深層学習に基づく 4 コマ漫画の感情推定とマルチモーダル化への検討

1　はじめに

近年, 深層学習を始めとする機械学習技術の大きな発展を受けて, 人工知能を用いた創作物理解が注目されている. しかし, 創作は高次の知的活動であるため, いまだに実現が困難なタスクである. 人の創作物の理解に関する分野の中でもコミック工学 [1] など漫画を対象とした研究は, 絵と文章から構成される漫画を対象とするため, 自然言語処理と画像処理の両方の側面を持つマルチモーダルデータを扱う分野である. コミック工学の分野では様々な研究が報告されているが, その多くは画像処理に基づいた研究であり, 自然言語処理による内容理解を目指した研究は少ない. その大きな原因のひとつとしてデータが十分ではないという点が挙げられる. 漫画に含まれるテキストには, 口語表現, 擬音語, 表記揺れといった漫画特有の言語表現を含み, これらの扱いについて考慮する必要がある. そして, 漫画が著作物であることに起因する研究用データの不足も課題となっている.

本研究では人工知能を用いた漫画の内容理解のために, まず自然言語処理を用いた漫画のセリフの感情を推定して, その上で漫画のコマの画像情報を加えたマルチモーダル化について検討する.

2　研究用コミックデータ

4 コマ漫画を対象としたデータセットとしては Manga 109 [2] が知られているが, 漫画に登場するキャラクタの感情は明示されていない. そのために人手によるアノテートでラベルを付与する必要があるが, アノテートされたラベルが漫画家の意図とは異なる恐れがある. そこで, 本実験では上野によって作られた 4 コマ漫画ストーリーデータセット [3] を用いる. この 4 コマ漫画ストーリーデータセットは同一プロットの下, 幾人かの漫画家によって描き下ろされた 4 コマ漫画で構成されており, 作者によって感情ラベルがアノテートされている. また, 上野は異なる作者によって描かれた 4 コマ漫画を, そのタッチを基にギャグタッチ, 少女漫画タッチ, 少年漫画タッチ, 青年漫画タッチ, 萌えタッチと分類している.

図 1: タッチごとのアノテート例

る. 図 1 にこのデータセットにおける 2 つのタッチのコマ画像の例を示す. 図 1 より, 同一プロットであっても, 作者の感性によってアノテートされたラベルが異なっていることが分かる.

3　要素技術

3.1　BERT

BERT (Bidirectional Encoder Representations from Transformers)[4] は 2018 年に Google が発表した言語モデルの一種である.

その名前の通り, Transformer により構成される双方向のエンコーダであり, 文書分類や質疑応答といった様々な自然言語処理の幅広いタスクにおいて公開時点での最高性能を達成している. 従来のニューラルネットワークを用いた自然言語処理モデルは, 特定のタスクに対して 1 つのモデルを用いてきたが, BERT は転移学習により, 1 つのモデルで, 様々なタスクに対応できる. また, BERT を転移学習したモデルは, 少ないデータを追加学習するのみで動作するため, 1 からモデルを構築しなくて済む. 本研究では, 京都大学が公開している日本語 Wikipedia より全 1800 万文を用いて事前学習させたモデル [5] を用いた.

3.2　illustration2vec

illustration2vec [6] は Saito, Matsui らが提案した画像のベクトル化手法であり, Danbooru と Safebooru から 100 万枚のイラストを用いて学習した事前学

習済みモデルが公開されている. illustration2vec が扱った問題として, イラストに対する画像認識の難しさがあり, 既存の画像認識モデルのほとんどが ImageNet などの実画像を評価対象にしており, アニメや漫画といったイラストに対して評価をしていなかった. illustration2vec はそれらと比較してイラストのより合理的なベクトル化が期待できる手法である. 本研究では筆者らが公開している事前学習済みモデルを使って 4096 次元のコマ画像のベクトルを獲得する.

4 実験

4.1 Data Augmentation

4 コマ漫画ストーリーデータセットの欠点として, データ数が少ないことがあげられる. そこで, 本実験では日本語 WordNet [7] のシソーラスを用いてテキストデータを拡張する. 分かち書きされたオリジナルのセリフに対して, 日本語 WordNet で類似語を持つ単語について類似語に置き換え, 文を生成することでテキストデータを拡張した. ただし, 文の中に類義語を持つ単語が複数あった場合, 類似語に置き換える単語は同時に 1 つまでとし, 英数字・記号のみで表されている類似語は除外した. 例えば, 5 つの単語からなる文章があり, 各単語が 5 つの類似語を持っている場合, その文からは新しく 25 文が生成されることとなる.

4.2 設定

本研究では, 各タッチについての感情推定を行った. 使用するデータセットには 7 種類の感情ラベル（ニュートラル, 驚愕, 喜楽, 恐怖, 悲哀, 憤怒, 嫌悪）と, アノテーション不備によるラベル不明 (以下, “UNK” とする) の全 8 種類が含まれているが, データ数と解析の難しさの問題から, 今回はこの “UNK” のデータは除いた全 7 種類の感情ラベルが付いたデータのみを扱い, 喜楽を正例, その他の感情ラベルを負例とする 2 クラスに分類した.

訓練用データは各タッチの前半 1 話から 5 話までの拡張されたセリフを用い, 評価用データは後半 6 話から 10 話におけるオリジナルのセリフのみを用いた. 表 1 に各実験で用いたデータ数を示す.

多くのタッチにおいて, 正例は負例に対してデータ数が非常に少ない不均衡データであることから,

表 1: データ数

	ラベル	ギャグ	少女	少年	青年	萌え
train&valid	喜楽	15	39	15	18	25
	その他	40	26	45	44	35
	喜楽 (拡張後)	1115	2575	940	998	1766
	その他 (拡張後)	2851	1391	3076	3145	2323
test	喜楽	10	38	12	14	22
	その他	56	29	52	51	42

損失関数に使うクラス重みとして各タッチについて, 訓練用データの各ラベルのデータ数の逆数を正規化したものを用いた. また, 訓練用データの内 20 % をサンプリングして検証用データとし, 検証用データにおける正例の F 値が最大となるエポックのものを評価用モデルとして採用してモデルの性能を測った.

4.3 実験 1：セリフ 1 文の感情推定

実験 1 では 1 つのセリフを入力し, 対応する感情ラベルを出力するような感情推定をした. 識別器としては 3 層 MLP を用いた. 表 2 に MLP で用いたパラメータ, そして表 3 に学習で用いたパラメータを示す. 学習率は Optuna [8] によって最適なパラメータを探索した. JUMAN++ [1] によって分かち書きされたセリフをさらにサブワード化したものを BERT への入力とした.

BERT の事前学習済みモデルの全ての重みを固定した場合 (BERT fixed) と最終層のパラメータだけをチューニングした場合 (BERT fine tuning) の 2 つの手法について感情推定をして, 各タッチ・手法の結果について比較し, fine tuning の有用性を確かめた.

表 2: 実験 1・3 MLP パラメータ

parameters	実験 1	実験 3
(入力次元, 隠れ層次元, 出力次元)	(768,30,2)	(4864,300,2)
活性化関数	tanh	
ドロップアウト率	0.5	

表 3: 学習パラメータ

parameters	実験 1・2・3
エポック数	200
バッチサイズ	16
損失関数	Cross Entropy Loss
最適化手法	Adam

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

4.4 実験 2：過去のセリフを考慮した感情推定

実験 2 では, 連続したセリフ n 文を入力し, 末尾のセリフに対応する感情ラベルを出力する感情推定をした. 図 2 に $n = 3$ とした時のネットワークの概略図を示す. $\{s_i\}$ はそれぞれ連続する n 文のセリフを BERT の 単語 ID 列に変換したものである. 単語 ID 列はセリフ内の最大単語長に合わせて, 後方パディングをしている. また, 4 コマ漫画ストーリーデータセットは 1 話につき 2 つの 4 コマを含んでいるが, 各 4 コマは時系列的に繋がっていないことを踏まえて, s_i のセリフの組み合わせとしては同一の 4 コマに属し, かつ連続しているものを扱う. 各 4 コマの序盤に現れるセリフには参照できる過去のセリフが無いため, 便宜上のセリフ “[PAD]” を置くことで対処した. また, BERT への入力時における 1 文目, 2 文目を表す ID は奇数番目の入力では 0, 偶数番目の入力では 1 となるようにした. そして, 末尾のセリフ以外はオリジナルのセリフのみから抽出し, すべての組み合わせを入力列とすることで, 実験 1 のデータ数と合わせた.

単語 ID 列長を w, バッチサイズを b とすると入力次元は $(b \times n \times w)$ となる. このままでは BERT の入力次元に対応していないので, まず $(n \times b \times w)$ へと軸を入れ替え, これを 1 次元目について各ベクトルに分解し, これら n 個の 次元数 $(b \times w)$ のベクトルをそれぞれ BERT への入力とし, BERT の出力から [CLS] トークンに相当するベクトルのみをそれぞれ抜き取り, 先と逆の手順を踏むことで次元数 $(b \times n \times 768)$ のテンソルを得る. これを識別器として用いる Bi-LSTM をベースとした Self-Attention への入力とすることで末尾のセリフの感情を推定するように学習を行う. 本研究では n を 2 から 5 に変えて実験した. また, BERT は最終層のみをチューニングした.

表 3 に学習で用いたパラメータ, 表 4 に識別器のパラメータを示す.

表 4: 実験 2 識別器パラメータ

parameters	value
Bi-LSTM 入力次元	768
Bi-LSTM 隠れ層次元	128
Self-Attention 入力次元	128×2
Self-Attention ヘッド数	3

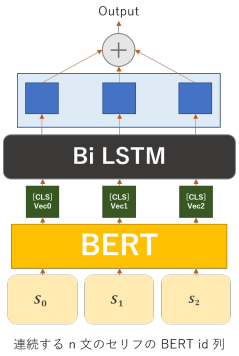

図 2: 提案ネットワーク $(n = 3)$

4.5 実験 3：マルチモーダルな感情推定の検討

実験 3 では, BERT から得た 768 次元のセリフベクトルと, 入力したセリフが含まれているコマ全体の画像を illustration2vec に入力して得た 4096 次元のコマベクトルを concat した 4864 次元のベクトルを入力ベクトルとして, 3 層 MLP に入力することでマルチモーダルな感情推定をした. また実験 1 と同様にして, BERT の事前学習済みモデルの全ての重みを固定した場合 (Multi fixed) と最終層のパラメータだけをチューニングした場合 (Multi fine tuning) の 2 つの手法について感情推定をして, 各タッチ・手法の結果について比較した.

5 実験結果および考察

実験 1, 2, 3 において, すべての出力が負例と推定された場合の評価指標の値をベースラインとして設定した. 以下, 表における P-Recall, P-F 値は正例の再現率, F 値を表し, Acc は全体の精度を表す. また, 5 タッチ総合の結果は各タッチにおける混合行列の和から算出している.

5.1 実験 1：セリフ 1 文の感情推定

表 5 に実験 1 の結果を示す. 表 5 より, Acc に関して BERT fine tuning のみがベースラインを超え, すべての評価指標において BERT fixed より BERT fine tuning の方が上だったことから, 本研究で扱っている 4 コマ画像におけるセリフの感情推定というタスクにおいても BERT の fine tuning が有用であることを確かめることができた.

5.2 実験 2：過去のセリフを考慮した感情推定

表 6 に実験 2 の結果を示す. 表 6 より, 実験 1 と比べ, 全体的に評価指標は下回ったことから, 過去のセリフを考慮することの優位性を示すことはできなかった. その理由としてはモデルの設計上, BERT から得た分散表現が文同士の繋がりを学習できておらず, 逆に BERT の強みを活かしきれていない可能性があることや, ネットワーク構造の複雑化によって過剰適合が進んだことが原因だと考えられる. 今後の改良のために学習率のスケジューラーを設定したり, 適宜ドロップアウト層を追加するなど, 様々な工夫をする必要があると考えられる. しかし, n が大きくなるに従って P-Recall, P-F 値は増加傾向にあり, P-Recall に関しては n が 3 以上の時, 実験 1 の結果を上回っていることから過去のセリフをより多く考慮することに有用性はあると推測できる.

5.3 実験 3：マルチモーダルな感情推定の検討

表 7 に実験 3 の結果を示す. 表 7 より, まず実験 1 と同様にすべての評価指標において Multi fixed より Multi fine tuning の方が上だったことから, BERT の fine tuning の有用性を確かめることができた. また, 実験 1 と比べて Acc は下回ったものの, P-Recall, P-F 値は増加し, 特に少女漫画タッチと少年漫画タッチにおいては実験 1 と比較すると最も高い P-Recall と P-F 値が得られており, マルチモーダルな感情推定の意味があると推測できる. しかし, コマの画像ベクトルがセリフの分散表現に対して次元数がかなり多いことを踏まえると, 全結合層などを追加し, 次元数を合わせる必要もあると考えられる.

5.4 実験全体に関する考察

まず, 本研究における結果は不均衡データに対して得たことから識別器としては無意味な負例のみを選んだ場合をベースラインとしたので, Acc に関してはベースラインを大きく超えることはできなかった. しかしながら, その他の評価指標である P-Recall, P-F 値に関しては良好な結果が得られたと考えられる.

次に, データ及びモデルの問題点について考察する. 本研究では Data Augmentation の手法における文法的意味の齟齬に関する問題については取り組んでおらず, この問題がどれだけ寄与しているかについての調査は今後の課題である. 具体的な解決策としては, 拡張されたセリフの分散表現とそれぞれに対応するオリジナルのセリフの分散表現とのコサイン類似度から閾値未満のデータを除外したり, 拡張する品詞を限定するといったことが考えられる. また, 本研究で用いた BERT の事前学習済みモデルでは, 'A' や 'B' 等のアルファベットは未知語として扱われている. 4 コマ漫画ストーリーデータセットにおいてはこれらの記号が主役であるキャラクタ名に充てられていることから, 主語の情報が抜け落ちてしまい, 会話の流れを上手く学習できていない可能性がある. その対策として, 発話者情報を加味したり, セリフ内のキャラクタ名を特殊なトークンに置き換えるなどの工夫が必要である. そして, アノテーション不備により "UNK" ラベルが付与されているデータの扱い方についても考える必要がある.

6 まとめと今後の課題

本研究では BERT で得たセリフの分散表現からセリフの感情の推定をした. 実験結果から過去のセリフを考慮することの優位性を示すには多くの課題があることが分かった. また, コマの画像のベクトルも考慮したマルチモーダルな感情推定もした. 1 つのセリフを入力とする実験では正例の識別能力は向上したが, 全体的な優位性を示す結果は得られなかった.

今後の課題として, 現在の 4 コマ漫画ストーリーデータセットのみでの訓練・テストでは精度の向上が困難であることから, データセットの拡張が急がれる. この問題の解決策としては Manga109 やその他データセットを併用した半教師あり学習や, 人手による新しいデータの作成が挙げられる. データ作成の際にはオリジナルのデータからセリフ部分の文字を白抜きにしたコマの画像に対して, 物語の一貫性などの制約を設けた上で新たなセリフを作成し, 対応する感情ラベルを付与してもらうことで, このデータセットの "作者による感情ラベルのアノテーション" という特徴を保ったままデータ拡張が可能であると考えられる.

また他の BERT の事前学習済みモデルとして, hottolink 社が公開している大規模日本語 SNS コーパスによって事前学習させた hottoSNS-BERT [9] について注目していく. 本研究で現在用いている事前学習済みモデルは主に文語的なコーパスを用いて学習している. hottoSNS-BERT は口語的な SNS の投稿

を扱っていることから漫画のセリフとの親和性が高く, より合理的な分散表現が得られることが期待できる. セリフから得たベクトルとコマの画像から得たベクトルの結合方法やネットワークの構造の最適化についても更なる工夫が必要である.

参考文献

[1] 松下 光範. コミック工学：マンガを計算可能にする試み. **日本知能情報ファジィ学会 ファジィ システムシンポジウム 講演論文集**, 29:199–199, 2013.

[2] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

[3] 上野 未貴. 創作者と人工知能: 共作実現に向けた創作過程とメタデータ付与 4 コマ漫画ストーリーデータセット構築. **人工知能学会全国大会論文集**, 2018.

[4] Chang M.-W. Lee K. Devlin, J. and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[5] 黒橋 禎夫 柴田 知秀, 河原 大輔. Bert による日本語構文解析の精度向上. **言語処理学会 第 *25* 回年次大会**, pages pp.205–208, 2019.

[6] Masaki Saito and Yusuke Matsui. Illustration2vec: A semantic vector representation of illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*, SA '15, New York, NY, USA, 2015. Association for Computing Machinery.

[7] Richard Fothergill Francis Bond, Timothy Baldwin and Kiyotaka Uchimoto. Japanese semcor: A sense-tagged corpus of japanese in the 6th international conference of the global wordnet association (gwc-2012), matsue. 2012.

[8] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[9] Sakaki, Takeshi, Sakae Mizuki, and Naoyuki Gunji. Bert pre-trained model trained on large-scale japanese social media corpus. 2019.

表 5: 実験 1 の結果 (評価用データ)

	ギャグ			少女漫画			少年漫画			青年漫画			萌え			5 タッチ総合		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
BERT fixed	0.712	0.200	0.174	0.567	0.632	0.623	0.766	0.083	0.118	0.692	0.643	0.474	0.594	0.591	0.500	0.666	0.510	0.473
BERT fine tuning	0.818	0.200	0.250	0.612	0.711	0.675	0.766	0.083	0.118	0.862	0.500	0.609	0.609	0.591	0.510	0.733	0.521	0.535
ベースライン	0.848	-	-	0.432	-	-	0.812	-	-	0.784	-	-	0.656	-	-	0.705	-	-

表 6: 実験 2 の結果 (評価用データ)

n	ギャグ			少女漫画			少年漫画			青年漫画			萌え系			5 タッチ総合		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
2	0.712	0.200	0.174	0.493	0.684	0.605	0.781	0.083	0.125	0.815	0.286	0.400	0.531	0.364	0.348	0.666	0.427	0.429
3	0.621	0.400	0.242	0.507	0.763	0.637	0.750	0.167	0.200	0.785	0.357	0.417	0.516	0.591	0.456	0.635	0.552	0.471
4	0.636	0.300	0.200	0.537	0.816	0.667	0.766	0.167	0.211	0.800	0.357	0.435	0.500	0.591	0.448	0.647	0.563	0.484
5	0.652	0.500	0.303	0.507	0.895	0.673	0.781	0.333	0.364	0.754	0.500	0.467	0.594	0.591	0.500	0.656	0.656	0.529
ベースライン	0.848	-	-	0.432	-	-	0.812	-	-	0.784	-	-	0.656	-	-	0.705	-	-

表 7: 実験 3 の結果 (評価用データ)

	ギャグ			少女漫画			少年漫画			青年漫画			萌え			5 タッチ総合		
	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値	Acc	P-Recall	P-F 値
Multi fixed	0.758	0.000	0.000	0.627	0.553	0.627	0.391	0.750	0.316	0.554	0.500	0.326	0.563	0.273	0.300	0.580	0.448	0.386
Multi fine tuning	0.773	0.200	0.211	0.687	0.763	0.734	0.703	0.417	0.345	0.769	0.643	0.545	0.641	0.500	0.489	0.715	0.583	0.546
ベースライン	0.848	-	-	0.432	-	-	0.812	-	-	0.784	-	-	0.656	-	-	0.705	-	-