Press Release

2023 年 12 月 19 日
東京工業大学
産業技術総合研究所

日本語に強い大規模言語モデル「Swallow」を公開

－英語が得意な大規模言語モデルに日本語を教える－

【要点】

○日本語能力に優れビジネスにも安心して活用できる大規模言語モデルを公開
○継続事前学習により大規模言語モデルの日本語能力を改善
○高度な日本語処理が求められる多くの場面で、生成 AI 技術の利活用を推進

【概要】

東京工業大学（以下、東工大） 情報理工学院 情報工学系の岡崎直観教授と横田理央教授らの研究チームと国立研究開発法人 産業技術総合研究所（以下、産総研）は、日本語能力に優れた生成 AI の基盤である**大規模言語モデル**（用語 1）「Swallow」を公開した（公開リンク）。本モデルは現在公開されている日本語に対応した大規模言語モデルとしては最大規模であり、オープンで商用利用が可能であるため、ビジネスに安心して用いることができる。

東工大と産総研の研究チームは、英語の言語理解や対話で高い能力を持つ大規模言語モデル（米 Meta 社 Llama 2）の日本語能力を拡張することで「Swallow」を構築した。拡張前の大規模言語モデルの高い言語処理能力を維持しながら日本語能力を強化するため、研究チームは言語モデルに日本語の文字や単語などの**語彙**（用語 2）を追加したうえで、新たに開発した日本語データを用いてモデルの構築を継続的に行う**継続事前学習**（用語 3）を行った。今回、パラメータ数が 70 億**パラメータ**（用語 4、7B）、130 億パラメータ（13B）、700 億パラメータ（70B）であるモデルを公開した。

公開リンク： https://tokyotech-llm.github.io/

●背景

米 OpenAI 社の ChatGPT や GPT-4、米 Google 社の PaLM 2 や Gemini など、大規模言語モデルの研究開発が急速に進展している。自然言語処理や人工知能の研究開発の推進、大規模言語モデルのメカニズムの解明、海外依存を理由とした安全保障上のリスク懸念、信頼できる人工知能の実現など、さまざまな動機により日本語に強い大規模言語モデルの開発が進められているが、日本語に強く、オープンかつ高性能な大規模言語モデルは少なかった。そこで、東工大と産総研は大規模言語モデルの開発に関する共同研究を開始した。この共同研究において、東工大は主にデータの語彙拡張によるモデル学習・推論効

率の改善に取り組み、産総研はモデル構築に必須である大規模計算資源として AI 橋渡しクラウド（ABCI: AI Bridging Cloud Infrastructure、図 1）を提供するとともに、主に継続学習によるモデルの日本語能力の改善を担当した。また、モデルの学習データとして、東工大が国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）のプロジェクトで開発した大規模な日本語ウェブコーパス（研究成果 3 を参照）を用いた。なお、本成果の一部は、大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（以下「NII」という）、産総研、東工大、NII が主催する勉強会 LLM-jp（NII、東北大学、東京大学、早稲田大学などが参加する LLM 研究開発チーム）が 2023 年 9 月に共同で提案して採択された、産総研 ABCI の一定部分（A ノードと呼ばれる高性能な計算ノード）を最大 60 日間占有利用する機会を提供する「大規模基盤モデル構築支援プログラム」によるものである。

図 1 産総研 AI 橋渡しクラウド ABCI

●研究成果

1. 継続事前学習により Llama2 の日本語の能力を大幅に改善

米 Meta AI 社が開発した Llama 2 シリーズはオープンかつ高性能な大規模言語モデルとして世界中から支持を集めている。また、日本語も含めて、複数の言語からなるデータで学習されているため、Llama 2 は日本語にも対応している。ところが、Llama 2 の事前学習データの約 90%は英語が占めており、日本語の割合は全体の約 0.10%に留まる。そのため、Llama 2 は英語で高い性能を示すにも関わらず、日本語の読み書きは苦手という弱点があった。

そこで、研究チームでは Llama 2 の 7B, 13B, 70B のモデルをベースに、大規模な日本語ウェブコーパスと英語のコーパスを 9:1 で混ぜたデータで継続事前学習を行い、元々の

言語モデルの能力を活かしながら日本語能力の改善を目指した。その結果、我々が採用した日本語に関するベンチマークデータにおいて、7B, 13B, 70B の全てのモデルはベースモデルよりも高い性能を示した。また、日本語コーパスのみで事前学習された同規模の日本語大規模言語モデルよりも高い性能を示すことから、継続事前学習の有効性が明らかになった。

2. 語彙拡張による大規模言語モデルの学習・推論効率の改善

Llama 2 では、**バイト対符号化**（用語 5）に基づいてテキストが**トークン**（用語 6）に区切られている。ところが、Llama 2 は英語を重視した多言語のモデルとして学習されているため、日本語の主要な単語や文字が語彙に含まれず、テキストが不自然な単位に区切られることがある。例えば、「吾輩は猫である」という 7 文字のテキストは、「<0xE5><0x90><0xBE><0xE8><0xBC><0xA9>は<0xE7><0x8C><0xAB>である」という、人間には理解しにくい 13 トークンに区切られる。これは、「吾」「輩」「猫」という漢字が語彙に収録されていないためで、**バイトフォールバック**（用語 7）により UTF-8 文字コードのバイト単位でこれらの漢字が表現されるからである。

日本語の語彙が不足している言語モデルは、日本語を不自然な単位で取り扱うことに加え、テキストをより多くのトークンで表現してしまうため、学習や生成の効率が低下する。大規模言語モデルの学習に必要な計算予算はトークン数に比例するので、逆に計算予算が一定である条件下では、テキストを少ないトークンで表現する方がより多くの情報を学習に詰め込める。また、大規模言語モデルがテキストを生成するのに要する時間はトークン数に比例するため、同じテキストを生成するのであれば、より少ない数のトークンで表現できる方が短時間で結果を出力できる。さらに、大規模言語モデルの入力や出力には、一度に扱えるトークン長の上限がある。入力をより少ないトークンで表現できる方が、タスクの指示や解き方（few-shot 事例）を多く詰め込めるので、下流タスクでの性能向上も期待できる。研究チームは Llama 2 のトークナイザに 16,000 件の日本語のトークンを追加することで、日本語テキストのトークン長を 56.2%に削減した。

3. 大規模な日本語のウェブコーパスの開発

大規模言語モデルの学習には膨大な言語データが必要である。その中でも、ウェブページを収集し、テキスト化したデータは、大規模言語モデルの構築の要である。従来、オープンな日本語大規模言語モデルの学習には、CC-100、mC4、OSCAR など既存のデータセットの日本語部分が用いられてきた。ところが、これらのデータセットでは、ウェブページの HTML をテキスト化する際のノイズが混入していることや、最新の情報や知識を収録していないという問題があった。また、これらは多言語のデータセットとして構築されているため、日本語に特化してデータの品質を高めるような工夫は取り入れられていない。

そこで、研究チームでは Common Crawl（用語 8）から配布されているアーカイブ（2020 年から 2023 年にかけて収集された 21 スナップショット分、約 634 億ページ）から日本語のテキストを独自に抽出・精錬し、約 3,121 億文字（約 1.73 億ページ）からなる日本

語ウェブコーパスを構築した。この規模は、CC-100 (約 258 億文字)、mC4（約 2,397 億文字)、OSCAR 23.10（約 740 億文字）を抜き、日本語の言語モデルの学習コーパスの中で、商用利用が可能なものとしては最大である。

●社会的インパクト

世界的に大規模言語モデルの大規模化が進むなか、日本語を扱う能力が高いものが少なかったところ、今回のモデル公開によって、高度な日本語処理が求められる日常生活・産業現場のより多くの場面で、対話システムなどの AI 技術の利活用を推進できる。なお今回公開する Swallow のライセンスは Llama 2 の LLAMA 2 Community License を継承しており、ライセンスに従う限りにおいては、研究および商業目的での利用が可能である。

Llama 2 ライセンスの公式情報 ：https://ai.meta.com/llama/license/

●今後の展開

公開された大規模言語モデルは学術と産業の両方に恩恵をもたらすと考えられる。学術分野では、日本語の大規模言語モデルの標準として研究開発に利用され、自然言語処理や人工知能分野で新たな研究成果が生み出される他、信頼できる人工知能の実現に向けた研究開発が促進される。産業分野では、API の使用などで外部の企業に依存することなく、自社で大規模言語モデルを運用できるだけでなく、特定のタスクに特化したモデルにチューニングができる。日本語に強くオープンな大規模言語モデルが登場したことで、日本における大規模言語モデルの研究開発・活用がさらに促進され、製品開発や技術革新が進むと考える。

●付記

産総研が構築・運用する AI 橋渡しクラウド（ABCI: AI Bridging Cloud Infrastructure）の「大規模言語モデル構築支援プログラム」、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の「次世代人工知能・ロボットの中核となるインテグレート技術開発」プロジェクト（JPNP18002)の「熟練者観点に基づき、設計リスク評価業務における判断支援を行う人工知能適用技術の開発」、その他の支援によって実施された。

【用語説明】

(1) **大規模言語モデル**：テキストの現れやすさをモデル化したもので、与えられた文脈（問いかけ）に対して続くテキスト（応答）を予測できる。

(2) **語彙**：言語モデルが扱えるトークンの集合のこと。

(3) **継続事前学習**：すでに学習されている大規模言語モデルに対し、追加で事前学習を行う手法。異なる言語やドメインで言語モデルを活用するときに用いられる。

(4) **パラメータ**：大規模言語モデルなどのニューラルネットワークの挙動を決定する数値の個数であり、ニューラルネットワークの規模を表す指標の一つ。

(5) **バイト対符号化**：言語データの文字や文字列の統計情報に基づき、指定された語彙数の範囲内で最適な語彙を求めるアルゴリズム。

(6)　**トークン**：言語モデルがテキストの入出力を行う時の単位。人間にとっては単語単位でトークンを構成すると分かりやすいが、全ての単語を予め把握しておくのは難しいため、大規模言語モデルでは単語よりも小さな部分単語や文字をトークンの単位とすることがある。

(7)　**バイトフォールバック**：扱いたい文字が語彙に含まれていないとき、その文字をバイト列（通常は UTF-8 のバイト列）として表現すること。

(8)　**Common Crawl**：ウェブサイトを巡回・収集（クローリング）し、そのアーカイブを無償で提供している非営利団体。

【問い合わせ先】
東京工業大学 情報理工学院 情報工学系　教授
岡崎　直観
Email: okazaki@c.titech.ac.jp
TEL: 03-5734-2186
FAX: 03-5734-3494

産業技術総合研究所
情報・人間工学領域 研究企画室
ith-liaison-ml@aist.go.jp

【取材申し込み先】
東京工業大学 総務部 広報課
Email: media@jim.titech.ac.jp
TEL: 03-5734-2975　FAX: 03-5734-3661

産業技術総合研究所
ブランディング・広報部 報道室
Email:hodo-ml@aist.go.jp