My research lies at the intersection of Programming Languages (PL) and Natural Language Processing (NLP). I aim to develop innovative programming languages and synthesis algorithms that combine the convenience and transparency of programming languages with the powerful capabilities of modern deep-learning techniques.

A key focus of my PhD research is to invent neurosymbolic programming techniques that make data insights more accessible to everyone. In this pursuit, my work addresses two main challenges: enhancing the efficiency of domain experts in data analytics and making data analytics more accessible for non-coders. For experts, I focus on designing new languages that blend traditional programming paradigms with the latest advancements in NLP. This combination enables more effective analysis of both semantic and syntactic data elements, simplifying complex tasks for these users. For non-coders, I focus on creating better task specifications and developing multimodal neurosymbolic learning methods. Combined with traditional formal method (FM) techniques, these algorithms are designed to maximize correctness and interpretability.

My research draws inspiration from PL, FM, and NLP and has been published in prestigious PL, NLP, and ML venues, such as PLDI, OOPSLA, NeurIPS, and ACL. My research is highly problem-driven: I first identify pain points people experience in the real world and then develop novel and principled algorithms for solving them.

Neurosymbolic Languages for Data Science

Data science is crucial in decision-making processes across various fields, yet it often demands high programming and technical expertise. In many data science tasks, extracting meaningful information from datasets is crucial. Such extraction typically requires balancing both syntactic and semantic understanding. However, traditional data science tools tend to focus on syntactic or semantic reasoning rather than both. This limitation hampers their effectiveness in scenarios where a comprehensive understanding of data is essential. Especially with data's increasing complexity and heterogeneity, this gap in tool capabilities poses a significant challenge for users striving to perform intricate data analytics tasks.

In response to this challenge, my research has focused on designing innovative neurosymbolic languages that merge symbolic methods for syntactic pattern matching with advanced Natural Language Processing (NLP) techniques for semantic data interpretation. This approach is applied to both semi-structured and unstructured data. For instance, with semi-structured data such as websites, I developed WebQA [1]. This language facilitates extracting information from websites by integrating tree operations for structured navigation with neural elements capable of deciphering complex website content. In the realm of unstructured text, I introduced Smore [2]. This language extends regular expressions with the ability to comprehend and interpret semantic concepts, like business entities and geographical names, utilizing the capabilities of Large Language Models (LLMs). These languages demonstrate the transformative impact of neurosymbolic DSLs in data science, enabling more nuanced and in-depth data analysis.

Program Synthesis for Automated Data Analytics

The development of neurosymbolic DSLs represents a significant advancement in making data science more accessible. However, to fully harness these languages in practical applications, it's crucial to automate the generation of programs based on user specifications. This process faces two primary challenges: firstly, enabling users to convey their intentions effectively, and secondly, learning to synthesize programs in these neurosymbolic languages efficiently.

Program Synthesis from Multi-Modal Inputs. Addressing the first challenge involves allowing users to communicate their desired tasks in a straightforward and comprehensive manner. Traditional methods like providing input-output examples can be burdensome, particularly for complex tasks. On the other hand, relying solely on natural language descriptions lacks a mechanism to verify if the synthesized programs align with user

intentions. My research has made significant contributions in interpreting user specifications from varied inputs, including natural language and input-output examples. By pioneering the use of multi-modal inputs, my work has focused on creating suitable intermediate representations, such as programming sketches or refinement types. These representations serve to efficiently incorporate information from different modalities into a unified form that the synthesizer can understand. For example, in the Regel [3, 4], we designed a hierarchical-sketch-based regex synthesis technique that combines natural language for outlining the intended program structure with examples to resolve ambiguities. Similarly, in Graphy [5], we combine natural language descriptions with data information to synthesize data visualizations, applying refinement types to parse user descriptions and using type-guided synthesis to generate contextually appropriate visualizations. This multi-modal approach significantly reduces the user's specification burden while ensuring the accuracy of the generated programs.

Learning Neurosymbolic Representations. With appropriate user specifications in place, the next challenge is efficiently learning program representations in neurosymbolic languages. The unique nature of these languages poses difficulties for both neural and symbolic synthesis techniques. The symbolic components are not inherently differentiable, and the neural components often yield only partially accurate outputs, complicating the task of finding exact matches. To address these issues, my research has introduced two innovative solutions. The first is the development of neural-guided synthesis algorithms. These algorithms combine formal methods and symbolic reasoning with machine learning-guided search. In the case of Smore [2], for example, a Large Language Model (LLM) first generates a program sketch, which is then refined using local search with user-provided examples. This approach allows the neural component to handle more apparent aspects of synthesis, while examples ensure the correctness of the final program. Such synergistic combinations of neural and symbolic reasoning also show their potential to alleviate the problem caused by errors in the neural techniques by having a way to prove the infeasibility of neural-generated outputs.

The second solution involves quantitative synthesis algorithms. Rather than seeking an exact match to user input, these algorithms identify a set of best-fitting programs when an exact match is unattainable. We employ a transductive learning-based program selection technique to select the program that best generalizes to unseen data. This method is pivotal in finding the most suitable program to adapt and perform effectively in various data contexts such as WedQA [1].

Future Research

My long-term research vision is to develop intuitive and powerful programming tools that cater to a broad spectrum of users. In the next 3-5 years, I aim to advance user-friendly neurosymbolic programming and synthesis techniques further. This pursuit encompasses three key components: broadening the scope and capabilities of neurosymbolic programming, generalizing domain-specific synthesis techniques, and designing life-long learning synthesis frameworks.

Programming support for multi-modal data analytics. A primary research direction will focus on programming support for multi-modal data analytics. Data in its natural environment often exhibits diverse modalities, including texts, images, tables, and graphs. For example, a typical Wikipedia page may feature textual descriptions, location maps, weather visualizations, and demographic statistics. While information retrieval has focused on these challenges for a decade, with systems developed such as DeepDive [6] and OpenRefine [7], these systems need to be more lightweight for quick task adaptation. They are also not on par with the latest developments in NLP and computer vision (CV). To address this, my plan involves leveraging the latest advancements in these fields, particularly vision-language models, combined with my specialty in domain-specific language (DSL) design. Building upon these technologies, I aim to develop innovative programming constructs that integrate and analyze these diverse data modalities seamlessly.

A notable aspect of multi-modal data is the various uses of domain knowledge within these modalities, underscoring the need for customizable analytics languages tailored to specific domains. For instance, a neurosymbolic language for stock market analysis in economics may involve a unique combination of time series data processing

and sentiment analysis for news articles. To enhance the usability of the analytics language in these contexts, I plan to collaborate with scientists across various fields so that the design of DSLs is finely tuned to meet the unique analytical needs of different domains. This collaborative approach will ensure that the developed languages are technically sound, practically relevant, and user-friendly for experts in those fields.

Domain-specific to domain-agnostic synthesis. In the forthcoming phase of my research, I am committed to developing a general neurosymbolic synthesis framework capable of being easily instantiated across multiple domains. This endeavor encompasses two key aspects. Firstly, existing general synthesis frameworks, like SyGuS, lack adequate support for both neurosymbolic languages and synthesis. This deficiency presents significant challenges, such as the absence of standardized benchmarks for neurosymbolic synthesis, elevating the development efforts required for each type of instantiation. Addressing this gap is crucial for streamlining the process of neurosymbolic programming and making it more accessible and standardized across various applications.

The second aspect regards the current tendency in program synthesis to create solutions overly tailored to specific data types or industry requirements. While this approach has its merits in providing highly specialized tools, it often overlooks the potential for generalization across domains. Many domain-specific techniques possess inherent versatility that, if harnessed correctly, could be applied more broadly. However, adapting these techniques to new domains typically involves significant ad-hoc effort. To mitigate this, I plan to develop more versatile specification languages designed explicitly for neurosymbolic programming alongside proper abstraction designed for different types of synthesis, such as search and pruning. This strategy aims to enhance the efficiency and applicability of synthesis techniques, enabling them to transcend domain-specific boundaries and become more universally applicable.

Human-centered synthesis framework. In my future research, I am committed to developing a human-centered synthesis framework that seamlessly integrates program synthesis with data analytic interfaces commonly used by end-users. This integration addresses two major shortcomings prevalent in current practices: the disconnection of program synthesis from user interfaces, which results in the loss of valuable context from past user interactions, and the additional cognitive load users face when adapting to an entirely new system.

One aspect of this integration is developing a backend synthesis engine that intelligently interprets user actions to infer programs using the contexts (such as underlying databases) and their past interaction with the system, thereby providing real-time recommendations and a more user-friendly experience. This approach simplifies user interaction with complex data and leverages their past actions to inform and optimize the synthesis process.

Moreover, a cornerstone of this framework will be incorporating a lifelong learning approach. Diverging from the static nature of traditional programming, where a new program is written every time there is new information, this framework will allow the system to evolve and adapt over time dynamically. It will continuously learn from newly updated data, user interactions, and feedback, ensuring the system remains relevant and effective in changing data and user requirements. Emphasizing human input as a fundamental system element aligns it more closely with user expectations and workflows, enhancing its practicality and efficiency. Realizing this vision requires significant advancements in underlying algorithms and reimagining interaction flows.

Summary

In conclusion, my research endeavors in the intersection of Programming Languages (PL) and Natural Language Processing (NLP) focus on crafting innovative neurosymbolic programming and synthesis algorithms that integrate the precision of programming with the versatility of modern deep learning. My PhD research aims to develop languages and synthesis techniques that enhance the efficiency and accessibility of complex data analysis for both expert and novice users. As I look to the future, my commitment is further to broaden the scope and applicability of neurosymbolic programming, design scalable and correct synthesis techniques, and foster a more inclusive and intuitive landscape in the programming world.

References

[1] **Qiaochu Chen**, Aaron Lamoreaux, Xinyu Wang, Greg Durrett, Osbert Bastani, and Isil Dillig. Web question answering with neurosymbolic program synthesis. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, PLDI 2021, page 328–343, New York, NY, USA, 2021. Association for Computing Machinery.

[2] **Qiaochu Chen**, Arko Banerjee, Çağatay Demiralp, Greg Durrett, and Işıl Dillig. Data extraction via semantic regular expression synthesis. *Proc. ACM Program. Lang.*, 7(OOPSLA2), oct 2023.

[3] **Qiaochu Chen**, Xinyu Wang, Xi Ye, Greg Durrett, and Isil Dillig. *Multi-Modal Synthesis of Regular Expressions*, page 487–502. Association for Computing Machinery, New York, NY, USA, 2020.

[4] Xi Ye, **Qiaochu Chen**, Isil Dillig, and Greg Durrett. Benchmarking multimodal regex synthesis with complex structures. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6081–6094, Online, July 2020. Association for Computational Linguistics.

[5] **Qiaochu Chen**, Shankara Pailoor, Celeste Barnaby, Abby Criswell, Chenglong Wang, Greg Durrett, and Işil Dillig. Type-directed synthesis of visualizations from natural language queries. *Proc. ACM Program. Lang.*, 6(OOPSLA2), oct 2022.

[6] Ce Zhang. Deepdive: A data management system for automatic knowledge base construction. 2015.

[7] OpenRefine. https://openrefine.org/. Accessed: 2023-11-22.