Machine Translation and Morphologically-rich Languages: Research Workshop of the Israel Science Foundation, University of Haifa, Israel, 23-27 January, 2011

Jason Eisner:
A Non-Parametric Bayesian Approach to Inflectional Morphology

We learn how the words of a language are inflected, given a plain text corpus plus a small supervised set of known paradigms. The approach is principled, simply performing empirical Bayesian inference under a straightforward generative model that explicitly describes the generation of

1. The grammar and subregularities of the language (via many finite-state transducers coordinated in a Markov Random Field).
2. The infinite inventory of types and their inflectional paradigms (via a Dirichlet Process Mixture Model based on the above grammar).
3. The corpus of tokens (by sampling inflected words from the above inventory).

Our inference algorithm cleanly integrates several techniques that handle the different levels of the model: classical dynamic programming operations on the finite-state transducers, loopy belief propagation in the Markov Random Field, and MCMC and MCEM for the non-parametric Dirichlet Process Mixture Model.

We will build up the various components of the model in turn, showing experimental results along the way for several intermediate tasks such as lemmatization, transliteration, and inflection. Finally, we show that modeling paradigms jointly with the Markov Random Field, and learning from unannotated text corpora via the non-parametric model, significantly improves the quality of predicted word inflections.

This is joint work with Markus Dreyer.