

マイクロブログを用いた地域における ホットピック検出手法の検討

石川 翔太^{†1} 荒川 豊^{†2}
田頭 茂明^{†2} 福田 晃^{†2}

マイクロブログサービスの流行に伴い時空間情報をもつテキストデータが爆発的に増加している。そして、そのデータを用いて、時間的または空間的にイベントを解析する研究が数多く行われている。これらの既存研究では検出対象のイベントを予め設定し、そのイベントの発生する時間また場所を解析することを目的としている。一方で、本研究ではある地域のある時間において発生するホットピックの検出システムに着目している。しかし、あるトピックに対する発言単語にはばらつきがあり、正確なトピック検出を行うには至っていない。従って、本論文ではその日本語のばらつきを吸収する、時空間を考慮した意味的辞書 (LocalWordNet) の構築手法について検討する。

A Detection Method for Hot-Topics in Local Areas Using Microblog

SHOTA ISHIKAWA,^{†1} YUTAKA ARAKAWA,^{†2}
SHIGEAKI TAGASHIRA^{†2} and AKIRA FUKUDA^{†2}

As microblog services become popular, temporal-spatial text data increases explosively. Also many studies have been proposed to analyze an event temporary and spatially. The goal of these studies is to extract the period and the location in which a specified topic frequently happened. In this paper, we focus on a system that detects hot topics in a local area and during a period. There is a variation in words even though the posted words are associated with the same hot topic. We propose a construction of a temporal-spatial meaning dictionary that mitigates the variation of posted words related the same topic in Japanese.

1. はじめに

近年、SNS・マイクロブログサービスの流行により、日常の生活や社会的な出来事などを短いメッセージとして伝える人々が増加している。中でも Twitter は 2006 年の始動以来ユーザ数は増加し続け、現在全世界中で 17.5 億アカウントが存在している (アクティブユーザ数は 5.6 億であると推定されている)^{*1}。また、スマートフォンの普及も広がっている。その 2 つが重なり、スマートフォンにおけるマイクロブログサービス向けのクライアントアプリケーションが数多く公開されている。クライアントアプリケーションには、単にユーザが言葉を投稿する機能だけでなく、撮影した写真、今聴いている音楽の情報や今自分がいる場所の位置情報などを付加する機能など、数多くの付加機能が実現されている。よって、ユーザは場所や時間の縛りなく容易に情報発信でき、様々な場所で頻繁にメッセージの発信を行っている。特に、日本においてユーザ数の成長が著しい Twitter において、その傾向は顕著に見られる。調査資料^{*2} 中でも、スマートフォン所有率と Twitter の利用率の相関を見ることができ、これ以後もスマートフォンの普及は進んでいくと考えられるため、マイクロブログのユーザ数はさらに伸びる可能性もある。

このような背景のもとで、一般的なブログでは情報として含まれなかった位置情報を持つ時空間情報テキストデータが爆発的に増加している。また手軽に投稿できるという観点から、Twitter の投稿は極めてリアルタイム性が高く、投稿内容を分析することでその時間における、その場所でのホットピックを検出することが可能である。発生するトピックの例として、TV でのサッカー中継であったり、自然災害情報等の社会的なイベントが挙げられる。また、先行研究¹⁾ において、ユーザの存在位置と入力する単語の関係性を分析した結果、特定の位置で頻繁に利用される単語が存在することが判明している。つまり、ある地域において頻度高く利用された単語が指し示す地域のホットピックに対して、同地域に存在するユーザも興味を持っているといえる。よって、ある地域におけるホットピックを抽出し、それをユーザに提示することはユーザの情報把握の観点からも有用であると考えること

^{†1} 九州大学大学院システム情報科学府

Graduate School of Information Science and Electrical Engineering, Kyushu University

^{†2} 九州大学大学院システム情報科学府

Faculty of Information Science and Electrical Engineering, Kyushu University

^{*1} <http://www.businessinsider.com/chart-of-the-day-how-many-users-does-twitter-really-have-2011-3>

^{*2} メディア環境研究所, 2011 年メディア定点調査, http://www.media-kankyo.jp/upload/files/article_128/teiten2011.pdf

ができ、検索ワードとして提示したり、日本語入力の省入力化²⁾等の人工知能サービスにも応用可能だと考えている。以上を動機とし、我々は Twitter を用いて、そのホットトピック検出システムの構築を行っている。しかし、元来日本語のツイートには日本語のゆれが存在しており、同一トピックに関する投稿単語であったとしても異なるトピックとして検知してしまうため、正確なトピック検出にはまだ至っていない。

従って、本論文ではホットトピックを正確に抽出するために日本語のゆれを吸収するための意味的辞書（以下、LocalWordNet）の構築手法を提案する。日本語のゆれには、意味的・時間的・空間的なゆれが存在する。例えば、“スタジアム”という単語が指し示すトピックが、ある場所においては野球に関するものであるが、他の場所ではサッカーに関するものであることが考えられる。また、ある場所において“祭り”という単語が指し示すトピックは時間的に変化することも考えられる（福岡地方において、5月ならば“どんたく”，7月ならば“山笠”）。本論文では、以上のように同一トピックに対して言及しているが日本語のゆれがある単語群を、1つのトピックとして分類する LocalWordNet として構築し、整理する。具体的には、この LocalWordNet を構成する単語群は、地方性の低い単語群を有す Wikipedia に登録されたデータに加えて、さらに地方性および時間依存性の高い単語群を有す Twitter の過去のデータを用いて補完することにより収集する。また、それらの単語の繋がりは、Wikipedia においては、上位概念からのパスを Infobox やスクレイピングを用いて定義、また関連度のある概念に対する説明単語の頻度により定義し、Twitter では各地域における頻出単語を基に作成する。

以下第2章では時空間情報を伴うテキストデータ分析における関連研究を紹介し、提案手法との差別化を行う。第3章ではホットトピック検出システムの概要および各要素での処理の流れを示し、第4章にて LocalWordNet の構築に関して述べる。そして、第5章にて本稿のまとめおよび今後の課題を示す。

2. 関連研究

時空間情報をもつテキストデータの分析に関する研究が多く行われている。これらの研究は、行動推定系、イベント検出系の2つに大別することができる。本研究ではその内でもイベント検出系に属するものである。

藤坂ら³⁾は Twitter が提供している API を用いてユーザの ID、時間情報、空間情報が付加された投稿を取得し、複数のユーザの移動パターン分析を行うことで地域社会的現象の発見を行っている。多くの人がある場所に集合したり、離散する度合いを検出し、お祭りの

ような社会的なイベント、特定の地域に依存した習慣的な傾向などを発見している。藤坂らの手法中では、Twitter API の制限を回避するために、地域を単一にグリッドで分割するという方法でなく、ツイート投稿数の密度に応じてグリッドの大きさを変化させている。

山田ら⁴⁾は Twitter で発信されるメッセージとそこに付加されたジオタグから、メッセージと行動の関係を条件付き確率により定式化し、過去の人々の行動履歴からユーザの未来の行動をベイズ推定によって予測する手法を提案している。藤坂らと山田らの提案手法は行動推定系に属するものである。

一方、イベント検出系に属するものでは以下の研究がなされている。山中ら⁵⁾は GPS 情報が埋め込まれたメールメッセージに対して、Sen^{*}により形態素解析を行い、SVM (Support Vector Machine)⁶⁾を用いてカテゴリに自動分類、前段階で決定したカテゴリとメッセージがもつ空間情報の2つを用いてクラスタリング、バースト検出を行うことにより観測地域でのイベントの発生(状況)を把握するシステムを提案している。バースト検出とは、対象データの時系列変化を調査し、メッセージ間の時間間隔が平常時よりも密になっているか否かを検出するアルゴリズムであり、Kleinberg⁷⁾が示したものである。山中らの手法では空間情報だけでなく、テキスト自体の情報を用いてクラスタリングするという点で上記の研究とは異なる。また、バースト検出により状況の変化を検出する点でも異なる。しかし、イベント検出に際し事前に必要な観点(例えば、公園での検出であれば「施設」「遊び」「鑑賞」「食事」「トイレ」の様なメッセージ分類子)を与えておき、その観点に沿って自動分類を行っているため、システムにおいてその観点を地域や場所の特性ごとに与えないといけない。よって、もし、どの場所にも適用するような形にするためには、そのインデックスをつける作業に手間がかかるという問題がある。

バースト性検出という点で対象を Twitter とした同類の手法⁸⁾が提案されており、ユーザが Twitter のホーム画面にアクセスした時点の直近の時間帯から 24 時間前の時間帯までのツイート投稿数の平均値を上回っているか否かにより判断を行うという方法によりバーストを検出している。ただし、空間的な側面を意識していなく、地理情報によらない統一的なモデルとなっている。

山中らと同様にイベント検出を題材とし、榊ら⁹⁾が提案を行っている。榊らは地震や台風をターゲットイベントとして想定し、各 Twitter ユーザをセンサ、彼らが投稿するツイー

*3 sen : <http://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sen.html>

ト投稿をセンサ値として見なしている。収集されたツイート投稿を MeCab^{*4}により形態素解析し、SVMを用いてターゲットイベントに関する投稿であるか判別し、イベントに関するツイート投稿数から算出するイベント発生確率により発生しているかどうかを判断する。また、位置推定でも用いられるカルマンフィルタやパーティクルフィルタを用いることにより、そのイベントの発生場所を推定するシステムを提案している。しかし、その災害に関連した単語にのみ注目しており、極めて一部分のイベントしか対象としておらず、イベント検出の汎用性に欠けている。上記に示した通り山中らと榊らの提案では事前にそのイベントに必要な観点を与えている

提案では、ある限られた範囲のイベント検出ではなく、広く全般的に検出を行う。これによりある分野のイベントに限られず、広く何が今起こっているかを把握することが可能となる。

また、上記のイベント検出において、事前に定義しているメッセージ分類子を意味的辞書を用いることにより代用することも考えられる。現在までに時空間をも加味した意味的辞書の作成という試みはされていないが、その要素として類似のものを以下に示す。WordNet^{*5} という意味的辞書が公開されており、単語概念自体の意味や他の概念間との関係が示され、シソーラス性が考慮されている。しかし、その関係性が示されているのは同一品詞間のみであり、“揺れる”から“地震”という概念を導くことは不可能である。

一方で、意味的辞書を作成するに当たり、手動ですべての意味概念を定義するには膨大な時間がかかるため、すでに幅広い分野に関する語彙を網羅している Wikipedia^{*6}や Folksonomyを用いたオントロジーの構築手法が提案されている¹⁰⁾¹¹⁾。Wikipediaではリダイレクトリンク、カテゴリツリー、Infoboxなど半構造化して作成されており、概念及び概念間の関係性の構築が可能である。さらに、地名の包含関係を加味した意味的辞書である、GeoNames^{*7}や GeoWordNet¹²⁾などのデータベースが構築されているのが、ここにおいても Wikipediaを用いた手法¹³⁾によるデータベースの構築が提案されている。上記のオントロジー構築に関する提案では、概念間の関係(is-a関係やa-part-of関係など)の定義を目的として行っているが、本研究ではあくまで概念間の関連を築くこと(上位概念からのパスの構築)を目的としている。

*4 MeCab : <http://mecab.sourceforge.net/>

*5 日本語 WordNet : <http://nlpwww.nict.go.jp/wn-ja/>

*6 <http://ja.wikipedia.org/>

*7 <http://www.geonames.org/>

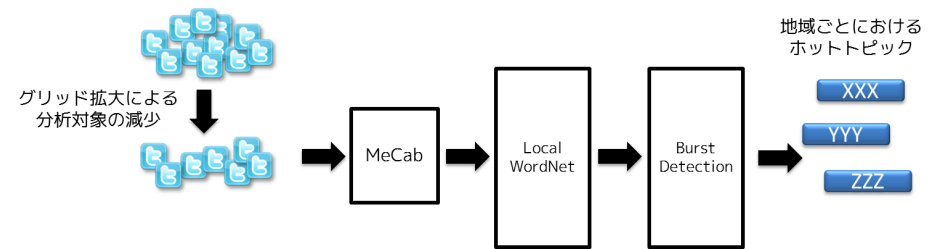


図1 システムの処理フロー

上記に示す通り、Wikipediaを知識獲得のリソースとして見たとき、それは非常に魅力的である。提案でも一部に Wikipediaを用いることにより意味的辞書の生成を行う。

3. 提案システム

我々が構築している、地域ごとのホットトピック検出システムについて述べる。具体的には、Twitter APIを用いて収集したツイート投稿を分析し、各地域ごとのホットトピックをリアルタイムに検出する。従来のモデルでもリアルタイムにトピック検出するものもあるが、極めて一部のイベントを対象としたものしか存在しない。本手法では同一意味の単語の差異を吸収し、すべてのトピックに適用できることを目指す。

まず分析の前処理として、foursquare^{*8}から投稿されたものは分析対象から外すものとする。これは、Streaming API^{*9}を用いてジオタグ付きツイートに絞って収集するが、foursquareから投稿されたものは大部分のツイート投稿に関して位置情報とURLのみしか含んでいないため、有用なサンプルとは言えないためである。

そして、ホットトピック検出に移行する。ここでは大きく分けて3つのフェーズを想定している。1. ツイート投稿分析の計算量削減のためのツイート投稿数による分析地域の選択(3.1節)、2. 単語を関係するトピックに分類するための LocalWordNetを用いたクラスタリング(3.2節)、3. バースト検出(3.3節)に頻出しているトピックを検出。本章にてその処理の詳細を述べる。

*8 Foursquare : www.foursquare.com

*9 http://dev.twitter.com/pages/streaming_api

3.1 ツイート分析エリアの選択および平滑化

前提として、図2の上部に表すように空間は同一間隔のグリッドによって分割されているものとする。単純にすべてのグリッドで分析を行うとすると計算量が莫大になる。応用先として考えているアプリケーションにおいて、処理に時間がかかり提示される変換候補がリアルタイム性に欠けたものになってしまうと、変換候補はユーザにとっての有用性が薄くなる可能性がある。よって、分析対象のグリッドの絶対数を削減することで、その計算量の減少を狙う。藤坂らの手法中でも空間分割法が示されていたが、その目的はAPI制限回避のためであった。しかし、本提案ではあくまで分析地域数の削減を目的としている。また、1章で引用した調査資料中において示されている通り、Twitterのユーザ利用率（つまりユーザ数）は地域によって大きく異なり、単純にグリッド分割して得た統計資料はサンプル数の粒度が揃っていないため、信頼性の欠けた分析結果となってしまう。例えば、ツイート投稿数が元来少ない地域においてホットトピック検出を行った場合、ある現象に対して（全国的に見ると）少ないツイート投稿数でもランキングに大きな影響を与えてしまいかねないということも考えられる。そこで予めグリッドに分割した地域が、ある一定のツイート投稿数に達しない場合には、藤坂らの空間分割法とは逆方向に、つまりある一定のツイート投稿数に達するまでグリッドを拡大していくことによりツイート投稿数の平滑化及び分析グリッド数の削減を行う（図2参照）。

ここで、その一定のツイート投稿数を N とすると、エリア E でのツイート投稿数が N 以上の場合、そのエリアはそのまま次のクラスタリングに移行する。ツイート投稿数が N 未満の場合には、あらかじめ状態数としてそのエリアが保持している隣接エリアのツイート数を基に、ツイート投稿数が N 以上を満たす隣接エリアの数を調べる。ツイート投稿数が N 以上を満たす隣接エリアの数が α 以上であれば、そのエリアでのホットトピック検出自体をとりやめ、最終結果のみを参照する。これは、トピックへの関心度は空間的に連続的であり、隣接エリアにおいてもその影響があると考えられるためである。一方、隣接エリア数が α 未満であれば、隣接エリアを結合し、再帰的にツイート最低投稿数 N との比較を行う。

以上のように、再帰的にエリアの拡大、また隣接エリアの結果を参照することで実際に分析するエリア数が減少し、計算量の減少が期待できる。

3.2 クラスタリングによるトピックへの分類

まず分析対象となったエリアの収集ツイート投稿は、日本語の文脈を品詞分解する必要があるため、MeCabにより形態素解析する。形態素解析により抽出した単語品詞の内、我々は名詞および動詞のみを候補として選択する。これは、形容詞・副詞などから修飾する対象

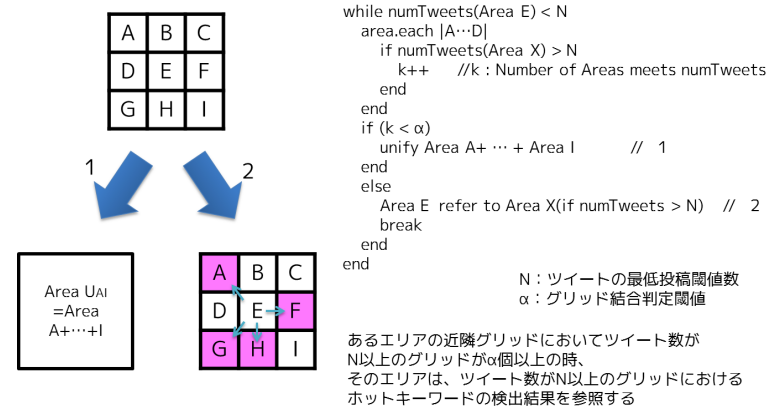


図2 空間拡大、他エリア参照モデル

語を導くことは難しいと考えられるためである。

元来日本語のツイート投稿には揺らぎがあり、同一トピックに関するものでも異なる単語を用いて表現する可能性が高い。例えば、“揺れる”や“震度”などの単語は“地震”というトピックに対してしている発言ものだと見なすことができる。この場合に用いられる意味概念は地域によらず統一的なものになっている。また、同一トピックに対して地域によって異なる単語を用いて表現したり、同じ単語でも違うトピックを指している可能性もある。例えば、福岡地方では“ホークス”、“ドーム”や“鷹”などはいずれもある野球球団に関することであり、単語同士の意味は互換性が存在する。しかし、他地方では“ドーム”という単語が別の球団や他のスポーツに関して言及している可能性がある。さらに、その地域の中でも出現する時期によって意味が異なるものがある。例えば、福岡地方において、“祭り”が指し示すトピックは、5月ならば“どんたく”、7月ならば“山笠”と考えることができる。

既存の研究⁵⁾⁹⁾では、ケーススタディとして様々な場所におけるイベント検出を行っている。それらの手法の中では、クラスタリングの手法としてSVM⁶⁾が用いられている。本研究の最終的な目的は汎用的なトピック検出であるが、SVMを用いて各トピックとの関連を図っているのは多大な時間がかかる。

よって提案では、次章に示す時空間を加味した意味的辞書であるLocalWordNetを構築することにより、この揺らぎを吸収し、各クラスタに分類する。具体的には、LocalWordNetにより、ある候補に対しての上位概念からのパス（概念階層の関係）を取得し、そのパスに

重みを付ける。ここで、Wikipedia の概念階層の深さは均一ではないので、得られた重みに対して正規化を行う。そして、修正された重みを用いて類似度を定義することにより、クラスタリングを行う。

3.3 トピック検出

先のクラスタリングにより得られたクラスタに対して、その時間的な頻発具合を検出する。その検出アルゴリズムとして、Kleinberg⁷⁾ が示したものをを用いる。Kleinberg の手法は時系列の文書（掲示板のスレッドや新着順のニュースの記事）の流れにおいて、文書が送られてくる間隔が平常状態よりも密となるバースト状態を検出するものである。

バースト検出では、まず 2 状態からなる確率的オートマトン A を定義する。この時、平常状態を q_0 、バースト状態を q_1 とおき、最初のメッセージが発信されてから最後 ($n+1$ 個目) のメッセージが発信されるまでの時間を T とする。メッセージがランダムに発信されると考えると、あるメッセージ i が発信されてから次のメッセージ $i+1$ が発信されるまでの間隔は指数分布に従うことになる。平常状態における、この発信時間間隔を x とおき、また $\alpha_0 = n/T$ とおくと、間隔 x で次のメッセージが発信される確率はポアソン分布に従い $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ となる。バースト状態時は平常状態よりも短い時間間隔でメッセージが発信されるため、間隔 x で次のメッセージが発信される確率は $f_1(x) = \alpha_1 e^{-\alpha_1 x}$ となる（ここで、 $\alpha_1 > \alpha_0$ ）。また、 $n+1$ 個のメッセージが発信される一連の間隔を $\mathbf{x} = (x_1, x_2, \dots, x_n)$ とおく（ただし、 $x_i > 0$ ）。さらに、各メッセージ間隔における状態を $\mathbf{q} = (q_{i_1}, q_{i_2}, \dots, q_{i_n})$ とすると、その確率密度関数は次式で示される。

$$f_{\mathbf{q}}(x_1, \dots, x_n) = \prod_{t=1}^n f_{i_t}(x_t) \quad (1)$$

つまり、時間間隔が \mathbf{x} の時に状態列が \mathbf{q} になる確率は次式のように示される。

$$Pr[\mathbf{q}|\mathbf{x}] = \prod_{t=1}^n f_{q_t}(x_t) \quad (2)$$

ここで、この確率が最も高くなるような \mathbf{q} が求める最尤状態（バースト状態）だと考えられる。これはすなわち、次の値を最小化することに相当する。

$$-\ln Pr[\mathbf{q}|\mathbf{x}] = \sum_{t=1}^n -\ln f_{i_t}(x_t) \quad (3)$$

最小化する状態 \mathbf{q} を求めることにより、クラスタのバースト検出が可能となり、ホットトピックの抽出が可能となる。

4. 時空間を考慮した意味的辞書 (LocalWordNet) の構築

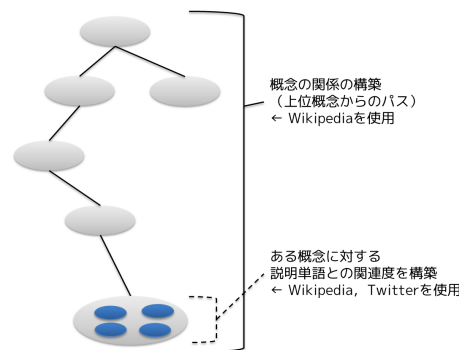


図3 概念パス及び、概念と単語の関連度構築

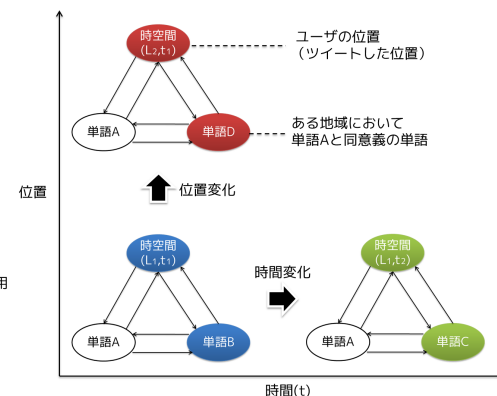


図4 LocalWordNet における時空間的な単語の意味変化

本論文で構築しようと考えている LocalWordNet はフリーテキストから作成することも可能であるが、多大な労力と時間がかかる。よって、前章の 3.2 節でも述べたように、既存の情報である Wikipedia と Twitter（過去のデータ）を用いて構築する。この 2 つの web コーパスに存在する単語群はそれぞれ次のような特徴をもつ。

- Wikipedia : 時間に依存しない、かつ地域性が低い
- Twitter : ユーザの状況を反映した、時間依存性や地域性が高い

まず、各コーパスから単語を収集する。次に、概念階層構造の構築のために収集した単語間の関連付けを行う。時空間において変動しない静的な概念階層をもつ Wikipedia では、ある概念に対するページ内での各単語の出現頻度を計ることにより、概念と単語間の関連を生成することが可能となる。また、Wikipedia で定義されている上位概念からのパスは、玉川らの提案¹¹⁾ のように、カテゴリ名と Infobox テンプレートの照合やスクレイピングにより取得する。

一方、Wikipedia で構築された概念間の関連付けを補完する役割である Twitter におい

では以下のように進める。ここにおいてもまず、グリッド拡大により分析するツイートを削減し、そこで収集したツイートを形態素解析にかけ、品詞レベルに分解し、分析対象である名詞と動詞を取り出す。そして、頻出する単語を抽出し各単語において、手動にて単語間の関連を生成する。また、収集した各ツイートにおける関連の空間的境界はグリッドの境界であるとし、その領域に対応する地域を GeoWordNet にて定義されているものと関連づけることにより、地理空間の包含関係を含むことも可能である。

以上のような工程を経ることにより、LocalWordNet では位置から単語を導くことができ、また単語からその位置（もしくは時間）において関連性のある単語を導くことが可能となる（図4参照）。

5. おわりに

特定の位置で頻りに利用される単語が存在するという事実を裏付けとし、地域ごとのホットトピック検出することに有用性があるという動機のもと、我々はホットトピック検出システムを構築している。しかし、あるトピックに対する単語にはばらつきがあり、正確にトピック検出するまでには至っていない。よって、本論文ではその日本語のゆれを吸収する、LocalWordNet の構築手法について検討した。

ホットトピックの検出シーケンスとしては、まずグリッドの拡大およびホットトピックの他エリア参照モデルにより分析ツイート投稿数の絶対数を削減し、LocalWordNet を用いてクラスタリングを行う。そして、そこで生成されたクラスタに対してバースト検出を行うことで、ホットトピックの検出が可能となる。

LocalWordNet の構築に関しては、現在の Web コーパスである Wikipedia（地方性および時間依存性の高い単語群を有す）と Twitter（地方性および時間依存性の高い単語群を有す）の過去のデータを用いることにより、階層構造をもつ時空間の変化に対応した意味的辞書構築の方法を示した。

また、今後の課題として、実装を進めるとともに、ある単語間の意味関連性がどの地域まで有効であるかという意味関連の空間的境界の決定や短縮語の扱い、類似度の重み係数の決定など手法の細部までの明確化が必要であると考えている。

参 考 文 献

1) 荒川 豊, 田頭 茂明, 福田 晃, “Twitter におけるコンテキストと単語の相関関係分析,” 情報処理学会研究報告 組込み技術とネットワークに関するワークショップ ETNET2010,

Vol. 2010-MBL-53 No. 50, pp. 1-7, 2010.
2) 荒川 豊, 末松 慎司, 田頭 茂明, 福田 晃, “コンテキストアウェア IME システムの提案と実装,” 情報処理学会 マルチメディア・分散・協調とモバイル (DICOMO2010) シンポジウム, No. 4D-1, pp. 914-922, 2010.
3) 藤坂 達也, 李 龍, 角谷 和俊, “マイクロプログラムの移動履歴を用いた地域特性分析,” 情報処理学会 データベース・システム研究会報告 Vol. 17, pp. 1-8, 2009.
4) 山田 和貴, 齊藤 裕樹, “地理位置情報履歴と発言コンテキスト解析による行動予測手法の提案,” 情報処理学会 マルチメディア通信と分散処理ワークショップ (DPSWS) 2010, pp. 71-72, 2010.
5) 山中 努, 田中 祐也, 土方 嘉徳, 西田 正吾, “時空間情報を伴うテキストデータを用いた状況把握支援システム,” 日本知能情報ファジィ学会論文誌 知能と情報, Vol. 22, No. 6, pp. 691-706, 2010.
6) T. Joachims, “Text categorization with support vector machines,” *Proc. ECML '98*, pp. 137-142, 1998.
7) J. Kleinberg, “Bursty and Hierarchical Structure in Streams,” *Proc. the 8th ACM Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pp. 91-101, 2002.
8) 大谷 友貴, 鈴木 政巳, 小林 亜樹, “Twitter における一連の事象へのつぶやき発見手法,” 電子情報通信学会 第 19 回 Web インテリジェンスとインタラクション研究会, <http://www.ieice.org/wi2/pastitem/past30.html>, 2011.
9) T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” *Proc. the 19th international conference on World wide web*, pp. 851-860, 2010.
10) 森田 武史, 山口 高平, “オントロジー学習の現状と動向,” 人工知能学会誌, Vol. 25, No. 3, pp. 354-365, 2010.
11) 玉川 奨, 桜井 慎弥, 手島 拓也, 森田 武史, 和泉 憲明, 山口 高平, “日本語 Wikipedia からの大規模オントロジー学習,” 人工知能学会論文誌, Vol. 25, No. 5, pp. 623-636, 2010.
12) F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta, “GeoWordNet: A Resource for Geo-spatial Applications,” *Proc. the ESWC 2010*, pp. 121-136, 2010.
13) 竹中 均, 吉岡 真治, 森田 武史, 山口 高平, “Wikipedia を用いた地名の包含関係情報の抽出,” 第 25 回人口知能学会全国大会, 2J3-NFC2-2, 2011.