6P-6

方言コーパスに基づく文章の地域性の推定

瀧本　恵理†　　奥村　紀之‡
香川高等専門学校　情報工学科†, ‡

1　はじめに

マイクロブログや SNS の流行により，他地域の人と交流が増加している．これに伴い，馴染みのない方言を目にすることが増えている．しかし，方言話者の多くは自身の使用する方言を完全には自覚していないため，意思疎通が困難になる場合がある．

そこで，馴染みのない方言が含まれる文を，自身の馴染みのある表現が使用されている文へ変換可能にすることで問題の解消を目指す．本研究では，多地域にわたる方言解釈システムのの基盤となる地域性推定に関する検証を行うとする．

2　関連研究

多数の地域を対象とした方言解釈システムの開発をするにあたり，方言の収集方法や方言同士の判別方法が問題となる．

小林らの研究では，長野県の方言に着目し，方言を含む文章を対象とした言語認識システムの開発を行っている[1]．長野県の方言に関するアンケートを実施し，長野県の方言についての特徴を得ている．独特な言い回しの方言については，方言を辞書に登録することで形態素解析を行うことができる．その結果，方言から標準語への変換が可能になる．標準語と同音の方言については，感情判断システムによって変換を行う必要の有無を判断している．

廣田らは検索エンジンを使用した方言コーパス収集システムを構築している[2]．ユーザに収集する地域の方言に特徴的な表現を複数入力させ，それらを検索クエリにすることで，Web からテキストデータを抽出している．

3　香川県の方言についての調査

本研究では，複数の地域の方言コーパスの収集を行う．香川県の方言コーパスを収集する際にはアンケート調査による方言の収集法を用いた．

アンケートは次の内容で実施した．

- 出身地域の方言を列挙させる
- 標準語文を方言文に変換させる(100 問)

それぞれのアンケートの回答例を表 1，表 2 に示す．

表 1:アンケート 1 の回答例

方言	標準語
てがう	遊んでやる
むつごい	油っこい

表 2: アンケート 2 の回答例

標準語	方言
これ片付けておいて	これ片付けとって
鳥が飛んでるね	鳥がとんびょる

男女 40 名の学生に対しアンケート調査を実施したところ，有効な回答を 29 部取得でき，以下のような特徴がみられた．

- 独特の言い回しが存在する
  - 「むつごい」，「こんこ」など
- 標準語と同音の方言が存在する
  - 「まける」，「こえる」など

これらの特徴は長野県の方言で見られた特徴と同様である．また，取得した香川県の方言と長野県の方言を比較したところ，上記の特徴に加え，以下のような特徴もみられた．

- 他地域の方言と同音の方言が存在する．
  - 「きんな」など
  -

「きんな」は長野県では「昨日」，香川県では「だから」という意味で使用されている．

4　香川県の方言テキストの収集

方言収集手法は廣田らの手法を用いている[1]．システムに検索クエリとして，アンケートから得た香川県の方言を与える．取得できた香川県の方言テキストの数からその方言の独自性を調査する．香川県の方言テキストの取得率が高い

「An Extimation for Sentences Including Areas Information Based of Dialect Corpus」
†「Eri TAKIMOTO」
‡「Noriyuki OKUMURA」
Kagawa National College of Technology, Department of Information Engineering

方言ほど，他地域の方言との重複が少ないため，独自性も高いと言える．香川県の方言テキストの取得率は次の計算式で行う．

$$\text{取得率} = \frac{\text{取得した香川県の方言テキストのセット数}}{\text{取得した文のセットの総数}} \quad (1)$$

1 セットは 5 文で構成されている．収集した方言テキストを香川県の方言と他地域の方言に分類を行った．その結果から得た香川県の方言テキストの，取得率が上位 5 語と下位 5 語の検索クエリを表 3 に示す．表 4 は香川県に関する記事を多く取得する検索クエリを示す．ここでは，標準語を他地域の方言として扱う．複数の地方の方言が混在するテキストは除外して集計を行った．

表 3:検索クエリごとの方言テキストの取得数

検索クエリ	香川県	他地域	総数	取得率
ちみきる	10	29	44	0.256
いっきょる	18	55	91	0.247
おもっしょい	17	52	91	0.246
やで	16	60	90	0.211
むつごい	15	57	87	0.208
…	…	…	…	…
いぬ	0	74	90	0
かまん	0	79	91	0
ほうる	0	74	92	0
たく	0	73	94	0
戻りし	0	75	94	0

表 4:香川県の記事を多く取得する検索クエリ

検索クエリ	香川県	他地域	総数	取得率
何がでっきょんな	3	22	28	0.120
じょんならん	9	70	92	0.113
ぴっぴ	3	72	93	0.040
がいな	3	76	89	0.038
こんぴらさん	1	76	90	0.013

表 3，表 4 ともに他地域の方言テキストの取得数の方が多い．これは博多や大阪の方言でも同様の傾向を得られた．

表 3 より「ちみきる」「いっきょる」「おもっしょい」「やで」「むつごい」は他の香川県の方言を検索クエリにした場合と比べ，香川県の方言テキストの取得数が多い．この結果から，これらの方言は独自性が高いことが分かる．したがって，他地域の方言と香川県の方言を判別する際に指標にすることが可能である．

また，「いぬ」「かまん」「ほうる」「たく」「戻りし」は独自性は高いものの解析のミスにより，各々の方言だけでは文章を香川県の方言であると断定することはできない．

表 3 より，「ぴっぴ」「がいな」「じょんならん」「こんぴらさん」などの方言では香川県の方言テキストを多くは取得できなかったが，香川県に関する記事を多く取得している．

そこで，独自性の低い方言と香川県の記事を多く取得した方言をペアで検索クエリにしテキストを取得した．「かまん」と「ぴっぴ」をペアにした場合，取得率は 0.48 と高く，ペアにすることで文章の地域を判別する際に指標として扱える．「かまん」「ぴっぴ」「じょんならん」と 3 つセットにした場合，テキストの取得はできなかった．独特の表現の方言を 2 つ含む場合が最も文章の地域の推定が容易である．

5 他地域の方言との関係

博多，大阪の方言テキストと香川の方言テキストの比較を行うと，「むつごい」「きょる」といった表現は香川の方言テキストのみで検出された．これに加え，「やきん」などの一部の独特な語尾表現でも同様な結果を得られた．語尾表現は他の方言より使用する回数が多いため，より少ない方言テキストでの地域の推定が可能である．また，方言ごとに同一のテキスト内で使用される頻度が異なるため，使用される頻度を調査する必要がある．

6 おわりに

本稿では，多地域にわたる方言解釈システムの地域性推定に関する検証について述べた．独特な表現の方言については，文章中からそれらの方言を見つけることで地域の推定が可能であることが分かった．それぞれの地域について独特の表現の方言を取得することで，独特の表現を含む文章の地域の推定を期待できる．

参考文献

[1] 小林聖也，奥村紀之:方言と標準語の違いを考慮した言語認識システムの開発，2009.
[2] 廣田壮一郎，笹野遼平，高村大地，奥村学:方言コーパス収集システムの構築，2013.