CS224u: Enhancing Compositional Generalization in Semantic Parsing via Few-Shot In-Context Learning

Akhilesh Veerapareddy
Stanford University
akhileshveerapareddy@gmail.com

April 13, 2025

1 General problem/task definition

Semantic parsing is the task of mapping natural language utterances to structured representations such as logical forms. A long-standing challenge in this domain is compositional generalization—the ability to interpret and generate novel combinations of familiar linguistic elements. While humans excel at generalizing to new combinations, contemporary language models often struggle when confronted with inputs that exhibit structural or syntactic configurations not seen during training.

Recent advancements in large language models (LLMs) have opened up new possibilities for improving compositional generalization through in-context learning. In particular, few-shot prompting strategies enable models to generalize from limited annotated examples by conditioning on demonstration examples at inference time. However, the extent to which in-context learning enhances compositional generalization in semantic parsing remains an open question.

This literature review explores foundational work on compositional generalization, semantic parsing, and few-shot learning. Our goal is to synthesize insights that will guide the design of few-shot prompting strategies to improve systematic generalization in semantic parsing, using fine-grained datasets such as ReCOGS.

2 Concise summaries of the articles

Lake and Baroni (2018) investigate compositional generalization using the SCAN benchmark. Their experiments demonstrate that sequence-to-sequence models

perform well on random splits but struggle on splits that require systematic recombination of elements. This work highlights a core limitation in neural models and emphasizes the need for approaches that better capture compositional structure.

Keysers et al. (2020) introduce the COGS dataset, which is designed to assess a model's ability to generalize syntactic constructions. Unlike SCAN, COGS focuses on natural language and systematically varies the syntactic structure while keeping lexical items constant. The authors report that transformer-based models trained on standard objectives exhibit sharp performance drops on generalization splits.

Gupta et al. (2022) present ReCOGS, a diagnostic benchmark for compositional generalization that builds on COGS. Each example is annotated with fine-grained subtasks, allowing for targeted evaluation of specific generalization phenomena such as subject-object inversion or nested structures. ReCOGS is well-suited for evaluating semantic parsers under compositional constraints.

Brown et al. (2020) introduce GPT-3 and demonstrate its ability to perform a range of NLP tasks using few-shot, one-shot, and zero-shot learning. By conditioning on a handful of demonstration examples, GPT-3 achieves competitive results without gradient updates. This work established in-context learning as a powerful mechanism for leveraging pretrained LLMs in low-data settings.

Zhou et al. (2022) study the role of prompt design in improving few-shot semantic parsing. They show that prompt composition—specifically the ordering and formatting of examples—can significantly impact performance. Their findings highlight the need for principled prompt engineering when applying in-context learning to structured prediction tasks.

3 Compare and contrast

The reviewed papers provide complementary perspectives on the challenges and opportunities in achieving compositional generalization in semantic parsing. Lake and Baroni as well as Keysers et al. focus on identifying limitations of standard sequence models through benchmark construction. Gupta et al. advance this direction by introducing ReCOGS, which enables more fine-grained analysis of compositional failures.

In contrast, Brown et al. and Zhou et al. explore solutions through few-shot in-context learning. While Brown et al. emphasize the broad capabilities of LLMs across tasks, Zhou et al. zero in on prompt design strategies specifically for semantic parsing. Together, these papers motivate a research agenda that combines diagnostic datasets like ReCOGS with targeted prompt engineering techniques.

4 Future work

Future research should investigate how in-context learning can be systematically optimized for semantic parsing tasks that require compositional generalization. This includes identifying which subtasks benefit most from few-shot prompting, developing methods for automatic prompt selection, and studying the robustness of prompting strategies across models.

Additionally, integrating ReCOGS-like subtask labels into the prompt design process could allow for more adaptive and interpretable few-shot learners. Another promising direction is combining few-shot prompting with modular reasoning frameworks, which may further improve generalization to novel structures.

Our project aims to build on these ideas by evaluating the effectiveness of prompt design strategies on compositional generalization in semantic parsing, using ReCOGS as the primary evaluation benchmark.

References

References

[Brown et al.(2020)] Brown, T., Mann, B., Ryder, N., et al. (2020). *Language Models are Few-Shot Learners*. In NeurIPS. https://arxiv.org/abs/2005.14165

[Gupta et al.(2022)] Gupta, A., Kheradpisheh, S. R., & Potts, C. (2022). *ReCOGS: A Fine-Grained Benchmark for Evaluating Compositional Generalization*. In EMNLP. https://aclanthology.org/2022.emnlp-main.375/

[Keysers et al.(2020)] Keysers, D., Schärli, N., Scales, N., et al. (2020). *Measuring Compositional Generalization: A Comprehensive Method on COGS*. In ICLR. https://arxiv.org/abs/2004.14076

[Lake and Baroni(2018)] Lake, B. M., & Baroni, M. (2018). *Generalization without Systematicity: Strong Generalization in Neural Networks Does Not Imply Strong Systematicity*. In ICML. https://proceedings.mlr.press/v80/lake18a.html

[Zhou et al.(2022)] Zhou, S., Xu, Y., Chen, X., & Lin, C. (2022). *Prompting Large Language Models for Few-Shot Semantic Parsing*. In Findings of ACL. https://aclanthology.org/2022.findings-acl.83/

Acknowledgments

I consulted OpenAI's ChatGPT to brainstorm ideas, clarify NLP concepts, and help organize the structure of this literature review. All summaries and analysis reflect my own understanding and judgment. The content was critically reviewed and written by me, Akhilesh Veerapareddy.