E-004

小説本文から抽出した人物情報の構造化手法の検討
A Method of Structuring Character Information Extracted from Novel Text

岡 裕二† 安藤 一秋†
Yuji Oka kazuaki Ando

1. はじめに

近年，電子小説や小説投稿サイトなどの発展により，時間や場所を問わずに小説を読めるようになった．一方，小説数の増加によって，個人の嗜好に合う作品を探すことは難しくなった．また，隙間時間を利用した読書が容易になったが，一冊を読み終えるまでの時間は長くなり，それまでの内容を振り返るための読み返しが増加している．我々は，物語の展開や登場人物の特徴に関する嗜好を用いた小説検索やあらすじ生成などにより，これらの問題を軽減できると考えている．

本研究では，小説テキストから抽出した人物情報を利用した検索支援やあらすじ生成を目的にする．本稿では，ファンタジー小説の本文から抽出した人物情報の構造化に向け，パターンマッチング手法および固有表現抽出手法に基づく人物属性の抽出手法の抽出性能を比較し，考察する．

2. 人物情報抽出手法の検討

筆者らは，先行研究[1]において，深層学習を用いた系列ラベリングモデルによって，ファンタジー小説の本文から人物情報を抽出する手法を検討した．本稿では，辞書や規則とのパターンマッチングを用いた既存手法[2]と，先行研究[1]で教師データに付与したタグを用いる提案手法それぞれでの抽出性能を評価する．以下，既存手法と提案手法について述べる．

2.1 既存手法

パターンマッチングによる人物情報の抽出手法として，馬場らが提案した人物情報の抽出手法[2]を利用する．馬場らの抽出手法は，小説テキストを文単位に分割し，形態素解析の結果に基づき人名を抽出する．そして，辞書と規則を用いて人名の周辺文脈（形式段落）から人物属性（性別，年齢，年代，職業，身体的特徴，性格）を抽出する．なお，馬場らの抽出手法の再現にあたり，形態素解析器は，データセットの都合上，ipadic を利用した MeCab を使用し，年代も考慮しない．また，抽出に利用した辞書も公開されていないため，再現可能な範囲での実装となる．

人名については，品詞に「人名」を含みかつ「接尾」を含まない形態素を人名として抽出する．そして，抽出した人名を含む形式段落から，人手で作成した規則と表層一致した表現を，その人物の人物属性として紐づける． 以下，人物属性（性別，年齢，職業，身体的特徴，性格）の抽出方法について説明する．

性別は，性別（男性 or 女性）を付与した性別辞書とのマッチングにより抽出する．同一人物に対して男性を表す語と女性を表す語の両方が出現する場合，テキスト全体で多く出現している性別を採用する．馬場らの論文[2]には，性別辞書の内容が明示されていないため，相当すると考えられる 75 語を辞書に登録した．

年齢は，「（数値表記）＋（歳，または，才）」といった規則により抽出する．数値表記として，半角数字，全角数字，漢数字に対応する．馬場らの論文[2]に従い，同一人物に対して出現する全ての年齢表記を年齢として付与する．

職業は，職業辞書とのマッチングにより抽出する．同一人物に対し，複数抽出された場合は全ての職業を採用する．なお，職業辞書は，馬場らの論文[2]に準じて，角川類語新辞典[3]に収録されている職業 2,261 語を参考に作成した．

身体的特徴は，CaboCha による係り受け解析の結果に対して，以下の 3 つの規則により抽出する．

- 「身体を表す語」＋(が｜は)＋({形容詞}｜{名詞})
- {形容詞}+「身体を表す語」
- {名詞}+の+「身体を表す語」

規則内の{形容詞}と{名詞}は，形態素解析結果の品詞を意味する．抽出規則に合致した単語が連用形の場合，身体ではなく，動作を修飾するため除外する．

性格は用語リストとのマッチングにより抽出する．ただし，係り受け解析の結果，動詞やサ変接続の名詞に係っているものは除外する．なお，用語リストは「基本的な性格表現用語」[4]を参考に作成した．

馬場らの論文[2]に準じて，人名を含む形式段落から抽出されたすべての人物属性（性別，年齢，年代，職業，身体的特徴，性格）を，その人物の人物属性として紐づける．

2.2 提案手法

提案手法については，我々の先行研究[1]で最良性能を記録した BiLSTM-CRF-pos10 モデル（抽出モデル）を利用して，人名と人物属性を抽出する．抽出モデルで用いるタグの種類と例を表 1 に示す．提案手法では，小説テキストの各文に抽出モデルを適用した後，NAME（人名）タグが付与されている形態素列を人名として抽出する．そして，馬場らと同様，抽出された人名を含む形式段落において，MF，AGE，STATE，PRO タグが付与されているすべての形態素列を，その人物の人物属性として紐づける．それ以外のタグは，対象外とする．また，MF（性別）タグでは性別を区別していないため，MF タグが付与された形態素列に対して，既存手法と同じ性別辞書を用いて分類する．

表 1. 提案手法で用いるタグの種類と例

タグ名	タグ付け対象	タグ付け例
NAME	登場人物の名前	西尾，太郎，シャルル・マーニュ
MF	性別表現	男，美男子，乙女
AGE	年齢表現	16歳，お婆さん，幼い
STATE	容姿・性格表現	白い髪，元気，高飛車
PRO	職業・立場表現	騎士，権力者，メンバー
AFF	組織・種族名	日本政府，討伐軍，エルフ
OTHER	その他の人物情報	神，気鋭，ペンギン
PLACE	地名・建物名	ムー大陸，日本，礼拝堂
REL	人物関係表現	兄，相棒，結婚
O	以上に当てはまらないもの	

† 香川大学, Kagawa University

3. 評価実験

3.1 正解データ

我々の先行研究[1]では，小説家になろうに掲載されているファンタジー小説を用いてデータセットを構築した．データセットに含まれる 2 作品（各 789 文と 394 文）の教師データを基に，本実験で利用する正解データを構築する．

先行研究[1]の教師データは，IOB2 形式のタグが付与された固有表現抽出用データであり，人物と人物属性を紐づける情報は付与されていない．そこで，NAME タグと MF，AGE，STATE，PRO タグを人手で紐づけて，正解データを構築した．具体的には，人物属性タグに，対象小説の中で当該人物を表す最長の人物名を付与することで構築した．

3.2 実験方法

2 作品に対して，馬場らの手法（既存手法）と提案手法（BiLSTM-CRF-pos10 モデル）を適用し，正解データとの一致率で各手法の性能を評価する．各人物属性の抽出性能は人名の抽出性能に大きく影響を受けるため，本実験では既存手法，提案手法ともに，正解データの NAME タグを利用し，人名が完全に抽出できた場合の結果で評価する．評価指標は precision，recall，F 値とする．また，人物が一人以上出現する形式段落内の全てのタグが形式段落内の人物と紐付けられた場合の理想値も算出して比較する．なお，ベースライン手法の身体的特徴と性格は，STATE タグの正解を基に判定する．

3.3 結果

理想値および 2 つの手法の評価結果を表 2 に示す． MF（性別）において，両作品で既存手法が提案手法を上回った．STATE（身体的・精神的特徴）においては，両作品で提案手法が既存手法を上回った．PRO（職業・立場）については，F 値で見れば，作品 1 で提案手法は既存手法を上回っているが，既存手法は recall が 100%を記録している．また，理想値では作品 1 において，MF，PRO タグで完璧に紐づけることができたが，作品 2 において，AGE，PRO タグで全く正解が導出できなかった．STATE タグでは，それぞれ 6~7 割程度の属性を紐づけることができた．

表 2. NAME タグを利用した場合の抽出性能

		理想値			既存手法			提案手法		
		Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
MF	作品 1	75.00	75.00	75.00	75.00	75.00	75.00	71.89	74.04	72.95
	作品 2	60.00	60.00	60.00	60.00	60.00	60.00	57.51	59.23	58.36
AGE	作品 1	46.15	75.00	57.14	0.0	0.0	0.0	40.35	73.59	52.12
	作品 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
STATE	作品 1	48.57	70.83	57.63	5.56	12.50	7.69	14.24	26.02	18.41
	作品 2	40.91	60.00	48.65	10.00	13.33	11.43	11.99	22.04	15.53
PRO	作品 1	14.29	100.0	25.00	3.03	100.0	5.88	6.67	56.85	11.94
	作品 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

4. 考察

表 2 より，既存手法，提案手法ともに STATE（身体的・精神的特徴）と PRO（職業・立場）の性能が低いことがわかる．特に precision が低い理由は，文中に出現する人物属性を同段落に出現する人物全員に付与していることが原因であると考える．理想値と比較すると，既存手法，提案手法ともに理想値を大きく下回っている．この問題については，人名との距離や係り受け，周辺での出現回数などの閾値を用いた別の紐付け手法を検討することで，性能向上を図ることができると考える．

MF（性別）において，既存手法は，提案手法を上回り，理想値と同等の結果を記録した．これは，性別表現として出現する語彙は種類が少なく，多様性もないため，パターンマッチングの優位性がでたと考える．

既存手法が AGE（年齢）属性をまったく抽出できなかった理由は，年齢を意図する表現が「少年」「17」「七つ」など，規則で記述しきれない多様性をもつことが原因と考えられる．また，規則に該当する表現があっても，対象となる人物が形式段落に存在しないことも要因として挙げられる．後者に関しては，提案手法でも影響を受けている．

理想値では，作品 2 において AGE，PRO の抽出性能が 0 となっていることから，作品によっては，該当する人名が全く存在しない形式段落にも人物属性が書かれていることがわかる．例えば，「彼」「彼女」「この男の子」などの表現は特定の人物を表す可能性があるが，代名詞的な用法で使われるために同じ形式段落内に該当する人名が出現しないことがある．さらに，代名詞に紐づいて他の属性が出現する場合もある．また，先行研究[2]では，青空文庫の英米文学の推理小説を評価対象に利用しているが，本実験で利用した小説は Web 小説である．Web 小説では，可読性向上のために頻繁に形式段落が区切られる（形式段落が短い）傾向にある．よって，本来は同一段落となる部分が別段落となり，人物名が出現しない段落に人物情報が出現する事例が増加したと考えられる．今後は，形式段落以外の文章の区切り方を検討する必要がある．

5. 終わりに

本稿では，ファンタジー小説の本文から抽出した人物情報の構造化に向け，パターンマッチングに基づく既存手法と固有表現抽出に基づく提案手法での抽出性能を比較した．実験の結果，固有表現抽出に基づく提案手法は， STATE（身体的・精神的特徴）について，パターンマッチングに基づく既存手法を上回る性能を記録した．一方，AGE（年齢）と PRO（職業立場）について，作品 1 では既存手法を上回ったが，作品 2 では既存手法と同様に全く正解が抽出できない結果となった．理想値およびこれらの結果から，形式段落に基づいて文中に出現する人物属性と人名の紐付ける手法の限界を確認した．また，MF（性別）については，パターンマッチングによる手法の性能が高くなり，多様性が少ない属性については，規則で抽出する優位性も確認できた．

今後は，人名と人物属性の距離や係り受け関係，人名周辺での当該人物情報の出現回数などを用いた紐付け手法や，ゼロ代名詞補完，人名の前後数文などの新たな文章の区切り方などについて検討し，人物情報を構造化する手法を完成させる．

参考文献

[1] 岡他, “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, (2021).
[2] 馬場他, “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, (2007).
[3] 大野晋，浜西正人著，角川類語辞典(1981).
[4] 村上宣寛, “基本的な性格表現用語の収集”, 性格心理学研究, Vol. 11, No. 1, pp. 35–49 (2002).