ビジネスのドメインに対応した日本語大規模言語モデルの開発

近江崇宏[1] 高橋洸丞[1] 有馬幸介[1] 石垣達也[2]
[1]ストックマーク株式会社 [2]産業技術総合研究所
{takahiro.omi,kosuke.takahashi,kosuke.arima}@stockmark.co.jp ishigaki.tatsuya@aist.go.jp

概要

この一年ほどで日本語に対応した大規模言語モデルが活発に開発されている。今回、我々はビジネスのドメインに対応した 130 億パラメータの日本語の大規模言語モデルの開発を行い、事前学習済みモデル及び指示学習済みモデルの公開をおこなった。本論文では、事前学習や指示学習の詳細や、モデルの評価、特にビジネスのドメインでの優れた質問回答性能について報告する。また、最新の情報の継続的学習に関する初期的な検証結果についても簡単に報告する。

1 はじめに

近年、大規模言語モデル (LLM) のユーザーの指示に従って柔軟に応答を行う能力や、少数の例の提示により新規な言語タスクを解く能力が著しく向上し、注目を集めている [1, 2]。ChatGPT の登場を契機に、一般社会でも LLM の活用が進みつつある。これまで開発されてきた LLM の多くは、英語を主な対象としてきたが[1, 3, 4]、この一年ほどで、日本語に対応した LLM も多く開発されるようになってきた[5, 6, 7, 8]。

これまで開発されてきた日本語の LLM のほとんどは、主に Wikipedia や mC4 などの Common Crawl 由来の幅広いドメインや話題を含むコーパスから学習を行った汎用型の LLM である[5, 6, 7, 8]。その一方で、特定のドメインのタスクを扱う際には、そのドメインの専門用語などの知識を獲得することが重要であり、英語のモデルでは科学、医学、金融などのドメイン特化型のモデルも開発されてきている[9, 10, 11]。

本研究ではビジネスのドメインに対応した LLM を開発する。そのために、既存の一般のドメインのコーパスに加えて、ビジネスドメインのコーパスを合わせたデータセットを作成し、これを用いて 130 億パラメータの LLM の事前学習を行った。そして、ビジネスのドメインでの質問回答の性能を評価するために、ビジネスに関する質問を 50 問作成し、ここで開発したモデルが既存のモデルに比べてビジネスのドメインの知識を獲得していることを確かめた。開発した事前学習および指示学習済みのモデルは stockmark-13b 及び stockmark-13b-instruct として、Huggingface Hub にて公開されている[12, 13]。

また、ビジネスのドメインでは最新の情報が重要であるため、LLM に対して最新の情報を継続的に学習させていくことも重要なテーマである。そこで、事前学習を行った後に収集されたデータを用いた継続的学習に関する初期的な検証結果についても簡単に報告を行う。

2 事前学習

本節では、事前学習についての詳細を記述する。

2.1 データセット

本研究では、ビジネスのドメインに対応した LLM を開発するために、日本語の LLM の事前学習で用いられる、一般ドメインの Wikipedia や CC100、mC4、Common Crawl などのコーパスに加えて、ビジネスドメインの公開されている Web ページ (2023 年 9 月まで) を収集し、クリーニングを行った Stockmark Web コーパス（非公開）と特許のコーパスを用いて、合計 2200 億トークンからなるデータセットを作成した。各コーパスのデータ量は表 1 にまとめられてい

表 1: 事前学習のデータセット構成

データセット	前処理後のトークン数 [billion]
Stockmark Web Corpus	9.1
特許	34.8
Wikipedia	1.0
CC100	10.9
mC4	53.2
Common Crawl	112.9

表 2: 日本語の指示学習のデータセット

データセット	略称	データ数	言語
databricks-dolly-15k-ja	dolly	15015	英語のデータセットを日本語に翻訳
oasst-89k-ja	oasst	88838	英語のデータセットを日本語に翻訳
alpaca_ja	alpaca	51716	英語のデータセットを日本語に翻訳
ichikara-instruct	ichikara	1003 (2023/11/10 時点)	日本語

る。

事前学習時には、コーパスに応じて重みを変えた。具体的には、1 エポックの中で、基本的には各データは 1 回のみ現れるが、Stockmark Web Corpus と Wikipedia のコーパスに関しては、それぞれのデータが 2 回現れるようにした。

2.2 モデル

本研究では、130 億パラメータの Llama モデルのアーキテクチャを用いて [4]、ゼロから事前学習を行った（本研究では Meta 社が公開している事前学習済み Llama モデルを利用した継続学習は行なっていない）。モデルのハイパーパラメータは Meta 社が公開している 130 億パラメータのモデルと同じである。

2.3 分散学習

事前学習には AWS が開発した Trainium と呼ばれるハードウェアアクセラレータを用いて事前学習を行った[14]。実際には、Trainium を搭載した trn1.32xlarge インスタンスを 16 node 用いた。また Trainium 上での分散学習を行うためのライブラリとして neuronx-nemo-megatron を用いた [15]。

2.4 その他

今回の事前学習では、2.2 節で説明したデータセットを用いて 1 epoch の学習を行い、学習には 30 日を要した。以下では、ここで開発された事前学習済みモデルを stockmark-13b と言及する。

3 指示学習

LLM にユーザーの指示に従った応答をさせるためには、事前学習の後に指示学習を行う必要がある。指示学習とは、さまざまな指示に対する望ましい応答を教師あり学習により学習を行うものである。これまでに、いくつかの日本語の指示学習のためのデータセットが開発されてきている（表 2）[16-19]。多くのデータセットは英語のデータセットを日本語に翻訳したものであるが[16-18]、日本では馴染みのない話題が含まれている、機械翻訳により日本語として意味が通じないデータも含まれているといったデータの質に関する課題がある。ichikara-instruct は、このような課題に対応するために、現在構築中の日本語のデータセットである[i][19]。その質は既存の英語のデータセットを翻訳したデータセットと比べると高いものになっている。

5 節では、それぞれのデータセットで指示学習を行い、用いるデータセットに応じて、下流タスクでの性能がどのように変わるかを調べた。指示学習では、LoRa tuning を用いた。

4 ベンチマーク

本研究で開発された事前学習モデルや指示学習モデルの性能を評価する、二つのベンチマークを説明する。

4.1 Stockmark Business Questions

本研究では、ビジネスのドメインに対応した LLM を開発することを目的としている。そのために、今回開発した stockmark-13b が既存のモデルに比べてビジネスのドメインの知識を獲得しているかを検証するために、ビジネスに関する知識を問う質問を 50 問作成した [20]。具体的には、表 3 にまとめられて

表 3: Stockmark Business Questions の例

カテゴリー	質問例
時事問題	2023 年 4 月にロシアのウクライナ侵攻を受けて NATO に加盟した国は？
企業活動	2022 年以降で、ペロブスカイト太陽電池の開発をしている日本のスタートアップ企業は？
社会課題	カーボンニュートラルとは？
トレンド	ダークストアとは？

[i] 本研究では、2023/11/10 時点で利用可能であった 1003 件のデータを用いた。

いるような、時事、企業活動、社会課題、トレンドに関する知識を問うような問題である。それぞれのLLMの応答は、人手で評価した。

4.2 lm-evaluation-harness

上のビジネスドメインでの質問回答の精度に追加して、一般の日本語の言語理解についても評価を行った。このために、lm-evaluation-harnessのライブラリを用いて、7 の言語タスクでの性能評価を行った[21]。一般に、LLMの性能はどのようなテンプレートを用いるのかに、大きく依存することがあり、性能を高めるためにプロンプトを調節することがよく行われている。そのため、今回は、それぞれのタスクに対して複数のテンプレートが用意されている場合には、それぞれのテンプレートで評価を行い、最も高いスコアを採用した。用いたタスクやテンプレートの詳細については付録を参照のこと。

5 結果

5.1 同規模のモデルとの比較

今回開発したstockmark-13bと同規模の100億パラメータクラスの日本語のLLMとの比較を行った。

まずは、Stockmark Business Questionsを用いた評価を行う。ここでは同規模のモデルに追加して、ChatGPT (gpt-3.5-turbo-0613)も評価対象に加えた。ChatGPT以外のモデルに関しては、事前学習のみを行ったベースのモデルに対して、一律alpaca_jaのデータセットを用いて指示学習したモデルを用いて評価を行った。これは、指示学習で用いるデータセットにより性能が大きく変わるので、条件を揃えるためである。

Stockmark Business Questionsに対するそれぞれのモデルの正解率は表4にまとめられている。この結果から、今回開発したstockmark-13bは、ビジネスドメインの学習データを多く学習したことにより、既存のモデルに比べて、ビジネスのドメイン知識を多く獲得していることが確認できた。参考のため、実際のLLMの応答例を付録の表A3に示す。

次に、lm-evaluation-harnessを用いて、一般の日本語の言語理解の能力の評価を行った。結果は表5にまとめられている。今回調べたモデルの中では、最も高いスコアを示した。各タスクのスコアなどの詳細は付録を参照されたい。

表4: Stockmark Business Question による評価

モデル	正解率
stockmark/stockmark-13b	0.80
llm-jp/llm-jp-13b-v1.0	0.40
pfnet/plamo-13b	0.38
matsuo-lab/weblab-10b	0.34
ChatGPT (gpt-3.5-0613)	0.42

表5: lm-evaluation-harness による評価

モデル	平均スコア
stockmark/stockmark-13b	0.527
llm-jp/llm-jp-13b-v1.0	0.469
pfnet/plamo-13b	0.457
matsuo-lab/weblab-10b	0.445

表5のスコアはそれぞれのタスクに対して複数のプロンプトのテンプレートで評価を行い、最も高いスコアを採用したものである。その一方で、最も低いスコアを採用した場合には、stockmark-13bは全体の中では低い値を取ることもわかった。このことは、プロンプトを適切に調節すれば高い性能を得られるが、プロンプトの違いに対する性能の揺れが大きいことを意味する。

5.2 指示学習

次に、指示学習において用いるデータセットに応じて下流タスクでの性能がどう異なるかを調べた（表6; lm-eval-harnessでの評価の各タスクのスコアは付録の表A2にまとめられている）。lm-evaluation-harnessとStockmark business questionsでの評価では、いずれの場合もichikaraデータセットを用いる場合に、一番スコアが高かった。特に、Stockmark business questionsでの評価では、用いるデータセットにより、正解率に差が見られた。英語のデータセットを翻訳して作成したalpaca, dolly, oasstなどのデータセットでは質の低いサンプルも含まれていることを先に述

表6:指示学習モデルの評価

データセット	lm-eval-harness	stockmark business questions での正解率
ichikara	0.547	0.86
alpaca	0.545	0.80
dolly	0.547	0.72
oasst	0.510	0.70

べた。そのようなデータセットに含まれるノイズは指示学習で LLM が内在する知識の表現を壊してしまうことが考えられる。そのため、Stockmark business questions のような知識を問うような問題では、alpaca, dolly, oasst のデータセットを用いた場合には、正解率が下がったことが考えられる。また、ichikara データセットで用いたデータ数は 1000 程度と他のデータセットに比べて少ないものの、高い性能が得られたことは、指示学習を行う際にはデータの質が重要であることが示唆される。

ichikara データセットで指示学習を行ったモデルは stockmark-13b-instruct として公開されている。

6 継続的学習

ビジネスのドメインでは最新の情報が重要であることから、最新の情報で LLM を継続的に学習することも重要なテーマである。この時、問題になるのが破壊的忘却と呼ばれる現象である。これは、最新の情報を追加で学習した時に、過去の学習で得た知識を忘れてしまうという現象である。このような問題に対して、先行研究では、追加データで学習を行うときに、過去の学習データを混ぜることで破壊的忘却を抑止できることが示されている [22]。

今回の事前学習で用いた Stockmark Web Corpus は 2023 年 9 月までのデータを含むが、その後に新たに得られた 2023 年 10 月〜11 月のデータを追加データとして、追加学習を 1 回のみ行う設定で、継続的学習に関する初期的な検証をおこなった。以下では、事前学習で用いた Stockmark Web Corpus を D_0、追加のデータを D_1 で表す。追加学習では、先行研究にならって、D_1 に D_0 からランダムにサンプルしたデータを加えて得られる D_1^*のデータセットを用いて、stockmark-13b に対して追加の学習を行う。ここで、D_0 から追加されるデータ量はパラメータ r によりコントロールされ、D_1 のデータ量に r をかけたものとする。例えば、$r = 0.0$ の時は追加データ D_1 のみを用いて追加学習を行うことに対応し、$r = 0.1$ の時には追加データ D_1 にそのデータ量の 1 割の D_0 からサンプルされたデータを加えて追加学習を行うことに対応する。

ここでは、r の値をどの程度にすれば、破壊的忘却を防ぎつつ、追加データから最新の情報に関する知識を獲得できるのかを検証する。r の値としては 0.0, 0.1, 0.3 を考えた。それぞれの r に対して作成された D_1^*のデータセットで追加の学習を行い、モデルの評価指標としては過去のデータ D_0 と追加データ D_1 に対する validation loss (val_loss)を用いた。

表 7: 追加学習の評価

	val_loss (D_0: 事前学習のデータ)	Val_loss (D_1: 追加のデータ)
(base)	2.11	2.25
$r = 0.0$	2.19	2.05
$r = 0.1$	2.14	2.05
$r = 0.3$	2.12	2.04

表 7 は結果をまとめたものであり、表中の base は事前学習のみをおこなったモデルを表したものである。r の値によらず、追加データ D_1 に対する val_loss は追加学習後に下がっており、追加学習により追加データに対するモデルの適合度が上がっていることが示されている。その一方で、過去のデータを追加学習に用いない時 $(r = 0.0)$ の時には、過去のデータ D_0 に対する val_loss が大きく上がってしまい、モデルの適合度が下がってしまっており、破壊的忘却に対応した現象が見られる。過去のデータを一定混ぜた $(r = 0.3)$ の時には、過去のデータに対する val_loss の値の減少を抑えることができている。このことは、先行研究で示された結果と同様に、追加学習において、過去の学習データを一定量混ぜることで、破壊的忘却を抑制できることを示している。

7 まとめ

本研究では、ビジネスのドメインに対応した日本語 LLM の開発を目指し、ビジネスドメインのコーパスを含めたデータセットで事前学習をおこなった stockmark-13b を開発した。そして、事前学習の詳細や、指示学習、継続的学習での検証結果を簡単に報告した。

継続的学習の検証では、初期的な検証として、事前学習後に得られた 2 ヶ月のデータを追加のデータとして、追加かの学習を一回のみ行う実験を行った。その一方で、実際には、追加の学習は一回だけ行うものではなく、一定の間隔で繰り返し行うことが想定される。そのような状況を想定したような検証は今後の課題である。また、今回の検証ではモデルの性能の指標として val_loss を用いたが、実際のユースケースに即したようなタスクを用いて評価することも今後の課題である。

謝辞

本研究の一部は AWS の LLM 開発支援プログラムの支援を受けました。

参考文献

1. *Language Models are Few-Shot Learners*. **T. Brown, et al.** Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020.
2. *Training language models to follow instructions with human feedback.* **L. Ouyang, et al.** Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 2022.
3. *PaLM 2 Technical Report.* **R. Anil, et al.** arXiv:2305.10403, 2023.
4. *Llama 2: Open Foundation and Fine-Tuned Chat Model.* **H. Touvron, et al.** arXiv:2307.09288, 2023.
5. cyberagent/calm2-7b: https://huggingface.co/cyberagent/calm2-7b
6. llm-jp/llm-jp-13b-v1.0: https://huggingface.co/llm-jp/llm-jp-13b-v1.0
7. pfnet/plamo-13b: https://huggingface.co/pfnet/plamo-13b
8. matsuo-lab/weblab-10b: https://huggingface.co/matsuo-lab/weblab-10b
9. *Galactica: A Large Language Model for Science*. **R. Taylor, et al.** arXiv:2211.09085, 2022.
10. *Large language models encode clinical knowledge*. **K. Singhal, et al.** Nature 620, 2023.
11. *BloombergGPT: A Large Language Model for Finance.* **S. Wu, et al.** arXiv:2303.17564, 2023.
12. stockmark/stockmark-13b: https://huggingface.co/stockmark/stockmark-13b
13. stockmark/stockmark-13b-instruct: https://huggingface.co/stockmark/stockmark-13b-instruct
14. https://aws.amazon.com/machine-learning/trainium/
15. https://github.com/aws-neuron/neuronx-nemo-megatron
16. databricks-dolly-15k-ja: https://github.com/kunishou/databricks-dolly-15k-ja
17.oasst1-89k-ja: https://github.com/kunishou/oasst1-89k-ja
18. alpaca_ja: https://github.com/shi3z/alpaca_ja
19. *ichikara-instruction: LLM のための日本語インストラクションデータの構築.* **関根聡, 安藤まや, 後藤美知子, 鈴木久美, 河原大輔, 井之上直也, 乾健太郎.** 言語処理学会第 30 回年次大会(2024).
20. Stockmark business questions: https://huggingface.co/datasets/stockmark/business-questions
21. lm-evaluation-harness: https://github.com/Stability-AI/lm-evaluation-harness/tree/effdbeaf742e74ea1787871e99272c12146ba346
22. *Fine-tuned Language Models are Continual Learners.* **T. Scialom, T. Chakrabarty, and S. Muresan.** In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022.

付録:

表 A1: タスクとプロンプトのバージョン

タスク	タスクのバージョン	プロンプトテンプレートのバージョン	Fewshot の提示数
JSQuAD	jsquad-1.1	0.1, 0.2, 0.3	2
JAQKET	jaqket_v2-0.2	0.1, 0.2, 0.3	1
JCommonsenseQA	jcommonsenseqa-1.1	0.1, 0.2.1, 0.3	3
XWinograd	xwinograd_ja	-	0
JNLI	jnli-1.3	0.2, 0.3	3
MGSM	mgsm-1.0	0.0, 0.3	5
XLSum	xlsum_ja-1.0	0.0, 0.3	1

表 A2: lm-evaluation-harness による評価

モデル	JSQuAD	JAQKET	JCommonsenseQA	XWinograd	JNLI	MGSM	XLSum	average
stockmark/stockmark-13b	0.785	0.860	0.747	0.734	0.392	0.096	0.078	**0.527**
llm-jp/llm-jp-13b-v1.0	0.690	0.752	0.645	0.679	0.395	0.016	0.108	0.469
pfnet/plamo-13b	0.679	0.698	0.562	0.675	0.445	0.036	0.103	0.457
matsuo-lab/weblab-10b	0.696	0.457	0.652	0.668	0.542	0.040	0.057	0.445

データセット	JSQuAD	JAQKET	JCommonsenseQA	XWinograd	JNLI	MGSM	XLSum	average
ichikara	0.790	0.863	0.803	0.733	0.485	0.076	0.077	0.547
alpaca	0.731	0.849	0.807	0.727	0.565	0.068	0.067	0.545
dolly	0.772	0.865	0.771	0.733	0.528	0.084	0.076	0.547
oasst	0.777	0.848	0.805	0.738	0.246	0.072	0.081	0.510

表 A3:「2023 年 3 月に経営破綻したアメリカの 2 つの銀行は?」という質問に対する各モデルの応答。この例では stockmark-13b のみが正しい答えを出力している。

モデル	応答例
stockmark/stockmark-13b	2023 年 3 月に経営破綻した米国の 2 つの銀行は、シリコンバレー銀行とシグネチャー銀行である。
llm-jp/llm-jp-13b-v1.0	2023 年 3 月に経営破綻したアメリカの 2 つの銀行は、ウェルズ・ファーゴとバンク・オブ・アメリカである。
pfnet/plamo-13b	2023 年 3 月に経営破綻した 2 つの銀行は、バンク・オブ・アメリカとシティ・グループでした。
matsuo-lab/weblab-10b	2023 年 3 月に経営破綻したアメリカの 2 つの銀行は、Bank of America と Wells Fargo です。
ChatGPT (gpt-3.5-0613)	2023 年 3 月に経営破綻したアメリカの銀行については、現時点では具体的な情報は提供されていません。(以下省略)