7K-07

類語の違いを考慮した地域推定に関する研究

山本雄平† 中村健二‡ 田中成典†‡ 坂本一磨‡‡ 中村竜也†‡
関西大学先端科学技術推進機構† 大阪経済大学情報社会学部‡ 関西大学総合情報学部†‡
関西大学大学院総合情報学研究科‡‡

１．はじめに

近年の CGM（Consumer Generated Media）の発達により，インターネット上には，膨大なデータが蓄積されるようになってきた．特に，SNS（Social Networking Service）においては，ユーザが多種多様なライフログを日々投稿している．ライフログは，ユーザの生活習慣や居住地（以下，地域）の文化や伝統に関連している．そこで，ライフログを分析し，地域の属性（以下，地域属性）を推定することで，エリアマーケティングや企業の販売戦略に役立つ有用なサービスの提供が可能[1]である．これを実現するために投稿内容に付与されたジオタグを用いた研究[2]，投稿に含まれる地名[3]や地域独自の方言[4]等を用いて地域属性を推定している研究がある．しかし，ジオタグは付与率が非常に低く，また，投稿に含まれた地名や方言などは都市部に集中するため，既存研究の方法では，正確に地域特性を分析することが困難である．そこで，本研究では，地域ごとの文化や伝統に応じて変化するや単語に着目して，投稿者の地域属性を推定する手法を提案する．

２．研究の概要

本システムの概要を図 1 に示す．入力データは，対象ユーザが投稿した時系列のライフログ（以下，投稿履歴）と対象の類語である．出力データは，対象の類語の地域ごとの使用頻度である．本システムは投稿履歴の分類機能，類語の集計機能，類語の使用頻度の算出機能により構成される．

Research for Reasoning Area Based on Synonym
†Yuhei Yamamoto
Organization for Research and Development of Innovative Science and Technology, Kansai University,3-3-35 Yamate-cho Suita City, Osaka 564-8680, Japan
‡Kenji Nakamura
Faculty of Information Technology and Social Science, Osaka University of Economics
†‡Shigenori Tanaka, Tatsuya Nakamura
Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki City, Osaka 569-1095, Japan
‡‡Kazuma Sakamoto
Graduate School of Informatics, Kansai University, 2-1-1 Ryozenji- cho, Takatsuki City, Osaka 569-1095, Japan

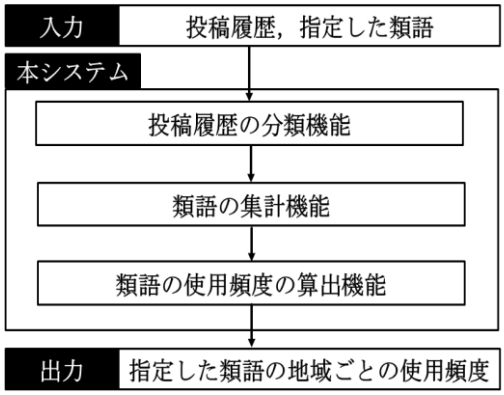

図 1　本システムの概要

２．１　投稿履歴の分類機能

本機能では，投稿履歴を地域区分ごとに分類する．まず，Twitter API を用いて，地域が判明しているユーザを取得し，投稿履歴をそれぞれ該当地ごとに分類する．なお，地域区分は北海道・東北地方，関東地方，中部地方，近畿地方，中国・四国地方と九州地方・沖縄県の計 6 地域とする．これにより，類語の出現回数の集計と使用頻度を地域ごとに算出できる．

２．２　類語の集計機能

本機能では，類語の使用頻度を算出するため，地域ごとの類語の出現回数を地域ごとに集計する．ただし，検索された類語が本来の意味と異なる場合はノイズとして取り除く．例えば，検証対象が「アホ」の場合，「アホウドリ」はノイズに該当する．この作業により，類語を正しく含む投稿のみの検索が可能となる．

２．３　類語の使用頻度の算出機能

本機能では，推定した類語の出現回数を正規化して地域ごとの使用頻度を算出する．出現回数の正規化数を式(1)に示す．

$$C_{Nor} = \frac{N_{Min} C}{N_A} \quad (1)$$

ここで，C_{Nor} は正規化の出現回数，N_{Min} は投稿履歴が最も少ない地域の投稿件数，C は集計した類語の出現回数を意味する．N_A は地域の投稿件数である．次に正規化した類語の出現回数を用いて，地域ごとに類語の使用割合を算出し，使用頻度とする．

３．実証実験

本提案手法の有用性を検証するため，マイクロブログの Twitter から収集した投稿を対象に地域ごとの類語の使用頻度と既存資料[5]に記載されている正解データを用いて地域特性を確認する．

３．１　実験内容

本実験では，2011 年から 2015 年の間に Twitter 上で地域を確認できる 1,514 ユーザの投稿内容 3,328,157 件を対象とする．類語は，既存資料[5]で示されているように，地域で使用頻度が顕著に異なる「アホ」と「バカ」を対象とする．本システムにて分析した結果を Microsoft 社の Bing Maps Platform API を用いて地図上に可視化して差異を確認する．

３．２　結果と考察

本実験で対象とした類語の出現回数と地域ごとの使用割合の結果を表 1 に，そして，その可視化結果を図 2 に示す．

表 1 から「アホ」は近畿地方の出現回数が 78 件（23.1%）であり，最も抽出数が多い結果を得られた．一方，「バカ」は北海道・東北地方の出現回数が 145 件（23.7%）で最も多い結果となった．このことから次に示す事項が明らかになった．

一点目は，類語を分析すると，地域の特性が明らかになることである．図 2 の可視化結果に示すとおり，地域によって類語の使用頻度が異なることがわかる．そのため，投稿者の投稿履歴に含まれる類語の使用頻度を分析することで，投稿者の地域属性を推定できる可能性があることがわかった．

二点目は，本提案手法で分析した結果が信憑性の高い結果であることである．表 1 の結果および既存資料[5]の結果に共通して，「バカ」が全国的に使用されている一方，「アホ」は近畿地方に限定して多く使用されている結果となった．そのため，本提案手法の結果は信憑性の高い結果であることがわかった．

三点目は，類語によって地域特性が顕著になる場合があることである．表 1 の結果から，「バカ」の出現回数の標準偏差が 23 であり，「アホ」の 17 に比べて高い結果となった．そのため，「バカ」の方がより地域特性が顕著であることがわかる．しかし，この結果は地域の分類粒度に影響されるため，今後は都道府県レベルなどに地方の分類を細分化し，地域特性をより詳細に分析することが必要である．

４．おわりに

本研究では，類語の違いを考慮し，マイクロブログの投稿から地域を推定する手法を提案した．本実験の結果より，類語の地域特性を分析すると，投稿者の地域属性を推定できる可能性があることを確認した．今後は，類語抽出に用いる投稿データの数を増やしつつ，分類する粒度を都道府県レベル等に細分化し，提案手法の汎用性を高める予定である．

表 1　類語抽出の結果

	アホ		バカ	
	出現回数（件）	地域割合（%）	出現回数（件）	地域割合（%）
北海道・東北地方	61	18.0	145	23.7
関東地方	44	13.0	114	18.6
中部地方	47	13.9	84	13.7
近畿地方	78	23.1	85	13.9
中国・四国地方	31	9.2	78	12.7
九州地方・沖縄県	77	22.8	106	17.3

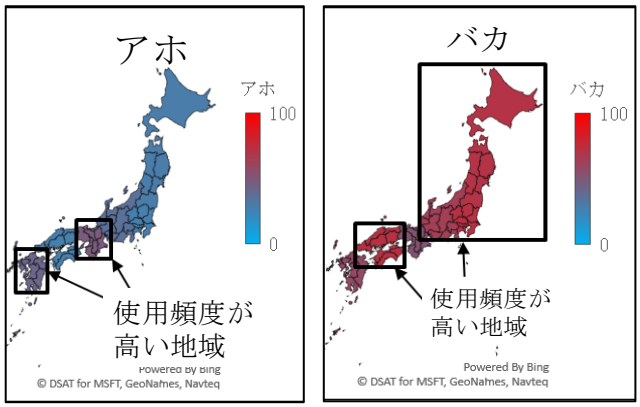

図 2　「アホ」・「バカ」の可視化結果

参考文献

[1] 鳥居大祐，赤塚隼，落合桂一，角野公亮：生活情報を提供するリアルタイム検索サービスの開発，NTT DOCOMOテクニカル・ジャーナル，Vol.20, No.4, pp.12-17, 2013.

[2] 斉藤祐樹，高山翼，山上慶，戸部義人，鉄谷信二：マイクロブログのジオタグと発言コンテキスト解析による行動予測手法，情報処理学会論文誌，情報処理学会，Vol.55, No.2, pp.773-781, 2014.

[3] 石田和成：地域特有の単語共起にもとづく位置推定と地域トピックの考察，情報処理学会研究報告，情報処理学会，Vol.2015, No.2, pp.1-6, 2015.

[4] 瀧本恵理，奥村紀之：方言コーパスに基づく文章の地域性の推定，情報処理学会論文集，情報処理学会，Vol.76, No.2, pp.2.193-2.194, 2014.

[5] 松本修 ：全国アホ・バカ分布考―はるかなる言葉の旅路，新潮社，1996.