata, and Saito, is entitled "GLR parsing in a hidden Markov model." It discusses an interesting continuous speech-recognition system based on hidden Markov models (HMMs), which uses phone units and is driven

by Tomita's algorithm. The basic idea is to save each state reached in the parsing analysis along with a probability array obtained by the HMM probability calculation process (the row of the trellis corresponding to the final state of the HMM associated with the predicted phone). In this way acoustic recognition can be resumed. The algorithm performs a breadth-first search using a threshold. Unfortunately, because of the nature of HMMs, it is no longer possible to directly apply the state-combining operation, which is one of the basic ideas in Tomita's algorithm.

As seen so far, this book presents a considerable range of applications in which Tomita's algorithm has been employed showing noteworthy performance. Therefore the book succeeds in providing a convenient reference for the researcher who plans to use Tomita's algorithm in practical natural language systems. At the same time, this book discusses important improvements to the original specification of the algorithm, and I especially recommend the implementation that is suggested by Kipps.

This said, I find that the overall picture of Tomita's algorithm that emerges through this book is not entirely satisfactory. This is mainly due to the fact that, beyond reporting some original ideas, many contributions fail to offer an adequate comparison with well-known alternative methods, or to relate at all the proposal that is advanced with the standard literature (parallel parsing and error recovery contributions are the most evident cases). I find also that, as a result of putting together disparate works, an adequate discussion of some important issues that are often alluded to in the individual chapters is missing from this book. One case is the issue of control. The original specification of Tomita's algorithm employs a breadth-first strategy in the analysis of ambiguous input sentences. Since the GSS can be manipulated in a nondestructive way, the question arises of how to weaken the control in order to have a more flexible method (compare for example with chart-parsing techniques and the use of the agenda data structure [Kay 1980]). Although some of the contributions mention the use of a depth-first strategy, the problem is never discussed in any detail.

A second very important (and much debated) issue, alluded to throughout this book but never adequately discussed, is the one of average-case parsing efficiency. I share the belief expressed by the editor that, in cases of natural language grammars used by "practical" systems, Tomita's algorithm performs better than methods using nonprecompiled grammars. This might be true not only because of the precompilation of rule predictions, as pointed out by Kipps, but especially because of the achieved compression of the input grammar. (However, I should mention that I do not know of any experimental comparison between Tomita's algorithm and very efficient parsing algorithms that only use a mild form of precompilation, as for example the method presented by Graham, Harrison, and Ruzzo [1980].) Given that, which is the most successful way of precompiling an input grammar? There is evidence that this question should be answered case by case. For example, the precompilation adopted by the already mentioned LL/LR automaton repairs the worst cases studied by Johnson in this book, but there are many cases in which an LR table is favorable, resulting in a sublinear representation of the input grammar (see also Schabes 1991 for the proposal of a "mixed" precompilation technique). Furthermore, if the productions of the input grammar can be favorably factorized, then the CNLR automaton proposed by Leermakers (1989) provides an even more succinct representation. Since I did not find a satisfactory discussion of this important issue in this book, I refer the interested reader to Billot and Lang (1989) for a general presentation of the problem and some experimental results.

I should also add that this book has not been edited well: there are many typographical errors and, more astonishingly, unprocessed LaTeX commands are found in a couple of mathematical expressions. The purpose of the book aside, I conclude by

warning the reader that the use of GSS in the simulation of nondeterminism is only one particular approach to the problem of nondeterministic LR parsing. Once a precompilation of the input grammar is achieved, in the form of the transition map of a pushdown automaton or in the form of a context-free grammar, general dynamic programming or memoizing techniques have successfully been employed in the simulation of nondeterminism; see for instance Lang (1974) and Leermakers (1989, 1991).

Acknowledgments

I am very grateful to Bob Frank, Fernando Pereira, and Yves Schabes for helpful comments on an early draft of this review. Thanks also to Diego Giuliani.

References

Billot, Sylvie, and Lang, Bernard (1989). "The structure of shared forests in ambiguous parsing." In *Proceedings, 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, B.C. 143–151.

Graham, Susan L.; Harrison, Michael A.; and Ruzzo, Walter L. (1980). "An improved context-free recognizer." *ACM Transactions on Programming Languages and Systems*, **2**(3), 415–462.

Jelinek, Frederick, and Lafferty, John D. (1991). "Computation of the probability of initial substring generation by stochastic context-free grammars." *Computational Linguistics*, **17**(3), 315–324.

Kay, Martin (1980). "Algorithm schemata and data structures in syntactic processing." Technical report CSL-80, Xerox Palo Alto Research Center, Palo Alto, CA. Reprinted in *Readings in Natural Language Processing*, edited by Barbara J. Grosz, Karen Sparck Jones, and Bonnie L. Webber, 35–70. Morgan Kaufmann, 1986.

Lang, Bernard (1974). "Deterministic techniques for efficient non-deterministic parsers." In *Proceedings, Second Colloquium on Automata, Languages and Programming*, edited by J. Loeckx, Saarbrücken, Germany, 1974. Lecture Notes in Computer Science, Springer-Verlag, 255–269.

Leermakers, René (1989). "How to cover a grammar." In *Proceedings, 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, B.C., 135–142.

Leermakers, René (1991). "Non-deterministic recursive ascent parsing." *Proceedings, Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany, 63–68.

Schabes, Yves (1991). "Polynomial time and space shift-reduce parsing of arbitrary context-free grammars." In *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, 106–113.

Sippu, Seppo, and Soisalon-Soininen, Eljas (1988). *Parsing Theory: Languages and Parsing*, Volume 1. Springer-Verlag.

Tomita, Masaru (1986). *Efficient Parsing for Natural Language*. Kluwer Academic Publishers.

Giorgio Satta received a Ph.D. degree in Computer Science from the University of Padua, Italy. He is currently on a postdoctoral fellowship at the Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA 19104, conducting research in natural language parsing; e-mail: gsatta@unagi.cis.upenn.edu