本日の資料は、2023/09/04にLLM勉強会(LLM-jp)で
発表した資料の若干更新版です.
https://llm-jp.nii.ac.jp/resources/

日英2ヶ国語対応の大規模言語モデル "Weblab-10B"の構築

東京大学大学院工学系研究科
技術経営戦略学専攻
松尾研究室 特任研究員 小島 武

自己紹介

- 略歴
  - 2023.3 東京大学大学院 工学系研究科 技術経営戦略学専攻 博士課程修了
  - 2023.4～ 東京大学大学院 工学系研究科 技術経営戦略学専攻 特任研究員
  - （以前はシステムエンジニアとして働いていました）
- 研究分野、研究テーマ
  - 深層学習、大規模言語モデル
  - 基盤モデルの効率的な知識転移に関する研究
    - 画像の基盤モデル（ViT）のテスト時適応の改善
    - LLMの思考の連鎖（CoT）による推論能力の改善
  - 最近：LLMの構築に興味

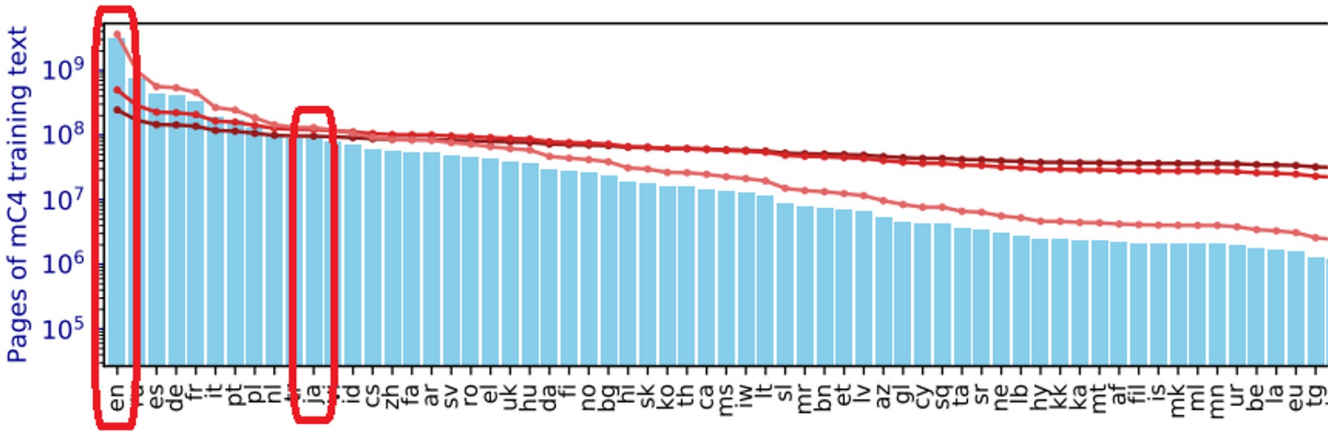
目次

- 背景
- 問題意識
- 実験
  - 予備実験（1K～100M）
  - 本実験（10B）
- Weblab-10Bの公開
- まとめと今後

背景

- 大規模言語モデル
  - 事前学習で大量のテキストデータを学習する.
    - 汎用性と高性能の源泉
    - インターネットから収集した大量のテキストデータを使う.
    - そのテキストデータの多くは一部の主要言語（例えば英語）で構成されており、それ以外の言語（例えば日本語）のテキストデータを大量収集することは現状では限界がある。

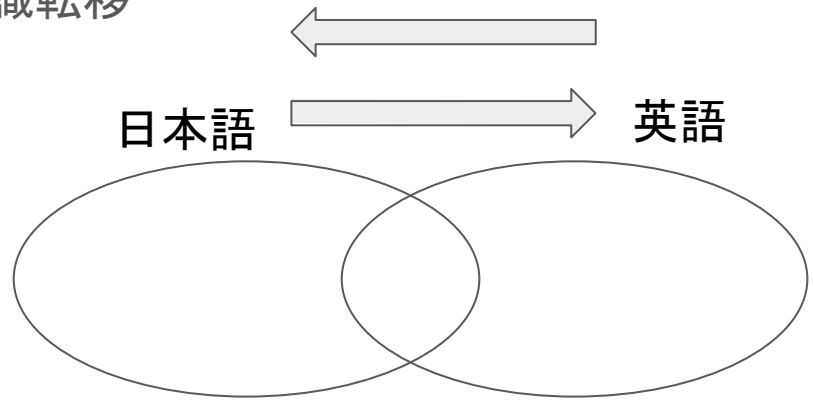
(mC4)
英語データは
日本語データ
の10倍以上.



https://arxiv.org/abs/2010.11934

問題意識

- 日本語を主体とした大規模言語モデルを構築する際に、日本語と英語の混合データで学習を行うことで、単体言語データでの学習よりも高いパフォーマンスを出すことができるか？
  - 考えられる根拠
    - 言語間の知識転移



実験

- 予備実験
  - モデルサイズ 1K～100M
  - **事前学習**における言語間の知識転移を検証
- 本実験
  - モデルサイズ 10B
  - 事前学習を行った後、**事後学習（ファインチューニング）**における言語間の知識転移を検証

予備実験（1K〜100M）

- シナリオ
  - **英語データでの学習**
    - THE PILE　* 利用実績：GPT-J, GPT-NeoX, Pythia
      - **約332B**トークン
      - マルチ言語だが、概ね英語（比率は非公開）
  - **日本語データでの学習**
    - Japanese-mC4　* 利用実績：OpenCALM(CA), Rinna GPT, Ricoh GPT
      - **約314B**トークン
      - 日本語（ただしサンプルを見ると英語もまま含まれている）
  - **混合データ（英語＋日本語）での学習**
    - THE PILEとJapanese-mC4を混合. トークン比で約１：１の混合比率.

予備実験（1K〜100M）

- 評価指標
  - 英語のValidation Loss
    - THE PILE
  - 日本語のValidation Loss
    - Japanese-mC4

予備実験（1K〜100M）

- ライブラリ：GPT-NeoX
  - モデル構造とトークナイザーはデフォルトのまま
- モデルサイズ
  - パラメータ数 1K ～ 100Mのオーダーの範囲で検証.
- 学習ステップ： 1 epoch * LLMの事前学習において一般的な設定.
  - 英語データでの学習：100,000 iteration
  - 日本語データでの学習：100,000 iteration
  - 混合データでの学習：200,000 iteration
- Misc.
  - Batch size: 1536, sequence len: 2048
  - Optimizer: Adam(lr=0.97e-4, min_lr=0.97e-5, warmup_with_cosine_decay)

予備実験（1K～100M）

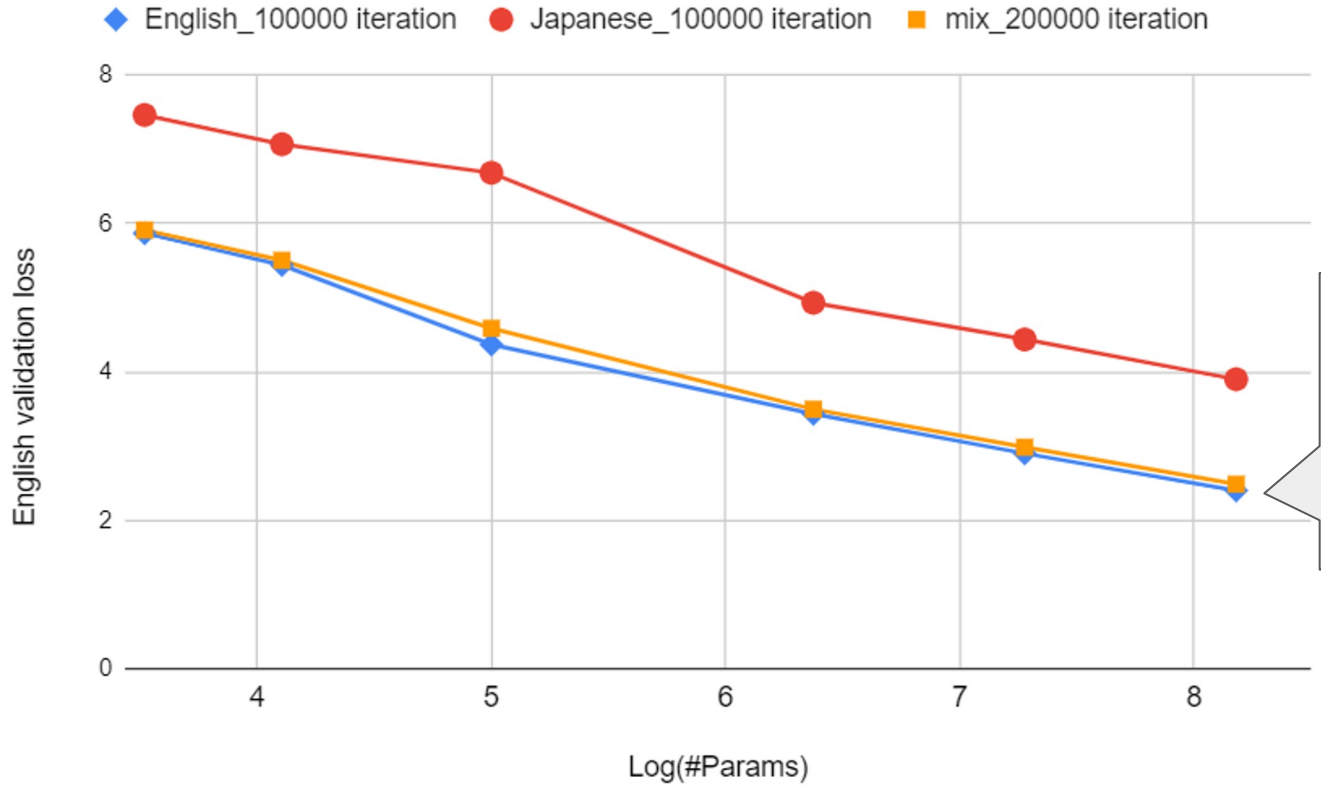
- 計算環境：東京大学情報基盤センター Wisteria

Model	# Params (w/o embedding)	# Params (w embedding)	# GPU (# node)	所要日数 (混合データ学習の場合)
E+3 (1K)	3,312	1,613,040	8 (1)	2.2
E+4 (10K)	12,768	3,232,224	8 (1)	2.2
E+5 (100K)	100,096	6,539,008	8 (1)	2.3
E+6 (1M)	2,369,792	28,125,440	8 (1)	2.8
E+7 (10M)	18,915,328	70,426,624	8 (1)	4.9
E+8 (100M)	302,311,424	405,334,016	32 (4)	4.1

* A100(40GB) GPU

予備実験（1K～100M）

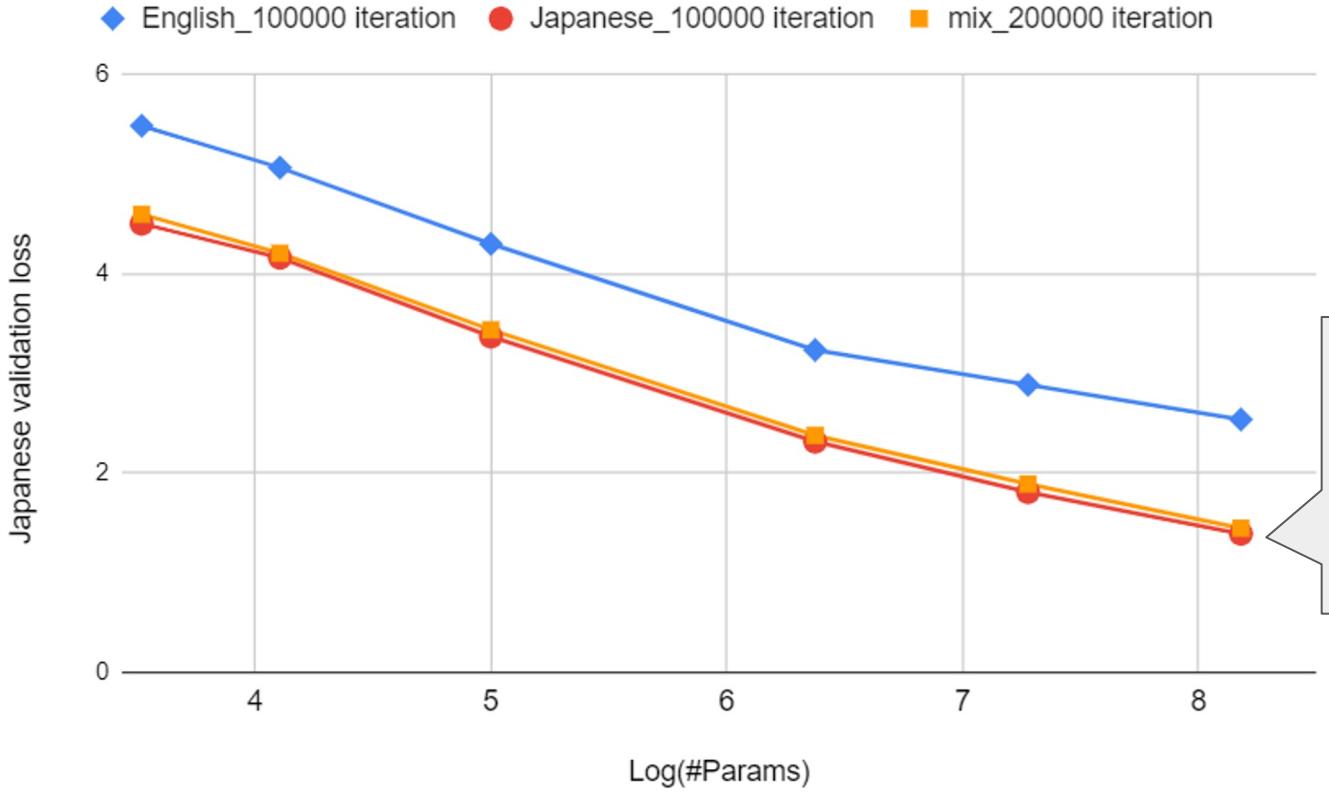
❑ 結果：英語のValidataion Loss



英語データでの学習と混合データでの学習でほぼ同じパフォーマンスを達成.

予備実験（1K〜100M）

❑ 結果：日本語のValidation Loss



日本語データでの学習と混合データでの学習でほぼ同じパフォーマンスを達成.

予備実験（1K〜100M）

- 考察
  - 日英混合データでの事前学習により、単一言語で事前学習させた場合とほぼ同じ性能（Validation Loss）を両言語において同時に達成することを確認できた. **ただし優位性までは確認できなかった.**

本実験（10B）

- 事前学習
  - ライブラリ：GPT-NeoX
    - モデル構造とトークナイザーはデフォルトのまま
  - 学習データ：日英混合データ
    - THE PILE（**約332B**トークン）とJa-mC4（**約314B**トークン）を混合.
    - トークン比で約１：１の混合比率.
- 事後学習
  - ライブラリ：Stanford Alpaca
  - 学習データ：以下のファインチューニング用データからサンプリング
    - Alpaca（英語＆日本語訳）
    - Flan 2021（英語）
    - Flan CoT（英語）
    - Flan Dialog（英語）

https://github.com/google-research/FLAN/tree/main/flan/v2#download

本実験（10B）

- モデル
  - 構造：GPT-NeoX
  - 活性化関数：gelu
  - コンテクスト長：2048
  - アテンションヘッド数：38
  - 隠れ層の次元：4864
  - 隠れ層の段数（ブロック数）：36
  - Vocabulary数：50277

  - 総パラメータ数：10,712,113,664（10,222,756,352 w/o embedding）

本実験（10B）

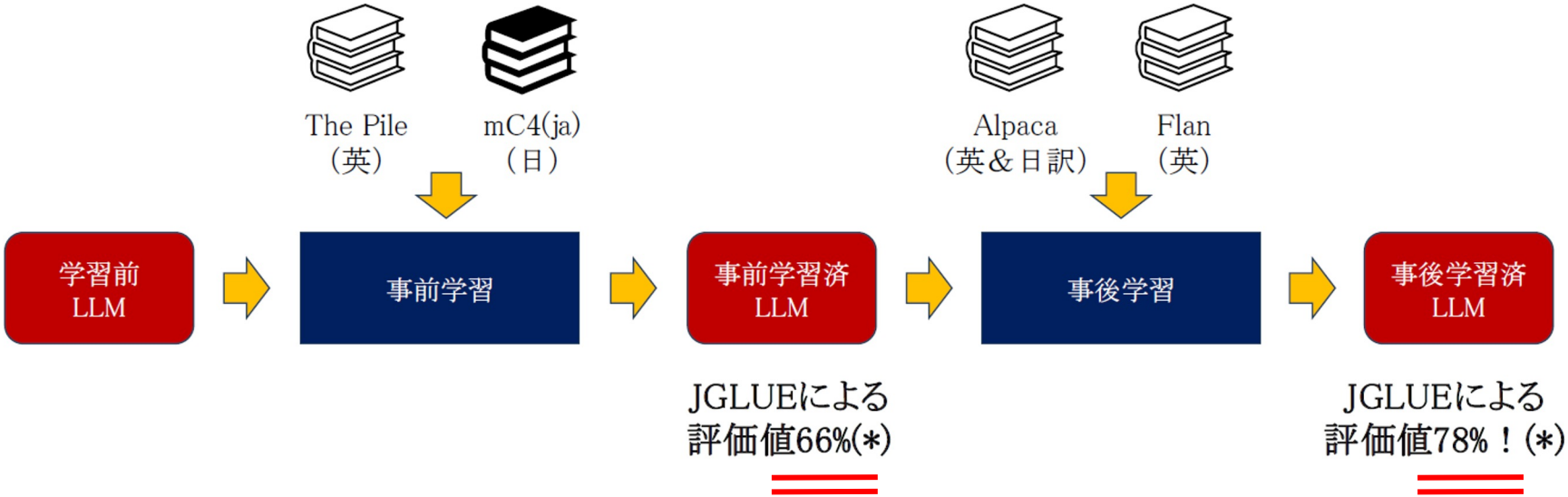
- 評価：JGLUE（4タスク） ＊日本語の評価タスク
  - JCommonsenseQA：知識問題（5択）
  - JNLI-1.1：NLI（3択）
  - MARC-ja-1.1：感情判定（2択）
  - JSQuAD-1.1：文章読解（単語抽出）

	英	日
Alpaca （一問一答のチャット）	✅	✅
FLAN2021 （分類系タスク多い）	✅	✗
FLAN_CoT （思考の連鎖タスク）	✅	✗
FLAN_Dialog （マルチターンのチャット）	✅	✗

JGLUE（4タスク）に類似するデータセットはFLAN2021だが、日本語訳は含まれず英語原文のみ学習に利用した。

本実験（10B）

- 実験結果：JGLUE４タスク



本実験（10B）

- 実験結果：JGLUE４タスク

model	average	jcommonsenseqa	jnli	marc_ja	jsquad
weblab-10b-instruction-sft	**78.78**	74.35	65.65	96.06	79.04
stabilityai-japanese-stablelm-instruct-alpha-7b	70.10	82.22	52.05	82.88	63.26
weblab-10b	66.38	65.86	54.19	84.49	60.98
stabilityai-japanese-stablelm-base-alpha-7b	61.03	33.42	43.34	96.73	70.62
rinna-bilingual-gpt-neox-4b-instruction-sft	61.97	49.51	47.08	95.28	55.99
rinna-bilingual-gpt-neox-4b-instruction-ppo	61.82	48.79	48.23	96.09	54.16
llama2-13b-chat	58.95	72.56	35.62	59.92	67.69
llama2-13b	52.98	74.89	21.98	38.89	76.14
rinna-japanese-gpt-neox-3.6b-instruction-ppo	59.87	44.06	54.19	89.61	51.62
rinna-japanese-gpt-neox-3.6b-instruction-sft-v2	57.20	40.57	53.45	89.88	44.91
rinna-japanese-gpt-neox-3.6b-instruction-sft	55.17	38.07	44.58	90.62	47.41
llama2-7b	56.33	52.64	28.23	86.05	58.4
rinna-japanese-gpt-neox-3.6b	47.20	31.64	34.43	74.82	47.91
llama2-7b-chat	58.72	55.59	29.54	90.41	59.34
rinna-bilingual-gpt-neox-4b	46.60	20.82	55.22	59.55	50.79
cyberagent-open-calm-7b	45.44	24.22	37.63	74.12	45.79
cyberagent-open-calm-3b	48.70	27.79	40.35	86.21	40.45
rinna-japanese-gpt-1b	46.62	34.76	37.67	87.86	26.18

本実験（10B）

Alpacaデータ(日英)を削って、英語データだけで事後学習を行っても、JGLUE評価値が改善した。

❑ 実験結果：JGLUE４タスク

model	average	jcommonsenseqa	jnli	marc_ja	jsquad
weblab-10b-instruction-sft	**78.78**	74.35	65.65	96.06	79.04
weblab-10b-instruction-sft-without-alpaca	75.42	77.30	59	95.28	70.1
stabilityai-japanese-stablelm-instruct-alpha-7b	70.10	82.22	52.05	82.88	63.26
weblab-10b	66.38	65.86	54.19	84.49	60.98
stabilityai-japanese-stablelm-base-alpha-7b	61.03	33.42	43.34	96.73	70.62
rinna-bilingual-gpt-neox-4b-instruction-sft	61.97	49.51	47.08	95.28	55.99
rinna-bilingual-gpt-neox-4b-instruction-ppo	61.82	48.79	48.23	96.09	54.16
llama2-13b-chat	58.95	72.56	35.62	59.92	67.69
llama2-13b	52.98	74.89	21.98	38.89	76.14
rinna-japanese-gpt-neox-3.6b-instruction-ppo	59.87	44.06	54.19	89.61	51.62
rinna-japanese-gpt-neox-3.6b-instruction-sft-v2	57.20	40.57	53.45	89.88	44.91
rinna-japanese-gpt-neox-3.6b-instruction-sft	55.17	38.07	44.58	90.62	47.41
llama2-7b	56.33	52.64	28.23	86.05	58.4
rinna-japanese-gpt-neox-3.6b	47.20	31.64	34.43	74.82	47.91
llama2-7b-chat	58.72	55.59	29.54	90.41	59.34
rinna-bilingual-gpt-neox-4b	46.60	20.82	55.22	59.55	50.79
cyberagent-open-calm-7b	45.44	24.22	37.63	74.12	45.79
cyberagent-open-calm-3b	48.70	27.79	40.35	86.21	40.45
rinna-japanese-gpt-1b	46.62	34.76	37.67	87.86	26.18

本実験（10B）

❑ 実験結果：JGLUE８タスクでの追加実験

model	average	jcommonsenseqa	jnli	marc_ja	jsquad	jaqket_v2	xlsum_ja	xwinograd_ja	mgsm
weblab-10b-instruction-sft	**59.11**	74.62	66.56	95.49	78.34	63.32	20.57	71.95	2
stabilityai-japanese-stablelm-instruct-alpha-7b	54.71	82.22	52.05	82.88	63.26	74.83	7.79	72.68	2
stabilityai-japanese-stablelm-base-alpha-7b	51.05	33.42	43.34	96.73	70.62	78.09	10.65	72.78	2.8
weblab-10b	50.74	66.58	53.74	82.07	62.94	56.19	10.03	71.95	2.4
rinna-bilingual-gpt-neox-4b-instruction-sft	47.75	49.51	47.08	95.28	55.99	61.17	5.51	64.65	2.8
rinna-bilingual-gpt-neox-4b-instruction-ppo	47.18	48.79	48.23	96.09	54.16	57.65	5.03	65.07	2.4
llama2-13b-chat	47.02	72.56	35.62	59.92	67.69	48.2	15.14	63.82	13.2
llama2-13b	46.32	74.89	21.98	38.89	76.14	67.7	18.11	62.88	10
rinna-japanese-gpt-neox-3.6b-instruction-ppo	46.32	44.06	54.19	89.61	51.62	50.95	6.63	69.13	4.4
rinna-japanese-gpt-neox-3.6b-instruction-sft-v2	45.23	40.57	53.45	89.88	44.91	52.84	6.14	71.22	2.8
rinna-japanese-gpt-neox-3.6b-instruction-sft	43.82	38.07	44.58	90.62	47.41	53.69	4.74	69.45	2
llama2-7b	42.97	52.64	28.23	86.05	58.4	38.83	9.32	64.65	5.6
rinna-japanese-gpt-neox-3.6b	41.79	31.64	34.43	74.82	47.91	68.38	5.16	70.8	1.2
llama2-7b-chat	41.31	55.59	29.54	90.41	59.34	17.96	2.34	66.11	9.2
rinna-bilingual-gpt-neox-4b	40.03	20.82	55.22	59.55	50.79	59.45	5.55	66.42	2.4
cyberagent-open-calm-7b	38.80	24.22	37.63	74.12	45.79	60.74	2.04	65.07	0.8
cyberagent-open-calm-3b	38.61	27.79	40.35	86.21	40.45	46.91	1.95	63.61	1.6
rinna-japanese-gpt-1b	36.92	34.76	37.67	87.86	26.18	37.03	5.34	64.55	2

special_tokens_map.json is modified to avoid errors during the evaluation of the second half benchmarks.
As a result, the results of the first half benchmarks became slightly different.
https://huggingface.co/matsuo-lab/weblab-10b-instruction-sft

本実験（10B）

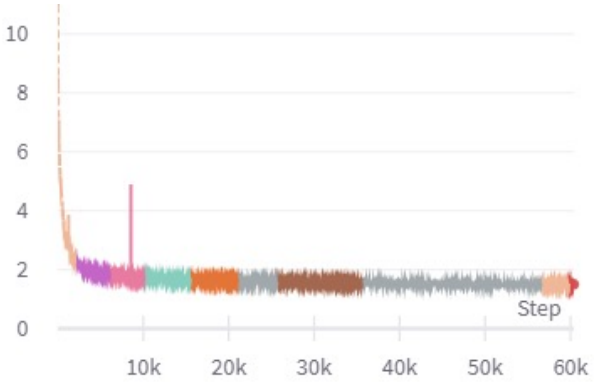
- 計算環境：ABCI
  - 事前学習
    - A100(40GB) GPU 256基 [32ノード] × 約20日間
    - 学習期間：2023/6/27 - 2023/7/22
      - トラブルシューティングやサーバ混雑のため、
        - 上記期間のうち６日間は学習ジョブ実行できず
        - チェックポイントからの学習再開回数：27回
      - 幸い学習のやり直しは発生せず.
    - 教訓：Baby Sitting部隊必要

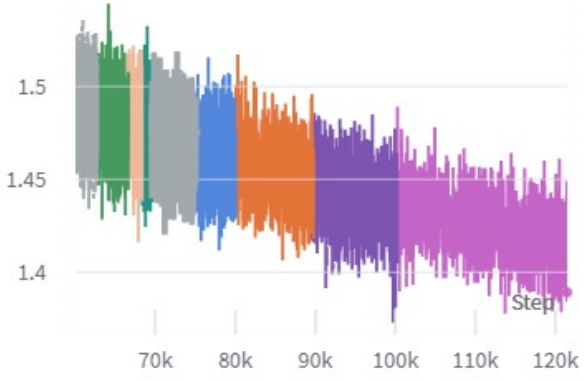
Training run babysitting[11]
Suchir Balaji, Mo Bavarian, Greg Brockman, Trevor Cai, Chris Hesse, Shantanu Jain, Roger Jiang, Yongjik Kim, Kyle Kosic, Mateusz Litwin, Jakub Pachocki, Alex Paino, Mikhail Pavlov, Michael Petrov, Nick Ryder, Szymon Sidor, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Chelsea Voss, Ben Wang, Tao Xu, Qiming Yuan

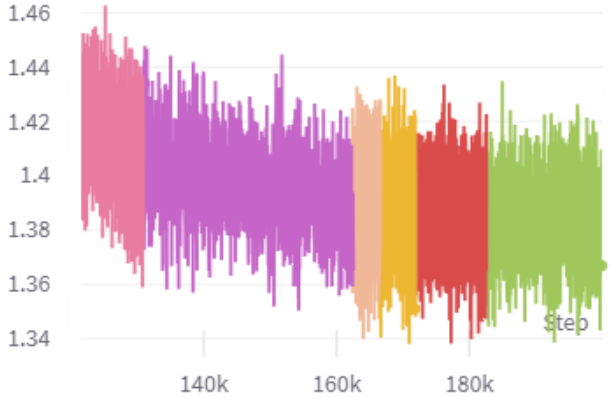
https://arxiv.org/abs/2303.08774

本実験（10B）

- 事前学習の学習曲線
  - チェックポイントからの再開が多すぎるため, 全曲線を一つの画面で表示させることができず、表示を三分割しています（そのため縦軸のスケールがバラバラ）.







Weblab-10Bモデルの公開

- 本実験で開発した10BモデルをHuggingFaceにて公開.
- 非商用ライセンス：cc-by-nc-4.0

事前学習済みモデル
weblab-10b

事後学習済みモデル
weblab-10b-instruction-sft

まとめと今後

- 日本語とマルチ言語（英語）の混合データセットで大規模言語モデルの学習を行うことにより、日本語の性能を高められるかどうか実験を行った.
- 予備実験（1K～100M）にて、混合データでの事前学習により、単一言語で事前学習させた場合とほぼ同じ性能（Validation Loss）を、両言語において同時に達成することを確認した. ただし優位性までは確認できなかった.
- 本実験（10B）にて、混合データでの事前学習を行った後、事後学習で英語データによるファインチューニングを行い、JGLUE評価値が大幅に向上したことを確認した. 言語間の知識転移を確認できた.
- 10Bのモデルを非商用ライセンス（cc-by-nc-4.0）でHuggingFaceにて公開.