

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

平野 正徳^{†,††} 鈴木 雅弘[†] 坂地 泰紀[†]

[†] 東京大学大学院工学系研究科 〒113-8656 東京都文京区本郷 7-3-1

E-mail: [†]research@mhirano.jp, b2019msuzuki@socsim.org, sakaji@sys.t.u-tokyo.ac.jp

あらまし 本研究では、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を構築した。本データセットは、約 840 万件のデータを含んでおり、翻訳タスクや、知識タスクなど、様々なタスクをチャット形式で含んだものとなっている。構築したデータセットの有効性を確認するために、既存の大規模言語モデルをチューニングし、性能向上を定性的に確認し、日本語における大規模言語モデルや言語資源の構築における課題を明らかにした。

キーワード 大規模言語モデル, データセット, 日本語, チャット

llm-japanese-dataset v0: Construction of Japanese Chat Dataset for Large Language Models

Masanori HIRANO^{†,††}, Masahiro SUZUKI[†], and Hiroki SAKAJI[†]

[†] School of Engineering, The University of Tokyo 7-3-1 Hongo, Bunkyo, Tokyo, 113-8656 Japan

E-mail: [†]research@mhirano.jp, b2019msuzuki@socsim.org, sakaji@sys.t.u-tokyo.ac.jp

Abstract This study constructed a Japanese chat dataset for large language models. The dataset contains approximately 8.4 million records and includes various tasks in chat format, such as translation and knowledge tasks. To confirm the benefits of our constructed dataset, we tuned an existing large language model and confirmed its performance qualitatively. Those results revealed challenges in building large language models and language resources for them in Japanese.

Key words Large Language Model, Dataset, Japanese, Chat

1. はじめに

大規模言語モデル (LLM) は、近年、著しい性能向上と汎化が進んでいる。特に、Transformer [1] ベースの大規模言語モデルである、BERT [2] や、GPT シリーズ [3]~[5] は、事前学習由来の高い性能を発揮している。さらにそれらを発展させたモデルとして、ChatGPT [6] や GPT-4 [7] などが有名であり、その性能は顕著である。また、Bard [8] や LLaMA [9], Dolly [10], BLOOM [11], Vicuna [12] などといったモデルも登場している。

言語モデルが乱立している一方で、多様な指示の入力への対応にはまだまだ難しさがある。例えば、オープンソースのモデルである LLaMA [9]^(注2) の場合、その応答の不完全性から、Alpaca [13] が登場している。

しかしながら、そうしたモデルの学習に使用されているデー

タセットも、英語以外の言語への対応は不完全である。例えば、Alpaca [13] の場合、追加のデータセットは英語のみで提供されているため、英語以外の言語で Alpaca のメリットを享受することは難しい。また、Alpaca の元となっている、LLaMA [9] であっても、英語以外の言語で書かれた一部の入力に対しては、英語で回答してしまったり、指示の意味を理解できていないケースも多々観測される。

これらの背景を考えると、モデル学習において、英語以外の言語のデータセットを増強することの必要性が考えられる。しかしながら、現状の学習モデルの開発競争を考えると、特定のモデルだけに固執することは良い戦略であるとは考えられず、数ヶ月後、あるいは 1-2 ヶ月後には、最新の学習モデルよりも性能の良いモデルが出てくることは容易に考えられる。そのため、学習モデル自体よりも、その学習を支えるデータセットの増強は、新しい学習モデルが出てきた場合でも使用できる可能性があり、有用性が高いとともに、新しい学習モデルを英語以外の言語に対応させることのハードルを下げるとも言える。

そこで、本研究においては、日本語における、約 840 万データ

(注^{††}): 責任著者

(注2): 厳密には、当初はオープンソースではなかったが、一定のライセンス下で提供されるようになった

を含むチャットデータセットを新たに構築し、そのデータセットの効用について簡単に示す。このデータセットは、オープンソースとして、公開を行っており、また、学習済みモデルも公開している。

- データセット: <https://huggingface.co/datasets/izumi-lab/llm-japanese-dataset>
- 学習済みモデル (LLaMA 1 epoch): <https://huggingface.co/izumi-lab/llama-13b-japanese-lora-v0-1ep>

詳細については、後述する。また、今後の課題として、更なるデータセット拡張とモデル学習を進める予定である。

2. データセットの構築 (izumi-lab/llm-japanese-dataset v0)

本研究では、日本語のチャットデータセットを作成した。このデータセット^(注3)には、8,393,726 個のデータを含んでいる。ここでは、その詳細について作成方法とともに説明する。

なお、データセットに関しては、以下の URL で公開している。

- Hugging Face: <https://huggingface.co/datasets/izumi-lab/llm-japanese-dataset>
- GitHub: <https://github.com/masanorihirano/llm-japanese-dataset>

また、学習時のチャットデータの使用フォーマットは下記のとおりであり、後述のデータセットの説明時には、冒頭部と改行を一部省略して示す。

学習時のチャットデータの使用フォーマット

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:

{Instruction}

Input:

{Input}

Response:

{Response}

2.1 翻訳タスク

多くの LLM においては、英語におけるパフォーマンスが良い一方で、日本語などの英語以外の言語に対する対応は未熟である場合が多い。この場合、英語においては、知識を十分に学習できている一方で、そこに対して、日本語などの言語をリンクさせるような学習ができれば、英語で学習されている知識を多言語に持つていくことができるのではないかと考えた。そのため、日本語から英語、英語から日本語の翻訳タスクを設定した。

(注3): 本論文では、v0, 特に、v0.1.0 について述べる。今後、アップデートを行う可能性がある。

翻訳タスクの例 1

Instruction:

次の日本語を英語に翻訳してください。

Input:

彼は忙しい生活の中で家族と会うことがない。

Response:

He doesn't see his family in his busy life.

翻訳タスクの例 2

Instruction:

次の英語の論文のタイトルを日本語に翻訳してください。

Input:

Odd viscosity in chiral active fluids

Response:

キラル活性流体の odd 粘度

なお、翻訳データは、下記のデータセットを加工することで獲得した。

- Coursera Corpus Mining and Multistage Fine-Tuning for Improving Lectures Translation [14]^(注4): 翻訳データをそのまま日英双方のデータとして活用 (例 1 の日英双方向)。ライセンスは、Apache License 2.0。作成されたデータセットサイズは 106,332。
- ParaNatCom — Parallel English-Japanese abstract corpus made from Nature Communications articles [15]: 論文に関する翻訳データのうち、タイトルとアブストラクトの対訳を例 2 のフォーマットで日英双方向で作成。ライセンスは CC BY 4.0。作成されたデータセットサイズは 1,732。
- Tab-delimited Bilingual Sentence Pairs^(注5): 日英の対訳データを例 1 のフォーマットで双方向で使用。ライセンスは CC BY。作成されたデータセットサイズは 202,398。
- Asian Language Treebank (ALT) Project [16]^(注6): 日英の対訳データのみならず、例 1 のフォーマットで双方向で使用。ライセンスは CC BY 4.0。作成されたデータセットサイズは 39,816。
- Tanaka Corpus^(注7): 日英の対訳データを例 1 のフォーマットで双方向で使用。ライセンスは CC BY。作成されたデータセットサイズは 298,026。
- Japanese-English Subtitle Corpus [17]^(注8): 日英の対訳データを例 1 のフォーマットで双方向で使用。ライセンスは CC BY-SA 4.0。作成されたデータセットサイズは 5,396,204。
- 日本語 WordNet [18]^(注9): 単語の説明と、使用例について、元の WordNet [19] における英語と対訳データセットとして

(注4): <https://github.com/shyyhs/CourseraParallelCorpusMining> より取得可能

(注5): <http://www.manythings.org/anki/> より取得可能

(注6): <https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/> より取得可能

(注7): http://www.edrdg.org/wiki/index.php/Tanaka_Corpus より取得可能

(注8): https://nlp.stanford.edu/projects/jesc/index_ja.html より取得可能

(注9): <https://bond-lab.github.io/wnja/> より取得可能

用いて、例1のフォーマットで双方向で使用。ライセンスはBSD like な独自ライセンス。作成されたデータセットサイズは367,936。

- やさしい日本語コーパス：「SNOW T15: やさしい日本語コーパス」[20] および「SNOW T23: やさしい日本語拡張コーパス」[21] に含まれている、日本語と英語の対訳を例1のフォーマットで使用した。ライセンスはCC BY 4.0。作成したデータセットサイズは168,600。

これらを全て合わせると、6,581,044個のデータとなる。これは、データセットのおよそ8割を占めるサイズであるが、前述のモチベーションから、そのまま採用することとした。

2.2 知識系タスク

続いて、日本語での知識を適切に学習させるために、知識系のタスクを学習データセットに取り込んだ。

データセット作成に使用したデータは以下のとおりであり、作成したデータセット例も続けて示す。

- Wikipedia^(注10)：Wikipediaのうち、タイトルと最初の段落を使用することで、単語を説明するタスクとして作成。ライセンスはCC BY-SA 3.0。作成したデータセットサイズは1,347,381。

- 日本語 WordNet [18]^(注11)：単語と単語の定義部分ももちいて、単語の説明を要求するチャットとして作成。ライセンスはBSD like な独自ライセンス。作成されたデータセットサイズは173,085。

- AI王 公式配布データセット Version 2.0^(注12)：クイズ形式のデータセットのため、そのままチャットデータセットとして組み込んだ。ライセンスはCC BY-SA 4.0。作成されたデータセットサイズは23,335。

- Japanese Movie Recommendation Dialogue (JMRD) [22]^(注13)：映画に関するメタデータの部分のみを取り出し、映画の説明とあらすじを示すチャットデータセットとして作成。ライセンスはCC BY-SA 4.0。データセットサイズは260。

- JCommonsenseQA [23]^(注14)：Japanese General Language Understanding Evaluation (JGLUE) に含まれるタスクの一部で、常識を問うQ&Aデータセットをそのまま取り込んだ。ライセンスはCC BY-SA 4.0。作成したデータセットサイズは10,058。

- 青空文庫^(注15)：青空文庫で公開されている文学作品等のうち、著作権が切れているものだけを対象^(注16)に、文章の冒頭部を問う質問と、文章冒頭部から作者とタイトルを問う質問を作成。作成したデータセットサイズは33,758。

これらを合計すると、1,587,877個のデータセットとなる。

(注10)：<https://dumps.wikimedia.org/> より取得可能。20230401 jawiki バージョンを使用

(注11)：<https://bond-lab.github.io/wnja/> より取得可能

(注12)：<https://sites.google.com/view/project-ai0/dataset> より取得可能

(注13)：<https://github.com/ku-nlp/JMRD> より取得可能

(注14)：<https://github.com/yahoojapan/JGLUE> より取得可能

(注15)：<https://www.aozora.gr.jp/>

(注16)：今回は<https://github.com/levellevel/AozoraTxt> を使用した。

知識系タスクの例: Wikipedia

Instruction:

入力されたワードを説明してください。

Input:

東京大学

Response:

東京大学(とうきょうだいがく)は、東京都文京区に本部を置く日本の国立大学である。略称は東大(とうだい)。

知識系タスクの例: 日本語 Wordnet

Instruction:

「可能」の意味を教えてください。

Response:

何かをするのに必要な方法、技能またはノウハウまたは権利を持っている

知識系タスクの例: AI王

Instruction:

織田信長、豊臣秀吉、徳川家康という3人の戦国武将の性格を表現するのに用いられる鳥は何でしょう?

Response:

ホトトギス

知識系タスクの例: JMRD (一部省略)

Instruction:

となりのトトロについて教えてください。

Response:

となりのトトロは、1988年に製作された映画で、監督は宮崎駿です。日高のり子、坂本千夏らが出演しています。あらすじは以下の通りです。

【あらすじ】

小学生のサツキと妹のメイは、母の療養のために父と一緒に初夏の頃の農村へ引っ越してくる。引越...

知識系タスクの例: JCommonsenseQA

Instruction:

主に子ども向けのもので、イラストのついた物語が書かれているものはどれ?

Input:

世界、写真集、絵本、論文、図鑑

Response:

絵本

知識系タスクの例: 青空文庫 (1)

Instruction:

ツルゲーネフの作品『はつ恋』の冒頭を教えてください。

Response:

「客はもうとうに散ってしまった。時計が零時半《れいじ

はん》を打った。部屋の中に残ったのは、主人と、セルゲイ・ニコラーエヴィチと、ヴラジーミル・ペトローヴィチだけである。主人は呼鈴《よびりん》を鳴らして、」です。

知識系タスクの例: 青空文庫 (2)

Instruction:

「二人の若い紳士《しんし》が、すっかりイギリスの兵隊のかたちをして、ぴかぴかする鉄砲《てっぽう》をかついで、白熊《しろくま》のような犬を二 | 疋《ひき》つれて、だいぶ山奥《やまおく》の、木の葉のかさかさし」で始まる文学作品の作者とタイトルを教えてください。

Response:

宮沢賢治の『注文の多い料理店』です。

2.3 その他のタスク

2.3.1 論文タイトルの作成とアブストラクトの想像

前出の ParaNatCom — Parallel English-Japanese abstract corpus made from Nature Communications articles [15] を用いて、論文に関連するタスクを作成した。詳細は下記のとおりである。なお、ライセンスは CC BY 4.0。作成されたデータセットサイズは 1,732。

タスクの例 1 (一部省略)

Instruction:

論文のアブストラクトからタイトルを作ってください。

Input:

超薄型ナノ構造、特に厚さが原子レベルのものは、例外的な光-物質相互作用により、通常他に類を見ない光学特性を示す。ここで我々は、原子レベルサイズ (厚さ 0.5 nm) と調整可能なリング内リング構造を有する硫黄ドープ型酸化モリブデン・ナノリングの簡単な合成戦略を報告する。これらの原...

Response:

フルスペクトル吸収と光応答特性を有する原子レベル酸化モリブデン・ナノリング

タスクの例 2 (一部省略)

Instruction:

論文のタイトルからアブストラクトを想像して書いてください。

Input:

表面プラズモンによって誘起される長波長光子の直接的な検出

Response:

ミリメートル及びテラヘルツ波の光検出器は、幅広い応用範囲のために長い間大きな関心が持たれているが、依然として検出性能における課題に直面している。ここで、我々は、アンテナ支援のサブ波長オーミック金属-半導体-金属 (OMSM) 構造における局所化された表面プラズモン・

ポラリトン (SPP) によって...

2.3.2 ニュースタイトルの生成

日本語版 Wikinews から記事を取得^(注17)し、記事本文とタイトルに分離し、記事本文からタイトルを生成するタスクとして設定。ライセンスは CC BY 2.5。作成したデータセットサイズは 4,265。

タスクの例 (一部省略)

Instruction:

次のニュース記事にタイトルをつけてください。

Input:

岩波書店は約 10 年ぶりの大改訂となる「広辞苑 第六版」を 1 月 11 日に発売した。J-CAST によると広辞苑は 1955 年に発売された初版も含めると累計 1100 万部を誇り、ベストセラー...

Response:

10 年ぶりに広辞苑が大改訂一第六版発売

2.3.3 抜き出しによる回答

知識と翻訳以外にも、与えられた文脈の中から回答を抜き出し形のタスクも採用した。元データとして、JSQuAD [23]^(注18)を使用した。JSQuAD も、Japanese General Language Understanding Evaluation (JGLUE) に含まれるタスクの一部で、そのまま取り込んだ。ライセンスは CC BY-SA 4.0。作成したデータセットサイズは 67,491。

タスクの例 (一部省略)

Instruction:

天然痘の大流行が起こったのはいつか。

Input:

聖武天皇：天平 9 年 (737 年) に天然痘の大流行が起こり、藤原四兄弟を始めとする政府高官のほとんどが病死するという惨事に見舞われ、急遽、長屋王の実弟である鈴鹿王を知太政官事に任じて辛うじて政府の体裁を整える。さらに、天平 12 年 (740 年) には藤原広嗣の乱が起こっている。...

Response:

天平 9 年 (737 年)

2.3.4 やさしい日本語データセット

通常の日本語をやさしい日本語に書き換えるタスクを取り込んだ。元データとしては、SNOW T15: やさしい日本語コーパス [20] および SNOW T23: やさしい日本語拡張コーパス [21] を加工して使用した。ライセンスは CC BY 4.0。作成したデータセットサイズは 84,300。

(注17) : <https://ja.wikinews.org/wiki/> より、クローリングにて取得

(注18) : <https://github.com/yahoojapan/JGLUE> より取得可能

タスクの例

Instruction:

以下の日本語をやさしい日本語に言い換えてください。

Input:

ビルには冒険心がまったくない。

Response:

ビルには危険なことをする気持ちが全くない。

2.4 既存のチャットデータセットの日本語翻訳版

そのほかに、すでに公開されている既存のチャットデータセットの日本語翻訳版も取り込んだ。取り込んだデータセットは以下の通りである。

- Japanese-Alpaca-LoRA^(注19): Alpaca [13] のデータセットを日本語に翻訳したもの。ライセンスは Apache License 2.0。データセットサイズは 52,002。

- databricks-dolly-15k-ja^(注20): Dolly [10] の学習に使用されたデータセットの日本語翻訳版。ライセンスは CC BY-SA 3.0。データセットサイズは 15,015。

3. データセットを用いたモデル構築と評価

LLM のスクラッチからの構築には大量の計算資源を必要とする。さらに、パラメータ数の多い LLM モデルは、事前学習に限らずファインチューニングにおいても GPU リソースを必要とする。本研究では、精度を大きく下げないまま大規模モデルをファインチューニングするための手法として LoRA [24] を用いる。LoRA では、LLM のパラメータの差分のみを小規模パラメータで更新する。実験の際に使用した主なパラメータを以下に示す。

- ベースモデル: LLaMA 13B [9]
- 学習率: 3e-4
- 入力長: 256 トークン
- バッチサイズ: 130
- エポック数: 1
- LoRA の r : 4
- LoRA の α : 16
- LoRA の Dropout 率: 0.05
- LoRA のチューニングパラメータ: q-proj, v-proj

実装には PEFT [25], DeepSpeed ZeRO 2 [26] を用いた。なお、このモデルは、<https://huggingface.co/izumi-lab/llama-13b-japanese-lora-v0-1ep> にて公開している。

なお、実験での再現性を高めるために、プロンプトの生成の温度パラメータは 0.0 とした。以下に、いくつか、性能を比較するために行った定性的な比較を示す。

(注19): <https://github.com/masa3141/japanese-alpaca-lora> より取得可能

(注20): <https://github.com/kunishou/databricks-dolly-15k-ja> より取得可能

応答の比較 (1)

入力:

あなたは誰ですか？

出力 (LLaMA+LoRA):

私は人間です。

出力 (LLaMA):

このアプリケーションは、Google のアイテムを使用しています。

応答の比較 (2)

入力:

東京から大阪の行き方を教えてください。

出力 (LLaMA+LoRA):

東京から大阪の行き方は、東京駅から東京駅から大阪駅までの列車で行きます。

出力 (LLaMA):

東京から大阪の行き方を教えてください。

応答の比較 (3)

入力:

東京大学はどこにありますか？

出力 (LLaMA+LoRA):

東京大学は日本の東京都にあります。

出力 (LLaMA):

東京大学は東京都文京区文京 1-1-1 です。

(著者注: 東京大学の住所は文京区本郷 7-3-1 です。)

応答の比較 (4)

入力:

電話が鳴りました。どうするべきですか？

出力 (LLaMA+LoRA):

電話を切ることです。

出力 (LLaMA):

電話が鳴りました。電話を受け取ると、電話を受け取った方が電話を受け取った方に電話を受け取るべきです。

応答の比較 (5)

入力:

京都の三大祭りと言えば何？

出力 (LLaMA+LoRA):

京都の三大祭りは、春の祭り、夏の祭り、秋の祭りです。

出力 (LLaMA):

京都の三大祭りと言えば何？

京都の三大祭りと言えば何？

京都の三大祭りと言えば何？

4. 考 察

本研究においては、izumi-lab/llm-japanese-dataset v0 という、

データセットの構築を実施した。このデータセットの構築においては、日英の翻訳タスクを中心に、約 840 万件のデータセットを構築した。この目的は、英語を中心に構築されている LLM に対して、追加で LoRA による学習を行うことで、英語の言語処理能力と日本語での入出力を接続できるような学習を行い、日本語における性能向上をさせることであった。

実際の学習後のプロンプトを確認すると、若干ではあるが、性能が向上しているようにも見える。今回の学習においては、LLaMA を 1 epoch のみ LoRA で学習させただけである。そのため、その学習による差分は少ない可能性があり、さらなる epoch 数の増加は今後の課題であるが、結果を見る限りでは、今回のデータセット構築は、有効である可能性がある。たとえば、応答の比較 (4), (5) で示したものは、答えは正しくないものの、LoRA を行うことにより、出力の日本語はより適切な出力となっている。また、応答の比較 (2), (3) についても、応答としてはたどたどしいものの、正しい応答をするようには変化している。

しかしながら、今回の定性的な比較だけでは、適切な比較ではなく、比較手法の確立も今後の課題であろう。より、正確な評価を行っていくためには、日本語に特化した評価データセットの構築が必要であり、Language Model Evaluation Harness [27] のような評価スキームの日本語版の必要性が高いのではないだろうか。

5. ま と め

本研究では、大規模言語モデルのチャット応答に対する学習のための日本語チャットデータセットの構築を行った。主に、日本語以外の言語で作成された大規模言語モデルに対して、チャット応答のチューニングを行うことが可能であり、実際に、このデータセットを用いて学習を行った場合の性能を定性的に確認した。データセットは、公開をしており、広く利用可能である。また、今回のデータセットの性能確認のためのモデル構築を通じて、日本語における大規模言語モデルの構築における課題も明らかになった。

謝 辞

本研究は JSPS 科研費 JP21K12010 および JST さきがけ JP-MJPR2267 の助成を受けたものです。

文 献

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol.30, pp.5999–6009, 2017.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp.4171–4186, Association for Computational Linguistics, 2019.
- [3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I.

- Sutskever, "Language Models are Unsupervised Multitask Learners," 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol.33, pp.1877–1901, 2020.
- [6] OpenAI, "ChatGPT," <https://openai.com/blog/chatgpt/>, 2023.
- [7] OpenAI, "GPT-4 Technical Report," 2023. <https://arxiv.org/abs/2303.08774>
- [8] Google, "Bard," <https://bard.google.com/>, 2023.
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., "LLaMA: Open and Efficient Foundation Language Models," 2023. <https://arxiv.org/abs/2302.13971>
- [10] Databricks, "Dolly," <https://github.com/databricks/dolly>, 2023.
- [11] T.L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A.S. Luccioni, F. Yvon, M. Gallé, et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," 2022. <https://arxiv.org/abs/2211.05100>
- [12] Vicuna, "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality," <https://vicuna.lmsys.org/>, 2023.
- [13] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T.B. Hashimoto, "Stanford Alpaca: An Instruction-following LLaMA model," https://github.com/tatsu-lab/stanford_alpaca, 2023.
- [14] H. Song, R. Dabre, A. Fujita, and S. Kurohashi, "Coursera Corpus Mining and Multistage Fine-Tuning for Improving Lectures Translation," *Proceedings of the 12th Language Resources and Evaluation Conference*, pp.3640–3649, May 2020.
- [15] M. Utiyama, "ParaNatCom — Parallel English-Japanese abstract corpus made from Nature Communications articles," 2019. <https://www2.nict.go.jp/astrec-att/member/mutyama/paranacom/>
- [16] H. Riza, M. Purwoadi, T. Uliniansyah, A.A. Ti, S.M. Aljunied, L.C. Mai, V.T. Thang, N.P. Thai, V. Chea, S. Sam, et al., "Introduction of the asian language treebank," 2016 *Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques*, pp.1–6, 2016.
- [17] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz, "JESC: Japanese-English Subtitle Corpus," *Language Resources and Evaluation Conference*, pp.1133–1137, 2018.
- [18] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki, "Development of the Japanese WordNet," *Language Resources and Evaluation*, pp.2420–2423, 2008.
- [19] C. Fellbaum, "WordNet," *Theory and applications of ontology: computer applications*, pp.231–243, Springer, 2010.
- [20] T. Maruyama and K. Yamamoto, "Simplified Corpus with Core Vocabulary," *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp.1153–1160, 2018.
- [21] A. Katsuta and K. Yamamoto, "Crowdsourced Corpus of Sentence Simplification with Core Vocabulary," *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp.461–466, 2018.
- [22] T. Kodama, R. Tanaka, and S. Kurohashi, "Construction of Hierarchical Structured Knowledge-based Recommendation Dialogue Dataset and Dialogue System," *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pp.83–92, 2022.
- [23] K. Kurihara, D. Kawahara, and T. Shibata, "JGLUE: Japanese General Language Understanding Evaluation," *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp.2957–2966, 2022.

- [24] E.J. Hu, yelongshen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” International Conference on Learning Representations, pp.1–13, 2022. <https://arxiv.org/abs/2106.09685>
- [25] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, and S. Paul, “PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods,” <https://github.com/huggingface/peft>, 2022.
- [26] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, “ZeRO: Memory Optimizations toward Training Trillion Parameter Models,” SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, pp.1–16, 2020.
- [27] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, “A framework for few-shot language model evaluation,” 2021. <https://doi.org/10.5281/zenodo.5371628>