Research Statement

Jennifer Gillenwater

My primary research interests are machine learning and natural language processing. Within these broad areas, I've been involved in a wide variety of relevant projects. In particular, my previous work includes posterior regularization for parsing, parser optimization using information retrieval measures, a partition conditional random field model for speaker diarization, and determinantal point processes subset selection tasks.

1 Posterior Regularization for Parsing

The initial research I conducted involved exploration of posterior regularization. This is an efficient, novel technique for incorporating side-information into the optimization of models that do not readily admit direct coding of such information. I co-authored several papers applying this technique to the natural language processing task of parsing. Details of these applications are given in the following two subsections. Additionally, I wrote the parsing portion of a posterior regularization toolkit, whose code is publicly available.

More recently, in 2013 I wrote a paper extending the posterior regularization technique to non-linear regularizers. Via this extension, I was able to shoe improvements in part-of-speech tagging and handwriting recognition accuracy. Details are given in the third subsection below.

1.1 Dependency Grammar Induction via Bitext Projection Constraints

Broad-coverage annotated treebanks necessary to train parsers do not exist for many resource-poor languages. The wide availability of parallel text and accurate parsers in English has opened up the possibility of grammar induction through partial transfer across bitext. We consider generative and discriminative models for dependency grammar induction that use word-level alignments and a source language parser (English) to constrain the space of possible target trees. Unlike previous approaches, our framework does not require full projected parses, allowing partial, approximate transfer through linear expectation constraints on the space of distributions over trees. We consider several types of constraints that range from generic dependency conservation to language-specific annotation rules for auxiliary verb analysis. We evaluate our approach on Bulgarian and Spanish CoNLL shared task data and show that we consistently outperform unsupervised methods and can outperform supervised learning for limited training data.

1.2 Sparsity in Dependency Grammar Induction

A strong inductive bias is essential in unsupervised grammar induction. We explore a particular sparsity bias in dependency grammars that encourages a small number of unique dependency types. Specifically, we investigate sparsity-inducing penalties on the posterior distributions of parent-child POS tag pairs in the posterior regularization framework. In experiments with 12 languages, we achieve substantial gains over the standard expectation maximization (EM) baseline, with average improvement in attachment accuracy of 6.3%. Further, our method outperforms models based on a standard Bayesian sparsity-inducing prior by an average of 4.9%. On English in particular, we show that our approach improves on several other state-of-the-art techniques.

1.3 Graph-Based Posterior Regularization for Semi-Supervised Structured Prediction

We present a flexible formulation of semi-supervised learning for structured models, which seamlessly incorporates graph-based and more general supervision by extending the posterior regularization (PR) framework. Our extension allows for any regularizer that is a convex, differentiable function of the appropriate marginals. We show that surprisingly, non-linearity of such regularization does not increase the complexity of learning, provided we use multiplicative updates of the structured exponentiated gradient algorithm. We illustrate the extended framework by learning conditional random fields (CRFs) with quadratic penalties arising from a graph Laplacian. On sequential prediction tasks of handwriting recognition and part-of-speech (POS) tagging, our method makes significant gains over strong baselines.

2 Parser Optimization Using Information Retrieval Measures

While at Microsoft in the summer of 2010, I completed additional parsing work. This work focused on adapting parsing as a tool for improving web search results. In 2013, I published a paper highlighting the positive results of applying this approach. Details are below.

Parsers have been shown to be helpful in information retrieval (IR) tasks. While previous work focused on using traditional syntactic parse trees, this work proposes a new approach where, unlike previous work, the relevance between a document and a query is modeled by the weighted tree edit distance (TED) between their parses, and the parser parameters are optimized directly for a non-convex and non-smooth IR measure. We evaluate our method on a large scale web search task consisting of a real world query set. Results show that the new parser is more effective for document retrieval than the baseline parser using traditional syntactic parse trees.

3 Discriminative Diarization in Video

For about a year I worked on the following task, and, while the proposed model seems like it should do well, I was ultimately unable to get it to perform better than a clustering baseline. Details are below.

We consider the task of speaker diarization ("who spoke when") for the novel domain of videos in the wild. Most current speaker diarization systems, such as those that compete in the NIST Rich Transcription evaluations, focus on speech data from meetings, broadcast news, and telephone conversations. Diarization for domains such as movies, TV episodes, and YouTube videos is more difficult as the data is generally noisier. On the other hand, videos offer additional visual and textual cues that can be exploited to compensate for the difficulty posed by an increased amount of noise; we can draw stabilizing features from the visual data by recognizing faces in addition to voices, and we can exploit textual clues in the video's closed captions. We propose a *discriminative model that seamlessly combines audio, visual, and textual cues.* In particular, we propose a conditional random field where each state partitions a sequence of several speech clips by speaker.

4 Determinantal Point Processes

Most recently, I've worked on several projects centered around determinantal point processes (DPPs). These processes, initially used in physics to model fermions, have in the past few years been further developed for a variety of machine learning problems. In general, what DPPs offer is an efficient means for selecting a diverse, high-quality subset of points. The following two subsection detail the DPP work with which I've been involved.

4.1 Discovering Diverse and Salient Threads in Document Collections

We propose a novel probabilistic technique for modeling and extracting salient structure from large document collections. As in clustering and topic modeling, our goal is to provide an organizing perspective into otherwise overwhelming amounts of information. We are particularly interested in revealing and exploiting relationships between documents. To this end, we focus on extracting diverse sets of threads—singly-linked, coherent chains of important documents. To illustrate, we extract research threads from citation graphs and construct timelines from news articles. Our method is highly scalable, running on a corpus of over 30 million words in about four minutes, more than 75 times faster than a dynamic topic model. Finally, the results from our model more closely resemble human news summaries according to several metrics and are also preferred by human judges.

4.2 Near-Optimal MAP Inference for Determinantal Point Processes

Many DPP inference operations, including normalization and sampling, are tractable; however, finding the most likely configuration (MAP), which is often required in practice for decoding, is NP-hard, so we must resort to approximate inference. Because the objective is log-submodular, greedy algorithms have been used in the past with some empirical success; however, these methods only give approximation guarantees in the special case of monotone objectives, which correspond to a restricted class of DPPs. In this paper we propose a new algorithm for approximating the MAP problem based on continuous techniques for submodular function maximization. Our method involves a novel continuous relaxation of the log-probability function, which, in contrast to the multilinear extension used for general submodular functions, can be evaluated and differentiated exactly and efficiently. We obtain a practical algorithm with a 1/4-approximation guarantee for a more general class of non-monotone DPPs; our algorithm also extends to MAP inference under complex polytope constraints, making it possible to combine DPPs with Markov random fields, weighted matchings, and other models. We demonstrate that our approach outperforms standard and recent methods on both synthetic and real-world data.