小説本文から抽出した人物情報の関係抽出

岡　裕二 [1]　安藤　一秋 [2]
[1] 香川大学大学院工学研究科 [2] 香川大学創造工学部
s20g460@stu.kagawa-u.ac.jp ando.kazuaki@kagawa-u.ac.jp

概要

本稿では，小説の本文テキストから抽出される人物情報を体系化することを目的に，同一文に出現する人名と人物情報のみを関係抽出の対象に設定し，ルールベースモデルと深層学習モデルの 2 つの関係抽出モデルの性能を検証する．性能評価の結果，ルールベースモデルと深層学習モデルは共に precision が recall より高い結果となった．micro-F1 については，ルールベースモデルが深層学習モデルを約 9.7 ポイント上回った．ルールベースモデルでは，近距離で頻出する関係しか抽出できないが，深層学習モデルでは遠距離の関係も抽出できることを確認した．

1　はじめに

電子書籍の普及や小説投稿サイトの増加などに伴い，膨大な数の小説が出版・投稿されるようになった．これは読みたい小説が増えることを意味する反面，膨大な小説の中から個人の嗜好に適した小説を見つけることが難しくなったともいえる．書籍を取り扱う EC サイトや小説投稿サイトでは，作者や各種タグなど，特定の情報に基づく検索機能を利用できるが，小説それぞれの内容に基づいた検索機能は実装されていない．また，携帯端末を利用して隙間時間に小説を楽しむ，複数の小説を同時に読むなど，個人毎の読書スタイルが多様化してきた．読書スタイルの多様化により，まとまった時間に 1 つの小説を読む場合と比べて，登場人物の特徴や関係などを失念したり，読書再開時に内容を思い出すために，一部を読み直す機会なども増えている．

個人の嗜好は，「メリーバッドエンド」や「仲間が敵になる」など，展開に関する嗜好と，「角が生えたちびっ子」や「金髪の用心棒」などの登場人物に関する嗜好に分けられる．小説テキストから登場人物に関する情報を抽出して利用できるようになれば，人物に関する嗜好に基づいて小説を検索できるようになる．また，読書スタイルの多様化に対しては，登場人物情報を利用して，読んだところまでのあらすじや人物相関図を生成・提示することで読書支援が可能になる．そこで本研究では，小説内の人物情報を抽出して体系化することで，人物情報を用いた小説検索，あらすじ生成，人物関係図生成などの実現を目指している．

著者らの先行研究 [1] では，商業小説のあらすじテキストで学習した系列ラベリングモデルを Web 小説の本文テキストに適用し，訓練データおよび提案モデルの小説本文に対する人物情報の抽出性能を検証した．本稿では，小説の本文テキストから抽出される人物情報を体系化することを目的に，同一文に出現する人名と人物情報のみを関係抽出の対象に設定し，ルールベースモデルと深層学習モデルの 2 つの関係抽出モデルの性能を検証する．

2　実験用データセットの構築

関係抽出手法の性能評価に利用する実験用データセットの構築法について述べる．

2.1　対象データ

本稿では，著者らの先行研究 [1] で収集した「小説家になろう」サイトに掲載されているファンタジー小説の人気上位 250 件から無作為に選んだ 2 作品と，馬場らの研究 [2] で使用された「青空文庫」に収録されている 4 作品，計 6 作品の本文を用いる．

「青空文庫」にて公開されている作品にはそれぞれ図書カードという Web ページが紐づいているため，図書カードページからルビつきのテキストファイルをダウンロードし，ルビや注釈など本文以外の要素を排除した．その後，我々の先行研究 [1] と同様の手順で固有表現タグを付与する．その後，人名と人物情報の正解関係を示す情報として，各タグに対して本文テキスト内の人名の中で最長のものを付与する．以下に手順と例を示す．

1.「青空文庫」から収集した小説の本文を 1 文ずつ形態素解析する．
2. 各形態素に対して，以下のルールでタグ付けする．タグの形式には IOB 2 タグ形式を用いる．
   - NAME: 名前
     例：西尾，信長，シャルル・マーニュ
   - MF: 性別
     例：男，美男子，美女，乙女，女の子
   - AGE: 年齢
     例：16 歳，少年，お婆さん，幼い，高校生
   - STATE: 容姿や特性
     例：白い髪，元気，高飛車，天才，職人気質
   - PRO: 職業や立場
     例：竜飼い，仙女，最高権限者，メンバー
   - AFF: 組織や種族名
     例：鳳凰学園杖術部，討伐軍，エルフ
   - OTHER: 以上に当てはまらない人物情報
     例：異星人，神，元凶，気鋭，ペンギン
   - PLACE: 地名や建物名
     例：ムー大陸，日本，礼拝堂，魔法学校
   - REL: 人物関係表現
     例：兄，親，敵，相棒，結婚
   - O: それ以外
3. IOB2 タグ形式で B タグが付与されているものについて，当該人物情報に紐づく本文テキスト内の人名の中で，最長のものを付与する．紐づく人名が本文テキスト内に存在しない人物情報には「不明」を付与する．

「小説家になろう」から選択した 2 作品については，我々の先行研究 [1] で上記の固有表現タグを付与したデータに上記の処理 3 で紐づく人名を付与したものを利用する．各作品の総文数と各タグの出現数，関係文数（人名とほか人物情報が同一文内に出現する文の数）を表 1 に示す．

2.2 データセットの構築

関係抽出モデルの性能評価に用いるデータセットの構築法を以下に示す．

1. 小説本文を 1 文ずつ形態素解析および係り受け解析する．係り受け解析には，後で述べる深層学習モデルの関係から安岡らの論文 [3] の Universal Dependencies(UD) に基づく係り受け解析器 [4] を利用する．なお，固有表現タグに関する情報は，上記の結果に準拠させる．
2. 深層学習モデルの訓練及び性能評価のため，1 文内に人物情報が 2 つ以上存在する文に対しては，人名と人物情報が同一人物に紐づく場合「exist」，紐づかない場合「no_relation」タグを追加し，人名と対象人物情報タグの最初と最後の形態素の位置を記録する．上記は，1 文内の人名と人物情報の組み合わせの数だけ作成する．

上記手順により作成されたデータ例を表 2 に示す．

3 関係抽出モデル

本稿では，*n*-gram によるルールベースモデルと，GCNs（Graph Convolutional Networks）の一種である AGGCNs（Attention Guided Graph Convolutional Networks）モデルを用いて，人名・人物情報間の関係を抽出する手法を検討する．なお，対象となる人物情報は，MF（性別），AGE（年齢），STATE（容姿や特性），PRO（職業や立場），AFF（種族や所属），OTHER（その他の人物情報）とする．

3.1 ルールベースモデル

人名と人名に紐づく人物情報が近距離に出現する事例から関係抽出するルールを検討するため，人名とそれに紐づく人物情報およびそれらの間に出現する形態素のパターンを *n*-gram で調査する．調査対象は，表 1 に示す関係文数の総計 290 文とする．

まず，データセット中に，人名とその人名に紐づく人物情報の間に形態素が存在しない事例（2-gram）を調査した結果，83 件存在することを確認した．次に，人名とその人名に紐づく人物情報の間に 1 形態素以上含む，*n*=3〜5 までの *n*-gram の頻度を調査した．その結果を表 3 に示す．表 3 より，3-gram の「の」，「、」，「な」以外は頻度が 1 件となり，極端に少ないことがわかる．以上の結果より，本稿では，人名に形態素を挟まず隣接する，または「の」，「、」，「な」が間に存在する人物情報を人名に関係のある情報として紐づけることにする．

3.2 深層学習モデル

ルールベースモデルでは，人名の近距離に出現する人物情報を抽出できるが，人名と人物情報が離れて出現する場合は，その関係を抽出できない．そこで，深層学習モデルでは，AGGCNs[5] を用いて，距離にかかわらず 1 文内に出現する 2 つの人物情報間に関係があるか否かを推定することを目指す．

表 1　各作品の文数と各タグの出現数

	文数	関係文数	NAME	MF	AGE	STATE	PRO	AFF	OTHER
モルグ街の殺人	884	36	110	127	14	63	93	59	3
白銀の失踪	709	89	278	53	14	102	144	11	0
黄色な顔	686	10	56	218	10	52	24	1	0
空き家の冒険	593	74	159	137	25	74	102	22	15
学園無双の勝利中毒者	753	66	298	52	64	51	44	0	10
異世界を救った少年 少女になって戻ってくる	382	15	101	18	16	37	12	5	2
総計	4,007	290	1,002	605	143	379	419	98	30

表 2　作成したデータセットの例

	黒崎	一護	は	死神	代行	だ	。
形態素番号	1	2	3	4	5	6	7
品詞	PROPN	NOUN	ADP	NOUN	NOUN	AUX	PUNCT
係り元の形態素番号	2	5	2	5	0	5	5
係り受けの依存関係	compound	nsubj	case	compound	root	cop	punct
固有表現タグ	NAME	NAME	O	PRO	PRO	O	O
人名・対象人物情報の終始	subj_start	subj_end		obj_start	obj_end		
関係タグ	exist						

表 3　人名・人物情報間の単語群の出現数

	人名・人物情報間の単語群の頻度順位					
	1	2	3	4	5	6
3-gram	の (16)	、(12)	な (4)	は (1)	いい (1)	・(1)
4-gram	の、(1)	—— (1)				
5-gram	として 有名な (1)	を着た (1)	の名は (1)	であった (1)		

AGGCNs は，文中の単語同士で構築される完全な依存木構造から，関係抽出に有用な関連部分構造に注目する方法を学習するソフトプルーニング手法である．AGGCNs のモデル図 [5] を図 1 に示す．AGGCNs はグラフを表現するノード埋め込みと隣接行列を入力とする M 個の三層ブロックから構成されている．Attention Guided 層では単語同士の依存関係から Multi-Head Attention を用いて Attention Guided 隣接行列を N 個生成する．Densely Connected 層では前層で得られた N 個の行列を別々の密結合層に入力し，最終層で線形結合する．

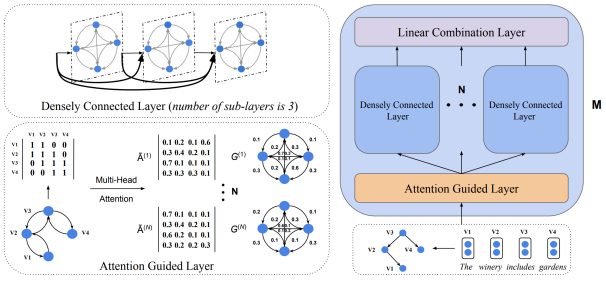

図 1　AGGCNs のモデル図（[5] より引用）

AGGCNs は，その手法を提案した著者が GitHub でコードを公開 [6] しているので，そのコードおよびパラメータを変更し，関係抽出モデルを実装した．ハイパーパラメータのうち，RNN は 2 層，バッチサイズは 4，各種隠れ層は 200 次元に変更し，他の設定はデフォルト設定を採用する．また，最適モデルを選ぶ指標は，「開発データに対する F 値が一番大きくなった場合」から「開発データに対する偽陽性と偽陰性の和が一番小さくなった場合」に変更し，品詞や依存関係表現の ID は日本語 Universal Dependencies に合わせたものに変更する．単語ベクトルとして用いる分散表現には，日本語 Wikipedia の本文全文で事前学習されたもの [7] を利用する．事前学習に用いたパラメータを表 4 に示す．

表 4　事前学習した単語分散表現のハイパーパラメータ

モデル	cbow
次元数	200
Window size	5
ネガティブサンプリング	5
ダウンサンプリング	0.001

4　評価実験

4.1　評価方法

ルールベースモデルの抽出性能は，2.1 節で付与した，各種人物情報に紐づく人名から正解辞書を作成し，ルールベースモデルでの抽出結果と正解辞書との一致率で評価する．深層学習モデルの抽出性能は，2.2 節で構築したデータセットの「exist」タグを正例，「no_relation」タグを負例として評価する．precision，recall，F 値（micro-F1）を評価尺度とし，深層学習モデルについては 10 分割交差検証で評価

する．データセットには，表 1 の関係文数の総計 290 文を対象に，2.2 節に基づいて構築した 542 文を，訓練データ：開発データ：テストデータ＝ 7:2:1 の割合で用いる．

4.2 実験結果と考察

実験結果を表 5 に示す．表 5 より，ルールベースモデルと深層学習モデルは共に precision が高いことが確認できる．また，micro-F1 は，ルールベースモデルが深層学習モデルを約 9.7 ポイント上回った．

表 5 6 作品に対する各手法の性能

	precision	recall	micro-F1
ルールベース	83.10	54.13	65.56
深層学習	60.96	51.48	55.82

本稿では，近距離の関係のみを抽出対象に設定し，データセット中に頻出するパターンをもとに関係抽出ルールを作成した．実験の結果，83.1%の precison が得られたことから，近距離で頻出する関係を正しく抽出できたといえる．しかし，ルールの網羅性の問題から recall は，54.1%に留まった．ルールベースモデルの抽出エラーを分析した結果，人名の近距離に人物情報が出現する場合でも，「ニコルズ博士、エピキュロス」のように一文内に二つ以上の人名が隣接する場合，「ニコルズ」，「エピキュロス」両名に「博士」が紐づくといったミスが確認された．これに関しては，人物情報と人名の物理的な距離を指標にすることで改善できる可能性がある．また，「ミニョー父子銀行（AFF）の行員（PRO）、アドルフ（NAME）」のように，人物情報が連続して出現する場合，ルールベースモデルでは，人名に近い人物情報しか紐付けられない．人物名と人物情報の記述スタイルには多様性があり，単純なルールベースでの限界も見えた．

深層学習モデルに関しては，ルールベースモデルでは抽出できない 4-gram 以上の距離がある人物情報に対して関係を抽出できることを確認できた．しかし，ルールベースモデルよりも recall が約 2.7 ポイント低いことから，近距離に対する抽出性能については，ルールベースモデルが深層学習モデルの性能を上回っていると考えられる．よって，3-gram 以内の近距離はルールベース主体，4-gram 以上の遠距離は深層学習モデルというハイブリッドモデルを構築することで，抽出性能を向上できる可能性がある．また，本稿での実験では，深層学習モデルに対して 542 文という少量のデータセットによる実験となったため，今後はデータセットを拡充した実験が必要である．

表 6 人名と紐づく人物情報が出現する場所と割合

	同一文内	別文
モルグ街の殺人	59.88	40.12
白銀の失踪	51.49	48.51
黄色な顔	18.99	81.01
空家の冒険	60.18	39.82
学園無双の勝利中毒者	47.11	52.89
異世界を救った少年 少女になって戻ってくる	22.52	77.48

本実験では，同一文に出現する人名と人物情報のみを関係抽出の対象に絞ったが，人名と人物情報が異なる文に存在する場合もある．そこで，人名と紐づく人物情報が同一文内に存在するのか，または別文に存在するのかの割合を調査した．その結果を表 6 に示す．表 6 より，人名と人物情報がそれぞれ別文に存在する場合が少なくとも約 4 割から約 8 割存在することがわかる．今後は，「彼」や「彼女」といった代名詞を手掛かりに，人名が出現しない文に出現する人物情報に対して，関係を抽出する手法やゼロ代名詞への対応について検討する必要がある．

5 おわりに

本稿では，小説の本文テキストから抽出される人物情報を体系化するため，同一文に出現する人名と人物情報のみを関係抽出の対象に設定し，ルールベースモデルと機械学習モデルの 2 つの関係抽出モデルの性能を検証した．性能評価の結果，ルールベースモデルと深層学習モデルは共に precision が recall より高い結果となった．micro-F1 については，ルールベースモデルが深層学習モデルを約 9.7 ポイント上回った．ルールベースモデルでは，近距離で頻出する関係しか抽出できないが，深層学習モデルでは遠距離の関係も抽出できることを確認した．

今後の課題として，まずデータセットを拡充する必要がある．また，人名ではなく代名詞が利用される文やゼロ代名詞の文などから関係を抽出する手法や，複数文にまたがる人名・人物情報の関係を抽出する手法について検討する．

参考文献

[1] 岡裕二, 安藤一秋. 小説あらすじを用いて学習した系列ラベリングモデルによる小説本文からの人物情報抽出の性能検証. 言語処理学会第 27 回年次大会発表論文集, 2021.

[2] 馬場こづえ, 藤井敦. 小説テキストを対象とした人物情報の抽出と体系化. 言語処理学会第 13 回年次大会発表論文集, 2007.

[3] 安岡孝一. 世界の universal dependencies と係り受け解析ツール群. 第 3 回 Universal Dependencies 公開研究会, 2021.

[4] UD 係り受け解析器, 2021. https://colab.research.google.com/github/KoichiYasuoka/deplacy/blob/master/demo/2021-06-22/supar-ja.ipynb.

[5] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, 2019.

[6] AGGCN の github リポジトリ, 2019. https://github.com/Cartus/AGGCN.

[7] 日本語 Wikipedia エンティティベクトル, 2007. http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/.