Proposal for Master Thesis in Computer Science

A Wide-Coverage Grammar and Parser for Swedish

Malin Ahlberg

September 12, 2011

Abstract

By using existing technologies and tools, we would like to implement a robust parser for Swedish. The goal is to be able to parse open domain language and the parser will be evaluated on an extensive Swedish treebank, Talbanken. A parser like this would be of great use for many natural language processing applications, such as translation and information retrieval as well as making semantic representations.

The project is accepted and funded by Center of Language Technology, Gothenburg University.

Description and future work

Making computers able of handling human language is a hard problem. The meaning of a sentence depends not only of which words it consists of, but also on their syntactic use, how they interact and relate to each other. For a computer to make sense of natural language, it needs to analyse this syntactic structure; it needs a good grammar and parser. In this project, a robust parser for Swedish will be implemented using the grammar formalism Grammatical Framework.

The libraries of GF provide a basic grammar for Swedish, covering the fundamental features of the language, such as morphology and commonly used syntax. This is suitable for building domain specific applications. In those cases the user is not allowed to freely compose sentences, but has to stay within the bounds of a controlled language. Both the vocabulary and grammatical structures are fixed, meaning that there only is a limit number of ways to write a sentence. Parsing open domain natural language is a much bigger task, since it involves handling both standard and non-standard grammatical constructions.

The future work will include extending and enhancing the existing Swedish GF grammar, importing lexicon and develop technics for handling unknown words and grammatical constructions, proper names, idioms, ellips etc, in order to make the parser robust.