JAIST Repository
https://dspace.jaist.ac.jp/

Title	言語モデルの信頼性向上にむけて -論理的同値性を理解できるか-
Author(s)	田中, 健史朗
Citation	
Issue Date	2024-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19359
Rights	
Description	Supervisor: 井之上 直也, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Abstract

Large Language Models (LLMs) are language models trained on huge amounts of data, achieving near-human accuracy in various benchmarks such as inference and translation. However, it has been reported that LLMs may produce different outputs depending on the variations in input expressions, even if the meaning remains the same. For example, when querying an LLM about Anne Redpath's place of death, the input prompts "Anne Redpath's life ended in" and "Anne Redpath passed away in" may yield different results, with the LLM responding with "London" and "Edinburgh," respectively. This issue raises concerns about the reliability of LLMs. Despite their high performance on benchmarks, doubts remain about whether LLMs can effectively utilize this high performance, making it difficult to trust their responses. This inconsistency in responses is not limited to vocabulary-level paraphrasing, as previous research has pointed out, but also extends to logically equivalent expressions, as revealed by preliminary investigations in this study.

To address the inconsistency in LLM responses, existing research suggests that fine-tuning LLMs with data extended through paraphrasing could potentially enhance their robustness. Additionally, there is existing research on data augmentation using equivalence transformations in propositional logic. While the generation way of the logical formulas is relatively simple, LLMs using the augmented data have shown improved accuracy on benchmarks, indicating a possible enhancement in robustness. However, propositional logic represents a relatively simple form of logic found in natural language sentences, and there is a need to improve robustness concerning more expressive predicate logic.

Therefore, this study proposes the construction of a language model that is robust to logically equivalent expressions in predicate logic. To achieve this, we aim to use a data augmentation method that guarantees logical equivalence to learn that the extended sentences have the same meaning as the original ones. This will enable the language model to implicitly learn logically equivalent relationships, thereby enhancing the robustness of such expressions. Regarding the data augmentation method that guarantees logical equivalence, we propose a paraphrase method LET, which transforms transforming input sentences into logical formulas, performing equivalence transformations using first-order predicate logic, and then converting them back into natural language sentences.

LET consists of three components: T2L (Text-to-Logic translation), L2L (Logical form-to-Logical form translation), and L2T (Logical form-to-Text translation). These components are connected in a pipeline to achieve data augmentation with logical equivalence.

For T2L, translating natural language sentences into logical formulas has been a

long-standing area of research, particularly focused on first-order predicate logic. This translation enables the use of theorem provers like Coq, which can apply logically grounded approaches to a wide range of natural language processing tasks, such as RTE and calculating sentence similarity. Traditionally, Text-to-Logic translation has been explored using rule-based methods. However, due to the complexity of natural language, extending these rule-based approaches to practical applications is challenging. With the success of large language models, there has been growing interest in methods that use LLMs, which offer a balance of translation accuracy and flexibility. The emergence of powerful LLMs like GPT-3.5 and GPT-4 has led to a new paradigm where LLMs are employed to handle most of the translation tasks. Nonetheless, recently, ccg2lambda has been proposed for semantic analysis and reasoning from the perspective of mathematical logic. ccg2lambda is a system that consistently performs CCG parse tree generation (syntactic parsing) using Combinatory Categorial Grammar (CCG), logical formula generation (semantic analysis) using higher-order logic, and automated reasoning using the Coq. One advantage of ccg2lambda over LLM-based methods is that, with correct syntactic parsing, it can generate unique logical formulas that faithfully preserve the meaning of sentences. Therefore, this study prioritizes the reliability of the generated logical formulas and adopts ccg2lambda for T2L.

For L2L, while there are many possible transformations for logically equivalent expressions, many of these transformations may not result in natural expressions when converted back into the same natural language or may not produce noticeable differences. Therefore, this study focuses on equivalence transformations in first-order predicate logic and adopts rules that ensure the resulting expressions are different when translated back into natural language.

For L2T, translating logical formulas into natural language sentences (Logic-to-Text) has also been a focus of research, particularly in the context of first-order predicate logic. Since logical formulas can be viewed as a language that translates natural language sentences into a specific theoretical framework, Logic-to-Text has traditionally been considered a simpler task than Text-to-Logic and has been approached using rule-based methods. With the advent of LLMs, researchers have explored using these models for Logic-to-Text, leveraging their high text generation capabilities and flexibility, which rule-based methods lack. Although this area has not been as actively pursued as Text-to-Logic, there have been studies utilizing LLMs for Logic-to-Text translation. However, because ccg2lambda generates formulas in higher-order predicate logic, existing translation methods cannot be directly applied. Therefore, this study aims to construct an LLM-based model that takes logical formulas as input and outputs natural language sentences.

To construct a language model that is robust to logically equivalent expressions,

we will use LET to extend the RTE (Recognizing Textual Entailment) problem and fine-tune the language model with the extended data. This approach aims to enable the model to implicitly learn logically equivalent relationships and improve its robustness to such expressions.

In experiments, to effectively verify the robustness of LLMs to first-order predicate logic equivalence, we focused on FOLIO, a natural language RTE dataset for first-order predicate logic. We used LET to paraphrase the premise sentences in FOLIO and extend the RTE problem. By fine-tuning the large language model with the extended RTE problems, we aimed to construct an LLM robust to logically equivalent expressions. The results confirmed that using manually extended data, which represents the theoretical upper bound of LET, improved accuracy on the RTE problems while achieving robustness on Llama2 (7B). This demonstrates that the proposed method is effective in enhancing LLM robustness to logically equivalent expressions without compromising inference ability. However, the effectiveness of the model fine-tuned with automatically extended data using LET could not be confirmed, indicating that the paraphrasing accuracy of LET is still insufficient. Additionally, varying the parameter size of the base model did not validate the effectiveness of LET even at its theoretical upper bound, suggesting that further investigation into the effectiveness of LET is needed.