

ichikara-instruction

LLM のための日本語インストラクションデータの作成

関根聡¹ 安藤まや² 後藤美知子¹ 鈴木久美¹ 河原大輔^{3,1} 井之上直也^{4,1} 乾健太郎^{5,6,1}

¹理化学研究所 AIP ²フリー ³早稲田大学 ⁴北陸先端科学技術大学院大学

⁵MBZUAI ⁶東北大学

satoshi.sekine@riken.jp

概要

GPT を始めとする対話型生成 AI は主に、事前学習、Supervised Fine Tuning (SFT)、強化学習を用いて構築された言語モデルで構成されている。特に SFT で利用されるインストラクションは非常に重要な役割を担っている。我々は 2024 年 3 月までに、高品質な日本語インストラクションデータを 1 万件作成する計画である。本論文では 2023 年 12 月 21 日に共同研究企業に向けて公開したデータの概要と、それに基づく評価結果、最終目標、データの作成手順、プロジェクトの形態と今後について紹介する [1]。

1. はじめに

GPT を始めとする生成 AI (LLM) が、数多くのタスクにおいて非常に高い精度で対話的な言語生成を行い、実用化が広がるなどの影響を与えている。特に GPT を開発している OpenAI は大きな影響を与えた代表格と言える。OpenAI は、その技術の詳細は公開していないが、GPT は事前学習、インストラクションに基づく Supervised Fine Tuning (SFT)、強化学習の 3 つのコンポーネントからなり、特に、SFT における高品質なインストラクションが重要な役割を担っていることを報告している [2]。しかしながら、高品質かつある程度の規模の日本語インストラクションデータは未だ存在しない。

LLM の開発は英語で特に進んでおり、インストラクションデータについても、公開されているものがある。特に Dolly [3], Open Assistant (OASST) [4] と呼ばれる大規模データは、日本語に自動翻訳されたデータも存在する。また、過去に日本語の言語処理タスクのために作られたデータを対話形式に加工し、インストラクションデータとして公開しているものもある [5] [6]。しかしながら、現状、それらの

データを用いて SFT された LLM では、GPT のような高品質な回答の生成には至っていない。したがって、GPT が実現しているような出力をする LLM を作成し、その透明性を実現するためには、単なるクイズや質問応答 (QA) のような単純なインストラクションだけでは不十分であり、下記の要件が満たされるインストラクションの作成が必要であると考えられる。

- 一般の人が LLM で聞くであろう自然で幅広い種類の質問がカバーされている
- 回答は流暢で読みやすく、豊富な情報量を含む
- 質問や回答の種類を認識し、LLM の透明性に関する研究のために、タスク、分野、回答タイプなどのタグを付与する

本稿では、このような高品質なインストラクションデータを作成することを目的とした「LLM のための日本語インストラクションデータ作成プロジェクト」を紹介する。本プロジェクトで構築されているデータは“**ichikara-instruction**”と呼ばれ、理研 AIP の言語情報アクセス技術チームが中心となり、2023 年 7 月に企業との共同研究という枠組みで作成を開始した。データの作成は、森羅プロジェクト [7] の教師データ作成などの経験がある 15 人のアノテーターと共に開始した。その後、アノテーターの人数を大幅に増やし、2024 年 1 月現在、40 名程度のアノテーターと 3 社の外注企業で、データ作成作業を行なっている。現在のところ、質問と回答は自然言語での 1 ターンで、一般的な分野を対象としているが、将来的には安全性対策、マルチターン、マルチモーダル、専門分野のデータを作成していく予定である。

本データは共同研究に参加された企業には随時共有し、安価な商用ライセンスも提供させて頂いている。完成時には研究目的 (CC-BY-NC-SA) での一般公開、有償での商用ライセンスの提供も計画している。

2. 関連研究

OpenAI の InstructGPT の論文[2]によると、GPT の学習は3段階からなる。

1. 事前学習：大量のコーパスを用いて大規模なパラメーター学習を行う
2. Supervised Fine Tuning (SFT)：インストラクションを用いて教師付き学習を行う
3. Reinforcement Learning from Human Feedback (RLHF; 強化学習)：多くのインストラクションに対してモデルを走らせ、多様な出力を得る。人間がモデルにその優劣を教え、モデルをより良いものにする。

論文では、この中で SFT の重要性が強調されている。InstructGPT は 14,428 件のデータによって学習され、非常に大きな精度向上が見られたとされている。作業者の選抜、作成方針、作成データの分布や例等も報告され、インストラクション付きの 1.3B の GPT-3 のモデルが、175B のモデルよりも高性能であったことが報告されている。

また、オープンな LLM を用いてインストラクションの影響を調べた論文[8]では、結論として下記のように報告されている。

- 特定の目的やジャンルのインストラクション学習はそれらの性能を大幅に向上させる
- インストラクション学習は、論文で報告されているすべての実験において性能を上げた
- この実験で使われたすべてのインストラクションを用いて学習されたモデルは、評価タスク全体の平均値としては最良だが、個別の評価タスクにおいて常に最良というわけではない
- オープンなモデルとデータを使うだけでは、ChatGPT に遥かに及ばない
- 出力が長いほど、評価結果が良くなる傾向が見られた

ここでも、インストラクションによる SFT の重要性が認められる。もちろん、インストラクションを用いた SFT ができるほどの高精度な LLM が構築されていることも重要ではあるが、ここで示されているような高品質のインストラクションを作成することが、LLM 構築の重要な鍵であることがわかる。

次に一般に公開されている2つの日本語インストラクションデータセットを検証する。

まず、DataBricks が提供している Dolly という 15,000 件からなるインストラクションデータを紹介する[3]。これは DataBricks の社員に英語で質問と回答を作成させそれを機械翻訳したものである。しかし、データには似た質問が数多くあり、回答もシンプルである。また、日本語翻訳のインストラクションデータ[5]には図1のような、翻訳による問題や言語や文化による問題が見られる。

<翻訳の問題>

"オランダとオランダの違いは何ですか？",

"What is the difference between Holland and The Netherlands?"

"雪見だいふく"

"what are some things you can do on a snowy day",

<文化の問題>

"曜日を開始時のアルファベットで分類する。日曜日、月曜日、火曜日、水曜日、木曜日、金曜日、土曜日",

"人気ミュージシャンを5人挙げる",

→ "テイラー・スウィフト、ヴァンス・ジョイ、ザ・スクリプト、コールドプレイ、レディー・ガガ"

図1. 問題のある Dolly 翻訳データの例

次に、Open Assistant[4]を紹介する。これはクラウド参加者に、ある形式で自由に対話を行わせ、そのデータを木構造のインストラクションとして構築したものである。主に対話を目的にしたデータであり、自然な会話を作るために有効であることは報告されているが、いわゆる、質問応答や要約や翻訳などのタスクに対して有効であるかは未知数である。

3. データの概要

インストラクションデータは 2023 年 12 月 21 日の時点で、質問回答ペア 4,802 件が完成しており、3 月末までに 1 万件が完成する予定である。4,802 件中、200 件の質問には 2 名が回答を作成し、別の 200 件の質問には 5 人が回答を作成している。3 件のインストラクションの具体例を付録に掲載する。

各インストラクションには LLM の透明性の研究に利用することを目的にタグを付与している。最終的には全データに付与するが、2024 年 1 月現在では 951 件にタグが付与されている。現状のタグ種類の定義および例と、「操作」タグの値の統計データを、表1、表2に示す。

| タグの種類 | 説明 | 値の例 |
|-------|--------------------------------|--|
| 操作 | 質問がどのような操作を求めているか | オープン QA、創作、定義など（表 2 を参照） |
| 主観客観 | 質問内容がどのような種類の回答を求めているか | 唯一解、客観、主観 |
| 時間依存 | 質問の回答が時間に依存しているか | あり、なし |
| 対象 | 質問を回答するのに必要な能力 | 知識、数学、言語知識、経験、創作、読解力 |
| 分野 | 質問内容の分野（一般的な教科の種類や新聞面などの粒度で定義） | 社会、人物、数学、物理、産業、経済、天文、気象、健康、グルメ、ゲーム、映画、音楽など |
| 回答タイプ | 回答の形式的タイプ | 文章、文、単語、固有名詞、地名、時間表現、YES/NO、詩、X のリスト、X のグループなど |

表 1. タグの種類

| 操作の分類 | 951 件での分布 | 1 万件での分布(予定) |
|----------|-----------|--------------|
| オープン QA | 406 | 3200 |
| クローズド QA | 60 | 500 |
| ブレスト | 173 | 1300 |
| 創作 | 78 | 300 |
| 定義 | 82 | 800 |
| 要約 | 11 | 400 |
| 抽出 | 20 | 400 |
| 例示 | 38 | 500 |
| 分類 | 40 | 500 |
| 選択 | 38 | 500 |
| 穴埋め | 5 | 300 |
| 書き換え | 0 | 200 |
| 校正 | 0 | 200 |
| 翻訳 | 5 | 500 |
| 数学 | 18 | 400 |

表 2. 操作タグと頻度

4. データ作成の手順

(1) 質問の作成

質問は自由な発想で作成する方法と、既存の質問を参考に、それにインスパイアされるような質問を作るという方法で作成した。既存の質問は、Dolly, Alpaca[9]などのインストラクションデータや、コミュニティ質問応答データ、ブログなどに載っている質問などを利用した。もちろん、著作権の問題からそのまま利用することは一切行っていない。また、一部の自由な発想で作ることが得意なアノテーターには、自由な発想の質問を積極的に作ってもらった。質問には誤字脱字が散見されるが、これは実

際のシステムでも起こり得ることと考え、修正していない（例えば、「AKB45」と書かれている質問があるが、回答はその点について「AKB48 のことだと思います」と指摘してから、AKB48 に関する回答を作成することを基準としている）。

(2) 回答の作成

上記の質問に対し、原則として異なるアノテーターが回答を作成している。これは、質問が自身の得意分野に偏らないようにするための配慮である。また回答は、GPT のような丁寧な回答になるよう、度重なる試行錯誤を経て作成方法をブラッシュアップさせた。回答に ChatGPT の出力を利用すると、データを LLM 開発に利用できなくなるため、アノテーターには極力、利用しないように指導している。利用した際にも、全く異なる回答を作成することとしている。また、アノテーターからの質問は作成要領の指針を決める貴重な材料となり、データを作る作業自体が学びの場であった。

(3) 回答のチェック

回答が日本語として適切か、回答作成の方針を逸脱していないか、Web ページや GPT などの出力そのままではないかなどのチェックを 2 人が独立に行い、高品質なデータ作成に努めた。

(4) 質問・回答に対するタグ付け

3 節で述べたタグを全データに付与する予定である。このタグ定義も試行錯誤を重ねており、現在のものが最終形ではなくなる可能性もある。タグ付けは揺れがないように、このようなタスクが得意なアノテーターの一人による作業として進めてきたが、全体にタグ付けするのは困難であるため、このアノテーターを中心に 6 名で作業を行なっている。

5. 評価

共同研究を行なっている企業が SFT で **ichikara-instruction** データを利用した LLM の評価結果を報告している。詳しくは引用論文やプレスリリースを参照して頂きたいが、簡単に紹介する。

株式会社マネーフォワードによる評価 [10]

rinna/youri-7b (base system) を ichikara-instruction で SFT 学習させた houou システムを Rakuda ベンチマーク[11]で評価した結果が表3の通りである。v1 は base システムを **“ichikara-instruction”** の 2,903 件で学習させたもの、v2 は 4,802 件で学習させたものである。ベースシステムや他のインストラクションに対する本インストラクションの優位性が示されている。また、Rakuda ベンチマークは GPT-4 による自動評価なため信頼性には疑問はあるが、この評価では GPT-3.5 よりも高い勝率を得ている。

| | 相手システム | houou 勝率 | 相手勝率 | 引き分け |
|----|--------------------|----------|------|------|
| v1 | base | 85.0 | 7.5 | 7.5 |
| | base+Dolly 日本語 | 82.5 | 12.5 | 5.0 |
| | base+OASST 日本語 | 77.5 | 15.0 | 7.5 |
| v2 | GPT-3.5 turbo 1106 | 67.5 | 30.0 | 2.5 |

表3. houou システムの評価結果

ストックマーク株式会社による評価 [12]

独自に 2,200 億トークンから事前学習したモデルを、15k の Dolly, 89K の OASST, 52K の Alpaca および 1,003 件の **“ichikara-instruction”** のそれぞれで追加学習し、Stockmark Business Question を用いて人手で評価したところ、データの規模は他のデータに比較して 1/15 以下であるにも関わらず、**“ichikara-instruction”** での正解率が 1 位であったことが報告されている。回答は単に質問の答えを述べるだけでなく、より詳細な情報が伴う傾向が見られたことも報告されている。

株式会社オルツによる評価 [13]

(株) オルツが公開した LHTM-OPT の追加学習に **“ichikara-instruction”** の 2,903 件のデータが利用され、Rakuda ベンチマークにおいて、当時の国内最高性能が達成したと報告されている。

これらの評価から、少なくとも **“ichikara-instruction”** は Dolly や OASST の日本語訳に比較して、有意に LLM の精度を向上させており、高品質なインストラクションデータが LLM の構築のために重要であることがわかる。

6. プロジェクトの形態と今後

本インストラクションデータの作成には約 2,000 万円の費用をかけている。この費用を捻出するために本プロジェクトは、企業との共同研究という形式をとった。研究目的のデータ公開(CC-BY-NC-SA)を前提としているものの、データの逐次公開と作成ノウハウの共有に対し 200 万円の共同研究としたところ、21 社からの応募があり、1 月 12 日現在、15 社との契約が成立している。LLM のための日本語インストラクションの作成に対して大きな期待があることがわかる。多額の費用を多くの参加者で分担し、共有財産を構築することで日本の LLM の一刻も早いキャッチアップを実現すると同時に、LLM の透明性の実現に貢献したいと考えている。このデータは、共同研究企業とそれ以外で費用の差はあるが、有償にて商用ライセンスをすべての企業団体向けに提供している。詳細はホームページ[1]を参照頂きたい。

今後は、まずこのデータを利用して日本語 LLM の底上げに貢献したい。インストラクション構築の今後に関しては、4つの方向性を考えている。それらは、マルチターン、マルチモーダル、安全性、ドメイン依存のインストラクションの構築である。詳細については別の機会に紹介していきたい。

7. まとめ

本論文では、LLM のための日本語インストラクションデータ **“ichikara-instruction”** を紹介した。高品質なインストラクションを丁寧に作成し、既に、複数の企業が利用して、LLM の精度が向上することが示されている。2023 年 12 月 21 日には 4,802 件のデータを共同研究企業に、2,903 件を研究目的で一般に公開したが、2024 年 3 月末までに 1 万件を完成させる予定である、研究目的には CC BY-NC-SA にて一般公開をしており、商用利用については有償でのライセンス提供を行なっている。

謝辞

本インストラクションデータは2024年1月初旬時点で、15の企業様との共同研究において構築されている。他に、4社との契約締結作業中である。すでに契約を締結した15の企業は以下の通りである。(順不同 敬称略)

- 株式会社日本総合研究所
- 三菱電機株式会社
- 株式会社マネーフォワード
- スtockマーク株式会社
- 株式会社レトリバ
- 株式会社オルツ
- 株式会社フィックスターズ
- ソフトバンク株式会社
- ファーストアカウンティング株式会社
- みずほリサーチ&テクノロジーズ株式会社
- 株式会社リコー
- TOPPAN デジタル株式会社
- 株式会社 JSOL
- 株式会社 ELYZA
- 他匿名1社

参考文献

1. ichikara-instruction. LLMのための日本語インストラクション作成プロジェクトホームページ:
<https://liat-aip.sakura.ne.jp/wp/llmのための日本語インストラクションデータ作成>.
2. Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin et al. Training language models to follow instructions with human feedback. NuerIPS 2022. arXiv:2203.02155v1 [cs.CL] 4 Mar 2022.
3. Dolly Dataset :
<https://huggingface.co/datasets/databricks/databricks-dolly-15k>
4. Andreas Kopf, Yannic Kilcher. “OpenAssistant Conversations – Democratizing Large Language Model Alignment. Open Assistant”.
<https://arxiv.org/abs/2304.07327>
5. Dolly日本語翻訳データ
<https://github.com/kunishou/databricks-dolly-15k-ja>
6. 平野正徳, 鈴木雅弘, 坂地泰紀. “llm-japanese-dataset v0: 大規模言語モデルのための日本語チャットデータセット構築”. Jxvi 383.
<https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/383>
7. 森羅プロジェクトホームページ <http://shinra-project.info>
8. Yizhong Wang et al. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. NeurIPS2023. arXiv:2306.04751v1 [cs.CL] 7 Jul 2023
9. Alpaca Homepage
<https://crfm.stanford.edu/2023/03/13/alpaca.html>
10. MFの論文 (TBA)
11. Rakudaベンチマーク :
<https://yuzuai.jp/benchmark>
12. Stockマークのプレスリリース :
https://tech.stockmark.co.jp/blog/202311_stockmark_13b_instruct/
13. オルツのプレスリリース :
<https://prtimes.jp/main/html/rd/p/000000053.000111359.html>

A 付録 (ichikara-instruction の例)

```
[
{
  "ID": "ichikara-instruction-003-001-0000002-001",
  "text": "「1匹=いっぴき」と数えるのに、「2匹→2びき」と数えない理由を答えなさい。",
  "output": "日本語には、「助数詞」と言って「数量を表現するのに用いる接尾語」があります。¥nこの助数詞は、「種類の多さ」と「音の変化=変音現象」が特徴です。¥n¥n紙→枚（まい）、車→台（だい）、列車→輛（りょう）、箸→膳（ぜん）、エンジン→基（き） ¥n¥n種類の多さもさることながら、もう一つの難しさが変音現象です。¥n例えば、カエルを数えるときに「1匹=いっぴき」と読むのに対して「2匹≠にびき」で2匹を「にびき」とは読まずに「2匹=にひき」となります。¥nどうしてこうなるのか。¥nここには、漢語の半濁音のルール、というものが発生します。¥nまず、半濁音はハ行にだけしか付きません。（はひふへほ→ばびぶべぼ） ¥nそして、連濁といって語と語が合体するときに後ろに来る語の頭の部分が清音から濁音に変化します。¥nハ行音の連濁のうち、前の語の最後に「つ」がくると、後ろの語の頭が半濁音に変わります。¥n¥n例）切（せつ） + 腹（ふく） = 切腹（せつぷく） ¥n ¥n助数詞も、数字の「一・六・八・十・百」が「ハ行音の助数詞」と結びつく時、促音便「っ（小さいっ）」を起し、後ろの助数詞が半濁音となります。¥n1杯（いち+はい→いっぱい）、6泊（ろく+はく→ろっぱく）、8袋（はち+ふくろ→はっぷくろ）、¥n10編（じゅう+へん→じっぺん）、100本（ひゃく+ほん→ひゃっほん） ¥n¥nこういった理由から、「2匹→2びき」とはなりません。",
  "meta": { "操作": "オープンQA", "主観客観": "客観", "時間依存": "なし", "対象": "言語知識", "分野": "国語", "回答タイプ": "文章" }
}, {
  "ID": "ichikara-instruction-003-002-0002166-001",
  "text": "「自由を手に入れた喜び」をポエムで表現してください。",
  "output": "「自由を手に入れた喜び」のポエムを作成するにあたって、まずは、それをどのように表現していくか、について考えてみます。¥n¥n¥n 自由とは、なにものにも束縛されず、強要されず、解放された状態を指しますので、「広い空間」を思わせるような世界観で描くとよいでしょう。¥n¥n それによって得られた喜びは「躍動感に富んだ表現」を用いることで、より鮮明に、より心躍る感じを出すことができます。具体的には、「短い言葉」や「反復法」を用いると良いでしょう。¥n¥n¥n¥n¥n¥n それでは、一例を挙げてみます。¥n¥n¥n~ポエム~¥n¥n¥n ぐんぐんぐんぐん スピード上げて¥n¥n ザブンッ 白い光の中に飛び出した¥n¥n グッと身体を突き出して¥n¥n 青いキャンバス突き破れ！ ¥n¥n 輝く光を手に入れろ！ ¥n¥n¥n¥n 長い長い線を描いて ¥n¥n どれだけ長く引けるかな¥n¥n ザブンッ 誰も描いたことのない¥n¥n 青いキャンバス突き破れ！ ¥n¥n 輝く自由を手に入れて！ ¥n¥n¥n ずっと向こうのアオの彼方へ",
  "meta": { "操作": "創作", "主観客観": "主観", "時間依存": "なし", "対象": "創作", "分野": "国語", "回答タイプ": "詩" }
}, {
  "ID": "ichikara-instruction-003-001-5000088-005",
  "text": "日本の観測史上、もっとも梅雨明けが遅かったのは何年のことですか?",
  "output": "気象庁発表の「昭和26年（1951年）以降の梅雨入りと梅雨明け（確定値）：関東甲信」によると、関東甲信で最も遅く梅雨明けしたのは1982年の8月4日です。この年は6月17日から8月4日までの101日間梅雨が続いていました。また、梅雨の期間が一番長かったのは、2020年の6月11日から8月1日までの174日間です。¥n¥n尚、梅雨明けの時期は、地域によって変わりますので、他の地域に関しては気象庁のHPをご確認ください。",
  "meta": { "操作": "オープンQA", "主観客観": "唯一解", "時間依存": "あり", "対象": "知識", "分野": "気象", "回答タイプ": "時間表現" }
}
]
```