概要

大規模言語モデル (LLM) は大量のデータによって学習を行った言語モデルであり, 推論や翻訳など様々なベンチマークにおいて人間の精度に迫る高い性能を発揮している. しかし, その一方で, LLM には同じ意味の入力でも表現の違いにより出力結果を変えてしまう問題が報告されており, LLM の信頼性を揺るがしている. 入力頑健性が低い現状では, 仮に LLM がベンチマークで高い性能が実現していたとしても, LLM がその高い性能を正しく活用できるかどうかという観点で疑問が残り, LLM の回答を信頼することはできない. このような回答の非一貫性は, 既存研究で指摘されていた語彙レベルの言い換えだけに留まらず, 論理的に同値な表現についても発生していることが本研究の予備調査で明らかとなった.

LLM の回答の非一貫性に対して, 既存研究では Paraphrase により拡張したデータを用いて LLM をファインチューニングさせることが LLM の頑健性を高める可能性があることを示唆している. また命題論理の同値変形によるデータ拡張についても研究がされており, 同値変形元となる論理式の生成が簡易的ではあるものの, 拡張したデータを利用した LLM はベンチマーク上で正答率を向上させており, 頑健性についても向上している可能性が示されている. ただし, 命題論理は, 自然言語文に現れる論理の中では比較的簡素なものであり, より表現力の高い述語論理に関して頑健性を向上させる必要がある.

そこで本研究では, 述語論理上で同値な表現に対して頑強な言語モデルの構築を提案する. これを実現するために, 論理的な同値性を保証するデータ拡張手法を使い, 拡張した文が拡張元の文と同じ意味であることを学習させることで, 言語モデルが暗黙に論理的に同値な関係性を学習し, 論理的に同値な表現に対して頑健性を向上させることを目指す. また, 論理的な同値性を保証するデータ拡張手法については, 入力文を論理式へ変形し, 一階述語論理を活用した同値変形を行い自然言語文に戻すことで, 論理的な同値性を担保する Paraphrase 手法 LET を提案する.

LET は, 自然言語文を論理式に変換する T2L (Text-to-Logic translation), ある論理式を論理的に同値な論理式へと変形する L2L (Logical form-to-Logical form translation), 論理式から自然言語文へと変換する L2T (Logical form-to-Text translation) の 3 つのコンポーネントからなり, これらをパイプライン状に繋げることで論理的に同値なデータ拡張を実現する. T2L には, 論理式の信頼性を重視して, 形式統語論と形式意味論の見地から高階述語論理を活用して意味解析・推論を行う ccg2lambda を採用する. L2L には, 古典一階述語論理における同値変形に注目し, 自然言語文に戻したときに異なる表現となるような規則を採用する. L2T には, 高階述語論理に対応するため, 論理式を入力として自然言語文を出力する, LLM ベースのモデルを構築する.

そして, 論理的に同値な表現に対して頑強な言語モデルの構築では, LET を使って RTE 問題を拡張し, その拡張データをファインチューニングすることで, 言語モデルに暗黙に論理的に同値な関係性を学習させ, 論理的に同値な表現に対して頑健性を向上させる.

評価実験では, 一階述語論理の同値性に対する LLM の頑健性を効果的に確認するため, 一階述語論理の自然言語文の RTE データセットである FOLIO に着目し, LET を使って FOLIO の前提文を言い換え RTE 問題を拡張した. そして, 拡張した RTE 問題で大規模言語モデルをファインチューニングすることで, 論理的に同値な表現に対して頑健な LLM の構築を目指した. 結果として, LET の理論上限である, 手作業で拡張したデータを用いた場合には RTE 問題に対する正答率が向上させながら頑健性を獲得していることを確認し, 本手法が推論能力を損なわずに論理的に同値な表現に対する LLM の頑健性向上に有効な手法であることを示した. しかし, LET により自動的に拡張したデータを用いてファインチューニングしたモデルについては有効性が確認できず, LET 自体の言い換え精度は未だ不十分である. また, ベースモデルのパラメタサイズを変更すると LET の理論上限でも有効性が確認できず, LET の有効性についてさらなる調査が必要となる結果となった.