Title	言語モデルの推論能力に関する研究
Author(s)	原口, 大地
Citation	
Issue Date	2024-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/18914
Rights	
Description	Supervisor: 白井 清昭, 先端科学技術研究科, 修士(情報科学)

Abstract

Reasoning is a fundamental capability of human intelligence, which allows us to make inferences or predictions about unknown things based on known information. In the field of natural language processing (NLP), a subfield of artificial intelligence that deals with language, the reasoning ability of NLP models is a key factor directly linked to their generalization performance in machine learning contexts. Despite the significant progress in the language capabilities of NLP models since the advent of Transformer, numerous challenges still exist in their reasoning abilities. Shortcut reasoning refers to the irrational reasoning of NLP models, which is a critical issue for reasoning ability. Specifically, shortcut reasoning involves a type of reasoning that performs well with input that follow the same distribution as the training data, but perform poorly on input with a different distribution from the training data. Previous studies have shown that this irrational reasoning can degrade the robustness of NLP models.

Recently, Large Language Models (LLMs) have attracted much attention for their language capabilities that surpass previous NLP models. Improvements have also been observed in the reasoning performance of LLMs. Previous studies have reported that attaching specific prompts to the input or combining LLMs with external algorithms for reasoning improves the performance.

However, Hallucination, the generation of counterfactual knowledge by LLMs, is known as a significant challenge. Even if it is possible to construct the reasoning process correctly, the final answer may contain errors when the knowledge used in the process is incorrect. Previous work has shown that hallucinations can snowball, where the errors generated during reasoning successively affect downstream reasoning processes. Given these issues, various solutions are being studied. One of them is a method called Retrieval-Augmented Generation (RAG). RAG is a paradigm that retrieves documents from a knowledge base in response to an input and generates an answer while referring to the retrieved documents. Many research has shown that this approach reduces hallucination and improves performance. However, there is still room for improvement in the retrieval mechanism of RAG. A typical example is that the retrieved fixed number of documents may contain unnecessary information or may not contain the necessary information. Unnecessary information can introduce noise and amplify hallucination. In addition, there are still unresolved issues, such as requiring retrieval for each reasoning step.

LLM's reasoning ability itself still has problems. Previous studies have shown that LLMs are good at single-step reasoning, but inherently struggle with compositional reasoning problems that require synthesizing partial reasoning results, such as computation problems like 3-digit multiplication. It has been revealed that

the seemingly excellent reasoning of LLMs we observe is due to pattern matching using the reasoning processes included in the massive pre-training data.

In view of the aforementioned issues related to the reasoning of NLP models and LLMs, we aim to realize a reasoning system with self-awareness and interpretability. Self-awareness refers to the ability of the system to recognize what information it possesses and what it lacks. Interpretability, on the other hand, pertains to its ability to explain how input questions are broken down and which knowledge is consulted during the reasoning process. This system outputs a final prediction by decomposing the input question, querying a knowledge base, and synthesizing the obtained results. The proposed system comprises three major modules: a knowledge base module, a reasoning module, and a central control module. The knowledge base module employs a self-aware LLM that stores and retrieves relevant knowledge related to a given question. The reasoning module adopts a probabilistic logical programming, which decomposes the question into sub-questions and refers to the knowledge base to obtain necessary information. Finally, the central control module integrates the retrieved information to generate a final prediction of answer.

To achieve this goal, we have conducted research on reasoning ability of NLP models, followed by a discussion toward developing a self-aware knowledge base module. Specifically, we worked on the following three themes.

(i) In automatic discovery of shortcut reasoning with generality, we addressed the challenges of existing methods and proposed a method for automatically detecting shortcut reasoning. The previous methods to discover shortcut reasoning within NLP models have several issues, such as pre-definition of shortcut reasoning, not using internal states of the models, or requiring human evaluation. Even though the latest work overcame those problems, its method still suffer from several limitations. As a result of experiments with our proposed method on NLP models trained on sentiment analysis and natural language inference tasks, we succeeded in detecting shortcut reasoning without human intervention and discovering unknown shortcut reasoning not revealed in previous studies as well as known ones.

(ii) In our research on a logical rationale-based machine reading comprehension model, we conducted experiments on shortcut reasoning in Explainer for the machine reading task. Explainer is a module of Explain-then-predict architecture. Explain-then-predict is an architecture generally used for Rationalization, where models are forced to output their rationale along with the predictions. As the rationalization architecture has explainer, which extracts necessary information for the right inference, previous work hypothesize the architecture can improve robustness by excluding unnecessary noise in the input documents. However, the results did not support their expectation. Thus, we hypothesized that the robustness improvement was not achieved since explainer was performing shortcut reasoning. The empirical experiment revealed explainer's shortcut reasoning by showing that accuracy did not drop even when the explainer's input was destructively corrupted.

(iii) In our discussion toward large language models as knowledge bases with self-awareness, we approached the relationship among confidence, accuracy and frequency in terms of LM as KB, which utilize language models as knowledge bases. In previous work, confidence, the probability that the prediction is correct, in prediction by LLMs is known to be well-calibrated. Correlations between (calibrated) confidence and accuracy, and between frequency and accuracy are revealed in several studies, but not for the correlation between frequency and confidence. Therefore, we analyzed the relation between frequency and confidence, and conducted preliminary experiments to verify our hypothesis that the confidence of LLM's knowledge reflects the frequency of its appearance in pre-training rather than its correctness. We employed PopQA, which contains triplets of knowledge (subject, relation, object) retrieved from Wikipedia and each of which is annotated with its popularity as a proxy of frequency the knowledge appeared in pre-training data. We used several LLMs for the experiments, and when using GPT-3.5-turbo, we obtained results showing that confidence and accuracy were positively correlated for knowledge with high frequency but that accuracy did not increase even when confidence increased for knowledge with low frequency. Although these results were intriguing, they did not lead to the verification of our hypothesis.