William Lewis, Fei Xia, & Dan Jinguji

Enriching Language Data through Projected Structures

This paper explores the potential for annotating and enriching data for minority or endangered languages via the alignment and projection of structure from annotated and parsed data for a resource-rich language such as English. The work presented here draws inspiration from the work of (Yarowksy and Ngai, 2001), who tested the methods for projecting linguistic annotations from one language to another, where the resulting projections could be used to train automated part-of-speech taggers and NP bracketers. However, unlike Yarowsky and Ngai, who sought to develop tools and resources for the 200+ major languages of the world, we seek to develop enriched, searchable resources for a larger number of the world's languages, most of which have no significant digital presence. We do this by tapping into the large body of Web-based linguistic data, most of which exists in small, analyzed chunks embedded in scholarly papers, journal articles, Web pages, and other online documents. By harvesting and enriching these data, we provide an automated means to search for them, facilitating a kind of structure-based, "construction" query. Further, the enriched data can be used to train and develop robust, statistically-based NLP tools, which can be used for the automated annotation and analysis of language data, especially that of resource-poor and computationally underrepresented languages.