4R-7

助詞を含む固有名詞の形態素の頻度情報による抽出

木村徹† 古宮嘉那子‡ 小谷善行‡

東京農工大学工学部情報工学科† 東京農工大学工学研究院‡

1. はじめに

インターネットの普及に伴い，クチコミサイトや Q&A コミュニティなどの CGM（Consumer Generate Media）の利用が多くなっている．CGM では人名，商品名などといった固有名詞を多く含む特徴がある．また，固有名詞は日々誕生している．そのため，新聞などの静的な辞書では対応できない場合があり，人手で常に整備するには時間や費用のコストが高い．そのため，CGM などのテキスト情報より自動的に固有名詞を抽出する研究が行われている（[1][2]）．しかし，従来の研究の多くは助詞を含む固有名詞を抽出することが困難であった．

そこで本論文では，固有名詞の中で「風と共に去りぬ」などの助詞を含むものを自動抽出する手法を提案する．固有名詞の抽出は，文章を形態素列に変換し，特定の箇所の先頭の形態素と末尾の形態素に注目して固有名詞であるかをスコアを付ける事により行う．スコア付けされた形態素列を形態素解析し，助詞を含む形態素列だけを選別する．最後に固有名詞であるかの確認を行い，固有名詞である妥当性が確認された形態素列が抽出された助詞を含む固有名詞である．

2. 助詞を含む固有名詞を抽出するシステム

この章では提案する，助詞を含む固有名詞を抽出する理論及びシステムについて説明する．システムは，入力としてタグなしコーパスを受け取り，出力としてそのタグなしコーパスに含まれる助詞を含む固有名詞を返す．システムは「形態素の頻度情報による固有名詞スコアの割り当て」「形態素列が助詞を含む判定」「助詞を含む形態素列が固有名詞であることの判定」の三つのモジュールで構成されている．

Extraction of Proper Nouns with Particles Based on Frequency of Morpheme

Toru Kimura†, Kanako Komiya‡, and Yoshiyuki Kotani‡

† Department of Computer and Information Sciences Faculty of Engineering Tokyo University of Agriculture and Technology

‡ Institute of Engineering Tokyo University of Agriculture and Technology

2.1 助詞を含む固有名詞を抽出する理論

固有名詞を形態素列中から抽出するために，固有名詞の前後に接続される形態素を発見できれば挟まれた部分は固有名詞になりうると仮定する．固有名詞の前後に接続される品詞は，形容詞や助詞など様々な品詞がある．各品詞に対応する語が固有名詞の前後に接続される形態素となる．そのため固有名詞の前後に繋がる各品詞の出現頻度は固有名詞に対して少なくなり，出現頻度の差を計算することで固有名詞の切れ目を発見する．

2.2 形態素の頻度情報による固有名詞スコアの割り当て

形態素列 $x_1 \ldots x_n$ に計算式(1) で付けられる固有名詞スコアを提案する．ただし，タイプミスを排除するため，コーパス中の出現頻度が閾値より小さい形態素列は固有名詞スコアを計算しないこととする．

$$Score(x_1 \cdots x_n, \delta) = \frac{f(x_1 \cdots x_n)}{f(x_1) f(x_2 \cdots x_n)} \cdot \frac{f(x_1 \cdots x_n)}{f(x_1 \cdots x_{n-1}) f(x_n)} \cdot \sum_{x_0 \cdots x_{n+1} \in \delta} \left(\left(1 - \frac{f(x_0 \cdots x_n)}{f(x_1 \cdots x_n)} \right) \left(1 - \frac{f(x_1 \cdots x_{n+1})}{f(x_1 \cdots x_n)} \right) \right) \cdots (1)$$

ただし f(x)は形態素 x の形態素頻度表での出現回数，δ はコーパス全体である．

計算式(1)について説明する．一つ目の分数の部分は，形態素 x_1 とその後に続く形態素列 $x_2 \ldots x_n$ の相互情報量をみることで，形態素 x_1 と形態素列 $x_2 \ldots x_n$ の共起しやすさを計算している．同様にして，二つ目の分数は，形態素 x_n が形態素列 $x_1 \ldots x_{n-1}$ の後ろに共起するか判定する．Σ の部分で固有名詞の切れ目を判別する．形態素列 $x_1 \ldots x_n$ がコーパスに出現するたびに一つ前の形態素 x_0 が形態素列 $x_1 \ldots x_n$ の前に出現する頻度と，一つ後ろの形態素 x_{n+1} が形態素列 $x_1 \ldots x_n$ の後に出現する頻度をそれぞれ形態素列 $x_1 \ldots x_n$ の出現頻度で割ることにより，前後の形態素 x_0，x_{n+1} が形態素列 $x_1 \ldots x_n$ に接続する確率を求める．求め

た確率の逆数が形態素列 $x_1 \ldots x_n$ に接続しない確率である．前後二つの接続しない確率の積をとることにより，形態素 x_1 と x_n が固有名詞の切れ目である確率を計算する．

一つ目，二つ目の分数の値が大きいほど形態素列 $x_1 \ldots x_n$ は一つの単語であり，Σの部分が大きいほど形態素 x_1 と x_n が単語の切れ目となる．したがって，固有名詞スコアの高いものほど固有名詞である可能性が高い．

2.3 形態素列が助詞を含む判定

次に，助詞を含む形態素列の抽出を行う.固有名詞スコアの付けられた形態素列に対して形態素解析を行い，形態素解析によって助詞を含むと判断されなかった形態素列を削除する．

2.4 助詞を含む形態素列が固有名詞であることの判定

最後に，抽出された形態素列が固有名詞である判定を以下の処理により行う．

(1) 固有名詞スコアが付けられた助詞を含む形態素列を固有名詞候補辞書として形態素解析器に登録する．
(2) 再度コーパスを形態素解析する．

形態素解析によって固有名詞と判定された形態素列を助詞を含む固有名詞として出力する．

3. 評価実験

システムの入力としてマイクロブログサービス Twitter[*1]に投稿された tweet を使用する．tweet の収集方法は，日本語タイムライン提供サービス me・you[*2]に表示される 20 件の最新 tweet を一分ごとに取得する．2 日間取得し続けた総計 50,322tweet を実験に使用する．

また，形態素解析器として茶筌[3]を利用し，品詞体系は IPA 品詞体系[4]を用いる．

固有名詞スコアの閾値を 2 とし，スコアの大きい上位 100 件を固有名詞候補辞書に登録する．形態素 3gram から形態素 7gram の助詞を含む固有名詞を抽出した結果を表 1 に示す．

表 1　助詞を含む固有名詞の抽出結果

F 値	0.057
適合率	0.160
再現率	0.035

「ゲゲゲの女房」（ドラマ作品名），「ハリー・ポッターと死の秘宝」（映画作品名），「のはなし」（書籍名）などの助詞を含む固有名詞を抽出することができた．一方，「2001 年宇宙の旅」（映画作品名）は抽出することができなかった．また，「ミツバチの羽音と地球の回転」という固有名詞の一部である「羽音と地球の回転」が誤って固有名詞として抽出されてしまった．

「のはなし」を抽出することができたが，これは『A の B』のように二つの名詞を接続したものではなく，先頭の形態素が助詞である．このような固有名詞は従来手法では抽出が困難であった．「のはなし」が抽出できた理由は，教師無し学習であるため出現する頻度が少ない品詞の組合せに対応できたためであると考えられる．

また，一回しか出てこなかった固有名詞を正解から削除した場合の再現率，F 値を表 2 に示す．

表 2　正解データを限定した場合の実験結果

F 値	0.199
適合率	0.160
再現率	0.262

重複して出現する助詞を含む固有名詞が少なかったため，F 値を大きくすることができた．しかし，抽出可能な固有名詞が少ないので，より多くの固有名詞を抽出できるように改善することが必要である．

4. おわりに

本論分では，助詞を含む固有名詞を自動抽出する手法を提案した．形態素列に、先頭の形態素と末尾の形態素に注目した固有名詞スコアを付け，形態素解析器による，助詞を含むフィルタリング，及び固有名詞判定によるフィルタリングを行った．その結果，F 値 0.199 で助詞を含む固有名詞を抽出することができた．

参考文献

[1] 中野桂吾, 平井有三: 日本語固有表現抽出における文節情報の利用, 情報処理学会論文誌, Vol.45, No.3, pp.934-941, 2004.
[2] 西川侑吾, 伊藤直之, 田村直之, 田中慶之, 中川修, 新堀英二: 形態素 n-gram を用いた助詞を含む固有名詞抽出, 言語処理学会第 16 回年次大会発表論文集, pp.12-14, 2010.
[3] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸:形態素解析システム『茶筌』version 2.3.3 使用説明書, 2003.
[4] 浅原正幸, 松本裕治: ipadic version 2.6.3 ユーザーズマニュアル, 2003.

*1 http://twitter.com/
*2 http://meyou.jp/