Advances in Cross-Lingual Syntactic Transfer

Ryan McDonald, Google Research, USA

The idea to use annotated resources from one language to learn models for another has been around for at least a decade. Typically these models have relied on access to parallel data. However, recent approaches have focused on "direct" cross-lingual transfer, and in particular, delexicalized transfer. Delexicalized parsing models are conditioned only on properties of the input that are available across languages, typically induced tags or clusters. Since these properties are universally available, it is possible to directly use a parser trained on English for every other language. This simple method has shown itself to be surprisingly effective and outperforms the best weakly-supervised models by a significant margin. However, the assumptions underlying these models are far to weak to obtain parsing accuracies at the level of monolingual supervised methods. In this talk I will focus on porting ideas from work on selective parameter sharing in multi-source direct transfer to highly accurate latent CRF parsing models. I will then present novel semi-supervised learning algorithms that relexicalize these models on unlabeled target language data to give significant improvements. The final model brings us one step closer to building robust syntactic parsers for all the world's languages.

Joint work with: Oscar Tackstrom, Slav Petrov, Keith Hall