Den stora grammatiken

Malin Ahlberg
Gothenburg University, Sweden
ahlberg.malin@gmail.com

om parsning of free language (take from old report) but for linguistics, make a wide coverage grammar, robust parser By using existing technologies and tools:use three resources, gf, talbanken, saldo 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 goal is to be able to parse open domain language and the parser will be evaluated on an extensive Swedish treebank, Talbanken.

This report describes the start of implementing a robust parser for Swedish, using the grammar formalism Grammatical Framework[**?**] (GF).

1 Introduction

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. Moreover, the parser needs a preprocessor, named entity recognizer and more tequnices to make it robust.

We are working on an extended implementation of a grammar from Swedish, allowing complicated constructions such as focusing different parts of the sentences, using complex relative clauses etc and constructions (byt ord) specific to Swedish.

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 techniques for handling unknown words and grammatical constructions, proper names, idioms, ellipses etc, in order to make the parser robust.

By using this, we hope to eventually implement a grammar able to handle free Swedish. It would be of great use for many natural language processing applications, such as translation and information retrieval as well as making semantic representations.

1.1 Grammatical Framework

GF[**?**] (Grammatical Framework) is a grammar formalism based on functional programming, and suitable for multilingual grammar applications. The key idea is to have one abstract grammar, modelling the

Submitted to:
LREC2012

structure of the domain. In addition, concrete grammars are implementing the abstract in different languages. Since all languages supported in an application shares a common abstract syntax, multilingual is easy to support.

In this way the grammar becomes independent, it may be reused in another project and the implementation may be changed without modifying the application utilizing the grammar.

GF provides resource grammars[**?**] implementing information about the morphological and syntactical rules for 18 different languages. By utilizing these, GF is an advantageous tool for writing special-purpose grammars.

The libraries of GF, the resource libraries[**?**], provides a common abstract and implementations of this in 18 languages. Since translation always should be possible, the abstract may only contain constructions that are common to all implemented languages.

But this also means that all —- added to the resources must be translatable to and expressable in all other language. therefor, the resources are very general, non language specific.

Language specific constructs may be given in the module `Extra.gf`. This may be stylistic changes, idioms, informal expressions etc. As an example, the Swedish `Extra` module contains a function for expressing sentences where the negation is put in focus: *"Inte var jag glad"* (*"Not was I happy"*).

The work to enhance and expand the Swedish GF grammar has already been started

2 Talbanken

Talbanken[**?**] is a Swedish treebank which was put together in the 1970s at Lund University, consisting of 6316 sentences. Each word is tagged with its part of speech, syntactic function and its position in a head dependent analysis. In 2005 Talbanken was mordenized [] and enriched with annotation for a full phrase structure analysis. Talbanken05 was used when training the data-driven parser Maltparser [], which var bra?

3 Saldo

SALDO[**?**] is an open source lexicon resource based on Svenskt Associations Lexikon (SAL). It is developed at Sprakbanken[] at Gothenburg University and intended to be used in language technology research. From SALDO, a large GF lexicon,
`DictSwe.gf` has earlier been extracted [], containing 50 000 entries. Using the same updated tecniques, we would like to reimport the lexicon in order to enlarge, enhance it. The importing-method should be fast and reliable enough to be able to always have an fresh version of the dictonary in GF. We are using parts of Talbanken for testing and development, and the parsetree generated by our grammar will eventually be compared to those from the treebank.

4 The lexicon tool

A tool for automatically acquisition has been created. Making use of the tags in Talbanken and the paradigms in the Swedish resource grammar, it interactivelly generates GF lexicons. The user verifies tho correctness of a guess from the program, and can either allow the word to be added to the lexicon, remove it or demand the program to make another guess. Altough using very simple tecniques, the test of the tool on verbs has a correctness rate of about 70-75%, and can easily ? be improved.

use together with Extract/FM,

5 Mapping of trees

The information from the tags in Talbanken can be used for many xx. The tools for lexical extraction can be enhanced if we get more information about which form the word is currently used in. We are currently working on a automatical mapping from Talbanken trees to trees in GF format. This would later enable us to extract possibilities for how often different functions are used, a feature (!) that would improve the parsing considerably. The evalution of the parser could also be accomplished by comparing the trees from the parser and from the mapping-process. brastartord there are differences in the notation of the sttet trden r uppbyggda p
exempel??

explation of the trees why we need the mapping : evaluation, adding words, finding missing constructions difficulties

6 New things and hard things in the grammar

Due to the ressemblemences between the Scandinavian languages, they share about 80% (?) of their resource code. This is good because blabla. However, for a bigger Swedish grammar, the languages need to be more separated. Some constructions may common in Norweigan and Danish and therefor set as the standard function altough they are very rare in Swedish. A number of grammatical constructions have been added the the Swedish grammar. Those include using the reflexive pronoun *sitt*:
Han sg sitt hus *He saw his (own) house*
as opposed to **Han sg hans hus** *He saw his (an other person's) house*
The possibility to put a part of a sentence in focus has also been added such as:
Glad var han inte *Happy was he not* Some ellipses are added, while there is work going on to add trickier ones (eller inte ellipser just nu dr, men vi skriver ngot fint hr med, ja-jaaaa!).

7 Future Work

robust parser, ner, preprocessing, adding new words, probabilties ellipses, chunk parsing

8 Conclusion

hard but nice why better than statistical