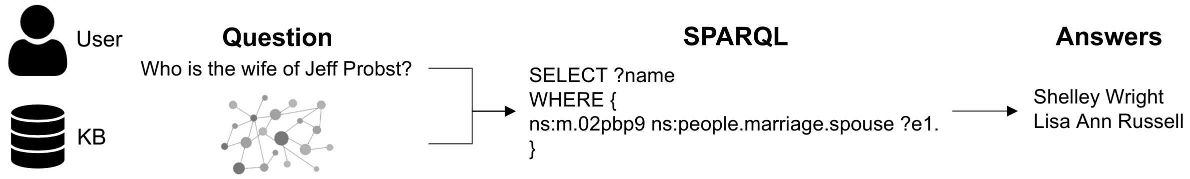
Yiheng Shu — Research Statement

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 systems and develop algorithms that can answer general natural language questions*. My motivation comes from the significance of QA for artificial intelligence. For example, the famous Turing Test, proposed in 1950 to test whether a machine can exhibit the same level of intelligence as a human, takes the form of QA. Among knowledge sources that a QA system can utilize, the knowledge base (KB) is a structured form of higher quality than textual data and facilitates logical reasoning. 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 accumulated several years of research experience for this goal, working with Professor Yuzhong Qu at Nanjing University and researcher Zhiwei Yu at Microsoft Research. Specifically, my works discuss KBQA approaches to question decomposition, multi-task learning, and multi-grained retrieval augmentation. They have been accepted for conferences including ISWC'21 [6], COLING'22 [7], and EMNLP'22 [9]. In these works, I participated in developing the relational linking and query generation components of EDGQA [6]. As the lead author, I designed algorithms, implemented a complete KBQA system named TIARA, and achieved SOTA on the GrailQA benchmark [5]. In addition, I participated in writing a survey [4] with Professor Guibing Guo at Northeastern University (China) in recommender systems, which is highly relevant to NLP techniques. I learned how to organize many publications and develop prototype algorithms in this work, which has been accepted by TOIS'20. Currently, I am attempting to address the challenges of semantic parsing-based KBQA, such as generalization and data scarcity. There are still many unresolved challenges in this field, and I believe there is plenty of room to explore in order to achieve the goal of *building more general and robust QA systems that can make AI more effective in helping humans in more areas*. To learn how to conduct long-term research for this goal, I would like to apply to a Ph.D. program in computer science. Based on above works, I think several related 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 is necessary in order to make QA systems serve more domains and scenarios. The robustness of QA systems 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. Thus, annotations are usually limited in quantity and confined to limited domains. Most methods assume the distribution of test data is the same as the training data, but this assumption does not hold in practice. Though my work TIARA [9] using multi-grained retrieval helps pre-trained lan-



guage model (PLM) improves performance on compositional and zero-shot generalization, this challenge is far from being solved. I suggest an intuitive approach to this problem is question generation, which expands the amount of training data and covers unseen domains. While question generation on KBs has gained some attention, few studies have evaluated the usefulness of the generated questions for KBQA, which is critical to the practicality of this approach in future works. With recent advances in Seq2Seq methods, open-domain question generation methods, and the development of large-scale KBQA datasets, I think question generation over KBs is promising to mitigate the challenge of generalization. In addition, PLMs such as GPT-3 [2] and Codex [3] have shown strong generalization capabilities. I would like to see how they can better generate logical forms using prompt or in-context learning methods.

Second, though PLMs are powerful, 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, 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 [9] 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 the limitations of their construction process. While QA is an essential application of KB, it should not be limited to KB. To leverage knowledge in a more comprehensive and real-time manner, I believe that building QA systems on rich knowledge sources and combining the advantages of different modalities is an issue that the open-domain QA, visual QA, text2SQL and KBQA communities should consider. First, for the KBs themselves, how to build a QA system that **incorporates multiple KBs** is still an open question. Freebase [1] with over 300 million facts is the usual KB used in experiments. However, it was no longer under maintenance after 2015 and not suiable for practical QA systems any more. I believe the community has recognized that Wikidata [10], which is under maintenance and has expanded in recent years, will be the successor to Freebase. But these two KBs are already very different in content and structure, and research on Wikidata is much less than that on Freebase. I think research on QA systems that can leverage information across KBs is worth exploring. Abstract query language across KBs is a potential solution, and PLM is a powerful tool for learning to reason on multiple KBs. In addition to these two large general KBs mentioned above, this research will also contribute to applying KBs in vertical areas such as finance and healthcare.

Second, QA systems incorporating **KB and other modalities** is an unexplored but important area. DecAF [11] provides an idea to decode the answer text and logical form using PLM jointly, but using text as the answer is likely to face the problem of incomplete answers. Therefore, I suggest the inclusion of text in the scope of semantic parsing may be a solution. 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. In addition, tables and KBs are both structured data. Text2SQL methods, such as Seq2Seq-based ones, provides ideas for integrating tables and KBs on QA tasks. Apart from the differences in retrieval methods, the idea of semantic parsing of relational tables is worth considering for migration to KBs.

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

[3] 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.

[4] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Trans. Inf. Syst.*, 39(1):10:1–10:42, 2020.

[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] 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.

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

[11] 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.