E-007

標準語用のパラメータを流用した岡山弁向け統計的形態素解析
Statistical Morphological Analyzer for Okayama Dialect using Parameters for Standard Japanese

福圓 琢真† 但馬 康宏† 菊井 玄一郎†
Takuma Fukuen Yasuhiro Tajima Genichiro Kikui

1. はじめに

音声対話システムの商用化や会議録に対する情報抽出[1]など，話し言葉に対する言語処理のニーズが高まっている．また音声由来ではないが，マイクロブログや SNS などの口語的なテキストに対する言語処理も活発化している．これらを実現する上で話し言葉に対する形態素解析処理は必須である．しかしながら地域特有の方言を含む文章については正しく処理できないことが多い．

近年主流となっている統計的形態素解析手法では処理対象の方言のタグ付きコーパスが大量にあれば原理的には実現可能である．しかしながら，方言のコーパスは限られており，タグ付きのものは更に少ない．一方で，文法および語彙において標準語と似通っている部分が多い方言については標準語のモデルパラメータを流用することにより，目的を達成できる可能性がある．このような考え方に基づき，本研究では HMM（Hidden Markov Model）を用いた形態素解析において標準語コーパスから学習したパラメータを辞書・規則を用いて流用することにより岡山弁の形態素解析を試み，解析精度を評価する．

2. 基本となる形態素解析

本研究では形態素解析手法として，クラス N グラムによる隠れマルコフモデル（HMM）を用いた手法[1]を利用する．具体的には下記で与えられる $\mathrm{p(W,C)}$ を最大化する $\mathrm{W}=w_1,w_2,\cdots,w_n$，$\mathrm{C}=c_1,c_2,\cdots,c_n$ の組みを求める．但し，w_i，c_i はそれぞれi番目の形態素表層形と形態素クラスを表し，$\mathrm{p}(w_i|c_i)$，$\mathrm{p}(c_i|c_{i-1})$ はそれぞれ c_i が与えられた時の w_i の事後確率（単語出現確率），c_{i-1} の後に c_i が出現する確率（品詞連接確率）として表す．なお当該言語における形態素表層形と形態素クラスの組みの集合は形態素解析辞書に定義されている。形態素クラスとしては通常、品詞や活用形を用いる．

$$\mathrm{p(W,C)}=\mathrm{P(W|C)P(C)}\cong\prod_i \mathrm{p}(w_i|c_i)\mathrm{p}(c_i|c_{i-1})$$

実際にはアンダーフローを抑止するために両辺の対数をとって利用する．

$$\log \mathrm{p(W,C)}\cong\sum_i(\log \mathrm{p}(w_i|c_i)+\log \mathrm{p}(c_i|c_{i-1}))$$

標準語用の形態素解析はこの式における $\mathrm{p}(w_i|c_i)$，$\mathrm{p}(c_i|c_{i-1})$ を標準語の形態素タグ付きコーパスから推定する．

3. 形態素解析から見た岡山弁の特徴

3.1 岡山弁特有の言語表現・語彙

岡山弁は全国の方言の中では比較的標準語に近い方であるが，異なっている部分も多い [2]．本研究ではこれらを大きく次の 4 つに分類する．

① 標準語の単語の表記が発音の「なまり」によってが変化したもの
例）「じゃ」←「だ」（断定の助動詞）
「ねー」←「ない」（打消しの助動詞）
「たけー」←「高い」（形容詞）
2 番目と 3 番目の例はいずれも ai という音が e-（長音）に変化している．このような，規則は文献[2]に整理されている．

② ①以外かつ標準語形態素 1 つと置換可能なもの
例）「ぼっけー」←「すごい」（形容詞）
当該岡山弁表現と意味的かつ統語的（品詞的）に置換可能なものである．単に品詞が同じではなくても副詞と動詞連用形のように文法機能が同じものと考える．なお，岡山弁側が厳密には複数形態素であっても固定的な場合は一語と考える．

③ ①と同様であるが標準語表現が連語になるもの
例）「ちーときま」←「少し（名詞）+の（名詞）+間（名詞）」
②の例と同様，岡山弁側が固定的な連語の場合は 1 語と考える．

④ 名詞と助詞の組み合わせによる縮訳
例）「ありゃー」←「あれ（名詞）+は（助詞）」
「あみゃー」←「飴（名詞：アメ）は（助詞）」
格助詞または副助詞が直前の名詞の最後の音と縮約するもので「曲用」とも呼ばれる[2]．本来，条件を満たせば任意の名詞と助詞の組みについて起こりうるが，最近は「あれ」，「これ」，「それ」，「どれ」，「わし」などの代名詞と副助詞「は」の組み合わせのみ縮訳する話者も多いため，本研究ではこれらに限定する．

3.2 岡山弁の表記法

話し言葉の形態素解析においては，入力をどのように表記するかが問題となる．日本語の通常の表記法は「漢字かな交じり」である．漢字で表記される単語については（なまりのため）読みは異なるかもしれないが標準語と同じ表記にするのが自然である．しかし，かなを含む単語は方言の発音に忠実であることが望ましい．さらに，送り仮名の場合については，たとえば「あけえ（赤い）」は「赤え」と記述し，漢字部分の発音（読み）を「アケ」とみなすような扱いが必要となる．また，長音表記をどこまで取り入れるかも検討の余地がある．現代仮名遣いでは長音として発音される場合でも二重母音で表記することがあること

† Graduate School, Okayama Prefectural University

（例：きょーと→きょうと）と整合的にするのが適切である．本研究においては漢字かな交じり表記を採用し，付属語や和語系の一般にかな書きされる単語は方言の発音に即したかな表記を採用した．また，送り仮名は先に述べた「赤え」といった表記を用いることとした．

4. 提案手法

2 章で述べた標準語形態素処理用いて，3 章で述べた岡山弁表現が扱えるようにするために，3 章で述べた①~④の各表現を辞書に追加するとともに，それぞれの単語出現確率$\mathrm{p}(w_i|c_i)$を付与する．

4.1 岡山弁表現の辞書への追加

岡山表現のうち①については規則性があるため，参考文献[2]，[3]の変換規則に基づいて MeCab 付属の辞書（ IPADIC）の項目から岡山弁の辞書項目を自動生成し登録する．但し，助動詞，カ変動詞，サ変動詞は不規則なものが多いため手作業で登録する．

②および③の表現については個別的かつ数が少ないので参考文献[2]，[3]をもと，手作業で辞書への登録を行う．

縮約の④については名詞全てを対象にする場合，単純な辞書登録では扱うことが難しいが，今回は 3 章の表現に限定しているので②，③と同様に手作業で辞書登録する．

なお，③および④については当該形態素の品詞を連語として辞書登録し，対応する全ての標準語形態素の出現形，品詞などの情報も一緒に登録する．

4.2 単語出現確率の付与

①と②については対応する標準語単語の単語出現確率をそのまま付与する．例えば，岡山弁形態素「あちー」の場合，この形態素に対応する標準語形態素「あつい」の単語出現確率をそのまま使用するものとする．

③および④の場合対応する標準語列を構成する各形態素の単語出現確率と，形態素間の品詞連接確率の和を付与する．例えば，岡山弁形態素「ゆーかす」の場合，この形態素に対応する標準語形態素は「言っ（動詞）て（助詞）聞か（動詞未）せる（動詞）」である．そのため岡山弁形態素「ゆーかす」の単語のコストは，「言っ」，「て」，「聞か」，「せる」，のそれぞれの単語出現確率と，それぞれの品詞連接確率の和となる．具体的な式を以下に示す．

$$
\begin{aligned}
&\log \mathrm{P}(\text{ゆーかす}|\text{連語}) \\
&\quad = \log \mathrm{P}(\text{言っ}|\text{動詞}) + \log \mathrm{P}(\text{助詞}|\text{動詞}) + \log \mathrm{P}(\text{て}|\text{助詞}) \\
&\quad + \log \mathrm{P}(\text{助詞}|\text{動詞}) + \log \mathrm{P}(\text{聞か}|\text{動詞}) + \log \mathrm{P}(\text{動詞}|\text{動詞}) \\
&\quad + \log \mathrm{P}(\text{せる}|\text{動詞})
\end{aligned}
$$

5. 評価実験

提案手法の有効性を検証するため，岡山弁テキストを形態素解析し人手で作成した正解データと比較した．解析結果と正解で，形態素の範囲および品詞の両方が一致したものを正解とし，適合率，再現率，F 値を求めた．また，MeCab（IPADIC）の解析，および本手法において形態素クラスに品詞だけを用いた手法の結果との比較を行った．

評価データとして総社市議会の議会録より岡山弁を含む文章 131 文，（6118 形態素）を抽出し利用した．

標準語用の形態素パラメータは 2007 年～2012 年の毎日新聞を MeCab（IPADIC）で形態素解析したものを学習データとして推定した．なお，ゼロ頻度問題に対応するために加算スムージング[1]を適用した．また，クラスとして品詞と活用系の組みを用いた．

表 1：岡山弁に対する解析結果

	適合率	再現率	F 値
MeCab	0.838	0.867	0.852
提案手法 1 (クラス:品詞のみ)	0.843	0.8525	0.847
提案手法 2 (クラス:品詞と活用形の組)	0.885	0.906	0.895

表 1 に各手法の適合率，再現率，および，F 値を示す．提案手法 2（クラス:品詞と活用の組み）について F 値が 0.895 と高い値を達成することができた．

MeCab（IPADIC）の解析精度が低いのは，岡山弁の表現が辞書にないためである．

提案手法 1（クラス:品詞のみ）は，岡山弁の表現には対応できているものの，活用形を考慮していないため，動詞未然形の後に終助詞が来るなどの解析ミスが見られた．

提案手法 2（クラス:品詞と活用形の組み）において多く見られた誤りは，「か（助詞）な（助詞）」が「かな（名詞）」になるなど，話し言葉でよくみられる表現が書き言葉の表現に解釈されるものである．これは，ベースとなる標準語用のパラメータを新聞コーパスから推定したため，話し言葉でよく見られる単語や品詞連鎖の確率が小さくなったためであると考えられる．対処法としては，学習コーパスに標準語圏内の議会の会議録やブログを，辞書として unidic を用いることなどにより口語のモデルを強化することが挙げられる．

なお，本論文で対象とした 2 章①～④の表現については概ね正しく解析できたが，③についていくつか誤りが見られた．例えば「～て（助詞）+いる（動詞）」に対応する「~とる」やその過去形の「~とった」が「とる（動詞）」や「とっ（動詞）+た（助動詞）」と解析誤りした．これは「とる（動詞）」の単語出現確率に対して「ている」の確率，特に，クラス連接確率が小さかったためと考えられる．この問題は単語連接確率の導入などで改善できる可能性がある．

6. まとめ

本研究では岡山弁を含む文字列をクラス N グラムによる隠れマルコフモデルを用いた形態素解析の手法について提案した．岡山弁表現を 4 つに分類し，それぞれに単語出現確率の付与を行い辞書登録することで標準語形態素処理において岡山弁表現を扱えるようにした．実験の結果，提案手法 2（クラス:品詞と活用の組）では F 値 0.895 と高い精度を達成した．

今後の課題として，ベースモデルのパラメータを口語に対応させること，単語Ｎグラムの利用などモデル自体の強化があげられる．

参考文献

[1] 高村 大也， “言語処理のための機械学習入門”， コロナ社，(2010). pp. 148-151. pp. 115-117
[2] 青山 融， “岡山弁会話入門講座”， 岡山若者新書 1， (1986).
[3] 虫明 吉次郎， “岡山弁あれこれ 1”，研文館 吉田書店，(1978)