

Research Statement

Noah A. Smith

July 2008

1 Introduction

Can computers understand human language? In some ways they already give the appearance of doing so—we have reasonably good dialog systems (in some domains) and very good Web search (from keywords)—but I believe there is at least a career’s worth of work to be done before computers achieve fluency and broad-domain competence with language. Hence my research goals span many aspects of the problem of **natural language processing** (hereafter, NLP).

The key to this problem, I believe, is the right combination of **learning** from available data, efficient algorithms and representations, and identifying problems that lead NLP research in the right direction. Here I will briefly discuss my current work and goals, organized as follows:

- Acquiring linguistic knowledge from data via **statistical grammar learning** (§2);
- Solving **formal** and algorithmic problems encountered in computational linguistics (§3);
- Bridging the gap between core methods like those above and real-world **applications** (§4); and
- Exploring new NLP techniques that **ground** text in non-linguistic data (§5) to aid human decision-making and understanding.

My earlier contributions related to these themes, as well as acquired and planned research funding supporting the work, will be noted.

2 Statistical Grammar Learning

Grammars¹ provide an elegant framework for defining and learning natural language processing components. My research spans many dimensions of the grammar learning problem: learning from raw data, scaling up to very large corpora, techniques applicable to data in many languages, efficiency, and moving from syntax to semantics. I address these directions in turn.

Past work Unlike programming languages, the rules of natural language grammars are difficult to write down succinctly and are seldom obeyed perfectly. Algorithms to syntactically parse natural languages are expensive and require considerable human effort in the form of manual linguistic annotations. My dissertation [13] and related papers [14, 15, 16, 17] developed several algorithms to learn syntactic grammars from *unannotated* text. That line of work comprises two themes: encoding prior domain knowledge within objective functions that generalize the usual maximum likelihood approach to unsupervised learning, and exploiting domain knowledge to design better numerical optimization routines. These methods represent the best known for learning robust grammars from unlabeled real-world text examples, though unsupervised grammar learning still lags far behind the best supervised methods in accuracy.

¹I intend broad scope: grammars are formal systems for describing discrete phenomena, not merely syntactic models.

Empirical Bayesian approaches More recently, I have begun exploring empirical Bayesian methods for unsupervised grammar induction together with my students Shay Cohen and Kevin Gimpel, and we have met with considerable success [1]. Apart from its elegant theoretical underpinnings, the Bayesian approach reduces dependence on hyperparameter tuning, one of the unsolved problems noted in my thesis. We are continuing this work with plans to learn grammars for multiple languages together (connecting them through a prior), and encapsulating the ideas of “bias” and “contrast” in my dissertation in a more Bayesian framework.

Large-scale learning Grammar induction is usually carried out on a few thousand short sentences only, for computational reasons. Kevin Gimpel, Shay Cohen, and I are considering the question of scaling grammar induction to more data, and to more complex data. Leveraging a 4,000-CPU supercomputer (“M45”) generously provided by Yahoo, my students and I have been experimenting with larger and more complex datasets than were feasible in the past. This involves distributed implementations of EM (and related Bayesian and non-Bayesian iterative algorithms). Expensive, iterative algorithms like EM and discriminative structured prediction methods (CRFs, MMMNs) are widely used in NLP but not perfectly suited for MapReduce-style parallelism. Preliminary results, using orders of magnitude more data and much longer sentences for grammar induction, are in review [6]. I believe that solving this iterative-training scaling problem is crucial for statistical NLP researchers to make good use of the ever-growing available data and ever more demanding machine learning algorithms. This work is supported by computational resources provided by Yahoo and NSF grant IIS-0836431.

Multilingual NLP Despite the fact that most people do not speak English, NLP is far more advanced for English than any other language. Indeed, many methods capitalize on the idiosyncrasies of English and are therefore inappropriate for other languages. For example, many languages have more complex word-formation (morphology) systems or more free word order than English. I have developed methods for morphological disambiguation [19], efficient learning of rich-featured models [20], cross-language learning [11], and most recently the integration of morphology and syntax [3]. My multilingual parsing efforts are supported by NSF grant IIS-0713265 and an IBM Faculty Award.

Efficient representations This problem of joint learning and joint prediction of *multiple* kinds of linguistic structure is a topic of continued interest [2]. My approach is to build expressive grammatical models by combining simpler ones (e.g., a weighted regular grammar for morphological disambiguation and a weighted context-free grammar for syntax). Predicting these joint structures can be computationally expensive. I have been exploring efficient approximations to deep NLP representations, such as “vine” grammar [5]. More recently students André Martins, Dipanjan Das, and I have been exploring models for dependency syntax that effectively break the standard “arc-factored” assumptions. We approximate richer features (for which parsing becomes NP-hard) using *stacking*, and give theoretical motivation; the technique gives state-of-the-art improvements on eight out of ten languages tested [9] while maintaining quadratic parser runtime.

Broad-coverage semantics I recently started a major project with collaborators Alan Black, Rebecca Hwa (U. of Pittsburgh), and Ric Crabbe (U.S. Naval Academy) to aggregate and visualize information in large newstext collections (titled RAVINE), with a focus on the semantic relationships among attributed quotes. This project is intended to provide a dataset and platform to help develop robust *semantic* representations and processing. Through data visualization and the user-oriented goal of identifying semantic relationships between public statements (e.g., paraphrase, contradiction, topical relatedness), we plan to develop lightweight semantic methods that move beyond lexical semantics, predicate-argument relations, and narrow-domain logic approaches to meaning, toward practical models of sentence meaning at the granularity of quoted utterances. This project is supported by DARPA grants HR-0010110013 and NBCH-1080004.

3 Formal Issues in Computational Linguistics

In both research and teaching, I advocate for high-level representations of complex NLP models and algorithms, for improved efficiency, modularity, understandability, and ease of implementation. This is especially important as NLP models richer phenomena and moves toward joint models, as discussed above. Building on my earlier work on the Dyna language [4], my students and I have been exploring how weighted logic programs can be used to design and implement very complicated multi-structure NLP algorithms [2]. As my students develop expertise with these algorithms, we are developing software infrastructure for continued research, as well as for eventual use in teaching.

I am also working on resolving theoretical questions about the expressive power of weighted formalisms [18] as they arise in our more applied research, and removing algorithmic obstacles, such as sum-product algorithms for nonprojective dependency representations [12].

4 Natural Language Processing Applications

While there are many interesting theoretical questions in NLP, it is at heart a branch of *applied* computer science, with the eventual goal of building useful applications, and I am actively contributing to several kinds of applications.

With Mengqiu Wang and Teruko Mitamura I recently developed a new approach to answer scoring based on syntactic representations and weighted grammars with rich features (including lexical semantics) in **question answering** [21]. This method far outperformed replicated competitive baselines. (The underlying grammar formalism, *quasisynchronous dependency grammar* was first proposed by David Smith and Jason Eisner.)

Kevin Gimpel and I have been using NLP on input data to **machine translation** to improve estimated phrase translation probabilities [7]. This gives substantial performance improvements in large-data scenarios when accurate NLP is available for the source language. Currently we are applying the quasisynchronous models we used in question answering [21] to machine translation. Use of these models for a *generation* task is especially challenging, because it requires us to find a translation among a very large set of possible grammatical derivations. Learning such a model from parallel corpora requires iterative training methods for very large datasets if the system is to be competitive (see §2), so this effort is closely connected to our efforts on improving NLP efficiency. My machine translation efforts are supported in part by a Google Grant, and I plan to seek NSF funding for this work with long-term collaborator David Smith (starting this fall as faculty at the University of Massachusetts). Further, in the immediate future I am seeking funding in collaboration with Stephan Vogel to develop open-source, distributed infrastructure for machine translation (on MapReduce architectures) through the NSF.

In a smaller side-collaboration, Danny Rashid and I used a straightforward language model and a simple model of relative key positions to build an “invisible” touch-typing interface that treats a touch-screen as a keyboard [10].

5 Grounding in Non-Linguistic Data

The advent of the world-wide Web has placed inconceivably large amounts of data within reach of many people, but the right information is rarely easy to find, consume, compile, or understand. Data is not information, and information is not understanding. Can we leverage ever-growing body of text to build tools that will help people to better understand the world and make more informed decisions? I am exploring ways to use statistical models of language to include *non-language* effects that can be observed in real-world data.

Predicting risk Together with Bryan Routledge (Tepper School of Business), Shimon Kogan (formerly at Tepper, now at U. Texas at Austin), and Jacob Sagi (Vanderbilt U.) and undergraduate Dimitry Levin, I have begun a preliminary study to use government-mandated annual financial reports published by publicly traded corporations to predict the volatility of those companies' stock prices (i.e., the standard deviation of the stock price over a fixed period of time, a measure of risk). We have found that a relatively simple text regression model can predict the volatility of the stock during the period after the report more accurately than the historical volatility, particularly reports published after the passage of the Sarbanes-Oxley Act of 2002, a congressional attempt to reform financial reporting after the accounting scandals that year [8]. This line of research shows promise both in the scientific question of how information affects human behavior and in the practical development of tools that will aid human decisionmaking.

Modeling political discourse Together with William Cohen and our student Tae Yano, I have begun studying text on political blogs. Text on blogs is different from text in news reports and government and scientific documents; it does not typically claim to report *facts*, and it varies widely in style and intent. We have been using unsupervised statistical modeling techniques to build predictors of how online communities will react (collectively and as individuals) to blog posts, as shown through comments left on the blog site. The goal of this work is to develop computational models of how language is used to convey non-factual information in a specific domain (politics), and how humans interact when communicating in this important new medium. This project has been supported by a competitive department-internal award; we plan to submit funding proposals to continue this work to the NSF (an initial proposal was highly rated but not successful).

References

- [1] Shay B. Cohen, Kevin Gimpel, and Noah A. Smith. In review (Neural Information Processing Systems).
- [2] Shay B. Cohen, Robert J. Simmons, and Noah A. Smith. In review (International Conference on Logic Programming).
- [3] Shay B. Cohen and Noah A. Smith. Joint morphological and syntactic disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 208–217, Prague, Czech Republic, June 2007.
- [4] Jason Eisner, Eric Goldlust, and Noah A. Smith. Compiling Comp Ling: Practical weighted dynamic programming and the Dyna language. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 281–290, Vancouver, BC, October 2005.
- [5] Jason Eisner and Noah A. Smith. Parsing with soft and hard constraints on dependency length. In *Proceedings of the International Workshop on Parsing Technologies*, pages 30–41, Vancouver, BC, October 2005.
- [6] Kevin Gimpel, Shay B. Cohen, Severin Hacker, and Noah A. Smith. In review (Conference on Empirical Methods in Natural Language Processing).
- [7] Kevin Gimpel and Noah A. Smith. Rich source-side context for statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 9–17, Columbus, OH, June 2008.
- [8] Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. In review (Conference on Empirical Methods in Natural Language Processing).

- [9] André F. T. Martins, Dipanjan Das, Noah A. Smith, and Eric P. Xing. In review (Conference on Empirical Methods in Natural Language Processing).
- [10] Daniel R. Rashid and Noah A. Smith. Relative keyboard input system. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 397–400, Canary Islands, Spain, January 2008.
- [11] David A. Smith and Noah A. Smith. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 49–56, Barcelona, Spain, July 2004.
- [12] David A. Smith and Noah A. Smith. Probabilistic models of nonprojective dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 132–140, Prague, Czech Republic, June 2007.
- [13] Noah A. Smith. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. PhD thesis, Johns Hopkins University Department of Computer Science, Baltimore, MD, October 2006.
- [14] Noah A. Smith and Jason Eisner. Annealing techniques for unsupervised statistical language learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 487–494, Barcelona, Spain, July 2004.
- [15] Noah A. Smith and Jason Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 354–362, Ann Arbor, MI, June 2005. **Nominated for best paper award.**
- [16] Noah A. Smith and Jason Eisner. Guiding unsupervised grammar induction using contrastive estimation. In *Proceedings of the International Joint Conference on Artificial Intelligence Workshop on Grammatical Inference Applications*, pages 73–82, Edinburgh, UK, July 2005.
- [17] Noah A. Smith and Jason Eisner. Annealing structural bias in multilingual weighted grammar induction. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics*, pages 569–576, Sydney, Australia, July 2006.
- [18] Noah A. Smith and Mark Johnson. Weighted and probabilistic context-free grammars are equally expressive. *Computational Linguistics*, 33(4):477–491, December 2007.
- [19] Noah A. Smith, David A. Smith, and Roy W. Tromble. Context-based morphological disambiguation with random fields. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 475–482, Vancouver, BC, October 2005.
- [20] Noah A. Smith, Douglas L. Vail, and John D. Lafferty. Computationally efficient M-estimation of log-linear structure models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 752–759, Prague, Czech Republic, June 2007.
- [21] Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. What is the Jeopardy model? a quasi-synchronous grammar for QA. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 22–32, Prague, Czech Republic, June 2007. **Nominated for best paper award.**