ABCI 3.0開発加速利用 （2024年度）成果概要（公開用）

課題名： LLMの事前学習のための学習環境の構築や環境検証のための小規模なモデルの開発	実施時期： 2025年1月29日～2025年3月30日 所属機関名：国立研究開発法人情報通信研究機構 代表者氏名： 呉　鍾勲

成果概要：

NICTでは、2025年度以降、大規模な日本語学習データを用いた日本文化、アイデンティーに忠実なLLMの学習を計画しているが、LLMの学習環境の動作確認、計算速度の推定、適切な学習データの構成等を事前に調査、確認するため、学習予定の大規模学習データの一部を用いて80億パラメータと160億パラメータの小規模なモデルの開発を行った。

成果のポイント：

NICTではこれまで、高性能な日本語特化型の大規模言語モデル（LLM）構築のため、長年に渡って収集してきた約381億ページの日本語Web文書から、22.9TBという日本語データとしては、我々の知る限り、世界最大の事前学習用データを構築した。このデータは、通常のプログラマが作り込んだフィルタープログラムによるものではなく、275万文の日本語Webテキストに対して、LLMの学習データとして有効と思われるテキストを人手で選択した学習データで訓練したBERTを、上記381億ページの日本語Web文書に適用し、さらに重複削除を行うことで構築されたものである。また、我々はその学習データの一部を用いてより適切な学習データの構築法やLLM学習時のハイパーパラメータの探索のために多数の日本語LLMを試作してきた。この過程で、事前学習用データの品質、量、さらには学習時のハイパーパラメータがLLMの性能を往々にして想定外な形で大きく左右することを確認している。本研究は、この一連の試行錯誤から得られた知見を活かし、より大きな学習データを用いてLLMを効率的に学習するための取り組みである。

より具体的には、NICTでは2025年度以降に上述した日本語事前学習用データを用い、日本の文化やアイデンティーに忠実な日本語特化型LLMを構築することを計画している。なお、この構築はLlama等のオープンなモデルに追加で学習を行わせるのではなく、完全にスクラッチから日本語データのみでの学習を予定している。本研究では、この大規模な学習に必要な学習環境の動作確認、さらに、計算速度や適切な学習データの構成等を事前に調査、確認するために、上記の日本語事前学習用データからサンプリングした約2.7TBのデータを使うこととし、80億パラメータと160億パラメータのLlamaのアーキテクチャを採用した、最大トークン長が8,192の比較的小規模な2つのモデルの学習を行った。上記学習のために、合計12台のマシン（8台は160億パラメータモデルの学習用、残りの4台は80億パラメータモデルの学習用にし、2つのモデルを同時に学習させた）を用い、約45～50日で２つのモデルの構築を完了した。今回の小規模モデルの学習を通して、学習環境の事前確認や整備の他、ABCIの運用に関わる問題点を明らかにしその対策を検討する等、2025年度以降の大規模LLM学習を円滑に行うための様々なノウハウを蓄積できた。

なお、NICTでは、日本の民間企業との共同研究の枠内でデータ、LLMの提供を開始しており、本研究で活用する巨大日本語学習データ、開発したLLM、その他の知見も民間企業等に提供していく予定であり、今後、これらは日本企業のLLMビジネス展開やより強力なLLMの開発に対する大きな貢献になると考えている。

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：

なし