大規模言語モデルの日本語能力の効率的な強化: 継続事前学習における語彙拡張と対訳コーパスの活用

水木 栄 [1,2†] 飯田 大貴 [1,3†] 藤井 一喜 [1] 中村 泰士 [1] Mengsay Loem[1]
大井 聖也 [1] 服部 翔 [1] 平井 翔太 [1] 横田 理央 [1] 岡崎 直観 [1]
[1] 東京工業大学情報理工学院 [2] 株式会社ホットリンク [3] 株式会社レトリバ

概要

英語を主体として学習ずみの LLM を元に日本語テキストを主体として継続事前学習する方法は，高性能な日本語 LLM を構築する有望なアプローチである．本研究ではまず継続事前学習の効果を分析し，特に日本語の質問応答で効果的であることを報告する．また LLM の能力を効率的に強化する方法を明らかにするため，日本語の語彙拡張の影響および対訳コーパスの有効性を調査した．その結果，語彙拡張による効率化は要約を除き性能への悪影響はないこと，および対訳コーパスの併用が翻訳能力を強化することを明らかにした．

1 はじめに

大規模言語モデル（LLM）は幅広い用途を支える基盤としての役割が期待されており [1]，特にわが国の知識や日本語の理解に優れた日本語 LLM を構築する方法の確立は重要な研究課題である．英語圏ではオープンな LLM を構築する試みが活発であり，これらの優れた学習ずみ LLM を元に日本語テキストを主として学習を継続する，いわゆる継続事前学習は，少ない計算予算で高性能な日本語 LLM を構築する有望なアプローチである [2]（図 1）．継続事前学習による日本語 LLM は複数の開発例[1)]があるが，その能力を効率的に強化する知見は乏しい．そこで本研究では，Llama 2 [3] を元に継続事前学習した日本語 LLM である Swallow [2] の開発を基礎として，語彙拡張の影響および対訳コーパスの有効性を報告する．

語彙拡張は，トークナイザおよび LLM に語彙を追加する手法である．Llama 2 が採用するバイト対符号化（BPE [4]）に基づくトークナイザは英語を重視した設計のため，日本語テキストをバイト単位の多くのトークンで表現してしまう．このため日本語の単語や文字を追加する語彙拡張は，トークン列を短縮して日本語テキストの学習・生成効率を改善する効果がある．しかし，語彙拡張が性能に及ぼす影響は明らかではない．追加した語彙を上手に扱えなければ直感的には性能を損なう可能性がある一方で，ドメイン適応では性能が改善するとの報告がある [5, 6] ほか，同一の計算予算で学習可能なテキスト量の増加が有利に作用しうるためである．

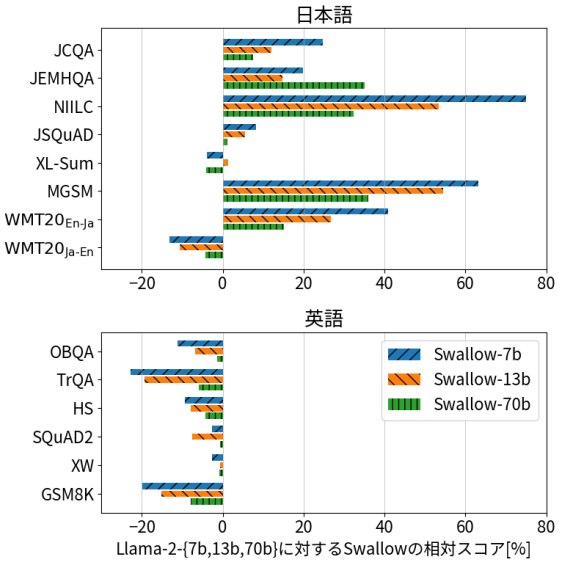

図 1 Llama 2 に対する Swallow の性能の変化．日本語タスク（上段．説明は表 1）のスコアが最大で約 70%向上．

対訳コーパスは，継続事前学習に日英対訳文を併用する手法である．多言語文埋め込みでは，対訳コーパスを併用した事前学習が言語間転移を促進すると報告されている [7, 8, 9]．また LLM では指示チューニングに翻訳タスクを用いておなじく言語間転移を促す試みもある [10, 11]．しかし，継続事前学習での有効性および効果的な用法は明らかでない．

本研究では，まず Swallow と Llama 2 の性能を比較して，継続事前学習による日本語および英語の知

† Equal contribution.
1) Stability AI Japan 社，ELYZA 社，rinna 社の事例がある．

識や能力の変化を考察する．次に，継続事前学習における語彙拡張の影響および対訳コーパスの有効性を調査する．語彙拡張については，拡張しない場合との性能を比較する．対訳コーパスについては，対訳文を学習する順序および学習形式を変えて効果的な学習方法を探る．本論文の貢献は以下のとおり．

- 継続事前学習は日本語能力の改善，特に知識を用いる質問応答に効果的であることを示した．
- 語彙拡張は大半のタスクで性能には影響せず，自動要約のみ性能が低下することを示した．
- 対訳コーパスの併用は翻訳能力を強化し，他のタスクには効果が波及しないことを示した．

2 関連研究

2.1 語彙拡張

文埋め込みモデルのドメイン適応では，対象ドメインの語彙の追加による性能の改善が知られている [5, 6]．これに対して LLM の継続事前学習における語彙拡張は対象言語での学習・生成効率の改善が主な動機であり，性能に関する知見は乏しい．日本語 LLM での語彙拡張の実施例[1)]は単一のモデルサイズおよび日本語能力の評価のみにとどまるほか，中国語 LLM [12] では語彙拡張を行った場合の性能のみが報告されている．多言語 LLM では各言語の語彙サイズと性能が相関するが，学習量とも相関するため語彙単独の影響は明らかでない [13]．

2.2 対訳コーパス

多言語文埋め込みモデルでは，対訳コーパスを用いた言語モデリングによる事前学習が隠れ状態ベクトルの言語個別性を緩和し，言語間転移を促進するとの報告がある [7, 8, 9]．また LLM への応用では対訳文を指示チューニングに用いることで多言語コーパスよりも効率的に翻訳能力を改善するとの報告がある [10, 11]．

3 実験設定

3.1 継続事前学習

本研究で用いる継続事前学習の方法は，特記なきかぎり Swallow [2] と同一である．学習元の LLM は Llama 2 base である．学習テキストは日本語と英語が 9 対 1 の多言語コーパスで，日本語は Swallow コーパス [29] および Wikipedia，英語は RefinedWeb [30] および The Pile [31] の arXiv サブセットである．

学習トークン数はすべての実験で 100B(illion) で統一する．すなわち，語彙拡張の実験ではそれぞれのトークナイザで得られる 100B 相当のテキストを学習する．なお語彙拡張をする場合の日本語テキスト量は，語彙拡張なしの約 1.8 倍となった．対訳コーパスの実験ではすべての対訳文を使用し，残りは多言語コーパスを足して 100B にする．

3.2 語彙拡張の影響

Swallow で採用した語彙拡張では，日本語語彙の構築，ベクトルの初期化，および文字列前処理の追加を実施している．詳細は付録 A.1 を参照．

日本語語彙の構築は，MeCab [32] と UniDic 辞書で単語分割した Swallow コーパスに対して BPE アルゴリズムによって語彙を構築（上限 16k）したうえで，ひらがな・カタカナ・漢字・長音記号で構成される 11,176 件のサブワードを Llama 2 トークナイザに追加した．追加後の語彙サイズは 43,176 となった．

追加したサブワードの埋め込み層と出力層のベクトルは，先行研究 [6] に倣い，Llama 2 トークナイザで分割したサブワード，すなわち Llama 2 が学習ずみのサブワードのベクトルの平均で初期化した．

文字列の前処理は，半角英数字記号の学習ずみ知識の活用を企図して，NFKC 正規化を追加した．

語彙拡張の影響を調べる実験（§ 4.2）では，語彙拡張をせずに継続事前学習した Swallow¬VE を Swallow と比較する．

3.3 対訳コーパスの有効性

本実験では，Swallow-7b¬VE をベースラインとして語彙拡張の影響を捨象したうえで，対訳コーパスを継続事前学習に併用する場合の性能を調べる．使用したコーパスは JParaCrawl 3.0 [33] であり，ウェブから抽出した約 2,200 万件の日英対訳文が含まれる．

対訳コーパスの用法は，学習の順序およびタスク形式が異なる 3 通り（表 3，詳細は付録 A.2）を試行する．学習の順序は，対訳コーパスを使い切ってから多言語コーパスに切り替える “先行” と，多言語コーパスと混合する “同時” の 2 種類である．タスク形式は，連結対訳文による次単語予測形式と，翻訳指示文とソース文の連結からターゲット文を予測する翻訳指示形式の 2 種類である．いずれもひとつ

2) https://github.com/Stability-AI/lm-evaluation-harness commit #9b42d41 を使用した．

表 1 日本語の評価．acc. は精度，EM は完全一致，JCQA は JCommonsenseQA [14]，JEMHQA は JEMHopQA [15] の略．

ベンチマーク	llm-jp-eval [16] (v1.0.0) 開発データ				JP LM Eval. Harness[2)]		LM Eval. Harness [17] (v0.3.0)	
評価タスク	択一質問応答	自由記述質問応答		機械読解	自動要約	算術推論	機械翻訳	
データセット	JCQA	JEMHQA	NIILC [18]	JSQuAD [14]	XL-Sum [19]	MGSM [20]	$\text{WMT20}_{\text{En-Ja}}$ [21]	$\text{WMT20}_{\text{Ja-En}}$ [21]
事例数	1,119	120	198	4,442	766	250	1,000	993
few-shot 数	4	4	4	4	1	4	4	4
評価指標	EM acc.	文字 F1	文字 F1	文字 F1	ROUGE-2	EM acc.	BLEU [22]	

表 2 英語の評価．OBQA は OpenBookQA [23]，TrQA は TriviaQA [24]，HS は HellaSwag [25]，XW は XWINO [26] の略．

ベンチマーク	Language Model Evaluation Harness [17] (v0.3.0)					
評価タスク	質問応答		機械読解	常識推論		算術推論
データセット	OBQA	TrQA	SQuAD2 [27]	HS	XW	GSM8K [28]
事例数	500	17,944	11,873	10,042	2,325	1,319
few-shot 数	8	8	8	8	8	8
評価指標	acc.	EM acc.	EM acc.	acc.	acc.	EM acc.

タスク形式	学習の順序	学習トークン数 [$\times 10^9$]
次単語予測	先行	5.6
次単語予測	同時	5.6
翻訳指示	先行	2.8

表 3 対訳コーパスの実験で試行した用法の一覧．

の対訳ペアから日 → 英と英 → 日の二方向を作る．

3.4 評価方法

日本語および英語の評価方法をそれぞれ表 1 および表 2 に示す．データセットは日本語 8 種類，英語 6 種類で，評価タスクは質問応答・読解・自動要約・推論・機械翻訳の few-shot 学習である．タスクの選定は LLM-jp の議論 [16] や Llama 2 論文の方法論 [3] を参考にしつつ，推論や文生成に関わるタスクを積極的に採用した．なお llm-jp-eval に含まれる自然言語推論タスクは，特に 7b・13b モデルでスコアが不安定であったため評価対象から除外した（付録 C）．

4 実験結果

4.1 継続事前学習の効果

Swallow および，その学習元である Llama 2 の評価を表 4 に，また Llama 2 に対する Swallow のスコアの増減率を図 1 に示す．Swallow の日本語タスク平均スコアは Llama 2 を約 7 ポイント上回る一方で，英語は 2–5 ポイント下回る．タスク別に見ると[3)]，日本語の質問応答（JCQA, JEMHQA, NIILC）は最大 75%，算術推論（MGSM）は 36–63%の顕著な改善が見られる．対照的に英語では質問応答（TrQA）および算術推論（GSM8K）で 6–23%の悪化が生じる．自動要約（XL-Sum）の増減は 5%未満である．機械翻訳は方向により対照的で，英日（En-Ja）は 15–41%改善，日英（Ja-En）は 4–13%悪化する．日本語の読解（JSQuAD）は Llama 2 のスコアが 0.8 超のため伸び代は小さく，10%未満の改善にとどまる．

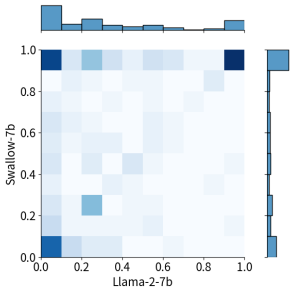

図 2 NIILC の各設問の採点（文字 F1，完全一致なら 1）に対する Llama 2 (X 軸) と Swallow (Y 軸) の同時分布．

継続事前学習がもたらす日本語の能力および知識の変化について考察する．算術推論（日：MGSM, 英：GSM8K）は，Llama 2 では英語優位（GSM8K>MGSM）であるところ，Swallow では MGSM が改善するも GSM8K が同水準まで悪化しており，英語での推論能力が日本語に転移したとは言いがたい．指示チューニングでの転移効果 [34] を踏まえると，インストラクションデータの併用が有効かもしれない．

知識については，質問応答の顕著な改善をふまえると，日本の知識の獲得が進んだことが示唆される．継続事前学習によって質問応答（NIILC）の各設問の採点に生じた変化を図 2 に示す．左上隅の色が濃いことから，誤答から正答に変化した設問が多く，その逆は少ないことがわかる．この傾向は，継続事前学習が，新たな知識を取り入れて間違いを修正するように機能したことを示唆している．

4.2 語彙拡張の影響

語彙拡張をしない Swallow¬VE に対する Swallow のスコアの増減率を図 3 に示す．日本語の能力に着目すると，総合的には語彙拡張による性能への影響は小さい．タスク別に見ると，質問応答は ±10%程度の増減が見られるも 7b と 70b で一致した優劣は認められない．よって，語彙拡張による学習テキスト量の増加（§ 3.1）は性能に表出していない．なお学

3) 各タスクのスコアの水準差が大きいため増減率で論じる．

モデル	日本語の評価									英語の評価						
	JCQA	JEMHQA	NIILC	JSQuAD	XL-Sum	MGSM	En-Ja	Ja-En	平均	OBQA	TrQA	HS	SQuAD2	XW	GSM8K	平均
Llama-2-7b	0.385	0.424	0.341	0.792	0.191	0.076	0.178	0.174	0.320	0.358	0.627	0.586	0.321	0.905	0.141	0.490
Swallow-7b	0.481	0.508	0.597	0.857	0.183	0.124	0.251	0.151	0.394	0.318	0.484	0.531	0.313	0.882	0.113	0.440
Llama-2-13b	0.700	0.442	0.417	0.853	0.214	0.132	0.215	0.198	0.396	0.376	0.726	0.615	0.368	0.914	0.240	0.540
Swallow-13b	0.784	0.506	0.640	0.901	0.217	0.204	0.272	0.177	0.463	0.350	0.585	0.566	0.341	0.908	0.204	0.492
Llama-2-70b	0.869	0.466	0.526	0.908	0.236	0.356	0.264	0.240	0.483	0.428	0.824	0.674	0.377	0.929	0.528	0.627
Swallow-70b	0.935	0.629	0.696	0.918	0.227	0.484	0.304	0.230	0.553	0.422	0.776	0.646	0.375	0.920	0.487	0.604

表 4　継続事前学習した Swallow と，学習元である Llama 2 の日本語および英語での評価．

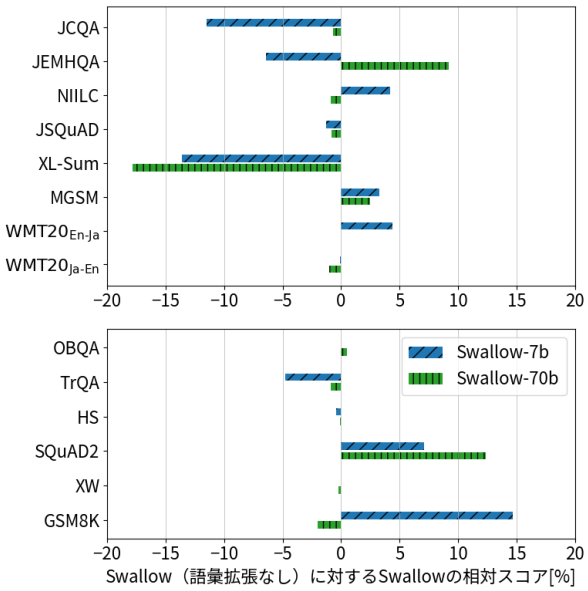

図 3　語彙拡張をしない Swallow¬VE に対する Swallow の性能の変化．

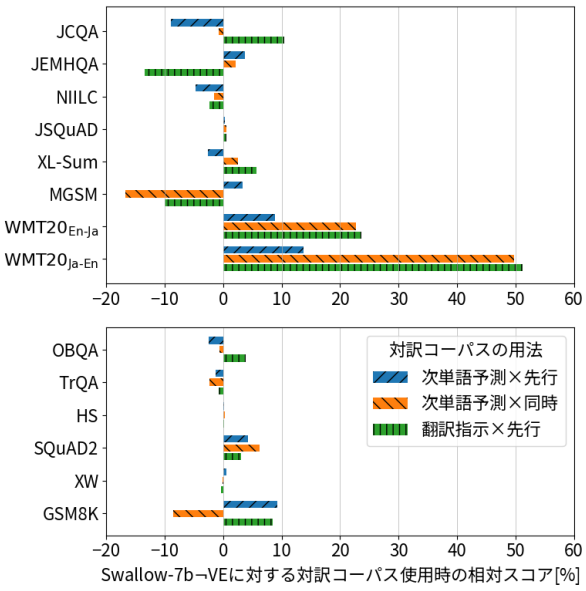

図 4　Swallow-7b¬VE（7b，語彙拡張なし）に対する，対訳コーパスの使用による性能の変化．

習曲線における収束特性にも差異は見られなかった（付録 B.2）．自動要約（XL-Sum）は 7b・70b ともに語彙拡張をすると約 15%悪化した[4]．長文を入力するタスクでは影響が顕在化しやすい可能性がある．

4.3　対訳コーパスの有効性

語彙拡張をしない Swallow-7b¬VE をベースラインとして，継続事前学習に対訳コーパスを併用した場合のスコアの増減率を図 4 に示す．翻訳能力は，En-Ja が 9–24%，Ja-En が 14–51%改善した．特に Ja-En の改善は対訳コーパスに特有の効果である．

コーパスの用法については，次単語予測形式で多言語コーパスと同時，または翻訳指示形式で多言語コーパスに先行する用法が有効であった．すなわち，対訳文を多言語コーパスに混ぜて継続事前学習を行うだけで，翻訳能力を効果的に改善できるとわかった．この知見は，LLM の翻訳能力は平文コーパスに散在する対訳文に由来するとの主張 [35] と整合的である．

翻訳以外のタスクのスコア増減は ±15%以内にとどまり，かつ一貫した優劣は見られない．したがって，対訳コーパスが言語間転移を促して翻訳以外の能力を改善する証拠は得られなかった．

5　結論と今後の展望

本研究では，日本語 LLM の能力を効率的に強化する方法を探るべく，継続事前学習の効果を分析するとともに，語彙拡張の影響および対訳コーパスの有効性を調査した．その結果，継続事前学習の効果は知識獲得による日本語の質問応答で顕著であること，語彙拡張による効率化は要約を除き性能への影響は小さいこと，対訳コーパスを混合するだけで翻訳能力（特に日英）を改善するも他タスクには効果が波及しないことを明らかにした．

英語能力の維持や推論能力の転移など，継続事前学習には多くの挑戦が残されている．多様な言語資源の活用，学習方法やアーキテクチャの工夫により，より優れた日本語 LLM の構築を目指したい．

4)　この傾向は指示チューニング実施後も不変だったため，指示追従能力の問題ではないと思われる．

謝辞

継続事前学習の実験は，国立研究開発法人産業技術総合研究所が構築・運用する AI 橋渡しクラウド（ABCI: AI Bridging Cloud Infrastructure）による「大規模言語モデル構築支援プログラム」の支援を受けました．学習した LLM の評価実験では、LLM-jp（LLM 勉強会）で開発されているデータや公開されている知見を活用しました．富士通株式会社の平岡達也氏には，語彙構築の設定について助言をいただきました．

参考文献

[1] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, ..., and et al. On the opportunities and risks of foundation models. arXiv:2108.07258, 2021.

[2] 藤井一喜, 中村泰士, Mengsay Loem, 飯田大貴, 大井聖也, 服部翔, 平井翔太, 水木栄, 横田理央, 岡崎直観. 継続事前学習による日本語に強い大規模言語モデルの構築. 言語処理学会第 30 回年次大会 (NLP2024), 2024.

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, ..., and Thomas Scialom. Llama 2: Open foundation and Fine-Tuned chat models. arXiv:2307.09288, 2023.

[4] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics**, pp. 1715–1725, 2016.

[5] Vin Sachidananda, Jason Kessler, and Yi-An Lai. Efficient domain adaptation of language models via adaptive tokenization. In **Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing**, pp. 155–165, 2021.

[6] Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In **Findings of the Association for Computational Linguistics**, pp. 460–470, 2021.

[7] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, ..., and Furu Wei. XLM-E: Cross-lingual language model pre-training via ELECTRA. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**, pp. 6170–6182, 2022.

[8] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In **Proceedings of the 37th International Conference on Machine Learning**, pp. 4411–4421, 2020.

[9] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics**, pp. 878–891, 2022.

[10] Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, ..., and Lei Li. Extrapolating large language models to Non-English by aligning languages. arXiv:2308.04948, 2023.

[11] Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. arXiv:2308.14186, 2023.

[12] Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. arXiv:2304.08177, 2023.

[13] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, ..., and Sunayana Sitaram. MEGA: Multilingual evaluation of generative AI. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 4232–4267, 2023.

[14] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, 2022.

[15] 石井愛, 井之上直也, 関根聡. 根拠を説明可能な質問応答システムのための日本語マルチホップ QA データセット構築. 言語処理学会第 29 回年次大会 (NLP2023), pp. 2088–2093, 2023.

[16] Namgi Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会 (NLP2024), 2024.

[17] Leo Gao, Jonathan Tow, Stella Biderman, Charles Lovering, Jason Phang, Anish Thite, ..., and silentv0x. EleutherAI/lm-evaluation-harness: v0.3.0, 2022.

[18] 関根聡. 百科事典を対象とした質問応答システムの開発. 言語処理学会第 9 回年次大会 (NLP2003), pp. 637–640, 2003.

[19] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, ..., and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In **Findings of the Association for Computational Linguistics**, pp. 4693–4703, 2021.

[20] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, ..., and Jason Wei. Language models are multilingual chain-of-thought reasoners. In **The Eleventh International Conference on Learning Representations**, 2023.

[21] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, ..., and Marcos Zampieri. Findings of the 2020 conference on machine translation (WMT20). In **Proceedings of the Fifth Conference on Machine Translation**, pp. 1–55, 2020.

[22] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, 2018.

[23] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2381–2391, 2018.

[24] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, pp. 1601–1611, 2017.

[25] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4791–4800, 2019.

[26] Alexey Tikhonov and Max Ryabinin. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In **Findings of the Association for Computational Linguistics**, pp. 3534–3546, 2021.

[27] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics**, pp. 784–789, 2018.

[28] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, ..., and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

[29] 岡崎直観, 服部翔, 平井翔太, 飯田大貴, 大井聖也, 藤井一喜, 中村泰士, Mengsay Loem, 横田理央, 水木栄. Swallow コーパス: 日本語大規模ウェブコーパス. 言語処理学会第 30 回年次大会 (NLP2024), 2024.

[30] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. arXiv:2306.01116, 2023.

[31] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, ..., and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. arXiv:2101.00027, 2020.

[32] 工藤拓. 形態素解析の理論と実装. 近代科学社, 2018.

[33] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6704–6710, 2022.

[34] Jiacheng Ye, Xijia Tao, and Lingpeng Kong. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. arXiv:2306.06688, 2023.

[35] Eleftheria Briakou, Colin Cherry, and George Foster. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 9432–9452, 2023.

[36] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, 2018.

[37] Tomoki Sugimoto, Yasumasa Onoe, and Hitomi Yanaka. Jamp: Controlled Japanese temporal inference dataset for evaluating generalization capacity of language models. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 57–68, 2023.

[38] Hitomi Yanaka and Koji Mineshima. Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In **Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP**, pp. 337–349, 2021.

[39] 川添愛, 田中リベカ, 峯島宏次, 戸次大介. 日本語意味論テストセットの構築. 言語処理学会第 21 回年次大会 (NLP2015), pp. 704–707, 2015.

[40] Hitomi Yanaka and Koji Mineshima. Compositional evaluation on Japanese textual entailment and similarity. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1266–1284, 2022.

A　実験設定の詳細

A.1　語彙拡張の方法

日本語語彙の構築は，単語分割した Swallow コーパスから乱択したサブセット（単語数 1.5B）を使用した．ただし記号は単独でサブワードにするために，記号を含む単語は記号の両端で分割した．BPE アルゴリズムの実装は，Llama 2 に倣って SentencePiece [36] を用いた．語彙サイズは予備実験にて 16k，32k，48k の 3 種類を試行したが日本語タスクの性能差がなかったため，最小である 16k に決定した．SentencePiece によって構築したサブワードの語彙に対しては，2 つの後処理を適用した．まず，SentencePiece が付与したスペースの特殊文字を削除した．これは，学習・推論時にトークン化する際には，語彙構築時とは異なり MeCab の単語分割を経由しないためである．次に，追加する語彙サイズが 8 の倍数になるように意図的に調節した．これはモデルの分散並列学習を容易にするための措置である．

追加したサブワードのスコアについては，SentencePiece が出力した BPE の結果を変更せず流用した．これは本来の語彙と追加した語彙のあいだでのマージルールの競合は極めて稀だと判断したためである．実際，Llama 2 トークナイザの本来の語彙には 2 文字以上の日本語のサブワードは存在しない．

A.2　対訳コーパスの用法

日英対訳文を，次単語予測形式（上）および翻訳指示形式（下）に変換するテンプレートを以下に示す．なお翻訳指示形式の場合は，翻訳先の文のみが学習（次単語予測）の対象である．

```
[和文]［英文］
[英文]［和文］
```

```
次の日本語を英語に翻訳してください。［和文］［英文］
Please translate the following English text into
Japanese.［英文］［和文］
```

対訳コーパスを使い切ってから多言語コーパスに切り替える “先行” を実験した意図は，対訳文によって英語主体から日本語主体の学習への切り替えが円滑化される可能性を想定したためである．なお翻訳指示形式かつ多言語コーパスと同時に使う用法は，LLM 学習ライブラリの制約により断念した．

B　語彙拡張の影響

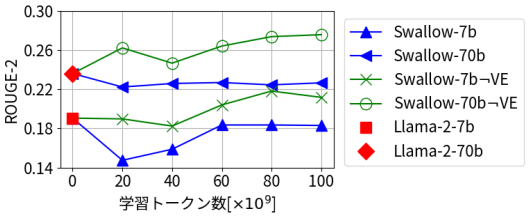

図 5　自動要約（XL-Sum）における Swallow¬VE および Swallow の学習曲線．

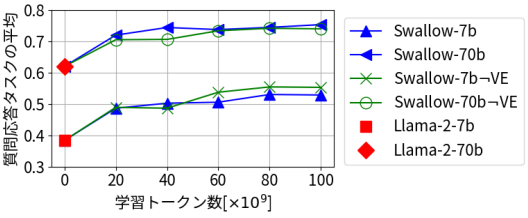

図 6　質問応答（JCQA, JEMHQA, NIILC）の平均スコアにおける Swallow¬VE および Swallow の学習曲線．

B.1　自動要約の学習曲線

語彙拡張の有無による自動要約（XL-Sum）の学習曲線の違いを図 5 に示す．語彙拡張をする Swallow は 20B トークン学習時点で Llama 2 より悪化するが，語彙拡張をしない Swallow¬VE は横ばいまたは改善する．したがって，語彙拡張時の要約能力の悪化は継続事前学習の開始時点で生じる模様である．

B.2　質問応答の学習曲線

語彙拡張の有無による質問応答の学習曲線の違いを図 6 に示す．学習曲線に顕著な差異はないことから，語彙拡張による日本語テキスト学習量の増加は，知識を用いるタスクの学習効率に特段影響しないことが示唆される．

C　自然言語推論タスク評価の課題

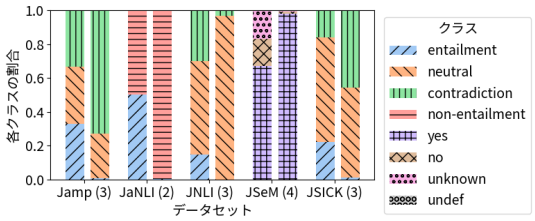

図 7　自然言語推論における各クラスの割合．各データセットの縦棒は左が正解，右が Swallow-7b による予測．括弧内の数字はクラス数．

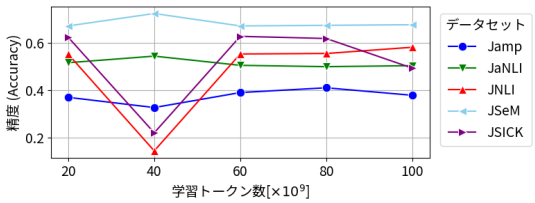

図 8　自然言語推論における Swallow-7b の学習曲線．

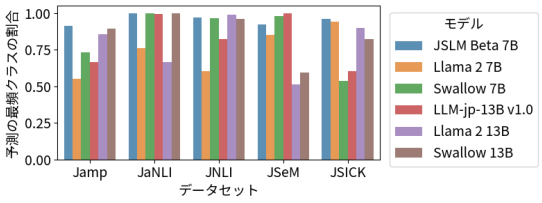

図 9　各種 LLM による予測の最頻クラスが占める割合．

llm-jp-eval に含まれる自然言語推論（Jamp [37], JaNLI [38], JNLI [14], JSeM [39], JSICK [40]）タスクは，クラスの不均衡に起因するスコアの乱高下を複数のモデルで確認した．まず，正解および Swallow-7b が予測したクラスの割合を図 7 に示す．正解・予測ともに偏りが大きく，特に予測は 3 つのデータセットで最頻クラスが全予測の 95%以上を占めている．このため，予測と正解の最頻クラスが偶然に一致するか否かでスコアに大差が生じる性質がある．次に Swallow-7b の学習曲線を図 8 に示す．2 つのデータセットで約 40 ポイントの乱高下が生じているが，これは予測の最頻クラスの遷移が原因であった．最後に，予測の不均衡は Swallow に特有ではなく他の 7b・13b モデルでも見られた（図 9）．以上をふまえて，自然言語推論の能力をスコアのみで論じることは困難と判断してベンチマークから除外した．本タスクの評価においては，正解クラスの不均衡を解消するなどの工夫が望ましいと考える．