Syntactic Simplification and Text Cohesion

By Advaith Siddharthan

University of Cambridge Computer Laboratory

August 2004

Presented by Assad Jarrahian and Kevin Bauer

Talk Abstract

This dissertation addresses a significant problem within the domain of Natural Language Processing of *syntactic simplification*, or the process of reducing the grammatical complexity of a natural language text, while retaining its information content and semantic meaning. The primary goal of this delicate process is to transform coarse, difficult to read text into a form that is easier to comprehend for human readers. This process can be extended to modify human readable text into a representation that is more suitable for parsing by a computer program. This thesis addresses how syntactic simplification can be achieved using shallow robust analysis, a small set of contrived simplification rules, and a more detailed analysis of the discourse-level aspects of syntactically re-writing text.

A potential difficulty in this process stems from relative clause and appositive attachment. Siddharthan presents a novel approach to these attachment problems and argues that these attachment decisions should not be made solely on syntactic information. This approach relies upon a shallow discourse model obtained from a lexical knowledge base. He further describes how clause and appositive information can be extracted from a sentence *reliably* using a decision procedure based upon the local context in question, represented by part-of-speech tags and noun chunks.

During the simplification process, Siddharthan presents a formal model that captures the interactions that occur between syntax and discourse. This is an important step in order to transform the original text into a form that is more accessible to a wider audience of readers, but it must be done in such a way as to preserve the text's cohesion. Techniques related to sentence-ordering and cue-word selection is discussed in the context of enhancing the transformed text's cohesion.

To address the problem of syntactic simplification while maintaining strict text cohesion, the author addresses several difficult problems within Natural Language Processing. These include clause and appositive identification and attachment, pronoun resolution, and referring-expression generation.

Siddharthan evaluates his new techniques to address these problems individually, and then evaluates the performance of the entire system as a whole.

In our talk, Assad and Kevin will present the primary contributions to the field of Natural Language Processing that are contained in this thesis. We will further discuss our opinions on the approach and presentation of the thesis. Finally, we will conclude by offering ideas as to how this thesis could be improved.