RINA:マルチモーダル情報を利用したキャラクターの感情推定

賴 展韜　　髙橋 誠史
株式会社バンダイナムコ研究所
z-lai@bandainamco-mirai.com m3-takahashi@bandainamco-mirai.com

概要

本研究では，テキストからキャラクターの感情を推定するアプローチについて考察した．マルチモーダル情報を含むゲームシナリオデータセットにおいて，本研究で提案する Reference Information Normalize Adapter (RINA) モデルを用いてマルチクラスの感情推定の精度が既存手法に比べて向上することを実験を行い示した．また，結合手法を変更することで，非テキスト情報を取り込む方法を比較し，Transformer の入力トークンとして直接追加するよりも，テキスト情報の分散表現ベクトルに対し補正を行う方が優位であることを示した．

1　はじめに

テキストから感情を推定することは自然言語処理における主要な課題のひとつである．BERT[1] をはじめ，さまざまな事前学習済みモデルを用いた感情推定の手法が対話システムやソーシャルメディアマイニングなどの分野で活発に研究されている [2]．

現実世界の人間に対する感情推定と同じように，ゲームなどのフィクション作品内のキャラクターに対する感情分析の研究も注目されている．近年の技術発展とともに，ゲームの表現は高度に多様になり．ゲームの開発コストが膨れ上がる結果となった．特に多数のキャラクターが登場するゲームにおいて，キャラクターのセリフにあわせて適切な感情表現や演出を指定するタスクが手作業で行われているため，現状多大な工数がかかっている．この課題を解決するには，自然言語理解によるキャラクターの感情推定の作業支援が必要である．

本研究では，フィクション作品内のキャラクターのセリフから感情推定について取り組む．現実世界の人間との違いとして，キャラクターに対する感情推定の特徴は以下 2 点が挙げられる．

一つ目は，キャラクターによる感情表現の分布が人間と異なることである．コメディ作品のキャラクターの感情表現が明るい傾向を示すことに対し，シリアスな作品のキャラクターはネガティブな感情を表現する場合が比較的多い．また，物語を通して多様な場面で活躍するメインキャラクターでは豊かな表現を持つことに対し，物語の一部でしか登場しないサブキャラクターはその個性のイメージをより早くユーザに定着させるため，感情表現が偏るステレオタイプである可能性が高い [3]．作品の基調，キャラクターの設定，または物語での役割に応じて感情表現の分布が変化することがあるが，逆にこれらの情報は事前に開示されることが多いため，補助情報として参照することで，キャラクターの感情表現をより正確に把握することができる．

二つ目は，キャラクターの感情を推定する際，文脈に対する依存度が人間より高い点である．本研究で指す「文脈」とは，感情推定の対象文前後のセリフなどのテキスト情報だけでなく，発話シーンの背景グラフィックや音楽の種類などの非テキスト情報を含む．依存度が人間より高い原因として，キャラクターのセリフは演出の都合や，画面で表示できる文字数などのインターフェイスによる制約があるため，現実世界の言葉より断片的であることが多い．感情推定の対象文として利用する場合，現実世界の人間の発話に比べて情報量が少ないセリフは，文脈情報を参照せず感情表現を正確に推定することが困難である．

本研究では，ゲームのシナリオスクリプトを題材として，テキストを主体とした対象文に，非テキストの参照情報を結合することにより，キャラクターの感情推定の精度を向上させることを試みた．その際複数の結合方法の効果について実験的に調査した．本研究で提案した Reference Information Normalize Adapter (以下 RINA という) モデルを用いて，テキスト情報の分散表現ベクトルに対し参照情報による補正を行うことで，従来手法より推定精度の向上を確認した．

2 関連研究

テキストにおける感情推定は，マーケティング，心理学などの分野における膨大な応用の可能性から，これまでさまざまな手法が提案されている [2].

Raskin ら [4] はフィクション作品に注目し，物語中の人物の心的状態を推定するためのデータセットとして，Story commonsense データセットを構築した．Story commonsense データセットを対象した感情推定に関する研究として，田辺ら [5][6] はコモンセンス知識を導入し，さらに推定対象文の先行文脈を対象文と連結させることによって，推定精度を向上させた．Gaonkar ら [7] は感情カテゴリ間の関係を考慮し，ラベル同士の共起関係の利用することでラベルなしデータによる半教師あり学習の手法を提案した．これらの研究は，感情推定の対象文と関連するテキスト情報の利用に焦点を当てており，非テキスト情報を利用していない．

非テキストを含むマルチモーダル情報を扱うことを目的としたモデルはこれまで様々なものが提案されている．VilBERT[8] や VLBERT[9] などのモデルは，BERT の構造をそのまま流用し，非テキストの画像情報を入力の追加トークンとして利用している．しかし，これらのモデルのタスクは本来テキスト情報のみ利用した BERT の事前学習タスクと大きく異なるため，画像とテキストのマルチモーダルデータを用いて改めて事前学習が必要である．上記課題の解決案として，Kiela ら [10] は画像情報を追加の入力トークンとして扱いながら，事前学習済みの BERT モデルを直接利用し，マルチモーダルデータをファインチューニング段階のみ利用する MMBT モデルを提案した．これまでのマルチモーダルデータセットではテキストと非テキスト情報の重みが対等であることが多かった．しかし，本研究の対象する感情推定において，補助という位置付けの非テキスト情報に過大の重みを付けると，事前学習済みモデルによる分散表現に悪影響を及ぼす可能性がある．これに対し Rahman ら [11] は画像・音声によるテキストにおける表現空間のベクトルを補正する手法を提案した．また，Gu ら [12] は複数のテキストと非テキスト情報を結合するモデル構造に対するベンチマークを行った．

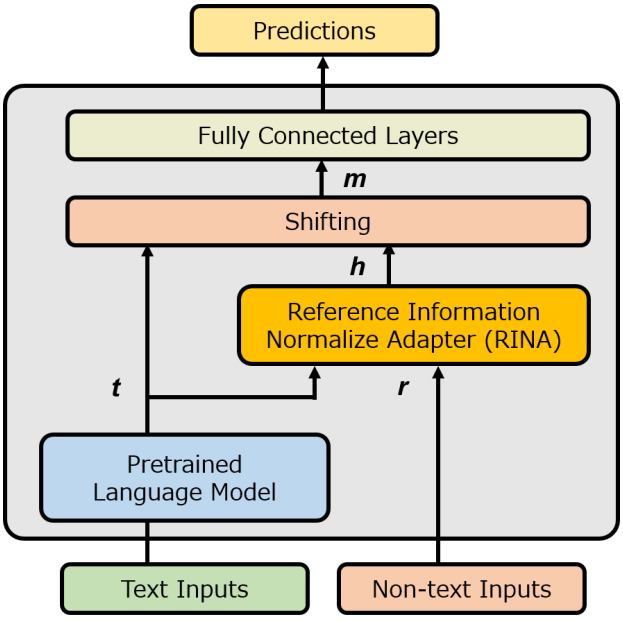

図 1 提案手法のモデル構造

3 提案手法

本研究の提案手法のモデル構造を図 1 に示す．まずテキスト情報同士を連結し，BERT などの事前学習済みモデルによるエンコードを行った．その後，RINA モデルを用いて非テキスト情報を結合させた後，複数の全結合層を経て出力し．出力されたシーケンスの [CLS] トークンのベクトルの分散表現を感情推定の対象として利用した．

3.1 テキスト情報同士の結合

テキスト情報同士の BERT など事前学習済みモデルへの入力形式は，先行研究 [5] を参考し，以下のように対象文と関連文を記号「|」により区切る形式とした．

$$[CLS] \text{ 対象文 | 関連文 } [SEP] \tag{1}$$

3.2 テキスト情報と非テキストの結合

本研究は以下 2 つのコンセプトに基づいて Reference Information Normalize Adapter (RINA) モデルを提案する．

マルチモーダル情報の結合は Transformer モデル出力の後に行う． 既存のマルチモーダルモデルの構造は Transformer の中間層にモダリティを跨ぐ Attention 機構を組み込むことが多いが，本研究では BERT[1] や XLNet[13] など Transformer の内部構造に

依存しない特性を重視し，Transformer の出力の後にモダリティの結合を行う Adapter 型の構造 [14] を利用した．

非テキスト情報を考慮しテキストの分散表現ベクトルを補正する． 感情推定の対象文の意味は分散表現としてベクトル空間で表現しているが，非テキスト情報を付随条件として与えられた際，対象文の意味の変化はベクトル空間上の変位として表現する．このテキストと非テキストの関係性を反映した変位ベクトルによって新しい分散表現を決定する．

$$
\begin{aligned}
\boldsymbol{m} &= \boldsymbol{t} + \alpha \boldsymbol{h} \\
\boldsymbol{h} &= \boldsymbol{g} \odot (\boldsymbol{W}\boldsymbol{r}) + b_h \\
\alpha &= \min\left(\frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{h}\|_2} * \beta, 1\right) \\
\boldsymbol{g} &= \mathrm{R}(W_g(\boldsymbol{r}||\boldsymbol{t}) + b_r)
\end{aligned}
\tag{2}
$$

ここで，t はテキスト情報の入力行列，r は非テキスト情報の入力行列，b はスカラーバイアス，W は重み行列，R は活性化関数を表す．β はゲーティング機構 [11] のスケーリング係数を表す．β を導入する目的として，クロスバリデーション段階でチューニングすることにより，非テキスト情報による補正の度合いを調整可能とする．

4 実験

4.1 データセット

これまで先行研究では Story commonsense データセット [4] を採用するものが多いが，非テキスト情報が付与されていないため，本実験では独自のゲームシナリオスクリプトをデータセットとして用いた．データセットでは 3 シーズンに分かれており．シーズン 1 からシーズン 2 を学習データとして使用し，シーズン 3 の約 50%をハイパーパラメータ最適化用にバリデーションデータ，残り 50%をテストデータとして評価を行った．データ量は学習データ 19199 件，バリデーションデータ 4512 件，テストデータ 4484 件である．データに含まれる説明変数について，テキスト情報では対象文となるセリフと文脈のセリフ計 2 種類，非テキスト情報ではキャラクターの名前，性別などの属性や，対象文が発生するシーンの ID 情報を含め計 9 種類を用いる．推定対象となる目的変数はゲーム内対象文が発話時に表示されるキャラクターの表情画像の種類を用いる．表情画像は「通常」，「怒り」，「悲しみ」，「驚き」，「笑い」計 5 種類である．

表 1 実験で利用する各手法のマルチモーダル情報の結合方式

	結合方式
BERT text only	$\boldsymbol{m} = \boldsymbol{t}$
BERT w/concat	$\boldsymbol{m} = \boldsymbol{t}\|\|\boldsymbol{r}$
BERT w/MLP	$\boldsymbol{m} = \boldsymbol{t}\|\|\mathrm{MLP}(\boldsymbol{r})$
BERT w/RINA (Ours)	$\boldsymbol{m} = \boldsymbol{t} + \alpha\boldsymbol{h}$

4.2 ベースライン

提案手法の比較対象とするベースライン手法は表 1 で示す．BERT text only はテキストのみ利用するモデルである．BERT w/concat は非テキスト情報を Transformer の追加トークンとして入力するモデルとして，BERT w/MLP は非テキスト情報を多層パーセプトロンに入力し，その出力を Transformer の出力と結合するモデルを表している．

4.3 訓練設定

事前学習済みモデルは東北大学が公開した BERT base Japanese (unidic-lite with whole word masking, jawiki-20200831)[1)] を使用した．事前学習で使用した Wikipedia データセットと今回の対象となるゲームシナリオのドメインが離れていたため，データセットに存在し，BERT の Tokenizer 辞書に存在しない頻出 n-gram の上位 179 件を辞書に追加した上，Gururangan ら [15] が提案した Domain-Adaptive Pretraining を用いて本データセットの全テキストに対して 5 Epoch の追加事前学習を行った．追加事前学習ではバッチサイズを 64，トークンのマスク率を 15%に設定した．

追加事前学習済みのモデルを用いてベースラインと提案モデルを実装する際のハイパーパラメータは以下で示す．

- Epoch 数: 10
- 学習率 α: 3×10^{-5}
- バッチサイズ: 64
- 最大入力長: 128
- Dropout 率: 0.2
- 活性化関数 R: ReLU
- ゲーティング機構のスケーリング係数 β: 0.4

1） https://huggingface.co/cl-tohoku/bert-base-japanese-v2

4.4 評価指標

マルチクラス分類タスクの評価指標として $F1_{micro}$ スコアと $F1_{macro}$ スコアを用いる．

4.5 実験結果

実験の結果を表 2 で示す．本研究が利用するシナリオスクリプトデータセットにおいて，$F1_{micro}$ は全体件数に対する推定精度，$F1_{macro}$ は各クラスに対する推定精度のバランスを反映した．全体のスコアの傾向として，データセット内の表情画像が「通常」と表示されるセリフは約 40%を占めているため，件数が少ない「怒り」，「悲しみ」，「驚き」，「笑い」クラスの推定精度は「通常」クラスより低い傾向を示している．また，各手法の $F1_{macro}$ と $F1_{micro}$ の変化は同じ傾向を示している．

各手法の非テキスト特徴を取りこむ有効性の比較において，非テキスト情報を利用しない BERT text only が最も推定精度が低いモデルである．これはデータセットの非テキスト情報がキャラクターの感情推定に寄与することを示唆している．

テキスト情報と非テキスト情報を結合させる手法の中，BERT w/MLP では $F1_{micro}$ において BERT w/concat を上回るが，$F1_{macro}$ においては BERT w/concat より低いことから，非テキスト情報を Transformer 入力段階のトークンとして追加する手法と，Transformer 出力で結合させる手法に大きな差が見られないことが明らかになった．

提案手法の BERT w/RINA について，$F1_{micro}$ と $F1_{macro}$ 2 つの指標が他の手法と比較して最も高いスコアを確認している．これは今回使用しているデータセットにおいて，RINA を利用したテキスト情報の分散表現ベクトルを補正する手法の有効性を証明した．

4.6 考察

本研究のタスクである感情推定と，本データセットのテキストが主体情報，非テキストが補助情報という特徴が，前節で述べた RINA のコンセプトに適しているが原因であるが，非テキスト情報がタスクにおける重要度がテキスト情報と同等もしくは以上の場合，他の手法がより推定精度が高い可能性があるため，今後の課題としてデータセットのラインアップを追加し検証したい．

表 2 各手法の表情クラス推定精度の比較

	$F1_{micro}$	$F1_{macro}$
BERT text only	0.6028	0.3698
BERT w/concat	0.6555	0.4294
BERT w/MLP	0.6564	0.4270
BERT w/RINA (Ours)	0.6718	0.4376

5 おわりに

本研究では，ゲームのシナリオスクリプトに注目し，テキストと非テキスト情報を含むマルチモーダルデータセットを対象として，セリフ発話時のキャラクターの感情を推定する手法の提案した．今回提案する RINA モデルを用いて，テキスト情報の分散表現ベクトルに対し参照情報による補正を行うことで，従来手法より推定精度の向上を確認した．

また，今回利用する非テキスト情報はカテゴリや数値形式のデータに限定しているが，今後の取り組みとして，表情画像を直接入力として利用するなど，多様なモダリティに対応できるよう RINA モデルの構造の改善を行いたい．

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. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. Emotion detection in text: a review. **ArXiv**, Vol. abs/1806.00674, , 2018.

[3] 金水敏. 役割語研究の展開. くろしお出版, 2011.

[4] Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. Modeling naive psychology of characters in simple commonsense stories. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2289–2299, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[5] Hikari Tanabe, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoshihiko Hayashi. Exploiting narrative context and a priori knowledge of categories in textual emotion classification. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 5535–5540, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[6] 田辺ひかり, 小川哲司, 小林哲則, 林良彦. コモンセンス知識を利用した物語中の登場人物の感情推定. 言語処理学会第 27 回年次大会発表論文集.

[7] Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjan Balasubramanian, and Nathanael Chambers. Modeling label semantics for predicting emotional reactions. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4687–4692, Online, July 2020. Association for Computational Linguistics.

[8] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[9] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. **CoRR**, Vol. abs/1908.08530, , 2019.

[10] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. **ArXiv**, Vol. abs/1909.02950, , 2019.

[11] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2359–2369, Online, July 2020. Association for Computational Linguistics.

[12] Ken Gu and Akshay Budhkar. A package for learning on tabular and text data with transformers. In **Proceedings of the Third Workshop on Multimodal Artificial Intelligence**, pp. 69–73, Mexico City, Mexico, June 2021. Association for Computational Linguistics.

[13] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. **CoRR**, Vol. abs/1906.08237, , 2019.

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019.

[15] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. **CoRR**, Vol. abs/2004.10964, , 2020.