Kefei Duan — Personal Statement for PhD in Computer Science

My name is Kefei Duan, a girl from China who will complete her undergraduate degree in July 2023. My research interests are mainly related to language and knowledge, like Natural Language Processing, Natural Language Understanding, Question Answering System/Chat bot, Knowledge Extraction, Human-Computer Interaction, etc. From my perspective, my interests for these things mainly come from my enthusiasm for robot – mainly chatting robot. It's so exciting to have machines understand human language and knowledge, to communicate and interact with a robot, and to see what knowledge computers can gain from text and language. In my opinion, the true artificial intelligence should be intelligence that can understand human language and can gain knowledge from it. This can take a long time to achieve and require many people to put their efforts.

The story of my research began a year and a half ago. When I was a sophomore, I was enrolled in a seminar guided by Prof.Ming Zhang at Peking University. It was in the seminar that I was first exposed to research in a real sense. In that seminar, I did research related to the field of text generation and completed a course project related to it, which is called Medical Concept Generation. I tried the classic Seq2Seq model, and also read and tried famous Transformer architecture. Owing to this seminar, I got a chance to meet Prof.Ming Zhang, who is kind and introduced me to research. I got permission from Prof.Ming Zhang to join her research group.

Prof. Zhang asked one of the PhD students in her research group, whose name is Zequn Liu, to supervise me. When I was ready to enter my junior year, I started working on a project about graph generation with Senior Zequn Liu. This project was under the guidance of Prof. Sheng Wang at University of Washington. In this project, we aimed to generate a whole Heterogeneous Information Network (HIN) on medical terms based on a partially observable graph, taking the name of nodes and edges into consideration. We did some research and read some papers on this topic. Due to the hierarchical structure of the medical data, we tried to generate the whole graph layer by layer. But when we tried some existent methods, we found that it was so hard for computer to generate the nodes in next layer based on nodes' name. The performance was just so poor. So we assumed that to generate nodes layer by layer based on name can be hard instinctively. We tried to use a Seq2Seq model to generate a son node's name given a father node's name, or given a father node's name and a grandfather node's name, in which we found that it was hard for the model to generate the son node's name. After a few months' attempts, we changed our focus to a different direction.

When doing research, we found there are some works utilizing metapaths to do graph generation, graph embedding or some other tasks on Heterogeneous Information Networks. Metapath is a sequence of node types and edge types, which can improve the performance of many tasks on graph. But many works generated metapath by manually curating, which can be time-consuming. Automated metapath generation approaches need to be developed. We did some research on metapath generation and found that existing metapath generation methods can not fully exploit the rich textual information in HINs, such as node names and edge type names. So we wanted to incorporate the textual information to help model to generate metapath. The key idea was to formulate metapath identification problem as a word sequence infilling problem, which can be advanced by Pretrained Language Models that bring textual information into metapath. From my perspective, using Pretrained Language Models to facilitate metapath generation is not only a good way to incorporate textual information, but also a more efficient way to find metapath than the search-based methods. We completed our project at June, 2022 and

submitted it to EMNLP 2022, in which we received good review scores. This paper is finally accepted by EMNLP 2022.

When I entered the junior year, I also got a chance to meet with Senior Meng Qu, who is a PhD student at Quebec Artificial Intelligence Institute and is supervised by Prof. Jian Tang. At that time, Senior Meng Qu was going to begin a project combining text and graph. More specifically, he stated that from text we can extract many triplets and construct a Knowledge Graph, but at the same time, existing Knowledge Graph can also help us to extract relationships among entities appearing in corpus. They are complement to each other actually. So, we mainly want to utilize KG to improve the performance of Relation Extraction, which aims to identify relationships expressed by a piece of text and is helpful to many downstream tasks. We did some research on Relation Extraction and found that most existing methods that utilize Knowledge Graph to facilitate Relation Extraction typically leverage the idea of distant supervision or construct additional objective function. But the structural information from the Knowledge Graph is under exploration. So we would like to combine contextual information from text and structural information from KG to predict relationships among entities, hoping to yield better performance. The basic idea is that the entities in corpus can be linked to a Knowledge Graph, where we can find some paths connecting entities. These paths can express some relationship between entities. Given these paths, we could combine text reasoning and path reasoning together to predict relation.

Considering the more complicated relations exhibited in a document, we first focused on sentence-level relation extraction to test the usefulness of structural information from KG. We used FewRel, a famous sentence-level RE dataset, to construct our dataset. In FewRel, each entity is annotated with a corresponding entity in Wikidata. So most of the entities in FewRel can be found on Wikidata5M, which is a large KG constructed from Wikidata and Wikipedia. We then found some paths on Wikidata5M and utilized the relational paths between two entities as additional information to predict their relationship. Preliminary experiments proved that the relational paths from KG can really be helpful. But there may exist a huge gap between the contextual information from text and structural information from KG, which may hinder the model's performance. So we further proposed a fusion module based on attention mechanism to fuse the information from text and KG, which yielded better results. The paper for this work also was submitted to EMNLP 2022, but received review scores that were not so good. It was finally rejected by EMNLP 2022. Taking the advice from reviewers into consideration, we would like to further do some attempts on entity linking, then our method can be applied to more datasets rather than the datasets with annotated entity linking results.

During the summer vacation of 2022, I began another project about Biomedical Document-level Relation Extraction under supervision of Prof. Sheng Wang at University of Washington. This project will be the topic of my thesis.

These research experiences have showed me what research is and taught me some basic methods to do research. They make me familiar with the procedure of a whole project, from the beginning idea to final results. I've learned how to find relevant works, how to think a problem critically, how to write a simple academic paper, and so on. My coding ability is also enhanced after these projects. Apart from these research experiences in research group, some of my course projects also improve my research skills, like a Machine Translation project, an analysis on Dream of the Red Chamber using machine learning, medical concept generation and so on. I value all these experiences so much.

I really would like to learn more about natural language processing and understanding,

question answering system, knowledge extraction and so on. I have the desire to know about the latest research progress. I hope to work deeper into the research, and try to figure out what machines can get from human's language, how machines can interact with human through language, where machines get knowledge, how machines can understand and utilize the knowledge taught to them, and so on. It must be exciting and wonderful. So I think a PhD degree is desirable for me.