# 漫画のテキスト分散表現と発話者情報の統合によるコマ順序推定

## 1　はじめに

近年, 多様な分野において人工知能や機械学習を用いる試みが盛んに研究されている．人の創作物理解にも機械学習が適用され, 大きな発展が見られている．特にコミック工学に代表される漫画を対象とした研究は, 画像処理と言語処理が密接に結びついたマルチモーダルな分野として注目されている．人間が漫画を読む際には, 画像から言語やキャラクターなどの様々な情報を抽出, 統合して読んでいると考えられる．漫画は画像データで提供されることが多いため, 情報の抽出に関する研究は画像処理に基づいて数多くなされてきたが, 情報の統合に関する研究は殆どなされてこなかった．人間はキャラクター同士の関係性や時代背景などを加味して漫画を読み進めていくため, 情報の統合は漫画の内容理解という観点からは重要であると考えられる．そこで, 本研究では, 漫画の内容理解を目的とし, 情報の統合を漫画の内容理解のタスクとして設定する．

統合される情報の組み合わせとして,「テキスト」と「発話者」が挙げられる．例えば, 会話文の順番は, 発話者が自然に入れ替わるようにテキストを読み進めると考えられる．また, 効果音などの特定の発話者が存在しないテキストは読み飛ばされる傾向がある [1] ことから, ストーリーへの影響力が小さいことが多いと言える．また, ナレーションに相当するテキストは話の大枠を説明する役割を担うため, ストーリーへの影響力が大きいと言える．このように, テキストと発話者の組み合わせはストーリーの進行を考える上で重要であると考えられる．

これを踏まえて, 内容理解に対する問題として「テキスト」と「発話者」の情報の統合による 4 コマ漫画の前後半コマの順序推定を設定する．

## 2　Manga109

一般に公開されている漫画のデータセットとして「Manga109」[2, 3] がある．これは漫画の研究のために相澤らにより作られたもので, 漫画 109 冊の画像データに加え, 登場人物の名前, 台詞文, 画像内における登場人物の顔, 全身, コマ, 台詞の座標などのアノテーションデータが含まれている．Manga109 ではすべての漫画の画像データは見開き 2 ページ 1 枚の形式で保存されており，画像上での台詞やコマを含む矩形の座標情報がアノテーションによって付与されている．画像上の座標平面は左上を原点とし，右向きに x 軸，下向きに y 軸が設定されている．座標情報は台詞, コマを囲う矩形の x 座標の最小値，y 座標の最小値，x 座標の最大値，y 座標の最大値の 4 次元で表される．

本稿では, Manga109 を対象としたアノテーションツールを開発し, 見開き 2 ページ内の台詞およびコマ全てに対して一意の通し順位と, それぞれの台詞, コマが 4 コマ漫画のものであるか否かのメタデータを付与した．また, 発話者情報としては, Manga109 の「セリフ発話者データセット」 [4] を使用した．

## 3　要素技術

### 3.1　LSTM

LSTM [5] は, 時系列性の有するデータを学習できる Recurrent Neural Network (RNN) の一種である．RNN は, 前回の出力を現在の入力に追加して再帰的に入力するものであり, 時間方向に展開すると静的なニューラル・ネットワークと見ることができる．RNN は, この時間方向に展開した静的なネットワーク上で誤差逆伝播を用いて学習を行うもであるが, 系列が長くなると重みが掛けられる回数が多くなり, 勾配が消失してしまう．これによって, 長期依存を学習できるなくなるという問題が生じる．これに対し, LSTM では重みを掛けずに誤差を逆伝播させることによって, 長期依存を学習できなくなる問題を解消している．

### 3.2　BERT

BERT [6] とは Transformer による双方向のエンコーダーであり、2018 年 10 月に Google が発表した言語モデルである．文章分類, 質問応答, 固有表現抽出等の多様なタスクで公開当時の最高性能を達成するといった大きな成果が報告されている．これまでの言語モデルでは特定の学習タスクに対して 1 つのモデルを用いてきたが、BERT は転移学習により, 1 つのモデルをチューニングすることで、様々な問題に対応することができる．さらに, 以前はモデル毎に語彙を 1 から学習させるため, 非常に多くの時間とコストがかかってい

表 1: 実験データ

| | OL Lunch | KoukouNoHitotachi | Akuhamu | TetsuSan | YouchienBoueigumi | total |
|---|---|---|---|---|---|---|
| 話数 | 201 | 227 | 185 | 136 | 91 | 840 |
| 最大テキスト長 $t$ | 8 | **13** | 8 | 9 | 10 | 13 |
| 発話者数 $m$ | 10 | **111** | 26 | 36 | 10 | 111 |

表 2: 実験条件 MLP

| | テキスト | 発話者 | テキスト+発話者 |
|---|---|---|---|
| 中間層の unit 数 | 97 | 75 | 27 |
| activation | sigmoid | relu | relu |
| optimizer | Adam | Adam | Adam |
| 学習率 | 7.24E-4 | 1.9E-3 | 8.21E-4 |
| epoch | 50 | 50 | 50 |

たが, BERT ではオープンソースで公開されている文脈を既に学習させた Pre-Training BERT モデル を使用することで短時間で学習ができる. 本稿では, 日本語 Wikipedia より 全 1,800 万文を用いて事前学習させたモデル[1]を利用した.

# 4 提案手法

## 4.1 前後半コマのテキスト分散表現

4 コマ漫画の各エピソードのコマを前 2 コマ, 後ろ 2 コマに分割し, それぞれを前半コマ, 後半コマとする. 前半コマおよび後半コマにそれぞれ 1 文以上テキストが存在するものとする前後半コマを対象として, それらのコマに所属するテキストを抽出する. 次に, 抽出したテキストを 1 文ずつ $n$ 次元で分散表現化し, 読み順に連結する. この時, 最大テキスト長 $t$ に合わせて後方ゼロパディングする. このようにして得られた $t \times n$ 次元のベクトルをそれぞれ前半コマテキスト分散表現, 後半コマテキスト分散表現とする.

## 4.2 発話者ベクトル

まず, 作品中の $m$ 人の登場人物に対して「その他」, 「ナレーション」, 「不明」の 3 つの項目を加えたものをカテゴリ変数としてキャラクターを one-hot 表現する. このときカテゴリ変数の並びは「その他」, 「ナレーション」, 「不明」+ 作品中の $m$ 人の登場人物を登場回数で降順に並べたものとする. 次に, テキストの発話者として記録されているキャラクターの one-hot ベクトルを足し合わせ, これを発話者ベクトルとする. 例えば, テキストの発話者が [1, 0, 0, 0, 0, 0] と [0, 0, 1, 0, 0, 0] の場合, このテキストの発話者ベクトルは [1, 0, 1, 0, 0, 0] となる.

## 4.3 前後半コマの発話者ベクトル

4 コマ漫画の前後半コマに含まれるテキストの $m$ 次元発話者ベクトルをテキストの読み順に連結する. この時, 最大テキスト長 $t$ に合わせて後方ゼロパディングする. このようにして得られた $t \times m$ 次元のベクトルをそれぞれ前半コマ発話者ベクトル 後半コマ発話者ベクトルとする.

# 5 数値実験

本実験では, 提案手法によって得られる前後半コマのテキスト分散表現と発話者ベクトルを特徴量として用いてコマの順序を識別し, 5 分割交差検証で精度を確認した. 実験には, manga109 の 4 コマ漫画作品から全 840 話のデータを使用した. 表 1 に使用したデータを示す. 提案手法で用いる最大テキスト長 $t$ と発話者ベクトルの次元数 $m$ はそれぞれ最大の要素を持つ作品のものとした. また, テキストは BERT を用いることで $n = 768$ の分散表現を獲得した.

## 5.1 実験 1: 単体特徴量によるコマ順序識別

テキスト分散表現, もしくは発話者ベクトルを単体で用いてコマ順序識別実験をした. まず, 前後半コマテ

[1] http://nlp.ist.i.kyoto-u.ac.jp

表 3: 実験 1: 結果

| | 正順 | | 逆順 | | 全体 | |
|---|---|---|---|---|---|---|
| | f1 | std | f1 | std | acc | std |
| テキスト | 0.681 | 0.032 | 0.682 | 0.040 | **0.683** | 0.021 |
| 発話者 | 0.588 | 0.061 | 0.595 | 0.065 | 0.596 | 0.041 |
| baseline | - | - | - | - | 0.500 | - |

表 4: 正解データ割合

| | | テキスト | | |
|---|---|---|---|---|
| | | ○ | × | 全体 |
| 発話者 | ○ | 0.426 | **0.170** | 0.596 |
| | × | **0.257** | 0.146 | 0.404 |
| | 全体 | 0.683 | 0.317 | |

キスト分散表現または前後半コマ発話者ベクトルを前 - 後 で連結したものを正順, 後 - 前 で連結したものを逆順とし, これを LSTM に入力する. LSTM の最終出力を識別器に入力して正順逆順の判定をした. 正順と逆順のデータ数の比率は 1:1 であり, 学習データとテストデータの比率は 8:2 とした. 識別器には 3 層 MLP を使用し, optuna によりパラメータを調整した. 表 2 に実験条件を示す.

表 3 に実験結果を示す. 表 3 より, テキスト分散表現を入力したモデルの方が精度が高いことが分かる. テキストの情報の方が発話者の情報に比べて多くの情報を持っていると考えられるため, これは人間の直感的にも妥当であると言える. 表 4 にそれぞれのモデルの正解データの割合を示す. ○は正解したデータを, ×は不正解だったデータを表しており, 各カラムはテキストのみを入力するモデルと発話者のみ入力するモデルの出力結果から, 正解もしくは不正解データの全体に占める割合を表している. 表 4 より発話者情報のみが正解するものが 17 %, テキスト情報のみが正解するものが 26% あることが分かる. このことから, テキスト情報に対して発話者情報を付加することでテキスト情報のみを特徴量として用いるよりも精度を向上させることができると考えられる.

## 5.2 実験 2: テキストと発話者情報の統合によるコマ順序識別

テキスト分散表現と発話者ベクトルの両方を用いてコマ順序識別実験をした. まず, 前後半コマテキスト分散表現と前後半コマ発話者ベクトルをそれぞれ正順または逆順で連結し, 各々を LSTM に入力する. LSTM の最終出力を concat した後に識別器に入力して正順逆順の判定をした. 正順と逆順のデータ数の比率は 1:1 であり, 学習データとテストデータの比率は 8:2 とした. 識別器には 3 層 MLP と SVM を使用し, パラメータは optuna で探索をした後に調整した. 表 2, 5 に実験条件を示す. なお, LSTM の重みは実験 1 で用いたもので固定し, 識別器のみを学習させた.

表 5: 実験条件 SVM

| | |
|---|---|
| kernel | rbf |
| C | 36.24 |
| gamma | 0.289 |

表 6 に実験結果を示す, 表 3 と表 6 より, テキストの分散表現を単体で入力するモデルよりもテキストと発話者の情報を組み合わせたモデルの方が精度が高くなっているが, 統計的な有意差は確認できなかったため, テキスト情報と発話者情報を組み合わせたことで精度が向上したとは言い難い.

ここで, 情報を統合した 2 つのモデルの傾向を調べるため, テキスト情報単体入力モデルに対して, 発話者情報を組み込んだと考えて, それぞれのテキスト+発話者モデルのテキスト情報単体入力モデルに対する正答したデータの割合の変化を考える. 表 7, 8 にテキスト単体入力モデルと発話者単体入力モデルの出力結果から, それぞれのモデルの正解もしくは不正解データの全体に占める割合を示す. テキスト情報および発話者情報単体モデルの識別結果が一致しており, それが反転した結果を出力した結果, つまり表 7, 8 で (○, ○) → × もしくは (×, ×) → ○ の変化をした結果を「誤差」とする. また, テキスト情報単体モデルで正解かつ発話者情報単体モデルで不正解した結果もしくはテキスト情報単体モデルで不正解かつ発話者情報単体モデルで正解した結果, つまり表 7, 8 で (○, ×) → × もしくは (×, ○) → ○ の変化をした結果を発話者情報の入力による影響と考えて「発話者影響」とする. 表 9 にテキスト情報単体モデルの精度を基準とした誤差と発話者影響による精度の変化を示す. この表の数値は全データに対する割合を示している. MLP モデルは誤差の影響が大きいことから, 発話者情報を加えたことによって向上したとは考えにくい. 一方で SVM モデルは発話者影響が正の値を取るため, 発話者情報を加えたことでテキストだけでは判定できなかったデータを正しく識別できるようになったと言える. 以上から, 単純に情報を統合するだけでは有意に精度を向上させられないが,

表 6: 実験 2: 結果

| | 正順 | | 逆順 | | 全体 | | | |
|---|---|---|---|---|---|---|---|---|
| | f1 | std | f1 | std | acc | std | loss | std |
| MLP | 0.682 | 0.029 | 0.685 | 0.040 | 0.685 | 0.030 | 1.003 | 0.160 |
| SVM | 0.706 | 0.031 | 0.663 | 0.041 | **0.687** | 0.034 | - | - |
| baseline | - | - | - | - | 0.500 | - | - | - |

表 7: 正解データ割合 MLP

| | テキスト+発話者 | |
|---|---|---|
| (テキスト, 発話者) | ○ | × |
| (○, ○) | 0.425 | 0.001 |
| (○, ×) | 0.251 | **0.006** |
| (×, ○) | **0.001** | 0.169 |
| (×, ×) | 0.007 | 0.139 |
| 全体 | 0.685 | 0.315 |

表 8: 正解データ割合 SVM

| | テキスト+発話者 | |
|---|---|---|
| (テキスト, 発話者) | ○ | × |
| (○, ○) | 0.376 | 0.050 |
| (○, ×) | 0.224 | **0.033** |
| (×, ○) | **0.050** | 0.120 |
| (×, ×) | 0.037 | 0.110 |
| 全体 | 0.687 | 0.313 |

統合したモデルに対して発話者情報の影響の度合いに差があることが分かった. ただし, 特徴量を単体で入力するモデルの学習 epoch を 50 から下げて実験をしたところ, テキストと発話者の情報を統合したモデルの精度は単体入力モデルに比べて有意に精度が向上することが確認できているため, 特徴量の組み合わせ方やその順番を調整することで精度の向上が見込めると思われる.

## 6 まとめと課題

本研究において, 発話者情報のベクトル化手法およびテキストと発話者の情報を用いたコマの順序識別手法を提案した. テキストと発話者ではそれぞれのコマの識別結果に差が見られることから, 情報を統合することで精度向上の可能性が考えられる. しかし, 単純な統合モデルでは優位に精度が向上しないため, 何らかの工夫が必要である. 今後の課題として以下の点が挙げられる.

表 9: 統合モデルの傾向

| 識別器 | 誤差 | 発話者影響 | 合計 |
|---|---|---|---|
| MLP | **+0.006** | -0.005 | +0.001 |
| SVM | -0.013 | **+0.017** | **+0.004** |

- BERT の fine tuning
- 各モデルの出力結果の定性的な解析
- 個々のテキストと発話者の対応付け

## 参考文献


[1] Allen.K, Ingulsrud.J. Strategies used by children when reading manga. 神田外語大学紀要. 2018

[2] Y.Matsui, K.Ito, Y.Aramaki, A.Fujimoto, T.Ogawa, T.Yamasaki, K.Aizawa, Sketch-based Manga Retrieval using Manga109 Dataset, Multimedia Tools and Applications, Springer, 2017

[3] T.Ogawa, A.Otsubo, R.Narita, Y.Matsui, T.Yamasaki, K.Aizawa, Object Detection for Comics using Manga109 Annotations, arXiv:1803.08670, 2018

[4] 阿部 和樹, 中村 聡史. 漫画における台詞発話者の自動判定に向けた技術的困難性による整理とデータセット構築手法の検討, 第 2 回コミック工学研究会発表会, pp.7-14, 2019.

[5] Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory, Neural Computation, Vol. 9, No. 8, pp. 1735 - 1780, 1997

[6] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805, 2018