Journal of Automata, Languages and Combinatorics **10** (2005) 2/3, 367–388

A CONSISTENT AND EFFICIENT ESTIMATOR FOR DATA-ORIENTED PARSING [1]

ANDREAS ZOLLMANN
School of Computer Science
Carnegie Mellon University, U.S.A.
e-mail: zollmann@cs.cmu.edu

and

KHALIL SIMA'AN
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
e-mail: simaan@science.uva.nl

ABSTRACT

Given a sequence of samples from an unknown probability distribution, a statistical estimator aims at providing an approximate guess of the distribution by utilizing statistics from the samples. One crucial property of a 'good' estimator is that its guess approaches the unknown distribution as the sample sequence grows large. This property is called *consistency*.

This paper concerns estimators for natural language parsing under the *Data-Oriented Parsing* (DOP) model. The DOP model specifies how a probabilistic grammar is acquired from statistics over a given training treebank, a corpus of sentence-parse pairs. Recently, Johnson [15] showed that the DOP estimator (called DOP1) is biased and inconsistent. A second relevant problem with DOP1 is that it suffers from an overwhelming computational inefficiency.

This paper presents the first (nontrivial) consistent estimator for the DOP model. The new estimator is based on a combination of held-out estimation and a bias toward parsing with shorter derivations. To justify the need for a biased estimator in the case of DOP, we prove that every non-overfitting DOP estimator is statistically biased. Our choice for the bias toward shorter derivations is justified by empirical experience, mathematical convenience and efficiency considerations. In support of our theoretical results of consistency and computational efficiency, we also report experimental results with the new estimator.

Keywords: Statistical parsing, data-oriented parsing, consistent estimator

1. Motivation

A formal grammar describes a set of sentence-analysis pairs, where the analysis is a syntactic construct, often graphically represented as a tree. A major problem with

[1] Full version of a submission presented at the Workshop on *Weighted Automata: Theory and Applications* (Dresden University of Technology, Germany, June 1–5, 2004).