

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	2411043	提出日	令和 8年 1月 19日
学生氏名	大熊 基暖		
論文題目	教育的観点を考慮したLLM-as-a-Judgeに基づく教師応答生成		
要旨			
本研究は、教育対話における教師応答生成を対象とし、教育的観点を明示的に組み込んだ LLM-as-a-Judge に基づく応答生成フレームワークを提案する. 従来、教師応答の評価には BLEU や BERTScore など参照応答との類似度に基づく自動評価指標が広く用いられてきたが、教育対話では望ましい応答が一意に定まらないことが多く、学習者支援としての有用性を十分に反映しにくい. そこで本研究では、教師応答の教育的質を評価する観点として Pedagogical Appropriateness (PA), Student Understanding (SU), Instructional Helpfulness (IH)を導入し、大規模言語モデルに 1〜5 の尺度で採点させる. 提案フレームワークは、(i) 学習者発話から複数の教師応答候補を生成し、(ii) それらを上記 3 観点でスコアリングし、(iii) Best/Worst の対比情報を用いて最終応答を再生成することで、評価と生成を統合する. 英語学習者と教師の対話コーパスを用いて実験を行い、生成応答を BERTScore, DialogRPT, および提案する教育的観点に基づく LLM 評価によって比較した. その結果、主要評価項目で平均スコアの観点では提案手法が既存ベースラインを一貫して上回る傾向は確認されなかった. 一方で、エラー分析により、教育的観点に基づく LLM 評価は、学習者発話との整合性の欠如、足場(例示・制約・段階化)不足のままの問い掛け、教育的文脈からの逸脱といった失敗パターンに敏感に反応することが分かった. 以上より、本研究は LLM-as-a-Judge に基づく教育的評価を、単なる性能比較の尺度としてではなく、教師応答生成における評価挙動と失敗構造を診断的に捉えるための枠組みとして位置付け、今後の教育対話システムにおける評価設計および生成制御の検討に資する知見を示す.			