日本語とマルチ言語の混合データセットにおける大規模言語モデルの構築

東京大学大学院工学系研究科
技術経営戦略学専攻
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
- 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

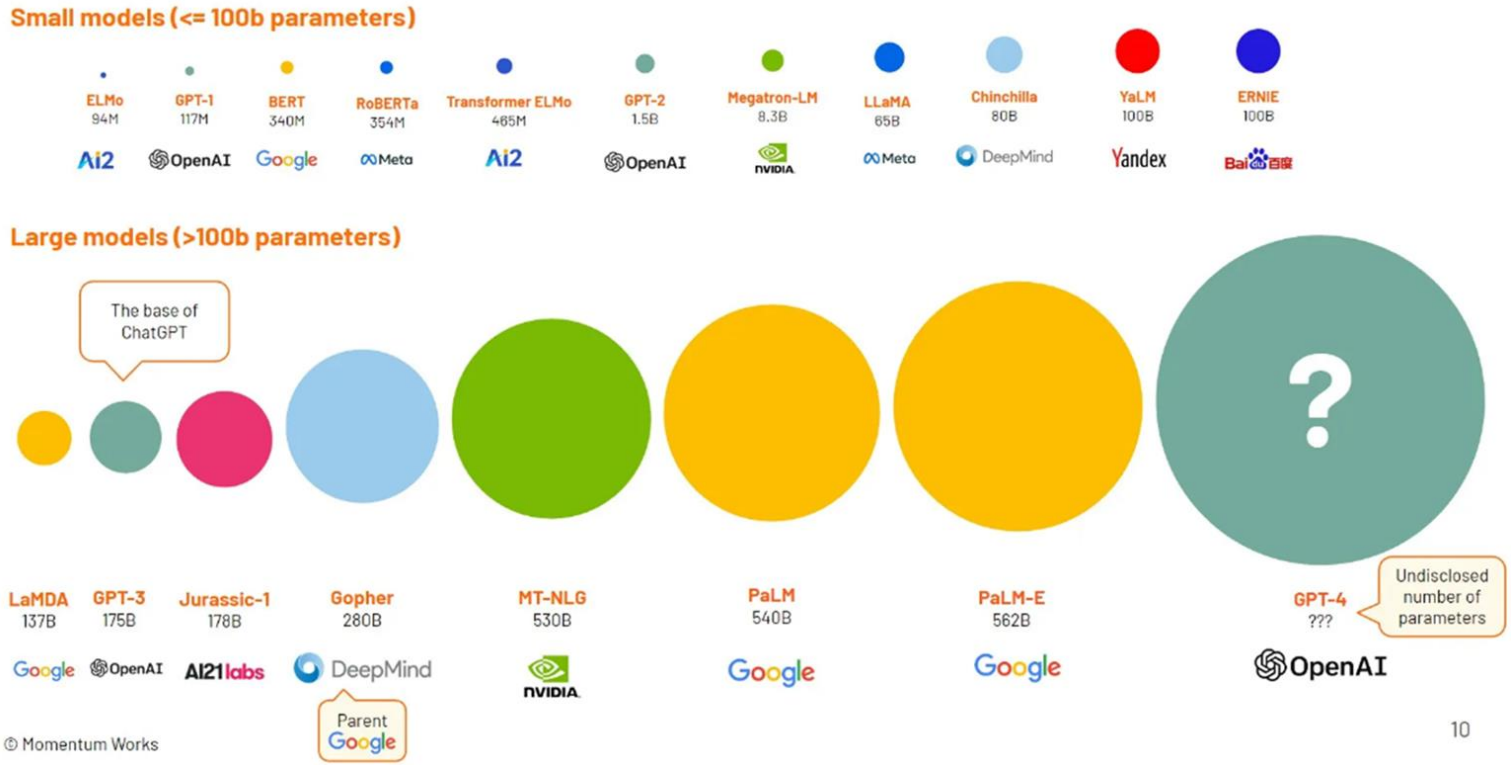
発表の概要

- 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。
- 実験の結果：
  - 日本語あるいは英語の単体データで学習を行った場合と同程度の性能を達成することが確認された（優位性までは確認されなかった）。

    ⇒ 単一言語で学習させた場合とほぼ同じ性能を達成し、かつ他の言語も同時に習得できることが確認できた。

背景

- 大規模言語モデル

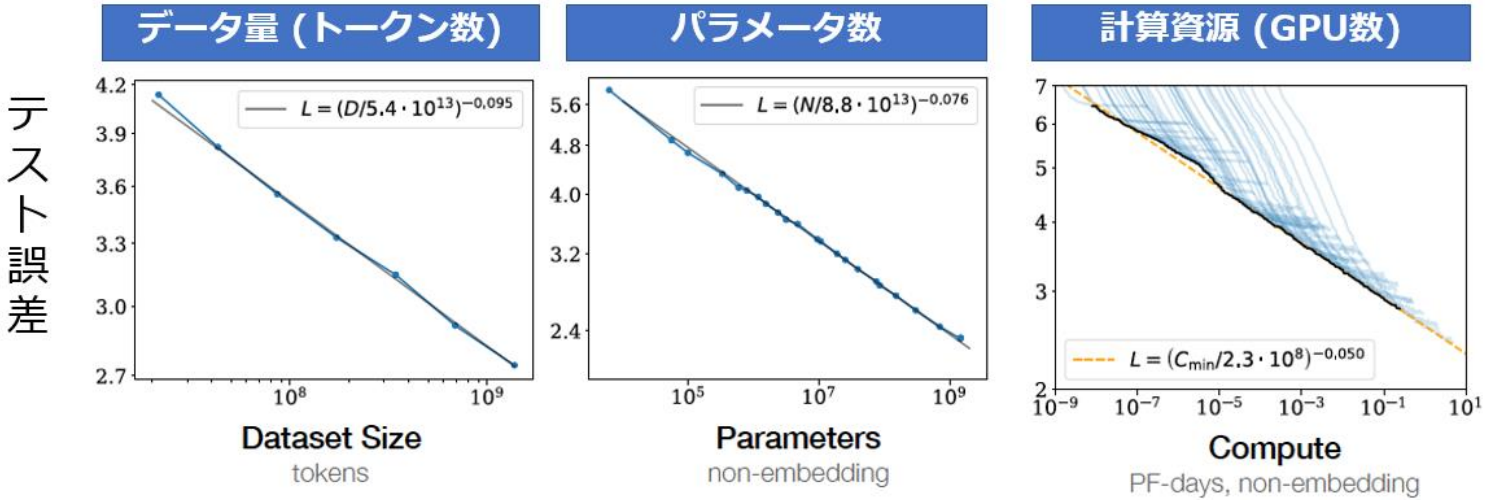


https://thelowdown.momentum.asia/the-emergence-of-large-language-models-llms/

背景

- 大規模言語モデル
  - べき乗則（スケール則）

大規模言語モデルの性能（テスト誤差，Test Loss）は，1. データ量，
2. パラメータ数，3. 計算資源 (GPU数）を増やすほどに改善

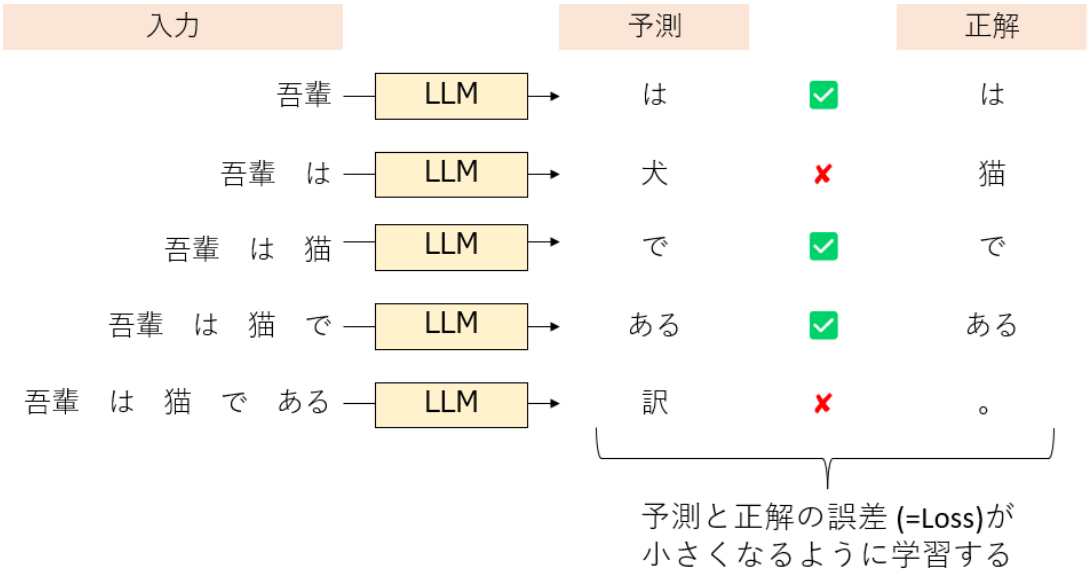


“Scaling Laws for Neural Language Models” [Kaplan+ 2020]

背景

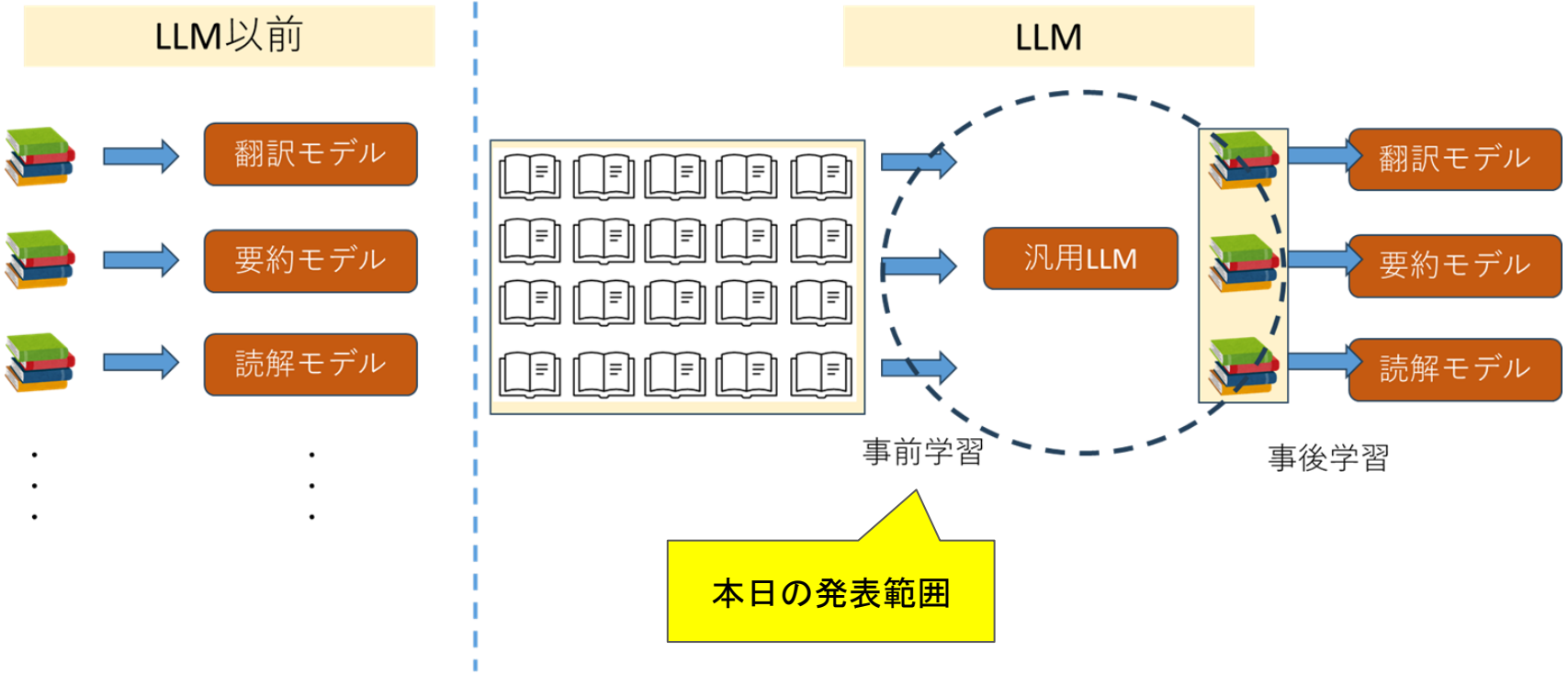
- 大規模言語モデル
  - Next Token Prediction (次のトークンをひたすら予測)による学習　＊自己教師あり学習の一種

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1})$$



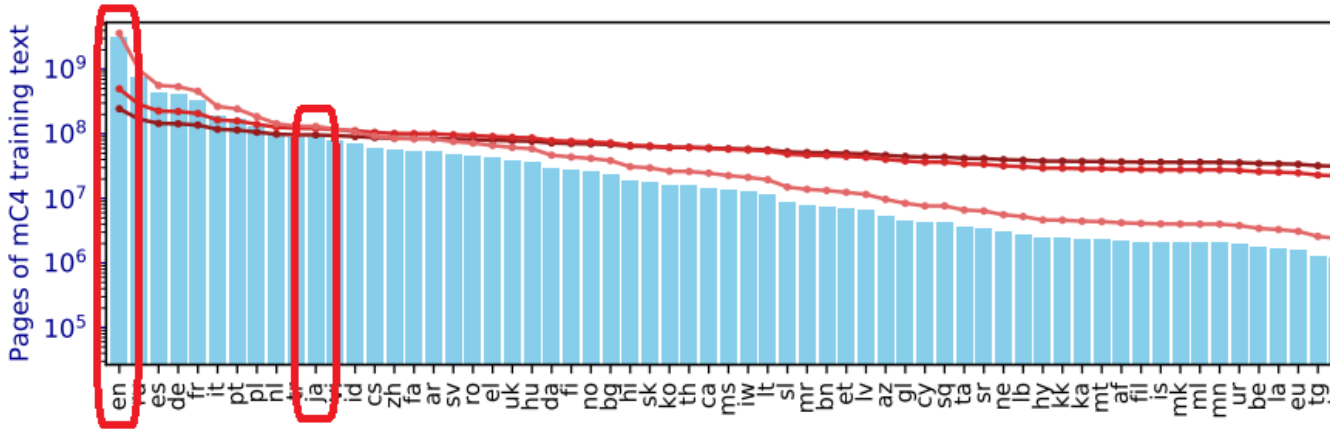
背景

- 大規模言語モデル
  - 事前学習→事後学習の二段階の学習



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し、汎用性を高める.
    - 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。



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、日本語とマルチ言語（英語主体）の混合データで事前学習を行うことで、単体言語データでの事前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
    - データの水増し

実験

- 大規模言語モデルにおいて、日本語とマルチ言語（英語主体）の混合データで事前学習を行い、単体言語データでの事前学習とのパフォーマンスを比較する.

実験

- 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.

実験

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

実験

- ライブラリ：GPT-NeoX
  - モデル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)

実験

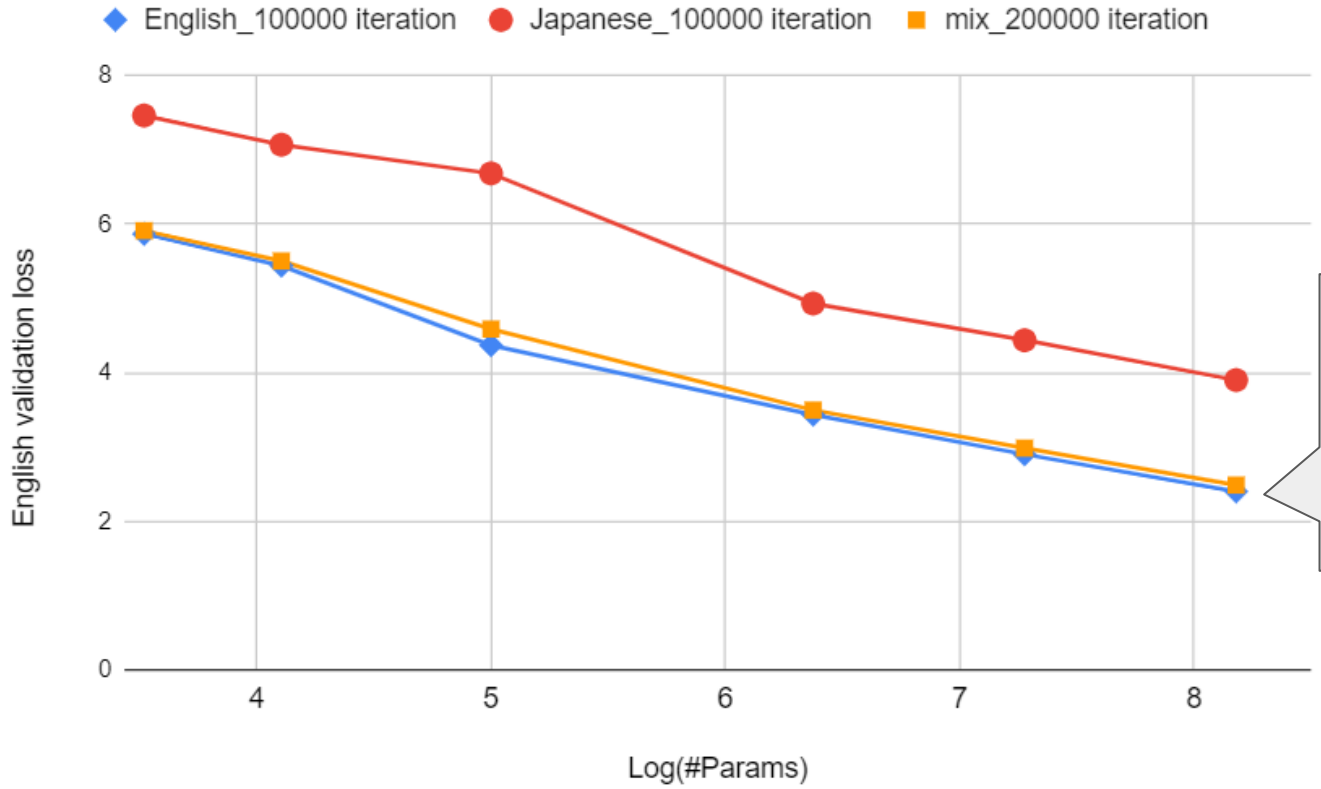
- 計算環境：東京大学情報基盤センター Wisteria

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

実験

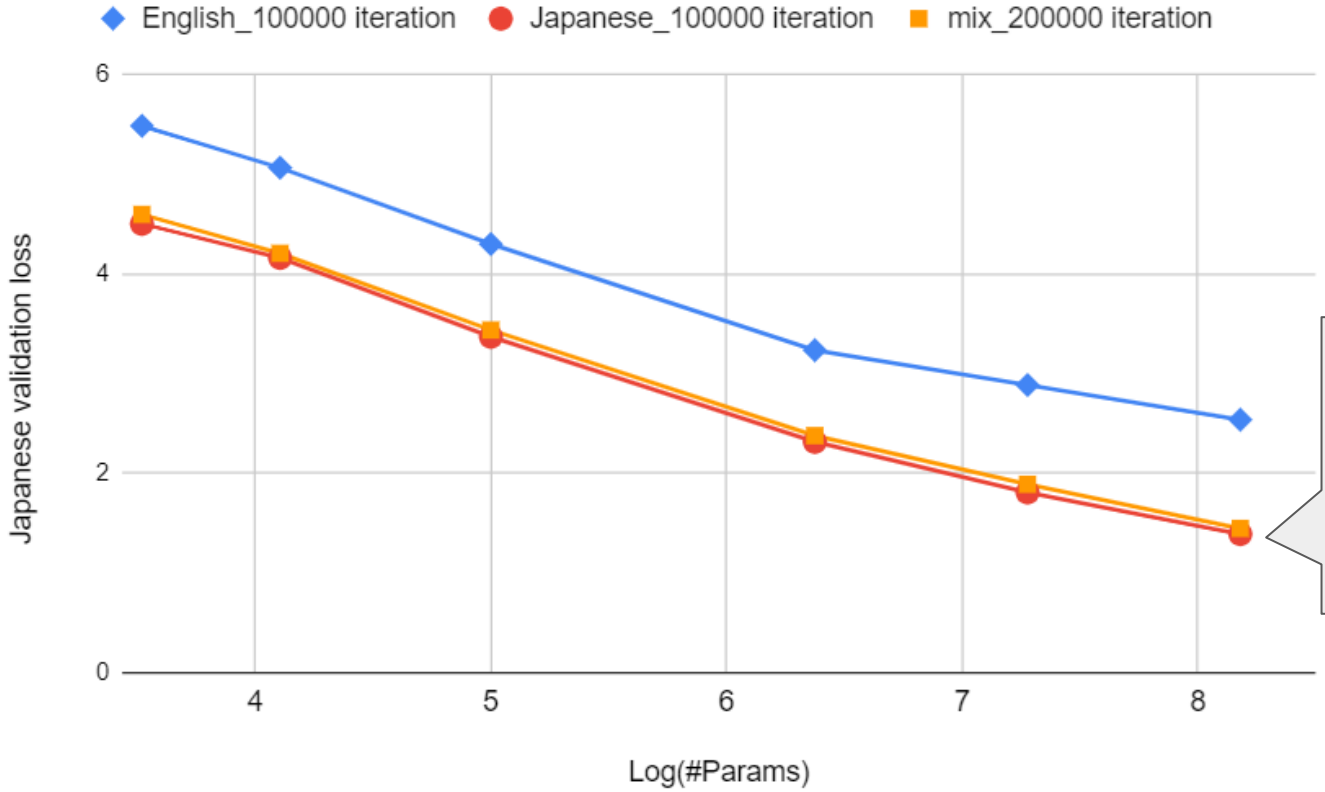
- 結果：英語のValidataion Loss



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.

実験

- 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.

実験

- 結果考察
  - 混合データでの事前学習が、英語もしくは日本語単体データでの事前学習に比べてValidation Lossによる評価軸での明確な優位性を確認することはできなかった.
  - 単一言語で学習させた場合とほぼ同じ精度を達成し、かつ他の言語も同時に習得できることが確認できた.
  - 混合データで事前学習を行うことにより、モデルパラメータサイズ10^3～10^8のオーダーの範囲で、両言語で同時にスケール則が成立することが実証できた.

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事前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.
- 実験の結果、混合データでの事前学習により、両言語において、単一言語で事前学習させた場合とほぼ同じ性能（Validation Loss）を同時に達成することが確認できたが、優位性までは確認できなかった.
- 今後、以下の検証を行う必要がある.
  - 言語の混合割合と性能の関係
  - 事後学習への影響
  - モデルサイズの更なる拡張による性能の変化

ご清聴ありがとうございました.