Research Statement

Yuekun Yao

May 13, 2024

My research uses machines to learn human intelligence through the lens of natural language processing (NLP). The main question I am interested in is *how to develop human-like cognitive capacities for machines to understand natural languages.* I investigate this question from both learning and evaluation perspectives. For learning, humans can understand and produce languages rapidly from learning a small amount of sentences of the language. To understand such capacities of machines, I study two questions: (1) *Are machines the same data-efficient learners for natural languages as humans?* and (2) *If not, how to make the language learning process of machines more effective?* For evaluation, humans review, evaluate and correct their solution for a given problem. Hence, I investigate (3) *How to use machines to perform evaluation on solutions generated by themselves?*

Learning languages in a data-efficient way. Humans are data-efficient language learners. Learning a small amount of sentences, they can understand and produce infinite sentences. This is due to the compositionality principle of human languages, which allows humans to understand the meaning of a sentence as a function of meanings of its parts and how they are combined. For example, knowing the meaning of *John loves the girl* immediately enables a speaker to understand the meaning of *the girl lover John.* Are machines also aware of the compositionality when learning languages?

My work [2] in this area focuses on studying compositional generalization, an ability to generalize to complex sentences after being trained on simple ones. To answer this question, I investigated the compositional generalization ability of sequence-to-sequence (seq2seq) and structured models across tasks that require different levels of understanding of languages (e.g. part-of-speech tags, syntax and semantics). We found that across different tasks, generalizing to more complex linguistic structures is consistently challenging for seq2seq but not structured models. This suggests that these models are still not effective enough to learn the underlying rules or structures of human languages, since they only understand sentences similar to what they encountered before. We also investigated the reason behind their failures and introduced multiple methods to incorporating human-like structural bias into the learning process. The results revealed that the failure is due to the inability of the decoder to predict complex structures, and this issue persists even with our introduced methods. This ineffective learning is also observed in LLM in another collaborated work

[1]. Hence, I regard how to make models effectively learn human languages in a data-efficient way as a key step to achieve human intelligence with machines.

Effective learning with general-purpose architectures. Due to the findings in [2, 1], it is natural to ask how to improve the model ability to learn human languages. Among different ways to achieve this, I am particularly interested in methods with general-purpose architectures (e.g. transformers [5]) as the backbone and focuses on improving strategies of training (e.g. training data, loss) or inference (e.g. decoding).

My research in this track still focuses on improving model performances on compositional generalization tasks. As mentioned in [2], I already explored informative features or training losses as the first step to approach the issue. Additionally, I employed data augmentation to improve model effectiveness in [4]. In this work, I adopted existing or hand-crafted probabilistic context-free grammars to sample unlimited targeted symbolic sequences (e.g. meaning representations) as additional training data. The sampled sequences are then back-translated into natural language sentences to be learned for seq2seq models. Our results show that this sample-and-backtranslate strategy is a useful method to improve the generalization ability on complex structures, especially when the grammar is able to cover most local structures in the targeted symbolic sequences. This suggests that controlling the complexity or difficulty of training data serves as a promising way to improve the effectiveness of learning process.

Self evaluation. When humans write or solve a problem, they usually generate a draft or strategy first, and then iteratively evaluate and refine their generations by themselves. Such ability allows humans to understand the problem better and thus avoid errors made in previous trials.

My research in this area mainly focuses on evaluating the model generations with the model itself [3]. Specifically, For existing models (e.g. parser) trained on a particular task, I trained an additional neural network (e.g. discriminator) to judge the correctness of the parser on unseen inputs across different out-of-distribution generalization tasks. By aggregating decisions from multiple discriminators with novel ensemble mechanisms, we predict the upper and lower bounds of the model accuracy on the test set. Our results show that the bounds predicted by discriminators can accurately capture realistic accuracy on the test set. This suggests that neural networks do have the capacity to judge their errors on unseen data by learning from errors in previous trials (e.g. training data). How to make the model effectively learn from errors and refine them, is still an open question for future work.

References

[1] Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, **Yao, Yuekun**, and Najoung Kim. SLOG: A structural generalization benchmark for semantic parsing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Lan-*

guage Processing, pages 3213–3232, Singapore, December 2023. Association for Computational Linguistics.

[2] **Yao, Yuekun** and Alexander Koller. Structural generalization is hard for sequence-to-sequence models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5048–5062, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[3] **Yao, Yuekun** and Alexander Koller. Predicting generalization performance with correctness discriminators, 2023.

[4] **Yao, Yuekun** and Alexander Koller. Simple and effective data augmentation for compositional generalization, 2024.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.