先端科学技術研究科　修士論文要旨

所属研究室 (主指導教員)	自然言語処理学 (渡辺　太郎 (教授))		
学籍番号	2311141	提出日	令和 7年 1月 21日
学生氏名	鈴木　刀磨		
論文題目	大規模言語モデルにおけるタスク特有の表層表現に起因する脆弱性の調査		

要旨

大規模言語モデル(LLM)はラベルなしデータを用いた事前学習により高いタスク汎化性能を達成しているが，指示テンプレートを用いて様々なタスクを学習するInstruction-tuningを適用することで，さらにその能力を高めることが可能である．Instruction-tuningでは過学習を回避するため，学習に使用する指示テンプレートの多様性を確保しなければならない．この点を踏まえ，FLANデータセットに代表されるような既存のInstruction-tuning用データセットではタスクごとに複数のテンプレートを提供している．その一方で，これらのテンプレートには対象とするタスクと密接に関連する単語といったタスク特有の表層表現が含まれている．このような指示テンプレートに含まれる語句の偏りは学習を通じてLLMに反映される可能性があり，その場合に特定の表層表現に対して性能低下を引き起こす原因となり得る．本研究ではこのような指示テンプレートに含まれるタスク特有の表層表現に起因するLLMの脆弱性の調査を行う．この調査のため，我々は指示文に対してタスクの観点から内容を維持しつつ対象とする単語を挿入する手法を提案した．FLANデータセットから抽出した単語を含むように作成した指示文を用い，ベンチマークデータセットであるMMLUとBBHを対象とした検証の結果，各タスクに強く関連する単語が指示文に含まれることで，文意と無関係に出力結果が大きく変化し得ることを明らかにした．この結果は指示テンプレートに含まれる表層的な単語表現がLLMの脆弱性を引き起こす可能性を示すものであり，Instruction-tuningをより頑健なものとする上で重要な知見である．