

教師あり機械学習を用いたツイート投稿時の ユーザ位置推定手法

杉谷 卓哉^{1,a)} 白川 真澄¹ 原 隆浩¹ 西尾 章治郎¹

概要: 本研究では、位置情報の付いていないツイートに対し、ツイート中に出現するスポット名を用いて投稿時のユーザの位置を推測する手法を提案する。提案手法では、スポット名を含む位置情報付きツイートから学習した特徴量を用いて、実際にユーザがそのスポットからツイートを投稿したか否かを SVM により判定する。また、位置情報サービス Foursquare から抽出したスポット名と緯度経度情報をもとに CRF を用いてツイート中のスポット名を検出し、提案手法によるツイートの位置情報付与を行うシステムを実装した。2012 年 5 月に投稿された位置情報付きツイートをテストセットとして提案手法の評価を行った結果、適合率 80%において再現率約 43%、また最大で適合率 92%を達成した。

1. はじめに

Twitter は、ツイートと呼ばれる 140 文字までの短いメッセージの投稿と、そのメッセージの共有を行うためのソーシャル・ネットワーク・サービスである。Twitter は、投稿が手軽であるため、ここ数年でユーザ数が急増しており、5 億人を超える (2012 年 7 月 30 日現在^{*1}) アカウントを持ち、全世界のユーザによる 1 日当たりのツイート数は 5 億に達している (2012 年 10 月 28 日現在^{*2})。この大量のツイートはリアルタイム性の高い情報を大量に含んでいるため、Twitter をマイニングの対象とした研究が多数行われている。また、Twitter では、携帯電話やスマートフォンを Twitter と連動させることにより、現在の位置情報 (緯度、経度) をツイートに付与して投稿できる。この位置情報付きツイートを利用して、地震や祭りなどのローカルイベントの検出 [8], [9], [10], [11] やユーザの行動予測 [7]、ニュースや話題に対する地域や国ごとの意見抽出 [6]、各地域のホットピック抽出 [4] など多数の研究が行われている。

これらの研究は位置情報付きツイートを用いてツイートと地理情報のマッピングを行っている。しかし、全ツイートに対して、位置情報付きツイートは少量しかなく、全世界

のツイートに対する位置情報付きツイートの割合は 0.77% (2012 年 7 月 30 日現在^{*1}) となっている。日本の場合、位置情報付きツイートの割合は、さらに少ない傾向にあり、橋本らの調査 [5] によると、2011 年 7 月 15 日午前 0 時から 2012 年 4 月 1 日午前 0 時までの Twitter Streaming API の sample メソッドによって取得できた日本語ツイートのうち、位置情報付きツイートの割合は約 0.18%であった。位置情報付きツイートを利用した研究では、このリソースの少なさが問題となる。

位置情報付きツイートの不足に対処するため、位置情報が付いてないツイート^{*3}に対して、そのツイートに関連した緯度経度情報を付与する研究 [2], [3], [9], [10], [11] が行われている。これらの研究では、地名を含むツイートに対してその地名の緯度経度情報を付与するという手法が用いられている。しかし、地名と同じ文字列がたまたまツイートに含まれていた場合や、地名がニュースや話題として出現している場合など、投稿時のユーザの位置情報と異なる緯度経度情報を付与している場合が存在する。ニュースや話題として地名が出現している場合、ユーザ自体がその場所になくても、その場所に関連した内容のツイートならば緯度経度情報を付与しても問題がない場合も多い。しかし、リアルタイムなユーザの位置を利用するアプリケーションでは、ツイート投稿時のユーザの位置情報と付与した緯度経度情報が同じである必要がある。

そこで本研究では、地名を含むツイートに対し、教師あり機械学習手法である SVM (Support Vector Machine) を

¹ 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

a) sugitani.takuya@ist.osaka-u.ac.jp

^{*1} http://semioast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

^{*2} http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/

^{*3} 本研究では以後、位置情報が付いていると明記していない限り、ツイートに位置情報は付いていないものとする。

利用して、その地名の緯度経度情報からツイートが投稿されたか否かを判定する手法を提案する。提案手法では、学習において、位置情報付きツイートを利用して、機械的に正解、不正解のラベルを付与する。そのため、人手を用いることなく教師データを大量に作成できる。大量の教師データを用いて学習を行うことで、高い精度で付与した緯度経度情報から投稿されたか否かを判別できると考えられる。また、提案手法による位置情報付与のシステムを実装した。このシステムでは、1つのツイートを入力とし、位置情報サービス Foursquare から抽出したスポット名と緯度経度情報をもとに、教師あり機械学習である CRF (Conditional Random Field) を用いてツイート中のスポット名を検出した後、提案手法によるツイートの位置情報付与 (判定) を行う。

以下、第2章で、ツイートに対する位置情報付与に関する先行研究について述べる。第3章で、提案手法について詳述し、第4章で、提案手法のシステム実装について述べる。第5章で、提案手法の評価及び考察を行う、第6章で本研究のまとめと今後の課題について述べる。

2. 関連研究

ツイートの位置情報を付与する研究はこれまでにいくつか行われている。渡辺ら [10], [11] の研究や榊ら [9] の研究では、位置情報に関連したイベント検出の一環として、ツイートに出現したスポット名から、そのスポットの緯度経度情報をツイートに結び付ける手法を提案している。渡辺らの研究では、位置情報サービス Foursquare^{*4} を利用し、スポット名と緯度経度情報の組合せからスポット情報データベースを構築する。次に、スポット情報データベースに登録してある一つのスポット名に対して複数の緯度経度情報が登録してある場合、地理的分散が大きければ、スポット情報データベースから除去する。そして、スポット情報データベースからスポット名の辞書を作成し、ツイートにスポット名が出現していた場合、そのスポット名の緯度経度情報をツイートに付与する。榊らの研究では、ツイートに対して、形態素解析を行った際の品詞情報やパターンマッチングから、スポット名を検出する。そして、Google Maps API^{*5} を用いて、スポット名を緯度経度情報に変換し、ツイートに付与する。これらの研究では、ツイートに結び付けた緯度経度情報がツイート投稿時のユーザの位置情報かどうかの判定を行っていないため、ツイート投稿時のユーザの位置情報を利用するようなアプリケーションでは問題となる。本研究では、この問題に対応するため、スポット名が地名表現として使われているかどうかの判定や、付与した緯度経度情報が投稿時のユーザの位置情報かどうかの判定を、教師あり機械学習により実現している。

^{*4} <https://ja.Foursquare.com/>

^{*5} <https://developers.google.com/maps/>

Cheng ら [1] は、ユーザが過去に投稿したツイートの内容を分析することで、ユーザの居住地を都市レベルで推定している。Sadilek らの研究 [7] では、友人ユーザの位置情報付きツイートと対象ユーザの位置情報を学習することで、各ユーザの行動モデルを予測し、また、ユーザの位置情報を推定している。本研究では、ユーザ単位の位置情報付与ではなく、ツイート単位の位置情報付与を目的としている点でこれらの研究とは異なる。

伊川ら [2], [3] の研究では、ツイートのテキスト情報を用いて、ツイート投稿時のユーザの位置情報を推定している。具体的には、位置情報サービスから投稿されたツイートに対して、ツイート投稿時のユーザの位置情報を特定し、そのツイートをもとに時間的に近接して投稿された同一ユーザによるツイートのテキストの表現と位置情報との関係性を教師あり機械学習で学習する。そして、学習した特徴量を用いて、ツイートのテキスト表現から、位置情報を付与する。この手法で提案されている、時間的に近接したツイートを利用する方法は、提案手法と組み合わせて使うことが可能である。

3. 提案手法

本研究では、地名を含むツイートに対し、その地名の緯度経度情報がツイート投稿時のユーザの位置情報かどうかを、SVM を用いて判定する手法を提案する。提案手法の大まかな流れを図1に示す。まず、学習フェーズとして、位置情報付きツイートに対してジオコーディング (地名を用いた緯度経度情報の付与) を行う。付与した緯度経度情報とツイートの実際の位置情報との誤差から、付与した緯度経度情報が投稿時のユーザの位置情報を指しているかを判定し、正解、不正解のラベル付けを行い、SVM の教師データとする。判定フェーズでは、学習フェーズで学習したツイートの特徴から、ツイートに付与された緯度経度情報が投稿時のユーザの位置情報かどうかを判定する。

学習フェーズでは、ジオコーディングによって付与された緯度経度情報が投稿時のユーザの位置情報として正しいかどうか判定するために学習を行う。提案手法で使用するジオコーディングは、ツイートにスポット名が含まれている場合に、そのスポットに紐づけられた緯度経度情報を付与する。学習する素性は、ツイートの文字数や URL、メンション、リツイートの有無、URL のタイプ、スポット名の文字数、スポットのカテゴリ、スポット名周辺の語や品詞など付与した緯度経度のスポットに対するものなど様々なものが考えられ、より多くの素性を利用すればそれだけ精度向上できる可能性が高くなる (なお、本研究で具体的に使用した素性については次章のシステム実装の 4.2 節で述べる)。一方で、素性が多数ある場合、教師データを大量に用意できなければ十分な学習が行えない。

そこで本研究では、位置情報付きツイートに対してジオ

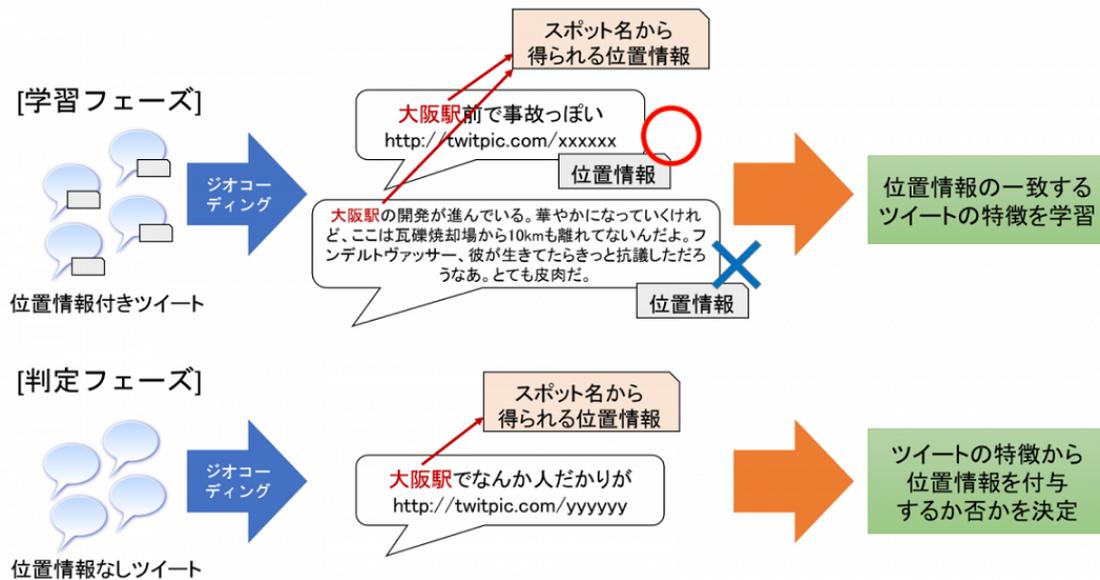


図 1 提案手法の流れ

コーディングすることによって、付与した緯度経度情報と投稿時のユーザの位置情報との誤差から、正解、不正解のラベル付けを機械的に行う。これにより、人手によるラベル付けを行う必要がなく、大量の教師データを用意できる。提案手法では、教師データを位置情報付きツイートから構築することになるため、位置情報が付いていないツイートとは異なる特徴を学習して、判定の精度が下がることが考えられる。そこで、位置情報付きツイートに特有の位置情報サービスやそれと連動したサービスから投稿されたツイートを除去することで、教師データとする位置情報付きツイートと一般的なツイートの特徴の差異を減らして学習を行う。

正解、不正解のラベル付けに用いる距離の誤差を決定するため、事前実験として、2011年10月25日から2012年4月30日までに投稿された日本全域の位置情報付きツイートに対して、ジオコーディングを行った。そして、投稿時のユーザの位置を示すスポット名が明示された、「<スポット名>なう」という定型文のツイートから、位置情報がスポット名の指す緯度経度の場所であると人手によって判断したツイート（正解ツイート）を1,000件抽出した。これらのツイートに対して、スポット名から付与した緯度経度情報とツイートの位置情報の実空間上の距離の誤差を求め、誤差ごとの正解ツイート数分布を調査した。図2は誤差ごとの正解ツイート数の分布である。この分布から、正解ツイートの大部分が1km以内の誤差に含まれていることが分かる。また、誤差が数km以内にもある程度の割合で正解ツイートが含まれているが、誤差が10km以上の場合、含まれる正解ツイートは1%未満となることが分かる。この結果より、提案手法では、ツイートの位置情報とスポット名の緯度経度情報との実空間上の距離が1km以

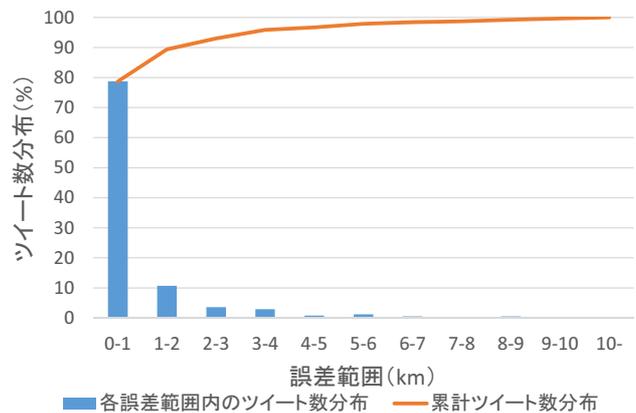


図 2 正解ツイートにおける誤差ごとのツイート数分布

下の場合は正解、10km以上の場合には不正解として、ラベル付けを行い、教師データを作成する。それ以外の場合は正解・不正解があいまいであるとみなし、その位置情報付きツイートを教師データとして利用しない。作成した教師データを用いてツイートの特徴をSVMで学習する。

判定フェーズでは、学習フェーズで学習したSVMによる判定器を利用して、ジオコーディングよりツイートに付与した緯度経度情報が投稿時のユーザの位置情報かどうかを判定する。付与した緯度経度情報が投稿時のユーザの位置情報であると判定された場合にのみ、そのツイートに位置情報を付与する。これにより、位置情報付きツイートの総量を増やすことができる。

4. システム実装

3章で提案した手法を用いて、地名を含むツイートの位置情報を推定するシステムを実装した。システムは大きく、地名を含むツイートに対して緯度経度情報の付与を行うジ

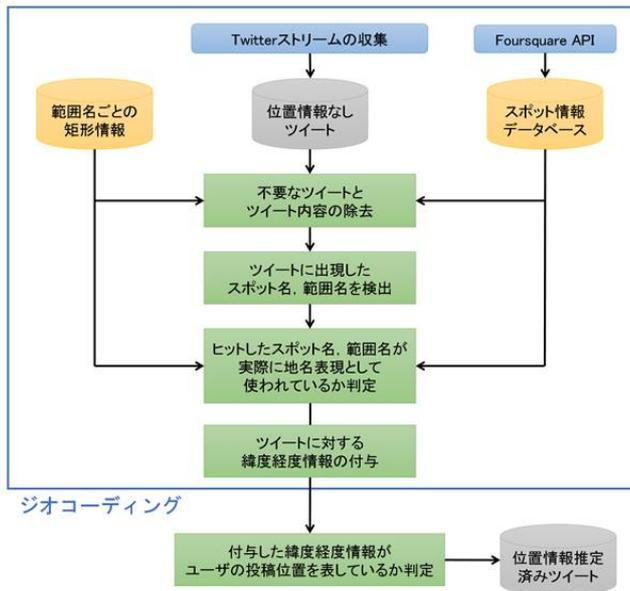


図 3 実装したシステムの流れ

ジオコーディングと、付与した緯度経度情報が投稿時のユーザの位置情報かどうかを判定する部分に分かれる。実装したシステムの大まかな流れを図 3 に示す。まず、位置情報サービスである Foursquare の API からスポット情報を抽出し、スポット情報データベースを作成した。また、都道府県名や市町村区名などの範囲を指す地名を範囲名と定義し、範囲名とその矩形情報の組合せを範囲情報データベースとして構築した。そして、Twitter の Streaming API を用いて取得したツイートから、データベースに登録したスポット名、範囲名を地名候補として検出する。その後、検出した地名候補が実際に地名表現として使われているかどうかを CRF を用いて判定する。スポット名が地名表現として用いられていると判定された場合、そのスポット名に該当する緯度経度情報を付与する。このとき、地名として判定された範囲名が同ツイートに存在する場合、範囲名に該当する矩形内のスポットに絞り込む。最後に、付与した緯度経度情報が実際の投稿時のユーザの位置情報と同じかどうかの判定を SVM を用いて行う。以下の節では、実装したシステムの各処理について詳述する。

4.1 ジオコーディング

ジオコーディングは大きく分けて、携帯端末以外から投稿されたツイートを除去するノイズ除去、スポット名と範囲名の候補を検出する地名候補検出、検出した地名候補が実際にテキスト内で地名表現として使われているかを CRF により判定する地名判定、地名表現であると判定されたスポット名と範囲名からツイートにスポットの緯度経度情報を付与する緯度経度情報付与の 4 つの処理から構成される。

4.1.1 ノイズ除去

本システムでは、ユーザがあるスポットにおいて、その

表 1 スポット情報一覧

スポット情報	説明
スポット名	スポットの名前
カテゴリ	スポットのカテゴリ
チェックイン数	スポットにチェックインした回数
チェックインユーザ数	スポットにチェックインしたユーザ数
公式による認証	公式により認証されたスポットかどうか
緯度	スポットの緯度
経度	スポットの経度

スポット名を含むツイートを携帯端末から投稿する必要が状況を想定している。そこで、携帯端末以外から投稿されたツイートをノイズとして除去する。ツイートのクライアント情報を利用することで、ツイートが携帯端末からのものであるかどうかを判別した。2011 年 10 月 25 日から 2012 年 4 月末までに投稿された位置情報付きツイートにおいて、投稿ツイート数の多いクライアント上位 200 件から、携帯端末からの手入力によるツイート投稿のみと考えられるクライアント 61 件を抽出し、利用するツイートはそのクライアントによって投稿されたものだけに絞り込んだ。リツイートの引用部分に出現するスポット名は投稿時のユーザの位置情報とは無関係であると考えられる。そこで、リツイートの引用部分の削除を行う。一般的なリツイートでは、RT あるいは QT に続いて引用元のユーザと引用文が記述され、コメントがある場合は RT、QT の前に記述される。そのため、リツイートの引用部分の判別は、RT、QT の出現により判断し、引用部分のみを除去する。そして、URL、メンション (ユーザに対する返信などに使う「@ユーザ ID」) といった地名が含まれない部分も除去を行う。

4.1.2 地名候補検出

地名候補を検出するため、あらかじめスポット名と範囲名のデータベースを構築した。まず、Foursquare の API を利用して、2012 年 8 月時点において Foursquare に登録してある日本全域の緯度経度情報をもつスポット情報 1,199,667 件を抽出し、スポット情報データベースを構築した。スポット情報とは表 1 に示すものである。チェックインとは、Foursquare を利用するユーザが、スポットの緯度経度情報が指す地点に赴き、携帯端末の GPS 機能から自分がそのスポットにいることを認証する操作のことを指す。カテゴリは、Foursquare で登録されているスポットのカテゴリで、arts entertainment, building, education, food, nightlife, parks outdoors, shops, travel, non category の 9 種類である。Foursquare の情報に加えて、国土交通省国土地理院の位置参照情報ダウンロードサービス*6から、範囲が狭い街区レベルの地名をスポットとしてスポット情報データベースに登録した。この際、スポットのカテゴリは town とし、チェックイン数、チェックインユーザ数は 0、公式による認証はされていないとした。また、スポット情

*6 http://nlftp.mlit.go.jp/cgi-bin/isj/dls/_choose_method.cgi

報データベースとは別に、国土交通省国土地理院の Web ページ*7から、範囲名とその東西南北端点の緯度経度情報の組合せを抽出し、範囲情報データベースを構築した。スポット名と同じ範囲名を登録する場合、スポット情報データベースからそのスポット名のスポットの情報を除去した。

構築したスポット情報データベース、範囲情報データベースからスポット名辞書、範囲名辞書をそれぞれ構築して、ツイートに出現したスポット名、範囲名を検出する。ここで、Foursquare のスポット名では、「東京駅 (Tokyo sta.)」といったような、日本語に続いてその英訳を括弧内で表現した形式が多用されているため、Foursquare に登録されているスポット名をそのまま用いると、ツイートからスポット名を検出できないと考えられる。そこで、スポット名登録の際に、そのままスポット名を登録するだけでなく、括弧を取り除いた表記と括弧内の表記（「東京駅 (Tokyo sta.)」の場合は「東京駅」、「Tokyo sta.」）についても登録した。また、「大衆食堂 あさちゃん」、「居酒屋 雪月花」といった、店のカテゴリとその店の名前との組合せのスポット名も頻繁に出現する。この形式のスポット名もそのままではツイート上にほとんど出現しないが、店の名前だけだとヒットするケースが多い。そのため、スポット名をスペースで分割して、それぞれの表記をスポット名辞書に登録した。分割したスポット名は分割スポット名辞書として構築し、分割を行っていないユニークな Foursquare のスポット名と区別した。また、範囲名では、都道府県名の「都」、「府」、「県」の表記、市町村名の「市」、「町」、「区」、「村」の表記の部分を削除し、都道府県名についてはローマ字、ひらがな、カタカナ表記でも範囲名を辞書に登録した。これらスポット名辞書、分割スポット名辞書、範囲名辞書を用いて、ツイートから最長一致でかつ検出する地名の数が最も少なくなる地名の組合せを地名候補として抽出する。

4.1.3 地名判定

抽出した地名候補が本当に地名表現として用いられているかどうかを CRF を用いて判定する。ツイートに出現した地名候補が一つの語となるように設定して、形態素解析を行い、その語と品詞、また地名だった場合はそれが範囲名、スポット名、分割スポット名のどれかをラベル付けし素性とした。形態素解析器には、MeCab*8を用いた。教師データとして、2011 年 10 月 25 日から 2012 年 4 月 30 日までに投稿された日本全域の位置情報付きツイートで、スポット名が検出できたもののうち、スポットの緯度経度情報とツイート投稿時のユーザの位置情報との誤差が 1km 以下のツイートにおいて、頻出上位 100 件のスポット名を抽出し、スポット名を含むツイートをそれぞれ 10 件ずつ選んだ。また、誤差が 10km 以上のツイートについても同様に、頻出上位 100 件のスポット名を抽出し、各スポット

名ごとに 10 件ずつツイートを取得した。これら合わせて 2,000 件のツイートに含まれる各語について、人手により地名表現のラベル付けを行い、CRF++*9を用いて学習を行った。

4.1.4 スポット名に対する緯度経度情報付与

地名表現と判定されたスポット名、分割スポット名がツイートに出現した場合、以下の手順により該当するスポットの緯度経度情報をツイートに付与する。まず、範囲名が出現している場合、範囲名に対応する矩形内の緯度経度情報をもつスポットのみに絞り込む。範囲名が複数出現した場合、範囲名に対応する矩形のマージを行う。ある範囲名の矩形が別の範囲名の矩形に包含される場合、より狭い範囲の矩形のみでスポットの絞り込みを行い、そうでない場合は、範囲名の矩形情報すべてについてスポットの絞り込みを行う。その後、スポット名、分割スポット名ごとに該当するスポットのマージを行う。スポット間の実空間上の距離が 2km 以下なら同一のスポットであるとし、カテゴリが town のもの、なければ最もチェックインユーザ数の多いスポットに併合する。そして、スポット名、分割スポット名に対して、スポットが一意に定まる場合のみツイートの緯度経度情報を付与する。これにより、チェーン店の名前など、複数のスポットが該当する場合に位置情報付与が行われないようにする。

スポット名、分割スポット名が複数ある場合は、それぞれのスポット名ごとに範囲名によるスポットの絞り込み、マージを行う。その後、異なるスポット名のスポットと実空間上の距離が 2km 以内ならば、それらのスポットのみに絞り込み、それぞれのスポットで緯度経度情報をツイートに付与し、そうでなければ、スポットが一意に定まるスポット名ごとにツイートに対して位置情報を付与する。ただし、分割スポット名の場合は、範囲名の絞り込みが行われた場合と、スポット名、分割スポット名が複数出現し、他のスポット名、分割スポット名のスポットと実空間上の距離が 2km となる場合のみ位置情報付与を行う。

4.2 投稿時のユーザの位置情報かどうかの判定

SVM を利用して、ジオコーディングでツイートに付与した緯度経度情報が、ツイート投稿時のユーザの位置情報かどうかの判定を行う。2011 年 10 月 25 日から 2012 年 4 月 30 日までに投稿された日本全域の位置情報付きツイートに対して、付与した緯度経度情報とツイートの位置情報との実空間上の距離が 1km 以内のツイートを正解、10km 以上のツイートを不正解とし、それぞれ 50,000 件で学習を行う。

SVM によって学習する素性は表 2 に挙げるものとした。学習はツイート単位ではなく緯度経度情報を付与したス

*7 <http://www.gsi.go.jp/KOKUJYOHO/center.htm>

*8 <https://code.google.com/p/mecab/>

*9 <https://code.google.com/p/crfpp/>

表 2 素性一覧

対象	素性
ツイート	ツイートの文字数
	ツイートの文字数 (URL, メンション, リツイートの引用部分を除去)
	ツイートの文字構成
	ツイートに対してジオコーディングできたスポット数
	非公式リツイートの有無
	ハッシュタグ出現回数
	メンションの出現回数
	URL の出現回数
	URL のタイプ
	スポット名の出現回数
	投稿日時の時間帯
	投稿日時の曜日
	投稿日時が平日か休日か
スポット	スポット名の文字数
	スポット名の文字構成
	分割スポット名かどうか
	スポットのカテゴリ
	チェックイン数
	チェックインユーザ数
	スポットが公式から認証されているか
	スポットの緯度
	スポットの経度
範囲	範囲名の文字数
	都道府県か市町村区村か
語句	最初から何語句目
	最後から何語句目
	スポット名から何語句前, 後
	語句の文字数
	語句の文字構成
	品詞
	もしスポット名なら分割スポット名かどうか
	もし範囲名なら都道府県か市町村区村か
自立語	語句自体

ポット単位とし、スポットごとに、特徴量を学習する。自立語は、2011年10月25日から2012年4月30日までに投稿された日本全域の位置情報付きツイートの中で、ジオコーディングできたものを対象とし、出現回数が1,000回以上で、かつ品詞が名詞、動詞、形容詞の語を採用した。ツイート、スポット名、語句の文字構成とは、ツイート、スポット名、語句のテキストが何文字のひらがな、カタカナ、漢字、数字、記号、その他からなっているかを表したものである。URLのタイプは、2011年10月25日から2012年4月30日までに投稿された日本全域の位置情報付きツイートにおいて、出現頻度の高いURL 250件をアプリケーション連動、ブログ、エンターテイメント、情報(Wikipediaなど)、ライフログ、位置情報サービス、動画共有、ニュース、写真共有、画像共有、ショッピング、URL短縮、SNS、店舗情報、該当なしの15種類に分類したものである。

位置情報付きツイートから自動的に教師データを作成し

た後、SVMのツールであるlibsvm^{*10}を用いて学習を行った。RBFカーネルによる学習を行い、正解、不正解の推定確率を算出するようにした。ジオコーディングの際に、一つのツイートに対して、複数のスポットの緯度経度情報が付与される場合があるが、それぞれのスポットごとに判定を行い、推定確率を求め、最も正解となる確率の高いスポットに対する判定をそのツイートに対する判定結果とする。

5. 評価実験

実装したシステムのうち、システムの地名判定とジオコーディング全体について評価を行い、システムの精度を検証した。そして、提案手法であるツイート投稿時のユーザの位置情報推定の有効性を検証するため、位置情報付きツイートをを用いて評価実験を行った。

5.1 地名判定

CRFを用いた地名判定について評価を行った。地名判定の教師データに含まれていないスポット名をランダムに1,000件選出し、2011年10月25日から2012年4月30日に投稿された日本全域の位置情報付きツイートのうち、ジオコーディングにより位置情報付与できたものから、これらのスポット名を含むツイートをそれぞれ1件ずつ抽出し、人手により地名表現のラベルを付与した。

性能評価の結果は、表3となった。適合率は、0.89%となり、高い精度で地名表現かどうかを判定できている。一方、再現率は0.78%と若干低くなっている。これは、教師データの少なさから、地名表現でない場合の特徴はある程度学習できたが、地名表現である特徴を学習しきれずに、地名表現を取り逃がしてしまっているためである。

5.2 ジオコーディング

2012年5月1日から2012年5月31日までに投稿された日本全域の位置情報付きツイート1,871,881件に対して、ジオコーディングを適用した。その際に、ノイズ除去でのクライアントによるツイートの絞り込みで832,298件(44.46%)となり、緯度経度情報を付与できたツイートは36,689件(1.96%)となった。なお、一つのツイートに対して複数のスポットの緯度経度情報が付与できる場合があるため、位置情報を付与した回数は45,367件(2.42%)となった。これら45,367件の緯度経度情報に対する実際のツイート投稿時のユーザの位置情報との距離分布は図4のようになった。付与した緯度経度情報とユーザ投稿時の位置情報との誤差が1km以内の場合に位置情報を正しく推定できたと判断すると、ツイート投稿時のユーザの位置情報を正しく付与可能なツイートは45,367件の約27%、すな

*10 <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

表 3 地名判定の評価結果

地名と判定したうち実際に地名だった件数	地名表現と判定した件数 (適合率)	テストセットに含まれる地名表現の総件数 (再現率)
1,043	1,175 (0.89)	1,338 (0.78)

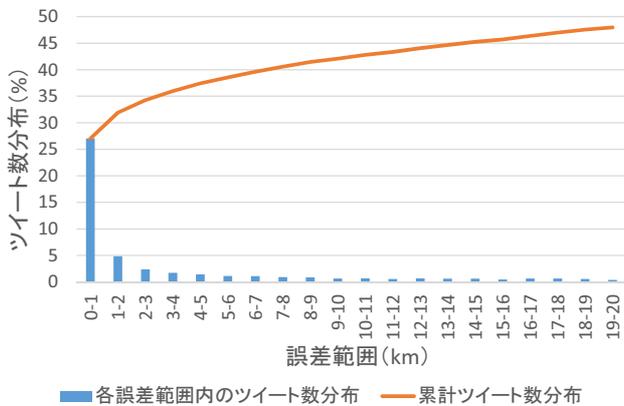


図 4 投稿時のユーザの位置情報と付与した緯度経度情報との誤差ごとのツイート数分布

わち 12,270 件となった。これは、全位置情報付きツイートの 0.66% に相当し、ツイート全体における日本語の位置情報付きツイートの割合が 0.18% であることを考慮すると、位置情報付きツイートを最大で 3 倍以上増やせる試算となる。なお、この試算は位置情報付きツイートにおいて算出したものであり、位置情報の付いていないツイートにおいて同様の割合で位置情報を付与できるかどうかについては、今後調査する予定である。

5.3 投稿時のユーザの位置情報かどうかの判定

2012 年 5 月 1 日から 2012 年 5 月 31 日までに投稿された日本全域の位置情報付きツイートに対して、前節のジオコーディングにより緯度経度情報を付与できた 45,367 件 (36,689 件のツイート) から、実際の位置情報との誤差が 1km 以内のもの 12,270 件を正解、10km 以上のも 26,268 件を不正解としてテストセットを作成し、提案手法の評価を行った。ベースライン手法として、Twitter 上での今現在の自身の場所を表す表現として用いられる「<スポット名>なう」、「~@<スポット名>」という形式のツイートを利用する手法を用いた。ジオコーディングされたスポット名が、このいずれかのパターンで出現していた場合、ユーザがそのスポットから投稿していたものと判定する。また、提案手法の投稿時のユーザの位置情報かどうかの判定での学習フェーズにおいて、提案手法の素性から自立語の素性を除外し、学習を行ったものについても評価した。

図 5 は、正解ツイートに対する再現率を横軸、適合率を縦軸にとったときの提案手法とベースライン手法の評価結果である。提案手法の各点は、libsvm の確率推定の機能により、正解に分類される確率を求め、その確率に対して閾値を 0.5 から 0.05 ずつ増加させた時の評価結果である。

提案手法は、全体的に精度が高く、適合率 80% においても、43% 程度の再現率を達成できている。これは、全位置情報付きツイート 1,871,881 件に対して 0.28% 付与できるということであり、全ツイートに対する位置情報付きツイートの割合 0.18% より高く、位置情報が付いていないツイートに対しても同じ割合で位置情報を推定できるならば、リソース量を 2.5 倍以上に増やすことができる。また、適合率を重視した場合、再現率は 12% と低くなるが、最大で 92% 以上の適合率を達成できた。

自立語の素性の有無について比較すると、性能がほとんど変わらず、自立語が素性として有効でないことが分かった。libsvm の学習したモデルのベクトルの重みから、素性として有効であったと考えられるものは順に、スポット名の文字構成、スポット名の文字数、スポットの緯度経度、メンションの出現回数、スポットが分割スポット名かどうか、スポットのカテゴリ、スポットのチェックインユーザ数、範囲が都道府県か市町区村か、語句の品詞、スポットが公式から認証されているか、ツイートの文字構成、範囲名の文字数、ツイートの文字数 (URL、メンション、リツイートの引用部分を除去)、語句の文字数、URL のタイプ、ツイートの文字数、投稿日時が平日か休日かであった。

ベースライン手法では、適合率 0.62%、再現率 0.04% となり、提案手法と比べると適合率、再現率ともに低くなっている。再現率が非常に低くなっていることから、投稿時の所在地をツイートに含める表現の中では、「<スポット名>なう」、「~@<スポット名>」といった形式は割合としては少ないことが分かる。提案手法と比較して、適合率が低くなる原因は、スポット情報データベース構築に利用した Foursquare のスポット情報のノイズの多さであると考えられる。例えば、「みぞれなう」というツイートがあった場合、スポット情報データベースにはみぞれというスポット名が登録されているため、ベースライン手法ではそのまま位置情報を付与する。一方、提案手法では、スポットの素性を学習しているため、スポットの文字構成、文字数、緯度経度、分割スポット名かどうか、チェックインユーザ数などからスポットをノイズと判定し、位置情報を付与しない。

一方で、スポット情報データベースのノイズの多さから、提案手法でも位置情報の推定に失敗するが多かった。具体的には、「目の前」、「リアル」といった一般語句でありながら、スポットとして登録されており、スポットの文字構成、文字数、緯度経度、分割スポット名かどうか、チェックインユーザ数などから判別ができない場合が多数を占める。また、「武道館」の緯度経度の指す場所が日本武道館で

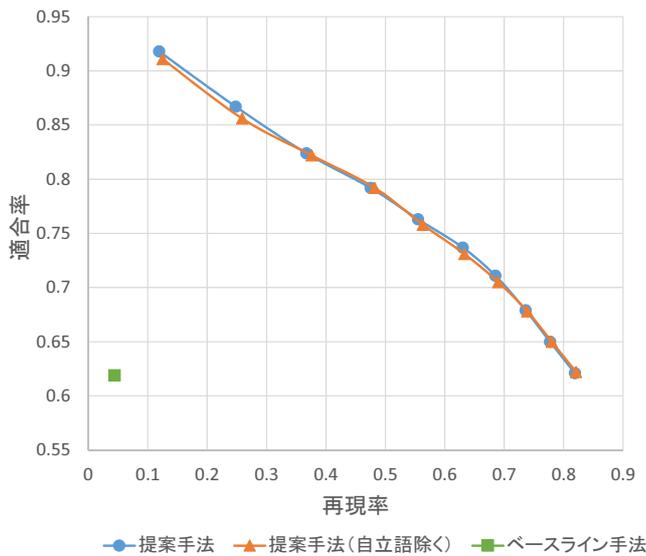


図 5 評価結果

はないといったようなスポット名から推測される場所と異なる場所にスポットの緯度経度が登録されている場合が挙げられる。さらに、「関西」、「新東名(高速)」といった範囲名や「新幹線」などの乗り物の場合も同様に、登録された緯度経度と実際の位置情報が大きく異なっていた。

6. まとめ

本研究では、位置情報が付いていないツイートに対して、ツイート投稿時のユーザの位置情報を推定することを目的として、ジオコーディングにより、付与した緯度経度情報が投稿時のユーザの位置情報かどうかを判定する手法を提案した。また、提案した手法を用いて、地名を含むツイートの位置情報を推定するシステムを実装した。

評価実験の結果、適合率 80%において再現率が約 43%、また、最大で適合率 92%を達成できることを確認した。「<スポット名>なう」、「~@<スポット名>」といった定型文を用いるベースライン手法と比較して、提案手法は非常に高い性能を達成できていた。これは、スポット情報データベースの構築の際に利用した Foursquare の情報にノイズが含まれており、提案手法では、スポットの素性を学習しているために、ノイズによる影響を防いでいることが分かった。しかし、実際に取得した失敗例から、提案手法においても Foursquare のノイズの影響による誤判定が多く起こっていることが分かった。

今後は、提案手法により学習した位置情報付きツイートの特徴が、位置情報の付いていないツイートにおいてどの程度利用できるかについて調査を行い、提案手法の妥当性を確認する予定である。また、ジオコーディングにおけるミスによる精度低下を減らすため、スポット情報データベースの整理や既存のジオコーディング手法の利用を検討する。

参考文献

- [1] Z. Cheng, J. Caverlee and K. Lee, "You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users," Proc. of ACM Conference on Information and Knowledge Management, pp.759-768, Oct. 2010.
- [2] 伊川洋平, 榎美紀, 立堀道昭, "マイクロブログのメッセージを用いた発信場所推定," データ工学と情報マネジメントに関するフォーラム, March. 2012.
- [3] Y. Ikawa, M. Enoki and M. Tatsubori "Location Inference using Microblog Messages," Proc. of the Workshop on Social Web and Disaster Management 2012, held in conjunction with the 21st international conference companion on World Wide Web, pp.687-690, April. 2012.
- [4] 石川翔太, 荒川 豊, 田頭茂明, 福田 晃, "マイクロブログを用いた地域におけるホットトピック検出手法の検討" マルチメディア通信と分散処理ワークショップ, vol.2011, pp.77-82, Sep 2011.
- [5] 橋本康弘, 岡 瑞起, "都市におけるジオタグ付きツイートの統計," 人工知能学会誌, vol.27, no.4, pp.424-431, 2012.
- [6] A. Marcus, M.S. Bernstein, O. Badar, D.R. Karger, S. Madden and R.C. Miller, "Twitinfo: Aggregating and visualizing microblogs for event exploration," Proc. of ACM Conference on Human Factors in Computing Systems, pp.227-236, May 2011.
- [7] A. Sadilek, H.A. Kautz, J.P. Bigham, "Finding Your Friends and Following Them to Where You Are," Proc. of International Conference on Web Search and Data Mining, pp.723-732, Feb. 2012.
- [8] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," Proc. of International World Wide Web Conference, pp.851-860, Apr. 2010.
- [9] 榎 剛史, 松尾 豊, "ソーシャルメディアからの人物目撃情報抽出システムの試作," 人工知能学会全国大会 2011 論文集, 2011.
- [10] 渡辺一史, 大知正直, 岡部 誠, 尾内理紀夫, "Twitter を用いた実世界ローカルイベント検出," 楽天研究開発シンポジウム, 2011.
- [11] K. Watanabe, M. Ochi, M. Okabe and R. Onai, "Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs," Proc. of ACM Conference on Information and Knowledge Management, pp.2541-2544, Oct. 2011.