系列ラベリングによる小説のあらすじからの人物情報抽出の検討

岡 裕二　　安藤 一秋
香川大学 工学部　香川大学 創造工学部
s16t215@stu.kagawa-u.ac.jp ando@eng.kagawa-u.ac.jp

1 はじめに

近年，小説の電子化や小説投稿サイトの発展などにより，小説が容易に読めるようになった．一方で，出版・投稿される小説数が増加することにより，個人の嗜好に合う作品を探すことが難しくなっている．ライトノベルなどの小説では，登場人物の個性や立場が重視される傾向にあり，それらを楽しむためにキャラ文芸というタイプの小説が誕生するほど，登場人物に関する情報は小説を探すための必要条件の 1 つになっている．しかし，「異世界で年老いた執事が活躍する小説」や「勇者と魔王が友達になる小説」という検索要求を満たす検索サービスは現在存在しない．

そこで本研究では，読者の求める登場人物情報や人物関係などに合致する小説を検索できるシステムの構築を目的とする．本稿では，検索システムの構築に向け，系列ラベリングによってファンタジー小説のあらすじから人物情報と人物関係表現を抽出する手法について検討し，その性能を評価する．

2 関連研究

小説からの登場人物の抽出や人物関係図を構築する研究はいくつか存在している．馬場ら [1] は，辞書や規則によるマッチングによって人物情報を抽出する手法を提案している．人物が発言したか否かの台詞情報，人物が特定の場面に存在するか否かの入退場情報及び登場人物同士の共起頻度を用いて人物間の重みを決定し，関係グラフを構築する．この関係グラフは，人物名が書かれたノードとそれを結ぶエッジで構成されているが，それぞれの関係がどのような性質（味方・敵，上司，部下など）を表現しているのかまでは表していない．また，4 つの小説に対する人物情報抽出の精度は 42.4%，再現率は 67.0%程度であると述べている．

米田ら [2] は，共起する述語情報や局所出現頻度を用いて小説から登場人物名を抽出する手法を提案している．登場人物は，主語として登場するという仮説の下，形態素および文節情報を利用して人物候補を抽出し，人物候補の主語としての局所出現頻度や人物候補が主語となった時の述語との対応を利用して，人物候補から登場人物の名前を判別する．open テストにおける結果は，適合率 60.3%，再現率 91.9%，F 値 71.5%であると述べている．

神代ら [3] は，発話文から話し手と聞き手の友好敵対関係，上下関係を推定する手法を提案している．発話文と話し手の相対的な位置を示すラベルを素性に機械学習で話し手を同定し，発話文中の呼びかけや発話文との相対的な位置などによって聞き手を同定する．話し手・聞き手を同定した会話の内容から人称表現などを素性として二者間の関係を友好敵対関係，上下関係を推定する．有効敵対関係，上下関係推定の各 F 値は約 40%であると述べている．

3 提案手法

本稿では，CRF（Conditional Random Fields）による系列ラベリング（CRF モデル）を用いて，人物情報を抽出する手法についてを検討する．抽出対象となる人物情報は，名前，性別，年齢，容姿や特性（性格・性質），職業や立場，所属組織とする．

3.1 あらすじの収集

あらすじの収集には，国立情報学研究所（NII）が提供する無料の情報サービスである Webcat Plus[4] を利用する．Webcat Plus では，全国の大学図書館 1,000 館や国立国会図書館の所蔵目録，新刊書の書影・目次 DB，電子書籍 DB など，本に関する様々な情報源を統合して，それらを本・作品・人物の軸で整理した形で提供している．作品ごとにページが存在しており，BOOK データベースなどで提供されている要旨が記載されている．BOOK データベースの要旨は，本の

あらすじと見なせるものが多いため，本稿では作品の要旨をあらすじとみなして利用する．

3.2 教師データの作成

次の手順で，系列ラベリングに利用する教師データを作成する．

1. Webcat Plus から 2 文以上あるあらすじを収集し，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．

- 名前に名前タグ（NAME）を付与
  例：西尾，信長，シャルル・マーニュ
- 性別表現に性別タグ（MF）を付与
  例：男，美男子，美女，乙女，女の子
- 年齢表現に年齢タグ（AGE）を付与
  例：16 歳，少年，お婆さん，幼い，高校生
- 容姿や特性表現に状態タグ（STATE）を付与
  例：白い髪，元気，高飛車，天才，職人気質
- 職業や立場表現に能力タグ（PRO）を付与
  例：竜飼い，仙女，最高権限者，メンバー，国王
- 組織・種族名に所属タグ（AFF）を付与
  例：鳳凰学園杖術部，日本政府，討伐軍，エルフ
- 以上に当てはまらない人物情報にその他タグ（OTHER）を付与
  例：異星人，神，元凶，気鋭，ペンギン
- 地名や建物名に場所タグ（PLACE）を付与
  例：ムー大陸，日本，パリ，礼拝堂，魔法学校
- 人物関係表現に関係タグ（REL）を付与
  例：兄，親，敵，相棒，結婚
- それ以外のものに O タグを付与

地名や建物名に関しては，直接的な人物情報ではないが，小説の舞台（現実世界，パラレルワールド，異世界など）を考慮するために同時に抽出する．また，人物関係表現は，人物関係図のエッジラベルに利用するために抽出を試みるが，人物情報の抽出後にも，係り受け関係や会話文，共起語などを利用して抽出する．

3.3 CRF で用いる素性とパラメータ

CRF は，単語単位でラベリングするモデルを採用し，windowsize は 2 とする．CRF の実装には CRFsuite[5] を用い，ハイパーパラメータはデフォルト値を用いる．また，素性としては以下を利用する．

- 表記
- 文字種
- 品詞＋品詞細分類
- 文字 uni-gram
- 文字 bi-gram
- 各タグに頻出する漢字上位 10 件と同じものが含まれている場合そのタグのフラグを立てる（以下，タグフラグと呼ぶ）

ベースラインは，表記，文字種，品詞＋品詞細分類を素性とした CRF モデルとし，提案素性である文字 uni-gram，文字 bi-gram，タグフラグは，素性を組み合わせて有効性を検証する．

4 評価実験

4.1 評価方法

CRF の素性の組み合わせを 8 パターンで実験し，それぞれの抽出性能を比較する．適合率，再現率，F 値を評価尺度とし，10 分割交差検証で評価する．人手でタグ付けした結果と CRF がラベリングした結果を比較し，完全一致した場合のみを正解と判断する．

4.2 実験データ

実験には，Wikipedia の日本の小説家一覧に登録されている小説家名を検索語として，Webcat Plus の一致検索でヒットした小説のあらすじを用いる．収集したあらすじの中で，Wikipedia の「日本のファンタジー作家一覧」に登録されている作家の作品のあらすじ，または “ファンタジー” という単語を含むあらすじからランダムに 1,008 件のあらすじを収集した．それらに含まれる約 5,000 文を MeCab[1] で形態素解析し，人手で人物情報タグを付与することで実験データを構築した．

[1] https://taku910.github.io/mecab/

4.3 実験結果

4.3.1 NAME を対象とした素性の組み合わせ評価

人物情報の抽出においては，その人物を特定する情報，つまり，NAME の抽出性能が重要といえる．そこで，NAME の抽出性能が高くなる素性を確認するための実験を行う．表記，文字種，品詞＋品詞細分類を素性としたベースラインに対して，文字 uni-gram，文字 bi-gram，タグフラグを組み合わせた場合の結果を表 1 に示す．表の項目は F 値で昇順ソートしており，太字はそれぞれの項目での最大値である．

表 1 から，文字 uni-gram をベースラインに加えた場合，適合率は 1.5 ポイントほど低下するが，再現率が約 13 ポイントと大幅に向上することで，F 値が約 7 ポイント向上していることが確認できる．さらに，文字 bi-gram，タグフラグを加えた場合，適合率，再現率ともに向上することが確認できる．本稿の実験では，以降，すべての素性の組み合わせを利用する．

表 1: 素性の組み合わせによる性能評価（NAME）

	precision	recall	f1-measure
baseline	0.799	0.594	0.681
+flag	0.808	0.610	0.695
+bi	0.811	0.660	0.731
+bi+flag	**0.817**	0.683	0.744
+uni	0.784	0.723	0.752
+uni+flag	0.791	0.732	0.760
+uni+bi	0.796	0.734	0.763
+uni+bi+flag	0.799	**0.741**	**0.769**

表 2: 全タグに対するラベリング評価結果

	precision	recall	f1-measure
NAME	0.799	0.741	0.769
MF	0.877	0.930	0.901
AGE	0.880	0.828	0.851
STATE	0.554	0.222	0.311
PRO	0.681	0.567	0.616
AFF	0.626	0.458	0.525
OTHER	0.571	0.355	0.436
PLACE	0.616	0.491	0.545
REL	0.796	0.660	0.720

4.3.2 全タグに対するラベリング評価

全素性を利用した CRF で，全タグに対するラベリング性能を評価した結果を表 2 に示す．

表 2 から，MF（性別表現）と AGE（年齢表現）の抽出では，F 値が 80%を超えることがわかる．一方で，STATE（容姿表現や性格表現）の抽出では，F 値が 40%を下回ることがわかる．

5 考察

5.1 関連研究との比較

登場人物名の抽出について，関連研究の論文 [1][2] で示された性能と，本稿で提案したモデルの性能を比較する．表 3 に各性能を示す．対象とするデータが異なるため単純比較はできないが，馬場らの手法 [1] に対しては，提案モデルが適合率，再現率，F 値すべてにおいて上回った．米田らの手法 [2] に対しても単純比較はできないが，再現率に関しては米田らの手法が上回ったが，適合率と F 値は提案モデルが上回った．以上の結果から系列ラベリングが名前抽出において有効である可能性を確認できた．

表 3: 先行研究との比較（NAME）

	precision	recall	f1-measure
提案モデル	0.799	0.741	0.769
馬場ら	0.424	0.670	0.512
米田ら	0.603	0.919	0.715

5.2 エラー分析

提案手法のラベリング結果についてエラー分析する．以下に，特徴的なラベリング誤りと例文を示す．なお，例文中の下線部分はラベリングを誤った箇所である．

1. 漢字 1 字を NAME（名前）と誤判定する．
   例）焔の魔術を操り、「冬」と戦う女戦士ゲルダ。
2. NAME（名前）を PLACE（場所）や AFF（所属）と誤判定する．またはその逆が起こる．
   例 1）そんな彼らの前に、マウゼル教の異教検察官を名乗るベルケンスが現れた。
   (answer:AFF → predict:NAME)
   例 2）劣勢を挽回しようとするタリオ大公は、総力戦をいどもうとするが…。
   (answer:NAME+PRO → predict:PLACE)

表 4: 各タグにおける系列長別の間違った系列の割合

	1	2	3	4	5	6	7	8	9
NAME	0.026(53/1991)	0.28(153/550)	0.49(59/120)	0.47(8/17)	1.0	-	-	-	-
MF	0.015(5/339)	0.67(2/3)	0.83(5/6)	1.0(1/1)	-	-	-	-	-
AGE	0.038(13/334)	0.29(15/51)	0.16(4/25)	1.0(3/3)	-	-	-	-	-
STATE	0.19(41/216)	0.75(47/63)	0.84(31/37)	0.89(16/18)	1.0	1.0	1.0	1.0	0.0
PRO	0.061(34/555)	0.31(109/353)	0.38(66/173)	0.45(27/60)	0.76	0.70	1.0	1.0	-
AFF	0.088(11/125)	0.34(60/177)	0.43(28/65)	0.43(13/30)	0.47	0.56	0.0	1.0	-
OTHER	0.099(49/495)	0.59(141/238)	0.65(71/109)	0.65(22/34)	0.92	0.80	1.0	1.0	1.0
PLACE	0.060(37/616)	0.39(119/307)	0.50(82/164)	0.59(20/34)	0.69	0.80	0.80	1.0	-
REL	0.037(21/565)	0.55(34/62)	0.58(25/43)	1.0(6/6)	-	1.0	-	-	-

1 に関しては，教師データに漢字 1 字の名前が存在し，教師データとテストデータでの出現パターンが類似していたことが原因と考えられる．2 に関しては，“タリオ大公” など（カタカナ名前）＋（漢字二文字の立場）に場所タグを誤付与したり，“マウゼル教” など（カタカナ＋漢字一字の所属）に NAME を誤付与している．これは，“アメリカ合衆国” のようにカタカナの後に国を表す漢字が付いて PLACE になるもの，“ヴラド卿” のようにカタカナの後に敬称の漢字が付いて NAME になるものからパターンを学習したことが影響していると考えられる．また，“タリオ大公” に関しては，文字 uni-gram と文字 bi-gram のみの場合，正解していることを確認した．これは，タグフラグにおける PLACE 候補中に「大」が含まれていることが影響したと考える．

表 2 から STATE（容姿や特性表現）の再現率が低いことが確認できるため，各タグの系列長ごとに間違った系列の割合を調査した．その結果を表 4 に示す．表中の横線は該当系列が存在しなかったことを表している．表 4 より，STATE は系列長に関係なく不正解率が他のタグより高いことがわかる．同等に不正解率の高い MF タグと比較すると，MF は，出現する系列数（分母）が少ないため，不正解率が高くなっていると推測できる．一方，STATE は，出現する系列数（分母）が一定数存在するにも関わらず不正解率が高い．このことから，STATE は，表現が多様であるため各系列長での事例が多岐にわたり，それぞれの出現数が教師データに少ないことが影響していると考えられる．さらに，表からは系列長が長くなるほど不正解率が高くなっていることがわかる．これは，CRF で注目できる系列長より正解の系列長が長い場合，ラベリングが難しいことを意味していると考えられる．

6 おわりに

本稿では，CRF を用いて，ファンタジー小説のあらすじから人物情報を抽出する手法を提案した．性能評価の結果，全素性を追加したモデルが最高性能を得た．関連研究との比較により，名前抽出には系列ラベリングが有効である可能性を確認した．また，エラー分析から文字種やタグフラグに含まれる候補によってラベリング誤りが発生すること，系列長の長い人物情報に対して不正解率が高いことを確認した．

今後は，長い系列に対応するため深層学習モデルによるラベリングについて検討する．その後，人物情報と人物名の紐づけや，登場人物間の関係を抽出する手法を検討し，小説検索システムを構築する．

参考文献

[1] 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, pp. 574–577, 2007.

[2] 米田崇明, 篠崎隆宏, 堀内靖雄, 黒岩眞吾. 述語情報を利用した小説の登場人物の抽出. 言語処理学会第 18 回年次大会発表論文集, pp. 855–858, 2012.

[3] 神代大輔, 高村大也, 奥村学. 物語テキストにおけるキャラクタ関係図自動構築. 言語処理学会第 14 回年次大会発表論文集, pp. 380 – 383, 2008.

[4] 国立情報学研究所（NII）. Webcat Plus. http://webcatplus.nii.ac.jp/.

[5] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.