6P-1

「文章からの人物関係抽出」

高本恵利花[†]　浦谷則好[†]
[†]東京工芸大学工学部コンピュータ応用学科

1　はじめに

近年はインターネットが普及し,誰もが情報を発信できるようになった.このことにより,紙媒体でしかありえなかった小説が,「電子小説」として販売されるようになった.また,「携帯小説」などアマチュア作家が気軽に小説を全世界へ公表することができるようになった.

こうして,プロだけでなくアマチュアの小説も気軽に読むことができるようなったことで,ユーザが読める小説は膨大な量となった.そこで,ユーザが読みたい小説を探すのに役立つように,本研究では文章から登場人物の情報の抽出およびその人間関係の自動作成を目指す.

2　関連研究

馬場,藤井は,英米文学の推理小説を対象にテキストから自動的に人物相関図を作成した[1].その手法を,図1に示す.

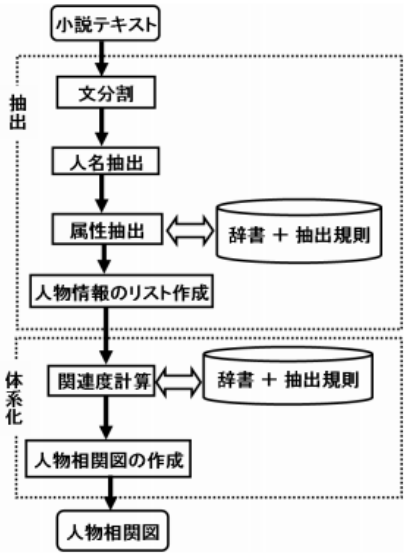

図1:馬場ら[1]の手法

馬場らの研究では人物相関図を作成することに成功している.しかしその出力画面は,関係線のみで人物間の関係の分類をしていない.情報としては,登場人物の人間関係の抽出も必要だと考える.また,人名抽出の精度を上げるために,登場人物名がカタカナである英米文学を対象としている.本研究では日本文学への対応も可能にする.

横野らは,登場人物の推定をカタカナ名に絞らず行っている[2].だがこの研究は物語要約を出力としているため,人物間の関係は出力していない.

3　提案手法

本研究のフローは,図2のようになる.

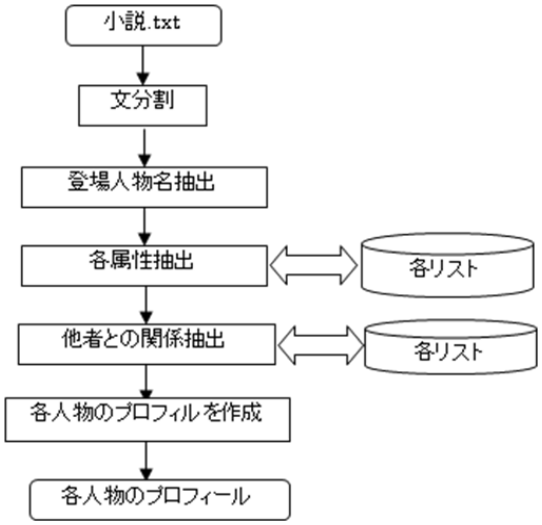

図2:本研究の手法

3.1　入力

本研究で対象とする小説として青空文庫に収録されている作品から以下を選定した.

小説Ⅰ　コナン・ドイル,『白銀の疾走』
小説Ⅱ　エドガー・A・ポー,『モルグ街の殺人』
小説Ⅲ　島崎藤村,『夜明け前　第一部上』

小説ⅠとⅡは,馬場らの研究で入力に使用しているため,抽出の精度と再現率の比較のために選んだ.

小説Ⅲは,登場人物が日本名の小説である.本研究の日本文学への対応を確認するため選定した.

3.2　文分割

小説によっては改行が少ないものもある.正しい係り受け解析を得るためにそのテキストを,句読点“.”またはかぎ括弧“」”を文末とし,改行を入れて文分割を行う.

3.3　登場人物名抽出

文分割で得た形態素解析の結果から登場人物名

Extraction of person relation from text.
Erika TAKAMOTO[†],Noriyoshi URATANI[†]
[†]Department of Applied Computer Science,Faculty of Engineering,Tokyo Polytechnic University

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.

を抽出規則によって抽出する.抽出規則としてＡ～Ｃの三通りを使用する.また,本研究では実際の登場人物を“二回以上登場している人物”とするため,考察にある馬場らの結果と数値が異なっている.

手法Ａ　形態素解析器の人名辞書を使用した手法.

Ａ１:解析結果が“名詞,固有名詞,人名,(一般|姓|名)”

Ａ２－１: 助詞(は|が)を後接した形で出現

Ａ２－２:複数回出現

上記Ａ１とＡ２のいずれかが適合した単語を人物名とする.

手法Ｂ　人名辞書に頼らない手法.

Ｂ１:解析結果が“名詞,(一般|固有名詞)”

Ｂ２:文節中で単語の前後が同じ表記
(ひらがな or カタカナ or 漢字)
もしくは“・”があった場合
これらの場合は前後の単語を結合

Ｂ３: 助詞(は|が)を後接した形で複数回出現

上記工程をすべて適合した単語を人物名とする.

手法Ｃ　敬称基づいた手法.

Ｃ１:敬称の前または後にある“名詞,(一般|固有名詞)”
(敬称リストによる)

Ｃ２:文節中で単語の前後が同じ表記
(ひらがな or カタカナ or 漢字)
もしくは“・”があった場合

Ｃ３: 助詞(は|が)を後接した形で複数回出現

これらの場合は前後の単語を結合した上で,上記工程すべて適合した単語を人物名とする.

3.4　文変換

抽出した人物名を基に,台詞の話し手と一人称に対し人物名を推定し付与する.

3.5　各属性抽出

性別(男:247 語,女:357 語),**年代**(277 語),**職業**(1294 語),**性格**(3182 語)は各リストの語句が文章中に出現した場合,その語句の係り受けから人物を推定し付与する.

身体的特徴(1725 語)の抽出は,
「身体的特徴語」(は|が)({形容詞}|{名詞})
という条件を満たした場合,抽出する.この時の人物推定は上記の性別等の抽出と同様である.

3.6　他者との関係抽出

登場回数,台詞回数から主人公の推定を行う.

また会話回数から緊密度を推定する.ここでの緊密度の計算は点数付けで求める.感情リストの語句があった場合,発言者が対象者に緊密関係にあるとし,点数を付けていく.相手への好意の語があればプラスを付与,敵対する語があればマイナス値を付与する.

４　考察

人物名抽出の精度(正解した単語数/抽出した単語数)と再現率(正解した単語数/本来の登場人物数)について検証した結果,表１のようになった.

表１:提案手法の精度・再現度

手法		小説Ⅰ	小説Ⅱ	小説Ⅲ
A	精度	75.00% (6/8)	33.33% (2/6)	31.31% (31/99)
	再現率	37.50% (6/16)	10.52% (2/19)	63.26% (31/49)
B	精度	32.60% (15/46)	8.16% (4/49)	10.93% (21/192)
	再現率	93.75% (15/16)	21.05% (4/19)	42.85% (21/49)
馬場らの結果	精度	40.6% (13/32)	53.3% (16/30)	
	再現率	72.2% (13/18)	55.2% (16/29)	

手法Ａは精度が高く,手法Ｂは再現率が高くなるといった結果が出た.しかし,小説Ⅲは,手法Ａの方が精度も再現率も高くなっている.

手法Ａでは辞書に載っていない名前は抽出できない.また登場人物名を形態素解析器が語の途中で区切ってしまい,再現率が落ちている.

手法Ｂによって,辞書に載っていない人物名も抽出することができた.しかし,抽出した単語に代名詞や,関係のない名詞まで出力されてしまい,精度が落ちてしまった.さらに,和名は敬称まで氏名に結合してしまった.また,『グレゴリ警部』のように敬称付けで書かれている場合,『グレゴリ』はＢ３の条件を満たせず抽出漏れが起きている.

そこで,手法Ｃを考案した.手法Ｃによって,敬称付きで登場している人物も氏名を確実に抽出することができると考える.手法Ｃはまだ未実施である.

５　おわりに

本研究では文章から人物相関図を抽出する手法を提案した.形態素解析器と係り受け解析の結果から人物名の抽出をし,本手法の精度と再現率について検証を行った.　小説によって結果が大きく変動することが判明した.今後は人物名の候補が「意思性のある動作文の動作主」となることなどを利用して,精度と再現率の向上を目指したい.

属性抽出と関係抽出の評価はまだ未実施であるが,最終的には各登場人物のプロフィールを作成し,このプロフィールから人物相関図上に人物間の人間関係を付与していきたい.

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 vol. 13, pp. 574-577, 2007.
[2]横野光:登場人物の感情表現に着目した物語要約,言語処理学会 vol. 13, pp. 1141-1144, 2007.
[3]大野晋,浜西正人:類語新辞典,角川書店,1981.
[4]中村明:感情表現辞典,東京出版,1993.

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.