Proposal for master thesis in computer science

Martin Agfjord, 19870503-4877

March 24, 2014

1 Initial description of the project

The working title of the master thesis is 'Interpreting natural languages into executable machine readable instructions'. This project is will be done for Findwise AB, a company with focus on search driven solutions. The task of the project is to develop a search engine which will take a query in a natural language as input and make a sophisticated query based on the input string.

A sophisticated query is created by parsing the input string with a grammar to retrieve important semantics of the string. The string will be parsed by using a grammar which will be developed for this project. The grammar will be developed in GF (Grammatical Framework). The development of the grammar is the most important part of this project.

In order to delimit the number of keywords supported by the parser, the final application shall be able to help the user by using a hybrid completion system. One completion part from the grammar and one part from the search platform.

After the query is parsed into a tree by the grammar, it will be translated into the query language which will be used to perform the search. The preliminary engine to handle the search is Apache Solr, an enterprise search platform based on Apache Lucene. We've also discussed a graph based database, but since Findwise do not use graph databases at the moment we have chosen to start with Apache Solr.

Another part of the project is to develop a website to perform the search. The website will integrate the grammar-system with Apache Solr. The integration will probably include re-configuration of Apache Solr.

To summarize, a research question for this project could be: How can a grammar be developed in order to translate text queries into sophisticated queries which can be used in established search systems?

2 Related courses

This project is closely related to the course *Programming language technology* in the sense that we want to translate a text into machine readable instructions. The difference is that instead of working with grammars for programming languages we want to develop a grammar for natural languages.

Another closely related course is *Artificial Intelligence*, where we in the Shrdlite project were given a grammar to translate natural language queries into parse-trees. We could use these parse trees to make decisions of what do do in the block world.

I also believe that *Finite automata and formal languages* can be useful for *information extraction*, i.e. extract words from a string. It is also very efficient to analyze formal languages with automatons.

3 Literature

Grammatical Framework: Programming with Multilingual Grammars by Aarne Ranta will be an important resource to develop a working grammar for this project.

Compilers: Principles, Techniques, and Tools by Alfred V. Aho et. al. can also be useful in the development of the grammar.

Artificial Intelligence: A Modern Approach by Stuart Russel et. al. is a complement to the GF-book. It has some chapters regarding natural language processing and many chapters about related subjects. It is considered one of the most up-to-date resource in artificial intelligence and is used in over 1200 universities around the world.