Beyond Grammar: An Experience-based Theory of Language

Rens Bod
(University of Amsterdam)

Stanford: CSLI Publications (Lecture notes number 88), 1998, xiii+168 pp; distributed by Cambridge University Press; hardbound, ISBN 1-57586-151-8, $59.95; paperbound, ISBN 1-57586-150-X, $19.95

Reviewed by
Michael Collins
AT&T Labs–Research

1. Introduction

Over the past few years, Rens Bod and other researchers have investigated Data Oriented Parsing (DOP) approaches to statistical parsing. This book gives theoretical background, algorithms, and evaluation of DOP models.

So what is DOP? The book's initial definition (page 6) is as follows:

> In accordance with the general DOP architecture outlined by (Bod 1995b), a particular DOP model is described by specifying settings for the following four parameters: (1) a formal definition of a well-formed *representation for utterance-analyses*; (2) a definition of the *fragments* of the utterance-analyses that may be used as units in constructing an analysis of a new utterance; (3) a set of *composition operations* by which such fragments may be combined; and (4) a *probability* model that indicates how the probability of a new utterance analysis is computed on the basis of the fragments that combine to make it up.

Bod goes on to say:

> We hypothesize that human language processing can be modeled as a probabilistic process that operates on a corpus of representations of past language experiences, but we leave open how the utterance-analyses in the corpus are represented, how fragments of these utterance-analyses may be combined, and what the details of the probabilistic calculations are.

These definitions are perhaps too general to be useful; in fact, they are probably general enough to include all statistical parsing models in the literature. Fortunately, an earlier passage (page 5) gives a better idea of the flavor of the approaches in the book and in previous work by Bod:

> We should not constrain or predefine the productive units beforehand, but *take all, arbitrarily large fragments of (previously experienced) utterance-analyses as possible units and let the statistics decide.*

This philosophy is what really distinguishes DOP from other approaches. Given a corpus, *all* subtrees seen in that corpus, regardless of size, are taken to form a grammar—thus the models are sensitive to counts of fragments that vary in size from single

context-free rules to entire sentence-tree pairs. The DOP methods share a common method for estimating probabilities attached to these fragments, and Monte-Carlo style parsing algorithms to search for the most likely tree. For the rest of this review, I'll take the term "DOP" to refer to this narrower definition.

2. Content

Chapter 2 describes a first model, DOP1. The underlying grammar is a Tree Substitution grammar (TSG) (a restricted form of Tree Adjoining Grammar [Joshi 1987]); the grammar is a set of elementary trees, with substitution used to combine trees to give a derivation for a complete parse tree. The grammar is made up of all subtrees seen in a treebank of sentence-tree pairs. The key innovation of DOP is to remain relatively agnostic about the derivation underlying a tree in the corpus: the approach assumes that all TSG derivations could have produced the tree, and that the probability of a tree is calculated by summing over all derivations underlying the tree. The result is that counts of tree fragments of a wide range of sizes are considered: everything from counts of single-level rules (as in a Stochastic Context-Free Grammar) to counts of entire trees (where a tree-sentence pair is derived in a single step). Thus the model has the potential to be sensitive to the frequency of large tree fragments, while remaining relatively robust, thanks to the smoothing effects of counts of small fragments. Chapter 3 describes some formal results regarding the relationship between DOP1 and SCFGs, and a more qualitative comparison to other models in the literature.

Chapter 4 describes parsing algorithms for the DOP1 model. Efficient parsing is difficult, for a couple of reasons. First, the inclusion of all tree fragments leads to a very large grammar. Second, calculation of a tree's probability requires summation over an exponential number of derivations for a tree, rather than a (simpler) dynamic programming search for a single most likely derivation. Bod describes a relatively efficient Monte-Carlo–style method that samples derivations—given sufficient samples, the highest probability tree will, with high probability, be sampled most often.

Chapter 5 goes on to evaluate the model on the ATIS sentences in the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993). Most importantly, the impact of a number of restrictions on the model is tested: the effect of searching for the single most likely derivation, rather than summing over derivations to calculate the probability of a tree; the effect of imposing a varying limit on the number of lexical items in any elementary tree; the effect of limiting the depth of trees; the effect of excluding low-frequency trees; and finally, the effect of excluding trees that do not include head words. All of the results suggest that any restriction on the elementary trees included in the grammar results in a decrease in parsing performance.

The next few chapters extend DOP1 in various ways. Chapters 6 and 7 describe two new models—DOP2 and DOP3—that extend DOP1 to parse sentences with unknown words. Unfortunately the approaches have some problems: there is a further increase in grammar size, which causes the size of tree fragments to be limited for the sake of parsing efficiency; and the methods do not take into account the affixes or other spelling features of unknown words—these features are well known to be useful when dealing with unknown words (Weischedel et al. 1993). Chapter 8 extends the approach to corpora that include lambda-calculus–style semantics. Chapter 9 describes an application to the OVIS dialogue domain: the method uses an approach similar to DOP1, but extended to treat trees with semantic annotation; the approach is also extended to parse word lattices that summarize multiple possible outputs from a speech recognizer.

Finally, chapter 10—joint work with Ronald Kaplan—describes a model for Lexical Functional Grammar (LFG-DOP). The approach assumes a corpus of LFG analyses

(each analysis consists of a c-structure, an f-structure, and a mapping ϕ between them). The chapter concentrates on the problem of how to break LFG representations down into smaller units to form a grammar, how to compose these units within a derivation, and how to define probabilities associated with LFG structures.

3. Criticism

A concern with the approach is the efficiency of parsing algorithms for DOP models. The algorithm runs in time $N \times O(Gn^3)$, where N is the number of derivations sampled per sentence ($N = 100$ in the ATIS experiments[1]), G is the size of the grammar, and n is the length of the sentence. The main problem is that G can be very large, as it comprises all distinct subtrees seen in the corpus; Bod reports that the method takes more than 18 hours to parse 75 ATIS sentences. In the later experiments involving unknown words (DOP2 and DOP3) efficiency considerations mean that trees have to be restricted to at most depth 3, a restriction which is shown to be suboptimal in Chapter 5. In the OVIS domain, where the method parses word lattices, the Monte-Carlo approach is abandoned altogether; instead a Viterbi search for the most likely derivation is carried out—even though OVIS sentences are an average of only 4.6 words in length. These problems raise a general question of whether the approach can be scaled to larger domains—in work on the Penn *Wall Street Journal* Treebank, for example, sentences are significantly longer, and the grammar will be vastly larger. The grammar size will be strongly related to the number of training sentences, and approaches on *WSJ* have typically used around 40,000 training data sentences; Bod uses 675 sentences of training in the ATIS domain.

Perhaps due to efficiency problems, in the major sections on evaluation of the models—Chapters 5, 6, and 7—different variations of the model are evaluated and compared using a single test set of 75 sentences. This raises questions about the statistical significance of the results: in many cases different configurations give results that differ by a few percentage points in accuracy, corresponding to a difference of only two or three parse trees.

Given the problems with efficiency, what advantages does DOP offer? Bod argues throughout the book that counts of large substructures are important, culminating in the conclusion that

> we emphasize the most important outcome, namely that any systematic restriction of the fragments seems to jeopardize the statistical dependencies that are needed for predicting the appropriate structure of a sentence. ... If this outcome is generally true, it has important consequences for linguistic theory.

Unfortunately there may be a simpler explanation for the effects that Bod describes. Results in Chapter 5 do show that parsing accuracy increases with increasing fragment depth. But this gain may be due to the model becoming sensitive to the influence of lexical heads higher in the tree (for example, most dependencies between headwords require fragments of depth 3 before they are counted; many would require depths of 4 or more). In this case, approaches that instead extend the influence of lexical heads by

1 Goodman (1998, Chapter 4) gives convincing arguments that N is likely to increase exponentially with the length of the sentence n, further exacerbating efficiency problems. Similar arguments would suggest that N is also exponential in the grammar size G. Goodman also describes an algorithm equivalent to Bod's that runs in $N \times O(Gn^2)$ time.

annotating nonterminals with headwords capture these effects while retaining efficient parsing algorithms.

The book claims that other methods fail to capture the influence of nonheadwords, but the experiments fail to isolate the cases where the DOP approach differs from other approaches.[2] Unlike other approaches, DOP does capture dependencies between nonheadwords such as *nearby* and *to* in *nearby airports to Atlanta*, but the book does not give experiments isolating the contribution of these kinds of dependencies. (The important experiment, it seems, is to try the model with the elimination of all fragments containing two or more nonheadwords.)

As a final point, I would be negligent if I didn't warn the reader that there is a fair amount of bombast to wade through: from the preface of the book ("It has been argued that this outcome has important consequences for linguistic theory, leading to an entirely new view of the nature of linguistic competence"); through the introductory sections ("The resulting model also offers a new view of the nature of linguistic competence and the relationship between linguistic theory and models of performance"); to the perhaps overstated title of the conclusion ("Linguistics revisited"). The problems with parsing efficiency, and the limited evaluation of the approach, raise questions about the importance of the work within the statistical parsing literature: I found that the over-hyping of the work's relevance to linguistics in general quickly became irritating.

4. Conclusions

I would recommend this book to readers who are interested in statistical parsing—the DOP approach is interesting and original. The book's main value is in the thorough discussion of the models, experiments, and examples where DOP differs from other approaches. The reader should, however, be wary of the limitations of the approach and its evaluation. For somebody familiar with the statistical parsing literature, the book should be informative, if sometimes clearly off the mark. For a newcomer to statistical parsing, parts may be badly misleading. I would recommend strongly that Chapter 4 of Goodman (1998) be read in tandem with the book: it offers further experimentation, useful algorithms, and a rather more critical look at DOP models.

Acknowledgments

I would like to thank Steve Abney and Adwait Ratnaparkhi for helpful comments on an earlier draft of this review.

References

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, AAAI Press/MIT Press, Menlo Park.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, pages 16–23.

Goodman, Joshua. 1997. Probabilistic feature grammars. In *Proceedings of the Fifth International Workshop on Parsing Technologies (IWPT-97)*, Boston.

Goodman, Joshua. 1998. *Parsing Inside-Out*. Doctoral dissertation, Department of Computer Science, Harvard University.

Joshi, Aravind. 1987. Introduction to tree adjoining grammars. In *Mathematics of Language* (ed. Alexis Manaster-Ramer), John Benjamins, Amsterdam.

2 Contrary to Bod's arguments, methods such as those of Charniak (1997), Collins (1997), and Goodman (1997) do include counts of "intuitively silly subtrees with just one non-head word": they appear in the backed-off statistics. These counts are likely to be important, allowing the models to make the generalization, for example, that a particular preposition typically modifies noun or verb phrases regardless of the identity of the head noun or head verb.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330.

Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics* 19(2): 359–382.

Michael Collins is a member of the Machine Learning and Information Retrieval research group at AT&T Labs–Research. He recently completed his Ph.D. thesis at the University of Pennsylvania on *Head-Driven Statistical Models for Natural Language Parsing*. Collins's address is: AT&T Labs–Research, Rm A-253, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932; e-mail: mcollins@research.att.com