Breaking the Resource Bottleneck for Multilingual Parsing

Rebecca Hwa, Philip Resnik, Amy Weinberg
Language and Media Processing Laboratory
Institute for Advanced Computer Studies
College Park, MD 20742

Abstract

We propose a framework that enables the acquisition of annotation-heavy resources such as syntactic dependency tree corpora for low-resource languages by importing linguistic annotations from high-quality English resources. We present a large-scale experiment showing that Chinese dependency trees can be induced by using an English parser, a word alignment package, and a large corpus of sentence-aligned bilingual text. As a part of the experiment, we evaluate the quality of a Chinese parser trained on the induced dependency treebank. We find that a parser trained in this manner out-performs some simple baselines inspite of the noise in the induced treebank. The results suggest that projecting syntactic structures from English is a viable option for acquiring annotation syntactic structures quickly and cheaply. We expect the quality of the induced treebank to improve when more sophisticated filtering and error-correction techniques are applied.

***The support of the LAMP Technical Report Series and the partial support of this research by the National Science Foundation under grant EIA0130422 and the Department of Defense under contract MDA9049-C6-1250 is gratefully acknowledged.