XLM-RoBERTa を利用した実データの英日評判分析

大井恵奈 [1]　古宮嘉那子 [1]　佐々木稔 [2]
[1] 東京農工大学生物システム応用化学府　[2] 茨城大学 工学部
s231186v@st.go.tuat.ac.jp　kkomiya@go.tuat.ac.jp
minoru.sasaki.01@vc.ibaraki.ac.jp

概要

本研究では，XLM-RoBERTa を用いて，実際に企業に寄せられたレビューの評判分析を行う．XLM-RoBERTa を用いた二値分類タスクの手法を元に，実際に日本企業に寄せられた多くの日本語レビューと少量の英語レビューを用いて Fine-tuning を行った．英語に日本語を追加し学習したモデルと英語のみで学習したモデル，日本語に英語を追加し学習したモデルと日本語のみで学習したモデルそれぞれについて AUC と正解率を算出し考察する．

1　はじめに

商品に対するユーザーの意見には様々存在し，ネガティブなレビューは商品を販売する企業にとっては商品開発の一助となることが期待される．レビュー文をポジティブなものとネガティブなものに分類する際，一般に理想的には商品ごとに一定の量のラベル付きデータが必要となる．

英語など話者が多く高リソースな言語では，EC サイトにおける 16 万件を超える単言語のレビューデータセットが公開されているが，一つの言語のレビューを分析するために大量のデータセットを用意することは困難である．そこで，本研究では，実際の企業のデータを用いて，様々な言語のレビューから新製品に寄せられたレビューの評判分析を行うことを目的とする．一つの言語のレビューを評判分析するとき，その言語のみを学習するだけのデータが足りない場合，別の言語のデータを用いて，fine-tuning を行う．その際に，言語横断的なタスクに高い性能を示す事前学習済みモデルである XLM-RoBERTa を用いることで，正解率が向上するか検討する．

2　関連研究

評判分析タスクには様々な手法が提案されている．例として，BERT[1] を用いた評判分析が挙げられる．三戸ら [2] は BERT において二つのデータセットにおいて共学習を行い，ネガティブなレビューのみを抽出する手法を提案している．すべてのレビューデータを文で区切り，レビューデータ全体の評点をそれぞれの文の評点として疑似的な文ラベルを付与した上で，日本語の二つのデータセットそれぞれから作成したモデルを共学習させる．この手法はデータセットの単純な五分割交差検定のベースラインの結果を上回った．

また，特に言語横断的な評判分析において，吉野ら [3] は三種類の BERT を用いて，訓練データが少ない言語の商品レビューの評判分析方法を比較している．具体的には，英語 BERT を大量の英語データによって Fine-tuning させたモデルに日本語のテストデータを機械翻訳したものを入力する手法，日本語 BERT を少量の日本語訓練データと大量の英語訓練データを日本語に機械翻訳したデータによって Fine-tuning させたモデルに日本語テストデータを入力する手法，M-BERT の 2 文入力タスクを応用して日本語と英語の対訳を訓練データとして Fine-tuning させたモデルに対訳の形のテストデータを入力する手法の 3 つが比較されている．これらの手法の中では，日本語 BERT を用いた手法がベースラインよりも高い精度を示している．Rusli ら [4] は XLM-RoBERTa を用いて評判分析を行っている．通販サイトでは評点が 1〜5 と設定されていることが多いが，評点の 1，2，3 を 0 とし，4，5 を 1 としてネガティブなレビューかポジティブなレビューかの二値分類として評判分析を行っている．このとき，一般に多くの量がある英語のデータを利用して，少量の日本語データ，インドネシア語のデータとともに Fine-tuning を行い，単言語モデルよりも精度が高

いことを示した．

3 XLM-RoBERTa について

本研究で用いる XLM-RoBERTa の前身となる XLM[5] は，2019 年に発表された言語横断的な事前学習済みモデルの一つである．まず，BPE 用いて全言語の を対象にサブワード分割を行う．この時，低リソース言語も高リソース言語と同程度のトークン数になるようにサンプリングする．こうすることで、多言語の事前学習を行うことができる．XLM-RoBERTa[6] は，2019 年に発表された，XLM を拡張した事前学習済みモデルである．日本語を含む 100 か国語，2.5TB のデータを使って言語モデルを学習している．XLM-RoBERTa では，テキスト分類や質問応答などの下流タスクで mBERT や XLM などの先行モデルよりも精度が向上している．このモデルの特徴として，100 の言語のうち例文が少ない言語と多い言語でサンプリングのパラメータを調整し，最適なパラメータを求めていることが挙げられる．

4 データセット

学習用データセットとして，工機ホールディング株式会社より頂いた実際のレビュー文セットを用いる．前処理として，実際の評点 1，2，3 は 0 に，4，5 は 1 に変更している．英語，日本語それぞれのレビュー文の件数を表 1，2 に示す．なお，metaboHPT，AmazonUSA，Amazon，楽天はそれぞれ EC サイトの名称であり，サイトに投稿されたドライバーに対するレビュー文をデータとして使用している．また，() 内はポジティブなレビューの件数である．

表 1 英語のレビュー文データセット

機種	metaboHPT	AmazonUSA	合計
A	293(288)	48(45)	341(333)
B	240(238)	15(14)	255(252)

5 実験

本研究では，日英の評判分析において，英語のレビューデータを対象に評判分析を行う場合と，日本語のレビューデータを対象に評判分析を行う場合の両方において，それぞれもう片方の言語のレビューが有効であるかを調べる．そのため，以下の 1〜4 の実験を行った．

表 2 日本語のレビュー文データセット

機種	Amazon	楽天	合計
C	-	150(147)	150(147)
D	-	53(52)	53(52)
E	150(124)	59(55)	209(179)
F	76(63)	57(54)	133(117)
G	107(90)	95(89)	183(179)
H	178(160)	608(474)	786(634)
I	-	502(442)	502(442)
J	70(66)	70(66)	140(132)

1. 英語を学習したモデルの英語のレビューの評判分析
2. 英語に加えて日本語データを学習したモデルの英語のレビューの評判分析
3. 日本語を学習したモデルの日本語のレビューの評判分析
4. 日本語に加えて英語データを学習したモデルの日本語のレビューの評判分析

train データ，validation データ，test データを表 3 のように設定し，学習率 [0.01，0.001，0.0001] それぞれの場合について 500 ステップごとに 2000 ステップまで AUC を計算する．また，AUC が最も高かったパラメータについては，test データを用いて AUC と正解率を算出する．また，プログラムは

表 3 実験に使用したデータ

実験 1 のデータ	
train	英語 (機種 A)
validation	英語 (機種 B/AmazonUSA)
test	英語 (機種 B/metaboHPT)
実験 2 のデータ	
train	英語 (機種 A)，日本語 (機種 C〜J)
validation	英語 (機種 B/AmazonUSA)
test	英語 (機種 B/metaboHPT)
実験 3 のデータ	
train	日本語 (機種 G)
validation	日本語 (機種 E/楽天)
test	日本語 (機種 E/Amazon)
実験 4 のデータ	
train	日本語 (機種 G)，英語 (機種 A，B)
validation	日本語 (機種 E/楽天)
test	日本語 (機種 E/Amazon)

XLM-RoBERTa を用いた二値分類プログラム [7] を参考に作成した．

6 結果

それぞれの実験の AUC を以下の表 4 に示す．

表 4 それぞれの実験の AUC

実験 1				
	Step			
学習率	500	1000	1500	2000
0.01	0.23	0.51	0.34	0.28
0.001	0.29	0.42	0.40	**0.77**
0.0001	0.34	0.53	0.41	0.41
実験 2				
	Step			
学習率	500	1000	1500	2000
0.01	0.65	**0.73**	0.72	0.44
0.001	0.58	0.56	0.59	0.59
0.0001	0.60	0.60	0.60	0.60
実験 3				
	Step			
学習率	500	1000	1500	2000
0.01	0.62	0.29	0.51	0.63
0.001	0.38	0.34	0.34	0.52
0.0001	0.53	0.60	0.64	**0.66**
実験 4				
	Step			
学習率	500	1000	1500	2000
0.01	0.55	0.73	0.74	0.75
0.001	0.42	0.48	0.62	0.63
0.0001	**0.84**	0.63	0.65	0.68

それぞれの AUC が最高値だったパラメータに対して test データを用いて正解率を算出した結果は表 5 の通りである．

表 5 実験 1〜4 のモデルの正答率

実験	train	validation	正答率 [%]
1	英	英	98.98
2	英，日	英	98.98
3	日	日	78.00
4	日，英	日	78.67

7 考察

7.1 英語に日本語を追加した実験の考察

実験より，英語のみで Fine-tuning を行った際の正解率と，それに日本語を加えて Fine-tuning を行った際の正解率は変化しなかった．今回使用した英語データセットのうち，ポジティブな評価のものは 341 件中 333 件である．また，テストデータの評判分析の結果としては全てポジティブと判定されている．以上より，これは訓練データの偏りが原因だと考えられる．正解率は 98.98% となっているが，ネガティブな文を抽出できていないということである．このような問題の改善点として，以下のようなものが考えられる．

- ネガティブなレビューとポジティブなレビューの偏りを少なくする． ネガティブなレビューとポジティブなレビューの偏りが少なくなるように訓練データをサンプ リングすることで，ネガティブなデータをモデルがより学習できるようにする．しかし，ネガティ ブな文自体のデータ数が少ない本実験の場合は，同じデータをモデルが何度も参照してしまうと いう問題がある．
- 評点を見直す，または評点の分類を増やす 今回の実験では，関連研究を参考に，レビューの評点 (星)1,2,3 は評点 0 に，評点 4,5 は評点 1 として二値分類を行った．この方法はネガティブ/ポジティブのみを分類し企業へのフィードバック とするためのものだが，今回の実験では訓練データの偏りが特に大きいため，1〜5 にレビューを 分類することでネガティブな文を抽出できるのではないかと考えられる．

7.2 日本語に英語を追加した実験の考察

実験より，日本語のみで Fine-tuning を行った際の正解率と，それに英語を加えて Fine-tuning を行った際の正解率は四捨五入し 0.7% 向上した．したがって言語横断的な機械学習によ り評判分析の精度が上がったといえる． 本実験では，テストデータのポジティブな文の割合が四捨五入して 82.67% であるのに対して 日本語のみの実験は 94.00%，日本語と英語のモデルは 96.00% の割合でポジティブな文章だと判定していた．英語のデータを追加することで，ポジティブな文と分類する割合は増えるが精度は高くなっていることがわかる． したがって，本実験においても，7.1 で考えられるような方法でより精度が向上するのではないかと考えられる．

8 おわりに

本研究では，実際のデータを用いて，XLM-RoBERTa による評判分析の精度を向上させることを目的とし，低リソースな英語のデータに追加をして高リソースな日本語のデータを学習する実験，低リソースな日本のデータに追加をして高リソースな英語のデータを学習する実験を行った．その結果，学習したデータに偏りが少ない対象言語においては，言語横断的な学習は評判分析に有効であった．今後は，偏りが大きい言語についても，モデルがネガティブな文を抽出するために学習したデータの偏りを減らすサンプリングや分類の方法を検討する必要がある．

謝辞

工機ホールディングス株式会社の西河智雅様，吉陽祐様，山口勇人様には，本研究のベースライン，評価データで使用するデータを準備していただきました．また，本研究は工機ホールディングス株式会社，茨城大学，東京農工大学の共同研究により行われたものです．この場を借りて御礼申し上げます．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[2] 三戸尚樹, 古宮嘉那子, 佐々木稔. 共学習によるデータ選別を利用した評判分析. pp. 1253–1257, 2022.

[3] 吉野弘泰, 古宮嘉那子. Bert を用いた言語横断型評判分析手法の比較. 人工知能学会全国大会論文集 第 36 回 (2022), pp. 3Yin226–3Yin226. 一般社団法人 人工知能学会, 2022.

[4] Andre Rusli and Makoto Shishido. On the applicability of zero-shot cross-lingual transfer learning for sentiment classification in distant language pairs.

[5] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. **arXiv preprint arXiv:1901.07291**, 2019.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. **arXiv preprint arXiv:1911.02116**, 2019.

[7] Dezs˝o Ribli. Train xlm-r large with tpu v2-8 on colab, 2020. https://www.kaggle.com/code/riblidezso/colab-train-xlm-r-large-with-tpu-v2-8-on-colab/data.