ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名： 深層学習による系列学習

実施時期： 2025年4月1日～2026年3月31日
所属機関名： 豊田工業大学
代表者氏名： 佐々木裕

成果概要：

言語データはトークンが全順序で並んだ系列データである．本研究はLLMが生成する文における**誤用**に注目した．例えば「車が逆走してきたが（　　）で避けられた」の穴埋めについて，LLMが「危機一髪」を生成する確率を測定すると59.8%であったのに対し，「危機一発」と誤用する確率は僅か0.36%であった．これは人間の誤用傾向と大きく異なる．LLMは膨大な文書から単語の生成確率を学習しているため，誤用を無視し，マジョリティである正しい表現に文生成が偏っている可能性がある．これが**LLMが生成する文の違和感**の原因なのではないかという点に着目した．そこで本研究では，独自に構築した日本語誤用データセットにより，ABCIを用いて様々な日本語LLMに大規模な追加訓練を行うことで，人間の誤用傾向にどの程度に近づけられるかの評価実験を行った．

成果のポイント：

本研究では，人間とLLMの誤用生成傾向を比較するために誤用データセットを作成した．本研究で対象とする誤用事例は，成人の日本語母語話者による書き言葉の誤りであり，キーボード操作などの物理的ミスではなく，認知や勘違いに起因する誤りであること，文字・単語レベルで誤り箇所が同定可能であること，そして十分な文脈長が求められる．ベースとして日本語Wikipedia入力誤りデータセット（Japanese Wikipedia Typo Dataset; JWTD）を用いた．JWTDは編集履歴から文単位の修正前後を収集した正誤ペアであり，本研究の枠組みに適合する．

一方で，JWTDには誤変換・表記揺れ・打鍵ミスなど，本研究の対象外の事例も多く含まれる．対象件数は約29万件に及ぶため，全件の手作業分類は困難である．また，誤用判定は文脈依存性が高く，単純な辞書・ルールでは高精度な分離が難しい．そこで，今回はLLMを用いたICL（In-Context Learning）により，各事例を二値分類問題として解かせた．具体的には，誤りの原因が認知的な誤用（CognitiveError）であるか単なる入力操作ミス（KeystrokeError）かを分類する．これを実施するために，開発用データとして200件のデータにラベル付けをして，複数のモデルとプロンプト構成の組み合わせから最も適合率が高いモデルを選択した．作成したデータセットにより，誤用しやすい表現についての人間とLLMの正解率を測定した．その結果を右の表に示す．

次に，作成した誤用データを用いて，日本語LLMに対してファインチューニングを行った．パラメータ数が30億～700億に上るLLMをファインチューニングするには，最新のGPUが有効であった．また，複数のGPUを並列で利用できたことで，訓練の効率を上げることができた．実験の結果，誤用傾向を人間に近づけられることが定量的に検証された．

誤	正	人間	sarashina 3b	ELYZA 8B	Swallow 70B Inst
一獲千金	一攫千金	0.340	0.945	0.986	0.974
(場内を) 湧かせた	沸かせた	0.435	0.679	0.984	0.920
若干	弱冠	0.481	0.818	0.755	0.245
木偶の棒	木偶の坊	0.583	0.947	0.971	0.818
激 (を飛ばし)	檄	0.603	0.018	0.005	0.026
(雪辱を) 晴らした	果たした	0.693	0.976	0.997	0.932
貪欲に	貪欲に	0.753	0.997	1.000	1.000
思いばかって	慮って	0.813	1.000	1.000	1.000

成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

中西 純, 牧野 晃平, 佐々木 裕, 日本語における LLM と人間の誤用傾向の差異の分析, Q2-14, 第32回言語処理学会年次大会, 3月, 2026.（第32回言語処理学会年次大会スポンサー賞受賞）