2B1-3

方言コーパス収集システムの構築

Construction of dialect corpus collection system

廣田 壮一郎 *1　笹野 遼平 *2　高村 大也 *2　奥村 学 *2
Soichiro Hirota　Ryohei Sasano　Hiroya Takamura　Manabu Okumura

*1東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

*2東京工業大学精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology

A lot of people use dialects on Web as spread of consumer generated media such as social network services and blogs. Accordingly, the demand for the natural language processing technology that can robustly deal with dialects has been growing. Although large-scale dialect corpora are useful for development of such technology, it is very costly to create such corpora. In this paper, we propose a dialect collection system based on active learning. Our experiment shows that our system can collect dialect sentences at low cost with high precision.

1. はじめに

ブログ等の CGM の普及により Web 上で方言が使用される機会が増えている．それに伴い，方言に対しても頑健な言語処理技術の必要性が高まっている．しかし，方言を含む文の処理を行おうとした場合，標準語とは異なる独自の表現を含むため標準語を想定して構築されてきた従来の言語処理技術をそのまま用いることはできない．たとえば，以下のような博多弁の文を MeCab*1 を用いて形態素解析を行った結果を図 1 に記す．

(1)　そげんこつ無か

「そげんこつ」は標準語の「そんなこと」,「無か」は標準語の「無い」にそれぞれ相当し，正しくは「そげん」,「こつ」,「無か」の 3 つに分割されるべきであるが,「そ」,「げんこつ」,「無」,「か」の 4 つの形態素であると解析されてしまう．

方言に対しても頑健な自然言語処理システムの構築法としては，まず，対象の方言で使用される単語を辞書に登録していく方法が考えられる．しかし，様々な方言に対し，その方言に特有な単語を網羅的に登録し，さらに，適切な品詞を付与するには専門的な知識が必要であり，一般のユーザが行うことは容易ではない．次に，対象の方言のコーパスを大量に準備し，方言用の言語モデルを自動で学習する方法が考えられる．方言ごとに十分な量のコーパスが準備できれば，入力テキストがどの方言で記述されているかを識別するシステム，および，特定の方言に最適化された解析システムの構築が可能であると考えられ，これらのシステムを組み合わせることで，方言を含むテキストに対しても頑健な言語処理システムの構築ができると考えられる．しかし，種々の方言の大規模なコーパスを人手で収集するコストは非常に大きいと言え，低コストで対象の方言を収集する技術が必要となる．

そこで本研究では，能動学習を用いて方言識別モデルを構築することにより，方言にも頑健な言語処理システムの構築に有用となる方言コーパスを低コストに構築する手法を提案する．

連絡先: 廣田 壮一郎，東京工業大学，神奈川県横浜市緑区長津田町 4259 R2-728，hirota@lr.pi.titech.ac.jp

*1　MeCab(https://code.google.com/p/mecab/) は Ver.0.994 を使用し，辞書としては IPADIC を使用した．以降，本稿における実験では同様の条件で形態素解析を行った．

形態素	そ	げんこつ	無	か
品詞	名詞	名詞	助動詞	助詞

図 1: MeCab による方言文の形態素解析の例

2. 関連研究

方言はもともと話し言葉として使用され，書き言葉としては使用されることは少なかった．このため，計算機を用いた方言の処理を目的とした研究は，音声認識の分野において多くみられる．小林ら [小林 09] は方言文を対象として言語認識システムの開発を行った．小林らは人手で方言の単語を辞書に登録し，入力した方言文を一度標準語文に変換した後，その標準語文を形態素解析することで元の方言文の解析精度を向上させている．その際，12 歳から 20 歳までの男女 200 人に方言に関するアンケートを実施し，実際に現在使われている方言, またその頻度を調査した．そのアンケートの中から有用だと思われる方言文を取得し，利用している．平山ら [平山 12] は方言の音声認識精度向上のために方言・共通語の対訳コーパスを用いている．標準語と方言の共起頻度を計算し，大規模な標準語コーパスから方言コーパスに訳す事で得られる方言文を方言コーパスとして利用している．中本ら [中本 12] はまず広島弁で書かれたブログから 658 文を収集し，続いて，日本語形態素解析システム ChaSen*2 に広島弁を形態素解析するために必要となる単語情報や活用情報を追加した上で，方言情報が追加された ChaSen を用いて言語モデルを作成し，広島弁の音声認識の精度を向上させている．

以上のように計算機を用いた方言文処理に関する研究は多く存在するが，それらの多くは研究の前段階として独自に方言データを用意しており，種々の方言に適用しようとすると大きなコストが必要となる．

3. 収集対象とする方言テキスト

Web 上での方言の使われ方は以下に挙げるようないくつかのタイプに分類できる．

*2　http://chasen-legacy.sourceforge.jp/

1. テキスト全体が方言で記述されている場合
   e.g. なんか新しかラーメ屋のあっぱい．昔はよ～来よったばってん今は昔んごとは中洲も歩かんですたい．だけん新しか店もよう知らんけん，少し寂しく感じるですたい．

2. 複数の人の会話文や複数の人の投稿から成るページで，会話文や一部の投稿のみが方言となっている場合
   e.g. 「中洲にうまかラーメン屋があったばい！」と親父が言うと．そこでお母さんも「またラーメンのこつばっか.」との一言．

3. 方言の解説や，方言の持つ印象を利用することを目的とし，特定のフレーズや単語のみが方言となっている場合
   e.g. う，うまかばってん!! がく？ ついつい謎の方言？も出てしまいます．たらーっ（汗）フワッと軽いバウムクーヘンは，ほんのり温かく，やさしい甘さでとてもおいしいです．

本研究では，入力テキストがどの方言で記述されているかを識別するシステムや，特定の方言の言語モデルの構築に利用できる方言コーパスの収集を目的とすることから，タイプ 1 のように全体が方言で記述されているテキストを収集対象とする．

4. 提案手法

本研究では，人手によるタグ付けにかかるコストを可能な限り小さくすることを目的とし，以下の 3 つのステップにより方言コーパスの収集を行う．

ステップ 1: 方言文の可能性がある文を効率的に収集するための初期方言識別モデルの構築

ステップ 2: 能動学習に基づく方言識別モデルの高精度化

ステップ 3: ステップ 2 までで構築された方言識別モデルを利用した方言コーパスの収集

いずれのステップにおいても，まず，対象の方言に特徴的な表現を検索エンジンにクエリとして与え，その検索結果を方言テキストの候補とし，人手による分類や学習した方言識別モデルに基づく識別を行い，方言テキストを収集する．本研究では検索結果に含まれる連続する 5 文を 1 セットとして方言テキストの収集を行う．

4.1 ステップ 1:初期方言識別モデルの構築

まず，方言文の可能性がある文を効率的に収集するための初期方言識別モデルの構築を行う．概要を図 2 に示す．

まず，ユーザに収集目的とする方言に特徴的な表現を複数入力してもらい，それらを検索エンジンのクエリとし，検索結果を取得する．検索結果の上位の各 Web サページからテキストを取得し，方言テキスト候補となる 5 文を抽出する．方言テキスト候補となる 5 文は，クエリを含む文に隣接する 5 文とする．たとえばクエリを「うまか」とした結果，図 3 に示すようなテキストが得られたとするとクエリが出現する 2 文目に隣接する 5 文，すなわち 3 文目から 7 文目の 5 文を方言テキスト候補として収集する．

このようにして得られた方言テキスト候補の集合から，明らかに目的とする方言テキストであると考えられるテキスト候補を 20 セットユーザに選択してもらう．その上で，選択された 20 セットを正例，事前に構築した標準語テキスト集合を負

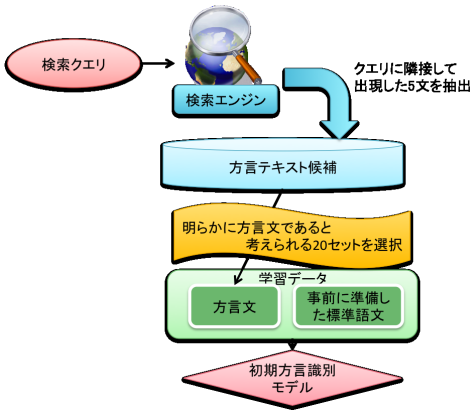

図 2: 初期方言識別モデルの構築

例とし，初期方言識別モデルを構築する．ここで，方言テキスト候補を方言文と標準語文の 2 値に人手で分類し，それらを学習データとすることも考えられるが，検索によって得られた方言テキスト候補は，クエリによって方言テキストの割合に大きな偏りがあるため，一律に方言文とそれ以外に分類するのは効率が悪いと考えられることから，明らかな方言文のみタグ付けすることでタグ付けコストを少なくしている．

本研究では，方言文の判別を行うために，Support Vector Machine (SVM) を使用した．SVM の素性には文字 2-gram，文字 3-gram，形態素 1-gram，形態素 2-gram を利用した．方言を含むテキストは，適切な辞書が存在しないため正しく形態素解析を行うことができない可能性があるが，学習データの方言文とテストデータの方言文を同様の基準で解析するため，形態素 n-gram も素性として一定の効果があると考え，形態素 n-gram も素性に加えた．

4.2 ステップ 2:能動学習に基づく方言識別モデルの高精度化

続いて，能動学習を用いて方言識別モデルの高精度化を行う．概要を図 4 に示す．

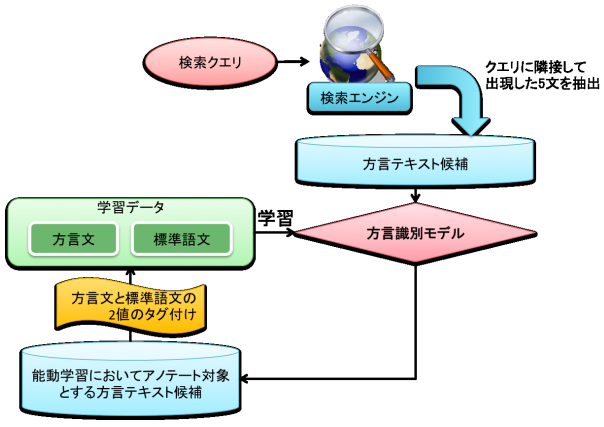

図 4: 能動学習に基づく方言識別モデルの高精度化

まず，初期方言識別モデルを構築した際と同様にユーザに収集目的とする方言に特徴的な表現を複数入力してもらい，それらを検索エンジンのクエリとし，検索結果を取得し，方言テキスト候補を取得する．続いて，取得した方言テキスト候補にステップ 1 で作成した方言識別モデルを適用し，識別境界に近

1	本場の中華料理より、日本で食べる中華料理のほうが美味しい
2	行っちるお店の悪かんかもしれんばってんね、どこさ行っちも日本んほうの うまか〜ばい
3	そげなん日本や考えられんけんばいね
4	うちは今までんなん回かよそ旅行に行ったこつのいるたい
5	タイはホテルん中で食べる食事はよかとやけど、外で食べるっちきは要注意やった
6	匂いのきつくよったまに腐っちいるもんもあっけんばい
7	そいな、っちゆうこつばい
8	１度お米不足ん時期のちゃて、日本にもアジアんお米の流れてきよったっちきのあったったいねね
9	残念なのらヨーロッパやアメリカ大陸へは行ったこつのなか

図 3: Web テキストの例

い候補 [Settles 10] のみ 20 候補を出力し，それらの候補を対象に人手で方言テキストであるか否かのタグ付けを行い，タグ付け結果を学習データに追加し，方言識別モデルの再学習を行う．この手順を 1 周期とし能動学習を行っていく．

4.3 ステップ 3:大規模な方言テキストの収集

ステップ 2 までで構築された方言識別モデルを用い，大規模な方言テキストの収集を行う．概要を図 5 に示す．

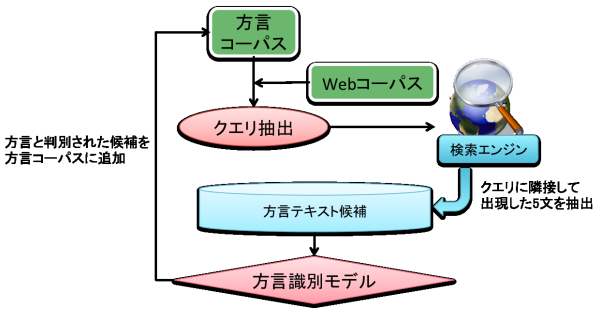

図 5: 大規模な方言テキストの収集

このステップでは，検索エンジンを用いて方言テキスト候補を収集し，それに方言識別モデルを適用することにより，方言テキストの自動収集を行う．ステップ 2 までで構築された方言識別モデルの再学習は行わない．方言テキスト候補を効率的に収集するためには，方言が含まれている Web ページをより多く検索できるような検索クエリを決める必要がある．そこで，収集した方言テキストに含まれる文字 2-gram，および，文字 3-gram をクエリの候補とし，以下の式 (1) によりランキングした結果，スコアが高かった表現を新たなクエリとして使用する [*3]:

$$score(q) = \frac{P_d(q)}{P_c(q)}. \quad (1)$$

ここで $P_d(q)$ は収集された方言テキスト集合に占める対象のクエリ候補 q を含むテキストの割合を表し以下の式 (2) により計算される:

$$P_d(q) = \frac{\sum_{d \in D} \delta(q \in d)}{|D|}. \quad (2)$$

D は方言文と分類されたテキストの集合，$\delta(q \in d)$ はテキスト d 内に単語 q が含まれたら 1，それ以外なら 0 となる関数である．また，$P_c(q)$ は Web 上におけるクエリ候補 q の出現率であり以下の式 (3) により計算される:

$$P_c(q) = \frac{\sum_{s \in C} count(q, s)}{\sum_{s \in C} \sum_{q'} count(q', s)}. \quad (3)$$

C は事前に用意した Web コーパスに含まれる文の集合，$count(q, s)$ は文 s 内に単語 q が出現する回数とする．Web コーパスとしては，検索エンジン基盤 TSUBAKI [Shinzato 08] で使用されていた 5 億文を使用した．

5. 評価実験

5.1 実験設定

実験では検索エンジンとして Google を使用し，検索対象をブログ記事に限定して実験を行った．また，本研究では博多弁と大阪弁を対象に方言の収集実験を行った．使用する SVM は LibSVM[*4] を使用して，カーネル関数に線形カーネルを用いた．

構築した方言識別モデルの評価を行うため以下の手順により評価用データを作成した．まず,「そげん」,「ばってん」,「とーと」,「げな」,「ばり」の 5 単語の全ての対となる組み合わせ 10 通りをクエリとして検索エンジンに与え，各検索結果から上位 20 件を取得し，各クエリごとにクエリに隣接して出現した連続する 5 文をアノテート対象とした結果，方言テキスト 93 セット，標準語テキスト 136 セットから成る評価データが作成された．ここで，10 通りのクエリにより検索された計 200 件の検索結果に対し，2 セットずつ評価セットが作られることから，すべて使用すると 400 セットから成る評価セットが構築されることになるが，重複して出現したセットは 1 つにまとめたため，229 セットから成る評価セットとなった．

ステップ 2 における能動学習は 2 回行い，識別モデルの出力が-0.2 から 0.2 の間となった方言テキスト候補を能動学習におけるアノテートの対象とした．初期方言識別モデルの構築の際にクエリとして「うまか」と「ばい」を，1 回目の能動学習におけるクエリとして「たい」と「ばってん」を，2 回目のクエリとして「たとよ」と「やろ」をそれぞれ使用した．能動学習の回数は 2 回であることから，実際に人手で行う必要がある作業は，合計 6 つのクエリ選択と，ステップ 1 において方言テキスト候補から対象の方言テキストを 20 セット選択する作業と，能動学習時に合計テキスト候補 40 テキストを方言テキストとそれ以外に分類する作業である．

5.2 実験結果

5.2.1 能動学習による精度の向上

まず，能動学習の効果を確認するため，初期方言識別モデルと，1 回目，2 回目それぞれの能動学習の後に得られた方言識

*3 ただし，実際には同一のフレーズが何度もクエリとして使用されるのを防ぐため，一度使ったフレーズのスコアは低くなるよう修正して使用している．

*4 LibSVM(http://www.csie.ntu.edu.tw/čjlin/libsvm/) は Ver. 3.17 を使用した．

表 1: 方言識別モデルの精度

モデル		適合率	再現率	F 値
初期方言識別モデル		0.85	0.43	0.57
能動学習 (識別境界付近をタグ付け)	1 回目	0.95	0.56	0.71
	2 回目	0.94	0.68	0.79
検索結果上位をタグ付け	1 回目	0.95	0.41	0.58
	2 回目	0.93	0.60	0.73

表 2: 検索クエリごとの得られた方言テキスト数

検索に利用したクエリ	得られた方言テキスト数
ち思う　ちゆう	176
こつば　ち思っ	171
けんん　っちゆ	147
...	...
たち猫　投資人	0
抹茶ん　なー	0
と一人　かん程	0

別モデルの精度を求めた．また，比較対象として，能動学習と同じクエリで検索した上位 20 件から得られた方言テキスト候補をアノテート対象とした場合の精度も求めた．結果を表 1 に示す．

能動学習を行うことにより F 値が大きく向上していることが確認できる．また，能動学習を用いなかった場合と比べても，同じアノテーション回数でより大きく精度が向上していることが確認できる．

5.2.2　収集された方言コーパスの精度

次にステップ 3 により自動収集された方言コーパスの精度を調べるため，自動収集された方言コーパス 1,000 セットから無作為に 100 セットを選択し，それらが博多弁であるかどうか人手で判定を行った．その結果，96 セットが博多弁であると判定され，高い精度で方言を収集できていることが確認できた．

5.2.3　検索クエリによる違い

ステップ 3 では自動生成される検索クエリによって得られる方言テキスト候補の数は大きく異なる．表 2 に多くの方言テキスト候補が得られた検索クエリのペア上位 3 ペアと 1 つも方言テキスト候補が得られなかったクエリ 3 ペアを示す．たとえば「ち思う」は以下のような博多弁の文で使用される表現であり，博多弁に特徴的な文字列がクエリとして選択された場合に多くの方言テキストが得られていると考えられる．

(2)　食べきらんっち思う てからくさ！

一方,「たち猫」や「投資人」などのように博多弁とは関係のない文字列がクエリとして選択された場合は方言テキストをほとんど収集することができず，より効率的に方言テキストを収集するためには，より効果的なクエリ選択が必要となると考えられる．

5.3　収集される方言テキスト数の変化

ステップ 3 において方言テキストを収集する際，同一の方言テキストは収集せず，また，時間が経つにつれて使用されるクエリも適切でないものの割合が増えてくると考えられることから，収集が進むにつれて 1 検索あたりの収集される方言テキスト数は減ってくると考えられる．このため 200 検索ごとに収集された方言テキスト数の変化を調査した．結果を図 6 に示す．収集が進むにつれて検索クエリあたりの収集される方言テキスト数が減っていることが確認できる．また，2,000 検索を行った結果，収集された総方言テキスト数は 4,341 セットであったが，検索回数が増えても一定の割合で新たな方言テキストを収集できていることから，さらに収集を続けることで，さらに多くの方言テキストを獲得できると考えられる．

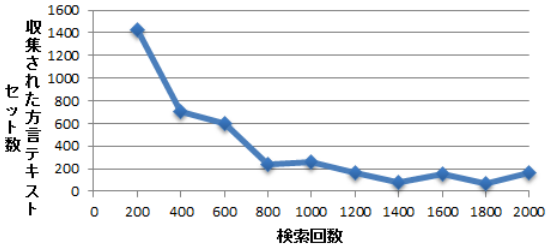

図 6: 200 検索ごとの収集された方言テキスト数

5.4　博多弁以外の方言の収集

提案手法が他の方言にも適用できるか確かめるため，大阪弁に対しても博多弁と同様の方法で収集実験を行った．この際，初期方言識別モデルの構築のためのクエリとして「あかん」と「おもろい」を，1 回目の能動学習におけるクエリとして「おおきに」と「なんぼ」を，2 回目のクエリとして「ほんま」と「かまへん」をそれぞれ使用した．自動収集された大阪弁コーパス 1,000 セットから無作為に 100 セットを選択し，それらが大阪弁であるかどうか人手で判定を行った結果，90 セットが大阪弁であると判定され，提案手法は博多弁以外の方言に対しても，有効であることが確認できた．また，大阪弁の場合，2,000 検索を行った結果，収集された方言テキスト数は 26,074 セットであった．

6.　まとめと今後の課題

本研究では方言コーパス作成のための低コストな方言収集システムを提案した．博多弁を対象とした実験の結果，6 つの博多弁クエリの入力と，約 60 方言テキスト候補文をタグ付けすることで 96%という高い精度で 4,341 セットの方言テキストを収集可能であることを確認した．しかし，1 検索ごとに収集される方言テキストの数は多いとは言えず，大規模なコーパスの構築のためには検索を繰り返し行うことが必要であるため，良い検索クエリの生成方法が今後の課題として挙げられる．

参考文献

[Settles 10] Settles, B.: Active Learning Literature Survey, in *Computer Sciences Technical Report 1648*, pp. 1–67 (2010)

[Shinzato 08] Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C., and Kurohashi, S.: TSUBAKI: An open search engine infrastructure for developing new information access methodology, in *Proc. of IJCNLP'08*, pp. 189–196 (2008)

[小林 09] 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, 第 23 回人工知能学会全国大会, pp. 1–4 (2009)

[中本 12] 中本 典子, 目良 和也, 黒澤 義明, 竹澤 寿幸 : 広島弁音声認識のためのコーパスと言語モデルの構築, 言語処理学会第 18 回年次大会, pp. 883–886 (2012)

[平山 12] 平山 直樹, 森 信介, 奥乃 博 : 方言音声認識のための話し言葉言語モデル構築, NLP 若手の会 (YANS) 第 7 回シンポジウム, pp. 1–7 (2012)