Manning: NSF award IIS-1514268 $1,100,000.00. June 1, 2015–May 31, 2018; no cost extension to May 31, 2019. RI: Medium: Deep Understanding: Integrating Neural and Symbolic Models of Meaning. (Dan Jurafsky, PI; Christopher Manning; Percy Liang)

Intellectual Merit: The results of our project have already had a large impact on our ability to use combinations of neural and symbolic approaches to do semantic understanding of human language, including novel models of discourse coherence, models that combine text and networks, novel models of interpretability of embeddings and neural models, new ways of combining adversarial learning with text processing, and novel neural models for conversational dialogue. Particular achievements include the following. We introduced two deep learning algorithms for the task of developing lexicons that predict outcomes like human preferences or actions from text as a way of making these models transparent and interpretable. We investigated the role of linguistic context in an LSTM language model, through ablation studies, analyzing the increase in perplexity when prior context words are shuffled, replaced, or dropped, in order to understand how neural models are able to use much longer context than traditional language models in NLP. By using neural models that combine language and social networks to investigate conflict in communities online, we show that such conflicts tend to be initiated by a handful of communitiesless than 1% of communities start 74% of conflicts. We developed 2 new domain-independent neural models of discourse coherence that are capable of measuring multiple aspects of coherence, a discriminative model that learns to distinguish coherent from incoherent discourse, and generative models that produce coherent text, including a novel neural latent-variable Markovian generative model that captures the latent discourse dependencies between sentences in a text. Our work achieves state-of-the-art performance on multiple coherence evaluations. We addressed the important problem of generating discourse-coherent utterances in dialog by combining reinforcement learning and adversarial training: the system was trained to produce sequences that are indistinguishable from humangenerated dialogue utterances.

Broader impacts: The project has trained 17 graduate students, postdocs, and undergraduates, including 6 women, who have received weekly mentoring from the PIs in various phases of the project, with training in research methodology, in career development, and in the research content described above. One woman postdoc from this project, Yulia Tsvetkov, has now begun her faculty career at CMU. Three graduate students have graduated: Will Hamilton is starting a faculty job at McGill this year. Raine Hoover received her MS and Jiwei Li received his PhD and both are now at startups. Finally, the undergraduate, Jon Gauthier, graduated and is now a PhD student at MIT.

The algorithms have been described in publications and talks to the community and in talks given by the PIs to computer science departments around the country. Several pieces of code have been incorporated into the Stanford

CoreNLP software, which is used by research groups and companies around the country.

Publications: Understanding neural models: Li et al. (2015); Khandelwal et al. (2018); coreference: Clark and Manning (2016a,b); summarization: See et al. (2016); neural language understanding: Bowman et al. (2016); Hamilton et al. (2016a,b); events: Huang et al. (2016); dialogue: Li et al. (2017); Muzny et al. (2017); Guu et al. (2017); neural sequence translation models: Luong and Manning (2016); Wuebker et al. (2016).

Research products: The algorithms have been described in publications and described in talks to the community and in talks given by the PIs to computer science departments around the country.

Several systems have been incorporated into the Stanford CoreNLP open source NLP software, which is widely used by research labs and companies around the country, including in particular Kevin Clark's statistical and neural coreference systems: `https://stanfordnlp.github.io/coref.html` and Grace/Felix Muzny's quote annotator: `https://stanfordnlp.github.io/quote.html`.

Several datasets have been made publicly available, including: The SCONE dataset for learning models of context-dependent executable semantics at: `https://nlp.stanford.edu/projects/scone/`; the network model of Reddit interaction at: `http://snap.stanford.edu/data/web-RedditNetworks.html`; and quote attribution data: `https://stanfordnlp.github.io/quote.html`.

References

Bowman, Samuel R., Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Association for Computational Linguistics (ACL)*.

Clark, Kevin, and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Empirical Methods on Natural Language Processing*.

Clark, Kevin, and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Association for Computational Linguistics (ACL)*.

Guu, Kelvin, Panupong Pasupat, Evan Zheran Liu, and Percy Liang. 2017. From language to programs: Bridging reinforcement learning and maximum marginal likelihood. In *Association of Computational Linguistics (ACL)*.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Association for Computational Linguistics (ACL)*.

Huang, Ruihong, Ignacio Cases, Dan Jurafsky, Cleo Condoravdi, and Ellen Riloff. 2016. Distinguishing past, on-going, and future events: The eventstatus corpus. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Khandelwal, Urvashi, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Association for Computational Linguistics (ACL)*.

Li, Jiwei, Minh-Thang Luong, Dan Jurafsky, and Eudard Hovy. 2015. When are tree structures necessary for deep learning of representations? In *Empirical Methods in Natural Language Processing (EMNLP)*.

Li, Jiwei, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Luong, Minh-Thang, and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Association for Computational Linguistics (ACL)*, Berlin, Germany.

Muzny, Grace, Michael Fang, Angel X. Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

See, Abigail, Minh-Thang Luong, and Christopher D. Manning. 2016. Compression of neural machine translation models via pruning. In *Computational Natural Language Learning (CoNLL)*.

Wuebker, Joern, Spence Green, Saša Hasan John DeNero, and Minh-Thang Luong. 2016. Models and inference for prefix-constrained machine translation. In *Association for Computational Linguistics (ACL)*, Berlin, Germany.