順序回帰ネットワーク CORN を用いた単言語学習による多言語強度推定

任晶　福田悟志　難波英嗣　庄司裕子
中央大学大学院理工学研究科

概要

多言語環境では、XLM-R や LaBSE などの多言語事前学習モデルを用い、各言語のラベル付きデータを組み合わせて学習する手法が一般的である。しかし、多言語データの収集・整備には多大なコストがかかるという問題がある。本研究では、日本語のみのラベル付きデータを用いて学習したモデルを、多言語データに適用する枠組みを提案する。具体的には、文埋め込みモデルによりテキストを固定長ベクトルに変換し、回帰モデルおよび順序回帰モデル（CORN）で感情強度を推定する。さらに、埋め込み次元数や過学習が多言語への汎化性能に与える影響について、実験的に分析する。

1 はじめに

本研究では、日本語のみのデータセットで推定モデルを学習し、そのモデルが他言語にどの程度汎化するかを検証する。モデル構成としては、文埋め込みモデルによりテキストを固定長ベクトルに変換し、回帰モデルおよび順序回帰モデルを用いて感情強度を推定する。さらに、学習済みモデルを多言語データセットに適用し、多言語への汎化性能を評価する。本枠組みにより、言語依存の特徴量設計を行わずに多言語汎化性能を評価できる。

本研究の主な貢献は以下の通りである。

1. 日本語のみの学習データで構築したモデルを、多言語データに適用する枠組みを示す。
2. 単一のモデルで複数言語の感情強度推定を行う。
3. 埋め込み次元数と過学習の観点から、多言語汎化性能への影響を実験的に分析する。

2 関連研究

Feng et al.(2022)[1]では、大規模な多言語翻訳対（parallel corpus）を用いた対照学習を採用し、ベクトル空間には翻訳ペアの文同士を近づけ、無関係な文同士を遠ざけるように学習する。LaBSE（Language-agnostic BERT Sentence Embedding）は異なる言語で意味的に同等な（対訳関係にある）文を同じベクトル空間に埋め込むことを目的とした文埋め込みモデルである。主な狙いは、翻訳文検索（bitext mining）や多言語情報検索、クロスリンガル類似度計算などのタスクを、高精度かつ言語非依存で実現することにある。

Wang et al.(2024)[2]では、弱教師あり（weakly-supervised）対照学習を採用し、大規模に収集されたテキストペアデータセット（CCPairs）を用いて、意味的に関連する文ペアと無関係な文を区別する形で学習し、文や文章を一つのベクトルに変換する設計で、ベクトル同士の類似度計算がそのまま検索や分類に使える。E5（Embeddings from bidirectional Encoder representations）が、単一ベクトルでテキスト（文章）を表現するためのモデルで、検索、クラスタリング、分類など幅広い NLP タスクに対応できる汎用的なテキスト埋め込みを提供する。

Enevoldsen et al.(2025)[3]では、MMTEB（Massive Multilingual Text Embedding Benchmark）というテキスト埋め込みモデルを多言語・多タスクで大規模かつ公平に評価するためのベンチマークを提案する。従来のベンチマークは言語やタスクの範囲が限定されていたが、MMTEB はその制約を大幅に解消し、モデルのクロスリンガル性能や低リソース言語での性能比較が可能になる。MMTEB は 250 以上の言語と 500 以上の評価タスクを含む非常に大規模なベンチマークである。タスクには、検索（retrieval）、分類（classification）、クラスタリング（clustering）、意味類似度（semantic similarity）、命令フォロー（instruction following）、コード検索（code retrieval）など多岐にわたっている。

Acharya et al.(2024)[4]では、デュアルエンコーダ構造（クエリエンコーダ＋ドキュメントエンコー

ダ）を採用し、異なる言語の文章を共通のセマンティック空間に埋め込む。NLLB エンコーダを多言語埋め込みに活用し、E5 モデルから知識蒸留（distillation）によって多言語検索性能を獲得する。多言語のラベル付きデータを必要とせず、英語の豊富なデータを利用したモデル蒸留で学習するため、低リソース言語にも対応可能である。NLLB-E5 は、多言語情報検索（Multilingual Retrieval）を効率的に行うために設計されたスケーラブルな検索モデルで、特に低リソース言語（例：インディック諸言語）への対応を重視し、多言語訓練データがなくてもゼロショット検索を可能にする。

本研究では日本語のみの学習データを用いて埋め込み表現に変換し、機械学習を通して多言語に対応する分類器を構築することを提案する。

3 提案手法

本章では、日本語のみのラベル付きデータから多言語対応の感情強度推定器を構築する枠組みを提案する。提案手法は、(1)日本語テキストを多言語文埋め込みにより固定長ベクトルへ変換し、(2)そのベクトルを入力として順序回帰により感情強度を推定する、という二段階からなる。多言語文埋め込みは言語間で表現空間を共有するため、日本語で学習した推定器を他言語へゼロショットで適用できることを狙う。

3.1 学習データと入力表現

学習には、日本語の感情強度推定データセット WRIME[5]を用いる。各テキストを多言語文埋め込みモデル$f(\cdot)$に入力し、固定長ベクトル$x=f(t)\in R^d$を得る。このxを推定モデルの入力とし、WRIME の教師ラベルに基づいて学習を行う。

なお、本研究の新規性は「特定の埋め込み（例：Qwen3）を用いること」それ自体に置くのではなく、多言語埋め込みを入力表現として用いることで、日本語のみの学習から多言語へ汎化可能な感情強度推定器を構築できるという枠組みの検証にある。そこで埋め込みモデルは単一に固定せず、Qwen3 Embedding[6]、E5、Linq-Embed-Mistral など複数の候補を比較対象として用い、埋め込みの性質（次元数等）の違いが多言語汎化に与える影響も併せて検討する。

3.2 順序回帰 CORN を用いた感情強度推定モデル

感情強度ラベルは「弱い→強い」のように順序をもつため、回帰（SVR/GBDT/ニューラル回帰など）や多クラス分類など複数の定式化が考えられる。本研究では、ラベルの順序性を明示的に利用できる順序回帰を重視し、CORN（Conditional Ordinal Regression for Neural Networks）[7]を用いる。

CORN は、順序付きラベルを「クラスがk以上か？」という複数の二値分類問題に分解して学習する。例えばラベルが 0, 1, 2, 3 の 4 段階であれば、「≥1」「≥2」「≥3」を順に学習し、それらの出力に基づいて最終クラスを推定する。本研究では、埋め込みベクトルxを入力とし、MLP を介して CORN の各二値分類器を出力する MLP+CORN 構成を基本とする。

3.3 多言語対応の評価方針

WRIME（日本語）のみで評価すると、多言語対応（クロスリンガル汎化）を十分に検証できない。そこで本研究では、多言語の感情強度推定データセット BRIGHTER[8]を用いて、多言語への汎化性能を評価する。BRIGHTER は 28 言語を含み、感情強度ラベルが付与されているのは 10 言語である。

評価では、BRIGHTER で追加学習は行わず、WRIME で学習した推定器を BRIGHTER の各言語テストデータにゼロショットで適用する。これにより、「日本語のみで学習したモデルが、どの程度他言語に一般化できるか」を直接評価できる。実験設定と評価指標の詳細は第 4 節で述べる。

4 実験

4.1 日本語データセットでの学習結果

4.1.1 実験条件

実験データ

感情分類器の構築には、日本語の感情強度推定のためのデータセットである WRIME を用いる。本実験では Kajiwara ら[5]の分割に従い、30,000 件の訓練用データ、2,500 件の検証用データ、2,500 件の評価用データを重複なく使用する。

表 1 感情分類器の評価結果

	提案手法		比較手法			
	Qwen3(8b)＋MLP+CORN	E5＋MLP+CORN	Linq-Embed-Mistral +SVR	Linq-Embed-Mistral +GBDT	E5＋MLP+EMD	Qwen3(8b)＋LP+EMD
joy	0.5657	0.5582	0.5714	0.5389	0.5369	0.5323
sadness	0.4515	0.4395	0.4685	0.3956	0.4169	0.4048
anticipation	0.4716	0.4184	0.4204	0.4266	0.3695	0.4109
surprise	0.3424	0.3261	0.3371	0.3162	0.0000	0.0029
anger	0.4636	0.4037	0.4441	0.3956	0.0343	0.0891
fear	0.3630	0.2975	0.3561	0.3119	0.0681	0.1520
disgust	0.4588	0.4220	0.4528	0.4502	0.1671	0.1827
trust	0.3119	0.2878	0.2462	0.2399	0.0753	0.0406
average	**0.4286**	0.3941	0.4121	0.3844	0.2085	0.2269

実験手法

多言語用のモデルを日本語コーパス WRIME で学習させ、さらなる精度の高い多言語用モデルを検討する。埋め込み表現 E5 を Linq-Embed-Mistral と qwen3-embedding に切り替え、学習モデルを MLP+CORN あるいは EMD に切り替え、また検証用データセットを K-Fold で 5 つに分け、しきい値を検討する。

提案手法および比較は以下の通りである。

- 提案手法
  - E5＋MLP+CORN
  - Qwen3＋MLP+CORN
- 比較手法
  - Linq-Embed-Mistral＋SVR
  - Linq-Embed-Mistral＋GBDT
  - E5＋MLP+EMD
  - Qwen3＋MLP+EMD

評価尺度

感情分類器の評価には Quadratic Weighted Kappa(QWK)を用い、感情ごとに評価を行う。QWK はラベル間に順序があるような場合に用い、今回の強度ラベルを考慮した評価に適している。

4.1.2 実験結果

評価結果を表 1 に示す。提案手法の評価結果については、Qwen3＋CORN +KFold が一番高い結果を示し、次は Linq-Embed-Mistral +SVR である。しかし、Linq-Embed-Mistral で埋め込み表現に変換する時間が長い。

4.2 多言語データセットへの対応結果

4.2.1 実験条件

実験手法

BRIGHTER の多言語用学習用データと検証用データで学習せず、WRIME で学習済みの Qwen3/E5+MLP+CORN モデルを評価用データに適用する。

評価尺度

Quadratic Weighted Kappa(QWK)による評価の他に、BRIGHTER で用いられている Pearson 相関係数でも評価を行う。Pearson 相関係数は 2 つの変数の線形な関係の強さと向きを表す指標である。

4.2.2 実験結果

表 2 に示すように、多言語データセットへの汎化性能は E5+MLP+CORN が QWK と Pearson の両方で最も高かった。一方、Qwen3+MLP+CORN は WRIME では高い性能を示すものの、BRIGHTER へのゼロショット適用では最も低い結果となった。これは、Qwen3 の埋め込み次元数が大きく（4096）、日本語データセット WRIME 上で過学習しやすい可能性があるためと考えられる。

表 2 多言語データセットへの対応結果

	QWK による評価			Pearson 相関係数による評価		
	E5+GBDT	E5+MLP+CORN	Qwen3(8b)+MLP+CORN	E5+GBDT	E5+MLP+CORN	Qwen3(8b)+MLP+CORN
joy	0.4070	0.4889	0.2998	0.5323	0.5995	0.4251
sadness	0.2490	0.2132	0.1593	0.3116	0.3321	0.2759
surprise	0.1294	0.1574	0.0890	0.2136	0.2584	0.1864
anger	0.2349	0.2458	0.1550	0.2655	0.3333	0.2578
fear	0.1561	0.1820	0.1257	0.2478	0.3279	0.2464
disgust	0.0762	0.0854	0.0854	0.1092	0.1793	0.1899
average	0.2088	**0.2288**	0.1524	0.2800	**0.3384**	0.2636

4.3 過学習への検討

4.2 と同様に BRIGHTER で追加学習は行わず、テストデータで評価した。Qwen3(8b)は次元数 4096、E5 は 1024 である。そこで、本研究では次元数が中間的な Qwen3(4b)（次元数 2560）を用い、過学習の影響を検討する。評価尺度は QWK に加えて Pearson 相関係数を用いる。

表 3 に示すように、Qwen3(4b)+MLP+CORN は BRIGHTER に対して最も高い結果を得た。これにより、Qwen3(8b)+MLP+CORN は WRIME 上で過学習している可能性が示唆された。また、Qwen3(4b)+MLP+CORN が E5+MLP+CORN を上回ったことから、少なくとも 2560 次元程度までは WRIME で過学習しにくい可能性がある。

表 3 Qwen3(4b)+MLP+CORN による実験結果

	QWK	Pearson
joy	0.5107	0.6033
sadness	0.2429	0.3676
surprise	0.1628	0.2430
anger	0.3434	0.4605
fear	0.2616	0.4100
disgust	0.1389	0.2238
average	0.2767	0.3847

5 結論

本研究では、日本語のみのデータセット WRIME で学習した感情強度推定モデルを、多言語データセット BRIGHTER にゼロショットで適用し、多言語への汎化性能を評価した。日本語データセットでの学習では Qwen3(8b)+MLP+CORN が最も高い結果を得た。一方、BRIGHTER への対応では Qwen3(4b)+MLP+CORN が最も高い結果を示し、Qwen3(8b)+MLP+CORN は相対的に低い結果となった。以上より、高次元の埋め込み（4096 次元）は単言語データ上で過学習し、多言語汎化性能を損なう可能性があることが示唆された。

参考文献

[1] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang: Language-agnostic BERT Sentence Embedding, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp.878-891, 2022.

[2] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei: Text Embeddings by Weakly-Supervised Contrastive Pre-training, arXiv:2212.03533v2 [cs.CL], 2024.

[3] Kenneth Enevoldsen, et al.: MMTEB: Massive Multilingual Text Embedding Benchmark, Proceedings of the Thirteenth International Conference on Learning Representations, poster, 2025.

[4] Arkadeep Acharya1, Rudra Murthy, Vishwajeet Kumar, and Jaydeep Sen: NLLB-E5: A Scalable Multilingual Retrieval Model, arXiv:2409.05401v1[cs.IR], 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. Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.2095-2104, 2021.

[6] Yanzhao Zhang, Mingxin Li, DingkunLong, Xin Zhang, Huan Lin, Baosong Yang, and PengjunXie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou: [6 Embedding: Advancing Text Embedding and Reranking Through Foundation Models, arXiv:2506.05176v3[cs.CL], 2025.

[7] Xintong Shi, Wenzhi Cao, Sebastian Raschka: Deep Neural Networks for Rank-Consistent Ordinal Regression Based on Conditional Probabilities, Pattern Analysis and Applications, Vol. 26, pp. 941-955, 2023.

[8] Shamsuddeen Hassan Muhammad, et al.: BRIGHTER: BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pp.8895-8916, 2025.