

## 部分形態素解析を用いたコーパスの品詞体系変換

松田 寛 桐山 和久 山田 悟史 吉野 圭一 松本 裕治

奈良先端科学技術大学院大学 情報科学研究科

〒630-0101 生駒市高山町 8916-5

e-mail: horosi-m@is.aist-nara.ac.jp

<http://cl.aist-nara.ac.jp/>

### あらまし

品詞体系の相違は、品詞タグ付きコーパスの再利用を阻む最大の障壁である。日本語の場合、品詞体系の相違は単語認定基準に影響を及ぼすため、変換規則は形態素列の対となる。変換規則の抽出は、形態素連接での用例比較が必要となり、その実現は容易ではない。より単純な方法として、変換先の品詞体系を持つ形態素解析システムにより、コーパス原文を再解析する方法もあるが、タグ情報を完全に無視して通常の形態素解析を行ったのでは、品詞タグ付きコーパスを用いる意味はない。本稿では、一部の語に関する形態素変換規則のみを用いて、部分形態素解析による品詞体系変換を行う方法について論じる。実験では、助詞に関する変換規則を用いた部分形態素解析を行い、変換精度の向上を確認した。

## A Part of Speech Tag Conversion Method for Corpora Using Partial Morphological Analysis

Hiroshi Matsuda Kazuhisa Kiriyama Satoshi Yamada Keiichi Yoshino Yuji Matsumoto

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara 630-0101, Japan

e-mail: horosi-m@is.aist-nara.ac.jp

<http://cl.aist-nara.ac.jp/>

### Abstract

The differences in part of speech tagset have prevented us from reuse of a tagged corpus but with a different tagset. Mapping from one tag set to another is difficult, since the mapping rules are tied with a definition of morpheme that each tag set employs. We note that some mapping rules, e.g. for postpositional particles, can be decided with high certainty. In this paper, we propose a 'tagged' lexical entry suitable for partial morphological analysis. We experiment with Japanese postpositional particles and obtain improvement in results.

## 1. はじめに

近年、いくつかの大規模日本語品詞タグ付きコーパスが公開され広く利用されるようになった。しかし、大規模とはいえ、一つのコーパスに全ての分野の用例をまとめることには限界があり、特定分野に特化したコーパスを大規模コーパスと組み合わせて用いることが多い。

一般に、コーパスはそれぞれ独自の品詞体系の下に構築されており、それらを組み合わせて用いるには何らかの方法で品詞体系の変換を行う必要がある。品詞体系を自動変換するための最も単純な方法は、変換先の品詞体系を持つ形態素解析システムを用いてコーパスの原文を再解析する、という手法である。最近では、統計ベースの形態素解析システムでも高い変換精度が得られており、変換先の品詞体系を持つコーパスから統計学習を行うだけで、変換元コーパスをかなりの精度で変換することができるようになる。

しかしながら、この方法では変換元コーパス中の品詞情報は全て無視されることになり、品詞タグ付きコーパスを用いる意味が無くなる。変換元コーパス中の品詞情報を利用するためには、何らかの方法で形態素変換規則を抽出・適用したうえで解析を行う必要がある。

そこで本稿では、助詞のような一部の頻出語について、予め人手により形態素変換規則を抽出しておき、変換元コーパスにそのような変換規則を

適用した結果を、形態素解析システムの入力に用いる、という方法で、変換元コーパス中の品詞情報を利用したコーパスの品詞体系変換を行った。また、部分的にタグ付けされた文を精度良く形態素解析するために、形態素解析システムの辞書に品詞タグを追加する、という方法を用いた。

## 2. 品詞体系変換

コーパスの原文には新聞記事や音声の書きおこしなど様々な分野のものがあり、品詞体系は各分野に最適化したものが用いられる。

日本語の場合、品詞体系の相違は単語の認定基準に波及するため、変換規則は一般に形態素列間での対応となる[1]。さらに、図1の「という」の例のように変換先を一意に決定できない場合もある。一般に、変換規則を適用した結果は曖昧性を持ち、図2のような形態素ラティスの構造となる。

品詞体系変換は変換規則の抽出と曖昧性の解消という2つのステージに大別することができる。以下では、先行研究がそれぞれのステージで用いている方法について概観する。

### ● 変換規則の抽出

田代[1]では、変換元コーパスの原文の一部(1000文)を人手によりタグ付けして変換先の品

変換元 (京大コーパス)	→	変換先 (RWCコーパス)
助詞-格助詞【と】 + 動詞-子音動詞ワ行 -基本形【いう】	→	助詞-格助詞-連語【という】
助詞-格助詞【と】	→	助詞-格助詞-引用【と】 助詞-格助詞-一般【と】 助詞-並立助詞【と】 助詞-副詞化【と】

図1 「という」に関連する変換規則

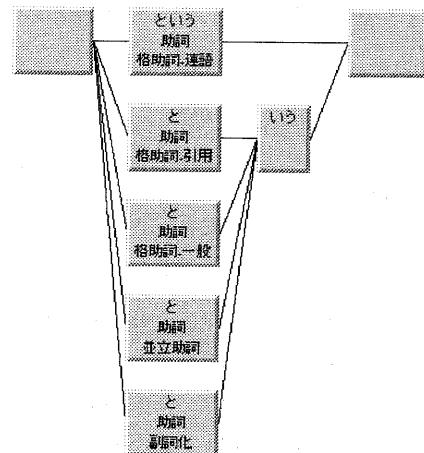


図2 形態素ラティス

詞体系で訓練コーパスを作成し、両方の体系で形態素境界が一致する部分を変換規則として抽出する。しかし、この方法では、訓練コーパスに出現しない語の規則は未定義となるため、そのような語が属するのと同じ品詞の語の規則で代用する、など規則の一般化を行う。

植木[2]では変換規則の抽出は行わず、文節内文法を人手により作成している。この際、曖昧性の解消に、変換元コーパスに付与されている係り受け情報を用いた構文解析を用いている。

乾[3]では変換元コーパスから各品詞を含む文節をいくつか切り出し、変換先品詞体系を持つシステムで形態素解析し、その結果を訓練コーパスと見なすことによって、品詞レベルの変換規則(約4000件)を抽出している。ただし、こうして自動抽出された変換規則群には大量の誤変換(約3300件)が含まれているので、人手による規則の洗練作業を別途行う。また、自動抽出できなかった助詞などの重要語についての規則(約60件)も人手により追加する必要がある。

### ● 曖昧性の解消

[1]では変換規則適用後の形態素ラティスに対して、訓練コーパスで学習した形態素辞書と品詞 bi-gram 連接表によるビーム探索を行っている。

[2]では構文解析の可否だけが述べられており、曖昧性の解消は議論されていない。

[3]では変換規則適用後、構文解析による曖昧性解消を文節内文法と文節間文法の2つのレベルで行っている。文節間文法を適用して一意に品詞列を決定できた文節数は85%となっている。

先行研究では、全ての形態素に対する網羅的な変換規則を作成して用いている。しかし、数千件にも及ぶ変換規則の抽出に、人手による何らかの作業を要していることは、品詞体系変換の効率を損ねる最大の要因となっている。そこで本稿では、変換規則を適用する対象を助詞に限定することで、変換規則数を減じている。従って、変換規則は文の一部にしか適用されないため、形態素ラティス

は途切れ途切れに構築される。次の節では、形態素ラティスの存在しない部分を、形態素解析システムを用いて高精度に部分解析する方法を述べる。

## 3. 部分形態素解析

茶筌[4]のような品詞 n-gram モデルを用いた形態素解析システムでは、文頭と文末に擬似的な形態素の存在を仮定して解析を行っている。これらの擬似的な形態素には『BOS/EOS』のような特別な品詞が振られており、文の両端から n-1 の範囲にある形態素は、図 3 のように、『BOS/EOS』との品詞連接コストを考慮して決定される。

これに対して、文の一部を切り出して解析する場合には、切り出した部分に隣接する品詞との連接コストを用いて、最適パスを求める必要がある。切り出した部分を単純に解析しただけでは、文頭と文末での品詞連接コストが正しく算定されないため、解析誤りを起こす要因となる。

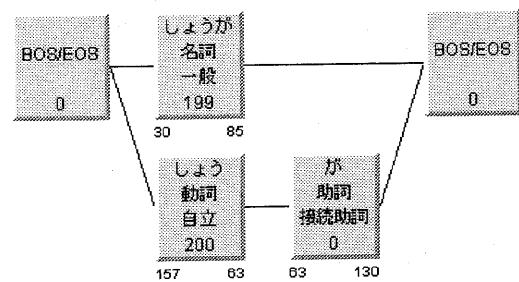


図 3 文頭と文末での品詞連接コスト

### 3.1. タグ化表現

一般に、形態素解析システムの辞書に、字面の長い語の生成コストを低くして登録すると、解析結果中のその語の再現率は高くなる。この性質を利用して、品詞 n-gram モデルによる解析時に、一部の語の品詞を指定する方法を本稿では提案する。

形態素辞書に、図 4 のような、他の形態素に比べて十分長く、ユニークな字面を持つ語(以下、タ

グ化表現と呼ぶ)を、指定したい品詞のエントリとして辞書に登録する。この辞書を用いて、タグ化表現を含む文を解析すると、タグ化表現を分割するパスのコストはそれを単一の形態素とするパスのコストに比べて大幅に高くなる(図5)。その結果、最適パス中のタグ化表現の部分は、形態素辞書でその語が登録されている品詞となり、期待された通りの結果が得られる。

予め、全ての品詞に対応するタグ化表現を辞書に登録しておけば、任意の語について品詞を指定することができるようになる。多くの場合、変換規則は語レベルの対応となるので、必要に応じて語レベルのタグ化表現も登録する。

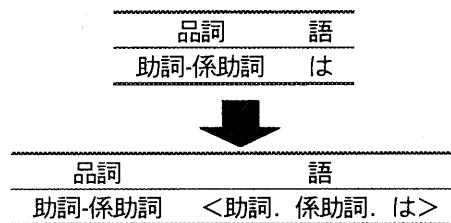


図4 タグ化表現への変換

### 3.2. 曖昧性の展開

部分形態素解析を用いて品詞体系変換を行うには、まず、変換規則適用後の形態素ラティスの曖昧性を全て展開する。次に、展開された形態素パスを、タグ化表現を含む文字列に変換して、部分形態素解析を行う。そして、全ての形態素パスを解析した結果を比較し、全文コストの最も低い形態素パスを最適解として採用する。

部分形態素解析の実行手順をまとめる。

#### 1. 形態素変換規則の抽出

変換規則は予め人手により抽出しておく。  
(詳細は次節を参照)

#### 2. タグ化表現の登録

変換規則の右辺(変換先)に出現する全ての品詞や語のタグ化表現を、形態素解析システムの辞書に登録する。

#### 3. 形態素変換規則の適用

変換元コーパスの各文に対し、全ての変換規則を適用して、形態素ラティスを取り出す。ラティスがない部分については、品詞情報を待たない单一の形態素から成るラティスに変換する。そして、これらを結合して文全体の形態素ラティスを作る。

#### 4. 曖昧性の展開

形態素ラティスから全ての曖昧性を展開して、パスの集合に変換する。

#### 5. 文抽出とエンコード

それぞれのパスから語の並びを取り出して、部分形態素解析の入力となる文を抽出する。この際、品詞情報を持つ形態素についてはタグ化表現に変換する。

#### 6. 部分形態素解析

2で作成した辞書を用いて、部分的にタグ化表現に変換された文を解析する。

#### 7. 最適解の選定

全てのパスの解析結果を比較し、全文コストが最小となるものを最適解として採用する。

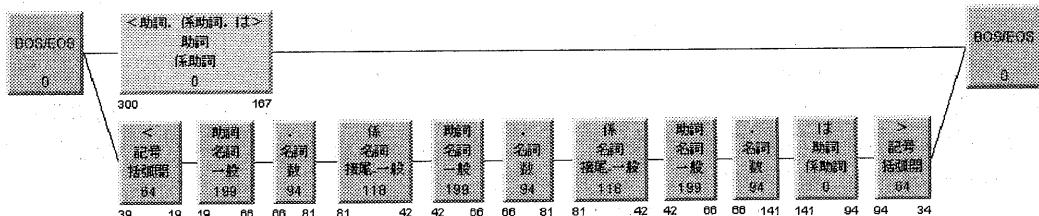


図5 パスの比較

## 8. デコード

最適解中のタグ化表現を、原文での表記に戻す。

タグ化表現を用いた部分形態素解析方法は、品詞 n-gram モデルの解析システム以外にも、コスト最小法を用いた解析システムであれば容易に実現できる。タグ化表現を実現するためのモジュールは、形態素解析ドライバモデル MACD<sup>1</sup>のモジュールとして実装されており、様々な形態素解析システムで利用可能となっている。

## 4. 実験

実験タスクは以下のように設定した。

- 変換元  
京大コーパス<sup>2</sup> (益岡・田窪文法) [5]
- 変換先  
RWC コーパス<sup>3</sup> (IPA 品詞体系) [6]
- 形態素解析システム  
茶筌 version 2.0b10 (IPA 品詞体系)

形態素変換規則の抽出は、IPA 品詞体系で助詞に分類されるものに限定した。助詞を用いた理由としては、

- 頻出語である
- 品詞認定の体系への依存度が低い
- 文節境界の特定に利用可能

が挙げられる。

### 4.1. 形態素変換規則の抽出

実験に先立ち、形態素変換規則を人手により抽出した。IPA 品詞体系で助詞に分類される全ての語について両コーパスでの用例を検索し、一意に対応可能と判定された形態素列の組を、変換規則と

<sup>1</sup> Morphological Analyzer Connectivity Driver Model.  
Java で作られた形態素解析ライブラリ。各種形態素解析システムの仕様を隠蔽し、統一されたインターフェイスを提供する。次の URL からソースコードを入手可能。  
<http://dahlia.aist-nara.ac.jp/macd/>

<sup>2</sup> version 2.0 を使用。毎日新聞 1995 年 1 月 1 日から 6 月 30 日までの記事。約 2 万文。

<sup>3</sup> 茶筌の学習用に最適化したものを使用。オリジナルの RWC コーパスとは品詞体系が一部異なる。

して採用した。

用例の検索は、変換元と変換先の両方のコーパスに対して、原文に対する部分マッチングを用いて行う必要がある。例えば、IPA 品詞体系の格助詞【に当たる】は、京大コーパスでは、格助詞【に】+動詞【当たる】の 2 語に分割されている。RWC コーパスでも、「宝くじに当たる」や「調整に当たる」などの用例では、格助詞【に】+動詞【当たる】の 2 語に分割する解釈が行われている。従って、「に当たる」の変換は一意には定まらない。今回の実験では、変換先が一意に定まらない語についての変換規則は不採用とした。

IPA 品詞体系では、連語を单一形態素として取り扱うことが多い。しかしながら、これには、助詞に関する変換規則が連語を誤まって分割する、という問題がある。このような問題を回避するには、

- 変換規則を削除する
  - 連語を生成する変換規則を追加する
  - 連語の変換を無効化する規則を追加する
- の 3 つの方法が考えられる。変換規則の減少を回避するには、連語を生成する規則を追加する必要があるが、連語の規則が誤ったまとめ上げを行う場合もあるので、単純にそのような規則を追加するという方法では解決できない。そのような場合には、変換の無効化で対処する。実験に使用した形態素変換規則を表 1 に示す。表の a の先頭に「まで」に関する規則がいくつかある。これらの規則は、助詞【まで】の変換規則による誤った分割を防ぐため、特定の語が連続した場合に変換を無効化するためのものである。

### 4.2. 品詞体系変換実験

京大コーパス全文(19956文)に対して、本稿の部分形態素解析による変換を行った結果と、通常の形態素解析による結果を比較した。異なる結果となった318文のうち、タグ付け誤りに起因するものと、他の品詞の連語に関する事例を取り除いて、残った290文について評価を行った。図6に結果を示す。当然ながら、これらの全事例において、部

種別	文数
品詞同定	183
形態素境界調整	107
合計	290

図6 実験結果

分形態素解析による結果が正解となっている。

## 5. おわりに

助詞に関する形態素変換規則を人手により抽出し、タグ化表現を用いた部分形態素解析を行うことで、コーパスの品詞体系変換を行うことができた。単純な原文の再解析と比べて、品詞同定と形態素境界調整の両方で変換精度の向上が見られた。

今後の課題としては、

- 正解コーパスとの比較による精度の測定
  - 助詞以外の品詞への適用
  - 曖昧性展開部の実装
  - 品詞の上位分類に関する変換ルールの適用
- などが挙げられる。

## 参考文献

- [1] 田代敏久, 森本逞.形態素情報付きコーパスの再構築手法. 情報処理学会論文誌, Vol.37, No.1, pp.13-22, 1 1996.
- [2] 植木正裕, 白井清昭, 徳永健伸, 田中穂積. 構造つきコーパスの共有化に関する一考察. 情報処理学会研究報告(98-NL-128)128-9, pp.61-66, 1998.
- [3] 乾健太郎, 脇川浩和. 品詞タグつきコーパスにおける品詞体系の変換. 情報処理学会研究報告(99-NL-132)132-12, pp.87-94, 1999.
- [4] 松本裕治, 北内啓, 山下達雄, 平野善隆. 日本語形態素解析システム『茶筌』version2.0 使用説明書. NAIST Technical Report, NAIST-IS-TR99008.  
<http://cl.aist-nara.ac.jp/lab/nlt/chasen.html>

[5] 黒橋 稔夫, 長尾 真. 京都大学テキストコーパス・プロジェクト. 人工知能学会全国大会予稿集, pp.58-61, 1997.

[6] テキストグループデータベースワークショッピ. RWC テキストデータベース報告書. Technical report, 技術研究組合 新情報処理開発機構, 3 1997.

表1 形態素変換規則

a) 変換ルールの適用を無効化するための規則

変換元	→	変換先
あく + まで	→	あくまで
飽く + まで	→	飽くまで
いつ + まで	→	いつまで
いま + まで	→	いままで
今 + まで	→	今まで
これ + まで	→	これまで
さ + まで	→	さまで
気 + が	→	気が
が + ない	→	がない
と + は + いえ	→	とはいえ
それ + に + は	→	それには
分 + の	→	分の
やむ + を	→	やむを

b) 変換先に曖昧性のない形態素変換規則

変換元	→	変換先
と 助詞-格助詞 + か 助詞-接続助詞	→	とか 助詞-並立助詞
どころ 助詞-副助詞 + か 助詞-接続助詞	→	どころか 助詞-接続助詞
か 助詞-終助詞	→	か 助詞-副助詞／並立助詞／終助詞
かい 助詞-終助詞	→	かい 助詞-終助詞
かしら 助詞-終助詞	→	かしら 助詞-終助詞
から 助詞-接続助詞	→	から 助詞-接続助詞
が 助詞-格助詞	→	が 助詞-格助詞-一般
が 助詞-接続助詞	→	が 助詞-接続助詞
くらい 助詞-副助詞	→	くらい 助詞-副助詞
ぐらい 助詞-副助詞	→	ぐらい 助詞-副助詞
けど 助詞-接続助詞	→	けど 助詞-接続助詞
けれど 助詞-接続助詞	→	けれど 助詞-接続助詞
けれども 助詞-接続助詞	→	けれども 助詞-接続助詞
こそ 助詞-副助詞	→	こそ 助詞-係助詞
さ 助詞-終助詞	→	さ 助詞-終助詞
さえ 助詞-副助詞	→	さえ 助詞-係助詞

し 助詞-接続助詞	→	し 助詞-接続助詞
しか 助詞-副助詞	→	しか 助詞-係助詞
すら 助詞-副助詞	→	すら 助詞-係助詞
ずっと 接尾辞-名詞性名詞接尾辞	→	ずっと 助詞-副助詞
ぞ 助詞-終助詞	→	ぞ 助詞-終助詞
だって 助詞-副助詞	→	だって 助詞-副助詞
だの 助詞-接続助詞	→	だの 助詞-並立助詞
って 助詞-副助詞	→	って 助詞-格助詞-連語
つつ 助詞-接続助詞	→	つつ 助詞-接続助詞
て 動詞 夕系連用テ形-連用形	→	て 助詞-接続助詞
と 助詞-接続助詞	→	と 助詞-接続助詞
ども 助詞-接続助詞	→	ども 助詞-接続助詞
な 助詞-終助詞	→	な 助詞-終助詞
なあ 助詞-終助詞	→	なあ 助詞-終助詞
なあ 助詞-終助詞	→	なあ 助詞-終助詞
ながら 助詞-接続助詞	→	ながら 助詞-接続助詞
など 助詞-副助詞	→	など 助詞-副助詞
なんか 助詞-副助詞	→	なんか 助詞-副助詞
なんて 助詞-副助詞	→	なんて 助詞-副助詞
ね 助詞-終助詞	→	ね 助詞-終助詞
ねえ 助詞-終助詞	→	ねえ 助詞-終助詞
ねん 未定義語-その他	→	ねん 助詞-終助詞
のに 助動詞 ナ形容詞-タ列基本連用形	→	のに 助詞-接続助詞
のみ 助詞-副助詞	→	のみ 助詞-副助詞
は 助詞-副助詞	→	は 助詞-係助詞
ばかり 助詞-副助詞	→	ばかり 助詞-副助詞
ばかり 助動詞 ナノ形容詞-語幹	→	ばかり 助詞-副助詞
ばっかり 助詞-副助詞	→	ばっかり 助詞-副助詞
へ 助詞-格助詞	→	へ 助詞-格助詞-一般
まで 助詞-格助詞	→	まで 助詞-副助詞
まで 助詞-接続助詞	→	まで 助詞-副助詞
まで 助詞-副助詞	→	まで 助詞-副助詞
や 助詞-終助詞	→	や 助詞-終助詞
やいなや 助詞-接続助詞	→	やいなや 助詞-接続助詞
やら 助詞-接続助詞	→	やら 助詞-並立助詞
よ 助詞-終助詞	→	よ 助詞-終助詞
より 助詞-格助詞	→	より 助詞-格助詞-一般
わ 助詞-終助詞	→	わ 助詞-終助詞