語彙置換継続事前学習による日英バイリンガルモデルの構築と評価

麻場直喜 野崎雄太 中島大 佐藤諒 池田純一 伊藤真也 近藤宏 小川武士 坂井昭一朗 川村晋太郎
株式会社リコー デジタル戦略部 デジタル技術開発センター 言語 AI 開発室
{naoki.asaba, yuta.nozaki1, dai.nakashima, ryo.sato4, j-ikeda, shinya.itoh, hiroshi.xx.kondoh, takeshi.t.ogawa, shohichiroh.sakai, shintaro.kawamura}@jp.ricoh.com

概要

ニューラルベースの LLM (Large Language Model) は，大量の英語データで事前学習が行われることが多い．これに英語以外の言語データで継続事前学習を行うことで，比較的少量の資源でその言語の高い性能を得ることが期待される．本稿では，オープンな LLM である事前学習済み Llama 2 13B Chat に対して日英 2 言語データで語彙置換継続事前学習を行い，LLM ベンチマーク 2 種で性能評価した結果について報告する．トークナイザは語彙置換により日英 2 言語に適応させ，カリキュラム学習により言語間の知識転移と学習効率化を図った．ベンチマーク結果より，日本語性能の向上を確認した．

1 はじめに

近年盛んに開発されている LLM の事前学習には大量の学習データが必要であり，豊富な言語資源を有する英語データが主に用いられることが多い．一方で特定の言語能力に特化した LLM，例えば日本語 LLM を開発する場合，単純には大量の日本語データを用いて事前学習を行う方法が考えられ，その方法で我々は昨年 6B パラメータの日本語 LLM を開発している[1]．しかし，英語と比べて日本語の言語資源は少ないため，日本語データのみを用いて英語 LLM と同等の言語処理能力を持たせることは難しい．そこで日本語データだけでなく日本語と英語の 2 言語データで事前学習を行うことで，言語の普遍的な特性や知識を英語データからも学習させつつ，それらの能力を日本語処理性能へと転移させる言語間知識転移が期待できる[2]．また，主に英語データで事前学習が行われた LLM に対して他の言語データで継続事前学習を行うことでも知識転移が期待できる[3]．日本語適応の事例として，主に英語データで事前学習が行われたオープンな LLM である Llama 2[4]に対して日本語データで継続事前学習を行った ELYZA-japanese-Llama-2-13b[5]は，Llama 2 の事前学習において日本語学習データが少なかったことを補いつつ，優れた英語能力を日本語能力に転移させていると考えられる．一方で，Llama 2 はトークナイザが主に英語に特化しており，日本語をトークナイズすると 1 トークンが平均 1 文字未満になるため，日本語は英語に比べて学習コストや推論コストが高く，一度に入出力可能な文章も少ないという課題がある．この対策として Llama 2 に対してトークナイザに日本語語彙を追加して継続事前学習を行った Swallow-13B[6]や ELYZA-japanese-Llama-2-13b-fast[5]がある．また，我々は元の語彙数を変えずに語彙を置換して継続事前学習を行う語彙置換継続事前学習を提案している[7]．これらの手法により日本語での学習・推論コストと入出力長を最適化することが可能である．

本稿では，事前学習済み Llama 2 13B Chat に対して日英 2 言語データを用いてトークナイザを日英 2 言語に適応させた上で，日英 2 言語の学習データで語彙置換継続事前学習を行い，LLM ベンチマーク 2 種で性能評価した結果について報告する．構築した日英 2 言語 LLM を以後 Ricoh-13B と表記する．

2 モデル構築

2.1 モデル設計

本稿で報告する Ricoh-13B は Transformer[8]の decoder に特化した自己回帰言語モデルであり，アーキテクチャ及びモデル初期重みには Meta 社の事前学習済み Llama 2 13B Chat[i][4]を採用した．表 1 に主なアーキテクチャ仕様を示す．

トークナイザは，語彙数に関する我々の関連研究[9]を元に，Llama 2 のトークナイザに対して前述の語彙置換継続事前学習の語彙置換を適用して，低頻

[i] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

度語彙を日本語語彙に置換した．置換先の語彙は Llama 2 と同じ仕様で SentencePiece[ii][10]を用いて BPE を構築して作成し，その学習データには日英 2 言語データを用いた．これにより主な英語性能は維持しつつ日本語への適応及び知識転移を図る．表 2 に Llama 2 と Ricoh-13B の語彙数の内訳と LPT (Length per Token: 文字列をトークナイズしたときの 1 トークンあたりの平均文字数) を示す．日本語の語彙が大幅に増加し，逆に英語の語彙が減少している．これにより日本語の LPT が約 2 倍となり，日本語の学習及び推論の速度向上と，入出力可能な日本語文字数の増大を実現している．

2.2 学習データとカリキュラム

Ricoh-13B の学習データは日英 2 言語の公開データとし，日本語は我々が昨年開発した 6B パラメータの日本語 LLM[1]と同様に Wikipedia (ja)，CC100 (ja)，OSCAR (ja)，mC4 (ja)，英語は RedPajama データ v1[iii][11]のうち Wikipedia (en), Book, Stack Exchange, C4 を用いた．学習順序には，カリキュラム学習に関する我々の関連研究[12]と同様に下記カリキュラム学習 2 種を採用した．具体的な学習データの学習順序と日英データ量の比率を図 1 に示す．

データ品質　一般にデータの品質と量にはトレードオフの関係があり，高品質なデータを大量に用意することは難しい．そこでデータ品質に応じた学習順序を設定する．具体的には，学習の序盤は少量の高品質データを学習して学習の安定化を図る．中盤は大量の低品質データを学習して知識獲得を図る．終盤は再度少量高品質データを学習して高品質なテキストの生成に適応させる．

言語間データ比率　Llama 2 13B Chat がもつ英語知識の日本語への転移と，継続事前学習の安定化を図るために，学習の前半は英語データの比率を高くする．学習の後半は逆に日本語データの比率を高くすることで，日本語の生成性能の向上を図る．

2.3 学習の詳細設定

Ricoh-13B の学習には Amazon EC2 Trn1 インスタンス (trn1.32xlarge) 64 ノードを用いた．学習フレームワークは Amazon Web Services, Inc.が公開している AWS Neuron Reference for NeMo Megatron の Llama 2 学習コードサンプル[iv]をベースとした．AWS Neuron SDK[v]のバージョンは 2.14 を用いた．

表 3 に，Ricoh-13B の学習における主なハイパーパラメータを示す．バッチサイズは Llama 2 と同様に 4M トークンとした．学習率は複数条件で予備実験を行い，Loss のスパイクが少ない 8.0×10^{-5} を採用した．カリキュラム学習における学習データ切り替えの際は都度学習率ウォームアップを実施して学習の安定化を図った．

表 1 Ricoh-13B の主なアーキテクチャ仕様

パラメータ	値
Context Length	4,096
Num Layers	40
Hidden Size	5,120
Num Heads	40
Vocab Size	32,000

表 2 トークナイザ語彙数の内訳と LPT の比較

モデル	語彙数内訳		LPT	
	日本語	英語	日本語	英語
Llama 2	1k	28k	0.85	4.01
Ricoh-13B	17k	14k	2.01	3.99

3 評価

2 節で述べた学習を行った Ricoh-13B に対して下流タスク性能の評価を行った．

3.1 評価方法

本評価においては公開評価データセットとベンチマークツール 2 種を用いた自動評価を実施した．具体的には，llm-jp-eval[vi]で日本語の性能を，lm-evaluation-harness[vii] で英語の性能を評価した．評価の一貫性を確保するため，評価条件や推論動作パラメータのほとんどは各ベンチマークツールのデフォルト設定を採用した．

[ii] https://github.com/google/sentencepiece
[iii] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T
[iv] https://github.com/aws-neuron/neuronx-nemo-megatron/blob/main/nemo/examples/nlp/language_modeling/llama_13b.sh
[v] https://awsdocs-neuron.readthedocs-hosted.com/
[vi] https://github.com/llm-jp/llm-jp-eval
[vii] https://github.com/EleutherAI/lm-evaluation-harness

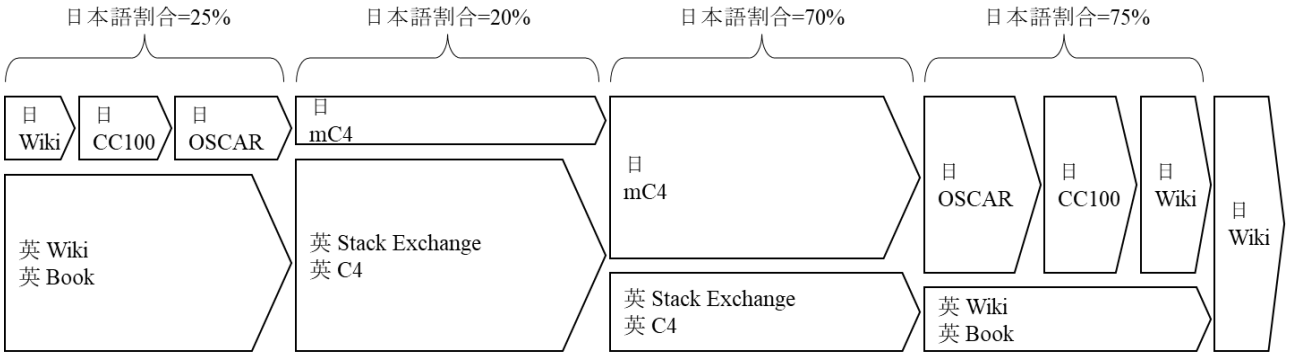

図 1 Ricoh-13B のカリキュラム学習における各学習データの順序と日英データ量の比率．ブロック矢印のように図の左から右の方向に学習が進み，ブロック矢印の太さや長さがデータ量の大小関係の概要を示す．

llm-jp-eval　LLM 勉強会(LLM-jp)によって開発された LLM 向けの日本語自動評価ツールである．ツールのバージョンは 1.0.0 を用いた．本評価では llm-jp-eval リーダーボード[viii]に準拠して，各評価タスクのスコアの他に MC (Multi-Choice QA)，NLI (Natural Language Inference)，QA (Question Answering)，RC (Reading Comprehension)の 4 つのカテゴリ別平均スコア及びそれらの平均スコアを算出した．主な推論制御パラメータを表 4 に示す．比較対象モデルはベースの Llama 2 13B 及び Llama 2 13B Chat と，公開されている日本語 LLM のうち 13B パラメータかつトークナイザを日本語適応している事前学習モデル 5 種とした．

lm-evaluation-harness　EleutherAI によって開発された LLM 向けの英語自動評価ツールである．ツールのバージョンは 0.4.0 を用いた．評価タスクは GLUE Leaderboard[ix]及び Hugging Face Open LLM Leaderboard[x]の評価タスクのうち，本ツールが対応している 8 タスク及び 6 タスクとした．推論制御パラメータの Few Shot 数を表 5 に示す．比較対象モデルはベースの Llama 2 13B 及び Llama 2 13B Chat と，Llama 2 に対してトークナイザを日本語適応した上で継続事前学習を行った Swallow-13B 及び ELYZA-japanese-Llama-2-13b-fast とした．

3.2 評価結果と考察

表 6 に llm-jp-eval を用いた日本語ベンチマーク結果を示す．ベースの Llama 2 と比較すると，我々の Ricoh-13B は Avg.スコアにおいて 8〜9 ポイント向

表 3 Ricoh-13B の学習における主なハイパーパラメータ

パラメータ	値
Precision	bfloat16 + Stochastic Rounding
TP: Tensor Parallel	8
PP: Pipeline Parallel	4
Batch Size (tokens)	4M
Learning Rate	8.0×10^{-5}
Optimizer	AdamW

表 4 llm-jp-eval での推論制御パラメータ

パラメータ	値
num_few_shots	4
top_p	1.0
top_k	0
temperature	0.1
repetition_penalty	1.0

表 5 lm-evaluation-harness での Few Shot 数

タスク	num_fewshot
GLUE 全 8 種	3
ARC	25
HellaSwag	10
MMLU, Winogrande, GSM8K	5
TruthfulQA	0

[viii] https://wandb.ai/llm-jp-eval/test-eval/reports/llm-jp-eval---Vmlldzo1NzE0NjA1?accessToken=s09hm7xrqg43ls8i25am6t0r7iwjpninwzeelqqgbx53zivlm9s04ixfpv3xgiwm

[ix] https://gluebenchmark.com/leaderboard

[x] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

表 6 llm-jp-eval 日本語ベンチマーク結果．スコアは%表記．9 種の評価タスクスコアの他，MC, NLI, QA, RC はそれらの評価カテゴリ別平均スコアを，Avg.はそれら評価カテゴリ別平均スコア 4 種の平均スコアを示す．Avg.に含まれない JSTS 評価スコアは付録 A の表 8 に添付する．Stockmark-13B において本ツールでは意図した形式の生成文が得られないタスクがあったため，当該タスクは評価対象外とした．

モデル	Avg.	MC	NLI	QA	RC	Jamp	JaNLI	JCommonseQA	JEMHopQA	JNLI	JSeM	JSICK	JSQuAD	NIILC
Llama 2 13B	58.9	67.7	39.6	43.2	85.1	41.1	54.3	67.7	44.4	15.8	55.9	30.7	85.1	42.0
Llama 2 13B Chat	57.9	62.5	47.3	37.9	84.0	**46.3**	54.6	62.5	43.8	31.4	65.7	38.4	84.0	32.0
PLaMo-13B	47.4	22.6	44.0	46.6	76.2	35.9	50.0	22.6	**51.9**	14.7	59.1	60.4	76.2	41.4
LLM-jp-13B	46.3	22.3	42.5	43.1	77.4	37.6	49.7	22.3	47.9	31.7	67.2	26.1	77.4	38.3
Stockmark-13B	-	22.2	-	-	73.0	-	49.6	22.2	-	-	-	-	73.0	50.2
Swallow-13B	66.2	**78.0**	39.7	**56.9**	**90.0**	36.2	50.0	**78.0**	50.5	34.0	49.7	28.4	**90.0**	63.3
ELYZA-japanese-Llama-2-13b-fast	62.5	74.3	44.9	45.0	86.0	35.9	52.9	74.3	45.3	34.1	**68.4**	32.9	86.0	44.6
Ricoh-13B (ours)	**67.0**	72.7	**50.4**	56.3	88.5	34.2	**56.9**	72.7	49.0	**34.2**	52.2	**74.6**	88.5	**63.6**

上し，評価カテゴリ別平均スコアにおいても全て向上している．これは英語を主とする Llama 2 に対して日本語適応の語彙置換継続事前学習を行うことで，日本語性能を向上できていると考えられる．日本語 LLM 各種と比較すると，Avg.スコアにおいて我々の Ricoh-13B が最も高いスコアであった．評価カテゴリ別平均スコアや各評価タスクスコアにおいては Swallow-13B や ELYZA-japanese-Llama-2-13b-fast と比べて大小それぞれあるがおおむね同水準である．一方で，ELYZA-japanese-Llama-2-13b-fast は LLM の指示追従性能や対話応答性能を図るための評価データセット ELYZA Tasks 100 において高いスコアが報告されている[5]．本評価で用いたベンチマークツールは複数の評価データセットでの評価を一括して実行することができるが，LLM の評価としてはより多面的な評価を実施することが望ましい．本評価ではあくまで LLM の性能の一部を評価しているという点に留意が必要であり，さらなる多面的な評価が今後の課題である．

表 7 に lm-evaluation-harness を用いた英語ベンチマーク結果を示す．ベースの Llama 2 と比較すると，ELYZA-japanese-Llama-2-13b-fast は同程度だが，Swallow-13B や Ricoh-13B はスコアが低下した．これは日本語を中心とした継続事前学習を行ったことで英語の忘却が起きていると考えられる．日本語をより多く学習することで英語をより忘却することは自然であるため，学習データが比較的少ない ELYZA-japanese-Llama-2-13b-fast のスコアが高く，学習データの多い Swallow-13B や Ricoh-13B のスコアが低いのは自然な結果であると考えられる．多言語モデルとしての言語間の最適な学習データの比率や量は今後の課題である．

表 7 lm-evaluation-harness 英語ベンチマーク結果．スコアは%表記．各タスクのスコアは付録 A の表 9 と表 10 に添付する．

モデル	GLUE 8 タスク 平均スコア	Hugging Face リーダーボード 平均スコア
Llama 2 13B	63.0	55.4
Llama 2 13B Chat	64.1	**57.9**
Swallow-13B	60.1	52.3
ELYZA-japanese-Llama-2-13b-fast	**65.3**	55.7
Ricoh-13B (ours)	62.1	51.9

4 おわりに

本稿では，オープンな LLM である事前学習済み Llama 2 13B Chat に対して日英 2 言語データで語彙置換継続事前学習を行い，LLM ベンチマーク 2 種で性能評価した結果について報告した．ベンチマーク結果より，英語を主とする Llama 2 に対して日本語適応の語彙置換継続事前学習を行うことで，日本語性能を向上できることを確認した．

今後の展望としては，本稿で報告した Ricoh-13B に対して Instruction tuning や Alignment tuning を行い，様々なユースケースに応じた LLM 群の構築を目指す．また，より大規模な 70B パラメータモデルの語彙置換継続事前学習に取り組む予定である．

謝辞

本研究のモデル構築にあたり，アマゾン ウェブ サービス ジャパン合同会社の AWS LLM 開発支援プログラムによる支援を頂きました．感謝いたします．

参考文献

[1] 麻場直喜，梅沢知紀，川村晋太郎．日本語に特化した 60 億パラメータ規模の GPT モデルの構築と評価．言語処理学会第 29 回年次大会, 2023.

[2] Fred Philippy, Siwen Guo1 and Shohreh Haddadan. Towards a Common Understanding of Contributing Factors for Cross-Lingual Transfer in Multilingual Language Models: A Review. arXiv preprint arXiv: 2305.16768, 2023.

[3] Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá: Portuguese Large Language Models. arXiv preprint arXiv: 2304.07880, 2023.

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288, 2023.

[5] 130 億パラメータの「Llama 2」をベースとした日本語 LLM「ELYZA-japanese-Llama-2-13b」を公開しました（商用利用可）．https://note.com/elyza/n/n5d42686b60b7, 2023.

[6] Swallow: LLaMA-2 日本語継続事前学習モデル. https://zenn.dev/tokyotech_lm/articles/d6cb3a8fdfc907, 2023

[7] 野崎雄太，中島大，佐藤諒，伊藤真也，近藤宏，麻場直喜，川村晋太郎．大規模言語モデルに対する語彙置換継続事前学習の有効性の検証．言語処理学会第 30 回年次大会, 2024.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, 2017.

[9] 中島大，野崎雄太，佐藤諒，麻場直喜，川村晋太郎. BPE を用いたトークナイザーの性能に対する，言語・語彙数・データセットの影響．言語処理学会第 30 回年次大会, 2024.

[10] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226, 2018.

[11] Together Computer. RedPajama: An Open Source Recipe to Reproduce LLaMA training dataset. https://github.com/togethercomputer/RedPajama-Data, 2023.

[12] 佐藤諒，麻場直喜，野崎雄太，中島大，川村晋太郎．事前学習済み Llama2 モデルを活用した言語間転移日英モデルの作成．言語処理学会第 30 回年次大会, 2024.

A 付録

表 8 llm-jp-eval 日本語ベンチマークの JSTS スコア

モデル	JSTS	
	pearson	spearman
Llama 2 13B	**63.3**	**57.9**
Llama 2 13B Chat	54.5	50.6
PLaMo-13B	-5.8	-6.0
LLM-jp-13B	0.3	-3.4
Stockmark-13B	-	-
Swallow-13B	55.7	51.4
ELYZA-japanese-Llama-2-13b-fast	42.3	30.4
Ricoh-13B (ours)	52.7	51.1

表 9 lm-evaluation-harness による英語ベンチマークのタスク別スコア：GLUE

モデル	Avg.	CoLA	SST-2	MRPC	QQP	MNLI	QNLI	RTE	WNLI
Llama 2 13B	63.0	42.9	87.7	76.1/66.9	56.0/61.8	47.5	58.6	69.3	**67.6**
Llama 2 13B Chat	64.1	26.8	**94.5**	79.0/69.6	**53.8/64.2**	53.6	**66.7**	73.3	64.8
Swallow-13B	60.1	33.6	91.3	80.4/71.1	45.0/65.3	46.4	56.8	68.6	53.5
ELYZA-japanese-Llama-2-13b-fast	**65.3**	42.7	**94.5**	**82.0/71.1**	51.5/51.2	**56.7**	62.0	**76.5**	62.0
Ricoh-13B (ours)	62.1	**44.4**	88.1	81.2/71.6	52.1/59.1	47.4	55.1	70.8	59.2

表 10 lm-evaluation-harness による英語ベンチマークのタスク別スコア：Hugging Face Open LLM Leaderboard

モデル	Avg.	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
Llama 2 13B	55.4	59.5	82.2	**55.2**	36.9	**76.0**	22.8
Llama 2 13B Chat	**57.9**	**60.3**	**82.4**	53.5	**44.0**	74.4	**33.1**
Swallow-13B	52.3	53.2	76.9	50.1	40.4	72.6	20.8
ELYZA-japanese-Llama-2-13b-fast	55.7	57.3	80.1	53.1	40.4	75.4	27.1
Ricoh-13B (ours)	51.9	52.1	81.2	48.9	38.4	74.9	16.1