E-008

格フレームと日本語 WordNet を用いた小説文中の登場人物抽出

Extraction of Characters in Novels Using Case Frames and the Japanese WordNet

加守田 侑† 上野 敦志† 田窪 朋仁†
Yu Kamoda Atsushi Ueno Tomohito Takubo

1. はじめに

電子書籍やオンライン小説により多くの小説が日々、電子化されている。しかし、小説の数が増えるにつれ読者は嗜好に合う小説を探すことが難しくなっている。このため小説を探す手がかりが必要となる。手がかり作成には小説の理解が必要であり、登場人物は重要な情報である。

登場人物を抽出する手法としては人名辞書、出現回数、文の係り受け関係を用いる手法などが考えられる。人名辞書を用いる手法は、小説固有の名前や人以外の動物が登場人物となる小説では汎用的ではない。そこで、本研究では最初に規則に基づいて人物候補を抽出し、人物候補の局所出現頻度と人物候補に係る動詞から登場人物の識別を行う。

2. 関連研究

馬場ら[1]の研究では、形態素解析辞書に西洋人人名辞書を追加し英米文学 4 作品に対して登場人物の抽出を行っている。実験結果として、適合率 35.3~53.3%、再現率 55.2~73.9%が得られている。

米田ら[2]の研究では、最初に登場人物は必ず主語として出現すると仮定する。訓練データにおいて、述語の動詞が登場人物に係る回数から、その動詞が人物に係る確率を算出し、登場人物の局所出現頻度とともに線形判別分析に利用してテストデータから登場人物を識別する。結果として適合率 60.3%、再現率 91.9%、F 値 71.5%が得られている。しかし、この手法では訓練データに含まれない動詞に関して人間に係る確率を得られないという問題がある。

3. 提案手法

提案手法は米田ら[2]の手法をベースとする。米田らは動詞が人間に係る確率を訓練データから算出していたが、提案手法では格フレームと日本語 WordNet を用いて動詞が人間を表す単語とどの程度係りやすいかを動詞の人間関連度として算出し、SVM で登場人物を識別する。提案手法の流れを図 1 に示す。

3.1 過分割修正

前処理として、本文を一文ごとに分割し形態素解析・係り受け解析・固有表現抽出を行う。その際、話し言葉は形態素解析精度が下がるため会話文は無視する。

形態素解析では形態素を過分割する場合がある。これは未知語の解析時に多く起こるため、人名の場合は起こりやすい。過分割により人名が複数の文節に分かれてしまうと抽出できないので、固有表現抽出を行い、登場人物になる可能性の高いと考えられる固有表現『PERSON』と判定された形態素列の一部(姓もしくは名と考える)と同じ文字列

† 大阪市立大学大学院 工学研究科 電子情報系専攻

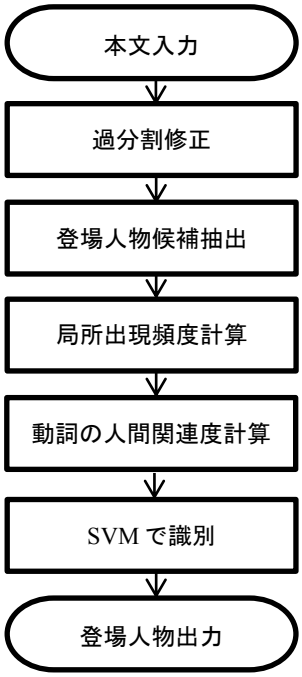

図 1　提案手法の流れ

が複数の形態素に分かれているとき、人名として一つの形態素に修正する処理を行う。

3.2 登場人物候補抽出

主語として 3 回以上出現した形態素列を登場人物候補として抽出する。その際、主語は最後の形態素が助詞『が』・『は』で終わる文節とし、末尾の助詞を削除したものを人物候補とする。また主語の一つ前の文節が並列助詞『と』で終わる際も主語の対象文節とする。抽出した人物候補の内、平仮名・片仮名一文字や『これ』『それ』など人物になりえないと思われる単語はストップワードとして除外した。また、固有表現『PERSON』と判定された形態素列において、主語として出現するものは出現回数に関係なく登場人物候補とした。

3.3 局所出現頻度

米田ら[2]の手法と同様に局所出現頻度を算出する。 登場人物は連続した文で多く出現すると考え、一定の連続した文(これを窓とよぶ)で登場人物候補が出現する頻度を算出する。窓を本文の先頭から終端まで一文ずつずらしていき、得られた出現頻度の最大値を局所出現頻度として算出する。また長い区間で見た際に多く出現する単語と特定の短い区間にのみ多く出現する単語は重要語である場合が多いので、窓は短い窓(文数 20)と長い窓(文数 100)の二つを用いる。

3.4 動詞の人間関連度

3.4.1 格フレーム

格フレームは用言と共起する名詞を格ごとに集めたものである。本研究では京都大学格フレーム[3]を用いる。登場人物は主語として出現すると仮定するので、格フレームのガ格名詞を利用する。登場人物候補に係る動詞の格フレームを取得し、共起数からそれぞれのガ格名詞と動詞の共起確率を求める。

3.4.2 日本語 WordNet

日本語 WordNet[4]は日本語用の概念辞書である。単語を概念ごとにグループ化し、それぞれの概念を上位・下位関係で結び付けた階層構造となっている。日本語 WordNet を用いて、ガ格名詞が含まれる概念と概念辞書の項目『人間』との意味的類似度を Wu ら[5]の手法で算出する。

3.4.3 動詞の人間関連度計算

動詞が人間を表す単語とどの程度係りやすいかを動詞の人間関連度とよぶ。動詞の格フレームで得られたガ格名詞それぞれに対し、ガ格名詞が含まれる概念と概念『人間』の意味的類似度を計算し、閾値以上のものを概念『人間』に近い意味を持つガ格名詞とする。閾値は 0.01~0.99 まで 0.01 ずつ増加させ、訓練データで最も F 値が高くなる値を用いる。

動詞の格フレームに含まれるガ格名詞の中で、概念『人間』に近い意味を持つ名詞の共起確率を足し合わせたものを、その動詞の人間関連度とする。

3.5 Support Vector Machine

SVM にはソフトマージン非線形 SVM を用いる。カーネル関数はガウシアンカーネルを使用する。SVM の入力素性は登場人物候補の局所出現頻度(二次元)と登場人物候補に係る動詞の人間関連度の平均(一次元)を合わせた三次元の素性とする。

4. 評価実験

実験には米田ら[2]と同じ青空文庫[6]の小説 30 作品を使用する。30 作品のうち 29 作品を訓練データ、1 作品をテストデータに分け、クロスバリデーションを行う。形態素解析器に MeCab ver0.996、 MeCab の辞書として新語辞書 mecab-ipadic-NEologd[7] ver0.0.2、構文解析器に CaboCha ver0.69 を用いる。 mecab-ipadic-NEologd は新語に対応した辞書であり、人名辞書も含まれている。

過分割修正や新語辞書を用いない場合との精度比較も含めた実験結果を表 1 に示す。MeCab の辞書に新語辞書を用いない場合は MeCab のデフォルト辞書である IPA 辞書を用いて解析する。表 1 の再現率において登場人物総数が異なっている理由は登場人物候補として抽出した中に含まれる登場人物数を正解の総数として計算しているためである。それぞれの場合で抽出された登場人物候補に含まれる登場人物とその他の内訳を表 2 に示す。また、正解データは抽出された登場人物候補を筆者が判別して作成した。

実験結果より、過分割修正と新語辞書をどちらか一方だけ用いると適合率が大きく下がる。しかし、過分割修正と新語辞書を両方使用した場合は適合率が高くなっているためそれぞれの欠点を補い合ったと考えられる。また、両方使用しない場合と比べて精度に大きな差がないため、取得された登場人物数の多い両方使用する場合が 4 つの中で最も良いといえる。

表 1　実験結果

手法	適合率	再現率	F 値
過分割修正 × 新語辞書 ×	74.4% (143/199)	72.2% (143/211)	71.9%
過分割修正 × 新語辞書 ○	67.8% (149/223)	72.1% (149/216)	68.2%
過分割修正 ○ 新語辞書 ×	71.2% (150/218)	74.3% (150/214)	71.6%
過分割修正 ○ 新語辞書 ○	74.2% (148/208)	71.9% (148/219)	71.4%

表 2　登場人物候補の内訳

手法	登場人物候補	登場人物	その他
過分割修正 × 新語辞書 ×	662	211	451
過分割修正 × 新語辞書 ○	669	216	453
過分割修正 ○ 新語辞書 ×	666	214	452
過分割修正 ○ 新語辞書 ○	669	219	450

関連研究の精度と比較する。馬場ら[1]の手法に比べると適合率・再現率ともに向上している。米田ら[2]の手法に比べると F 値にほぼ差はないが提案手法の方が適合率は高く、再現率は低い。これは提案手法の方が登場人物を正しく抽出することに特化していることを意味する。つまり、小説の理解のための登場人物抽出という目的を考えた際、提案手法に優位性があるといえる。

5. おわりに

本研究では格フレームと日本語 WordNet を用いて小説文中の登場人物抽出を行った。また、提案手法は既存手法に比べ適合率において優位性があることを示した。課題として、主語として出現しない登場人物や発話文のみに現れる登場人物を抽出できないことが挙げられる。今後は、得られた登場人物を用いて人物の属性情報などを抽出する手法の検討を行い、小説を探す手がかりとなる情報の作成を行う。

参考文献

[1] 馬場 こづえ, 藤井 敦, 石川 徹也, “小説テキストを対象に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).
[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 (2012).
[3] 河原 大輔, 黒橋禎夫, “高性能計算機を用いた Web からの大規模格フレーム構築”, 情報処理学会, vol.2006, No.1 (2006).
[4] Francis Bond, Timothy Baldwin, Richard Fothergill, Kiyotaka Uchimoto, “Japanese SemCor: A Sense-tagged Corpus of Japanese”, 6th International Global Wordnet Conference (2012).
[5] Zhibiao Wu, Martha Palmer, “Verb Semantics and Lexical Selection”, in ACL’94 (1994).
[6] 青空文庫, http://www.aozora.gr.jp/
[7] Toshinori Sato, “Neologism dictionary based on the language resources on the Web for Mecab”, https://github.com/neologd/mecab-ipadic-neologd (2015)