Unsupervised Evaluation of Parser Robustness

Johnny Bigert[1], Jonas Sjöbergh[1], Ola Knutsson[1], and Magnus Sahlgren[2]

[1] KTH Nada, 100 44 Stockholm, Sweden, {johnny,knutsson,jsh}@nada.kth.se
[2] SICS, Box 1263, 164 29 Kista, Sweden, mange@sics.se

Abstract. This article describes an automatic evaluation procedure for NLP system robustness under the strain of noisy and ill-formed input. The procedure requires no manual work or annotated resources. It is language and annotation scheme independent and produces reliable estimates on the robustness of NLP systems. The only requirement is an estimate on the NLP system accuracy. The procedure was applied to five parsers and one part-of-speech tagger on Swedish text. To establish the reliability of the procedure, a comparative evaluation involving annotated resources was carried out on the tagger and three of the parsers.

1 Introduction

Automatic parsing of text is a popular field of research. Many of the applications where parsing is used, such as parsing human input to a computer system, handle text that is not proofread. Depending on the application, the text can be relatively error free (e.g. parsing newspaper articles from the internet) or contain large amounts of errors (e.g. using a parser as a tool for second language learners when writing essays). If the intended use of a parser is domains with many errors, it must be robust enough to produce useful output despite noisy input. It is not sufficient to achieve a good performance on error-free text. Usually, the accuracy of a parser on error-free text is known, but the accuracy on texts containing errors is often unknown.

Carroll and others give a comprehensive overview of different parser evaluation methods and discuss some shortcomings [1]. Evaluation of parsers is usually carried out by comparing the parser output to a manually annotated or manually corrected version of a test text. Manual work is expensive, and not necessarily error free. If the NLP system is under development, the evaluation has to be carried out repeatedly. Thus, very large amounts of annotated resources may be required to avoid data exhaustion. Many languages have no large manually annotated resources at all, and those existing often contain only error-free texts.

Manual annotation is not only expensive, but often hard to reuse when evaluating a new parser. Generally, it is non-trivial to map the output of one parser to the output of another [2]. Thus, the effort of manually annotating text with one type of parse information is not generally reusable for other parsers.

To carry out the evaluation of NLP system robustness while avoiding the above-mentioned drawbacks, we propose a procedure that requires no manual work or annotated resources. There are, as pointed out by Menzel [3], many types of robustness. Robustness in this context is defined as the system's reluctance to change its output when the input becomes increasingly noisy and ill-formed. The only requirements of