マルチラベル付き日本語レビュー文章の分類

1 はじめに

昨今の AI の発展は目まぐるしく, 様々な応用分野でその有用性が示されている. しかしその一方で, 機械学習手法はその判断根拠が不明な点が問題視されているため, 近年では人工知能に判断根拠の説明を示させる「説明可能な人工知能」に関する研究が盛んである. 自然言語処理の分野においても先述の研究は盛んであるため, このような研究に取り組むことを目標とし, 本研究ではラベルが付与された日本語レビュー文章の 2 値分類や多値分類を通して BERT モデルの理解や周辺知識の習得に努めた.

2 要素技術

2.1 BERT

BERT-[1] は 2018 年 10 月に Google の Devlin らの論文で発表された自然言語処理モデルである. Transformer というアーキテクチャを組み込み, 文章を双方向から学習することによって文脈を読み取ることが実現された事前学習モデルである. ファインチューニングをすることで様々な自然言語処理タスクに対応することが出来る汎用性の高さが注目を集めた. 本研究の実験では, 東北大学の乾・鈴木研究室によって公開されている BERT 日本語 Pretrained モデルを使用した.

3 関連研究

3.1 アスペクトベースの感情分析

アスペクトベースの感情分析-[2] では文章中に含まれるアスペクト情報を利用することで, その文章がどの様な事柄について書かれたものかを分析することが出来る. アスペクト情報-[3] とはその文章のカテゴリを表しており, 文章が何を対象としているかを示すエンティティと, その対象のどの属性について言及しているかを示すアトリビュートによって定義される. アスペクトベースの感情分析では 3 つのステップで分析を行う. 1 つ目は与えられたアスペクトカテゴリに文章を分類する. 2 つ目では文中に含まれるアスペクトカテゴリに対するフレーズの位置を推定する. 3 つ目でフレーズの極性を分析し, その精度を向上させることを目指す.

4 データセット

4.1 評判分析用チェックデータ

実験 1 では日本語レビュー文章とそれぞれの文章のラベルが与えられたデータを用いた. ラベルは, レビュー文章がネガティブなら 0, レビュー文章がポジティブなら 1 として与えられている. 総データ数は 6000 であった.

4.2 楽天トラベルレビュー : アスペクトセンチメントタグ付きコーパス

実験 2 では日本語レビュー文章とそれぞれの文章のラベルが 12 個与えられたデータを用いた. 総データ数は 76624 で, そのうち全ラベルが 0 であるデータは 28255 であり, 今回はこれらを除くことで少なくとも 1 つのラベルが立っているデータのみを用いた. その結果として総データ数は 48369 となった.

5 実験

本研究ではラベルが複数あるデータに対して多値分類をする為に, まずは 2 値分類をして分類手法の確認やシンプルな分類でどれほど正しく分類できるかの確認をした. 実験 1 では評判分析用チェックデータを用いた 2 値分類をした. そして実験 2 として, 楽天トラベルレビューのアスペクトセンチメントタグ付きコーパスを用いて多値分類をした. BERT モデルの末尾にネガポジ分類のための全結合層を追加し、出力として 2 クラス分類 [ネガティブ (0) or ポジティブ (1)] を出力するモデルを用いた. クラス分類には入力した文章データの 1 単語目 [CLS] の特徴量を利用した. また, BERTLayer の最終層と全結合層のみ fine-tuning を行った.

5.1 実験 1

実験 1 では評判分析用チェックデータを用いた 2 値分類をした．BERT の最大入力長は 512 トークンなのでそれを超える文章データは取り除いた．表 1 に該当するデータを取り除いた後の実験に用いたデータの内訳を示す．表 2 にそれぞれのデータに含まれるポジティブラベル数とネガティブラベル数を示す．表 3 に実験時のパラメータについて示す．

表 1: 2 値分類に用いたデータの内訳

総データ数	訓練データ数	テストデータ数
5638	4511	1127

表 2: ポジティブ, またはネガティブのラベルが付与されているデータの数

	ポジティブ数	ネガティブ数
訓練データ	2350	2161
テストデータ	573	554

表 3: 2 値分類タスクに用いたパラメータ

値	パラメータ
入力層の次元数	768
出力層の次元数	2
バッチサイズ	32
最適化関数	Adam
学習率	1e-4
損失関数	CrossEntropyLoss

クラス分類には文章データの 1 単語目 [CLS] の特徴量を識別器の入力として用いた．上記の訓練データを用いて 5 分割検証を行い, 5 個のモデルを作成した．表 4 には 5 個の中で最も正解率の高かったモデルを用いて表 1 のテストデータでの正答率, 再現率, F1 値, を求めた結果を示す．そして, 表 5, 図 1 には 5 個の中で最も正解率の高かったモデルとテストデータを用いた実験におけるコンフュージョンマトリクスを示す．表 6, 図 2 には交差検証を行わず, 訓練データをそのまま学習した場合のコンフュージョンマトリクスを示す．図 1 と図 2 を比較すると, 交差検証を行って正解率の高かったモデルを用いる方がより正しくデータを分類できていることがわかる．ここまでの手順やプログラムを踏まえて多値分類をした．

表 4: 最も正解率の高かったモデルを用いたテストデータでの正解率, 再現率, F1 値

正解率	再現率	F1 値
0.8669	0.8865	0.8713

表 5: 5 分割検証をした場合でのコンフュージョンマトリクス

		予測の分類結果	
		ポジティブ	ネガティブ
実際の分類結果	ポジティブ	469	85
	ネガティブ	65	508

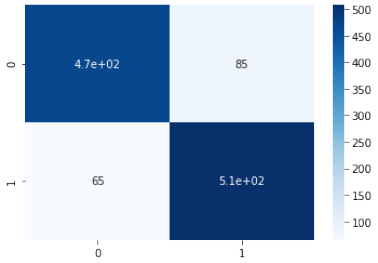

図 1: 5 分割検証をした場合でのコンフュージョンマトリクス

表 6: 訓練データをそのまま用いた場合でのコンフュージョンマトリクス

		予測の分類結果	
		ポジティブ	ネガティブ
実際の分類結果	ポジティブ	414	140
	ネガティブ	54	519

5.2 実験 2

実験 2 では楽天トラベルレビュー：アスペクトセンチメントタグ付きコーパスを用いた多値分類をした．表 7 に実験で用いたデータの内訳を示す．また, 表 8 に実験で用いたパラメータを示す．

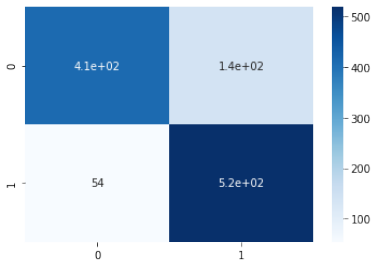

図 2: 訓練データをそのまま用いた場合でのコンフュージョンマトリクス

表 7: 多値分類タスクに用いたデータの内訳

総データ数	訓練データ数	テストデータ数
1200	960	240

表 8: 多値分類タスクに用いたパラメータ

値	パラメータ
入力層の次元数	768
出力層の次元数	12
バッチサイズ	12
最適化関数	Adam
学習率	1e-4
損失関数	BCEWithLogitsLoss

表 7 の総データを用いて 5 分割検証を行い, 5 個のモデルを作成した. ただし, ここでの正解率とは各クラスの正解率の平均値としている. また, 図 3, 4 に最も正解率の高いモデルにおける訓練データとバリデーションデータの損失と正解率の推移を示す. 図 3 の訓練データやバリデーションデータの損失はエポックが進むごとの小さくなっているため, 学習は進んでいると考えられる. また, 図 4 からは正解率もエポックが進むごとに向上していることがわかる. クラス分類には文章データの 1 単語目 [CLS] の特徴量を識別器の入力として用いた. 表 7 の通りにデータを分割し, 訓練データで学習してテストデータでの予測ラベルを求めた. 表 9 に全テストデータに対して, 全ラベルが正解したデータ数, 一部ラベルが正解したデータ数, 全ラベルが不正解であったデータ数とそれぞれの割合を示す. 全ラベルを正しく予測することに関して精度がかなり低くなっていることや, 立っているラベルを 1 つも正しく予測できなかったデータ数が全体の 4 割もあることから, 学習時のパラメータとしてより適切な値があると考えられる.

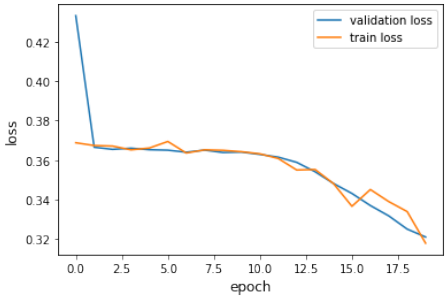

図 3: 訓練データとバリデーションデータの損失の推移

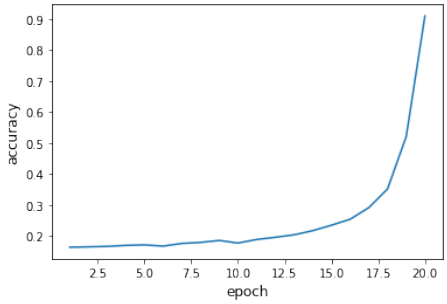

図 4: 正解率の推移

表 9: 予測データにおける立っているラベルの正解データ数

対象データ	データ数	割合
全テストデータ	240	1
全ラベルが不正解	85	0.3542
一部ラベルが不正解	59	0.2458
全ラベルが不正解	96	0.4

表 10 に正解率及び立っているデータに対してのみ適合率, 再現率, F1 値を示す. そして表 11 にコンフュージョンマトリクスを示す. 表 10 からは, 夕食 negative, 風呂 positive / negative, 立地 negative, 部屋 positive / negative の 6 クラスに関しては正解率が高く出ていることがわかる. しかし, 表 11 のコンフュージョンマトリクスでのその 6 クラスでの値を見ると, 夕食, 風呂, 立地, 部屋 negative の 4 クラスにおいて予測ラベルが 0 に寄っていることがわかる. 再び表 10 に戻り, F1 値を参照すると, これらの 4 クラスの値が非常に小さくなっていることがわかる. これらの原因として, 現時点ではこのクラスに属するデータ数の少なさが問題であったと考えられ

表 10: 予測データにおける各ラベルでの正解率, 適合率, 再現率, F 1 値

	夕食 po	夕食 ne	風呂 po	風呂 ne	サービス po	サービス ne	立地 po	立地 ne	設備 po	設備 ne	部屋 po	部屋 ne
正解率	0.8583	0.9583	0.9125	0.9583	0.8167	0.8333	0.8750	0.9833	0.8125	0.8792	0.9083	0.9208
適合率	0.8182	0.2500	0.5676	0.1818	0.6486	0.4324	0.7037	0	0.2895	0.4231	0.6829	0.3333
再現率	0.5806	0.3333	0.8076	0.6667	0.7273	0.4571	0.4634	0	0.3793	0.4400	0.7568	0.1875
F 1 値	0.6792	0.2857	0.6667	0.2857	0.4444	0.5588	0.5588	Nan	0.3284	0.4313	0.7179	0.2400

表 11: コンフュージョンマトリクス (多値分類の予測ラベルの分類)

	夕食 po	夕食 ne	風呂 po	風呂 ne	サービス po	サービス ne	立地 po	立地 ne	設備 po	設備 ne	部屋 po	部屋 ne
1 を 1 と当てた	36	2	21	2	48	16	19	0	11	11	28	3
0 を 0 と当てた	170	228	198	228	148	184	191	236	184	200	190	218
1 を 0 と間違えた	8	6	5	9	26	21	8	3	27	15	13	6
0 を 1 と間違えた	26	4	16	1	18	19	22	1	18	1	9	13

るが, 実際にどのようなフレーズからどのような極性が得られたのかを今後の研究で調査するべきであると考えた.

6 まとめと今後の課題

本研究ではアノテーションされたアスペクトベースのデータを用いて 2 値分類と多値分類をした. 2 値分類の正解率は 0.8669 で, F1 値は 0.8713 であった. 2 値分類においてはより精度の向上が目指すことが出来るのではないかと考えた. また, 多値分類におけるすべてのクラスの正解率の平均は 0.8930 であり, クラス 1 のみの F1 値の平均値は 0.4437 であった. ほとんどのデータにおいてラベルが立っていないクラスが多いため、予測が 0 に寄ってしまう問題があった. 多値分類における今後の課題として, 全体的な精度の向上とその手法の模索に取り組むこと, 今回用いたデータセット以外での実験を試みること, 一部正解データや全不正解データについて, 関連研究であるアスペクトベースの感情分析の手法を用いて分析することなどが挙げられる.

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] Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 486–495, 2015.

[3] 三浦義栄, 赤井龍一, 渥美雅保. 文中の複数アスペクトのセンチメント分析のための自己注意ニューラルネットワーク. 人工知能学会全国大会論文集 第 34 回全国大会 (2020), pp. 3Rin441–3Rin441. 一般社団法人 人工知能学会, 2020.