日本語性能に優れた NECの大規模言語モデル（LLM）

小山田 昌史　秋元 康佑　董 于洋　矢野 太郎　竹岡 邦紘　槇尾 純太

要 旨

NECは日本語性能に優れた独自の大規模言語モデル（LLM）を開発し、社内活用や事業展開を進めています。本モデルはGPU1基でも動作するコンパクトな設計にもかかわらず、多量の高品質なデータによる長時間の学習、ロバストなアーキテクチャ、綿密な指示チューニングなどの工夫によって、世界トップレベルの日本語性能を有しています。また、「事業で使える」をモットーに、20万字を超える長文の処理、高速な推論など、実用面で言語モデルに必要となる要素を見極め、重点的に強化した、その設計思想、開発プロセス、性能について概説します。

KeyWords　大規模言語モデル／LLM／生成AI／汎用

1. はじめに

NECは2023年の初めに日本語性能に優れた独自の大規模言語モデル（以下、LLM）を開発し、7月に公表、その後は性能や機能面での強化を実施しながら、社内活用や事業展開を積極的に進めています。LLMは「ユーザーからの指示」に対して、「高い精度でレスポンスを返す」枠組みとなっています。**図1**は、そうしたLLMの開発プロセスを概観したものです。

本稿はまずモデルの設計について説明したのち、このプロセスに従って「事前学習用のデータ準備」「指示データによるChat-Tuning」「人間によるフィードバック（Alignment）」、そして得られたモデルの評価、について順に説明します。

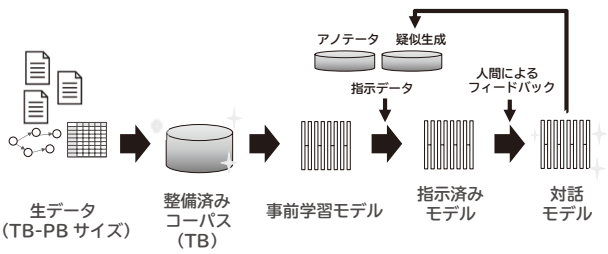

図1　大規模言語モデル（LLM）の開発プロセスの概要

2. モデル設計・事前学習

現在のほとんどのLLMは、Transformerと呼ばれるアーキテクチャに基づいています[1]。NECもTransformerに基づき、モデルのアーキテクチャを設計しました。

（1）モデルアーキテクチャ

モデルのサイズは性能と速度のバランスを鑑みて、パラメータ数が約130億（13B）となるように設定しました。レイヤ数、隠れ次元数、ヘッド数の3つのモデル形状パラメータは40、5120、40としています。これは LLaMA 13B[2]などと同じ値です。

（2）トークナイザー

トークナイザーは、日英合計10GBのコーパスを用いてBPEアルゴリズム[3]により学習したのち、トークンとして不適切なものを後処理で除外しました。

（3）事前学習

事前学習にはMegatron-DeepSpeed[4]の実装を利用しました。また、計算機としてはA100 80GBを8枚搭載したサーバを64基使用しています。バッチサイズは4Mトークン、学習率は10^{-4}から10^{-5}ま

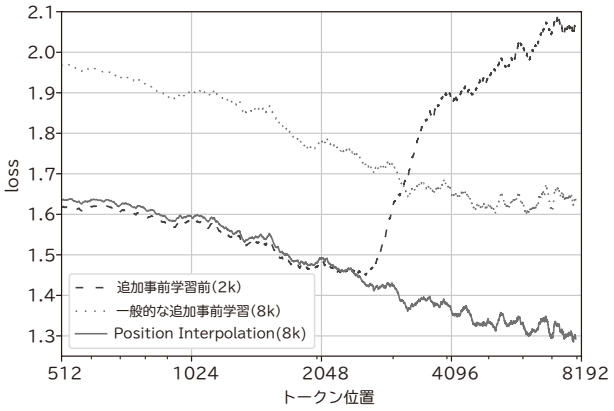

図２　学習方法の違いによる言語モデルの文章予測誤差（loss）の変化

でcosineスケジューラーで変化させました。このように、事前学習時の系列長は2048でした。より長い系列長に対する拡張は、次に述べます。

（4）長文適応

言語モデルに長文に対する推論能力を獲得させるため、NECは前述した事前学習を行った後に長文コーパスを用いた追加事前学習を行いました。一般的な追加事前学習手法を使って長文に対する性能を改善しようとすると、短文に対する性能が低下してしまうというトレードオフの関係が観測されたため、NECではPositional Interpolation[5]と呼ばれている手法をNECの言語モデル向けにカスタマイズすることによって、短文と長文どちらの性能も向上させることに成功しました（**図2**）。ここでは、トークン位置が大きい、もしくは小さい部分が、それぞれ長文または短文に対応しています。

3. 事前学習用のデータ準備

学習データについて、13Bというサイズで高い性能を実現するためには、極めて大規模かつ高品質なコーパスを準備する必要があります。そのために、さまざまな分野の知識が満遍なく含まれるよう、Web上のデータを中心にデータを大量に収集し、加工をほどこすことで、独自のコーパスを作成しました。このとき日本語だけでコーパスを構成するのではなく、英語やソースコードの情報を一定の比率で配分することによって、全体的な言語性能の向上を狙いました。また、LLMが品質の良い文章を生成できるようになるには、データの配分比率だけでなくデータを綺麗に加工することも必要となります。そのため、日本語のデータは、「ルールベースのフィルタリング」と「学習ベースのフィルタリング」によって多段階なクレンジングを実施しました。学習ベースのフィルタリングについては、汚ないデータと綺麗なデータを判別するような機械学習モデルを独自に学習したものを用いました。

4. 指示データによるChat-Tuning

事前学習直後のFoundation Modelは与えられた文章の続きを生成することはできますが、そのままでは適切に会話する能力を持ちません。ChatGPTのようにさまざまな指示に対して対話形式で対応可能にするためには、Chat-Tuningと呼ばれる学習が必要となります。Chat-Tuningはユーザーとアシスタントの会話を複数ターン用意し、対話形式での応答を学習させることで、ユーザーからの問い合わせに対してアシスタントが適切に回答したり、ユーザーとアシスタント間の対話履歴に基づいて回答を生成することが可能になります。

Chat-Tuning時の工夫として、NECはOpenAIによって開発されたChat Markup Language（以下、ChatML）というマークアップ言語を応用し対話データを半構造化しました。

ChatML とは、システム入力、ユーザーからの入力、アシスタントの入力から次のように半構造化する方法です。

<|im_start|>system
あなたは人の役に立つAIアシスタントです。
<|im_end|>
<|im_start|>user
NECの社長を教えてください。
<|im_end|>
<|im_start|>assistant
日本電気（NEC）の社長は、森田隆之氏が務めています。森田氏は、2021年4月1日に就任しました。
<|im_end|>

また、対話データとしてはNEC社員をはじめとする人間によって作成された大規模な対話データに加え、より複雑な指示や多様な質問に対する回答能力を向上させるため、NECが開発したLLM自身によって会話データを半自

動的に増幅する仕組みを構築し、活用しています。

5. 人間によるフィードバック（Alignment）

Chat-Tuningによって、モデルはユーザーからの指示に従うようになります。しかし、Chat-Tuningのみでは、有害な出力や役に立たない回答を確率的に生成してしまいます。Reinforcement Learning from Human Feedback（以下、RLHF）[6]は、このようなLLMの好ましくない回答を抑制し、好ましい回答を促す学習手法です。RLHFは通常、Chat-Tuningの後にreward modelの学習と、Proximal Policy Optimization（以下、PPO）[7]の2段階学習を行う必要がありますが、PPOのハイパーパラメータの選び方によって学習が安定しない課題があります。この課題を解決するために、近年Direct Preference Optimization（以下、DPO）[8]が注目されています。DPOは学習が1段階で安定しており、性能も良いため、NECではAlignmentのためにDPOを使用しました。

DPOは、ある与えられたプロンプトに対して、好ましい回答（正例）と好ましくない回答（負例）の3つ組 <プロンプト、正例、負例>のデータ（preference data）を使い、正例の尤度を上げ、負例の尤度を下げるように、直接モデルパラメータを最適化するアルゴリズムです。このように学習することで、モデルは好ましい出力のパターンを好ましくない出力のパターンよりも高い確率で出力するようになります。

preference dataは、NEC社員と外部企業に委託して集めたプロンプトに対してChat-tuningされたモデルで回答候補を出して作成しました。

こうしたAlignmentのプロセスを踏むことによって、LLM回答の人間評価の品質向上や有害な出力の抑制などの効果が確認されています。

6. ベンチマーク

LLMの性能の評価は多角的ですが、ここでは最も一般的な観点として、「日本語能力」と称される常識推論や文書読解など「基盤モデルとしての推論・情報処理能力」と、複合的な能力が必要とされる「対話モデルとしての情報処理能力」の2つの観点で評価します。

6.1 基盤モデルとしての評価

基盤モデルを評価するために、一般的に使われる日本語ベンチマークであるJGLUE[9]のなかからJSQuADとJCommonsenseQA（以下、JCQA）を採用しました。JSQuADは、与えられた文書から質問に対応する回答文字列を抽出するタスクであり、2-shotのin-context learningを用い、評価スコアは推論結果が正解文字列と完全に一致するかどうかで評価します。JCQAは、常識的な知識を問う選択式QAタスクであり、5つの選択肢から回答を選択します。推論時には、3-shotのin-context learningを用い、正解率で評価します。比較対象のモデルは、グローバルでトップレベルとなる海外モデルと、2023年10月末時点でNECが利用可能な日本語のモデルのうち性能が高かったもの（A，B，C，…と表記）としています。

評価結果を、**図3**に示します。JSQuADでは比較したモデルのなかで最高性能、JCQAでは国内モデルではトップの性能で海外モデルを含めても2番目の性能を達成しました。

6.2 対話モデルとしての評価

次に対話モデルとしての性能を測る日本語ベンチマークとして、RAKUDA[10]を採用しました。RAKUDAのタスクは前述のような知識や読解だけでなく、対話として妥当な回答を求めるものです。RAKUDAは、40個の日本語の自由形式のクエリに対して回答するタスクであり、回答の質はGPT-4によって評価します。また、回答を2つのモデルで生成したうえで、どちらの回答が良いかという観点で評価する方式をとります。ここでは、NECの開発したLLMが、NECが利用可能な日本語のモデルに対してどの

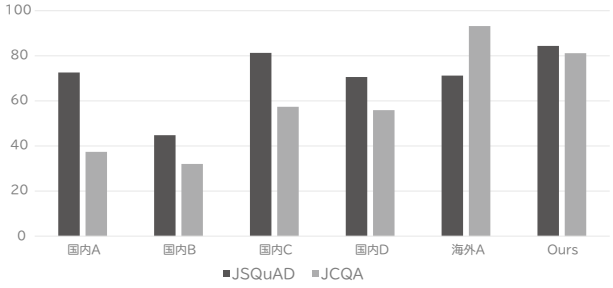

図3　JSQuAD、JCQAの実験結果（NECが開発したLLMをOursと表記）

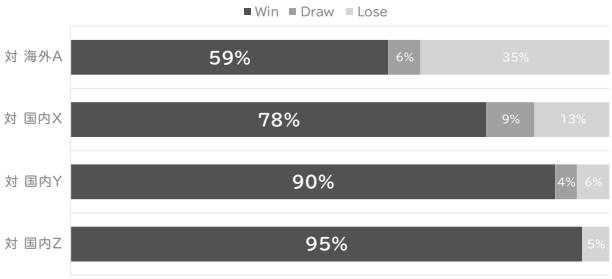

図4　RAKUDAベンチマークにおいて他モデルに対するNECの開発したLLMの勝率

程度の勝率になるかという観点で評価します。2023年10月末時点でNECが利用可能な日本語モデルのうち性能が高かったものを、X, Y, Zと表記しています。

図4にRAKUDAベンチマークにおける評価結果を示します。NECが開発したLLMは、グローバルでトップレベルとなるモデルを含めて比較しても高い性能を達成しており、対話性能として十分な能力を持っていると言えます。特に日本語特化のLLMと比較すると大きく勝率で上回っています。ただし、RAKUDAベンチマークは1ターンの対話かつ少数データでの評価のため、対話性能の一側面しか測れていないことには注意が必要となります。

7. むすび

本稿では、NECの開発した13BのLLMについて紹介しました。今後NECは本モデルを皮切りに、さまざまな生成AI技術を開発し、世の中に貢献していきます。

* ChatGPTは、米国OpenAI社の商標です。
* その他記述された社名、製品名などは、該当する各社の商標または登録商標です。

参考文献

1) Ashish Vaswani, et al.: Attention Is All You Need, 2017
https://arxiv.org/abs/1706.03762
2) Hugo Touvron et al.: LLaMA: Open and Efficient Foundation Language Models, 2023
https://arxiv.org/abs/2302.13971
3) Rico Sennrich, Barry Haddow, and Alexandra Birch: Neural Machine Translation of Rare Words with Subword Units, 2015
https://arxiv.org/abs/1508.07909
4) GitHub: Megatron-DeepSpeed
https://github.com/bigscience-workshop/Megatron-DeepSpeed
5) Shouyuan Chen et al.: Extending Context Window of Large Language Models via Positional interpolation, 2023
https://arxiv.org/abs/2306.15595
6) Long Ouyang et al.: Training language models to follow instructions with human feedback, 36th Conference on Neural Information Processing Systems, 2022
https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
7) John Schulman et al.: Proximal Policy Optimization Algorithms, 2017
https://arxiv.org/abs/1707.06347
8) Rafael Rafailov et al.: Direct Preference Optimization: Your Language Model is Secretly a Reward Model, 2023
https://arxiv.org/abs/2305.18290
9) Kentaro Kurihara et al.: JGLUE: Japanese General Language Understanding Evaluation, LREC 2022, pp. 2957-2966, 2022
http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.317.pdf
10) GitHub: Japanese-llm-ranking
https://github.com/yuzu-ai/japanese-llm-ranking

執筆者プロフィール

小山田 昌史
データサイエンスラボラトリー
主席研究員・研究グループ長

秋元 康佑
データサイエンスラボラトリー
主任研究員

董 于洋
データサイエンスラボラトリー
特別研究員（主任）

矢野 太郎
データサイエンスラボラトリー
主任

竹岡 邦紘
データサイエンスラボラトリー
特別研究員（主任）

槇尾 純太
データサイエンスラボラトリー

NEC 技報のご案内

NEC技報の論文をご覧いただきありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報（日本語）　NEC Technical Journal（英語）

Vol.75 No.2　ビジネスの常識を変える生成AI特集
～社会実装に向けた取り組みと、それを支える生成AI技術～

ビジネスの常識を変える生成AI特集によせて
生成AI技術への取り組み ～基盤から応用、ルール作りまで～

◇ 特集論文

急速に広まる生成AIの市場適用
NECにおける生成AIの取り組みについて
電子カルテと医療文書の作成支援による医師業務効率化
映像×LLMによる報告書作成業務の自動化
映像分析と生成AIによるリアル世界の行動理解
サイバー脅威インテリジェンス生成自動化
生成AIの社内活用を進めるNEC Generative AI Service（NGS）
ソフトウェア・システム開発への生成AIの活用
LLMとMIで革新する素材開発プラットフォーム
LLMと画像分析を活用した被災状況の把握

生成AIの可能性を高める基盤技術
日本語性能に優れたNECの大規模言語モデル（LLM）
NECの生成AIを支える国内企業で最大規模のAIスーパーコンピュータ
より安全な大規模言語モデル（LLM）を目指して
データを秘匿したまま連携できる連合学習技術とLLMへの適用可能性
大規模言語モデル（LLM）によるFew-shotクラスタリングの可能性
オープンドメイン常識推論のための知識拡張型プロンプト学習
AI連携とオーケストレーションのための基盤ビジョンLLM
クエリを考慮した新規手法により関連する企業データを減らしLLM APIの使用コストを最適化

AI技術が社会へ浸透するために
AI標準化・ルールメイクに関する動向とNECの取り組み
人権尊重に向けたNECのAIガバナンスの取り組み
RCModelを用いたAIリスクマネジメントのための人材育成事例

◇ NEC Information
2023年度C&C賞表彰式典開催

Vol.75 No.2
（2024年3月）

特集TOP