Alternative Approaches for Generating Bodies of Grammar Rules*

Extendend Abstract

Gabriel Infante-Lopez Maarten de Rijke
Informatics Institute, University of Amsterdam
{infante,mdr}@science.uva.nl

N-grams have had a great impact on the state of the art in natural language parsing. They are central to many parsing models [2, 3, 6, 4], and despite their simplicity n-gram models have been amazingly successful. Modeling with n-grams can be viewed as an induction task. Given a sample set of strings, the task is to guess the grammar that produced that sample test. Grammar induction is a problem that consists of two parts: choosing the class of languages amongst which to search and designing the procedure for performing the search. By using n-grams for grammar induction one addresses the two parts in one go. In particular, the use of n-grams implies that the solution will be searched for in the class of probabilistic regular languages. However, the class of probabilistic regular languages induced using n-grams is a proper subclass of the class of all probabilistic regular languages.

Besides N-grams, there is a variety of general methods capable of inducing *all* regular languages [5, 1, 7]. Their relevance for natural language parsing is that regular languages are used for describing the bodies of rules in a grammar. Consequently, the quality and expressive power of the resulting grammar is tied to the quality and expressive power of the regular languages used to describe them. And the quality and expressive power of the latter, in turn, are influenced directly by the method used to induced them. These observations give rise to a natural question: can we gain anything in parsing from using general methods for inducing regular languages instead of methods based on n-grams? Specifically, can we de-

*The full version of this paper appeared in Proc. of the ACL04, pages 452–462. 2004 scribe the bodies of grammatical rules more accurately and more concisely by using general methods for inducing regular languages?

In our paper, our main question is aimed at understanding how different algorithms for inducing regular languages impact the parsing performance with those grammars. A second issue that we explore is how the grammars perform when the quality of the training material is improved, that is, when the training material is separated into part of speech (POS) categories before the regular language learning algorithms are run.

Our experiments support two kinds of conclusions. First, they suggest that modeling rules with algorithms other than n-grams not only produces smaller grammars but also better performing ones. Second, the procedure used for optimizing the automata reveals that some POS behave almost deterministically for selecting their arguments, while others do not. These findings suggests that splitting classes that behave non-deterministically into homogeneous ones could improve the quality of the inferred automata. We saw that lexicalization and head-annotation seem to attack this problem. Obvious questions for future work arise: Are these two techniques the best way to split non-homogeneous classes into homogeneous ones? Is there an optimal splitting?

References

[1] R. Carrasco and J. Oncina. Learning stochastic regular grammars by means of state merging method. In *ICGI'94*, LNAI, pages 139–150. 1994.

[2] E. Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proc. of the* 14^{th} *National Conference on Artificial Intelligence*, pages 598–603, Menlo Park, 1997. AAAI Press/MIT Press.

[3] M. Collins. Three generative, lexicalized models for statistical parsing. In *Proc. ACL97/EACL97*, pages 16–23, Spain, 1997.

[4] M. Collins. Discriminative reranking for natural language parsing. In *ICML-2000,Stanford, Ca.*, 2000.

[5] François Denis. Learning regular languages from simple positive examples. *Machine Learning*, 44(1/2):37–66, 2001.

[6] J. Eisner. Three new probabilistic models for dependency parsing: An exploration. In *Proc. of COLING-96*, pages 340–245, Denmark, 1996.

[7] F. Thollard, P. Dupont, and C. de la Higuera. Probabilistic DFA inference using Kullback-Leibler divergence and minimality. In *ICML 2000*, Stanford, Ca., 2000.