E-026

品詞結合規則と外部辞書データを用いた複合名詞の生成

Generation of Complex Noun using Combination Rules and Dictionary Data

伊藤 直之† 西川 侑吾† 田村 直之† 中川 修† 新堀 英二†

Naoyuki Ito Yugo Nishikawa Naoyuki Tamura Osamu Nakagawa Eiji Shinbori

1. はじめに

テキストマイニングにおけるキーワード抽出の精度向上を目的とし，品詞結合規則と外部辞書データを用いて，形態素解析器の出力結果から複合名詞を抽出する手法を提案する．複合名詞は 2 つ以上の形態素から構成される名詞であり，全ての複合名詞を解析用の辞書データにあらかじめ記述しておくことはできない．複合語となる際のパターンを単語ごとに作成しておく手法も，パターン数の大きさから考えると現実的でない．また，品詞ごとの結合規則をシソーラスから生成して複合名詞を抽出する手法[1]があるが，形態素解析によって適切な品詞が付与されるとは限らず，抽出結果が形態素解析の精度に依存する．

提案手法では，経験則により作成した品詞結合規則によって特定の形態素どうしを結合させて形態素シーケンス(複合名詞候補)を作成した後，外部辞書データの見出し情報から学習した文字列ごとの複合名詞らしさのスコアを用いて，形態素シーケンスの中から複合名詞を判定する．提案手法により，事前に複合名詞リストや詳細なルールを記述することなく，複合名詞を抽出することができる．形態素解析ツールとして茶筌を，外部辞書データとして日本語語彙大系[2],Wikipedia 日本語版を用いる．

2. 提案手法

対象テキスト中に形態素解析器の辞書データに存在しない語が出現した場合，形態素解析によって正しい形態素区切りや品詞情報が付与されない．そのため，形態素解析結果を用いた複合名詞抽出を下記の２段階により行う．

① 品詞結合規則による形態素シーケンスの作成

② 文字列スコアによる複合名詞の判定

人手で作成した品詞結合規則によって，複合名詞の候補となる形態素シーケンスを作成し(①)，外部辞書データから学習した文字列ごとの複合名詞らしさのスコアにより，形態素シーケンスから複合名詞を判定する(②)．

2.1. 品詞結合規則による形態素シーケンスの作成

対象のテキストデータに対して形態素解析処理を行い，分割された各形態素に付与される品詞情報を利用して形態素シーケンスを作成し，複合名詞の候補とする．形態素解析結果に対し，名詞または未知語と判定された形態素が 2 つ以上連続で出現した箇所について，茶筌の出力する品詞下位分類(IPA 品詞体系[3])に基づき，経験則により設定した品詞結合規則(表 1)と照合して，形態素の連結ないし非連結を決定し，形態素シーケンスを生成する(図 1)．

表1 品詞結合規則

規則	品詞
○	名詞-一般
○	名詞-固有名詞-一般
○	名詞-固有名詞-人名-一般
○	名詞-固有名詞-人名-姓
○	名詞-固有名詞-人名-名
○	名詞-固有名詞-組織
○	名詞-固有名詞-地域-一般
○	名詞-固有名詞-地域-国
×	名詞-代名詞-一般
×	名詞-代名詞-縮約
×	名詞-副詞可能
○	名詞-サ変接続
○	名詞-形容動詞語幹
×	名詞-ナイ形容詞語幹
○	名詞-数
○	名詞-非自立-一般

規則	品詞
×	名詞-非自立-副詞可能
×	名詞-非自立-助動詞語幹
×	名詞-非自立-形容動詞語幹
×	名詞-特殊-助動詞語幹
○	名詞-接尾-一般
○	名詞-接尾-人名
○	名詞-接尾-地域
○	名詞-接尾-サ変接続
×	名詞-接尾-助動詞語幹
○	名詞-接尾-形容動詞語幹
×	名詞-接尾-副詞可能
○	名詞-接尾-助数詞
○	名詞-接尾-特殊
×	名詞-接続詞的
×	名詞-動詞非自立的
○	未知語

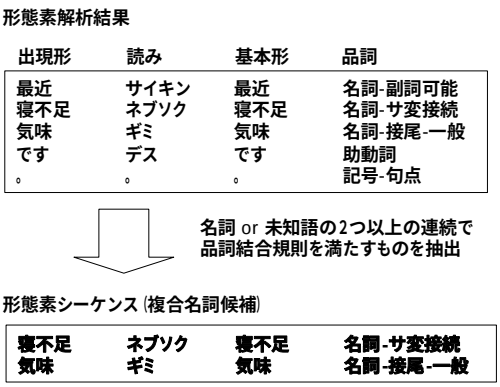

図1 形態素シーケンスの作成

2.2. 文字列スコアを用いた複合名詞判定

2.1 で作成した形態素シーケンスは，形態素解析器の品詞情報のみに依存するため，そのまま複合名詞と判定するには問題がある．例えば ”前回充電池を交換したときには” というテキストを形態素解析した結果に品詞結合規則を適用した場合，形態素シーケンスとして[前回 / 充電 / 池]が得られる．”前回充電池”を複合名詞と判定し，キーワードとして利用するのは適当でなく，”充電池”のみを複合名詞として用いるのが望ましいといえる．”前回”という形態素は複合名詞候補から取り除く、というように語ごとに詳細なルールを設定することも考えられるが，膨大にある単語に対して人手でルールを付与するのはコストが大きい．

†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 ユビキタスメディア研究所

2.2.1. 文字列スコアの算出

提案手法では，”前回”という文字列は複合名詞の先頭文字列として不適である確率が大きいということを示す文字列ごとのスコアを外部辞書データの見出しから統計的に算出する．見出し情報は，Wikipdeia の場合は記事の内容が一目で分かるように付けられた題名であり，日本語語彙大系の場合は一般名詞・固有名詞である（本手法では”用言データ”は用いない）．見出し文字列の先頭，末尾を起点として各 N-gram の出現頻度を求め，各 N-gram が見出し文字列の先頭に出現する頻度情報から前方スコアを，末尾に出現する頻度情報から後方スコアを算出する．頻度情報からスコアを算出する際には頻度の偏差値を用いる（図 2）．

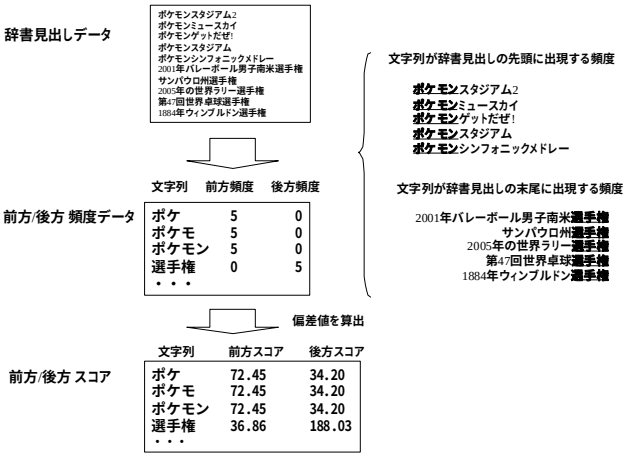

図 2 外部辞書データからのスコア生成

2. 2. 2. 複合名詞の判定

2.1.で作成した形態素シーケンスから，2.2.1.で算出した文字列スコアによって複合名詞を判定処理を行う．前方/後方スコアについて，それぞれ結合しきい値を設定し，形態素シーケンス中の先頭の形態素の前方スコア，末尾の形態素の後方スコアとしきい値との大小関係を調べる．先頭の形態素の前方スコアと末尾の形態素の後方スコアが結合しきい値より大きかった場合は，形態素シーケンス全体を複合名詞として判定し処理を終了する．また，先頭の形態素の前方スコアまたは末尾の形態素の後方スコアが，結合しきい値よりも小さい場合，当該の形態素を形態素シーケンスから削除する．複合名詞が判定されるか，形態素シーケンスの形態素数が 1 になるまで，上記の処理を再帰的に実行する（図 3）．

3. 評価実験

提案手法の有効性を確認するために Wikipedia の本文データを用いて複合名詞の抽出実験を行った．品詞結合規則のみによる抽出手法(従来手法)と，品詞結合規則と文字列スコアを用いた抽出手法（提案手法）での抽出結果の差の例を示す．とくに，形態素シーケンス末尾に”付近(名詞-一般)”や”主催(名詞-サ変接続)”といった，キーワードの一部として不要なものが出現する場合に，これらの語を正しく分割できていることがわかった．一方，”総理府(名詞-固有名詞-組織)”という，キーワードの先頭として適当なものであっても，外部辞書データの見出しでの出現頻度が少ない語については，誤った分割が発生することがわかった．

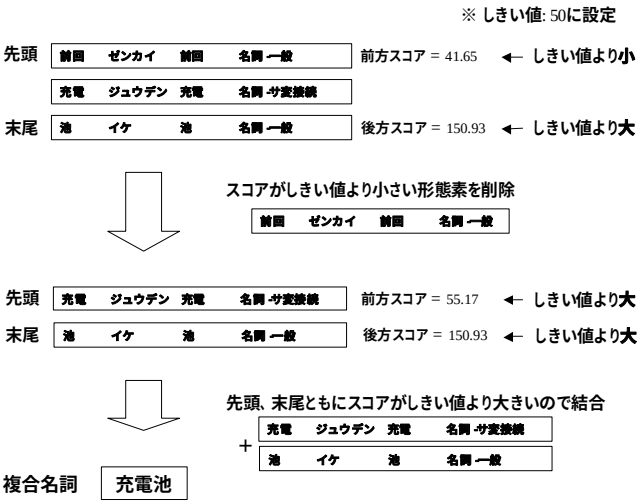

図 3 形態素シーケンスからの複合名詞判定

（従来手法）日本商工会議所主催
（提案手法）日本商工会議所 / 主催

（従来手法）大津インター付近
（提案手法）大津インター / 付近

（従来手法）総理府男女共同参画室長
（提案手法）総理府 / 男女共同参画室長

本手法では，文字列に対するスコア算出を，辞書見出しにおける出現頻度のみで行っているため，辞書見出しに登場しない固有名詞について，適切なスコア付与ができず，複合名詞判定の際に誤った分割が発生している．名詞の種類（一般，サ変，固有名詞など）によって，判定の際にスコアのしきい値を変更するなどの対策を行うことで，精度を向上していきたい．

4. まとめ

外部辞書データの見出し情報を用いて，形態素解析器の出力結果から複合名詞を抽出する手法を提案した．今後，詳細な性能評価実験を行うとともに,品詞ごとの文字スコアしきい値の設定，辞書見出しの内部構造を考慮したスコア算出について研究を進める．

参考文献

[1] 筏井勝，横山晶一，佐久間一弘 シソーラスを用いた複合名詞の生成・解析 情報処理学会第 52 回全国大会 pp.7-8

[2] 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.

[3] 5.品詞体系 pp.16-22 ipadic ユーザーズマニュアル 浅原正幸，松本裕治