日本語平易化への Task Arithmetic の応用とその検証

小西修平[1]
[1]日本放送協会
konishi.s-fq@nhk.or.jp

概要

大規模言語モデルは平易化においても一定の性能を持つことが知られており，平易化コーパスによる学習を行うことでより高性能な平易化モデルが作成できることが期待できるが，日本語平易化データセットは質・量ともに十分ではない．この問題を解決するため，モデルの重みの加減算によりモデルの能力を転移する技術である Task Arithmetic の応用を検討したが，平易化へ適用した事例は少ない．本研究では日本語平易化コーパス SJNC を用い， 平易化における Task Arithmetic の有効性を検証した．自動評価の結果スコアの向上が確認され， Task Arithmetic が日本語の平易化に対しても有効である可能性が示された．

1 はじめに

訪日外国人数の増加から，意味を保持しつつ難解な表現を平易な表現に変換する平易化の重要性は増している．ChatGPT などの大規模言語モデルは多くのタスクで高い性能を示しており,平易化においてもその効果が期待されている．Transformer ベースのモデルと ChatGPT を比較した研究[1]では，難易度の制御に課題がありつつも文意解釈性・文意正確性の観点で高品質な文章を生成できることが確認されている．そのため平易化コーパスを用いて教師あり学習を行うことで,より高度な平易化能力をもつモデルを作成できると考えられる．

すでにいくつかの平易化コーパスが提案されているが,日本語の平易化コーパスは非日本語資源に比べて量が少ない.また[2]では既存の平易化コーパスの一部について,原文に含まれない情報の追加や重要な情報の省略など原文への忠実性が低いことが指摘されており,質の観点でも分析の余地がある.さらに平易化コーパスの作成にはアノテーターの基準の統一や専門性が求められるためコストが高くなり，大規模なコーパスを作成することは難しい．

この問題に対処するために Task Arithmetic を応用することを考える．Task Arithmetic は，モデルの重みを加減算することでその性能を異なるタスクに転移する手法である．この手法は特にチャット能力や数学的能力の向上において効果が確認されている．

著者はこの Task Arithmetic を平易化タスクに応用することを考えた．英語の豊富な平易化コーパスで教師あり学習を行うことで平易化性能を向上させ，日本語の継続事前学習を行ったモデルに平易化機能を転移することで，日本語平易化コーパスを作成せずに平易化の性能を向上させることが期待できる．しかし Task Arithmetic を日本語平易化に適用した事例は少ない．

そこで本研究では日本語の平易化タスクにおける Task Arithmetic の有効性を検討した．朝日新聞社が作成したテキスト平易化コーパス「SJNC」を用いて Llama3 モデルにインストラクションチューニングを施し，平易化の学習を行った．次に学習によって変化したモデルの重みを計算し，Llama3 に日本語の継続事前学習を行った Swallow モデルに加算した．重み加算後のモデルの平易化能力を SARI と BLEU を用いて評価したところ，加算前に比べて評価値が改善した．また加算する程度を変化させたところ，評価値の向上を確認した．これらの結果から，Task Arithmetic が平易化においても有効である可能性が示された．

2 関連研究

2.1 テキスト平易化と平易化コーパス

平易化は機械翻訳などと同様に，系列変換タスクとして扱われる．近年では機械翻訳などと同様に Transformer[3]や BERT を用いた平易化[4]が提案されてきた．[5]はさらに大規模な Llama モデルを用いた語彙平易化を提案し，既存手法よりも高い精度を示した．

日本語平易化においても BART を用いた平易化

[6]や Transformer, BART および ChatGPT の比較研究[1]などがある. また平易化コーパスとしては５万文を人手で書き換えた SNOW T15 コーパス[7]や専門家によって平易化された記事による MATCHA コーパス[8]などが知られている. また原文への忠実性に考慮したコーパスとして朝日新聞の記事から構築された SJNC コーパスがある[9].

2.2 Task Arithmetic

Task Arithmetic[10]はモデルの重みの差分を利用してタスク固有の能力を転移する手法である.まず基準となるモデルに対して特定のタスクで微調整を行い, 微調整前後でのモデルの重みの変化量を計算する.この変化量を, 同じ構造をもつ別のモデルに加算することで, 微調整を行わずとも対象タスクの能力を向上させることができる. [10]では画像分類タスクや文章生成タスクにおいて Task Arithmetic が有効であることを示した. また[11]では Llama2 モデルに同手法を適用し, 対話タスクにおいて Task Arithmetic が有効であることを示した. この結果から, 英語データセットを用いて継続事前学習されたモデルと, そのモデルの対話能力を向上させたモデル, そして日本語の継続事前学習されたモデルという同じ構造をもつ3つの異なるモデルを用意すれば, 日本語の対話能力が向上したモデルを作成できることが知られている.

3 提案手法

3.1 平易化用微調整

平易化の能力を向上させた言語モデルを作成するために, 事前学習済みモデルに対して平易化コーパスを用いた微調整を行う. Huggingface ライブラリの SFTTrainer クラスを使用し, Instruction Tuning の形式でモデルの学習を実施する. 原文を入力とし, 出力として平易化された文を生成するよう学習を行う.

3.2 Simplification Vector

実験の概略図を図 1 に示す. 2.2 節で述べたように, Task Arithmetic は微調整されたモデルの重みから事前学習モデルの重みを引き, それを別のモデルに足し合わせることによって性能を転移する. 本実験では, 平易化の微調整による重みの変化を *Simp Vector*, 微調整前後のモデルをそれぞれ θ_{origin}, θ_{simp} とする. これは以下のような式で表される.

$$V_{simp} = \theta_{simp} - \theta_{origin} \quad (1)$$

ここで, $\theta_{origin}, \theta_{simp}, V_{simp} \in R^d$ である.また d は言語モデルの重みの次元数である.

次に, *Simp Vector* を日本語継続事前学習されたモデル θ_{new} に加算する.

$$\theta_{simp-new} = \theta_{new} + \lambda * V_{simp} \quad (2)$$

ここで $\theta_{simp-new}, \theta_{new} \in R^d$ である.また $\lambda \in R$ は *Simp Vector* を制御するスケールファクターである.

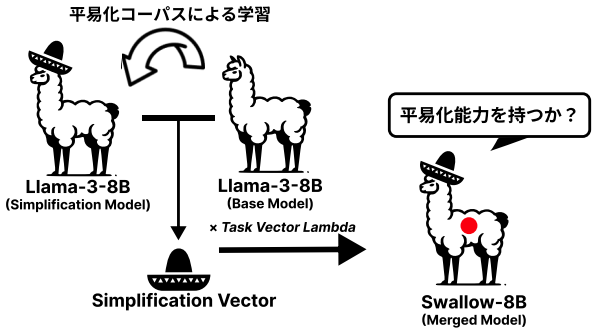

図 1 実験の概略図

4 実験手法

4.1 データセット

トレーニングデータには朝日新聞社が公開している SJNC[9]を用いた. SJNC は朝日新聞の記事から構築されており, 可読性と原文への忠実性のトレードオフを考慮して作成されている. SJNC のサンプルデータを表 1 に示す. kijiid を元に学習用・評価用・テスト用にデータを 8:1:1 に分割した.

表 1 SJNC のデータ例

idx	182
comp	"今後は各市で持ち帰り、具体的な連携の実現に向けて、やり方などを検討する。"
simp	"今後は 各市で 持って帰って、 具体的な 協力の 実現に向けて、 やり方などを 考えます。"
kijiid	"A1001120221107MTKE-1A-001"

4.2 評価指標

平易化の評価指標として一般的に用いられる BLEU[12]と SARI[13]を用いた. 機械翻訳の評価指標として使われる BLEU は n-gram の一致度を元に評価を行う. また平易化の指標として用いられる SARI は入力文・出力文・参照文の比較から評価を行

う. BLEU および SARI の実装には evaluate[a]ライブラリを用いた．また評価のための前処理として，fugashi[b]による単語分割をおこなった．

4.3 使用モデル

微調整用のモデルとして meta-llama/Llama-3.1-8B-Instruct[c], Simp Vector の適用先モデルに tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2[d]を用いた．

5 実験結果

5.1 微調整による平易化の学習

SJNC コーパスによって学習する前後の Llama3-8B モデルの出力例を表 2 に示す．微調整後のモデルは，難解な表現をより平易な表現に変換できている．例えば「獲得して」を「票を得て」，「圧勝」を「大きな勝利」と言い換えており，より理解しやすい表現に変換することができている．また微調整前後の評価指標の変化を表 3 に示した. BLEU, SARI ともにスコアが改善している．これらの結果から，文法性・流暢性に改善の余地があるが，微調整により平易化を学習させることができた．

表 2　微調整前後のモデルの出力例

原文
"党員らによる投票で約８５％を獲得して圧勝した。"
参照文
"党員たちによる投票で 約８５％を 勝ち取って一方的に 勝ちました。"
微調整前
"文脈に基づくと、パラグラフは政治的な結果に関連していますが、具体的にどの国の政治について言及しているわけではありません。ただし…"
微調整後
"党員たちが投票して約８５％の票を得て大きな勝利を得ました。"

表 3　微調整前後の評価指標の変化

	BLEU	SARI
微調整前	3.90	28.40
微調整後	36.41	48.46

5.2 Simp Vector の適用

微調整により計算した *Simp Vector* を Llama-3.1-Swallow-8B-Instruct-v0.2 に適用した. *Simp Vector* を制御するスケールファクターλを 0.00 から 3.00 に 0.25 刻みで変化させ，その効果を検証した．実験結果として以下の特徴が観察された:

i. λ=0.0: Swallow モデルの基本出力のみで，平易化の特徴はみられない
ii. λ=0.25~1.75: 文の意味を保持しつつ平易化が行われている
iii. λ=2.00~3.00: 原文に含まれない情報の追加や英単語の出力など破綻した文章が生成される割合が増加した

出力の一例を表 4 に示す．また参考として Llama-3.1-Swallow-8B-Instruct-v0.2 に 5.1 節と同じ条件で微調整を行った場合の出力も示す．
λ=0.0 では入力文に対する考察が出力され，平易化のような言い換えはみられなかった．λ=0.25~1.75 では概ね破綻のない文章が生成された．例えば「圧勝」が「圧倒的な勝利」などの単語に置き換えられるなど，一定の平易化が行われていた. λ=2.00~3.00 では原文に含まれない情報や，英単語の混在など破綻した文章が生成される傾向が見られた．

図 2 に BLEU と SARI による評価結果の変化を示す．λ=0.0 では BLEU,SARI はそれぞれ 6.20,32.09 でありスコアは低い．λを増加させるとスコアの改善がみられ，λ=1.50 が最良のスコアとなった．またλ=2.00 を超えるとスコアの逓減がみられた．

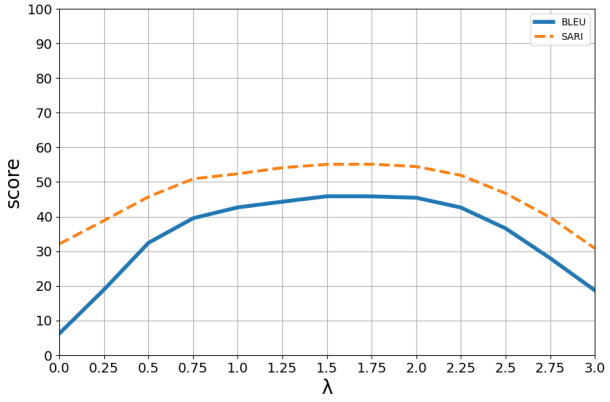

図 2　Simp Vector による評価指標の変化

[a] https://github.com/huggingface/evaluate
[b] https://github.com/polm/fugashi
[c] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[d] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2

ただし評価指標上は高品質な出力を観察すると，流暢性・文法性に課題のある言い換えが散見された．より詳細な評価のためには，人手による評価やLLM-As-a-Judge などの導入が必要と考えられる．

表 4　Simp Vector の適用例

λ	出力例
0.0	"これは、ある政党の選挙での得票率についての情報です。具体的には、その政党が党員からの投票で約 85%を獲得し…"
0.25	"党員らの投票の結果、約８５％の票を得て圧倒的な勝利を収めた。"
0.50	"党員らによる投票で約 85%を獲得し、圧倒的な支持を受けて再選されました。|end|>"
0.75	"党員たちが投票を行い、約８５％の票を獲得して圧倒的な勝利を収めました。"
1.00	"党員たちが投票して、約８５％の票を得て、圧倒的に勝利しました。"
1.25	"党員らによる投票で約８５％の票を集めて、圧倒的な差をつけて勝ちました。"
1.50	"党員たちが投票して、８５％以上の得票率で、圧倒的な勝利を収めました。"
1.75	"党員らによる投票で、約８５％の支持を得て、圧倒的に勝利しました。"
2.00	"党員らによる投票で約８５％の票を得て、おそろしい強さで、とうとう、圧倒的に勝利しました。"
2.25	"党員たちが選挙で約８５％の票を得て、李氏は圧倒的な勝利を収めました"
2.50	"党員らによる投票で、約８５％の 圧倒的な 得票で、 圧"
2.75	"党員たちが, その人に 投票 して, 圧 (あっこう) 勝 (しょうしょう) I won a landslide"
3.00	"党員らによる投票で、約８５％の<|user|>"
(参考)	"党員たちが投票をして、約８５％の票を得て、圧勝しました"

おわりに

本研究では Task Arithmetic の平易化タスクにおける有効性を検証した．平易化コーパス SJNC を用いて Llama-3.1-8B-Instruct を微調整し，平易化能力の獲得を確認した．この微調整による重みの変化(*Simp Vector*)を，同じ構造をもつ日本語継続事前学習モデル Llama-3.1-Swallow-8B-Instruct-v0.2 に加算し，平易化性能を調査した．その結果，平易化用の微調整を行っていない Swallow モデルにおいても，平易化の能力が向上したことを確認した．さらに *Simp Vector* を制御するパラメータを変化させることで, BLEU および SARI のスコアが向上した．以上の結果から，平易化タスクにおいても Task Arithmetic が有効である可能性を示した．ただし高スコアを示した出力文についても，文法性・平易性に課題のある出力がみられた.より詳細な評価を行うためには,自動評価以外の手法を検討する必要があると考える．

謝辞

本研究では,実験データとして朝日新聞社が作成した SJNC コーパスを利用させていただきました．貴重なデータを提供していただいたことに深く感謝いたします．

参考文献

1. 樽本空宙, et al. "系列変換タスクにおける ChatGPT の日本語生成能力の評価." *第 22 回情報科学技術フォーラム* (2023): 329-336.
2. Devaraj, Ashwin, et al. "Evaluating Factuality in Text Simplification." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 7331–345.
3. Vaswani, A. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).
4. Sun, Renliang, and Xiaojun Wan. "SimpleBERT: A Pre-trained Model That Learns to Generate Simple Words." arXiv preprint arXiv:2204.07779 (2022).
5. Baez, Anthony, and Horacio Saggion. "LSLlama: Fine-tuned LLaMA for lexical simplification." Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability. 2023.
6. 堀口航輝, et al. "日本語医療テキスト平易化の訓練用データセットの構築." 人工知能学会全国大会論文集 第 38 回 (2024). 一般社団法人 人工知能学会, 2024.
7. Maruyama, Takumi, and Kazuhide Yamamoto. "Simplified corpus with core vocabulary." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

8. 宮田莉奈, et al. "MATCHA: 専門家が平易化した記事を用いたやさしい日本語パラレルコーパス." 自然言語処理 31.2 (2024): 590-609.
9. Urakawa, Toru, et al. "A Japanese News Simplification Corpus with Faithfulness." *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024.
10. Ilharco, Gabriel, et al. "Editing models with task arithmetic." arXiv preprint arXiv:2212.04089 (2022).
11. Huang, Shih-Cheng, et al. "Chat Vector: A Simple Approach to Equip LLMs With New Language Chat Capabilities." arXiv preprint arXiv:2310.04799 (2023).
12. Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
13. Xu, Wei, et al. "Optimizing statistical machine translation for text simplification." Transactions of the Association for Computational Linguistics 4 (2016): 401-415.
14. 白藤大幹, et al. Task Arithmetic に基づく言語モデルにおけるバイアス低減手法の検討. 人工知能学会第二種研究会資料, 2024, 2024.AGI-028: 06.