

UMEÅ UNIVERSITET

Best Trees Extraction and Contextual Grammars for Language Processing

Anna Jonsson

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av teknologie doktorsexamen framläggs till offentligt försvar i MA316, MIT-huset,
Fredagen den 11:e juni, kl. 10:00.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Dr. Hendrik Jan Hoogeboom,
Universiteit Leiden, Leiden, Nederländerna.

Institutionen för Datavetenskap

Organization
Umeå University
Department of Computer Science

Document type
Doctoral thesis

Date of publication
21 May 2021

Author
Anna Jonsson

Title
Best Trees Extraction and Contextual Grammars for Language Processing

Abstract
In natural language processing, the *syntax* of a sentence refers to the words used in the sentence, their grammatical role, and their order. *Semantics* instead concerns the concepts represented by the words in the sentence and their relations - the meaning of the sentence. While a human can easily analyse a sentence to figure out its grammatical construction and meaning, this is a difficult task for a computer that lacks a language model in the form of a human brain. To start with, the computer needs data structures that can represent syntax and semantics. Then, the computer must have some information about what is considered correct syntax and semantics -- this can be provided in the form of human-annotated corpora of natural language. There are several ways to capture the correctness aspect of a corpus in a computer-based language model, but here we consider rule-based formalisms only. Rule-based formalisms use rule sets to define *formal languages* and are in that sense very similar to the grammatical rules of natural language. The goal is thus to model natural languages using formal languages.

Trees are commonly used to represent syntactic analyses of sentences, and graphs can represent their semantics. Examples of rule-based formalisms that define languages of trees and graphs are *tree automata* and *graph grammars*, respectively. When used in language processing, these formalisms are normally assigned weights, meaning that the trees and graphs in the languages defined by the formalisms have weights; this assignment is done to allow us to rank the trees and graphs by comparing them to the information extracted from the corpora of correct linguistic data. There are numerous systems that combine rule-based formalisms with algorithms working on their rule sets to solve language processing tasks.

Since natural language is very complicated to model, there are many small gaps in the research of natural language processing to dive into. The research of this thesis regards two separate problems: First, we have the *N-best problem*, which is about finding a number *N* of top-ranked hypotheses given a hypothesis space represented by a rule-based formalism; the hypotheses can for example have the form of syntax trees. The second problem is that of *semantic modelling*, whose aim is to find a formalism powerful enough to define languages of semantic representations. This model can however not be too powerful since we still want to be able to efficiently compute answers to language processing tasks.

This thesis is divided into two parts according to the two problems introduced above. The first part covers the *N*-best problem for weighted tree automata. In this line of research, we develop and evaluate multiple versions of an algorithm that efficiently computes the *N* best (top-ranked) trees in a tree language given by a weighted tree automaton. In the second part, we study how rule-based formalisms can be used to model graphs that represent meaning, i.e., semantic graphs. We investigate an existing formalism for this purpose and thereby learn what properties of that formalism are necessary for semantic modelling. Finally, we use our new-found knowledge to develop an alternative formalism, and argue that it is better suited for the task of semantic modelling than existing formalisms.

Keywords
rule-based formalisms, weighted tree automata, graph grammars, hyperedge replacement, contextual hyperedge replacement, semantic modelling, meaning representations, algorithms, natural language processing

Language	**ISBN**	**ISSN**	**Number of pages**
English	print: 978-91-7855-521-5 PDF: 978-91-7855-522-2	0348-0542	58 + 6 papers