2023 年度年次報告書
社会変革に向けた ICT 基盤強化
2023 年度採択研究代表者

曹 洋

北海道大学 大学院情報科学研究院
准教授

大規模言語モデルのための新しい信頼性向上技術

研究成果の概要

本年度の研究結果として、大規模言語モデル（LLM）のファインチューニングとユーザープライバシー保護、生成コンテンツの識別方法に関する研究を行った。この研究は二つの主要成果を挙げている。

一つ目は、LLM の差分プライベートファインチューニングに関するものである。特に、差分プライベート確率的勾配降下法（DPSGD）の使用に焦点を当て、勾配をクリッピングしてからノイズを注入する手法を採用している。この方法は、勾配がサブガウス分布に従うと仮定されている既存の研究に基づいているが、実際には勾配に重尾特性が見られることが多く、これにより既存の手法ではクリッピング損失が過剰に発生してしまう。そこで我々は、「Discriminative Clipping (DC)-DPSGD」という新しい手法を提案した。これは、勾配の重尾部分とそうでない部分を区別し、それぞれに異なるクリッピングしきい値を適用することで、クリッピング損失を減少させるものである。

二つ目の研究は、LLM による日本語生成コンテンツの識別方法に関するものである。LLM の性能向上により、その応用範囲は広がっているが、同時にフェイクニュースや詐欺メール生成などの問題も生じている。このため、生成された文を検出する手法が求められており、私たちは日本語における生成文の検出方法を開発した。この研究では、まず人間が書いた文と LLM による文を比較するデータセットを作成し、そのデータセットを用いて既存の検出手法の精度を評価した。この初期の成果は DEIM2023 で発表され、今後さらに改良を加えて国際会議での発表を目指している。

これらの研究成果は、LLM の安全性と効果性を高めるための重要なステップであり、今後の研究開発に向けた新たな指針を提供するものである。

【代表的な原著論文情報】

1) 丸井渚生, 曹洋, 中村篤祥. 日本語における大規模言語モデルの生成文検出. 第 16 回データ工学と情報マネジメントに関するフォーラム（第 22 回日本データベース学会年次大会）.