単語音素のベクトル化による言語地図作成

近藤泰弘
青山学院大学
yhkondo@cl.aoyama.ac.jp

概要

同じ意味の単語について，方言によって語形が異なることが一般的だが，その細かな語形を分類して，その方言形の採取地点の地図上に対応するアイコンを配置したものを，「言語地図」という．これまでは，言語地理学の分野で，「言語地図」は単語の語形を見ながら，研究者がそれを分類して，アイコンの色や形を決めていたが，本研究では，語形をその音素の分布のベクトルに変換し，機械学習によって分類することで，アイコンの種類の決定を自動化する手法を提案した．単語のベクトル化は．文脈から意味の分散表現による埋め込みベクトルを作成するのが一般的だが，言語地図のための方言の語形には，そのような意味による文脈は存在しないため，今回は，文書ベクトルの BoW(Bag of Words）の手法に習い，語形の音素の BoC（Bag Of Characters）表現を用いて，音素の頻度ベクトルを用いて分類を行った．そして，音素の unigram の BoC のベクトルによる方法に実用的な優位性があることを示した．

1 はじめに

社会言語学の研究の中では，方言語形を地点ごとに配置するいわゆる「言語地図」を作ることが一般的である．そのような方法は「言語地理学」と言われることが多い．その初期には手作業でスタンプを地図に押印する方法がとられたが，近年では，GIS の手法を用いて，電子的に表現する方法が一般的である．「言語地図」の研究手法は，古く，柳田国男の『蝸牛考』[1] で知られるように（例えば，列島の中心部で「デンデン」，その回りに「マイマイ」さらにその周辺部で「カタツムリ」，そして「ナメクジ」があるように見えることから，周圏分布などと呼ばれた）日本語の古語の語形の歴史的な流布順序などを推定するのに役立つとされている．方言語形は非常に多様で．「かたつむり」の方言形のひとつである「ナメクジ」系にしても，「ナメクシ・マメクジ・ナメクジリ・ナメラ・ナメクジナ」など極めて多くのバリエーションがある．これらをひとまとまりとして．「マイマイ」系や「デンデンムシ」系などをまとめたものと，はっきりと区別するように表示する必要がある．しかし今回，テストデータとして用いた「日本言語地図」データなどでは全国に 2800 地点分程度の方言語形があり，それぞれが微妙なバリエーションを示している．したがってこれを全部，研究者の判断で分類して，特定のアイコンに結びつけることは非常に手間が掛かり，言語地図を作るうえでの障害であったため，その自動化の手法を提案したい．

今回はそれを自動化するために，各単語ごとの音素（ローマ字 1 字に相当）の頻度を測り，Bag of Characters あるいは，Bag of character N-gram (n=1 または 2) のようにして，各単語について 26 次元のベクトルを作成した．実際は，その中で今回は使わない，音素に含まれない文字，例えば X などもあるが，今後の拡張も考えて 26 個とした．

2 関連研究

日本語の言語地図をコンピュータで作成する試みは既になされている．古くはメインフレームの時代の荻野綱男の GLAPS，[2] PC になってからの福嶋（尾崎）秩子の SEAL [3] などがある．大西拓一郎も，Adobe イラストレータや GIS 系のアプリを駆使して言語地図を発表している．[4] Python を用いたものも，松浦年男 [5]，加藤幹治 [6] のものなどが公開されている。

類似した発想によるものとして，語形（音形）のレーベンシュタイン距離を用いて，方言間の差異を計測した研究があり，参考になる [7]．[8] 日本語においても，方言分類を論じる場合に使われた．しかし，それらで既に示されているように，音素（ローマ字と同一と考える）のレーベンシュタイン距離では，母音字と子音字とが区別なく扱われるため，母音交替や子音交代が頻繁に起きる日本語の実状にあ

表 1　DENDENMUSI の語形音素ベクトル

0	0	0	2	2	0	0	0	1	0	0	1	2	0	0	0	0	1	0	1	0	0	0	0	0	0
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z

わない部分がある．

また，モーラの unigram を用いて，方言間の距離を計算して，方言を分類するという研究もある [9]．これも母音・子音の交代を反映できない問題がある．また，方言区画を計算する研究もある．[10]

以上のように，語形の距離を測るという方向での研究は多いが，本研究のような手法を用いたものは管見では，世界的に存在しない．

3　提案手法

1 節で述べた通り，本研究では，単語の語形のローマ字表記をデータとして用いた．ローマ字表記は，ほぼ音素表記と同等のものと考えることができる．それをベクトル化し，その後に，教師なし機械学習で 2 次元への圧縮を行い，クラスタに分類することで，その 1 クラスタごとにアイコン（今回は色）を与えて，アイコン付与の自動化を実現した．

3.1　モデル

表 1 のように、アルファベットの 26 次元のベクトルを用意し、それぞれに語形内（この例は「デンデンムシ DENDENMUSI」）の各文字の使用頻度を入れる。つまり、このベクトルはある単語の語形（の構成音素）の持つ母音・子音の分布と頻度情報を持っていることになる．ここでは N = 1 (unigram) なので、字の順序情報は入っていない．

この語形音素ベクトルを地図ごと（地図により、それぞれ 1500 から 2800 余の語形がある）に、機械学習で適宜次元圧縮して、さらにクラスタに分けて、音素ベクトルの特徴量によって語形アイコンの色分けを行った．

なお，比較のため，N = 2 (bigram) のものも作成した．その場合は，aa から始まり zz までの 676 次元のベクトルを作成するので，DENDENMUSI の場合は，DE,EN,ND,DE,EN,NM,MU,US,SI となり，DE と EN が頻度 2，他が 1 となる．

3.2　学習

各地点の音形（ローマ字）を語形ごとに音素の頻度を計算し，BoC の形にベクトル化した．その後，そのベクトル化したデータを，学習データとして，Python のライブラリの scikit-learn の PCA を用い，26 次元ベクトルのまま学習させた．また，それによって次元圧縮し，いくつかの次元のデータを作成した後，最適な次元を決定し，またその低次元データを K-Means で学習させることで，クラスタに分類した．クラスタ数の最適化については，語形分類が実用的にできるかどうかを主として考え，エルボー法などで結果を裏付けた．最終的な地図の見やすさの評価については，今後の課題とすべき点が残っている．

3.3　分類結果の表示

クラスタに分類した語形を，アイコンの色に置き換えて，方言の採取地点の緯度経度によって指定された位置に配置する．地図のスケールの動的な変更を用いることによって，従来の静的な言語地図よりも見やすい表示を実現する．

4　実験と考察

4.1　実験設定

提案手法の評価には，代表的な日本語方言の言語地図用のデータセットである「日本言語地図データベース」（国立国語研究所編）[11] を用いた．公開されている「日本言語地図」データには各種あるが，ここでは，熊谷康雄が開発したものを用いた．データは，CSV 形式で，語形（ローマ字音素表記），地点番号，緯度，経度，からなっている．

日本言語地図のすべての語形についてではなく，その一部についてのみ公開されているが，今回はその中から「かたつむり」と「くるぶし」のデータを用いた．それぞれ，2800 件のデータ数である．処理プログラムの実装には Python 3.8.16 と scikit-learn 1.0.2 を用いた．地図へのプロットには動的地図を作成できるため，folium 0.14.0 を用いた．

4.2　結果

実験結果を表 2 に示す．このように，BOW の音素の単位を unigram にしたばあいと bigram にした場合を比較すると，どちらの語でも unigram の方がよい成績を示した．bigram では出現順情報の一部を捉えられるが，ベクトルがかなりスパースになるため，データ量が少ない今回のようなものでは精度が出ないものと考えられる．また，図 3 に 7 クラスタでの散布図を示した（★印は中央点）．図 4 から図 6 が，実際の言語地図の例である．

表 2　主成分の分散の寄与度 (第 2 成分まで)

かたつむり	unigram	0.49	0.12
かたつむり	bigram	0.41	0.17
くるぶし	unigram	0.31	0.23
くるぶし	bigram	0.20	0.16

表 3　語形と色の対照（合わないものの一部）

BEKOCUBU	gray
CUNODASICUBU	gray
CIDAMI	gray
SUDAMI	gray
PARUUNNA	gray
NAMEKUZIRI	gray
NAMEKUZIRI	gray
NAMEZUKURI	gray
NAMEKUZI	gray
NAMEKUZI	gray

5　クラスター数の調整と性能

K-Means には，クラスター数決定の難しさがある．今回もどれだけのアイコンの種類（色）にするかという点で決定的な問題がある．方言語形は非常に多様であるので，そのすべてを完全に分類することは原理的に不可能であり，おおよそ，代表的な語形が同じ分類になり，頻度の少ない語形については，ある程度，多種の語形が同じ分類になっても実用上はかまわないという点もある．そこで，まず基本的な数値を把握するために，通例のようなエルボー図（図 1 と図 2）を作成したところ，クラスター数 4 あたりで十分と見られる．エルボー図はあくまでも概要を捉えるに過ぎないので，いくつかの値によって作成した言語地図を作り，実際の分類状態を観察してみた．示したのは，クラスター数 7 という値，つまり，アイコンの種類が 7 ということになる．7 の場合，主要語形の「カタツムリ」「デンデン」「ダイロ」系はほぼ完全にそれぞれ 1 つのアイコン色に当てはまり分離ができるが，それ以外の語形の「ナメクジ」系などの資料が少ない語形は表 3 のように同色に配当されてしまう．これはこの方法の限界であるが，地図を見る場合には，色の異なりで，「主要語形とは違う」ことがわかるため，参考資料として使うには問題ない部分がある．

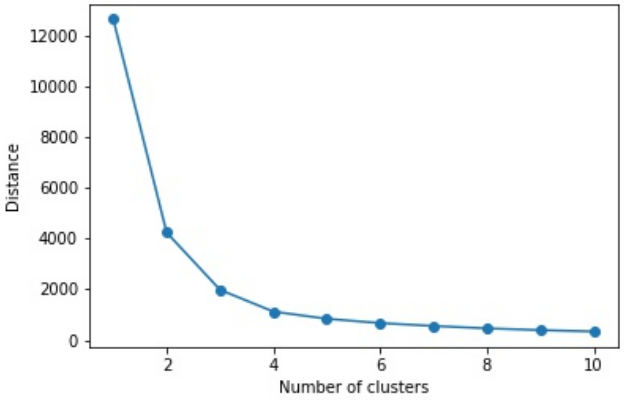

図 1　「かたつむり」のクラスター数調整

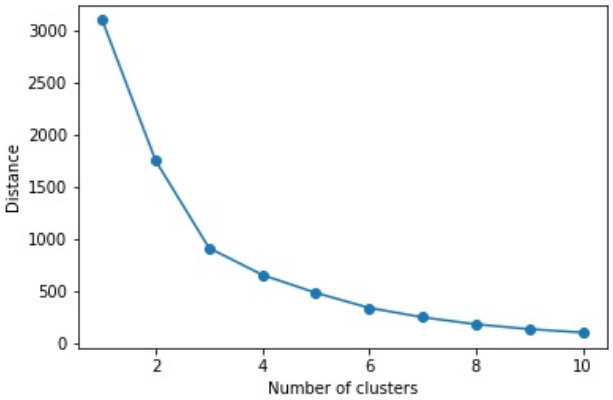

図 2　「くるぶし」のクラスター数調整

6　結論

本研究では，従来不可能だった，言語地図における語形のアイコンへの適切な割り当てを行う手法を提案し，ほぼ当初の目的通りの自動化を実現した．そのために，単語の音素情報のみによって作る BoW 方式の音素ベクトルを作成して機械学習によって分類するモデルを考案し，それによって既存の言語地図データセットを学習することで，語形をある程度分類することが可能であることを示した．また，性能的には，音素の unigram と bigram によるものを比較し，この場合，unigram の方が分類性能が高いことを示した．

今後，語形ベクトルを演算することで，分布の比較の数量化や，方言語形の移動の計算なども行うことを検討したい．

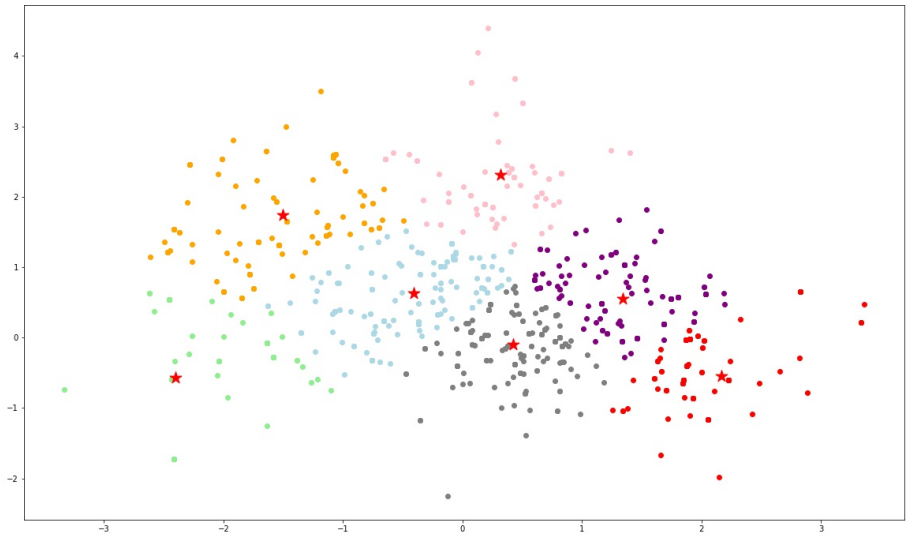

図 3　「かたつむり」unigram 散布図

図 4　「かたつむり」関西

図 5　「かたつむり」東北 (拡大図)

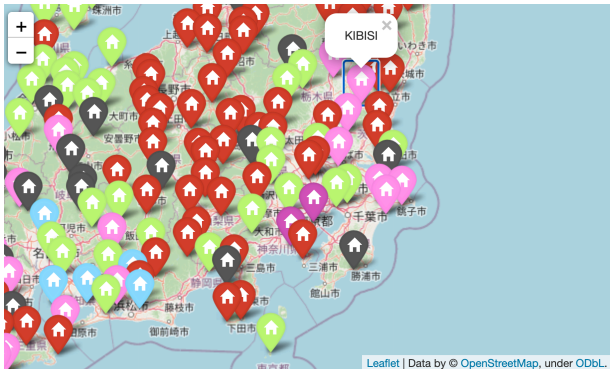

図 6　「くるぶし」関東 (拡大図)

謝辞

本研究の評価データセットに用いた「『日本言語地図』データベース」を開発した，国立国語研究所，熊谷康雄氏，サイトの保守にあたっている大西拓一郎氏に感謝申し上げます．

参考文献

[1] 柳田国男. 蝸牛考. 刀江書院, 1830.
[2] 荻野綱男. コンピュータ言語地理学. 言語研究, Vol. 1978, No. 74, pp. 83–96, 1978.
[3] 福嶋秩子. 言語地理学のへや, 2022. https://www.unii.ac.jp/chitsuko/inet/lg7.html.
[4] 大西拓一郎. 言語地図作成の電算化. 日本語学, Vol. 21, No. 11, 2002.
[5] 松浦年男. Python でプロットした地図を出力する, 2020. https://note.com/yearman/n/n69fa3f2d583d.
[6] 加藤幹治. Python の geopandas を使って，複数の言語特徴に基づいた地図を出力する，2020. https://qiita.com/hamident0/items/1baabe5f7034df838589?fbclid=IwAR1EnwTQwidJDDWoR-5UHzwTSM0dFxJtepnmbKNxEoXW65xRiGDpe4w-6PY.
[7] 鑓水兼貴. 共通語化過程の計量的分析. 東京外国語大学博士論文, 2009.
[8] Wolfgang Vierreck. The computer developed linguistic atlas of england, volumes 1 (1991) and 2 (1997). **International Computer Archive of Modern English: ICAME journal**, Vol. 1997, No. 21, pp. 79–90, 2015.
[9] 金明哲・中村靖子. 文学と言語コーパスのマイニング. 岩波書店, 2021.
[10] INOUE Fumio. Computational dialectology. **Area and Calture Studies**, Vol. 52, No. 1, pp. 68–100, 1996.
[11] 国立国語研究所. 『日本言語地図データベース』, 2016. https://www.lajdb.org/TOP.html.