Zhaofeng Wu zfw7@cs.washington.edu

The natural language processing (NLP) community has recently seen an exponential increase in the size of language models (LMs), already stepping into the trillion-parameter realm (Fedus et al., 2021). Despite their "superhuman" performance on many benchmarks, they are often brittle in domain and compositional generalization. Moreover, the correlation of model size and quality leads to an inflated cost of high-quality models, hurting the inclusiveness in NLP. These issues motivate my two primary research interests: **to better analyze and improve the generalizability and interpretability of NLP models, and to increase their efficiency to be more accessible and green**. I am fortunate to have conducted some initial research in these areas under the guidance of Prof. Noah Smith and Prof. Fei Xia alongside many wonderful collaborators, and I am passionate to continue studying these problems in my Ph.D.

Generalizability and Interpretability

My training in linguistics has led me to appreciate generalizable and interpretable theories and systems that robustly predict with arbitrary linguistic input consistently with human judgments, and with a traceable reasoning process such that they can be easily amended when the predictions are wrong.

Current models, however, usually work best on well-edited Wikipedia text in Standard American English and fail to generalize to new domains, languages, and compositional structures. I grew interested in generalization when I **discovered pretrained LMs' significant degradation in the biomedical domain** for the text entailment task (Wu et al., 2019; BioNLP Workshop @ ACL). In response, I proposed unsupervised LM-finetuning using documents in the same domain, which helped bridge the gap. I also intend to contribute to **developing representations that can easily adapt to many languages** with little supervision. Existing multilingual models show substantially variable adaptability for languages with various amounts of resources (Wu and Dredze, 2020), and they do not generalize well to the world's thousands of languages unseen during pretraining (Chau et al., 2020). While data availability affects adaptability, studies have also shown that even unsupervised models exhibit biases toward features of English (Dyer et al., 2019). I am curious to explore efficient models and training methods that can robustly adapt to low-resource or even unseen languages by capturing some notion of "linguistic universal."

The opacity of deep models also obscures a thorough understanding of their true linguistic capability and generalizability. Motivated by this challenge, I proposed to **analyze the mainstream coreference model**, a complex task that still requires task-specific architectures in addition to pretrained transformers (Wu and Gardner, 2021; CRAC Workshop @ EMNLP). I designed analyses such as oracle experiments to inspect the interaction of model components. They allowed us to identify factors key to a high performance yet often overlooked, such as mention detector precision and anaphoricity decisions.

Having seen how model analysis can reveal its shortcomings and lead to possible improvements, I proposed and am leading an ongoing project to **study the limits of current LM objectives** that underlie prevalent pretrained models. An often-unstated assumption of these objectives is that meaning can be acquired via training on linguistic forms alone, whose possibility has been challenged (Bender and Koller, 2020; Traylor et al., 2021). Nevertheless, Merrill et al. (2021) proved its possibility on strongly transparent formal languages.[1] Curious to see how their theoretical analyses realize empirically, I investigate if existing LM objectives attain this capability. Inspired by my linguistics background, I also study their understanding of "referential opacity,"[2] a closely related *natural language* phenomenon. Our early results suggest that current LM objectives can only capture the meaning of restricted classes of formal languages but not natural language with its full complexity. I am excited to further probe the limits of LMs and leverage the key language properties revealed in this study to ground LMs to expand their capacity.

To learn more generalizable and interpretable representations, I am interested in **leveraging linguistic structures as a useful inductive bias**. I took an initial step in this direction in Wu et al. (2021b; TACL). While studies had found the emergence of syntax in pretrained LMs (Hewitt and Manning, 2019), inspired by my learning in linguistics, I was curious to **inspect LMs' knowledge of semantics**, which is closer to language understanding. Through a controlled probe, I discovered that they significantly less readily surface predicate-argument semantics than syntax. Furthermore, by explicitly incorporating semantic structures, I improved their task performance and demonstrated greater compositional generalizability.

[1] That is, all valid expressions in the language have context-independent denotations.

[2] That is, contexts that yield *different* truth-conditions when embedding expressions that co-refer to the same entity.

In contrast to using a standalone parser in this project, I am also interested in learning with latent structures which does not require expert annotation, is less susceptible to cascading errors, and has yielded superior performance, mostly with previous RNN models (Yogatama et al., 2017; Bisk and Tran, 2018, *i.a.*). This paradigm, however, is often complicated by the discreteness of structured prediction. To equip myself with more theoretical knowledge to work with structures, I studied substantial related work and **wrote a survey on latent structure learning**, categorizing many approaches to circumvent this optimization obstacle (Wu, 2021). This knowledge enables me to synthesize ideas to facilitate, on top of recent models, more robust compositional and domain generalization by avoiding the over-reliance on surface-level correlations and enhanced interpretability by offering inspectable structures.

Efficiency

I still remember my shock seeing the wide internal availability and adoption of TPUs while interning at Google, and when my ~~Facebook~~ Meta colleagues said it would be easy to human-evaluate the image captioning system that I built there. Indeed, recent NLP models have led to drastically increased financial burden, inequity, and carbon footprint to participate in NLP research and to use NLP technology.

It can be prohibitive for researchers to finetune massive pretrained models with modest resources. By training only a few embeddings, prompt tuning has become an efficient alternative to full model tuning. Despite its advantages, its performance still sometimes trails behind traditional finetuning. This gap drives me to lead another ongoing project that **uses meta-learning to find a more instructable model** that is efficient and can robustly generalize to new tasks. Massive NLP models also lead to inaccessible consumer technologies. Though model shrinking methods exist, maintaining multiple shrunk models for different device capacities is an expensive overhead. To more efficiently leverage resources on edge devices, I **proposed dynamic sparse models** for Google Assistant where a single trained model can adjust its sparsity at runtime according to resource availability (Wu et al., 2021c; ICASSP). These past works of mine can and have already contributed to the accessibility of NLP technology.

Another source of inefficiency is the models' asymptotic complexity. Though it is the backbone of most state-of-the-art models, transformers consume quadratic time and memory and are challenging to apply to long sequences such as documents. In Wu et al. (2021a), I proposed to leverage random feature attention (Peng et al., 2021b), a linear time and space approximation to the transformer, and **accelerated document-level machine translation** by up to 40% while improving the BLEU score. In Peng et al. (2021a), we **unified efficient transformer variants** under a single abstraction by viewing its attention module as accessing a bounded memory. This abstraction allowed us to introduce a learned memory organization that offers a better quality-efficiency trade-off and enables better interpretability.

My interest in incorporating structural inductive biases into NLP models also faces efficiency challenges. For example, many inference algorithms for structure learning are expensive, such as the cubic-time CKY algorithm, if not intractable. This inefficiency often precludes large-scale pretraining with latent structures. I am excited to address these challenges by seeking methods to **efficiently combine recent advances in representation learning with previous structurally-informed models** such as (Unsupervised) Recurrent Neural Network Grammars (Dyer et al., 2016; Kim et al., 2019) and Structured Attention Networks (Kim et al., 2017). This direction, along with prompt tuning and many other methods, has the potential to **improve generalizability and interpretability efficiently**, all of which I would be passionate to pursue in my Ph.D. study, equipped with my past knowledge and experiences.

Having followed the work of the NLP groups at MIT, I believe it would be an ideal place for me to continue my exploration. I am interested in the research of Prof. Jacob Andreas on compositionality and language grounding and Prof. Yoon Kim on structured models. I believe structurally-inspired methods can contribute to improved compositionality and grounding in NLP models. It would also be constructive to work with faculty including Prof. Regina Barzilay, Tommi Jaakkola, and others. With an interdisciplinary background, I also firmly believe in the value of collaborating with and learning from adjacent communities at MIT, such as the Center for Brains, Minds and Machines and the Computational Cognitive Science group. I hope MIT can be the launchpad for my journey in further exploring and contributing to NLP research.

References

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proc. of ACL*, 2020.

Yonatan Bisk and Ke Tran. Inducing grammars with and for neural machine translation. In *Proc. of the 2nd Workshop on Neural Machine Translation and Generation*, 2018.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of EMNLP*, 2020.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In *Proc. of NAACL*, 2016.

Chris Dyer, Gábor Melis, and Phil Blunsom. A critical analysis of biased parsers in unsupervised parsing. *arXiv*, 2019.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv*, 2021.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proc. of NAACL*, 2019.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In *Proc. of ICLR*, 2017.

Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. Unsupervised recurrent neural network grammars. In *Proc. of NAACL*, 2019.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 2021.

Hao Peng, Jungo Kasai, Nikolaos Pappas, Dani Yogatama, **Zhaofeng Wu**, Lingpeng Kong, Roy Schwartz, and Noah A. Smith. ABC: Attention with bounded-memory control. 2021a. **Currently Under Review**.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *Proc. of ICLR*, 2021b.

Aaron Traylor, Roman Feiman, and Ellie Pavlick. AND does not mean OR: Using formal languages to study language models' representations. In *Proc. of ACL*, 2021.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proc. of the 5th Workshop on Representation Learning for NLP*, 2020.

Zhaofeng Wu. Learning with latent structures in natural language processing: A survey. *arXiv*, 2021.

Zhaofeng Wu and Matt Gardner. Understanding mention detector-linker interaction in neural coreference resolution. In *Proc. of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, 2021.

Zhaofeng Wu, Yan Song, Sicong Huang, Yuanhe Tian, and Fei Xia. WTMED at MEDIQA 2019: A hybrid approach to biomedical natural language inference. In *Proc. of the 18th BioNLP Workshop and Shared Task*, 2019.

Zhaofeng Wu, Hao Peng, Nikolaos Pappas, and Noah A. Smith. Modeling context with linear attention for scalable document-level translation. 2021a. **Currently Under Review**.

Zhaofeng Wu, Hao Peng, and Noah A. Smith. Infusing Finetuning with Semantic Dependencies. *Transactions of the Association for Computational Linguistics*, 2021b.

Zhaofeng Wu, Ding Zhao, Qiao Liang, Jiahui Yu, Anmol Gulati, and Ruoming Pang. Dynamic sparsity neural networks for automatic speech recognition. In *Proc. of ICASSP*, 2021c.

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. Learning to compose words into sentences with reinforcement learning. In *Proc. of ICLR*, 2017.