Vol.2017-GN-101 No.3
2017/3/10

情報のマッピングを目的としたツイート語義曖昧性解消

遠藤岳[†1]　六瀬聡宏[†2]　内田理[†1]

概要：近年，ソーシャルメディアのデータを活用して実世界の出来事の抽出や予測を行う研究が盛んである．特にTwitterは速報性が高く利用者も多いメディアであるため，リアルタイムな状況把握への活用が期待されており，近年は災害時の活用を指向した研究も活発である．ところで，ツイートに地理情報が付与されていれば地図上にマッピングすることが可能となる．したがって，災害時に被害の状況や支援のニーズなどを詳細に把握することが可能となる．しかし，ジオタグが付与されたツイートは極めて少数であることが知られている．そのため，ツイート本文中から地名を正確に抽出する必要がある．しかし地名と人名が同一表記となることが想定される名称場合，それを一意に特定することが困難な場合もある．そこで本研究では，情報のマッピングを目的としてツイート中の単語の曖昧性解消手法を提案し，実験によりその有用性を検証する．

キーワード：ソーシャルメディア，Twitter，曖昧性解消

A Word Meaning Disambiguation Method for Tweet Texts that Aims for Information Mapping

GAKU ENDO[†1]　TOSHIHIRO ROKUSE[†2]
OSAMU UCHIDA[†1]

Abstract: In recent years, studies on extraction and prediction of real world events utilizing information of social media are extensively conducted. In particular, since Twitter is a medium with high real-time property and a lot of users, it is expected to be utilized for real-time situation prediction. Due to this, studies on Twitter utilization at the time of disaster has conducted actively in recent years. By the way, when geographical information is given to a tweet, mapping the tweet becomes possible. Then, it becomes possible, for example, to grasp the damage situation, support needs, and so on, in more detail in the time of disaster. However, it is known that the number of geotagged tweets is extremely small. Therefore, it is necessary to extract the toponym in the tweet text accurately. However, when there is a possibility that the same proper noun may represent a place name and a person's name, there are cases where it cannot be uniquely specified. Then, in this study, we propose a word meaning disambiguation method for tweet texts that aims for information mapping and verify the usefulness of the proposed method through experiments.

Keywords: social media, Twitter, disambiguation

1. はじめに

近年，ソーシャルメディアのデータを活用して実世界の出来事の抽出や予測，例えば災害時の被害状況の把握や鉄道運行情報の抽出[1], 感染症の流行予測[2]などを行う研究が盛んである．特にTwitterは速報性が高く利用者も多いメディアであるため，リアルタイムな状況把握への活用が期待されており，近年は災害時の活用を指向した研究も活発である[3]-[6]．しかし，Twitter をはじめとするソーシャルメディアの情報は一般に情報量が膨大であり，さらに災害時における活用を考えると被災者支援や避難行動の意思決定等に必ずしも有用でないものも多く，情報の内容を精査する必要がある．また，有用な情報であったとしても地理情報が付与されていなければ活用が難しいケースも多い．例えば,「市役所前の道路が陥没しています」といったツイートではどこの市役所であるかが一意に特定できず，災害対応等に活用することができない．情報に地理情報が付与されていれば，地図上にマッピングすることが可能となるため，例えば災害時に被害状況や支援ニーズ等をより詳細に把握することなどが可能となる．Twitterには投稿にジオタグ（経緯度）を付与する機能があり，これらの情報を利用して地図上に可視化を行う事例も存在する．しかし，Chengら[7]によるとジオタグが付与されたツイートは全ツイート中 0.42%と極めて少数であるため，ツイートにジオタグが付与されていることを前提とすることは実用的ではない．そのため先行研究ではオープンデータなどを用いて作成した地名辞書を利用し，テキスト中に出現した地名とマッチングさせることによって実空間と対応付ける手法などが検討されてきた[8]．しかしこれらの手法では 2 つのタイプの地名曖昧性が問題となる[9]．1 つは Geo/Non-geo 曖昧性と呼ばれ,「松本」や「山口」など地名と同一表記の人名の存在によって生じる曖昧性である．もう 1 つはGeo/Geo曖昧性と呼ばれ,「日本橋」など表記が同一の地名が複数存

†1 東海大学
Tokai University
†2 データセクション株式会社
Datasection Inc.

ⓒ 2017 Information Processing Society of Japan

在する（東京の日本橋，大阪の日本橋）ことにより生じる曖昧性である．

本研究では特に Twitter のテキスト情報を対象に，地図上への情報マッピングを目的として，Geo/Non-geo 曖昧性の解消を行う．具体的には，Twitter に投稿されたテキストから地名候補となる文字列の前後に出現する単語情報を利用することによる Geo/Non-geo 曖昧性の解消法を提案し，実験によりその有用性を検証する．

2. 関連研究

2.1 地名曖昧性

地名と実空間の座標を対応付ける際に発生する曖昧性は，Geo/Non-geo 曖昧性と Geo/Geo 曖昧性の 2 種類が存在する[9]．本研究で扱う Geo/Non-geo 曖昧性とは地名と同一の表記で地名以外の意味を持つ単語に生じる曖昧性である．例えば，「松本」や「山口」という表記は地名としても人名としても用いられる．また，「松本」という表記には例えば「松本山雅」というサッカークラブを指している場合もあり，このような場合これを地名として扱うことは必ずしも適切ではない．このようなテキスト中の地名表記が含む語義の曖昧性は，周辺の文脈などにより解消する必要がある．一方，Geo/Geo 曖昧性とは「青葉区」や「日本橋」といった同一表記の地名が複数個所存在する場合の曖昧性であり，テキスト中に共起する地名を利用した曖昧性解消[8]などが提案されている．

2.2 固有表現抽出

Geo/Non-geo 曖昧性の解消のための単語の語義推定に関連する研究として，山田ら[10]の固有表現抽出に関する研究が挙げられる．山田らは周辺の単語，品詞，文字種など様々な特徴を素性とし， SVM を用いることで高い精度での固有表現抽出に成功した．しかし，この研究で扱ったデータは毎日新聞の記事データであり，文法的に整った文章である．一方，本研究の対象となる Twitter の文章はくだけた表現や省略形，未知語が多いため，同様の手法が適用できないことが予想される．

3. 提案手法

本研究では山田ら[10]の手法を参考に曖昧性解消を行う必要がある単語の前後の単語の情報を素性として特徴ベクトルを生成し，機械学習による学習を行う．

3.1. 形態素解析の適用

曖昧性解消の前処理として，対象となる曖昧性を含む文字列を含むツイートに形態素解析を適用し，前後に出現する単語の情報を取得する．本研究では形態素解析器には MeCab[11]を利用し，辞書としては mecab-ipadic-NEologd [12]を用いた．

3.2. 特徴ベクトルの作成

3.2.1 形態素ベクトルの生成

地名候補文字列は地名の場合とそうでない場合では前後の形態素や品詞などが異なると考えられる．例えば，地名には「駅」や「市」といった形態素が直後に出現しやすく，人名には「さん」や「先生」といった形態素が出現しやすい．そこで本研究では，特徴ベクトルとして曖昧性解消対象の形態素の前後 2 形態素を対象とし，素性の値は形態素が出現したか否かの 2 値とした．

3.2.2 品詞ベクトルの生成

地名前後 2 形態素の品詞情報を加える．例えば「○○／に／行く」といった場合，直後に助詞と動詞が出現したという情報を特徴ベクトルに加える．品詞は細分類（名詞・一般や名詞・固有名詞など）を利用する．

3.2.3 文字種ベクトルの生成

地名前後の文字種を特徴ベクトルに加える．例えば「渋谷ヒカリエ」や「横浜スタジアム」などの施設名は直後にカタカナが出現しやすいという特徴を考慮する．文字種はひらがな，カタカナ，漢字，ローマ字，数字，記号，その他，のいずれであるかを特徴とする．文字種が複数存在する場合，出現したすべての文字種を素性とする．

4. 予備実験

形態素解析は文章を最小単位である形態素に分割し，品詞を判別する．地域名を表す場合「名詞，固有名詞，地域，一般」といった形で判別される．事前の調査として，ツイート文に形態素解析を適用させた際，結果がどの程度信用できるか調査を行った．今回は対象とする文字列として Geo/Non-geo 曖昧性を持つ 4 か所（松本，宮崎，山口，石川）を含むツイートを 100 件抽出し形態素解析を適用させ，該当する形態素の判別が正しいかどうか手動で判定した．形態素解析器には MeCab[11]を利用し，辞書としては mecab-ipadic-NEologd[12]を用いた．判定精度は 86%と高かったが，形態素に分割した際に「松本城」や「松本潤」といった抽出に用いた文字列とは異なる曖昧性の含まない形態素が抽出された．これらのツイートは曖昧性を含まない形態素のため語義の推定を行う必要はないと考えられる．このような曖昧性の含まない形態素を含むツイートの割合はおよそ半分を占めており，これらを対象外とし，抽出を行った文字列と同一の形態素のツイート精度は 75%ほどであった．そのため本研究では曖昧性を含む地名候補文字列と同一の形態素を含むツイートを実験対象とする．

5. 評価実験

提案手法の有用性を検証するため，評価実験を行う．まず実験対象となるツイートを収集し，手動で正解ラベルを付与した．これらのツイートをデータセットとして分類機

ⓒ 2017 Information Processing Society of Japan

を生成し，有用性を検証する．

5.1 実験対象となるツイートの取得

実験対象の地域名称として「山口」「松本」「千葉」「石川」「宮崎」の 5 つを設定し，各地名候補文字列を含むツイートを実験対象とする．ツイートは 2016 年 11 月 1 日～30 日の間に投稿されたもので，データの重複や偏りを減少させるために URL(“http://”)を含むツイートと“RT”を含むツイートは実験対象から除外した．なお，これらのツイートデータはデータセクション株式会社より提供いただいた（全体の約 10%サンプリング）．

5.2 曖昧性を含むツイートの設定

予備実験より，取得したツイートに形態素解析を適用し，該当する地名と同一の形態素を含まないツイートを処理の対象から除外する．例えば「松本城は晴れている日に行きたい」という文章に形態素解析を適用させると「松本城」という形態素が抽出される．この場合,「松本城」という形態素は曖昧性を持たないため処理対象からは除外する．処理対象となるツイートを 700 件抽出し, 手動で「地名」,「地名以外」,「不明」の 3 種類のラベルを付与する．

5.3 実験方法

実験には「地名」と「地名以外」のラベルが付与されたツイートを利用する．今回は実験用ツイートの 9 割を学習データ, 残り 1 割をテストデータとする. 評価方法として，それぞれ適合率，再現率，及び適合率と再現率の調和平均である F 値を算出する．本研究では分類器に SVM を使用し，カーネル関数には多項式カーネルを用いた．

5.4 実験結果と考察

実験結果を表 1, 2 に示す．地名を表すツイートとそうでないツイート双方に対して高い精度で曖昧性が解消できることを示しており，地名曖昧性解消の有用性が検証できたと考えられる．

失敗したツイートに関しては前後 2 形態素のみでは情報が少なく，分類が困難であったと考えられる（表 2(1)）．このようなツイートを適切に分類するためには前後の形態素以外の特徴を用いる必要があると考える．また，（表 2(2)）のように前後 2 形態素からでは地名と人名のちらにも判断できてしまうため，分類に失敗してしまうと思われるツイートも存在した．このようなツイートの分類には人手での分類でも事前知識が必要となるため，自動分類は困難であると考えられる．しかし，中には明らかに地名を表しているにもかかわらず地名でないと判断されたツイートも存在した（表 2(3)）．これらは，直前の空白や記号などが大きく影響してしまったものと考えられる．また，特定の人名が地名と誤分類されたツイートも存在した（表 2(4)）．このようなツイートは形態素解析の際に用いる辞書を拡張することで解決できると考えられる．

6. まとめと今後の課題

本研究では地名候補の前後の形態素情報を用いることでツイート中に出現する地名表記の曖昧性解消を行った．実験の結果, 少ない情報でも高い精度で分類ができており，その有用性を示すことができた．

今後の課題として，曖昧性解消が必要なツイートを投稿したユーザの過去のツイートの利用や, 前後2形態素以外に含まれている地名や人名の利用などが挙げられる．また，今回は地名と人名がどちらも出現しやすい地名を対象として実験を行ったため，ツイートに出現する頻度が地名と人名で偏ってしまう名詞の場合では精度が下がってしまうことが考えられる．さらに，情報マッピングの主たる応用先として災害時におけるクライシスマッピングがあるが，特に大規模災害時では投稿文に含まれる情報が平時とは異なることが想定される．そこで，災害時に投稿されたツイートに対しても提案手法が有用であるか，今後検証する予定である．

表 1 実験結果

	適合率	再現率	F 値
地名	0.89	0.87	0.88
地名以外	0.87	0.89	0.88
全体	0.88	0.88	0.88

表 2 分類に失敗したツイート例

(1)	@username 大阪か！ い、石川……()
(2)	山口出ていくならオリ行け、結果残せば金払いはいいし横浜と対戦することも無いから絶対にヤジられないぞ
(3)	@username 松本観光したいけどなかなかできてなくて…
(4)	宮崎秋人がかわいすぎるんで とりあえず発表終わったらインスタ登録します

参考文献

[1] 長野伸一, 上野晃嗣, 長健太, “ソーシャルセンサからの鉄道運行情報検出システムの開発”, 電子情報通信学会論文誌 D, Vol.J96-D, No.10, pp.2262-2273, 2013.

[2] 荒牧英治, 増川佐知子, 森田瑞樹, “文章分類と疾患モデルの融合によるソーシャルメディアからの感染症把握”, 自然言語処理, Vol.19, No.5, pp419-435, 2012.

[3] H. Wilensky, “Twitter as a Navigator for Stranded Commuters during the Great East Japan Earthquake”, Proc. 11th International Conference on Information Systems for Crisis Response and Management, pp.695-704, 2014.

[4] O. Uchida, T. Rokuse, M. Tomita, Y. Kajita, Y. Yamamoto, F. Toriumi, B. Semaan, S. Robertson, M. Miller, “Classification and

Mapping of Disaster Relevant Tweets for Providing Useful Information for Victims During Disasters", IIEEJ Transactions on Image Electronics and Visual Computing, Vol.3, No.2, pp.224-232, 2015.

[5] R. Kitajima, R. Kamimura, O. Uchida, F. Toriumi, "Potential Information Maximization: Potentiality-Driven Information Maximization and Its Application to Tweets Classification and Interpretation", International Journal of Computer Information Systems and Industrial Management Applications, Vol.8, pp.42-51, 2016.

[6] R. Kitajima, R. Kamimura, O. Uchida, F. Toriumi, "Identifying Important Tweets by Considering the Potentiality of Neurons", IEICE Trans. Fundamentals, Vol.E99-A, No.8, pp.1555-1559, 2016.

[7] C. Zhiyuan, J. Caverlee, K. Lee, "You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users", Proc. of the 19th ACM international conference on Information and knowledge management, 2010.

[8] 六瀬聡宏, 内田理, 鳥海不二夫, "大規模災害時の情報提供を目的とした地名の曖昧性解消", 言語処理学会第 21 回年次大会, 2015.

[9] 落合桂一, 鳥居大祐, "時間変化する特徴語によるマイクロブログ地名曖昧性解消", 情報処理学会論文誌, データベース, Vol.7, No.2, pp.51-60, 2014.

[10] 山田寛康, 工藤拓, 松本裕治, "Support Vector Machine を用いた日本語固有表現抽出", 情報処理学会論文誌, Vol.43, No.1, pp.44-53, 2002.

[11] MeCab, http://mecab.googlecode.com/

[12] mecab-ipadic-NEologd, https://github.com/neologd/mecab-ipadic-neologd

ⓒ 2017 Information Processing Society of Japan