1Q-8

物語文における登場人物名抽出

山﨑 堅寛† 浦谷 則好†

†東京工芸大学大学院工学研究科

1． はじめに

筆者らは物語文からのアニメーション自動生成の研究を進めている。このため，まず登場人物の抽出が必要となる。

登場人物を特定するための情報として，辞書の利用，登場人物が存在すると考えられる文の特徴，本文中での出現回数，前後文での関連語の利用などが考えられる。

物語においては動物などが登場人物として扱われていることが多いため，人名辞書を用いてもすべての登場人物を特定することはできない。そこで，本研究では辞書と文の特徴，出現回数を用いて登場人物の抽出を行う手法を提案する。

2. 関連研究

馬場らの研究[3]では「8 万人西洋人名よみ方綴り方辞典」から人名を収集し ChaSen の辞書に追加し，形態素解析を行っている。対象は青空文庫に収録されている英米文学の推理小説 4 件としている。ChaSen により形態素解析した結果，品詞が「名詞-固有名詞-人名-一般」，「名詞-固有名詞-人名-姓」，「名詞-固有名詞-人名-名」と解析された形態素を人名として抽出する。実験結果として精度 55.2％～73.9%，再現率 35.3～53.3%が得られている。

3. 登場人物の抽出手法

本研究における登場人物抽出の手法は図 1 に示す通りである。

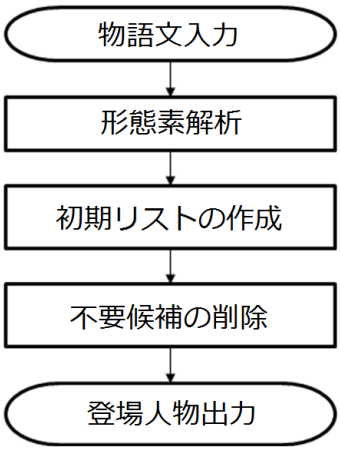

図 1 登場人物抽出の流れ

Extraction of the characters in narrative
†Takahiro Yamazaki and †Noriyoshi Uratani,
Graduate School of Engineering,
Tokyo Polytechnic University

登場人物は，物語などに登場する人物や動物といった存在である。本研究ではそのなかで「会話」または「動作」を行なっているものに限定する。

3.1 初期リストの作成

登場人物を抽出するために，物語文から単語を切り出す。そこから，再現率を下げないように登場人物の候補となり得る単語列をすべて取得する。

すなわち初期リストは，物語文を形態素解析器により単語区切りし，順番を変えずに文末まで連結して作成する。それを一単語ずつずらしてすべてのパターンを作成する。

例えば「いなかのねずみが町へ行きました。」という文の場合は，12 形態素で構成されているので 78 通りの候補が得られる。そして，作成された候補の本文中での出現回数を取得する。

最終的に，候補と出現回数，形態素解析の結果の 3 種類の情報をまとめて初期リストとする。(表 1)

候補名	回	形態素情報
いなか	7	いなか 名詞,一般,*,*,*,*,いなか,イナカ,イナカ
いなかの	6	いなか 名詞,一般,*,*,*,*,いなか,イナカ,イナカ の 助詞,連体化,*,*,*,*,の,ノ,ノ
いなかのねずみ	6	いなか 名詞,一般,*,*,*,*,いなか,イナカ,イナカ の 助詞,連体化,*,*,*,*,の,ノ,ノ ねずみ 名詞,一般,*,*,*,*,ねずみ,ネズミ,ネズミ

表 1 初期リスト例

3.2 不要候補の削除

登場人物になり得ない単語列を削除するために以下の処理を行う。

登場人物を確定する重要な手がかりとして，以下の 2 つを制約とする。

・候補の最後が名詞か名詞接尾である。

・物語文全中で登場人物名候補の後ろに「の」以外の助詞を持っている。

この理由は，抽出する登場人物は人間か動物であり，抽出する登場人物の条件で，登場人物は会話または動作を行なっているからである。

上記の条件を抽出の鍵とし，その上で残る不要候補を以下の登場人物名としてふさわしくないと考えられる条件を用いて，候補から削除する。

(1)登場人物名候補の最初の品詞が
・助詞
・助動詞
・接続詞
・接尾
の場合，候補から削除する。
(2)登場人物名候補のもつ全ての素性のいずれかに
・記号
・日付
・接続助詞
・名詞-副詞可能
・算用数字
・フィラー
・係助詞
・オノマトペ
が含まれている場合，候補から削除する。
(3)登場人物名候補の最後が
・代名詞
・「気持ち」など漢字とひらがなで構成される名詞の場合，候補から削除する。
(4)その他の素性情報による削除
・登場人物名候補の中に「の」以外の助詞があって，それより後ろの単語に動詞か助詞の「の」が含まれていないもの。
・文をまたぐもの（句点などを含むもの）や読点を含むものを候補から削除する。
・ひらがな，カタカナが一文字のみのものを候補から削除する。
・「さん」や「さま」など，名詞-接尾が単独のものを候補から削除する。
(5)同一候補による重複の削除
３つの方法で重複候補を削除する。
・全く同じ候補を削除する。
・同じ語を含み出現回数が同じなら形態素数が少ないものを削除する。
・同じ語を含む場合は，出現回数が少ないものを削除する。
(6)出現回数を用いた除去
・出現回数が1回以下のものを削除する。

3.3 日本語語彙体系を用いた削除

抽出する登場人物は動物なども含まれているが，共通して「意思を持つことができる」という要素を持っている。そこでこの辞書の一般名詞意味属性体系の意味属性のうち，意思を持てる「人」「組織」「生物」の 3 属性のいずれかが含まれていない場合はその候補を削除する。

日本語語彙体系による削除の対象を候補の主体に絞った場合の実験も行った。

4. 実験および結果

物語は福娘童話集[1]から手動で取得した 31 話を対象とした。

抽出した登場人物候補を筆者自らが作成した正解登場人物リストと比較し，性能を評価した。

ガ格の前にあるものが登場人物である傾向があることから，比較のための実験も行った。

表 5　追加実験を元にした比較

手法	精度	再現率
形態素＋ 出現回数２回以上	15.7% (95/604)	73.2% (90/123)
ガ格による制約	21.2% (68/321)	55.3% (68/123)
語彙体系による 削除	17.4% (75/430)	61.0% (75/123)
語彙体系による 削除（主体語）	19.2% (74/385)	60.2 (74/123)

この結果から，本手法はガ格の前に限定するよりも，より多くの登場人物を取得することができることが判明した。

5. おわりに

馬場らの研究結果と比べると精度で劣っているが，抽出対象となる登場人物が人間に限定されていないなど対象が異なるため，単純に比較することはできない。

本研究では形態素解析で得た品詞情報を主に用いて登場人物の抽出と，シソーラスとして日本語語彙体系の意味属性を用いて，登場人物候補が意思を持てるものか否かの判別を行うことで精度の向上を行なった。さらに同一登場人物で複数候補得られるものを同一化する手法を導入することで重複の削減も行なった。

現在は登場人物名そのものが持っている素性などの情報と直後の助詞の情報をメインに抽出を行なっている。しかし，それ以外にも登場人物名候補にかかっている動詞にも，それが人物であるという情報を持つものがあるため，動詞の格要素を利用することで性能の向上を図りたい。

参考文献

[1] 福娘童話集 http://hukumusume.com/douwa/
[2] NTT コミュニケーション科学研究所，「日本語語彙体系」，岩波書店
[3] 言語処理学会 第 13 回年次大会　馬場 こづえ 小説テキストを対象とした人物情報の抽出と体系化
[4] 第 11 回 情報科学技術フォーラム　山﨑 堅寛 物語文からの登場人物抽出