形態素 n-gram を用いた助詞を含む固有名詞抽出

西川 侑吾† 伊藤 直之† 田村 直之† 田中 慶之† 中川 修† 新堀 英二†
Yugo Nishikawa Naoyuki Ito Naoyuki Tamura Yoshiyuki Tanaka Osamu Nakagawa Eiji Shinbori
†大日本印刷株式会社 情報コミュニケーション研究開発センター 〒141-8001 東京都品川区西五反田 3-5-20
†E-mail:{Nishikawa-Y5, Itou-N12, Tamura-N5, Tanaka-Y51, Nakagawa-O, Shinbori-E}@mail.dnp.co.jp

1. はじめに

近年、Blog を用いたマーケティングやアンケートによる顧客満足度調査など、企業活動に CGM(Consumer Generated Media)を活用することが一般的となっており、精度の高い自然言語処理技術(テキストマイニング技術)が求められている。CGM は人名・地名・商品名などといった固有名詞を多く含むという特徴を持つが、それらの語は時と共に増え続けるため、新聞記事などから作成された静的な辞書データでは語数が足りず、またそれらの辞書を人手で整備するには高コストである。そこでテキスト中より自動で固有表現を抽出する技術に関し、様々な研究が行われている([1][2][3])。しかし、従来研究の多くが品詞情報を用いた形態素単位・文字単位での機械学習によるチャンキングであるため、名詞や未知語で構成される固有名詞は抽出できるが、書籍や映画などのタイトルなどに多い、助詞を含んだ固有名詞を抽出することは困難である。

そこで本研究では、大量の CGM データより n 個の形態素のつながり(形態素 n-gram)を作成し、ルールによるフィルタリング、頻度によるスコアリングを行うことにより固有名詞を抽出する手法を提案する。本手法により、形態素単位・文字単位のチャンキングによる固有名詞抽出手法では困難だった、助詞を含む固有名詞を抽出することができる。

2. 提案手法

本手法ではテキストに付随するカテゴリ情報をもとに、コーパスを"抽出対象コーパス"と"その他のコーパス"とに分け、"抽出対象コーパス"より、固有名詞の候補を抽出する。固有名詞の抽出は、下記 3 段階により行う。

1　形態素 n-gram 頻度集計

各コーパスを形態素解析し、人手で作成した抽出条件を満たす形態素 n-gram を作成、頻度集計を行う。

2　固有名詞候補のフィルタリング

作成された形態素 n-gram より 2 種類のフィルタリングルールにより固有名詞候補を抽出する。

3　固有名詞スコアの算出

頻度をベースに固有名詞スコアの算出を行い、固有名詞を獲得する。

本章では各処理の詳細について述べる。

2.1. 形態素 n-gram 頻度集計

"抽出対象コーパス"と"その他のコーパス"の両コーパスに対して形態素解析処理を行い、抽出条件(表 1) を満たす形態素 n-gram を作成し、それぞれのコーパスでの登場回数と、登場する文書数を集計する。

表 1 形態素 n-gram 抽出条件

条件	除外される n-gram の例
形態素 n-gram の最初の形態素が、以下の品詞でない。 (*助詞*、*助動詞*、*名詞-接尾*、*動詞-接尾*、*動詞-非自立*、*形容詞-接尾*、*形容詞-非自立*、*記号-句点*、*記号-読点*、*記号-一般*)	• にして下さい：助詞 • 的には、：名詞-接尾 • 、良いできだ：句点
最後の形態素が *接頭詞* でない。	• ご自分でお(調べ下さい)
同一形態素の 4 つ以上の連続を含まない。	• 無駄無駄無駄無駄 • ♪♪♪♪
n/2 個以上の記号を含まない。	• (・∀・)人
括弧の整合性が取れている。	• 亀」「キン肉マン
名詞-非自立、*名詞-数*以外の名詞もしくは未知語を最低 1 形態素含んでいる。	• 出ないのは • することが出来る
"(形態素+)*助詞*(形態素+)"の形を取っている。 ((形態素+):形態素の 1 回以上の繰り返し)	• 発売されました • ポートピア連続殺人事件

なお、形態素解析器は『茶筌』[4]、品詞体形は IPA 品詞体系[5]を用いる。

2.2. 固有名詞候補のフィルタリング

2.1 で作成した形態素 n-gram には、複数の形態

素からなる一語の固有名詞だけではなく、それぞれの形態素が意味を持つ、文の一部も大量に含まれている。そこで、抽出された形態素 n-gram に対し、"類語形態素 n-gram フィルタリング"と"形態素 n+1gram フィルタリング"の 2 種類のフィルタリングルールを適用することにより、形態素 n-gram より固有名詞候補のみを抽出する。

2.2.1. 類語形態素 n-gram を用いたフィルタリング

抽出した形態素 n-gram が、それぞれの形態素が意味を持つ、文の一部だった場合、形態素 n-gram を構成する形態素の一部を別の形態素に置き換えた別の形態素 n-gram も文書内に現れると考えられる。

そこで本研究では日本語語彙大系[6]を用い、形態素 n-gram 中に登場する固有名詞を除いた名詞と同一概念に属する名詞を取得、形態素 n-gram 内の形態素を置き換えることで、類語形態素 n-gram を作成し、その頻度が閾値を超えるものは文の一部と判断し、固有名詞候補から除外する。

例えば"逆襲のシャア"という形態素 3-gram があった場合、一般名詞である"逆襲"の類語を、日本語語彙大系より取得し、それらを用いた形態素 3-gram("反撃のシャア"、"復讐のシャア"など)を作成する。作成した類語形態素 n-gram の"抽出対象コーパス"内の登場回数が閾値以下だった場合、"逆襲のシャア"は固有名詞候補となる。

一方、"将棋の世界"という形態素 3-gram の場合、類語形態素 n-gram として"囲碁の世界"、"チェスの世界"、"将棋の世の中"などを作成、それらの"抽出対象コーパス"内登場回数が閾値($tf_{threshold}$)以上だった場合、"将棋の世界"は文の一部とみなし、固有名詞候補から除外する。(図 1)

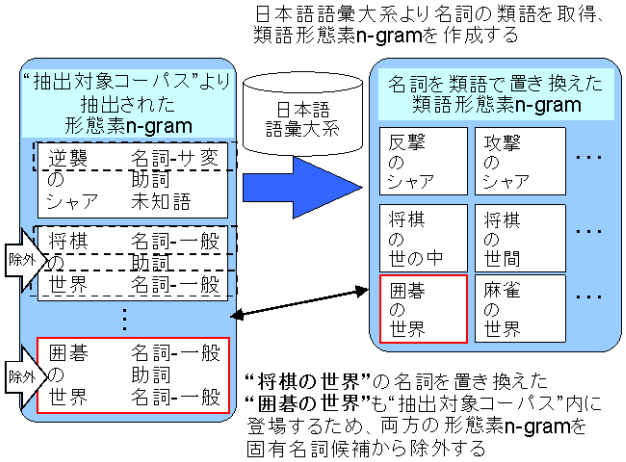

図 1 類語形態素 n-gram を用いたフィルタリング

2.2.2. 形態素 n+1-gram を用いたフィルタリング

コーパス中に登場する固有名詞が n 個の形態素からなる固有名詞だった場合、1-gram、2-gram、…、(n-1)-gram にもその固有名詞の一部が固有名詞候補として抽出されることになる。

そこで固有名詞候補として抽出された形態素 n-gram に対し、その形態素 n-gram を含む形態素 n+1-gram を照合する。形態素 n-gram から形態素 n+1-gram への生起確率が 100%である場合(形態素 n-gram と形態素 n+1-gram の登場回数が同一だった場合)、形態素 n-gram を固有名詞候補から除外する。

例えば"ジョジョの奇妙な"という形態素 4-gram があった場合、"ジョジョの奇妙な"を含む形態素 5-gram を検索し、"ジョジョの奇妙な冒険"を得る。それぞれ形態素列のコーパス内の登場回数を比較し、同一であった場合("ジョジョの奇妙な"から"ジョジョの奇妙な冒険"への生起確率が 100%だった場合)、"ジョジョの奇妙な"は"ジョジョの奇妙な冒険"の一部とみなし、固有名詞の候補から除外する。

2.3 固有名詞スコアの算出

2.2 節より得られた各形態素 n-gram に対し、式(1)により固有名詞スコアを算出する。

$$NE\_Score = \frac{(tf_{category} + 1)}{(tf_{others} + 1)} \cdot \mathrm{H}(df_{category}) \quad \cdots (1)$$

$$\mathrm{H}(df_{category}) = \begin{cases} 0 & (df_{category} \le df_{threshold}) \\ 1 & (df_{category} > df_{threshold}) \end{cases}$$

ただし、$tf_{category}$ は形態素 n-gram の"抽出対象コーパス"内での登場回数、tf_{others} は"その他のコーパス"内での登場回数、$df_{category}$ は"抽出対象コーパス"内での登場文書数、$df_{threshold}$ は登場文書数の閾値をそれぞれ表すとする。

n 毎に固有名詞候補をスコア順にソートし、各 n の上位 20 件の形態素列を固有名詞として抽出する。

3. 評価実験

提案手法の有効性を確認するために、Yahoo!知恵袋の回答データ(2004 年 4 月～2004 年 6 月までの全 603,845 件)を用いて固有名詞の抽出実験を行った。

“抽出対象カテゴリ”として、『携帯用ゲーム』、『ゲーム』、『テレビゲーム』、『オンラインゲーム』、『話題の本』、『本、雑誌、コミック』、『コミック』、『雑誌』、『アニメ』、『邦楽』、『テレビ、ラジオ』、『映画』、『バラエティ、お笑い』、『ラジオ』の 14 カテゴリを設定した。

また、各種パラメータは以下のように設定した。

$$\begin{cases} tf_{threshold} = 1 \\ df_{threshold} = 1 \end{cases} \quad \cdots (2)$$

上記条件のもと、n=3～6 の間で抽出した固有名詞の適合率と、抽出例を表 2 に示す。

表 2(a) 形態素 3-gram、4-gram の抽出結果

		3-gram	4-gram
適合率	上位 10 件	70%	50%
	上位 20 件	65%	45%
抽出例		・花とゆめ ・赤ちゃんと僕 ・リングにかけろ	・羊たちの沈黙 ・たけしの挑戦状 ・耳をすませば

表 2(b) 形態素 5-gram、6-gram の抽出結果

		5-gram	6-gram
適合率	上位 10 件	60%	60%
	上位 20 件	55%	35%
抽出例		・ジョジョの奇妙な冒険 ・二人のムラサキ東京	・風の谷のナウシカ ・ショーシャンクの空に

表 2 のとおり、『赤ちゃんと僕』(漫画作品名)、『耳をすませば』(映画作品名)、『二人のムラサキ東京』(曲名)などといった、従来手法では抽出が困難だった助詞を含む固有名詞を、上位 10 件に関し適合率 50%以上の精度で抽出できることが確認できた。

さらに、類語形態素 n-gram を用いたフィルタリングにより、"漫画が好き"(3-gram)、"原作を読んでから"(5-gram)といった"抽出対象カテゴリ"に高頻度で登場する、1 語の固有名詞ではない形態素 n-gram を、形態素 n+1-gram を用いたフィルタリングにより 3-gram、4-gram に登場する"ジョジョの奇妙"、"ジョジョの奇妙な"を、それぞれ効果的に除外できていることが確認できた。

また、n の値が小さい時には、上位 10 件の適合率と上位 20 件の適合率との差が小さいのに対し、n=6 の場合など、n の値が大きくなると、上位 20 件までの適合率が大きく下がる。これは形態素 6-gram の固有名詞が少なく、また抽出すべき固有名詞をスコアリングにより正しく上位に持ってきているためと言える。同時に n=3、4、5 の場合、上位 20 件以降にも抽出すべき固有名詞が登場しており、充分に抽出しきれていないと考えられる。よって、本手法で固有名詞候補をスコア順にソートした後、n の値や全体のスコアの分布により、固有名詞と判定する範囲を動的に定める手法を組み合わせる必要がある。

加えて、本手法では形態素 n-gram 内の品詞情報のみを用い、形態素 n-gram 抽出条件を定めたが、形態素 n-gram の前後に登場する形態素の情報も用いて固有名詞候補を除去することにより、抽出された形態素 n-gram の適合率を向上させることができると考えられる。

4. まとめ

コーパスより作成した形態素 n-gram を 2 種類のルールによりフィルタリングし、頻度をベースにスコアリングを行うことで、助詞を含んだ固有名詞を抽出する手法を提案し、実データを用いた実験を行い 3-gram において適合率 65%の精度で抽出できることを示した。

今後、固有名詞の抽出範囲を動的に設定する手法や、形態素 n-gram の前後の情報を用いたフィルタリングを行うことで、固有名詞抽出の再現率と適合率の両方を向上させる研究を進めていきたい。

5. 謝辞

本研究の実施にあたっては、ヤフー株式会社が国立情報学研究所に提供した Yahoo!知恵袋データを利用致しました。

参考文献

[1] 浅原正幸,松本裕治. 日本語固有表現抽出における冗長的な形態素解析の利用. 情報処理学会研究報告, 自然言語処理研究会, 2002-NL-153, pp.49-56, 2003.

[2] 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.

[3] 福島健一, 鍜治伸裕, 喜連川優. コーパスからの固有表現辞書の自動構築. 人工知能学会 第 79 回 知識ベースシステム研究会, 2007.

[4] 松本裕治, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』 version 2.4.0 使用説明書, 2007.

[5] 浅原正幸, 松本裕治. ipadic version 2.7.0 ユーザーズマニュアル, 2003

[6] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系— 全 5 巻—. 岩波書店, 1997.