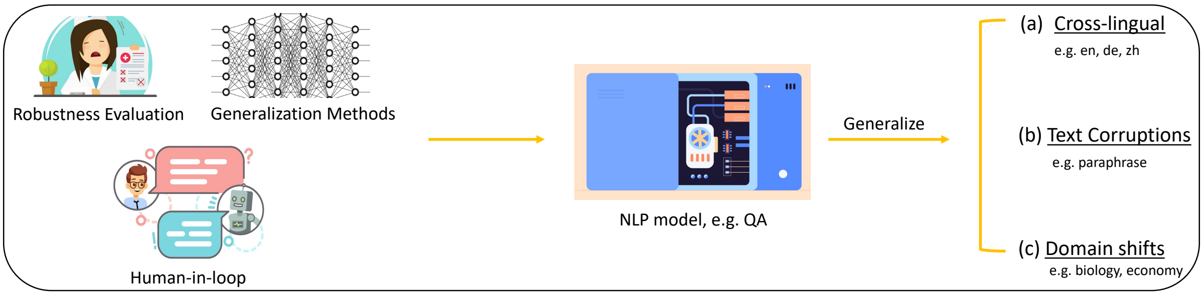
Research Statement

Hai Ye
National University of Singapore
hye.me@outlook.com

My desired research direction lies in trustworthy machine learning and its development for Natural Language Processing (NLP). Trustworthy machine learning is a broad term which studies the topics of explainable, fair, privacy-preserving, causal and robust. From these topics, I would pay much attention to ***robustness***. *In general, I am aiming to build NLP systems with focusing on not only the accuracy but more importantly the robustness.* In NLP, I would care more about ***Natural Language Understanding*** (NLU) tasks including Question Answering and Semantic Parsing. The models are built with large-scale pre-trained language models. To achieve my research goal, I would explore the following more specific problems:



- **Robustness Evaluation**: How can we effectively find out the weakness of modeling methods for certain tasks? How the robustness of the model is to text corruptions? Corruptions are applied to the model text input such as the grammatical errors, paraphrase, substitutions and etc.
  - **Task-oriented Corruptions.** Surrounding interested tasks such as question answering, find out certain task-specific corruption types to attack the model. I will focus on conversational question answering systems (CQAs) which understand dialogue to answer questions, and semantic parsing models that translate human questions into executable programs for computers to understand.

    Under this topic, I have one work on analyzing question rewriting systems (sub-task of CQA) to be submitted to ACL 2022 (Ye et al., 2021a).
  - **Automatic Corruption Discovery.** It is time-consuming for humans to design and test corruption types to attack certain tasks or models. And there is a concern whether the found corruptions can cover new data/domain in the future. Considering the two issues, I would explore how to automatically capture and generalize the corruption types mined from the data on which the model fails in the test time, through which, we can also cheaply enhance the model robustness to newly found corruptions by applying these corruptions to do data augmentation on the original training data and updating the model on the augmented data.
- **Domain Generalization**: How do we improve the model generalization ability to deal with long-tail phenomenon/ out-of-domain data? Here, I mainly care about three generalization aspects in NLP

which are cross-lingual, text corruptions, and the shifts between similar domains. I have studied unsupervised domain adaptation in this topic (Ye et al., 2020).

  - **Life-long/Online Model Adaptation.** I would like to explore how to adapt the model to new domains quickly and cheaply in the dynamic environment after model deployment. For domain generalization, we cannot always know the data distributions counted in the test time, so we have to actively adapt the model whenever new domains come. In the test time, the data can arrive in a stream manner. And accordingly, the model should be adapted in an online and life-long manner. How do we quickly and effectively adapt the model also in an online style?
  - **Restricted Domain Adaptation.** The data for building source model cannot always be available because of the privacy issue. And the data in the target domain can be very limited. By considering the issue of data privacy or low resource, how to effectively make domain adaptations? Specific topics can be studied are domain adaptation without source data and model personalization (adapt the central model to meet personalized requirements).
  - **New Benchmarks.** I would focus on creating more realistic and challenging benchmarks for studying domain generalization, such as test-time domain adaptation which cares more about the type of test data coming in an online manner.

- **Human-in-Loop.** How do we involve humans to assist the model to make more reliable decisions? And how to create proper interactions between the human and the model?

  - **Prediction Loop.** The first dimension of human-in-loop happens when the model making predictions, but how to design smoothing human-model interaction mechanisms for certain tasks. I have interest in studying interactive semantic parsing, for example, when the user asks the computer one ambiguous question, how the computer can feedback to the user to request more detailed information of the question.
  - **Adaptation Loop.** When the model needs to be adapted to deal with out-of-domain data, how the human can help to adapt the model by using human annotations in a low cost way by methods such as active learning?

Beside the above topics, I used to focus on methods to learn models with low-resource data (Ye et al., 2019; Ye and Wang, 2018; Ye et al., 2017, 2020; He* et al., 2021). I would also like to explore other directions under the big topic of trustworthy machine learning such as privacy and causal, and interesting ML problems such as learning with noisy labels (Ye et al., 2021b). And any other applications except for NLP can also become my research interest, such as legal systems (Ye et al., 2018).

Though the above is what I am interested currently, I can quickly switch into any other research topics in the future if needed.

References

Ruidan He*, Linlin Liu*, Hai Ye*, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of ACL. (* indicates equal contribution).*

Hai Ye, Wenhan Chao, Zhunchen Luo, and Zhoujun Li. 2017. Jointly extracting relations with class ties via effective deep ranking. In *Proceedings of ACL.*

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of NAACL-HLT.*

Hai Ye, Wenjie Li, and Lu Wang. 2019. Jointly learning semantic parser and natural language generator via dual information maximization. In *Proceedings of ACL.*

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of EMNLP.*

Hai Ye and Lu Wang. 2018. Semi-supervised learning for neural keyphrase generation. In *Proceedings of EMNLP.*

Hai Ye et al. 2021a. title. In *submission*.

Hai Ye et al. 2021b. title. In *submission*.