jh221004

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

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)

該当せず

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】分野特化型日本語言語モデル構築（国立情報学研究所（相澤彰子、金澤輝一、菅原朔)、総合研究大学院大学（壹岐太一)、東京大学（杉本海人、鈴木淳平)、奈良先端科学技術大学院大学（荒牧英治））
- 【課題２】汎用型大規模日本語言語モデルの構築（早稲田大学（河原大輔、井手竜也、栗原健太郎、榮田亮真、吉田あいり、Ritvik Choudhary、笠原智仁、伊藤俊太朗、清水博文、今井咲良、Rachel Ung)、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)、名古屋大学（笹野遼平、山田康輔、王億祥、韓毅、塚越駿、平子潤)、東北大学（鈴木潤））

2　研究の目的と意義

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 による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である。一方で言語モデルをめぐる国際的な研究開発の動きは速く、より優れたモデルの実現に向けて多岐にわたる課題への取り組みが同時進行している。これらは英語や中国語を中心に進められており、産官学における活動を束ねても日本語についての対応

は十分とはいえない。日本語への適用では、分かち書きなど日本語固有の処理を踏まえてモデルを構築する必要があり、日本語の専門用語辞書などの活用も期待されることから、本研究を通して知見を獲得・共有する意義は大きい。

以上を踏まえて、本研究では以下の課題に取り組む。

【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

複数の学術ドメインを対象として論文テキストを収集して言語モデルを構築して、単語予測や文書分類などの基本的な言語タスクを使って性能を評価する。また、学術論文からの知識抽出に関する共通タスク（たとえば NTCIR16 Real-MedNLP）や学術知識グラフの構築のための著者同定処理などに適用して有効性を検証する。構築した言語モデルは公開する。

【課題２】汎用型の日本語言語モデルの構築と性能評価

Wikipedia やウェブ文書などを用いて汎用的な大規模言語モデルを構築し、日本語言語理解ベンチマークで性能を評価する。また、既存の言語リソース (辞書や知識グラフなど) を言語モデルに統合することを検討し、言語モデルの精度向上を目指す。なお多くの言語リソースをこれまでに構築している研究者との連携を想定している。

3　当拠点公募型研究として実施した意義

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することによって、深層学習による日本語言語モデルを構築し、日本語言語処理を必要とする研究者や技術者に公開した。とりわけ 2022.11 に ChatGPT が公開された後は、大規模言語モデル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、今後の展開のためにも重要である。深層学習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。本研究は、言語モデルを使う幅広いユーザを支援するだけでなく、言語モデルに関する最先端の研究課題に取り組むための基盤構築にも資する。

4　前年度までに得られた研究成果の概要

該当せず

5　今年度の研究成果の詳細

本テーマでは以下に示す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でいる。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ている。

【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

[*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., Zettle-

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。

【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]。

6　今年度の進捗状況と今後の展望

【課題１】分野特化型日本語言語モデル構築

本年度の予定通り、医学系のドメインを対象として、日本語論文テキストから言語モデルを構築して性能を評価し、モデルを公開することができた。

また抄録への索引付与タスクにおいて生成型の言語モデルが有効である見込みが得られた

moyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.

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

[*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

ため、RoBERTa と同様に代表的な言語モデル構築法である GPT-2 を用いてモデルを構築した。ここで、医学系論文の抄録だけではテキストの分量が不足していることが判明しており、2023 年度では全分野の論文を用いて訓練を行う予定である。

今後について、分野特化型の代表例である医学系言語モデルについて、英語では新たな大規模モデルが次々と発表され、医学系の QA タスクで既存のモデルを大幅に上回る正答率を上げている。日本語では、訓練用の論文テキストや外部知識リソースが限られること、そもそも学術情報は言語や分野の壁を越えて流通するべきものであることを踏まえると、学術分野に関しては英語のリソースやモデルを日本語処理でも活用する転移手法を検討する必要がある。以上を踏まえて今後は、言語モデルの分野オントロジーへの対応付け方法や英語への対応についても検討を進める予定である。

【課題２】汎用型大規模日本語言語モデルの構築

本年度は計画通り、汎用型大規模日本語言語モデルを構築し、公開することができた。構築したモデルを基盤とし、既存言語リソースの融合研究やアプリケーション開発などを行うことができる。

汎用型の言語モデルは、英語においては大規模化の流れが留まるところを知らず、GPT-3 (パラメタ数 175B、2020.05) の後に、Gopher (パラメタ数 280B、2021.12)、PaLM (パラメタ数 540B、2022.04) などが構築され、難易度の高いベンチマークにおける精度向上が続いている。また、大規模言語モデルに対して、辞書や知識グラフなどの外部知識を統合する研究も盛んに行われており、DictBERT や ERNIE 3.0 などが構築され、有効性が示されている。今後は、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、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。また、汎用型言語モデルを課題１に提供することによって、分野ごとのテキストで継続的に訓練し、高性能な分野特化型モデルを構築することも検討する。

7　研究業績一覧（発表予定も含む）

国内会議発表　（査読なし）

(1) 植田暢大, 大村和正, 児玉貴志, 清丸寛一, 村脇有吾, 河原大輔, 黒橋禎夫: “KWJA：汎用言語モデルに基づく日本語解析器.” 情報処理学会 第 253 回自然言語処理研究会, 2022-NL-253 (2), pp.1-14, 2022.09.

(2) 黒橋禎夫: “ニューラル自然言語処理の進展と社会課題への取り組み.” IDR（情報学研究データリポジトリ）ユーザフォーラム 2022（招待講演）, 2022.11.

(3) 相澤彰子: “大規模言語モデルと汎用性.” 「AGI 研究第 3 の波」ワークショップ, 人工知能学会第 22 回汎用人工知能研究会（口頭発表）, 2022.11.

(4) 杉本海人, 壹岐太一, 知田悠生, 金沢輝一, 相澤 彰子: “JMedRoBERTa: 日本語の医学論文にもとづいた事前学習済み言語モデルの構築と評価.” 言語処理学会第 29 回年次大会, 2023.03.

(5) 児玉貴志, 植田暢大, 大村和正, 清丸寛一, 村脇有吾, 河原大輔, 黒橋禎夫: “テキスト生成モデルによる日本語形態素解析.” 言語処理学会第 29 回年次大会, 2023.03.