大規模言語モデル houou (鳳凰): 理研 ichikara-instruction データセットを用いた学習と評価

小島淳嗣 [1] 北岸郁雄 [1]
[1] 株式会社マネーフォワード
kojima.atsushi@moneyforward.co.jp

概要

supervised fine-tuning (sft) によって大規模言語モデルの指示追従性を向上させるには、prompt と completion のペアで構成される インストラクションデータが必要となる。このようなインストラクションデータの作成は、GPT-4 などの学習済みモデルからの出力を利用する方法を除くと、人手で prompt と completion を記述する必要があり、アノテーションコストが高い。それゆえ、日本語のインストラクションデータは、英語で作成されたインストラクションデータを日本語に翻訳することで 得るアプローチが大半であった。本稿では、日本語によってフルスクラッチから作成されたインストラクションデータセットである ichikara-instruction を用いてモデルを学習することで、自動翻訳で作成されたデータによって学習されたモデルに比べて高い日本語生成能力が獲得できることを示す。実験では、Rakuda Benchmark を用いて、翻訳データによって学習されたモデルと日本語インストラクションデータで学習されたモデルの性能を GPT-4 によって比較した。その結果、gpt-3.5-turbo-1106 、日本語に自動翻訳した dolly と OpenAssistant Conversations によってそれぞれ学習された sft モデルに対する勝率はそれぞれ 67.5%、82.5% 、70% となり、日本語による高品質なインストラクションデータが LLM の日本語生成能力向上に重要であることが示された。

1 はじめに

large langugage model (LLM) の指示追従性を効率的に向上させる学習手法として supervised fine-tuning (sft) がある。sft モデルは、prompt と completion のペアで構成される インストラクションデータを用意し、事前学習モデルに対して fine-tuning することで学習される。

このようなインストラクションデータの作成は、GPT-4 などの学習済みモデルからの出力を利用する方法を除くと、人手で prompt と completion を記述する必要があり、アノテーションコストが高い。それゆえ、日本語のインストラクションデータは、英語で作成されたインストラクションデータを日本語に翻訳することで 得るアプローチが大半であった。

本稿では、日本語によってフルスクラッチから作成されたインストラクションデータセットである ichikara-instruction [1] を用いて sft モデルを学習することで、翻訳によって作成されたデータを用いて学習されたモデルに比べて、高い日本語生成能力が獲得できることを示す。

実験では、Llama2 [2] 7B において日本語データを用いて継続事前学習 [3] を行ったモデルに対して、ichikara-instruction で sft を行うことで、 鳳凰 (houou) [1)] を学習した。評価では Rakuda Benchmark [2)] 、JGLUE、ELYZA-tasks-100 [3)] において、他の日本語インストラクションデータで学習された sft モデルと houou の性能を比較した。その結果、Rakuda Benchmark を用いた評価では、gpt-3.5-turbo-1106 、日本語に自動翻訳した dolly [4)] と OpenAssistant Conversations (OASST) [5)] によってそれぞれ学習された sft モデルに対する勝率はそれぞれ 67.5%, 82.5 % と 70 %となり、日本語による高品質なインストラクションデータが LLM の日本語生成能力向上に重要であることが示された。

1) https://huggingface.co/moneyforward/houou-instruction-7b-v2
2) https://huggingface.co/datasets/yuzuai/rakuda-questions
3) https://huggingface.co/datasets/elyza/ELYZA-tasks-100
4) https://huggingface.co/datasets/databricks/databricks-dolly-15k
5) https://huggingface.co/datasets/OpenAssistant/oasst1

2 大規模言語モデル houou (鳳凰) の学習

インストラクションデータは、理研 AIP によって作成され、共同研究企業に提供された ichikara-instruction を用いる。このデータセットにおいて、prompt は dolly や Alpaca などの既存のインストラクションデータや、様々なドメインの Q&A サイトの質問を参考にして作成され、completion はアノテータが日本語でフルスクラッチから作成された。詳細は理研の ichikara-instruction の論文 [1] を参照いただきたい。データは継続的にリリースされており、それぞれのバージョンのデータセットには、1003 件, 2903 件, 4802 件のインストラクションデータが含まれる。実験では、2023 年 12 月 21 日に提供された 4802 件のデータを用いた実験結果について報告する。

houou の学習について述べる。事前学習モデルには Llama 2 7B [6] に対して、日本語テキストを用いて継続事前学習 [3] を行ったモデル [7] を用いた。図 1 に sft におけるモデルの入力フォーマットを示す。prompt と completion の境界を区別するために特殊な文字列として「### 応答:」を使用し、学習時には prompt を構成するトークンの loss を 0 にして、completion を構成するトークンの loss のみ最小化することでモデルを学習した。学習条件を表 1 以下に示す。

以下は、タスクを説明する指示です。要求を適切に満たす応答を書きなさい。

指示:
面接で使える面白い質問をおしえてください。なぜ面白いのかもおしえてください。

応答:
面接で使える面白い質問として、次のような質問はいかがでしょうか？

「動物占いでは、あなたは何の動物ですか？」

この質問がなぜ面白いのかというと、動物占いは生年月日と性別から人間の個性を12種類の動物に当てはめて分類する占い方法ですが、それぞれの動物の特徴がその人に当てはまっているのか、当てはまっていないのか聞くことにより、その人が自分の性格をどのように自己分析しているのか知ることが出来るからです。

(略)

図 1 sft における入力フォーマット

表 1 sft モデルの学習条件

Parameter	Value
Learning rate	2×10^{-5}
Optimizer	AdamW
Number of epochs	30
Context length	2048
Batch size	256

3 実験

3.1 実験条件

評価では、Rakuda Benchmark、JGLUE [8]、ELYZA-tasks-100 を用いて houou の性能を他の日本語 LLM と比較した。ELYZA-tasks-100 では、100 の質問に対するそれぞれのモデルの回答に対して 5 段階のスコアを付与し、その平均を比較した。Rakuda Benchmark と ELYZA-tasks-100 では、gpt-4 による自動評価を行った。

3.2 結果

図 2 に Rakuda Benchmark における結果を示す。houou は、日本語に翻訳された dolly と OASST によってそれぞれ学習された sft モデル、及び dolly や FLAN を混ぜて学習された sft モデル [9] の性能を上回った。さらに、gpt-3.5-turbo-1106 との比較においても、houou は、Rakuda Benchmark の 40 の質問のうち、67.5% に対して gpt-3.5-turbo-1106 よりも優れた出力を行った。なお、人間による評価と gpt-4 による自動評価結果との比較と分析については、関根らの論文 [4] を参照されたい。

表 2 に JGLUE の結果を示す。表において、JCommonsenseQA は accuracy、JNLI と MARC-ja は balanced accuracy、JSQuAD は F1 の値を示す。houou は、JGLUE の 4 つのタスクのうち、3 つにおいて、日本語に翻訳された dolly と OASST によってそれぞれ学習された sft モデルよりも高い精度となった。

表 2 JGLUE の結果

Model	JCommonsenseQA 3-shot	JNLI 3-shot	MARC-ja 0-shot	JSQuAD 2-shot
houou	**80.25**	**45.3**	**91.48**	65.26
SFT trained by dolly	78.73	41.77	80.32	**76.37**
SFT trained by OASST	69.7	37	76.0	72.28

6) https://huggingface.co/meta-llama/Llama-2-7b
7) https://huggingface.co/rinna/youri-7b
8) Stability-AI/lm-evaluation-harness に基づき評価した。
9) rinna/youri-7b-instruction

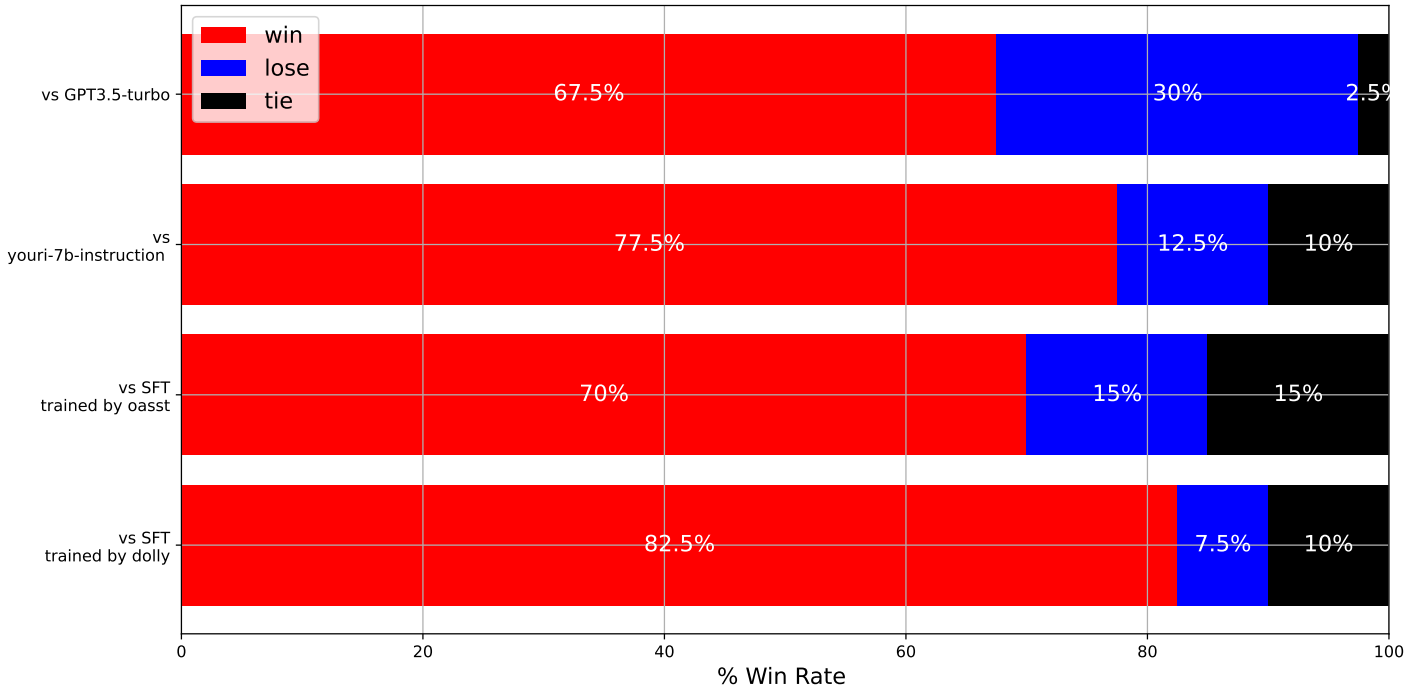

図 2　Rakuda Benchmark における houou の勝率

表 3 に ELYZA-tasks-100 の結果を示す。houou は同じパラメータ数である他の日本語 7B モデルである youri-7b-chat や japanese-stablelm-instruct-beta-7b、ELYZA-japanese-Llama-2-7b-instruct 、Xwin-LM-7B スコアを上回るスコア (2.63) を得た。また、houou よりもパラメータ数の多い、dolly と OASST によって学習された 13B の lm-jp-13b-instruct-full-dolly-oasst-v1.0 の性能も上回った。

表 3　ELYZA-tasks-100 の結果

Model	Score average
japanese-stablelm-instruct-beta-7b	1.28
llm-jp-13b-instruct-full-dolly-oasst-v1.0	1.86
youri-7b-chat	1.93
calm2-7b-chat	2.26
ELYZA-japanese-Llama-2-7b-instruct	2.3
Xwin-LM-7B-V0.1	2.32
houou-instruction-7b-v2	**2.63**

インストラクションデータの量の精度への影響を調査した。ここでは、これまでバージョンを分けてリリースされた 1003 件, 2903 件, 4802 件の ichikara-instruction データセットをそれぞれ用いて sft モデルを学習し、Rakuda Benchmark を用いて gpt-3.5-turbo-1106 と性能を比較した。表 4 に結果を示す。この結果より、インストラクションデータの量が増えると houou の勝率が向上することがわかる。また、2903 件から 4802 件に増やすことで gpt-3.5-turbo-1106 の性能を上回った。

表 4　sft データ数と勝率の関係

# of data	houou 勝率	gpt-3.5-turbo 勝率	引き分け
1003	27.5	57.5	15
2903	37.5	50	12.5
4803	67.5	30	2.5

3.3　Rakuda Benchmark における houou の勝因と敗因の分析

Rakuda Benchmark における houou と gpt-3.5-turbo-1106 の勝敗判断において、 gpt-4 によって出力された最終的な判断理由に該当する文章を用いて、houou の勝因と敗因の分析を行った。分析は、アノテータ 1 名が判断理由を以下の 4 つのカテゴリーに主観で分類することで実施された。

1. fluency (日本語の文法や語彙の正しさ、文章の円滑さ、よみやすさ)
2. accurate (情報の正確性)
3. detailed (詳細な情報、多く有用なデータを提供しているか、有用な情報を提供しているか)
4. relevance (質問に対する回答になっているか)

図 3 に houou の勝因の分布を示す。勝因の約 8 割は、detailed であった。fluency は 1 サンプルを除いて、語彙や文法の正しさでなく、読みやすさや文章の円滑さが評価されていた。また、accurate では、

gpt-3.5-turbo-1106 に勝利した 2 サンプルはどちらも地名などを列挙させる prompt (e.g., 北海道の主要な都市 5 つを挙げよ) であった。

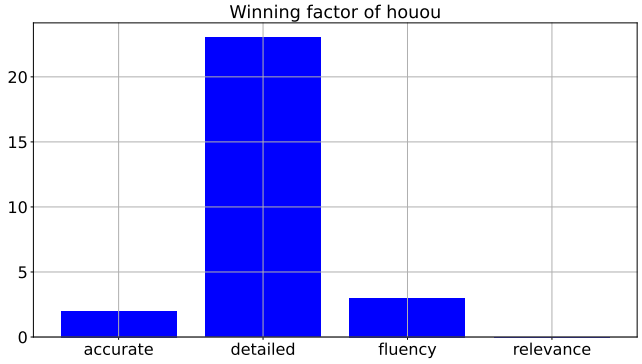

図 3 houou の勝因の分布

図 4 に houou の敗因の分布を示す。accurate に関しては、4 つの実際に houou が誤った情報を出力して敗北したことを確認した。detailed に関しては、いくつかのサンプルにおいて houou が、prompt に含まれる全ての条件を満たすような出力ができずに敗北したことが分かった。例えば、理由も合わせて説明せよ、といった prompt であるにもかかわらず、理由を出力できない、といった傾向が見られた。

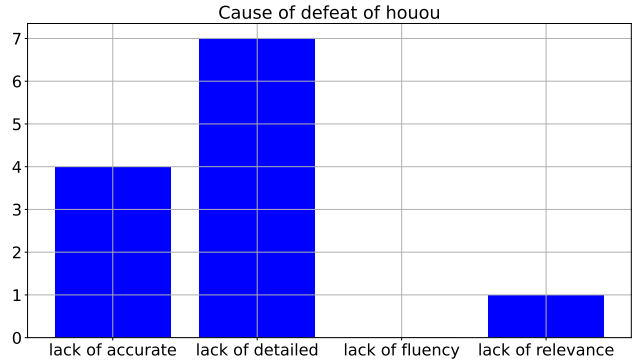

図 4 houou の敗因の分布

さらに、Rakuda Benckmark において、houou の勝利要因の約 8 割が detailed だったことに関して、houou と gpt-3.5-turbo-1106 の completion のトークン数を比較した。tokenizer には 、OpenAI の tokenizer である tiktoken を用いた。結果を図 5 に示す。図より、gpt-3.5-turbo-1106 に比べ、houou の completion は、長くなる傾向があることがわかる。なお、houou と gpt-3.5-turbo-1106 の平均トークン数は、それぞれ 513.125 と 371.375 であった。この結果から、gpt-3.5-turbo-1106 に比べて、houou は多くの情報を出力できたため、 detailed の観点で高く評価されたと考えられる。

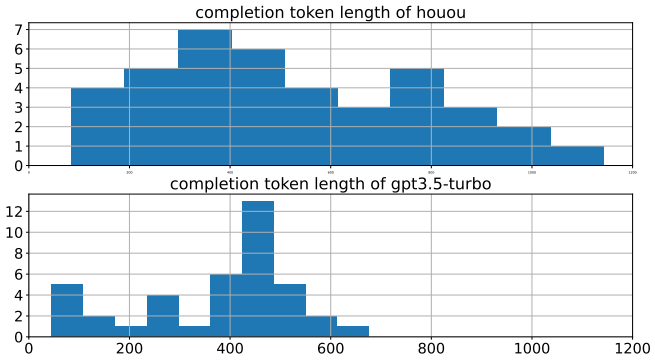

図 5 Rakuda Benchmark における houou と gpt-3.5-turbo-1106 の completion のトークン数比較

4 おわりに

本稿では、日本語によってフルスクラッチから作成されたインストラクションデータを用いてモデルを学習 することで、自動翻訳で作成されたデータによって学習されたモデルに比べて高い日本語生成能力が獲得できることを示した。実験では、Rakuda Benchmark を用いて、翻訳データによって学習されたモデルと日本語インストラクションデータで学習されたモデルの性能を gpt-4 によって比較した。その結果、Rakuda Benchmark を用いた評価では、gpt-3.5-turbo-1106 、日本語に自動翻訳した dolly と OASST によってそれぞれ学習された sft モデルに対する勝率はそれぞれ 67.5%, 82.5% と 70 %となり、日本語による高品質なインストラクションデータが LLM の日本語生成能力向上に重要であることが示された。

参考文献

[1] 関根聡, 安藤まや, 後藤美知子, 鈴木久美, 河原大輔, 井之上直也, 乾健太郎. 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 -. 言語処理学会第 30 回年次大会発表論文集.

[2] GenAI, Meta. 2023. Llama 2: Open foundation and fine-tuned chat models. CoRR, *arXiv preprint arXiv:2307.09288*, 2023.

[3] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342-8360.

[4] 関根聡, 小島淳嗣, 貞光九月, 北岸郁雄. LLM の出力結果に対する人間の評価分析と GPT4 の自動判断の比較分析言語処理学会第 30 回年次大会発表論文集.

A gpt-4 による自動評価に用いた prompt

```
あなたは、回答の質をチェックするための審判員です。

[質問]
<prompt>

[アシスタント1の回答の開始]
<LLM1のcompletion>
[アシスタント1の回答の終了]

[アシスタント2の回答の開始]
<LLM2のcompletion>
[アシスタント2の回答の終了]

上に表示されたユーザーの質問に対する 2つのAIアシスタントのパフォーマンスについて、あなたのフィードバックをお願いします。回答の有用性、関連性、正確性、詳細度、日本語能力を評価してください。まず、アシスタントの有用性、関連性、正確性、詳細度、日本語能力の評価を提供してください。評価の包括的な説明も提供してください。ユーザーは日本語しか話さないので日本語で書かれていない回答には低評価をつけてください。 偏見を避け、回答の提示された順序があなたの判断に影響を及ぼさないことに気をつけてください。両方の解答を慎重に評価した後、評価が高い方のアシスタントの解答を選び、アシスタント 1の回答であれば1を、アシスタント 2の回答であれば2を、そしてアシスタント1とアシスタント2の間から選べない場合は3を最後の行に出力してください
```

図 6 Rakuda Benchmark において gpt-4 に与えたプロンプト

```
あなたは採点者です。

問題, 正解例, 採点基準, 回答 が与えられます。

採点基準と正解例を参考にして、回答を1,2,3,4,5の5段階で採点し、数字のみを出力してください。

# 問題
<prompt>

# 正解例
<reference>

# 採点基準
基本的な採点基準
- 1点: 誤っている、指示に従えていない
- 2点: 誤っているが、方向性は合っている
- 3点: 部分的に誤っている、部分的に合っている
- 4点: 合っている
- 5点: 役に立つ

基本的な減点項目
- 不自然な日本語: -1点
- 部分的に事実と異なる内容を述べている: -1点
- 「倫理的に答えられません」のように過度に安全性を気にしてしまっている: 2点にする

問題固有の採点基準
<eval_aspect>

# 回答
<completion>
```

図 7 ELYZA-tasks-100 において gpt-4 に与えたプロンプト