

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* concerns the concepts represented by the words in the sentence and their relations, i.e., the meaning of the sentence. While a human can easily analyse a sentence in a language they understand to figure out its grammatical construction and meaning, this is a difficult task for a computer. To analyse natural language, the computer needs a *language model*. First and foremost, the computer must have data structures that can represent syntax and semantics. Then, the computer requires some information about what is considered correct syntax and semantics – this can be provided in the form of human-annotated corpora of natural language. Computers use *formal languages* such as programming languages, and our goal is thus to model natural languages using formal languages. There are several ways to capture the correctness aspect of a natural language corpus in a formal language model. One strategy is to specify a formal language using a set of rules that are, in a sense, very similar to the grammatical rules of natural language. In this thesis, we only consider such rule-based formalisms.

Trees are commonly used to represent syntactic analyses of sentences, and graphs can represent the semantics of sentences. Examples of rule-based formalisms that define languages of trees and graphs are *tree automata* and *graph grammars*, respectively. When used in language processing, the rules of a formalism are normally given weights, which are then combined as specified by the formalism to assign weights to the trees or graphs in its language. The weights enable us to rank the trees and graphs by their similarity to the linguistic data in the human-annotated corpora.

Since natural language is very complicated to model, there are many small gaps in the research of natural language processing to address. The research of this thesis considers two separate but related problems: First, we have the *N-best problem*, which is about finding a number N of top-ranked hypotheses given a ranked hypothesis space. In our case, the hypothesis space is represented by a weighted rule-based formalism, making the hypothesis space a weighted formal language. The hypotheses themselves can for example have the form of weighted syntax trees. The second problem is that of *semantic modelling*, whose aim is to find a formalism complex enough to define languages of semantic representations. This model can however not be too complex since we still want to be able to efficiently compute solutions 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 efficient algorithm that solves the problem in question. Since our algorithm is the first to do so, we theoretically and experimentally evaluate it in comparison to the state-of-the-art algorithm for solving an easier version of the problem. 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 and through this work learn what properties of that formalism are necessary for semantic modelling. Finally, we use our new-found knowledge to develop a more specialised 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
English

ISBN
print: 978-91-7855-521-5
PDF: 978-91-7855-522-2

ISSN
0348-0542

Number of pages
60 + 6 papers