On bootstrapping of linguistic features for bootstrapping grammars

Damir Ćavar
University of Zadar
Zadar, Croatia
dcavar@unizd.hr

Abstract

We discuss a cue-based grammar induction approach based on a parallel theory of grammar. Our model is based on the hypotheses of interdependency between linguistic levels (of representation) and inductability of specific structural properties at a particular level, with consequences for the induction of structural properties at other linguistic levels. We present the results of three different cue-learning experiments and settings, covering the induction of phonological, morphological, and syntactic properties, and discuss potential consequences for our general grammar induction model.[1]

1 Introduction

We assume that individual linguistic levels of natural languages differ with respect to their formal complexity. In particular, the assumption is that structural properties of linguistic levels like phonology or morphology can be characterized fully by Regular grammars, and if not, at least a large subset can. Structural properties of natural language syntax on the other hand might be characterized by Mildly context-free grammars (Joshi et al., 1991), where at least a large subset could be characterized by Regular and Context-free grammars.[2]

Ignoring for the time being extra-linguistic conditions and cues for linguistic properties, and independent of the complexity of specific linguistic levels for particular languages, we assume that specific properties at one particular linguistic level correlate with properties at another level. In natural languages certain phonological processes might be triggered at morphological boundaries only, e.g. (Chomsky and Halle, 1968), or prosodic properties correlate with syntactic phrase boundaries and semantic properties, e.g. (Inkelas and Zec, 1990). Similarly, lexical properties, as for example stress patterns and morphological structure tend to be specific to certain word types (e.g. substantives, but not function words). i.e. correlate with the lexical morpho-syntactic properties used in grammars of syntax. Other more informal correlations that are discussed in linguistics, that rather lack a formal model or explanation, are for example the relation between morphological richness and the freedom of word order in syntax.

Thus, it seems that specific regularities and grammatical properties at one linguistic level might provide cues for structural properties at another level. We expect such correlations to be language specific, given that languages qualitatively significantly differ at least at the phonetic, phonological and morphological level, and at least quantitatively also at the syntactic level.

Thus in our model of grammar induction, we favor the view expressed e.g. in (Frank, 2000) that complex grammars are bootstrapped (or grow) from less complex grammars. On the other hand, the intuition that structural or inherent properties at different linguistic levels correlate, i.e. they seem to be used as cues in processing and acquisition, might require a parallel model of language learning or grammar induction, as for example suggested in (Jackendoff, 1996) or the Competition Model (MacWhinney and Bates, 1989).

In general, we start with the observation that

[1] This article is builds on joint work and articles with K. Elghamri, J. Herring, T. Ikuta, P. Rodrigues, G. Schrementi and colleagues at the Institute of Croatian Language and Linguistics and the University of Zadar. The research activities were partially funded by several grants over a couple of years, at Indiana University and from the Croatian Ministry of Science, Education and Sports of the Republic of Croatia.

[2] We are abstracting away from concrete linguistic models and theories, and their particular complexity, as discussed e.g. in (Ristad, 1990) or (Tesar and Smolensky, 2000).

Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference, pages 5–6,
Athens, Greece, 30 March 2009. ©2009 Association for Computational Linguistics

natural languages are learnable. In principle, the study of how this might be modeled, and what the minimal assumptions about the grammar properties and the induction algorithm could be, could start top-down, by assuming maximal knowledge of the target grammar, and subsequently eliminating elements that are obviously learnable in an unsupervised way, or fall out as side-effects. Alternatively, a bottom-up approach could start with the question about how much supervision has to be added to an unsupervised model in order to converge to a concise grammar.

Here we favor the bottom-up approach, and ask how simple properties of grammar can be learned in an unsupervised way, and how cues could be identified that allow for the induction of higher level properties of the target grammar, or other linguistic levels, by for example favoring some structural hypotheses over others.

In this article we will discuss in detail several experiments of morphological cue induction for lexical classification (Ćavar et al., 2004a) and (Ćavar et al., 2004b) using Vector Space Models for category induction and subsequent rule formation. Furthermore, we discuss structural cohesion measured via Entropy-based statistics on the basis of distributional properties for unsupervised syntactic structure induction (Ćavar et al., 2004c) from raw text, and compare the results with syntactic corpora like the Penn Treebank. We expand these results with recent experiments in the domain of unsupervised induction of phonotactic regularities and phonological structure (Ćavar and Ćavar, 2009), providing cues for morphological structure induction and syntactic phrasing.

References

Damir Ćavar and Małgorzata E. Ćavar. 2009. On the induction of linguistic categories and learning grammars. Paper presented at the 10th Szklarska Poreba Workshop, March.

Damir Ćavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, and Giancarlo Schrementi. 2004a. Alignment based induction of morphology grammar and its role for bootstrapping. In Gerhard Jäger, Paola Monachesi, Gerald Penn, and Shuly Wintner, editors, *Proceedings of Formal Grammar 2004*, pages 47–62, Nancy.

Damir Ćavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, and Giancarlo Schrementi. 2004b. On statistical bootstrapping. In William G. Sakas, editor, *Proceedings of the First Workshop on Psycho-computational Models of Human Language Acquisition*, pages 9–16.

Damir Ćavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, and Giancarlo Schrementi. 2004c. Syntactic parsing using mutual information and relative entropy. Midwest Computational Linguistics Colloquium (MCLC), June.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.

Robert Frank. 2000. From regular to context free to mildly context sensitive tree rewriting systems: The path of child language acquisition. In A. Abeillé and O. Rambow, editors, *Tree Adjoining Grammars: Formalisms, Linguistic Analysis and Processing*, pages 101–120. CSLI Publications.

Sharon Inkelas and Draga Zec. 1990. *The Phonology-Syntax Connection*. University Of Chicago Press, Chicago.

Ray Jackendoff. 1996. *The Architecture of the Language Faculty*. Number 28 in Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Aravind Joshi, K. Vijay-Shanker, and David Weird. 1991. The convergence of mildly context-sensitive grammar formalisms. In Peter Sells, Stuart Shieber, and Thomas Wasow, editors, *Foundational Issues in Natural Language Processing*, pages 31–81. MIT Press, Cambridge, MA.

Brian MacWhinney and Elizabeth Bates. 1989. *The Crosslinguistic Study of Sentence Processing*. Cambridge University Press, New York.

Eric S. Ristad. 1990. Computational structure of generative phonology and its relation to language comprehension. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 235–242. Association for Computational Linguistics.

Bruce Tesar and Paul Smolensky. 2000. *Learnability in Optimality Theory*. MIT Press, Cambridge, MA.