E-020

物語文からの登場人物抽出

Extraction of the characters in narrative

山﨑 堅寛† 浦谷 則好†
Takahiro Yamazaki Noriyoshi Uratani

1． はじめに

物語文からのアニメーション自動生成などのために、登場人物の抽出の研究が行われている。

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。

既に、人名辞書を用いた登場人物の抽出が小説を対象として行われている。しかし、物語文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、人名辞書を用いた抽出ではすべての登場人物を特定することができない。そこで、本研究では文の特徴を利用して登場人物の抽出を行う手法を提案する。

3. 関連研究

馬場らの研究[1]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、形態素解析を行う。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。実験結果は精度 55.2%～73.9%、再現率 35.3～53.3%である。

3. 本文からの取得

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。

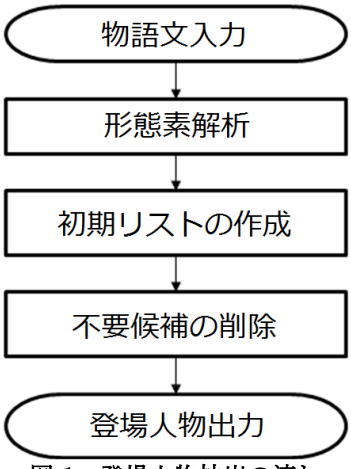

図 1 登場人物抽出の流れ

本研究では登場人物を以下に示す２つの条件のいずれか、もしくは両方を満たすものとする。

「発話を行なっているもの」

「動作を行なっているもの」

対象を物語としているので、登場人物にはウサギやネズミなども存在しており必ずしも人とは限らないが、それらも登場人物として抽出する。

なお、「いなかのねずみ」と「いなかに住んでいるねずみ」があった場合は、どちらも取得する。

3.1 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取ったリストを作成し、そこから登場人物になり得ない情報を利用して、不要候補を削除する。できる限り登場人物を落とさずに、再現率を確保するため、この手法を採る。

初期リストは、物語文を形態素解析器により単語区切りにし、順番を変えずに文末まで連結していく。それを一文字ずつずらしてすべてのパターンを作成する。

１２語からなる以下の例文の場合１０５通りの候補が得られる。表 1 は 5 形態素以内で「、」と「。」を含むものを除いたものであり、４０通りとなる。

例文：中国のテングは、偉そうな態度で言いました。

表 1 初期リストの例

中国	中国の	中国のテング	中国のテングは	
の	のテング	のテングは		
テング	テングは			
は				
偉	偉そう	偉そうな	偉そうな態度	偉そうな態度で
そう	そうな	そうな態度	そうな態度で	そうな態度で言い
な	な態度	な態度で	な態度で言い	な態度で言いまし
態度	態度で	態度で言い	態度で言いまし	態度で言いました
で	で言い	で言いまし	で言いました	
言い	言いまし	言いました		
まし	ました			
た				

†東京工芸大学大学院

3.2 不要候補の削除

以下のような登場人物名としてふさわしくないと考えられるものを、候補から削除する。

・文をまたぐもの（句点などを含むもの）や読点を含むもの
・「さん」や「さま」など、名詞-接尾が単独のもの
・ひらがな，カタカナが一文字のみの登場人物名
・オノマトペが含まれているもの
・フィラーが含まれているもの
・記号及び算用数字が含まれているもの
・日付が含まれているもの
・「気持ち」など形態素解析で「名詞-一般」となる漢字とひらがなで構成されるもの
・代名詞が末尾にあるもの
・先頭に接続詞，助詞，助動詞，名詞-接尾があるもの
・物語文中で登場人物名候補の後ろに「の」以外の助詞がないもの
・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

3.3 実験および結果

対象とする物語は福娘童話集[2]から手動で取得した 31 話とし、登場人物候補を自動で抽出した。

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。

現段階では精度 5.5%、再現率 86.1%が得られた。抽出精度が低い理由は主に「足音」や「木の下」など登場人物になりえないものや位置、掛け声なども「名詞-一般」として形態素解析器で判別されているためである。

4. 追加実験

4.1 形態素以外での不適切候補の削除

形態素情報以外に本文中での出現回数によって不要候補の削除が行えないかを、純粋な出現回数と出現割合を用いた場合とで、それぞれ調べた。

表 2、表 3 の結果から出現回数が低いものを削除することで、精度を高めることができることが分かった。本研究において再現率は重要な要素なので、再現率をできるかぎり落とさずに精度を高めることを考えると、出現回数が 2 回未満のものを削除するのが適切といえる。

表 2 候補の出現回数と精度

1:	*1230/	*2 1941*3	P= 11.85% R= 90.16% F= 20.95%
2:	127/	604	P= 21.03% R= 81.15% F= 33.40%
3:	106/	383	P= 27.68% R= 72.13% F= 40.00%
4:	97/	288	P= 33.68% R= 66.39% F= 44.69%
5:	86/	217	P= 39.63% R= 58.20% F= 47.15%
6:	73/	166	P= 43.98% R= 49.18% F= 46.43%
7:	66/	138	P= 47.83% R= 44.26% F= 45.98%
8:	61/	117	P= 52.14% R= 40.98% F= 45.89%
9:	50/	87	P= 57.47% R= 34.43% F= 43.06%
10:	46/	73	P= 63.01% R= 31.15% F= 41.69%
11:	42/	62	P= 67.74% R= 27.87% F= 39.49%
12:	33/	49	P= 67.35% R= 21.31% F= 32.38%

*1 出現回数, *2 正解候補数, *3 候補数

表 3 候補の削除割合と精度

20:	*4	604*5	P = 21.03%	R = 81.15% F = 33.40%
30:		467	P = 23.98%	R = 73.77% F = 36.20%
40:		314	P = 32.80%	R = 70.49% F = 44.77%
50:		198	P = 41.92%	R = 57.38% F = 48.44%
60:		135	P = 49.63%	R = 45.90% F = 47.69%
70:		77	P = 62.34%	R = 31.97% F = 42.26%
80:		34	P = 73.53%	R = 15.57% F = 25.70%

*4 削除割合. *5 候補数

4.2 題名からの抽出

物語において登場人物名が題名に利用されていることが多いことに着目し、初期リストを題名のみに絞った場合の検証も行った。

表 4 題名からの抽出結果

	精度	再現率	F 値
全パターン	5.9%	17.5%	8.8%
不要候補除去後	24.7%	15.8%	19.3%

この結果から、題名から取得された初期リストからの登場人物抽出に重みを負荷することで、精度が向上することが予想される。

4.3 題名からの抽出

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、比較のために追加実験を行った。

表 5 追加実験を元にした比較

	精度	再現率
形態素	11.8%	90.1%
形態素＋出現回数２回以上	21.0%	81.1%
ガ格の前	31.5%	52.8%

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た。

5. おわりに

馬場らの研究結果と比べると数値で劣っているが、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を行なった。そこで、シソーラスなどを用いて意味解析を行い、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が図れると考えられる。さらに同一文で複数候補得られるものを同一化する手法を検討することで精度の向上を図れると考えられる。

参考文献

[1] 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
[2] 福娘童話集 http://hukumusume.com/douwa/