Research statement of Jonathan May

My research goal is to understand the computational mechanisms of human language. I want to use these mechanisms to build applications to generate, analyze, and transform natural language input. These applications range in scope from optical character recognition, to automatic translation, to brain-to-brain communication. A complete understanding of human language mechanisms will allow both replacement of tedious manual tasks with automated systems, and the enhancement of current human communication.

My Ph.D research focuses on improving the state of the art in natural language processing (NLP) systems from the unified framework of weighted finite-state tree transducers and automata. Previous work has demonstrated that complicated models of language processing can be effectively represented as cascades of weighted finite-state transducers [4]. String-based transducer toolkits enable rapid development of systems that perform tasks such as transliteration, phrase-based translation, speech recognition, and optical character recognition. These tasks are all performed by generic operations over transducer cascades.

My work uses the formalism of tree automata to capture current work in NLP that is focused on tree-based syntactic structure. To do this I designed an algorithm for determinization of weighted tree automata [5] and co-designed algorithms for tree automata minimization with colleagues in Germany and Sweden [1, 2]. I demonstrated the applicability of this formalism, along with a novel algorithm for unsupervised EM training of tree transducers and automata, by replicating a model of syntactic machine translation in a tree transducer framework [3]. I also applied the framework and learning algorithms in the creation and training of a new model of syntax machine translation. Using this new model I obtained statistically significant improvements in Arabic-English and Chinese-English translation over a state-of-the-art system baseline [7].

As a consequence of this work I produced Tiburon, a finite-state tree automata and transducer toolkit [6]. Tiburon includes the above algorithms as well as implementations of other useful algorithms such as k-best item generation, Earley parsing, and transducer composition. In addition to its use in research, Tiburon is used to effectively teach empirical NLP in the classroom. To date, it has been downloaded over 400 times in 31 countries and used in at least four separate research projects, not counting my own.

We can improve our current models of language by a cycle of error-analysis followed by model adjustment. Using Tiburon as a common framework for competing models accelerates this cycle and clarifies the analysis task. We can also easily extend a model in new directions, simply by changing the rules and states of the representative transducer. In future work, I will replicate competing models of translation and language modeling with the express purpose of detailed analysis, and use that analysis in rapid consideration of new models. I will then apply the same learning methods and models to large data sets in non-linguistics domains such as protein folding, musical composition analysis, and athletic performance prediction.

While my primary research goals are in the applied field of NLP, I intend to keep a foot in theory as well. My colleagues in formal language studies are very interested in multi-site collaboration to develop algorithms with a clear empirical task in mind that benefits both fields. Andreas Maletti and I plan to continue research into appropriate formalisms and corresponding algorithms that go beyond the tree automata work from my thesis. Johanna Högberg and I plan to explore grammar induction research via investigations into learnability theory.

I am emboldened in my claims by evidence that there is support in the research funding environment for this type of work. The recent NSF Cyber-Enabled Discovery and Information solicitation indicates a strong desire for concentrated effort on exploration of the best models to exploit learning in large data sets. High-risk proposals for completely new models become less risky when they can be prototyped in a day and begin returning results in a matter of weeks. And a solid background in theory is essential to deeply understanding the computational power of proposed formalisms and the viability of desired operations.

Ultimately, I hope to apply the models and techniques learned via this research to the next great unexplored corpus — that of the brain. The data produced by the brain as it processes and generates language are still largely unknown to us due to difficulties in measurement. In the long view these are engineering challenges to be solved in the world of neuroscience, and there is current, active study in this field. The same combinations of model and data that led to great advances in translation and speech recognition today may well be applied to the brain-level intercept tomorrow, ushering in a new world of post-telephonic communication and effective treatment for autistic and otherwise communication-impaired patients. Such accomplishments are still speculative, but it is important to begin the research efforts on the data that does exist, so that when brain corpora become more robust we can make rapid advances.

References

[1] Johanna Högberg, Andreas Maletti, and Jonathan May. Backward and forward bisimulation minimisation of tree automata. In Jan Holub and Jan Žďárek, editors, *Proc. 12th Int. Conf. Implementation and Application of Automata*, volume 4783 of LNCS, pages 109–121, Prague, 2007. Springer.

[2] Johanna Högberg, Andreas Maletti, and Jonathan May. Bisimulation minimisation for weighted tree automata. In Tero Harju, Juhani Karhumäki, and Arto Lepistö, editors, *Proc. 11th Int. Conf. Developments in Language Theory*, volume 4588 of LNCS, pages 229–241, Taipei, 2007. Springer.

[3] Kevin Knight, Jonathan Graehl, and Jonathan May. Training tree transducers. *Computational Linguistics*, 34(3):391–427, September 2008.

[4] Kevin Knight and Jonathan May. Applications of weighted automata in natural language processing. In Manfred Droste, Werner Kuich, and Heiko Vogler, editors, *Handbook of Weighted Automata*. Springer-Verlag, 2009.

[5] Jonathan May and Kevin Knight. A better n-best list: Practical determinization of weighted finite tree automata. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 351–358, New York City, USA, June 2006. Association for Computational Linguistics.

[6] Jonathan May and Kevin Knight. Tiburon: A weighted tree automata toolkit. In Oscar H. Ibarra and Hsu-Chun Yen, editors, *Proceedings of the 11th International Conference of Implementation and Application of Automata, CIAA 2006*, volume 4094 of *Lecture Notes in Computer Science*, pages 102–113, Taipei, Taiwan, August 2006. Springer.

[7] Jonathan May and Kevin Knight. Syntactic re-alignment models for machine translation. In Jason Eisner and Taku Kudo, editors, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 360–368, Prague, Czech Republic, June 28 – June 30 2007. Association for Computational Linguistics.