

東京大学松尾研究室

100 億パラメータサイズ・日英 2 ヶ国語対応の

大規模言語モデル “Web1ab-10B” を公開

——公開済みの日本語大規模言語モデルで最高水準（注 1）——

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（教授：松尾 豊、以下「松尾研」）は、この度日本語・英語の 2 ヶ国語に対応した 100 億パラメータサイズの大規模言語モデル（Large Language Model ;LLM）を事前学習と事後学習（ファインチューニング）により開発し、モデルを公開しましたのでお知らせします。今後も、Web1ab-10B のさらなる大規模化を進めるとともに、この資源を元に、LLM の産業実装に向けた研究を推進して参ります。

松尾研は、知能の謎を解くことを目的に人工知能の研究に取り組む研究室です。現在はテキスト生成で注目されることの多い LLM の技術ですが、今後は画像組み込みなどのマルチモーダル化、ブラウザ・ソフトウェア・ロボット等の行動制御の実装に発展し、人工知能研究を加速させると期待されます。また、LLM の開発競争が世界で激化する中、技術を理解した人材の育成も重要です。本研究開発は、上記の通り研究室の人工知能の研究を加速させるとともに、研究開発から得られた知見を講義開発等に生かすことで、大学における教育活動に資することも意図しています。

発表の詳細

近年の大規模言語モデルは、インターネットから収集した大量のテキストデータを学習に用いますが、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には現状では限界があります。そこで松尾研は、日本語だけでなく英語のデータセットも学習に用いることで学習データ量を拡張し、言語間の知識転移を行うことで日本語の精度を高めることを目的とした 100 億パラメータサイズの大規模言語モデル “Web1ab-10B” を開発し、公開しました。

日英 2 ヶ国語対応の大規模言語モデル開発にあたり、事前学習には代表的な英語のデータセット The Pile および日本語のデータセット Japanese-mC4 を使用しました。事後学習（ファインチューニング）には、Alpaca（英語）、Alpaca（日本語訳）、Flan 2021（英語）、Flan CoT（英語）、Flan Dialog（英語）の 5 つのデータセットを使用しました。事後学習の日本語データ比率が低い（注 2）にも関わらず、日本語のベンチマークである JGLUE 評価値が事前学習時と比べて大幅に改善（66→78%）し、言語間の知識転移を確認しました。この精度は、国内の公開モデルとしては最高水準であり（注 1）、海外の公開モデルと比較してもひけをとりません（下記公開モデル比較表を参照）。

今後、世界と伍する事のできるさらに大きな日本発の大規模言語モデルの開発に挑戦します。大規模言語モデルの開発に強い意思と情熱を持っている方は以下の採用ホームページをご確認ください。

■松尾研採用ホームページ

<https://web1ab.t.u-tokyo.ac.jp/join-us/jobs-2/>

■開発モデルの公開 URL

今回開発された Weblab-10B の事前学習済みモデル・事後学習済みモデルは、商用利用不可のモデルとして公開します。（下記 Hugging Face のページを参照）今後、商用利用可のモデル開発も検討していきます。

- ・事前学習済みモデル

<https://huggingface.co/matsuo-lab/weblab-10b>

- ・事後学習（ファインチューニング）済みモデル

<https://huggingface.co/matsuo-lab/weblab-10b-instruction-sft>

■公開モデル比較表（注1，注3）

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-calml-7b	45.44	24.22	37.63	74.12	45.79
cyberagent-open-calml-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

■生成サンプル文

😊: 日英2ヶ国語対応の大規模言語モデル開発にあたり、事前学習には代表的な英語のデータセットThe Pileおよび日本語のデータセット Japanese-mC4を使用しました。事後学習（ファインチューニング）には、Alpaca（英語）、Alpaca（日本語訳）、Flan 2021（英語）、Flan CoT（英語）、Flan Dialog（英語）の5つのデータセットを使用しました。事後学習の日本語データ比率が低い（※2）にも関わらず、日本語のベンチマークであるJGLUE評価値が事前学習時と比べて大幅に改善（66→78%）し、言語間の知識転移を確認しました。

質問：このモデルの特徴的なところは？

🗨️: このモデルは、英語と日本語の2つの言語間での知識転移を示すことができました。

😊: Translate the text into English.

🗨️: The model is able to demonstrate knowledge transfer between two languages, English and Japanese.

😊: 評価に利用したベンチマークは何ですか？また、事後学習後の評価値はいくつですか？

🗨️: The benchmark used for evaluation is JGLUE, which is a Japanese language evaluation tool. The evaluation score after post-training is 78%.

■モデル名 “Weblab-10B” の由来

Web 工学の研究から始まった松尾研が、研究室立ち上げ当初から長く使用し、メンバーからも親しまれているドメイン名を利用しました。Web という言葉には「蜘蛛の巣状の」という意味もあります。さまざまな人が交錯しながら価値を生み出していく松尾研のあり方にも重なるものでもあり、今回開発した 100 億パラメータ (10 Billion) の LLM が、新たな研究との結節点となることを期待しています。

今後も松尾研は大規模言語モデルの研究・開発を切り口に、人工知能技術の社会実装、大学における実践的な教育活動に取り組んでまいります。

注釈

(注 1) JGLUE 評価での実績 (2023 年 8 月 16 日時点)

JCommonsenseQA-1.1、JNLI-1.1、MARC-ja-1.1、JSQuAD-1.1 の平均値を使用。本モデルにおける評価は、Stability-AI/lm-evaluation-harness ライブラリを用いて、float16 でモデルロードを行い、template version 0.3 で few-shot in-context learning による評価を実施した (few-shot 数は 3, 3, 3, 2)。

(注 2) 事後学習には、以下のデータセットを使用 : Alpaca (英語)、Alpaca (日本語訳)、Flan 2021 (英語)、Flan CoT (英語)、Flan Dialog (英語)。JGLUE タスクに類似したクラス分類 / 文章読解を含むデータセットは Flan 2021 (英語) だが、日本語訳が含まれていないにも関わらずパフォーマンスが改善した。

(注 3) 本モデル以外の評価結果は、Stability-AI/lm-evaluation-harness から直接引用 (2023 年 8 月 16 日時点)

<https://github.com/Stability-AI/lm-evaluation-harness/tree/2f1583c0735eacdfdfa5b7d656074b69577b6774>

※このたび OSI (Open Source Initiative) の定義に倣い、Weblab-10B については商用利用不可のため “オープンソース” の定義に当てはまらないものとして、2023 年 8 月 18 日発出のリリースの一部文言を訂正しました。