Thesis for the Degree of Doctor of Philosophy

Multilingual Abstractions: Abstract Syntax Trees and Universal Dependencies

Prasanth Kolachina

Presentation:
June 14th, 2019, 10:00 CEST
Room EA, EDIT Building
Hörsalsvägen 11,
Chalmers University of Technology, Campus Johanneberg
Also reachable from Rännvägen 6.

Faculty opponent:
Joakim Nivre
Professor of Computational Linguistics
Department of Linguistics and Philology
Uppsala University, Sweden

UNIVERSITY OF GOTHENBURG

The thesis is available at:
Department of Computer Science & Engineering
Chalmers | University of Gothenburg
Gothenburg, Sweden, 2019

Phone: 031 - 772 10 00

Abstract

This thesis studies the connections between parsing friendly representations and interlingua grammars developed for multilingual language generation. Parsing friendly representations refer to dependency tree representations that can be used for robust, accurate and scalable analysis of natural language text. Shared multilingual abstractions are central to both these representations. Universal Dependencies (UD) is a framework to develop cross-lingual representations, using dependency trees for multlingual representations. Similarly, Grammatical Framework (GF) is a framework for interlingual grammars, used to derive abstract syntax trees (ASTs) corresponding to sentences. The first half of this thesis explores the connections between the representations behind these two multilingual abstractions. The first study presents a conversion method from abstract syntax trees (ASTs) to dependency trees and present the mapping between the two abstractions – GF and UD – by applying the conversion from ASTs to UD. Experiments show that there is a lot of similarity behind these two abstractions and our method is used to bootstrap parallel UD treebanks for 31 languages. In the second study, we study the inverse problem i.e. converting UD trees to ASTs. This is motivated with the goal of helping GF-based interlingual translation by using dependency parsers as a robust front end instead of the parser used in GF.

The second half of this thesis focuses on the topic of data augmentation for parsing – specifically using grammar-based backends for aiding in dependency parsing. We propose a generic method to generate synthetic UD treebanks using interlingua grammars and the methods developed in the first half. Results show that these synthetic treebanks are an alternative to develop parsing models, especially for under-resourced languages without much resources. This study is followed up by another study on out-of-vocabulary words (OOVs) – a more focused problem in parsing. OOVs pose an interesting problem in parser development and the method we present in this paper is a generic simplification that can act as a drop-in replacement for any symbolic parser. Our idea of replacing unknown words with known, similar words results in small but significant improvements in experiments using two parsers and for a range of 7 languages.

Keywords

Natural Language Processing, Grammatical Framework, Universal Dependencies, multilinguality, abstract syntax trees, dependency trees, multilingual generation, multilingual parsers

ISBN 978-91-7833-508-4 (PRINT), 978-91-7833-509-1 (PDF)