新聞記事からの人物情報の抽出＊

1L－5

江原 貴彦†　絹川 博之†
東京電機大学大学院　工学研究科

1．はじめに

近年、インターネットなどのコンピュータネットワークの発達などにより様々な情報を容易に得ることができるようになり、個人の扱うことのできる情報量は飛躍的に増加した。そして、その膨大な情報の中から必要な情報を抽出する技術が重要になってきている。そこで、本研究ではある人物がどのような人物か、どのようなことを行ったかなど、人物の情報を新聞記事から抽出するシステムを構築したいと考える。そこで、まず記事中から人間の姓名を認識できなくてはならないと考え、姓名を抽出する方式を考案した。

2．形態素解析を用いた姓名抽出方式

2.1 姓名の出現パターン

新聞記事に人名が出てくるパターンは以下のようになっている。

＜人名＞：＝＜姓名＞｜＜姓名前置語＞＜姓名＞
　　｜＜姓名前置語＞＜姓名＞＜姓名後置語＞
　　｜＜姓名＞＜姓名後置語＞
＜姓名＞：＝＜姓＞｜＜名＞｜＜姓＞＜名＞
　　｜＜姓＞＜空白＞＜名＞｜＜姓＞＜記号＞＜名＞
＜姓名前置語＞：＝＜職業＞｜＜接頭語＞＜地位＞
｜＜組織名＞＜地位＞｜＜組織名＞＜接頭語＞＜地位＞
＜姓名後置語＞：＝＜職業＞｜＜地位＞
　　｜＜接頭語＞＜地位＞｜＜組織名＞＜地位＞
　　｜＜組織名＞＜接頭語＞＜地位＞｜＜敬称＞

2.2 姓名抽出処理の流れ

2.2.1 形態素解析

形態素解析は奈良先端大学院大学の茶筌を用いて行った。形態素解析とは文章を単語単位に分割し品詞付けを行うことである。

2.2.3 姓名抽出

まず始めに形態素解析結果に対し姓名抽出テーブルを参照し、パターンマッチを行い、パターンに当てはまる語を姓名候補とする。

次に姓名候補に対し姓・名の判別を行う。これは姓・名の片方だけ姓と認識され名が姓名候補とされた場合や、同じ文書中で特定の場所だけ認識されなかった場合など、明らかに判別できる場合を除き姓名辞書とのマッチングにより判別を行う。

2.2.4 姓名抽出テーブル

これは姓名出現パターンにあわせ、例えば姓名前置語なら姓名前置語となりうる語を羅列してありこれと形態素をマッチングすることにより姓名の認識を行う。

2.2.5 姓名辞書

これは人名の姓・名それぞれを羅列してある辞書である。これとマッチングすることにより姓・名の判別を行う。

2.2.5 出力結果

以上の処理により姓名と認識された語に対し抽出を行った。

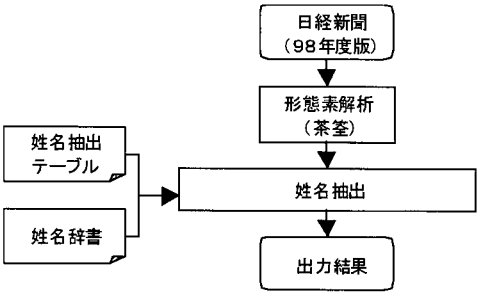

図1　姓名抽出プログラム構成

Information Extraction about Person from Newspaper's Articles
†Takahiko EBARA、Hiroshi　KINUKAWA
Graduate School of Engineering,Tokyo Denki University

3．実験と評価

3.1 実験データ

今回の実験では９８年度版日本経済新聞の記事から２００件をランダムに選択し姓名を抽出した。

3.2 評価方法

以下のように適合率と再現率を求めた。

$$精\quad度=\frac{抽出された適合姓名数}{抽出された姓名候補数}$$

$$再現率=\frac{抽出された適合姓名数}{抽出されるべき姓名数}$$

3.3 実験結果

実験結果を以下の表にまとめる。

表１　実験結果

	抽出されるべき姓名数	抽出された姓名数	抽出された適合姓名数	精　度［%］	再現率［%］
茶筌	２５４	２２６	２０３	89.8	79.9
茶筌＋姓名抽出	２５４	２３４	２２２	94.9	87.4

4．考察

4.1 精度の分析とその改良案

誤認識した１２件の内訳を以下に示す。

①姓の一部を名と誤認識：２件（16.7%）

②名の一部を姓と誤認識：１件（8.3%）

③姓名の一部が欠落：９件（75.0%）

①②は姓名辞書を用いることにより解決できる。

③は４パターンに細分することができる。

Ⅰ.姓を認識出来ず名だけ認識：４件（33.3%）

Ⅱ.名を認識出来ず姓だけ認識：１件（8.3%）

Ⅲ.姓が複数の形態素に分割され前部分が姓以外の語として誤認識：３件（25.0%）

Ⅳ.名が複数の形態素に分割され後部分が名以外の語として誤認識：１件（8.3%）

③の誤抽出は、形態素解析結果において姓名出現パターンに無い名詞として認識されてしまっている。よって、現在の姓名抽出テーブルは姓名出現パターンを元に作成したが、姓名の前後にこない形態素のパターンを加えることにより解決できる。

4.2 再現率の分析とその改良案

抽出出来なかった３２件の内訳を以下に示す。

①外人の姓名が誤分割：６件（18.8%）

②記事を書いた記者の姓名など前後に手がかり語がなく出現：７件（21.9%）

③力士や芸能人など一般的な名前ではなく、しこ名、芸名などで前後に手がかり語がなく出現：８件（25.0%）

④姓名抽出テーブルに含まれない手がかり語が出現：１１件（34.4%）

①については、カタカナ語の連続を一語と見なすことによって認識できる。

②記者の姓名が出現するのは記事の最初か最後と限られているので、そのことを利用すれば認識できる。

③については、芸名やしこ名など特殊な名を集めた辞書を作成し、それを参照する事によって認識できる。

④については、分析を進めテーブル情報を増やすことにより認識できる。

4.3 精度、再現率についてのまとめ

以上、4.1、4.2で示した改良案を組み込むことにより、現在の誤認識、未認識の半分以上は解決出来ると考えている。

5．おわりに

本研究では、新聞記事からパターンマッチにより姓名を抽出した。その結果

精度＝９４．９%、再現率＝８７．４%

で抽出することが出来た。今後、考察で述べた改良案を組み込むことで、さらなる再現率、精度の向上を目指し、それを元に人物情報の抽出を行う。今回は日経新聞９８年度版にて姓名抽出を行ったが NTCIR 等のテストコレクションにて同様の実験を行う予定である。

6．参考文献

1）岩波講座ソフトウェア科学１５　自然言語処理
長尾 真　編
岩波書店

2）言語と計算ー５　情報検索と言語処理
徳永 健伸　著
東京大学出版会