Reconciling Abstract Structure and Concrete Data in Statistical Natural-Language Processing

(Article begins on next page)

Stuart M. Shieber. Reconciling Abstract Structure and Concrete Data in Statistical Natural-Language Processing. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition*, Arden House, Harriman, New York, December, 1991, pages 57–59.

Reconciling Abstract Structure and Concrete Data in Statistical Natural-Language Processing

Stuart M. Shieber
Aiken Computation Laboratory
Harvard University
Cambridge, Massachusetts

In this paper, I present a research program for combining the most robust lessons learned from generative and statistical approaches to linguistics.[1] The most stable conclusion from generative linguistics, and perhaps the only truly uncontroversial discovery of the field, is that utterances of natural language have hierarchical structure. The observation dates back at least twenty-two centuries to Panini. It is most crisply formalized in the notation of context-free grammars. The field of statistical modeling of natural language well predates its modern era, which arguably starts with Shannon's study of the entropy of English. In this brief tenure, the lesson that has been learned most forcefully is the primacy of the primitives. Statistics are most useful when they involve directly the primitive parts of the utterance. In speech recognition, the phoneme may be primitive, in text-based modeling, the word. This lesson is exemplified by the importance of Markov models in statistical natural-language processing.

Unfortunately, these two lessons are at odds with each other. Hierarchical structure calls for definition of abstract classes of natural-language expressions, but statistical modeling is best applied to the concrete data. Yet the feeling that the two approaches must be combined is increasingly prevalent. The concreteness of corpus-based statistical natural-language processing leads to a robustness that grammatical methods cannot match, while modeling of the abstract structure of language is necessary for specifying a notion of context appropriate for language disambiguation tasks.

A statistical natural-language model, then, would ideally respect the hierarchical nature of language and the lexical sensitivity of parameters. Models based on n-grams, though lexically sensitive, are not hierarchical. Conversely, probabilistic context-free grammars are hierarchical but not appropriately lexically sensitive, because context-free grammars are not a *lexicalized* formalism, as defined by Schabes (1991a). The simplest formalism that can serve as the basis for lexicalizing a context-free grammar — that is, for manifesting both its hierarchical structure and its lexical basis — is the formalism of lexicalized tree-adjoining grammars. Thus, probabilistic lexicalized tree-adjoining grammars (PLTAG) fall out naturally in an attempt to preserve the best aspects of grammatical and statistical approaches to natural-language processing.

This observation concerning the utility of tree-adjoining grammars in combining grammatical and statistical approaches was made independently by Schabes (1991c) and Resnik (1991). In order to make use of PLTAGs for statistical natural-language processing, two key issues must be resolved: First, a precise definition of PLTAGs must be given, such that the parameters of a PLTAG specify cooccurrence relationships that accord with the structure of the language. Second, algorithms must be defined that

[1] This research program is being pursued by myself and others, especially Yves Schabes, Aravind Joshi, and Fernando Pereira. The original intuition behind forming a probabilistic variant of TAGs seems to have been due to Joshi. Schabes and Philip Resnik independently noted the motivation for lexicalized TAGs as an infrastructure for statistical language modeling.

allow PLTAGs to serve as the basis for statistical natural-language processing. In particular, efficient methods for determining the probability of a string as defined by a PLTAG and for inducing the parameters of a PLTAG from a corpus must be provided. The remainder of this paper describes the current work on these two issues.

Definition of PLTAGs

Schabes (1991c) has defined several variants of PLTAG, the most promising of which provides a probability distribution for each node in an elementary tree specifying, for each auxiliary tree, a probability for its adjunction at that node. (In addition, a probability for no adjunction is given as well.) This model is most simply presented by associating probabilities with each production in a linear indexed grammar (LIG) that is derived from the TAG.

The utility of the model depends crucially on the appropriateness of the parameters of the PLTAG model. The parameters in a PLTAG are intended to specify the cooccurrence relationships of lexical items of the language in a way that accords with the structure of the language. This contrasts with Markov models, which model cooccurrence but ignore hierarchical linguistic structure. However, under current definitions of TAG, certain natural linguistic analyses violate the intuitive notion of appropriate cooccurrence relationships. For instance, consider the analysis of a sentence with multiple verb phrase modifiers such as "I drove from Cambridge to Harriman." Intuitively, the most significant cooccurrence relationships hold between the verb "drove" and the heads of the two adverbials "from" and "to".[2] The relationship between the two adverbial heads is of secondary interest. Under standard definitions of TAG derivation, however, the elementary trees for "from" and "to" are not both adjoined into the tree for "drove"; such a derivation would be noncanonical (as defined by Vijay-Shanker (1987)), hence invalid. Rather, the canonical derivation would have the "to" tree adjoined at the root of the "from" tree, which in turn is adjoined into the tree for "drove". Thus, the lexical relationships directly manifested in the grammar's parameters would be the relationships between "drove" and "from" and between "from" and "to". TAGs model modification relationships with adjunction. A modification relation holds between an auxiliary tree and the elementary tree that it is eventually adjoined to. Unfortunately, TAG derivations do not always show this relationship directly, manifesting instead an artificial relationship in cascaded adjunctions.

In recent work, Schabes and I have defined an alternative notion of derivation for TAGs that allows multiple modifier trees to adjoin at the same node in a tree. Predicative auxiliary trees retain their cascaded behavior. A method for compiling TAGs into LIGs can be specified that characterizes the updated notion of derivation.

Algorithms for PLTAGs

The typical method for computing the probability of a string given a grammar is to piggyback the computation on a dynamic programming recognition method for the nonstochastic variant of the grammar formalism. This is the principle behind the calculation for Markov models (based on finite-state recognition) and probabilistic context-free grammars (based on the Cocke-Kasami-Younger algorithm). Schabes (1991a) has provided several recognition algorithms for TAGs including an extension of Earley's algorithm that applies to the LIG equivalent of a TAG. Furthermore, he has generalized the CKY-based algorithms for probabilistic Chomsky-normal-form context-free grammars to arbitrary CFGs by using Earley's algorithm as the basis for the algorithms (Schabes, 1991b). Combining these two results yields al-

[2] Intuition is an appropriate guide here, as the idea is to set up a linguistically plausible framework on top of which a lexically-based statistical model can be built. In addition, suggestive (though certainly not conclusive) evidence along these lines can be gleaned from corpora analyses. For instance, in an experiment in which medium frequency triples of the form "⟨*adjective*⟩ ⟨*adjective*⟩ ⟨*noun*⟩" were examined, the mean mutual information between the first adjective and the noun was larger than that between the two adjectives. The statistical assumptions behind the experiment do not allow very robust conclusions to be drawn, and more work is needed along these lines.

gorithms for arbitrary probabilistic TAGs (Schabes, 1991c). Recognition with respect to the revised notion of derivation mentioned above is also possible. We have devised an Earley-style algorithm for the LIGs resulting from the updated compilation mentioned in the previous section. Such an algorithm can also be used to base algorithms for PLTAGs on, but with the added benefit that the statistical parameters for adjunction of modifiers are more appropriately specified.

Conclusion

The lessons of grammatical and statistical linguistics argue for formalisms that are hierarchical yet lexical in character. Lexicalized tree-adjoining grammars hold a unique place in the convergence of these two criteria, in that they are the weakest formalism that is both hierarchically structured and lexically sensitive. The formalism can be made probabilistic in a natural fashion, and algorithms have been devised for using it as the basis for statistical natural-language-processing tasks.

Acknowledgements

The research presented in this paper was supported in part by Presidential Young Investigator award IRI-91-57996 from the National Science Foundation. I wish to thank Aravind Joshi for his support of the research and Yves Schabes and Fernando Pereira for their collaborative aid on this project. I am indebted to David Yarowsky for aid in the design of the experiment mentioned in footnote 2 and for its execution.

References

Resnik, Philip. 1991. Lexicalized tree-adjoining grammar for distributional analysis. In *Penn review of linguistics*.

Schabes, Yves. 1991a. *Computational and Mathematical Studies of Lexicalized Grammars*. Manuscript in preparation based on the author's PhD dissertation (University of Pennsylvania, August 1990).

———. 1991b. An inside-outside algorithm for estimating the parameters of a hidden stochastic context-free grammar based on Earley's algorithm. Manuscript.

———. 1991c. Stochastic lexicalized tree-adjoining grammars. Submitted for publication.

Vijay-Shanker, K. 1987. A study of Tree Adjoining Grammars. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.