Measuring the Divergence of Dependency Structures Cross-Linguistically to Improve Syntactic Projection Algorithms

Ryan Georgi[†], **Fei Xia**[†], **William D. Lewis**[*]

University of Washington[†], Microsoft Research[*]
rgeorgi@uw.edu, fxia@uw.edu, wilewis@microsoft.com

Abstract

Syntactic parses can provide valuable information for many NLP tasks, such as machine translation, semantic analysis, etc. However, most of the world's languages do not have large amounts of syntactically annotated corpora available for building parsers. Syntactic projection techniques attempt to address this issue by using parallel corpora between resource-poor and resource-rich languages, bootstrapping the resource-poor language with the syntactic analysis of the resource-rich language. In this paper, we investigate the possibility of using small, parallel, annotated corpora to automatically detect divergent structural patterns between two languages. These patterns can then be used to improve structural projection algorithms, allowing for better performing NLP tools for resource-poor languages, in particular those that may not have large amounts of annotated data necessary for traditional, fully-supervised methods. While this detection process is not exhaustive, we demonstrate that important instances of divergence are picked up with minimal prior knowledge of a given language pair.

Keywords: Multilinguality, Translation Divergence, Syntactic Projection

This paper will appear in the proceedings of LREC 2012.