Jason Naradowsky - Section G: Addendum

http://www.cs.umass.edu/~narad/
narad@cs.umass.edu

Department of Computer Science
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003

Research Experience

I've had a variety of research experience over the past six years, both in industry and academia, and across a few different fields. My first research experience was as a research assistant in both the computer science and psychology departments at SUNY Oswego, focusing on human-computer interaction. More specifically, I investigated the role of implied tone (emotion, or attitude) in the feedback dialog of a Java critiquing system, and developed a web application for testing culturally-specific aspects of usability (focusing on the particular contrasts between Chinese and American user preferences).

At SUNY Buffalo I moved more into natural language processing. I did a couple corpus studies, and examined the cause of significant performance hits when parsing out-of-domain text with lexicalized, Collins-style parsers. We found that unseen events were rarely the source of these errors, which implies that attempts to improve portability by simply adding more data have little direct effect and are slow to address the real problem: inappropriate ratios between well-represented head/parent events for the new domain. This was joint work with Doug Roland.

During my masters degree at the University of Edinburgh, I worked with Sharon Goldwater on improving unsupervised models of morphology induction by incorporating what loosely corresponds to phonological information. By positing a more sophisticated model of the latent variables that generate each word form, we learned small, context-sensitive transformations in which a character might be inserted or deleted. This proved to be useful on English inflectional morphology where this phenomenon is widespread (step + ing = stepping, rate + ing = rating). This received high marks in review and lead to an IJCAI paper in the following year. Each of these three research components comprised approximately 30% of the degree requirements.

Also while at Edinburgh I participated in the Google Summer of Code doing additional research and implementation of dependency parsing models for the Natural Language Toolkit. I was supervised by Sebastian Riedel and Jason Baldridge.

At UMass I've worked on several projects. I've dabbled in the use of factor graphs for semantic role labeling, and for morphological tagging (using the CoNLL feature set), and I'm preparing a paper for submission on some efficient parsing and joint inference work along these lines. I've also worked in the more Bayesian generative modeling paradigm, where I worked with David Mimno and Hanna Wallach to submit an EMNLP paper on ”polylingual topic models”: topic models that share topic parameters across loosely parallel sets of documents where each document is in a different language. In addition to pure CS research, I've also spent time in the linguistics department, and have developed some models of phonology learning, and studied the use of certain constructions on scalar implicatures.

Most recently I also did a summer internship with Kristina Toutanova at Microsoft Research, where we developed a somewhat novel word alignment model for machine translation. A typical word alignment model induces a pairing of words between parallel sentences, but this can be problematic for morphologically rich languages where a word can correspond to many different English words. To find a more refined alignment we learn a segmentation over the morphologically complex language jointly with the alignments, using the HMM alignment model as our baseline. This system not only improved significantly on state of the art monolingual segmentations, but the alignment model outperforms a very competitive system in that task as well.

Work that is partially done and may be in submission this semester deals with factor graph parsing, named entity recognition, phonology constraint learning, topic models, and incremental parsing.

Abstracts

In Submission, 2010

This paper describes an unsupervised dynamic graphical model for morphological segmentation and bilingual morpheme alignment for statistical machine translation. The model extends Hidden Semi-Markov chain models by using factored output nodes and special structures for its conditional probability distributions. It relies on morpho-syntactic and lexical source-side information (part-of-speech, morphological segmentation, dependency analysis) while learning a morpheme segmentation over the target language. Our model outperforms a competitive word alignment system in alignment quality. Used in a monolingual morphological segmentation setting it substantially improves accuracy over previous state-of-the-art models on three Arabic and Hebrew datasets.

Polylingual Topic Models, EMNLP 2009

Topic models are a useful tool for analyzing large text collections, but have previously been applied in only monolingual, or at most bilingual, contexts. Meanwhile, massive collections of interlinked documents in dozens of languages, such as Wikipedia, are now widely available, calling for tools that can characterize content in many languages. We introduce a polylingual topic model that discovers topics aligned across multiple languages. We explore the model's characteristics using two large corpora, each with over ten different languages, and demonstrate its usefulness in supporting machine translation and tracking topic trends across languages.

Improving morphology induction by learning spelling rules, IJCAI 2009.

Unsupervised learning of morphology is an important task for human learners and in natural language processing systems. Previous systems focus on segmenting words into substrings (taking => tak.ing), but sometimes a segmentation-only analysis is insufficient (e.g., taking may be more appropriately analyzed as take.ing, with a spelling rule accounting for the deletion of the stem-final e). In this paper, we develop a Bayesian model for simultaneously inducing both morphology and spelling rules. We show that the addition of spelling rules improves performance over the baseline morphology-only model.

Improving Morphology Induction with Phonological Rules, Edinburgh MSc thesis.

Recent research in computational approaches to natural language learning have lead to the development of nonparametric Bayesian models capable of inducing simple morphological structure. These models differ from previous work by placing a stronger emphasis on the interaction between grammar components and the learning biases inherent in the system. However, it can be shown that these models will lead to spurious classifications in situations where phonological constraints operate. Previous work has shown that phonological rules can, themselves, be learned via unsupervised methods, and it is our hypothesis that these rules can be used to further refine inference in a Bayesian morphology induction model. We show that an induction model augmented to account for phonological deletion transformations can yield better performance than an unaugmented baseline system.

The Effect of Frequencies and Unseen Events on Parser Portability, University of Buffalo MS thesis.

Previous work in statistical parsing has shown substantial drops in average precision and recall when parsing out of domain text. While the extent of this performance degradation has been thoroughly documented, precisely what aspects of the parsing model lead it to incorrect conclusions on unmatched data has gone unanswered. We train two parsing models on the Brown and Wall Street Journal portions of the Penn Treebank, and attempt to answer this question by examining the sentences that fail to parse correctly in the ported model. By analyzing the frequencies of the events used to define each model, we classify each incorrect parse based on whether or not these events were observed in both corpora. We find that relative frequencies, particularly in the events defining head-word and head-parent relationships, carry significantly more weight than unseen events in steering the parser to incorrect conclusions when using the out of domain parsing model.

Publications

Refereed Conference Proceedings

[1] Mimno, D., Wallach, H., Naradowsky, J., Smith, D., and McCallum, A. Polylingual topic models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2009).

[2] Naradowsky, J., and Goldwater, S. Improving morphology induction by learning spelling rules. In *International Joint Conference on AI* (2009), pp. 1531–1537.

Workshop Proceedings

[1] David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. Polylingual topic models. In *The Learning Workshop (A.K.A. The Snowbird Workshop)*, 2009.

Other Publications

[1] Naradowsky, J., Pater, J., Smith, D., and Staubs, R. Learning hidden metrical structure with a log-linear model of grammar. In *Computational Modelling of Sound Pattern Acquisition* (Edmonton, 2010), pp. 59–60.

Honors & Awards

2005 Oebele Van Dyk Outstanding Senior in Computer Science Award
State University of New York at Oswego
(Though I cannot find documentation - the website is under renovation.)