表記情報による物語文の登場人物の推定*

6R－6

鈴木理恵　鈴木由理　山口昌也　乾伸雄　小谷善行　西村恕彦
(東京農工大学　工学部　電子情報工学科)

1　はじめに

テキストからのキーワード抽出の研究は、データベース構築など応用範囲が広いため、多くの研究が行われている。[1]では、未知語を許容する製品紹介記事のキーワードの抽出法を提案しているが、従来のキーワード抽出は、形態素解析、構文解析、意味解析など、様々な解析処理が必要であり、未知語に弱いとされている。

これらの解析処理をできるだけ少なくして、未知語も許容できるキーワード抽出はできないだろうかという考えから、形態素解析だけを行った物語文を入力とし、表記情報から登場人物を推定して、さらに主人公を抽出するシステムを作成した。

2　手法について

2.1　変数の割当て

本システムは、形態素解析で人名詞、未知語、一般名詞と判断された単語を物語文から拾い出す。ここでいう未知語は、人名詞または名詞になりうると判断されたものを指す。これらを、主人公候補として扱い、それぞれの単語に対して、次のような変数を割り当てる。

t ····単語

$n(t)$ ····単語 t の文章中における出現回数

$l(t)$ ····「彼」、「彼女」という表現が出現した場合に、その前文に出現している候補の単語の出現回数に 0.5 を加算したときの総数

$Z(t)$ ····単語 t の接続する語による評価値計算

$$Z_0 = 0, \qquad (i = 1, 2, \ldots, n(t))$$

$$Z_i = \begin{cases} Z_{i-1} + 1 & \text{(接続すべきでない品詞と接続したとき)} \\ Z_{i-1} + 0.5 & \text{(あまり接続しない品詞と接続したとき)} \\ Z_{i-1} & \text{(どちらとも接続していないとき)} \end{cases}$$

$$Z(t) = Z_{n(t)}/n(t)$$

接続すべきでない品詞例：形容詞の活用語尾など

あまり接続しない品詞例：代名詞など

$Y(t)$ ····単語 t が、役職を表す名詞と接続した回数の $n(t)$ に対する割合

$S(t)$ ····単語 t が、人間につく接尾語と接続した回数の $n(t)$ に対する割合

$J(t)$ ····単語 t の助詞との接続を手がかりとした評価値

計算

$$J_0 = 0, \qquad (i = 1, 2, \ldots, n(t))$$

$$J_i = \begin{cases} J_{i-1} + 0.5 & \text{(「が」と接続したとき)} \\ J_{i-1} + 0.7 & \text{(「は」と接続したとき)} \\ J_{i-1} + 0.4 & \text{(「も」と接続したとき)} \\ J_{i-1} & \text{(どれとも接続しないとき)} \end{cases} \tag{1}$$

$$J(t) = J_{n(t)}/n(t)$$

$I(t)$ ····単語 t に係る動詞の意味素性の数による評価値

計算

動詞の格辞書で、単語 t に一番近い動詞と文の終わりから一番近い動詞を調べたとき、単語 t の意味素性が人間である確率をそれぞれ $p1$、$p2$ とする。

$$I_0 = 0,$$

$$I_i = I_{i-1} + p1 + p2 \qquad (i = 1, 2, \ldots, n(t))$$

単語 t に対する述語について、辞書で調べることができた回数を $m(t)$ とする。

$$I(t) = I_{n(t)}/m(t)$$

$R(t)$ ····単語 t が一回以上出現した段落の数

$K(t)$ ····単語 t の「が」、「は」、「も」との接続による評価値

計算

$$K_0 = 0, \qquad (i = 1, 2, \ldots, n(t))$$

$$K_i = \begin{cases} K_{i-1} + 0.25 & \text{(「が」と接続したとき)} \\ K_{i-1} + 0.5 & \text{(「は」と接続したとき)} \\ K_{i-1} + 0.1 & \text{(「も」と接続したとき)} \\ K_{i-1} & \text{(どれとも接続しないとき)} \end{cases} \tag{2}$$

$$K(t) = K_{n(t)}/n(t)$$

このなかの、式(1)は人物絞込み、式(2)は主人公決定と用途が違う。これらの式の加算する数値は、実験対象外の七編の物語に対して、「が」、「は」、「も」に

*Inference of Main Characters in a Story by Literal Information ,

Rie SUZUKI, Yuri SUZUKI, Masaya YAMAGUCHI, Nobuo INUI, Yoshiyuki KOTANI, Hirohiko NISIMURA
Tokyo University of Agric. and Tech., Dept. of Computer Science

接続した単語が、人間である割合と、そのなかでも主人公である割合を出した結果をもとに決定した。

2.2 登場人物絞込み

登場人物絞込みには、前述した $Z(t)$、$Y(t)$、$S(t)$、$J(t)$、$I(t)$ を利用する。方法は、$Z(t)$ について昇順、$Y(t)$、$S(t)$、$J(t)$、$I(t)$ について降順に候補を並べ替え、それぞれ順位づけをする。その順位の平均値 $H(t)$ がしきい値よりも小さい候補を人物であると判断する。このとき、未知語または人名詞のときは、係数(ここでは0.7)をかける。これは、これらが名詞よりも人である可能性が高いという調査結果から、これらを有利にするためである。ここでいうしきい値は、予備実験として、七編の物語に対して $H(t)$ を算出し、それぞれの物語の人物のなかで $H(t)$ が最大のものを拾い出して、その平均値とした。

2.3 主人公の決定

主人公は、前節で絞られた人物の中から決定する。主人公を決めるには、$l(t)$、$R(t)$、$K(t)$ を用いる。これらのどの数値も、大きいほど主人公の可能性が高いのではないかと考えられるので、各数値について降順に各候補を並べ替えて、それぞれ順位づけをする。その順位の平均値が最小のものを主人公であると決定する。

3 実験

3.1 実験方法

まず、東京農工大学西村・小谷研究室で作成されたJMORHACK(Japanese MORphological HACKer)によって形態素解析された物語文のファイルを作成して、そのファイルを作成したシステムに入力する。出力として、システムが登場人物と判断した単語と主人公を得る。

3.2 結果

実験は、十一編の物語に対して行った。その一例を表1に示す。

表1 「金のくびかざり」([3])の実験結果

登場人物の数	7
システムが人物と判断した数A	19
Aに含まれた登場人物の数	7
候補数に対するAの割合(%)	7.2
主人公は抽出できたか	○

この例では、候補数に対するシステムが人物と判断した数の割合は7.2 %であるが、十一編の物語の平均は約11 %であった。主人公抽出に関しては、一人に絞れるものに関しては9割成功した。また、人によって主人公の解釈が異なるものについては、候補の一人は抽出できた。

4 考察

主人公抽出は、9割成功したので、その鍵として、出現頻度、出現のまんべんなさ($R(t)$)、「は」、「が」、「も」との接続を用いることは正しかったといえる。

登場人物絞込みについては、主人公候補の単語のなかで、人物は $H(t)$ が小さい方に集まる傾向があったので、手がかりの収集の点は成功したといえる。しかし、物語によって人物を落としてしまったり、絞りきれなかったりとばらつきが大きかったので、しきい値の決め方に問題があったといえる。

5 おわりに

本研究は、主人公抽出の方法は正当性が認められたが、登場人物の絞込みの点でしきい値の決め方などの課題を残している。今後の課題として、物語中における人物表現の特徴の調査を、さらに視野を広げて行うことがあげられる。

参考文献

[1] 松尾比呂志：抽出パターンの階層的照合に基づく内容抽出法, 情報処理学会, 情報処理学会研究報告 94-NL-99, pp.9-16(1994).

[2] 小野 浩：金のくびかざり, 赤い鳥傑作集 坪田譲治編, 新潮文庫, pp.246-252(1955).