X のポストデータに対するレーティング予測

森廣 勇樹 [1]　南條 浩輝 [2]　馬 青 [1]
[1] 龍谷大学理工学研究科　[2] 滋賀大学データサイエンス学部
[1]t22m005@mail.ryukoku.ac.jp
[2]hiroaki-nanjo@biwako.shiga-u.ac.jp
[1]qma@math.ryukoku.ac.jp

概要

本研究では，複数の深層学習モデルで X のポストデータに対するレーティング予測を行った．ポストデータは話し言葉表現に近いことに着目し，事前学習モデルとして日本語話し言葉 BERT を導入した．他のモデルとの比較の結果，日本語話し言葉 BERT を用いた場合に高いレーティング予測精度が得られた．これはポストデータ特有の話し言葉表現が適切に捉えられているためと考えられる．次に，学習データである amazon レビューデータと X のポストデータの性質（文字数とテキストスタイル）が異なることから，学習データの書き換えを検討した．具体的には，ChatGPT と要約ツールを用いて amazon レビューデータを X のポストデータに近いデータに書き換えを検討したので，その報告もあわせて行う．

1　はじめに

本稿では，日本語の X（旧 Twitter）のポスト（旧ツイート）データに対してのレーティング予測について述べる．

商品レビューは消費者の正直な感想であるため広告より信頼性が高いが，ヤラセ・サクラなど，レビュー の悪用や利用者増加に伴う役に立たないレビューの増加で，真に有用なレビューが埋もれている．推薦・推薦文においてこのようなレビュー等が主に使われており，それらの結果に疑問が残るのではないかと考える．それに比べ，X のポストなどの生の声は，レビュー文よりも実際の気持ちを表していると予想される．

我々はこれまでに，多言語 amazon レビューコーパス [1] の日本語レビュー文に対し，評価点（レーティング）がどの程度正確に予測できるかを調査している [2]．BERT と RoBERTa の 2 つのモデルに対し合計 360 通りのハイパーパラメータの組み合わせでグリッドサーチを行い，いずれのモデルでもレーティング予測の傾向自体の学習はできていることを確認した．本研究では，レビューデータではなく，レビューそのものを目的としていないユーザ生成テキスト，具体的には X のポストデータに対してレーティング予測精度の調査を行ったのでその結果を報告する．その際，ポストデータは話し言葉表現に近いことから，日本語話し言葉 BERT を導入したのでその結果も報告する．さらに，学習データ（amazon レビューコーパス）のレビューテキストをポストデータの表現に近いテキストに書き換えることも検討したのでその結果についても述べる．

2　関連研究

勝又ら日本語話し言葉に特化した BERT モデル [3] を開発している．従来の書き言葉に基づくモデルとは異なり，この研究は日本語の話し言葉コーパス（Corpus of Spontaneous Japanese: CSJ）を活用している．BERT モデルの特定層を話し言葉データで追加学習することと，分野適応の手法を用いて追加学習するという 2 つの手法を導入し，依存関係解析，文境界推定，キー文抽出などのタスクでモデルの有効性が評価され，構文タスクでは高い性能を示し，意味タスクでは限定的な効果を示した．本研究で日本語話し言葉 BERT を使用する理由として，X のポストデータはレビュー文に比べ話し言葉が使用されていることが多く本研究のレーティング予測に適していると考えたためである．

3　実験データ

3.1　使用した事前学習済みモデル

本研究では先行研究 [2] で使用した BERT と RoBERTa に加え，新たに 3 種類の日本語話し言葉

表 1　使用した話し言葉 BERT

モデル名	モデルの説明	本研究での表記
1-6_layer-wise	CSJ で 1-6 層のみを fine-tuning したモデル	CSJ1
tapt512_60k	CSJ で fine-tuning したモデル	CSJ2
dapt128-tap512	国会議事録データと CSJ で fine-tuning したモデル	CSJ3

表 2　amazon レビューコーパスの星評価の内訳

星評価＼データ	学習データ	検証データ	テストデータ
★	37,000	4,000	1,000
★★	37,000	4,000	1,000
★★★	37,000	4,000	1,000
★★★★	37,000	4,000	1,000
★★★★★	37,000	4,000	1,000

BERT を含む合計 5 つのモデルを使用した．使用した BERT は東北大学乾・鈴木研究室が Wikipedia で訓練した日本語 BERT モデル（cl-tohoku/bert-base-japanese-v2）[4] であり，RoBERTa は Language Identification データセットに基づいて fine-tuning された xlm-roberta-base モデル（papluca/xlm-roberta-base-language-detection）[5] である．日本語話し言葉 BERT は公開されている 3 種類を使用した（表 1）．内訳は，CSJ で 1-6 層のみを Fine-Tune したモデルである 1-6_layer-wise（以降 CSJ1），CSJ で Fine-Tune したモデルである tapt512_60k（以降 CSJ2），国会議事録データと CSJ で Fine-Tune したモデルである dapt128-tap512（以降 CSJ3）となっている．

3.2　学習データ

学習データには多言語 amazon レビューコーパスを用いた．データセットには，英語，日本語，ドイツ語，フランス語，中国語，スペイン語の合計 6 言語のレビューが含まれておりその中で日本語のみのレビューを使用した．データセットの各レコードには，レビューテキスト，レビュータイトル，星評価，匿名のレビュアーが含まれている．各言語ごとに合計 210,000 のレコードがあり，185,000 を学習データ，20,000 を検証データ，5,000 をテストデータに分割し，先行研究で評価データとして使用したテストデータを省く 205,000 を学習データとした．

表 3　X のポストデータの星評価の内訳

星評価	データ数
★	12
★★	33
★★★	38
★★★★	105
★★★★★	102

すべてのレビューは 2,000 文字を超えると切り捨てられ，すべてのレビューは少なくとも 20 文字の長さである．本研究で使用する amazon レビューコーパスの星評価の内訳を表 2 に示す．コーパスは星評価（星 1〜星 5）でバランスが取れており，各星評価はレビューの 20 %で構成されている．

3.3　テストデータ

本研究では，テストデータとして X のポストデータを使用した．データの収集方法は，「レビュー」のタグが付けられたポストから商品に対する評価を行っているものを選定するというものである．

テストデータの作成においてそれぞれのポストデータに対し，4 人の被験者が 5 段階の評価スケール（星評価）を用いて評価を行い，これらの評価の平均値を四捨五入した値を最終的な評価値として定め，テストデータセットとして利用した．この方法により，客観的かつ統一された基準に基づく評価データを生成することができていると考える．本研究で使用するテストデータの星評価の内訳を表 3 に示す．

4　日本語話し言葉 BERT を用いたポストデータのレーティング予測

4.1　実験設定

4.1.1　評価指標

本研究で評価指標として Quadratic Weighted Kappa (QWK) を用いる．これは予測値と実際の値に対す

表 4　各モデルのファインチューニング用ハイパーパラメータ

パラメータ ＼ モデル名	BERT	RoBERTa	日本語話し言葉 BERT（CSJ1, 2, 3）
最適化アルゴリズム	RMSProp	RMSProp	RMSProp
学習率	1e-05	5e-06	1e-05
バッチサイズ	4	8	4
エポック数	2	4	2

表 5　各事前学習モデルによるレーティング予測結果（QWK）

テストデータ ＼ モデル名	BERT	RoBERTa	CSJ1	CSJ2	CSJ3
amazon データ	0.816	**0.848**	0.814	0.819	0.816
ポストデータ	0.683	0.735	**0.763**	0.750	0.761

る一致度を偶然の一致を考慮して計算する評価指標である．順序データにおいて遠いクラスに誤られたものに大きなペナルティをかける指標となっている．QWK は式（1）で与えられる．

$$\mathrm{QWK} = 1 - \frac{\sum_{i,j} w_{i,j} o_{ij,}}{\sum_{i,j} w_{i,j} e_{i,j}} \tag{1}$$

ここで，$w_{i,j}$ はクラス i のデータを j と分類したときの重み（$w_{i,j} = (i-j)^2$），$o_{i,j}$ はクラス i のデータを j と分類した数，$e_{i,j}$ はランダムに分類したときにクラス i のデータを j と分類する数（期待値）である．

4.1.2　ハイパーパラメータの決定

本研究では BERT と RoBERTa の 2 つのモデルで最適なハイパーパラメータを決定するためにグリッドサーチを行った．考慮されたハイパーパラメータには最適化アルゴリズム（AdamW，SGD，RMSProp の 3 種類），学習率（アルゴリズムごとに 3 種類），バッチサイズ（32，16，8，4 の 4 種類），エポック数（1 から 10 の 10 種類）が含まれ合計 360 通りの組み合わせとなった．そのパラメータの中で最も高い QWK を示したパラメータの組み合わせを使用している．得られたハイパーパラメータを表 4 に示す．

日本語話し言葉 BERT に関しては基礎モデルの類似性や，計算効率の観点から BERT と同様のハイパーパラメータを使用した．

4.2　実験結果

amazon レビューコーパスと X のポストデータに対してレーティング予測を行った．結果を表 5 に示す．

BERT と RoBERTa は，書き言葉である amazon データに対して高い QWK スコアを示している．しかし話し言葉表現に近い X のポストデータでは性能が低下している．それに対し日本語話し言葉 BERT（CSJ1，CSJ2，CSJ3）は BERT や RoBERTa に比べ，それぞれ高い QWK を示している．これは書き言葉と話し言葉の違いがモデルの性能に影響を与えており，話し言葉特有の言語のニュアンスや文脈が日本語話し言葉 BERT によってより適切に捉えられていると考えられる．

5　学習データの書き換えの検討

amazon レビューデータと X ポストデータで適した事前学習モデルが異なることがわかった．この原因として，amazon レビューデータと X ポストデータそれぞれのテキストの特性に着目した．これらのデータセット間では文字数とテキストスタイル（書き言葉と話し言葉）に違いがある．

amazon データのレーティング予測に用いたモデルの学習データとポストデータのレーティング予測に用いたモデルの学習データは，ともに amazon データである．ポストデータに対する予測精度が amazon データに対する予測精度よりも低いのは学習データと評価データの不一致によると考えらる．そこで，学習データを評価データに近い形に書き換えることを検討した．具体的には，文字数に関する

表 6　書き換えた学習データによるモデル学習と X ポストデータに対するレーティング予測（QWK）

学習データの書き換え ＼ モデル名	BERT	RoBERTa	CSJ1	CSJ2	CSJ3
なし（amazon データ）	0.683	0.735	**0.763**	0.750	0.761
要約	0.581	0.654	0.687	0.712	0.663
ChatGPT	0.644	0.738	0.711	0.734	0.724

違いに対処するためより長いテキストを要約する方法を検討した．次にテキストスタイルの違いに対応するため ChatGPT を使用して書き言葉を話し言葉に変換する手法を採用した．これにより X ポストデータに類似したデータセットを作成し，それがモデルの学習に与える影響を評価した．

5.1　要約による書き換え

amazon レビューデータの平均文字数は 101.3 文字であったのに対し，文字数制限のある X のポストデータの平均は 72.3 文字であった．この文字量の差異が学習結果にどのような影響をもたらすかを分析し，結果を検証するために amazon レビューから平均文字数を超えるデータに対して要約処理を行い学習データとした．

5.2　ChatGPT による書き換え

書き言葉から話し言葉への変換手法として，自然言語処理モデルである ChatGPT を使用した．このモデルは書かれたテキストの言語的特徴を解析し，それを話し言葉の文体に再構成する能力を有していると考えられる．変換過程では文法的に正しい書き言葉を，口語的な表現や省略形日常会話で一般的に用いられる構造へと変換し，X のポストデータのようなデータとした．

5.3　実験結果

学習データの書き換えによって得られたデータで学習を行った結果を表 6 に示す．書き換えられたデータセットに基づいてモデルを学習しても精度の向上は見られなかった．amazon レビューデータの書き換えが QWK スコアの向上につながらなかったことは，元のデータセットの品質が既に高かったこと，変換後のテキストがモデルにとって有用な情報とならなかった可能性などが考えられる．変換プロセスの質を向上させることやモデルが変換後のテキストの特徴をより効果的に捉えるためのアプローチを行うことで精度向上につながるのではないかと考える．

6　終わりに

日本語のポストデータに対するレーティング予測を行った．日本語話し言葉 BERT を含む複数の事前学習モデルを用い，amazon レビューデータでファインチューニングして評価をおこなった．日本語話し言葉 BERT が他のモデルと比較してポストデータのレーティング予測精度が高かった．これはポストデータ特有の言語のニュアンスや文脈をより適切に捉えられているためと考えられる．次に，学習に用いた amazon レビューデータと X のポストデータには，文字数とテキストスタイルの違いがあることから，学習データの書き換えを検討した．具体的には，要約と ChatGPT を用いた変換を通じて新たな学習データの作成を行った．今回は，学習データの書き換え効果は見られなかった．今後は，学習データの書き換え方法をさらに検討していきたい．

謝辞

本研究は JSPS 科研費 19K12241 の助成を受けたものです．

参考文献

[1] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, 2020.

[2] 森廣勇樹, 南條浩輝, 馬青. 日本語レビューに対するレーティング予測の精度比較. 言語処理学会第 29 回年次大会, pp. 1686–1689, 2023.

[3] 勝又智, 坂田大直. CSJ を用いた日本語話し言葉 BERT の作成. 言語処理学会第 27 回年次大会, pp. 805–810, 2021.

[4] 東北大学乾・鈴木研究室 BERT モデル. cl-tohoku/bert-base-japanese-v2. https://huggingface.co/cl-tohoku/bert-base-japanese-v2.

[5] RoBERTa モデル. papluca/xlm-roberta-base-language-detection. https://huggingface.co/papluca/xlm-roberta-base-language-detection.