コーパス中の特徴と文法的意味的情報を統合的に用いた新聞記事中の固有名詞認識

大石 巧　黒橋 禎夫　長尾 眞

京都大学大学院 工学研究科 電子通信工学専攻

{ooishi,kuro,nagao}@pine.kuee.kyoto-u.ac.jp

1 はじめに

現在の形態素解析はかなり高精度であるが[1]，その誤りの多くは固有名詞が原因と考えられ，形態素の単位で固有名詞を正しく認識することが求められている．また形態素解析以後の解析では文節を単位にすることが多く，したがって，文節内にある複合語が全体として固有名詞かどうか，それが人名であるか地名であるか等の情報も有用である．

従来，固有名詞の認識にはルールによるパターンマッチング手法が多く採られてきたが，ルールの作成にかかる人間の手間が問題となる．また，固有名詞はその性質(企業名などはいくらでも生成可能)から辞書登録という方法だけではすべてをカバーすることができない．さらに，地名から人名が転成することなどに見られるように，曖昧性の問題も存在する．同じ語が人名でも地名でもあり得ることに対して辞書は何も答えてはくれない．これに対して人間は文章中のさまざまな情報から，たとえ未知の語でも正しく固有名詞を認識することができる．

そこで，本稿では隣接する語，名詞句「A の B」，動詞の格，人名複合語の末尾語，辞書，並列名詞句，同一表記の語，縮約した語，などさまざまな情報を統合して固有名詞を認識する方法を提案する．

2 固有名詞認識の枠組み

固有名詞の認識は形態素解析(JUMAN[2]を用いた)と文節認識(KNP[3]を用いた)の処理後に行う．新聞記事を対象として用い，固有名詞としては人名，地名，組織名の3つを考え，扱う固有名詞の単位は形態素と文節内で最長の複合語とした．形態素での固有名詞認識の対象は名詞とした(ただし，「もの」「こと」などの形式名詞，副詞的名詞はのぞく)．複合語での固有名詞認識の対象は形態素解析で名詞，名詞接頭語，名詞接尾語と判定された形態素列のうち，最も長いものとした(「・」は含む)．「ファン・ボーベン氏が...」という文節では「ファン・ボーベン氏」を複合語と考える．

そして，形態素，複合語単位で人名か，地名か，組織名か，あるいはそれ以外の品詞であるか，を認識することを目的とする．以下これらをあわせて固有名詞認識と呼ぶ．また，人名，地名，組織名の3つをあわせて以下単に細分類と呼ぶ．

固有名詞認識は次のように行なう．

1. まず形態素について認識を行なう
   - 各形態素にいろいろな手がかりを与える
   - 決定木を用いて形態素の固有名詞認識を行なう
   - 形態素の固有名詞認識の後処理を行なう
2. 次に複合語について固有名詞認識を行なう
   - ルールによる複合語の固有名詞認識を行なう
   - 複合語の固有名詞認識の後処理を行なう

3章で形態素，4章で複合語の固有名詞認識の処理について述べる．

3 形態素の固有名詞認識

3.1 決定木のための属性の獲得

ここで決定木の構築のため用いた属性について述べる．属性にはコーパス(京大コーパス[4]を用いた)から抽出した特徴量と辞書から得られる品詞，および動詞の格を用いた．

コーパスから抽出した特徴量は次の4種類である．

- 隣接語
- 名詞句「A の B」
- 人名複合語の末尾語
- 形態素の文字種

以下それぞれについて詳しく説明する．

隣接語　直前，直後に固有名詞が現れやすい語をコーパスから抽出し，手がかりとして利用する．このような手がかり語の振舞いは文字種によりいくぶん異なるので，漢字の固有名詞の直前直後とカタカナ語の固有名詞の直前直後では別の手がかりとして扱う．固有名詞直前の手がかり語を左隣接語，直後の手がかり語を右隣接語とする．左隣接語およびその値の抽出は以下の要領で行なう．

1. コーパス中で品詞が，人名，地名，組織名と割り当てられた語の直前にある語 a をすべて収集し，その集合を A とする．

2. コーパス中で $a_i \in A$ の直後にくる語 x をすべて収集し，その集合を x の文字種により，$X(\text{漢字})_i$, $X(\text{カタカナ})_i$ とする．

3. 集合 $X(\text{漢字})_i$, $X(\text{カタカナ})_i$ において，それぞれの集合内で品詞の頻度統計をとり，人名，地名，組織名についてその百分率を計算し，手がかり語 a_i の各細分類に対する値とする．

4. 集合 A に属するすべての単語について上記の値を求める．

右隣接語の場合は上記のアルゴリズムの前後逆転したものをそのまま用いる．

隣接語の属性としては左右で2種，文字種が漢字かカタカナの2種，細分類が3種，かけ算して計12種類となる．属性の値は，上記で計算した頻度統計の百分率をそのまま用いる．

名詞句「A の B」　名詞句「A の B」は前後にある名詞の間に何らかの関係をもつ．よって，固有名詞認識の手がかりとして用いることができると考えられる．間に「の」があることを除けば，隣接語と考えかたはまったく同じであるので，そのコーパスからの抽出法は「隣接語」のときと同様である．

名詞句「A の B」の属性としては，隣接語のときと同様に12種類となる．

人名複合語の末尾語　複合語で人名となる場合，多くは所属や役職などの肩書をともなう．「クリントン 米大統領」では「大統領」は「米」だけでなく「クリントン」とも結びついており，「クリントン」がどの程度人名になりやすいかという手がかりともなるはずである．つまり複合語の末尾の語が，その複合語の先頭の語が人名かどうかの手がかりとなると考えられる．

この特徴量のコーパスからの抽出法は「隣接語」と同様であるが，複合語の先頭の語の文字種は考えず，細分類も人名だけを考える．

そして，得られた値が40を超えたものを人名複合語の末尾語とし，属性とした．

形態素の文字種　固有名詞認識の対象である形態素の文字種を属性とし，その値はカタカナ，アルファベット，それ以外の3種とした．カタカナ語やアルファベット列は外来語を表現し，特に新聞記事では固有名詞が多いのではないかと考えられるので属性とした．

辞書　辞書から得られる品詞情報を属性とする．属性は辞書人名，辞書地名，辞書組織名の3つである．なお，JUMAN の固有名詞辞書の語数は約33,600語である．

動詞の格　動詞の格要素は次のように固有名詞の認識において利用できると考えられる．「ラベンナに住む」いとう文があった場合．「住む」のニ格には場所を意味する名詞のみが入り得る，ということがわかれば，「ラベンナ」が人名と組織名であることはなく，地名の可能性があるという手がかりとなる．

本来，格要素と動詞が関係するかどうかは文の構造がわかって初めて明らかになることであり，本稿で固有名詞認識を行なう段階ではそのような構造はまだ明らかでない．しかし日本語文においてはほとんどの場合，格要素はその後で最初に現れる動詞に関係する．そこで，文の構造がわかっていなくてもこの性質を利用することで用言意味辞書の手がかりを用いることができる．

用言意味辞書として NTT の日本語語彙大系 [5] を以下のように用いた．日本語語彙体系では格要素に対して約 2,700 の意味属性による記述が与えられているが，そのうち 7つの意味属性 (以下の 〈 〉 で囲んだもの) を次のような対応関係で用いた．

- 人名 — 〈人〉,〈動物〉
- 地名 — 〈社会〉,〈場所〉,〈建造物〉,〈場〉
- 組織名 — 〈組織〉

ここで処理を簡単にするため，各動詞についてそれぞれの表層格ごとに，対応する細分類が1種類のみで

あるようなもののみを用いた．よって，属性「動詞の格」の値は人名，地名，組織名のいずれかである．

3.2 決定木の構築と決定木による固有名詞認識

前節で述べた特徴を属性として京大コーパス中の各形態素に与え，決定すべきクラスは人名，地名，組織名，それ以外の4つとして，C4.5[6] のアルゴリズムを用いて決定木を構築した．

テストコーパスの各形態素にも同じ属性を与え，決定木でクラスを判定するが，その際にクラスが決定されるノードでの予想される誤り率を100から引いて確信度とした．

3.3 形態素の固有名詞認識の後処理

同一表記の参照 表記が同じ形態素のうちに，固有名詞認識の結果細分類が付与されたものがある場合，それらのうち確信度が最大の細分類を付与する．

ルールによる固有名詞認識の誤りの修正 京大コーパスを調査したところ，人名，地名については同じ細分類が連続する傾向がかなり強いことがわかった．この傾向を用いて，決定木による固有名詞認識の結果の誤りを修正することができる．ルールは以下の3つである．

- 認識された結果，複合語の先頭から地名，人名の順になっている場合，その確信度の大きい方に統一する．
- 認識された結果，複合語の先頭から人名，地名の順になっている場合，その複合語の末尾が人名複合語の末尾語でなければ，人名，人名に統一する．この人名複合語の末尾語は，3.1 節で述べた人名複合語の末尾語そのものである．
- 「カタカナ語・カタカナ語」で一方が人名，一方が未決定の場合，人名に統一する．

4 複合語の固有名詞認識

4.1 ルールによる固有名詞認識

複合語については，コーパス中から人名，地名，組織名の複合語を抽出し，これを一般化してパターンを作成した．特に企業名については別に大量に (4,926語) 収集して同じようにパターンを作成した．

パターンの作成は以下の要領で行った．

- 複合語を形態素解析する
- 形態素解析結果に対して次のように一般化する
  - 品詞が人名，地名，組織名であるものを「固有名」に
  - 品詞が数詞であれば「数詞」に
  - カタカナ語を「カタカナ」に
  - アルファベット列を「アルファベット」に
  - かっこや空白，「・」の記号を「特殊」に

  上記以外のものは，単語の字面 (表層表現) そのままとする

例えば，組織名の「京都 (地名): 銀行 (普通名詞)」からは「固有名: 銀行」というパターンが得られる．

このほか，人名が連続する複合語は人名，地名が連続する複合語は地名とするルールなども用いた．

そして，これらパターンとテストテキストとのパターンマッチにより複合語の固有名詞認識を行なった．

4.2 複合語の固有名詞認識の後処理

これには形態素の場合と同じく同一表記の参照のほか，以下のものがある．これらは処理が漸進的に進むので，新たに固有名詞の認識が行なわれなくなるまでこのステップは繰り返される．

並列名詞句 助詞「や」で結ばれる名詞句の並列があった場合，それらの固有名詞細分類は多くの場合同一である．そこである複合語の細分類が認識されておらず，直前または直後の並列関係にある文節内の複合語に対して細分類が認識されている場合，その文節内にある複合語の細分類を伝搬させる．

縮約形 複合語の固有名詞の場合には縮約形が用いられることがある．例えば，国際連合に対して国連という縮約形が用いられるような場合である．このような場合，一方の固有名詞認識が成功しても他方の認識は行なわれないことが多い．これらについても一文章全体を調べて，その情報を伝搬させる処理を行なう．

5 実験

実験に使用したのは1995年1月10日の1日分の新聞記事で，190記事，1519文，72586文字[1]であった．正解データとしてこれに対応する1995年1月10日の1日分の京大コーパスを用いた．決定木の学習には京大

[1] 2バイトで1文字と数えている

表 1: 形態素の固有名詞認識結果

	形態素解析のみ			固有名詞認識後		
	適合	再現	F	適合	再現	F
人名	87.9	58.9	70.5	87.2	78.6	82.7
地名	83.1	84.6	83.8	88.1	85.7	86.9
組織名	87.0	74.8	80.4	90.7	70.7	79.5
合計	93.5	71.4	78.2	88.1	79.6	83.6

表 2: 複合語の固有名詞認識結果

	適合率	再現率	F メジャー
人名	78.5	78.5	78.5
地名	54.0	67.0	59.8
組織名	56.9	47.6	51.8
合計	69.3	67.9	68.6

コーパスの 1995 年 1 月 1 日から 9 日までの 8 日分，9206 文を用いた．

評価は再現率と適合率，この両者から計算される F メジャーで行なった．F メジャーは適合率の逆数と再現率の逆数との算術平均の逆数である．

プログラムが出力した細分類と正解テキストの細分類とを比較し，両者が同じとき正解とした．ただし，形態素解析の段階での分割誤りにより，単語の区切りが異なるものについては評価の対象とはしていない．

形態素の固有名詞認識の実験結果を表 1 に示す．F の欄は F メジャーを示したものである．この表は，固有名詞認識を行う前と後でその精度を比較したもので，固有名詞認識を行なう前とは，形態素解析が終了しただけの状態である．全体的にみると精度が改善したことがわかる．なお，正解テキスト中にあった固有名詞の数は，人名 926，地名 1000，組織名 335 であった．

複合語の実験結果は表 2 である．正解テキスト中の複合語の数は，人名 442，地名 113，組織名 188 であった．

6 考察

形態素の固有名詞認識に用いた決定木では，辞書，文字種，隣接語，名詞句「A の B」の手がかりが複雑にからみあってあり，同じルールを人手で作ることがきわめて困難だと思われた．それに対し，動詞の格や人名複合語の末尾語の属性などほとんど出現しない属性も存在した．これは，学習事例にこれらの情報が少なかったためと考えられるが，人間が固有名詞の認識を行なうときにはこれらの情報は有用と考えられる．このように数は少なくても有用と思われる手がかりをもっと反映させれば精度はより改善するものと思われる．また，複合語に関しては組織名が最も精度が低かった．今回用いたパターンはコーパスから 173，企業名を収集したものから 393 個得られた．しかしコーパスから得られた組織名のパターンは 27 個と少なく，また，テスト記事中に企業名よりも区役所，警察署といった役所や政党名，業界団体が多かったため組織名の精度が悪かったと考えられる．今後は学習コーパスの量を増やすことが必要である．

7 おわりに

本稿では新聞記事について固有名詞認識を試み，それにより形態素解析の精度が改善することを示した．今後はより多様な文章について実験すること，また複合語での固有名詞認識の精度向上が課題である．

参考文献

[1] 山地治, 黒橋禎夫, 長尾真, 連語登録による形態素解析システム juman の精度向上, 言語処理学会第 2 回年次大会発表予稿集, (1996), pp. 73–76.

[2] 黒橋禎夫, 長尾真, 日本語形態素解析システム juman version 3.4, (京都大学大学院工学研究科, 1997).

[3] 黒橋禎夫, 日本語構文解析システム knp version 2.0 使用説明書, (京都大学大学院工学研究科, 1997).

[4] 黒橋禎夫, 長尾眞, 京都大学テキストコーパス・プロジェクト, 言語処理学会第 3 回年次大会発表予稿集, (1997), pp. 115–118.

[5] 池原悟ほか（編）, 日本語語彙体系 5, (岩波書店, 1997).

[6] J.Ross Quinlan, AI によるデータ解析, (トッパン, 1995).