# 方言コーパスに基づく発話者の地域推定

瀧本 恵理　奥村 紀之
香川高等専門学校 情報工学科

i10295@sr.kagawa-nct.ac.jp,okumura@di.kagawa-nct.ac.jp

## 1　はじめに

マイクロブログや SNS の流行により，異なる地域の人との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機会が増えている．しかし，方言話者の多くは自身の使用する方言の自覚が薄く，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基盤となる地域性推定に関する検証を行う．

## 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基盤となる地域性推定に関する検証を行うにあたり，方言の収集方法や方言同士の判別方法が問題となる．

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 [1]．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，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．

平らは Support Vector Machine を用いたテキスト分類における属性選択手法について述べている [2]．最適な属性選択を，相互情報量を基準とした属性選択と品詞を基準とした属性選択の比較で調査した．調査では品詞によるフィルタリングのみを行い，全単語を入力として用いることで高い分類精度を得られている．

本研究では，形態素情報と文字 N-gram による方言分類器を構築する．

## 3　方言コーパスの作成

多地域に渡った方言の地域性の推定を行う上で，方言の表現が似通った地域同士での方言の判別は，全く異なる地域の判別に比べ困難であると考えられる．そこで，本研究では方言の性質が似ていると思われる地域間での方言の分類を行う．方言の性質が似ている地域として，香川・大阪・博多に着目して方言テキストの収集を行った．方言収集手法には廣田らの手法を用いた [1]．収集した方言テキストの中から各々の地域の方言であると断定できるものを，各々の地域出身者やその周辺地域の出身者の協力を用いて，それぞれ 100 セットずつ抽出した．

## 4　SVM による分類

各々の地域の方言と断定できる純粋な方言テキストを正しい地域の方言として分類できるか機械学習を用いて検証を行う．純粋な方言テキストを SVM により学習する．SVM は「LIBSVM-3.19」を使用する．

方言テキストを学習するにあたり素性が必要となる．素性には，廣田らの研究で使用されている素性を利用している [1]．形態素 1-gram，形態素 2-gram，文字 2-gram，文字 3-gram を用いる．これらの素性で 3 地域の素性ベクトルを作成する．素性ベクトルを作成した後，各セットの素性値を求める．素性値は，正しい地域の方言である場合を 1，その他の 2 つ地域の方言である場合を-1 とする．分類器は香川・大阪・博多の各々が正例となるものをそれぞれ 1 つずつ作成する．分類器を 3 つ用いることで各々の地域の方言テキストがどのように分類されるかを調査する．

## 5　分類器の評価

分類器の素性として形態素 N-gram と文字 N-gram を採用している．そのため，方言の特徴的な表現が学習された場合，分類が容易になる．そこで，未知の表現を含む方言の発話であっても正しく分類できるか検証を行う．検証は 3 地域のうち，ひとつの地域の方言テキストからその地域特有の特定の方言を含む方言テキストを除いたものを学習データとして用いる．取

り除いた方言テキストはテストデータとして学習を行う．香川は「むつご (い)」を含む方言テキスト，大阪は「さかい」を含む方言テキスト，博多は「ばってん」を含む方言テキストを除き，学習データを作成した．また，方言テキストを除かない場合の学習データも同様に作成し,「むつご (い)」「さかい」「ばってん」を含むテストデータを正しく分類できるか検証を行った．これを用いて，特定の方言を除いた場合と除かない場合の分類の比較を行う．

## 5.1 香川の方言の抽出

学習する方言テキストから「むつご (い)」を除いた場合の分類結果を表 1 に示す．

表 1: 香川「むつご (い)」を除いた場合

| | 香川 | 大阪 | 博多 | 不明 |
|---|---|---|---|---|
| 分類数 | 20/24 | 0/24 | 0/24 | 4/24 |
| 正確度 | 83.33334% | 100% | 100% | / |

香川の例では 24 文書中 20 文書の分類に成功している．大阪，博多に分類された方言テキストはなかった．どの地域にも分類されなかったテキストは 4 件である．どの地域にも分類されなかったテキストの例を次に示す．

超〜〜〜〜〜〜おいしかった
生クリームも予想を裏切られてん！
甘くないんよ !!! まじで !!!
ほんまにちょ〜〜〜〜〜どいい甘さなん！
でも半分を超えたあたりから、めっちゃむつご〜
なって残してしまったんやけどな涙

分類されなかったテキストは「むつごい」の活用形であるという点で共通している．

次に，学習する方言テキストから「むつご (い)」を除かない場合の分類結果を表 2 に示す．

表 2: 香川「むつご (い)」を除かない場合

| | 香川 | 大阪 | 博多 | 不明 |
|---|---|---|---|---|
| 分類数 | 10/10 | 0/10 | 0/10 | 0/10 |
| 正確度 | 100% | 100% | 100% | / |

表 2 から分かるように，テストデータは全て香川へ分類されている．表 1，表 2 から学習データに「むつご (い)」を含まない場合より，香川への分類の正確度が高いことが読み取れる．

## 5.2 大阪の方言の抽出

学習する方言テキストから「さかい」を除いた場合の分類結果を表 3 に示す．

表 3: 大阪「さかい」を除いた場合

| | 香川 | 大阪 | 博多 | 不明 |
|---|---|---|---|---|
| 分類数 | 1/16 | 12/16 | 0/16 | 3/16 |
| 正確度 | 93.75% | 75% | 100% | / |

大阪の例では，香川には 1 文書が誤って分類されている．大阪へは 16 文書中 12 文書の分類に成功した．博多への分類は 0 件であった．香川に誤って分類された方言テキストの例を以下に示す．

最初の方、読んでみるさかい
これでも君とは長いつきあいや、ちょっと読んでみたら、君がウソついてんのかわかるはずや
えーよえーよ、なんぼでも読んでみて
では、えーっとなになに……最初は、の章、これ自伝か？やすし
そうや、俺の生きてきた証、そのもんじゃ

この例では,「つきあいや」や「はずや」など「〜や」といった表現がよく使用されている．このような表現は，香川の方言テキストでは「〜やけん」,「〜やきん」といった形で複数使用されているため，誤って分類されたと考えられる．MeCab での香川の学習データの分かち書きの結果をみたところ「やけど」という方言が「や」と「けど」として分かち書きされていたため，分類がより困難となったと思われる．

また，どの地域にも分類されなかったテキストは 3 件存在する．分類されなかったテキストの例を以下に示す．

今は法律もしっかりしてるさかいに
組合そのものに存在理由なんかあらへん
悪いこと言わんさかいに
不安定な雇用環境がいややったら
派遣、パート、バイトをやめるこってすわ

分類されなかった方言テキストの共通点は「〜ねん」という表現が出現していない点である．学習する方言テキストを N-gram に分割した場合に「〜ねん」は出現頻度が高い．これらのテキストは「〜ねん」という表現が出現しなかったため，正しい分類が行えなかったと考えられる．

次に，学習する方言テキストから「さかい」を除かない場合の分類結果を表 4 に示す.

表 4: 大阪「さかい」を除かない場合

| | 香川 | 大阪 | 博多 | 不明 |
|---|---|---|---|---|
| 分類数 | 0/10 | 9/10 | 0/10 | 1/10 |
| 正確度 | 100% | 90% | 100% | / |

「さかい」を除かない場合，10 文書中 9 文書が正しく分類された．どの地域にも分類されなかったテキストは 1 件である．大阪の例でも同様に「さかい」を除いた場合より除かない場合の方が正しく分類できていることがわかる.

### 5.3 博多の方言の抽出

学習する方言テキストから「ばってん」を除いた場合の分類結果を表 5 に示す.

表 5: 博多「ばってん」を除いた場合

| | 香川 | 大阪 | 博多 | 不明 |
|---|---|---|---|---|
| 分類数 | 1/43 | 0/43 | 41/43 | 1/43 |
| 正確度 | 97.67442% | 100% | 95.3488% | / |

博多の例では，香川へは 1 文書分類されたが，大阪へ分類された方言テキストはなかった．博多へは 43 文書中 41 文書の分類された． 香川に誤って分類された方言テキストを次に示す.

> オレんこつば治しゅんは簡単ばい
> だけん、オレん意思でしきるがらばい
> オレんこつな、オレん考え方いっちょで、
> どげんにばってんなるたい
> だけん、環境ば整えればどげんにばってんなり、

大阪の「さかい」を除いた場合の香川に誤って分類された例と同様に，MeCab で分かち書きを行う際,「オレん」という博多の方言の表現が「オレ」と「ん」に分類されていた．そのため，博多への分類が困難であったのではないかと考えられる.

次に，学習する方言テキストから「ばってん」を除かない場合の分類結果を表 6 に示す.

「ばってん」を除かない場合，香川，大阪への分類は 0 件であった．博多への分類は 10 文書中 9 文書であった．どの地域にも分類されなかったテキストは 1 件検出された.

表 6: 博多「ばってん」を除かない場合

| | 香川 | 大阪 | 博多 | 不明 |
|---|---|---|---|---|
| 分類数 | 0/10 | 0/10 | 9/10 | 1/10 |
| 正確度 | 100% | 100% | 90% | / |

## 6 分類結果の考察

以上から，方言を除かない場合は除いた場合と比べて大阪、香川の結果では分類の正確度は高いことがわかる．また，除かない場合は他の地域への誤った分類はどの地域の分類結果でも見られなかった．しかし，大阪と博多の検証結果ではどこの地域にも分類されない方言テキストが検出された．これらは，方言を除いた場合は香川へ分類されている．したがって，学習データに地域特有の方言が多く含まれている場合の方が誤った分類を減らすことが可能であると考えられる.

## 7 おわりに

本稿では，多地域にわたる方言解釈システムの基盤となる地域性推定に関する検証について述べた．以上より，機械学習を用いた分類では，多くの方言テキストは正しく分類されていることが読み取れる．大阪の検証での問題点は，方言の辞書への登録を行い，形態素解析の際に方言を形態素として分かち書きすることを可能にすることによって正しい地域への分類を期待できる．また，今回の検証ではテストデータが少数であったため，正確度にモデルごとのばらつきが生じている．十分なテストデータを用いて検証を行うことで，より正確な分類の正負の判断が期待できる.

## 参考文献

[1] 廣田壮一郎，高村大地，奥村学 (2013)．方言コーパスの効率的な収集システムの作成

[2] 平博順，春野雅彦 (2000)，Support Vector Machine によるテキスト分類における属性選択

[3] 瀧本恵理，奥村紀之 (2013)．方言コーパスの基づく文章の地域性の推定