jh231006

大規模な日本語モデル構築・共有のためのプラットフォームの形成

相澤彰子（国立情報学研究所）

概要

言語モデルの大規模化に伴い、その構築には多くの GPU 計算資源が必要とされるようになっている。本研究では、(1) 汎用型の日本語言語モデル、(2) 学術分野に特化した日本語言語モデル、の２種類のモデルの構築および評価を通して、大規模言語モデルに関する研究開発を進めるとともに、得られた技術的な知見を公開・共有した。

1　共同研究に関する情報

1.1　共同研究を実施した拠点名

- mdx

1.2　課題分野

- データ科学・データ利活用課題分野

1.3　共同研究分野 (HPCI 資源利用課題のみ)

1.4　参加研究者の役割分担

本課題では、東京大学情報基盤センターとの共同研究のもと２つのサブ課題を設定して、事前学習済言語モデルの構築および利活用を進している。

- GPU 計算基盤構築を利用した大規模言語モデル構築に関する技術的サポート（東京大学情報基盤センター（田浦健次朗））
- 【課題１】分野特化型の日本語言語モデルの構築と学術分野への適用（国立情報学研究所（相澤彰子、金澤輝一、菅原朔)、東京大学（知田悠生、江俊鋒））
- 【課題２】汎用型の日本語言語モデルの構築と性能評価（早稲田大学（河原大輔、笠原智仁、伊藤俊太朗、清水博文、今井咲良、太田聖三郎、王昊、小林俊介、村田栄樹、植松拓也、近藤瑞希)、京都大学（黒橋禎夫、村脇有吾、Chenhui Chu、清丸寛一、植田暢大、大村和正、児玉貴志））

2　研究の目的と意義

言語は学術、教育、ビジネスなど含むあらゆる知的活動の基盤であり、計算機による日本語の言語処理は、日本の社会全体のデジタル化や AI イノベーションの根幹となる情報技術である。現在の自然言語処理は、深層学習による「事前学習済み言語モデル」を中核として進展しているが、この言語モデルの学習には多くのノウハウと計算資源が必要で、単一の研究室では人材や資源の確保が困難となっている。この状況を打開するために、関連研究者が組織横断的に連携して、最新の研究成果を反映した日本語モデルを戦略的かつ迅速に構築・共有する必要がある。

上記を踏まえて本研究では、大規模情報基盤 mdx 上の GPU リソースを効率的に活用して、深層学習による日本語言語モデルの構築および公開に資する研究開発に取り組む。具体的には、【課題１】においては、汎用型言語モデルの継続事前学習による高性能な分野特化型モデルの構築について検討する。【課題２】においては、日本語における汎用型言語モデルの大規模化、高性能化を検討する。特に研究期間中で、汎用型言語モデルに対して日本語の外部知識を統合する方法を確立することを目指す。

3 当拠点公募型研究として実施した意義

大規模言語モデルは急速な進展の途上にあり、その複雑さからモデル自体のふるまいも未解明であるなど、解決するべき問題が多い。とりわけ 2022.11 に ChatGPT が公開された後は、大規模言語モデル（Large Language Models, LLMs）の社会的な影響は大きく、アカデミアの視点でモデル構築のデータやノウハウを共有しつつオープンな形で研究に取り組むことは言語モデルを使う幅広いユーザを支援するだけでなく、今後のアカデミアとしての研究基盤の維持や研究者育成のために重要である。

以上の背景のもと本課題では、大規模情報基盤 mdx 上の GPU リソースを効率的に活用した大規模言語モデルの構築に先行して取り組み、そのノウハウを関連する研究者や技術者に共有した。

4 前年度までに得られた研究成果の概要

以下に示す２つのサブ課題を設定して、それぞれ言語モデルの構築に取り組んだ。

【課題１】分野特化型日本語言語モデル構築

課題１では、医学系の学術ドメインを対象として、日本語の論文テキストを収集して言語モデルを構築した。

まず､日本語医学系論文の抄録[*1]から約 1,160 万文（1 文あたり平均 54.9 文字、約 1.8 GB）を抽出し、代表的な言語モデル構築手法である RoBERTa（A Robustly Optimized BERT Pretraining Approach）[*2] を適用し、必要となる GPU 数を同一サーバ上の A100 8 GPU と見積った。これは、１つの言語モデルの学習に１週間程度を要するものである。

次に、言語モデルのパラメタ調整と評価のために、基本的な単語予測タスク、医学系テキスト（カルテ）からの情報抽出に関する共通タスク（NTCIR -MedNLP）に加え、新たに医学系テキスト（論文）を対象とした分野分類、索引への機能ラベル付与、および索引付与タスクを定義して、実行環境を整備した。

さらに、上記をを用いて、言語モデルの構築における専門用語辞書の利用やトークン化の手法、語彙サイズの影響を評価し、最終的に４通りの組み合わせで言語モデルを構築した。現在、自然言語処理で広く普及している深層学習フレームワークである Hugging Face Hub[*3]上で公開した。

[*1] 論文抄録は、JST AIP 日独仏 AI 研究、JPMJCR20G9 の実施のため、科学技術振興機構 (JST) から提供を頂いた

[*2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.

[*3] https://huggingface.co/

【課題２】汎用型大規模日本語言語モデルの構築

課題２では、一般ドメインを対象として日本語言語モデルを構築している。一般ドメインのテキストとしては、日本語 Wikipedia (約 2,000 万文) および日本語 CC-100 (約 6 億文) を結合して用いる。CC-100 は、ウェブクロールデータ Common Crawl から抽出された各言語のテキストである[*4]。

言語モデルとしては、トークナイゼーションの単位が異なる以下の二つをターゲットとした。

一つ目は、文字単位の言語モデルとして、RoBERTa-large (パラメタ数 330M) を構築した。モデルをより頑健にするために、Whole Word Masking (WWM) を適用した。これは、ある文字をマスクして予測する場合に、同じ単語内の残りの文字もマスクする方法である。このモデルを mdx の A100 16 GPU 上で深層学習最適化ライブラリ DeepSpeed[*5]を用いて約 1 か月で学習した。学習したモデルは Hugging Face Hub 上で公開した[*6]。

二つ目は、単語単位の言語モデルとして、GPT2-XL (パラメタ数 1.5B) を構築した。単語単位の RoBERTa-large は、我々のグループですでに構築、公開[*7]しており、有効性を確認済みであるため、RoBERTa と同様に代表的な言語モデル構築法である GPT-2 (Generative Pre-Training)[*8]を用いてパラメタ数が大きなモデルを構築することとした。mdx 上の A100 8 GPU を用いて、1 エポックに 1 週間をかけて学習した。先行研究を参考に、少なくとも 10 エポックの学習を行う必要があると考え、約 2.5 か月をかけて学習を完了した。学習したモデルは Hugging Face Hub 上で公開した[*9]。

5　今年度の研究成果の詳細

前年度に引き続き、以下に示す２つのサブ課題に取り組んだ。２つのサブ課題を中心とする参加研究者の間では、必要に応じてミーティングを実施して、日本語の分かち書き（トークナイゼーション）方法、言語モデル構築における GPU の有効活用法、構築した言語モデルの公開方法などについて、互いに情報共有をした。

【課題１】分野特化型の日本語言語モデルの構築

前年度では BERT をベースにしたエンコーダー系のモデルを扱っていたが、近年の大規模言語モデルの主流は GPT に代表されるデコーダー系モデルであることから、比較的小規模なモデルである GPT-2 をベースに、学術系テキストを用いた追加学習による言語モデルの構築および評価に取り組んだ。昨今の技術動向を踏まえて、LLM の事前学習ではなく、トークナイズや生成モデルによる専門用語の抽出手法などを中心に調査や検討を進めた。

また、これまでの mdx 公募研究で得られた知見に基づき、医療分野については別途 mdx 計算資源を確保して、オープンな LLM である LLaMA（7B パラメタ/13B パラメタ）をベースに日本語医療分野論文テキストを事前追加学習したモデルを構築して、代表的な医療分野のベンチマークセットを用いて性能評価を行っ

[*4] https://data.statmt.org/cc-100/
[*5] https://github.com/microsoft/DeepSpeed
[*6] https://huggingface.co/ku-nlp/roberta-large-japanese-char-wwm
[*7] https://huggingface.co/nlp-waseda/roberta-large-japanese-seq512
[*8] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
[*9] https://huggingface.co/nlp-waseda/gpt2-xl-japanese

て、有効性を確認した。さらに、mdx 基盤を利用して別途構築された LLM-jp（13B パラメタ）についても、同様に追加学習を適用して、医療分野のベンチマークセットを用いた性能評価を行った[*10]。

【課題２】汎用型の日本語言語モデルの構築

一般ドメインを対象として、現在公開されている汎用型言語モデルより高性能な日本語モデルの構築に取り組んだ。前年度はエンコーダー系の RoBERTa ベースのモデルについて、文字単位と単語単位の日本語モデルを構築し公開した。今年度は、RoBERTa より高性能と言われている DeBERTa (v2) に注目し、この文字単位および単語単位の日本語モデルを構築した。学習コーパスには、インターネットアーカイブである Common Crawl 由来の日本語 Web コーパスおよび日本語 Wikipedia を用いた。これらは base サイズのモデル (約 1 億パラメータ) であり、A100 8 GPU を用いて、それぞれ 3 週間程度で学習を完了した。文字単位のモデルについては統合的日本語解析器 KWJA において、単語単位のモデルについては日本語言語理解ベンチマーク JGLUE において性能を評価したところ、それぞれ RoBERTa よりも高性能であることを確認した。これらのモデルは Hugging Face Hub で公開している[*11]。

さらに、上記の DeBERTa モデルの高性能化を検討したところ、DeBERTa v3 が有望であったため、DeBERTa v3 の単語単位の日本語モデル (base サイズ) を構築した。学習コーパスには、上記の日本語コーパスに加え、英語コーパスおよびプログラミングコードを含めた。このモデルの学習は、A100 16 GPU を用いて 2 週間程度で完了した。JGLUE を用いて性能を評価したところ、残念ながら DeBERTa v2 と同等のレベルであったが、英語やコードに関するタスクにも適用可能であるという利点がある。このモデルは Hugging Face Hub で公開している[*12]。

6　今年度の進捗状況と今後の展望

【課題１】においては、当初パラメタ数が 1B 以上の大規模言語モデルを mdx 上で動かして検証や活用を行うことを目標としていた。本年度は mdx 上に別プロジェクトを立てて計算資源を確保することにより、13B の追加学習までを実施しており、十分に目標を達成することができた。一方で、学術分野における専門分野オントロジーへの対応付けについては、言語モデルの急激な大規模化の影響により、実際に日本語モデルを構築・公開することはできなかったが、生成型の LLM を用いた学術文献からのキーフレーズ抽出の有効性を確認し、今後に向けた知見とすることができた。

【課題２】においては、日本語における汎用型言語モデルの高性能化に成功し、モデルを公開しており、目標を達成することができた。一方で、生成型の LLM を含む汎用型言語モデルに対して外部知識を統合する方法を調査・検討したが、これについてはモデルの構築には至らなかった。この研究は後述の LLM-jp 等の活動において継続していく予定である。

本提案では、言語モデルを構築するためのノウハウを共有し、計算機による日本語の処理を幅広く支援するための今後の方策を探ることを

[*10] これらの学習には多くの資源が必要となるため、公募研究とは別枠でプロジェクトを設定して実施した。

[*11] https://huggingface.co/ku-nlp/deberta-v2-base-japanese-char-wwm, https://huggingface.co/ku-nlp/deberta-v2-base-japanese

[*12] https://huggingface.co/ku-nlp/deberta-v3-base-japanese

長期的な目標としていた。近年の大規模言語モデルの急速な社会への普及を受けて、2023 年 5 月に LLM-jp（大規模言語モデル勉強会）が立ち上がり、mdx 上での大規模言語モデル構築の試みがはじまった。本共同研究のメンバーは LLM-jp 立ち上げの中核メンバーでもあり、昨年度から継続している本共同研究における知見は、LLM-jp での活動に生かされている。LLM-jp では 2023 年度に mdx 上の計算資源を活用して 130 億パラメタの日本語汎用モデルを構築・公開している。医療モデルについては SIP 統合ヘルスケアシステムの活動の一環としての取り組みがスタートした。さらに 2024.4 には国立情報学研究所に大規模言語モデル研究開発センターが設置されるなど、LLM をめぐる動きは活発化しており、当初の公募における目的は果たしたといえる。

7　研究業績一覧（発表予定も含む）

学術論文 （査読あり）

国際会議プロシーディングス　（査読あり）

- Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi: KWJA: A Unified Japanese Analyzer Based on Foundation Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pp.538–548, 2023.7.

国際会議発表　（査読なし）

国内会議発表　（査読なし）

- 相澤彰子:「自然言語処理による学術論文の解析：現状と展望 」人工知能学会誌 特集：「研究評価と学術情報分析」 38(3) 375-383 2023 年 5 月
- Akiko Aizawa: Natural Language Processing for Scientific Paper Analysis. Seventh International Workshop on SCIentific DOCument Analysis (SCIDOCA 2023) 、招待講演、2023 年 6 月 6 日
- 相澤彰子：2023 年度人工知能学会全国大会 特別企画セッション「日本は生成 AI を起爆剤にできるのか？」パネリスト、2023 年 6 月 6 日
- 相澤彰子：大規模言語モデルの構築とドメイン適応、第 43 回医療情報学連合大会（第 24 回日本医療情報学会学術大会)、大会企画 1「生成 AI の医療への応用」パネリスト、2023 年 11 月
- 河原大輔：日本語大規模言語モデルと言語理解ベンチマークの共進化、第 43 回医療情報学連合大会（第 24 回日本医療情報学会学術大会)、共同企画 7 人工知能学会「医学医療における AI 応用」パネリスト、2023 年 11 月

公開したライブラリ等

その他（特許，プレス発表，著書等）