pfgen-bench: 日本語事前学習モデルのための文章生成性能評価ベンチマーク

今城 健太郎 [1,2,*]　平野 正徳 [1,2,*]　鈴木 脩司 [1,2]　三上 裕明 [1,2]
[1] 株式会社 Preferred Networks　[2] 株式会社 Preferred Elements
imos@preferred.jp　research@mhirano.jp　{ssuzuki, mhiroaki}@preferred.jp

概要

本研究では、日本語事前学習モデルの文章生成性能を評価するためのベンチマークである pfgen-bench を提案する。本ベンチマークは Fluency(流暢さ)、Truthfulness(真実性)、Helpfulness(有用性) の 3 つの評価軸から構成される。まず、日本の小中高の学習指導要領を参考に、13 科目 50 問からなる、日本語圏特有の常識問題集を作成した。さらに、複数の LLM とルールベースのフィルタリング手法を用いて、高品質な参照回答群を構築した。その上で、さまざまなモデルの回答と参照回答群の近さをはかる 3 つの評価軸を設計し、生成結果の評価が可能なベンチマークを構築した。このベンチマークを使用した評価結果に基づくと、事前学習モデル間の性能差を明確に示し、LLM による従来の評価とも一致する点が確認された。構築したベンチマークは公開し、一般に使用可能である。

1　はじめに

近年、事前学習された大規模言語モデル (LLM) は、翻訳や文章生成、質問応答などの自然言語処理タスクにおいて高い性能を示している。特に、Llama[1, 2] や Mixtral[3] などのモデルは、英語圏を中心に広く活用され、その汎用的な能力が評価されている。一方で、これらのモデルは主に英語データを基に訓練されており、日本語のような非英語圏の言語に対して同様の性能を発揮するかどうかは十分に検証されていない。

従来の日本語生成能力の評価手法にはいくつかの問題点が存在する。多くのベンチマークが選択肢形式の問題を採用しているが、この形式ではモデルの日本語生成能力が不十分であっても、日本語の読解力や英語の理解を駆使して正答を得ることができるため、生成能力を適切に評価することが難しい。また、Japanese MT-bench[1)] のような日本語の生成能力を計測するベンチマークでも、英単語や外来語を多用した日本語として不自然な回答に対しても減点があまり行われず、日本語としての流暢さや文脈適合性を十分に評価できないという課題がある。また、生成能力の評価にあたっては、非常に性能の高い言語モデルを評価者として用いる必要もあり、計算コストの観点でも課題が存在する。

本研究の目的は、日本語事前学習モデルの生成能力を小規模なモデルから大規模なモデルまで様々な規模のモデルをより高い分解能で評価するためのベンチマーク手法を提案することである。本ベンチマークでは、Fluency(流暢さ)、Truthfulness(真実性)、Helpfulness(有用性) の 3 つの評価軸を採用し、日本語モデルがどの程度正確で流暢な応答を生成できるかを評価する。これにより、日本語生成モデルが日本語の文脈に適応し、適切な回答を生成できるかを客観的に評価することが可能となる。

本研究で構築した pfgen-bench は https://github.com/pfnet-research/pfgen-bench で公開している。

2　提案ベンチマーク: pfgen-bench

提案ベンチマークは、事前学習モデルにおける、日本語の生成能力および日本語圏特有の常識の習熟度を計測することを目的としている。

提案ベンチマークの構築にあたっては、大きく分けて以下の 3 つのステップが存在する。

- 問題・回答例構築
- 参照回答群構築
- モデルの評価値計算

* Equal contribution
Full paper: https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/1008

1) https://github.com/Stability-AI/FastChat/tree/jp-stable

本ベンチマークで構築した問題と参照回答群は所与のものとして扱い、評価対象のモデルの回答をすべての問題に対して生成し、その生成文と参照回答群間で複数の指標計算を行うことでベンチマークスコアを計測可能である。本章では、これらの 3 つのステップのすべてについて説明を行うが、実際にモデルの性能評価を行う際には、評価値計算のみを行えばよいことに注意されたい。

2.1 問題・回答例構築

問題の構築にあたって、日本語圏特有の常識として、日本の小中高の学習指導要領を参考に、下記の 13 科目から全 50 問の割り当てを決めた。

- 国語：4 問
- 社会（地歴、公民、環境[2]）：各 4 問
- 算数：4 問
- 理科（生物、化学、物理、地学）：各 4 問
- 芸術、文化：各 4 問
- 保健：4 問
- 情報：2 問

なお、数学は日本語圏特有のものが比較的少ないことから除外し、他の実技科目についても除外をしている。これらの分類に基づいて、比較的端的に回答できる問題を中心に問題を 50 問作成した。また、人手により、回答例も構築を行った。回答例の構築にあたっては、公用文などの文体に近い 100 文字程度の文を目安として構築を行った。詳細な問題・回答例については、公開レポジトリを参照されたい。

2.2 参照回答群構築

前節で問題と回答例を人手で構築したものの、どの質問も答え方は一つではなく、様々な回答があり得ることから、参照用に参照回答群を構築する。この参照回答群は、日本語圏における、比較的尤度の高い回答を多く集めることで、回答分布を定義することを目的としている。

多くの回答を多くの日本語圏の日本語話者により人手で作成することが理想ではあるが、ここでは日本語特化かつ高性能な LLM で代替することとした。

日本語特化であり、比較的大きいモデルであることという基準に従いかつモデルファミリーのバランスも考慮し、開発時点で利用可能であった、stockmark-100b[3]、PLaMo-100b、Swallow-MX-8x7b-NVE-v0.1[4] を採用した。

参照回答群を構築するにあたって、複数のステップで回答群を作成した。各ステップは以下のとおりである。

- すべてのモデルで many-shots による各問 100 万回答を作成
- ルールベースでのハルシネーションの除去
- 頻度が極端に低い回答の除去を通じた稀なハルシネーションの除去
- 回答長と代表性を考慮に入れた各問 1000 回答への絞り込み

これらのフェーズは、代表的でかつ正確性の高い回答を幅広くとるための設計となっており、テイル事象を除去することに重きを置いている。

2.3 モデルの評価値計算

前節で構築した参照回答群を用いて、評価対象のモデルのベンチマークスコアを計算する。

生成した評価対象のモデルの回答に対して、前節で構築した参照回答群を用いて以下の 3 つの指標を計算する。

- Fluency: 文字レベルの 10-gram の出現割合の内積
- Truthfulness: 出現頻度 0.5%以上の文字レベルの 3-gram の割合
- Helpfulness: 手動で作成したルールによる評価

Fluency は参照回答群の平均評価スコアが 1.0 になるように各問ごとにスケールを行う。そのうえで、3 つのスコアの平均を取ることで最終的なベンチマークスコアとする。

以下では、これらの計算について詳しく説明する。

2.3.1 Fluency

Fluency とは、問題が与えられたときに、自然な日本語で回答できているかどうかを確認する指標であり、本研究で新たに定義する。

ここで、Fluency は、文字レベルの 1-gram から 10-gram のウィンドウに区切って文章を見たときに、参照回答群において尤度の高い文章になっているかを確認する。そこで、評価対象の回答文からすべ

2） 厳密には環境は学習指導要領上の科目として存在しないが、地歴・公民に分類できない問題が多いことから別の区分として設けた。

3） https://huggingface.co/stockmark/stockmark-100b

4） https://huggingface.co/tokyotech-llm/Swallow-MX-8x7b-NVE-v0.1

ての文字レベルの 1-gram から 10-gram を抜き出し、それらの参照回答群における出現頻度を計算し、足しこんでいく。回答は 100 文字程度を想定しているが、出現頻度を足しこんでいく方式であると、文章長が長い場合にスコア上有利になってしまうため、100 文字から減衰を開始し、150 文字でスコアが 0 になる線形減衰によるスコアディスカウントを導入した。

2.3.2 Truthfulness

Truthfulness とは、正確な情報をどの程度答えられているかどうかに関する指標として設計をした。Truthfulness では、極端な出現頻度の低さの閾値として 3-gram における 0.5%を採用し、0.5%以上の出現頻度の 3-gram の確率を計算することとした。ただし、ハルシネーションの有無を示す指標であることを鑑みて、句読点や鍵かっこなどの記号は頻度計算の対象から除外することした。さらに、Fluency と同様に 100 文字を超えた場合のディスカウントについても設定する。ただし、Truthfulness においては、100 文字を超えた場合については、それ以降をカットしてしまうことを認めることとする。つまり、100 文字を超えた場合については、Truthfulness スコアが最も高い文字数でカットをした場合のスコアを採用する。小さいモデルは文章を適切な文字数で出力することが難しい傾向があるが、ディスカウントを含めた最大値をとることで小さいモデルの出力のランダム性を吸収し、スコアを安定させる効果があるため、このような設計とした。

また、Fluency と異なり、出力長に依存しない評価値となっている。

2.3.3 Helpfulness

Helpfulness は、評価対象の文章がどれだけ必要情報を適切に含んでいるかを評価する、人手で構築したルールベースの指標である。各問ごとに、含むべき重要単語を定義し、それがどれだけ含まれているかを割合を計測する。例えば、「つるかめ算について教えて。」という質問に対して、「合計」、「それぞれ」、「各々」の 3 つのうち 1 つの単語と「算数」の計 2 単語を必須重要単語として設定している。このように、and/or を用いて必須単語の条件を定義している。なお、原則として、これらの必須重要単語は等価に扱うこととしているが、重みづけを変えるケースも存在する。ただし、100 文字を超えている場合については、100 文字までで打ち切った場合の評価値と、$I > 100$ となる I 文字目までで打ち切った場合の評価値に $\left(1 - \frac{\max(I-100,0)}{50}\right)$ を乗じた値をすべて計算し、最も高い値を Helpfulness score として採用する。これにより、文字数オーバー部分のディスカウントしてでも加算したほうがよい単語を考慮に入れることができるため、安定的な評価ができる。

3 実験

3.1 実験 1: 提案ベンチマークの特性分析

まず、提案ベンチマークが適切に機能しているかを確認するために、その特性を検証する。図 1–3 に、Fluency, Truthfulness, Helpfulness の関係性についてプロットした。

この図によると、Truthfulness、Fluency、Helpfulness の順に容易なタスクであることがわかる。

一方で、それぞれのモデル間での順序性は、多少の揺らぎがあるものの、どの指標においても概ね一貫しているようにも見える。また、Fluency、Truthfulness に関しては、現状ですでに 1.0 にちかいスコアを達成しているモデルが存在しており、限界が近い一方で、Helpfulness はまだまだ 1.0 に近いスコアを達成できていないものが多い。

3.2 実験 2: LLM-as-a-judge との比較

続いて、提案ベンチマークと GPT-4o による LLM-as-a-judge [4] の比較を行う。

ここでは、LLM-as-a-judge に GPT-4o の API の利用金額が大幅にかかることから、前節のモデルからさらに主要なものに絞った 50 個のモデルに対して比較を行った。

図 4 にその結果を示す。結果として、0.9896 というとても高い相関性を示しており、ベンチマークとしては充分に機能していると言える。今回の実験では、提案手法におけるスコアが 1.0 を上回る領域 (GPT-4o による LLM-as-a-judge で 9 を上回る領域) や、0.2 を下回る領域 (GPT-4o による LLM-as-a-judge で 0 点付近の領域) でのスコアの挙動を観測できていないものの、現状の実用のレベルでは既存手法である GPT-4o による評価と同等程度に分解能を持っており、充分であると言えるのではないだろうか。

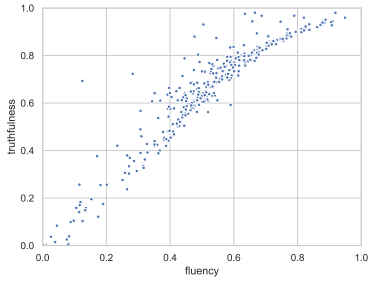

図 1: Fluency と Truthfulness

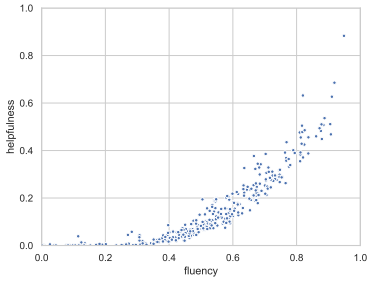

図 2: Fluency と Helpfulness

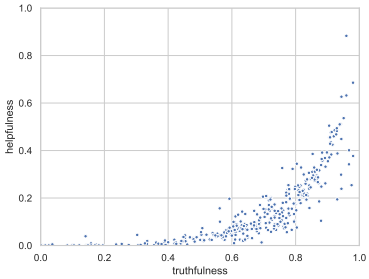

図 3: Truthfulness と Helpfulness

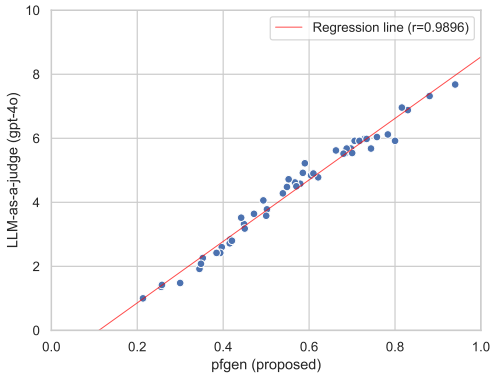

図 4: 提案ベンチマークと GPT-4o による LLM=as-a-judge のスコアの比較

4 考察

実験の結果、提案ベンチマークは一定のレベルで有効なベンチマークとして一定レベルで機能していることを確認できた。特に LLM-as-a-judge との相関性の確認は非常に高い相関を示しており、LLM-as-a-judge という計算コストが大きい手法を代替できる可能性が示唆されていると考えられる。

ここでは、なぜこのベンチマークが有効であるかどうかという点について議論する。

まず、Helpfulness に関しては、人手でルールを構築しているため、常識をどの程度学習できているのかどうかを計る指標となっていると言える。一方で、Fluency と Truthfulness は、参照回答群との乖離度を測る指標になっていると解釈できると考えられる。

その場合、Fluency と Truthfulness が十分に機能するためには、参照回答群の性能に依存するようにも見えてしまうが、今回の結果を分析すると、参照回答群を作成する際に使った LLM や参照回答群自体のベンチマークスコアを上回るスコアを出しているモデルも存在した。具体的には、anthropic/claude-3-5-sonnet-20240620 と openai/gpt-4o が、参照回答群をモデルの回答としてベンチマークを計測した場合よりも良い性能を発揮していた。個別の指標で見ても、anthropic/claude-3-5-sonnet-20240620 は Fluency で、openai/gpt-4o は Truthfulness で参照回答群のスコアを上回っていた。つまり、今回の提案手法は、参照回答群の生成結果の性能以上のものを分析できる可能性を示唆している。

5 まとめ

本研究では、LLM の対話における生成の良さを計るベンチマークとして、pfgen-bench を提案した。このベンチマークは、日本語圏特有の常識を問う 13 科目 50 問の問題からなり、評価対象の LLM がこれらの問題に対する回答を 100 文字程度で作成したのちに、スコアリングが行われる。スコアリングにあたっては、事前に構築された参照回答群を用いて、様々な LLM に対するベンチマークに対する Fluency、Truthfulness の 2 指標を n-gram に着目して計算し、さらに、人手で作成したルールベースのアルゴリズムからなる Helpfulness を組み合わせて計算される。参照回答群の構築にあたっては、主に 3 つの日本語の大規模言語モデルを用いて構築を行った。実験の結果、提案ベンチマークは、GPT-4o を用いた LLM-as-a-judge とほぼ同様の評価が行えることが確認でき、既存の Japanese MT-bench などのベンチマークとも高い相関性を持つことが確認できた。

Declarations

著者らは、pfnet/plamo-100b の開発元である、株式会社 Preferred Networks/Elements に所属しているが、本研究におけるモデル選定などにおいては、客観的根拠を以って公平な評価を行うように努めている。また、透明性の確保のために、ベンチマークの計測コードを公開している。

謝辞

データセットを作成するにあたり、片岡俊基氏に回答の正しさについて校閲を行っていただきました。

参考文献

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. **arXiv**, 2023. https://arxiv.org/abs/2302.13971.

[2] Hugo Touvron, Louis Martin, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. **arXiv**, 2023. https://arxiv.org/abs/2307.09288v2.

[3] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. **arXiv preprint arXiv:2310.06825**, 2023.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46595–46623. Curran Associates, Inc., 2023.