Sequences of Part of Speech Tags vs. Sequences of Phrase Labels

How Do They Help in Parsing?

Gabriel Infante-Lopez[1] and Maarten de Rijke[2]

[1] FaMAF, Universidad Nacional de Córdoba, Córdoba, Argentina
gabriel@famaf.unc.edu.ar
[2] Informatics Institute, University of Amsterdam, The Netherlands
mdr@science.uva.nl

Abstract. We compare the contributions made by sequences of part of speech tags and sequences of phrase labels for the task of grammatical relation finding. Both are used for grammar induction, and we show that English labels of grammatical relations follow a very strict sequential order, but not as strict as POS tags, resulting in better performance of the latter on the relation finding task.

1 Introduction

Some approaches to parsing can be viewed as a simple context free parser with the special feature that the context free rules of the grammar used by the parser do not exist a priori [**?**,**?**,**?**]. Instead, there is a device for generating bodies of context free rules on demand. Collins [**?**] and Eisner [**?**] use Markov chains as the generative device, while Infante-Lopez and De Rijke [**?**] use the more general class of probabilistic automata. These devices are induced from sample instances obtained from tree-banks. The learning strategy consists of coping all bodies of rules inside the Penn Tree-bank (PTB) to a bodies of rules sample bag which is then treated as the sample bag of an *unknown* regular language. This unknown regular language is to be induced from the sample bag, which is, later on, used for generating new bodies of rules.

Usually, the induced regular language is described by means of a probabilistic automata. The quality of the resulting automata depends on many things; the alphabet of the target regular language being one. At least two such alphabets have been considered in the literature: Part of Speech (POS) tags and grammatical relations (GRs), where the latter are labels describing the relation between the main verb and its dependents; they can be viewed as a kind of non-terminal labels. Using one or the other alphabets for grammar induction might produce different results on the overall parsing task. Which of the two produces "better" automata, that produce "better rules," which in turn lead to "better" parsing scores? This is our main research question in this paper.

Let us provide some further motivation and explanations. In order to obtain phrase structures like the ones retrieved in [**?**], the dependents of a POS tag should consist