日本語新聞記事からの固有名詞情報抽出

7L－3

竹元義美† 山田洋志† 若尾孝博‡
(†NEC 情報メディア研究所 ‡シェフィールド大学)

1 はじめに

電子化されたテキストが容易に入手できるようになり、近年はむしろ大量のテキスト情報が氾濫していると言ってもよい。そこで大量のテキストから重要な情報を検索し、抽出する技術が求められている。とくに新聞記事は内容がポピュラーで、最近はCD-ROMで市販されたりWWWで公開されたりするようになり、検索・抽出技術の応用が期待されている。

テキストからの重要情報抽出として、新聞記事からの固有名詞の抽出技術を検討した。固有名詞は、新聞記事において重要な5W1H情報の要素となる。例えば、Who情報に人名・組織名が、Where情報に地名がなり得る。そこで固有名詞を人名・組織名・地名の3種類に分類して抽出することにした。

固有名詞抽出の従来技術として前後の単語との係り受けや共起情報を手がかりに推定する方法が研究されている[1, 2, 3]。例えば「XX社長」で、「XX」が未登録語または他品詞や他固有名詞と曖昧性を持つ語であっても「社長」(人名共起語)の前の「XX」は人名と推定できる。筆者らも基本的にこの手法を用いて、政治・経済に関する100の新聞記事から人名・組織名・地名の推定ルールを作成した。本稿では、ルールの汎用性を評価するため、政治・経済以外の社会・エッセイ・スポーツなどの新聞記事を対象に評価を行った。

2 固有名詞抽出処理の概要

図1に固有名詞抽出処理の流れを示す。まず入力テキストに形態素解析(Amorph)を実行する。形態素解析では基本辞書を用いて入力テキストを単語単位に分割する。解析結果に固有名詞辞書および共起語辞書を参照して、各単語に固有名詞および共起語の情報を付与した後、固有名詞抽出ルールを実行する。ルール処理は、固有名詞辞書情報よりも共起語情報を優先するように働く。ルール処理部では学習処理も行う。学習処理は、未登録語がルールにより固有名詞と認定されたときに、その未登録語を固有名詞として学習する。

以上の流れから固有名詞を人名・組織名・地名の3種類に分けて抽出する。以下、3節で辞書、4節で抽出ルールおよび学習処理について説明する。

3 辞書

固有名詞辞書は、21,205件(人名：8,223件、組織名：2,544件、地名：10,438件)のものを用いた。共起語辞書は、736件(人名：351件、組織名324件、地名61件)のものを用いた。共起語は、新聞記事(約6万文字：ルール作成に用いたものとは異なる)の分析をもとに固有名詞の前後に共起する単語およびそれから連想される単語を集めた。共起語の例を以下に示す。

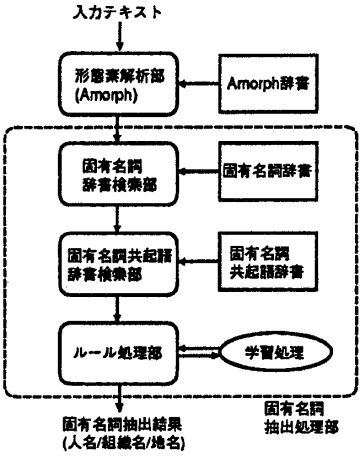

図 1: 固有名詞抽出処理の流れ

- 人名共起語：
  「～氏」「～大統領」「～社長」「故～」
- 組織名共起語：
  「～委員会」「～社」「～建設」「～組合」
- 地名共起語：
  「～県」「～市」「～郡」「～州」

4 固有名詞抽出ルールと学習処理

固有名詞抽出ルールは、基本的には共起語とその前後の字種をもとに固有名詞の認定範囲を決定する。例えば「金大統領(人名共起語)」では人名共起語の前の漢字列/片仮名列の「金」を人名とする。また、「東京都行政改革委員会(組織名共起語)」「ゴラン高原(地名共起語)」では共起語を含めて「東京都行政改革委員会」を組織名、「ゴラン高原」を地名と認定する。

基本ルールで対処できない例について詳細ルールを作成した。詳細ルールは100の新聞記事(約4万文字)の分析をもとに26種類のパターンを作成した。本節では主な抽出ルールと学習処理について述べる。

4.1 人名抽出ルール

人名抽出ルールでは、人名共起語の前の人名でない語の認定が必要である。人名共起語の前には人名だけでなく、「村山富市前首相」「ムバラクエジプト大統領」「金子尚志ＮＥＣ社長」の下線部のように接辞

The Extraction of Proper Names from Japanese News Text
Yoshikazu Takemoto†, Hiroshi Yamada†, and Takahiro Wakao‡
†Information Technology Research Labs, NEC Corp.
‡University of Sheffield

語・国名/県名・組織名などが付く。人名抽出ルールでは共起語からこれらの語をとばして人名を認定する。

4.2 組織名抽出ルール

組織名は共起語を含まない語も多いため、前後のパターンを手がかりにすることが必要である。とくに、次のような新聞記事特有の表現に注目し、下線部を組織名と認定する。

1. 漢字列/片仮名列＋組織名共起語＋“（”＋英字列
   (例1) パレスチナ解放機構（ＰＬＯ）
2. 漢字列/片仮名列/英字列＋“（本社”
   (例2) アコム（本社東京）
3. 漢字列＋未登録語の”委”/”取”/”審”
   (例3) 特別委

例1では、「機構」が組織名共起語で「パレスチナ解放機構」および「ＰＬＯ」を組織名と認定する。例2では、”（本社”というパターンの前の片仮名列「アコム」を組織名とする。例3の「特別委」は「特別委員会」の省略形であり「委」は未登録語と解析されるので組織名と認定する。同様に、「財政審」(財政審議会)、「香港証取」(香港証券取引所)などを組織名と認定する。

4.3 地名抽出ルール

地名認定には、基本ルールの他に「日」「米」などの国名の省略表現の認定が必要である。形態素解析で「米/大統領」「米/政府」のように切れる場合は、「米」と「大統領」および「政府」の共起関係から「米」を地名と認定する。

4.4 学習処理

学習処理は、ルールで固有名詞と認定された未登録語を固有名詞として学習する。新聞記事では、組織名の未登録語が最初に出現するときは4.2節の例1,2のように説明付きのパターンとなることが多い。この場合ルールを適用できる。しかし、2度目以降は組織名だけで出現するのでまた未登録語と認定されてしまう。

今回は片仮名列/英字列の組織名のみを学習の対象とした。組織名抽出ルールの例1,2では、「ＰＬＯ」「アコム」を組織名として学習する。

5 評価

固有名詞抽出精度を再現率R(Recall)、適合率P(Precision)により評価した。つまりシステムの出力結果と人手で作成した正解とを比較して検出洩れおよび検出誤りの件数を数えた。表1は前述のルール作成に用いた政治・経済に関する新聞記事、表2は社会・エッセイ・スポーツなどに関する32の新聞記事([4],約1.7万文字)の評価結果である。

人名・地名抽出は同等の好結果を得た。人名抽出で分野に限らず難しいのは、「歴代社長」の「歴代」のように人名共起語が付いても人名となり得ない語の認定である。地名抽出では、「福岡市・天神」の下線のように共起語のない地名の出現パターンの検討を要する。

組織名抽出は、3つの中で最も難しく分野が変わると精度が大きく落ちている。基本ルール以外の、政治・経済に頻出した特有表現に対するルールが効かなかったためである。また、今回選んだ記事には「AA、BBなど各社の」のように構文レベルまで解析が必要な例が多かった。

表 1: 評価結果(政治・経済記事：ルール作成用)

	実際の正解数	システム結果数	システム正解数	R (%)	P (%)
人名	298	276	240	81	87
組織名	578	469	410	71	87
地名	605	534	496	82	93

表 2: 評価結果(社会・エッセイ・スポーツ記事)

	実際の正解数	システム結果数	システム正解数	R (%)	P (%)
人名	79	84	71	90	85
組織名	127	83	57	45	69
地名	176	156	140	80	90

6 おわりに

政治・経済関連の新聞記事から固有名詞抽出ルールを作成し、別の分野の新聞記事で評価した。人名・地名抽出ルールは分野によらず有効なことを確かめた。組織名抽出ルールは構文レベルの解析が課題である。

今後は新聞記事以外のテキスト(ネットニュース、ホームページなど)での評価も行いたい。製品名・時間表現などの重要情報抽出も検討し、5W1H情報抽出ないし要約技術へと発展させたい。

謝辞 共起語辞書作成に協力してくださったNEC関西C&C研究所 宮部隆夫主任に感謝します。またルール作成にはARPAの情報抽出プロジェクトの研究用に提供された新聞記事を使わせていただいたことを感謝します。

参考文献

[1] “係り受け解析を用いた複合語の自動分割法”,宮崎,情処論文誌,Vol.25,No.6,1984

[2] “日本語処理における固有名詞実在性検定方式の検討”,高木・安田・島崎・池原,情処35全大,6S-3,1987

[3] “固有名詞の特定機能を有する形態素解析処理”,木谷,情処NL研,92-NL-90,1992

[4] 日経全文記事データベース日本経済新聞CD-ROM版,1994