1W-02

大規模言語モデルによる日本語感情分析の性能評価

近藤 里咲† 大塚 琢生† 梶原 智之†‡ 二宮 崇† 早志 英朗‡ 中島 悠太‡ 長原 一‡
愛媛大学† 大阪大学‡

1 はじめに

近年，大規模言語モデル（LLM: Large Language Model）の文脈内学習[1]が様々な自然言語処理タスクにおいて活用されている．日本語においても，GPT-NeoX[2]や LLaMA2[3]をベースに LLM の開発が盛んに進められている．現状，日本語の自然言語処理に活用できる LLM は，以下の 4 種類に大別できる．

① 日本語データのみを用いて訓練
② 日本語と英語の両方のデータを用いて訓練
③ 英語モデルを日本語データで追加訓練
④ 英語を中心とするデータを用いて訓練

LLM の性能は，英語においても日本語においても，言語理解や言語生成などの様々なタスクにおいて評価されている[4]．その中で本研究では，日本語の SNS テキストを対象とする感情極性分類[5-6]に焦点を当て，広く日本語 LLM の性能を評価する．感情極性分類は，テキスト分類における代表的なタスクのひとつであり，古くから盛んに研究されている．しかし，代表的な日本語 LLM ベンチマークのひとつである llm-jp-eval リーダーボード[1]には，感情極性分類が含まれていない．本研究では，感情極性分類タスクにおける日本語 LLM の振る舞いを分析し，今後の LLM 研究開発や活用のための知見を得る．

2 実験設定

2.1 データセット

本研究では，日本語の感情分析コーパスの WRIME[2] [5-6]を用いて，日本語 LLM の性能を評価する．本コーパスには，日本語の SNS 投稿テキストと，その投稿に対する書き手の感情極性をはじめとする複数の感情ラベルが付与されている．本研究では，5 段階の書き手の主観的な感情極性ラベルを推定するタスクに取り組む．

Evaluation of LLMs on Japanese Sentiment Analysis
Risa Kondo † (kondo@ai.cs.ehime-u.ac.jp)
Takumi Ohtsuka † (ohtsuka@ai.cs.ehime-u.ac.jp)
Tomoyuki Kajiwara †‡ (kajiwara@cs.ehime-u.ac.jp)
Takashi Ninomiya † (ninomiya@cs.ehime-u.ac.jp)
Hideaki Hayashi ‡ (hayashi@ids.osaka-u.ac.jp)
Yuta Nakashima ‡ (n-yuta@ids.osaka-u.ac.jp)
Hajime Nagahara ‡ (nagahara@ids.osaka-u.ac.jp)
† Ehime University ‡ Osaka University

[1] http://wandb.me/llm-jp-leaderboard
[2] https://github.com/ids-cv/wrime

2.2 日本語の大規模言語モデル

HuggingFace Transformers[3]にて公開されている日本語 LLM の中から 30 モデルを評価する．表 1 に示すように，①日本語データのみで訓練した LLM の中から 8 モデル，②日本語と英語の両方のデータで訓練した LLM の中から 7 モデル，③英語モデルを日本語データで追加訓練した LLM の中から 8 モデル，④英語を中心とするデータで訓練した LLM の中から 7 モデルを評価する．

比較手法として，BERT[4]および RoBERTa[5]を WRIME でファインチューニングして評価する．

2.3 評価方法

LLM はテキスト生成モデルであるため，LLM を用いて分類タスクを解くにあたっては特別な工夫が必要となる．本研究では，lm-evaluation-harness[6]の枠組みを用いて，LLM が各ラベルを生成する確率を算出し，最高の生成確率を持つラベルを出力として採用する．

LLM には Alpaca 形式の指示を与え，ネガティブ，ややネガティブ，どちらでもない，ややポジティブ，ポジティブの 5 段階の感情極性ラベルのうちのひとつを応答させる．文脈内学習の事例は WRIME の訓練用および検証用データから無作為抽出し，事例を与えない 0-shot および 10 件の事例を与える 10-shot の実験を比較する．

WRIME の感情極性分類は順序分類問題であるため，評価指標には重み付きカッパ係数（QWK: Quadratic Weighted Kappa）を用いる．評価用データの全 2,500 件に対する QWK を評価する．

3 実験結果

表 1 に実験結果を示す．I 列は指示チューニング[7]の有無を表す．比較手法の性能は，BERT が QWK=0.524，RoBERTa が QWK=0.594 である．

まず，全体的な傾向を見ると，③の英語モデルを日本語データで追加訓練するアプローチが最も有効である．指示チューニングと文脈内学習を組み合わせると，BERT や RoBERTa のファインチューニングを上回る性能を達成できる．

[3] https://huggingface.co/transformers
[4] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
[5] https://huggingface.co/nlp-waseda/roberta-base-japanese-with-auto-jumanpp
[6] https://github.com/Stability-AI/lm-evaluation-harness

表 1： QWK による評価結果（赤背景：0.1 未満，青背景：0.3–0.5，緑背景：0.5 以上，太字：10-shot で改善）

	モデル	I	0-shot	10-shot
①日本語データのみで訓練	line/japanese-large-lm-3.6b		0.162	0.068
	line/japanese-large-lm-3.6b	✓	0.169	0.022
	rinna/japanese-gpt-neox-3.6b		0.001	**0.158**
	rinna/japanese-gpt-neox-3.6b-v2	✓	0.305	**0.378**
	rinna/japanese-gpt-neox-3.6b-ppo	✓	0.227	**0.396**
	cyberagent/open-calm-7b		0.018	**0.307**
	stockmark/stockmark-13b		0.010	**0.061**
	stockmark/stockmark-13b	✓	0.063	**0.227**
②日本語と英語の両方のデータで訓練	cyberagent/calm2-7b		0.140	**0.366**
	cyberagent/calm2-7b	✓	0.026	**0.392**
	matsuo-lab/Weblab-10b		0.001	**0.111**
	matsuo-lab/weblab-10b	✓	0.008	**0.127**
	llm-jp/llm-jp-13b-v1.0		0.047	**0.257**
	llm-jp/llm-jp-13b-full-jaster-v1.0	✓	0.285	0.179
	llm-jp/llm-jp-13b-full-jaster-dolly-oasst-v1.0	✓	0.368	0.286
③英語モデルを日本語データで追加訓練	elyza/ELYZA-japanese-Llama-2-7b		0.041	**0.077**
	elyza/ELYZA-japanese-Llama-2-7b	✓	0.006	**0.204**
	rinna/youri-7b		−0.001	**0.193**
	rinna/youri-7b	✓	0.529	**0.552**
	stabilityai/japanese-stablelm-beta-7b		0.066	**0.394**
	stabilityai/japanese-stablelm-beta-7b	✓	0.002	**0.336**
	stabilityai/japanese-stablelm-beta-70b		0.132	**0.432**
	stabilityai/japanese-stablelm-beta-70b	✓	0.487	**0.641**
④英語を中心とするデータで訓練	HuggingFaceH4/zephyr-7b-beta	✓	0.172	**0.498**
	meta/Llama-2-7b-hf		0.039	0.036
	meta/Llama-2-7b-chat-hf	✓	0.066	**0.268**
	meta/Llama-2-13b-hf		−0.010	**0.410**
	meta/Llama-2-13b-chat-hf	✓	0.010	**0.453**
	meta/Llama-2-70b-hf		0.078	**0.107**
	meta/Llama-2-70b-chat-hf	✓	0.110	**0.479**

③の次に有力なのは④の英語モデルである．過半数の英語モデルは 0.4 を超える QWK を達成するが，日本語データをベースとする①や②の LLM がこの水準に到達することは難しい．

指示チューニングは，全体的に有効である．指示チューニングの有無を比較すると，0-shot と 10-shot の両設定にて，12/15 モデルが改善した．

文脈内学習も，全体的に有効である．0-shot 設定では 18/30 モデルが QWK=0.1 を下回る（赤）が，10-shot 設定で 25/30 モデルが改善（太字）．

大きいモデルが高性能というわけではないが，訓練データが同じなら大きいモデルの性能が高い場合が多い．例えば，①では open-calm-7b が stockmark-13b より高性能であり，②では calm2-7b が Weblab-10b より高性能である．しかし，③では stablelm-7b より stablelm-70b が高性能である．

4 おわりに

日本語 LLM を感情極性分類で評価した結果，英語モデルを追加訓練する手法が強力であり，指示チューニングや文脈内学習も有効であった．

参考文献

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah et al. Language Models are Few-Shot Learners. In Proc. of NeurIPS, pp.1877-1901, 2020.

[2] Sidney Black, Stella Biderman, Eric Hallahan et al. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In Proc. of BigScience, pp. 95–136, 2022.

[3] Hugo Touvron, Louis Martin, Kevin Stone, … and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288, 2023.

[4] 樽本空宙，畠垣光希，宮田莉奈，梶原智之，二宮崇．ChatGPT の日本語生成能力の評価．自然言語処理，Vol.31, No.2, 2024.

[5] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima and Hajime Nagahara. WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations. In Proc. of NAACL, pp. 2095–2104, 2021.

[6] Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima and Hajime Nagahara. A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In Proc. of LREC, pp. 7022–7028, 2022.

[7] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu et al. Finetuned Language Models Are Zero-Shot Learners. In Proc. of ICLR, 2022.