Scalable Feature-Rich Structured Prediction with Little Supervision

Waleed Ammar
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
wammar@cs.cmu.edu

The goal of my thesis is to improve structured prediction in low-resource scenarios. In particular, I am interested in methods to scalably leverage unlabeled data and use rich representations of a structured input-output pair.

The first part introduces the conditional random field autoencoder models [1] for unsupervised learning problems. Each input’s latent representation is predicted conditional on the observable data using a feature-rich conditional random field. Then a reconstruction of the input is (re)generated, conditional on the latent structure, using models for which maximum likelihood estimation has a closed-form. Representing two copies of observable examples, conditionally independent given the latent structure, enables the flexibility of defining features with global scope in one copy (which we only condition on) while maximizing the likelihood of the other copy. We use this framework for POS induction, word alignment and grammar induction.

The second part extends the CRF autoencoder framework in various partial supervision scenarios:

- when a small number of structured label examples are available, the CRF [4] part of the model can learn weights for a few features which characterize input-output correlations. By penalizing divergence from those weights, we can use the unlabeled data to train the full model without washing away label supervision.

- when domain knowledge can be specified in terms of expectations over feature functions, we show an elegant integration of the posterior regularization framework [2] with the CRF autoencoder models.

- when cross-lingual type-level supervision is available, we show how supervision in the high-resource language can be used as side information to further enrich feature functions in the CRF model. Examples include unsupervised POS induction when parallel data is available for training.

- when human annotated dependency parses used for training are underspecified. The fragmentary unlabeled dependency grammar (FUDG) [6] allows annotators to ignore uncertain or irrelevant attachments to improve their throughput.

The third part of the thesis is application-oriented. We study the problem of representation learning for statistical machine translation. For example, preprocessing heuristics have been shown to substantially improve translation quality from/to morphologically rich languages [5, 3]. However, learning an appropriate subword lexical representation for translation rules remains an open problem. In this project, we consider joint modeling of bitext word alignment and lexical representations. We generate alternative lexical representations by optional removal, splitting, and concatenation of morphemes.

The plan is to propose in August 2014 and defend in August 2016.
References


