Making Natural Language Reasoning Explainable and Faithful

Xinya Du

Computer Science Department
University of Texas at Dallas
xinya.du@utdallas.edu

Neural models, including large language models (LLMs), achieve superior performance on logical reasoning tasks such as question answering. To elicit reasoning capabilities from LLMs, recent works propose using the chain-of-thought (CoT) mechanism to generate both the reasoning chain and the answer, which enhances the model's capabilities in conducting reasoning. However, due to LLM's uninterpretable nature and the extreme flexibility of free-form explanations, several challenges remain: such as struggling with inaccurate reasoning, hallucinations, and not aligning with human preferences. In this talk, we will focus on: (1) our design of leveraging **structured information** (that is grounded to the input context), for the explainable complex question answering and reasoning; (2) our **multi-module interpretable** framework for inductive reasoning, which conducts step-wise faithful reasoning with iterative feedback.

First, in the context of multi-hop question answering (Figure 1), we will cover our recent efforts in pushing forward the explainability of **deductive reasoning** systems (Li and Du 2023). Multi-hop question answering (QA) involves answering questions that require reasoning over multiple pieces of information, often spread across different parts of a document or multiple documents. It is like "hopping" over various facts to arrive at a conclusion or answer. Mainstream methods conduct generations in an end-to-end way, which is not a strict "symbolic derivation" and leads to inconsistent and incomplete reasoning.

To tackle this challenge, we propose leveraging structured and grounded knowledge in multiple steps: firstly, constructing the semantic graph structures (blue elements in the Figure) with information extraction (IE) and then leveraging this symbolic information (including entities and semantic relations) for strictly guiding the model's reasoning process. The extracted graphs from the IE step naturally provide fully grounded interpretations of the given context. Plus, it helps the model's further generation of free-form explanations, by providing prior structured knowledge or potential reasoning path. Empirically, we will demonstrate that leveraging semantic graphs in the prompting process helps models (1) generate higher-quality and more faithful reasoning chains; as well as (2) answer questions more accurately.

Second, **Inductive reasoning** requires generating likely rules from a set of facts. Past research investigates Inductive Logic Programming (ILP), which suffers from systematic disadvantages such as low coverage and manual cost. While LLMs can directly transform natural language facts into possible rules, the process is uninterpretable and prone to hallucinations. We propose decomposing complex natural language inductive reasoning into multiple steps, which are tackled by specific modules in an interpretable fashion (Yang et al. 2022). More specifically, we design a framework including: (1) a raw rules generation module; (2) clarity, reality, and novelty checking modules for filtering out qualified rules and providing explainable feedback; and (3) an iterative mechanism for refining and filtering the rules with the feedback. Further, we will introduce the application of our work to hypothesis discovery, which is of fundamental influence on scientific society including AI research.

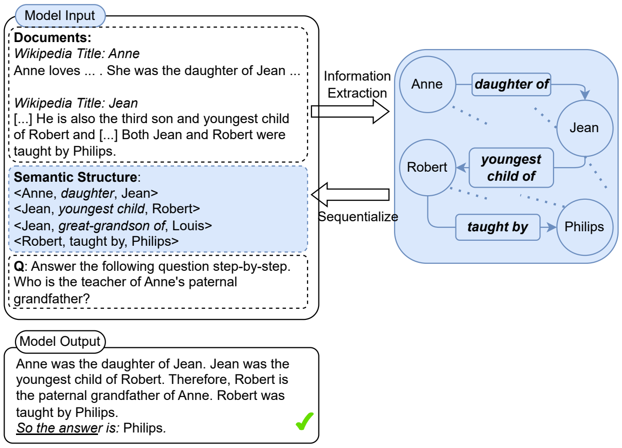
Model Input
Documents:
Wikipedia Title: Anne
Anne loves ... . She was the daughter of Jean ...
Wikipedia Title: Jean
[...] He is also the third son and youngest child of Robert and [...] Both Jean and Robert were taught by Philips.
Semantic Structure:
<Anne, *daughter*, Jean>
<Jean, *youngest child*, Robert>
<Jean, *great-grandson of*, Louis>
<Robert, taught by, Philips>
Q: Answer the following question step-by-step. Who is the teacher of Anne's paternal grandfather?
Information Extraction
Sequentialize
Anne
daughter of
Jean
Robert
youngest child of
taught by
Philips
Model Output
Anne was the daughter of Jean. Jean was the youngest child of Robert. Therefore, Robert is the paternal grandfather of Anne. Robert was taught by Philips.
So the answer is: Philips.
✓

Figure 1: Leveraging the extracted semantic graph for accurate question answering and faithful reasoning.

References

Li, R.; and Du, X. 2023. Leveraging Structured Information for Explainable Multi-hop Question Answering and Reasoning. In *EMNLP (Findings)*.

Yang, Z.; Dong, L.; Du, X.; Cheng, H.; Cambria, E.; Liu, X.; Gao, J.; and Wei, F. 2022. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*.