Statement of Purpose

Van Quyet DO (vqdo@connect.ust.hk)

My professional goal is to become a researcher who can make valuable contributions to advancing the capabilities of computers in understanding, generating, and manipulating human language. This aspiration has motivated me to pursue a doctoral program in Computer Science, specializing in Natural Language Processing (NLP). My areas of research focus encompass neural-symbolic knowledge representation and reasoning systems.

My interest in neural-symbolic systems derives from the existing drawbacks of using the current framework of NLP in achieving Artificial General Intelligence (AGI). The study of NLP has witnessed significant advancements in recent times, leading to anticipation that machines could achieve AGI within the coming decade. Nevertheless, the existing development framework of NLP models for achieving AGI has sparked several discussions and criticisms, supported by empirical evidence. Two key concerns that have been raised are the limited explainability of these models and the large data-computation consumption. First, regarding models' explainability, a majority of the State-of-the-Art models are pre-trained and deployed solely in an autoregressive manner. It virtually lets language models (LM) memorize all of the data rather than connect pieces of knowledge to achieve understanding, as opposed to humans. In contrast, a neural-symbolic approach can emulate human thoughts by, for instance, structuring models' knowledge as a graph which apparently serves the learning and reasoning process as in the human mind, thus enhancing the explainability of language models. Second, in regards to data-computation consumption, the training process of current large language models (LLM) poses a large demand for data and computation. A branch of studies aims to tackle this issue with parameter-efficient transfer learning. One approach (known as prompt-tuning) has demonstrated significant promise by achieving comparable results to the standard finetuning, however, its combination with task-relevant subnetworks or symbolic layers tuning remains to be explored. Since neural-symbolic systems show great potential to address two existing challenges in the current NLP development, I am motivated to go into the research area.

Over the past two years, I have been actively engaged in the study of commonsense reasoning, with some of my work intimately intertwined with the concepts explored in the ATOMIC project of the Allen Institute of AI. My first research is PseudoReasoner, which adopts pseudo-labeling as the data augmentation method to address the out-of-domain problem in a Commonsense Knowledge Bases (CSKB) reasoning benchmark. Further exploring the topic of commonsense reasoning, I was exposed to a study from my research collaborators on the application of conceptualization, a perceptual ability related to the K-line theory proposed by Prof. Marvin Minsky, into the Winograd Scheme Challenge. The K-line theory represents human’s mental states in a tree structure, in which conceptualization maps the states in lower-level nodes (associated with more specific information) to the states in higher-level nodes (associated with more abstract information). The study was very interesting as it demonstrated how humans often reason via conceptualization and how researchers can model such a technique in LMs to solve reasoning problems. It inspired me to adopt that technique to solve the CSKB reasoning benchmark. Though there are many instances in the benchmark that prove the applicability of the idea, overall consistent improvement is not yet achieved. Based on my practical experience, I understand that modeling human cognition in language models is challenging, yet appealing to explore.

Shaped by the aforementioned research experiences, my overarching objective is to develop neural-symbolic systems that effectively address various reasoning challenges. As a first step towards this goal, my most recent research attempts to provide a straightforward symbolic augmentation framework to enhance the performance of language models in CSKB reasoning. While LLMs have demonstrated satisfactory performance in numerous well-known commonsense reasoning tasks, prior research and my own findings indicate that LLMs, despite being combined with different prompting techniques like in-context learning and chain of thought, still struggle with acquiring explicit relational constraints present in CSKBs. Consequently, I proposed a plug-and-play component on top of LLMs that explicitly processes the constraint information of relations and augments the final output of LLMs. My research has been directly applied to another project of my research group, in which the framework is used in a link prediction task to improve the quality of ASER - an eventual knowledge base developed and maintained by our group.

As part of my forthcoming research endeavors, I am motivated by the publication titled "Language Models with Rationality" by the Allen Institute of AI. I intend to investigate the utilization of external resources, specifically existing knowledge graphs such as ConceptNet, ATOMIC, ASER, and others, in order to facilitate the identification and rectification of erroneous beliefs within language models. Besides, I am also very interested in examining the true reasoning capability of LLMs. Instead of doing the evaluation at the prompt level as in many previous studies, I want to conduct the evaluation through models' parameters, which I strongly believe can provide insights into whether language models engage in genuine reasoning or merely rely on memorization.

I am actively seeking PhD opportunities all over the world. Although I am open to a variety of research, I believe that I can maximize my contribution to the NLP community if working on neural-symbolic reasoning knowledge representation and reasoning systems.