Introduction. Tasks where the target output is natural language, such as translation, summarization, and dialogue, are the basis of some of the most broad reaching applications in natural language processing (NLP). As we begin to apply expressive, data-driven models to these tasks, evaluation has become a central methodological concern. My research interest is in the practical application of natural language understanding (NLU) to evaluate generation (NLG), and more generally within both NLG and NLU.

In this direction, I evaluated the syntactic properties of language generated from sequence-to-sequence (seq2seq) models by parsing with a rule-based grammar of English. Current work aims to develop MT metrics based on semantic parsing, which has implications for NLG and NLU. My objective is to pursue a Ph.D. in the UW NLP group to complement the expertise in language generation and its evaluation, language understanding, and semantic parsing where I may pursue this project while developing new projects within these fields. This independent research demonstrates my technical skills to pursue future work.

Previous research. My most recent work is in the general research direction of deep learning interpretability, specifically for seq2seq language generation models.[1] [6] Seq2seq models have the flexibility to output any sequence of tokens, and are increasingly used in tasks where the target output is natural language. However, these models are not explicitly designed to generate with respect to syntax or semantics, and the properties of the language generated from these models are poorly understood.

My research demonstrates the potential of evaluating language generation models in the setting of a language-like, rule-based, HPSG grammar, which provide several advantages over existing literature. Previous methods either evaluated models in artificial language settings to deeply understand the model, or with real language on constrained linguistic phenomena to gain a shallow understanding. My methodology strikes a balance on this spectrum - we can have a non-shallow understanding of our model on nearly real language, all while evaluating output it produces in practice.

Current direction. One main challenge of my research in neural network analysis/interpretability is in providing value to other areas. For this reason, my current efforts are in developing machine translation (MT) metrics (e.g. BLEU [5]) in the WMT'17 metrics shared task setting. [3] MT metrics assign scores to a machine translation based on the reference, which aim to correlate with human judgment of translation quality. Successful research in metrics would allow systems to be readily compared and deployed. Consequently, automatic metrics have the potential to accelerate development for an entire field.

Refer to my NSF GRFP proposal (2 pages).[2] Central to my metric is the semantic parsing of machine and reference translations. By comparing graph-based meaning representations of a machine and reference translation segment, it may be possible to achieve high sentence-level correlation with human judgments (as opposed to BLEU, which only achieves high correlation on the system level by aggregating scores over a test set). Existence of such a metric raises research questions with implications to both NLG and NLU:

- Can this metric based on semantic parsing be adopted for other NLG domains? The notion of a general purpose NLG metric is appealing, and BLEU has, at times inappro-

[1] Available at: `http://aclweb.org/anthology/W18-5432`

[2] Available at: `https://johntzwei.github.io/pdfs/NSF_GRFP_Research_Proposal_final.pdf`

priately, been used in various domains. With deep linguistic features from semantic parsing, can we generalize to other domains (e.g. image captioning, dialogue, etc.)?

- How may we effectively utilize fine-grained error analysis? Such a metric will be able to provide detailed error analysis of each test example by aligning semantic representations of the reference and machine translations, and presenting the deviations.
- Can this metric be used as a training signal for seq2seq? Previous research on alternative objectives were limited to metrics that are noisy on sentence-level scoring.
- Can this metrics setting be used to compare semantic formalisms? There are many competing graph-based semantic formalisms (AMR, MRS, SDP, etc.) that can be features for an MT/NLG metric. This setting tests a formalism's adequacy of linguistic description and the effectiveness of parsers and their training resources.

Department fit. To further my metrics research, I hope to work with Prof. Hajishirzi and and Prof. Choi to complement their expertise in language understanding and generation, respectively. During this work, I would also like to collaborate with Prof. Zettlemoyer and students working in semantic parsing, and Prof. Bender to explore the evaluation of semantic formalisms. I am also interested in generally contributing to both NLG and NLU. The UW NLP group span these areas, where I hope to begin new threads of research.

Technical skills. My independent research highlights some of the technical abilities that I will carry forth as a graduate student. They are listed below:

- My coursework consists of high-level undergraduate courses in pure mathematics, statistics, linguistics, psycholinguistics, and basic courses in computer science. This helps me understand and implement new concepts in NLP.
- Deep learning in NLP and its analysis/interpretability was introduced to me during my summer at UW. In this area, I have conducted literature review, and psycholinguistics gives me perspective. I am able to formulate research questions and experiments.
- My interest and exposure to parsing technologies was developed while working on parsing and language identification for dialectical English. [1, 2] I am familiar with the application of parsing, which is the focus of my current work.
- Technical skills to parallelize jobs on computing clusters were acquired when I dealt with 200GB+ bibliography datasets and collected search results at scale. [4] I am able to run large computationally intensive parsing jobs, which my recent work [6] required.

Purpose. Throughout my future studies at UW, my first hope is to contribute to tasks requiring capability in both NLG and NLU, which are the basis of broad reaching applications in NLP. My second hope is to mentor undergraduates. In my undergraduate career, I have taken on mentoring and teaching positions, and hope to continue as a graduate student. My goal after completing a Ph.D. is to pursue a career in research and teaching. As faculty, I hope to focus on scientific outreach to undergraduates, both through involvement in my own research and through university outreach programs. Thank you for your consideration!

References

[1] Blodgett, S. L., **Wei, J.**, O'Connor, B., "A Dataset and Classifier for Recognizing Social Media English". In: *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*. Association for Computational Linguistics, 2017, pp. 56–61. URL: https://aclanthology.info/papers/W17-4408/w17-4408.

[2] Blodgett, S. L., **Wei, J.**, O'Connor, B. T., "Twitter Universal Dependency Parsing for African-American and Mainstream American English". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Association for Computational Linguistics, 2018, pp. 1415–1425. URL: https://aclanthology.info/papers/P18-1131/p18-1131.

[3] Bojar, O., Graham, Y., Kamran, A., "Results of the WMT17 Metrics Shared Task". In: *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*. Association for Computational Linguistics, 2017, pp. 489–513. URL: https://aclanthology.info/papers/W17-4755/w17-4755.

[4] Chang, H., Abdurrahman, M., Liu, A., **Wei, J. T.-Z.**, Traylor, A., Nagesh, A., Monath, N., Verga, P., Strubell, E., McCallum, A., "Extracting Multilingual Relations under Limited Resources: TAC 2016 Cold-Start KB construction and Slot-Filling using Compositional Universal Schema". In: *Proceedings of TAC* (2016).

[5] Papineni, K., Roukos, S., Ward, T., Zhu, W., "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pp. 311–318. URL: http://www.aclweb.org/anthology/P02-1040.pdf.

[6] **Wei, J.**, Pham, K., O'Connor, B., Dillon, B., "Evaluating Grammaticality in Seq2seq Models with a Broad Coverage HPSG Grammar: A Case Study on Machine Translation". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 298–305. URL: http://aclweb.org/anthology/W18-5432.