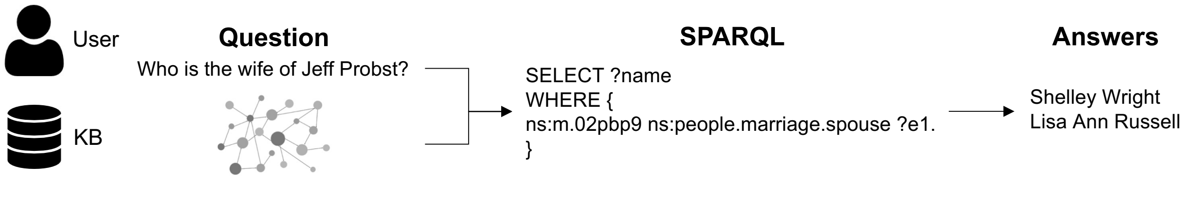
Yiheng Shu — Research Statement

I'm Yiheng Shu, a third-year graduate student in computer science. My research interests mainly relate to natural language processing (NLP), including question answering (QA) and knowledge graph (KG). I intend to *design algorithms and develop systems that can answer general natural language questions*. QA system aims to give users the desired answers to their verbal questions, and it avoids the time-consuming search in person, which is critical for improving productivity and quality of life. Among the knowledge sources of a QA system, a structured knowledge base (KB) has higher quality than textual data. It facilitates logical reasoning, such as multi-hop reasoning and numerical calculation. Therefore, I believe the research on QA and KG, especially Knowledge Base Question Answering (KBQA), is a promising direction for both academia and industry.

I have gained some experience for this goal, working with Professor Yuzhong Qu at Nanjing University and researcher Zhiwei Yu at Microsoft Research. My works mainly discuss semantic parsing-based KBQA methods, which convert natural language questions into executable formal queries (logical forms) over KBs. These methods can formally express the user's intent and give explainable answers, which is essential for building trustworthy AI. Specifically, I participate in developing the relational linking and query generation components of a **question decomposition**-based method named EDGQA [6] (ISWC'21). I also participate in the design of GMT-KBQA based on **multi-task learning** [7] (COLING'22). As the lead author, I design algorithms, implement a complete KBQA system named TIARA [10] (EMNLP'22) using **multi-grained retrieval** augmentation, and achieve the highest Exact Match score and the best zero-shot performance on the GrailQA benchmark [5]. There are still many unresolved challenges in building general and robust QA systems. To learn how to engage in long-term research for this goal and to qualify as a faculty candidate, I would like to apply to a Ph.D. program in computer science. Based on the above works, I think several topics can be my Ph.D. research directions in the coming years and will potentially gain the attention of the research community [8].

Robust KBQA KBs are more suitable for rigorous symbolic reasoning than text, and studying the **robustness** makes QA systems serve more domains and scenarios. The robustness of QA encompasses several aspects, of which **generalization** and **controlled generation** are two obvious issues. First, though interpretable and supporting complex reasoning, semantic parsing-based KBQA requires expensive data annotation, usually limited in quantity and confined to limited domains. However, most methods assume the distribution of test data is the same as the training data but ignore that practical distribution may vary. Though my work TIARA [10] using multi-grained retrieval helps pre-trained language model (PLM) improves performance on compositional and zero-shot generalization, this challenge is far from being solved. I suggest an intuitive solution is question generation, which expands the amount of training data and covers unseen domains. But few studies have evaluated the usefulness of the



```
SELECT ?name
WHERE {
ns:m.02pbp9 ns:people.marriage.spouse ?e1.
}
```

generated questions for KBQA, which is critical to the practicality of this approach in future works. Recent advances in Seq2Seq methods, open-domain question generation methods, and the development of large-scale KBQA datasets [3, 5] makes me believe question generation over KBs is promising to mitigate the challenge of generalization. In addition, PLMs such as GPT-3 [2] and Codex [4] have shown strong generalization capabilities. I want to explore how they can perform better on zero-shot scenarios using prompt or in-context learning methods.

Second, though PLMs are powerful on textual data, they are not initially trained for KBs or logical forms and cannot understand the logical form syntax well. Other semantic parsing tasks, e.g., text-to-SQL [9], have exhaustively studied constrained decoding techniques. Controlled text generation is also received much attention from the NLP community. However, it remains a challenge for semantic parsing on large-scale KBs. Though TIARA [10] proposes using prefix trees to constrain the generation of schema items for uncontrolled PLMs, many other generation errors still break the logical form syntax. I suggest designing more complete rules for complex KB structures and attaching them to PLM to significantly improve performance without additional training data.

Multi-KB and multi-modal QA KBs can be untimely and incomplete due to their construction limitations. Thus, while KB is an essential source of information, QA should not simply rely on it. To leverage knowledge in a more comprehensive and real-time manner, I believe that building QA systems on rich knowledge sources is a direction that deserves open-domain QA, visual QA, text2SQL, and KBQA communities to explore together. First, each KB usually contains a limited number of topics and incomplete facts. Using **multiple KBs** can alleviate this problem, but how to incorporate them remains an open question. Freebase [1] is often used in experiments, but it is no longer under maintenance and not adequate for practical QA systems. I believe the community has recognized that Wikidata [11] will be the successor to Freebase, but related research is much less. To leverage information across KBs, abstract query language across KBs is a potential solution, and PLM with strong generalization capabilities is a powerful tool for reasoning on multiple KBs. The research about multiple KBs will also contribute to applying them in vertical areas such as finance and healthcare, where data is more difficult to access.

Second, QA systems incorporating **KB and other modalities** is an unexplored but interesting area. DecAF [12] jointly decodes the answer text and logical form using PLM, but using text as the answer is likely to face the problem of incomplete answers. Therefore, I argue that extending the scope of semantic parsing requires more sophisticated explicit reasoning over text than simply learning from question-answer pairs. As Wikidata is well-tied to Wikipedia, there is a rich and real-time data source to mine for the fusion of KB and text.

References

[1] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Lan-

guage models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. KQA Pro: A large diagnostic dataset for complex question answering over knowledge base. In *ACL'22*, 2022.

[4] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[5] Yu Gu, Sue E. Kase, Michelle T. Vanni, Brian M. Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond i.i.d.: Three levels of generalization for question answering on knowledge bases. In *Web Conference*, 2021.

[6] Xixin Hu, Yiheng Shu, Xiang Huang, and Yuzhong Qu. Edg-based question decomposition for complex question answering over knowledge bases. In Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam M. Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, editors, *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*, volume 12922 of *Lecture Notes in Computer Science*, pages 128–145. Springer, 2021.

[7] Xixin Hu, Xuan Wu, Yiheng Shu, and Yuzhong Qu. Logical form generation via multi-task learning for complex question answering over knowledge bases. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1687–1696, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[8] Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. Complex knowledge base question answering: A survey. *arXiv preprint arXiv:2108.06688*, 2021.

[9] Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[10] Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje F. Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. Tiara: Multi-grained retrieval for robust question answering over large knowledge bases, 2022.

[11] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[12] Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Wang, Zhiguo Wang, and Bing Xiang. Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases. *arXiv preprint arXiv:2210.00063*, 2022.