ARG WI2 No.10, 2017
WI2-2017-7

地域型イベントの収集とその分類手法の検証

福馬 智生, 鳥海 不二夫
東京大学大学院工学系研究科システム創成学専攻
tomoki@torilab.net

概要 花火大会や近所のお祭りといった人が集まるイベントは、その周辺のコンビニなどの店舗に予期せぬ需要を生む可能性がある。そのような予告地域イベントを WEB や Twitter からの収集し、「開催場所」「開催期間」を特定する手法について検討する。また，得られたデータからどのようなイベントなのかカテゴリの推定を、教師ありもしくは半教師ありのトピックモデルを用い行った。教師ありトピックモデルを用いることにより、約 77 %でカテゴリの推定ができた。また半教師ありトピックモデルを用い、正解データのうち一部のラベルの情報を使わず学習を行い、学習に用いなかったラベルが、抜くカテゴリによっては再現されることを確認した。

キーワード Twitter, イベント検出, トピックモデル応用, 文書分類

1 はじめに

花火大会や近所のお祭りといった人が集まるイベントは、その周辺のホテルやコンビニなどの店舗に予期せぬ需要を生む可能性があり、イベント情報を事前に把握しておくことは、店舗運営の手助けになると考えられる。実際、天気や気温から需要予測を行う研究は行われているが [1]、イベントの有無はそれらに無視できない影響を及ぼし、事前に把握することで予期せぬ需要に備え潜在的損失を防ぐことが可能になる。

本研究では、Web と Twitter から、イベント情報を収集し、開催場所、開催期間、イベントのカテゴリを抽出することを目的とする。本稿ではまず予告地域イベントの収集の収集手法について述べる。それらのデータから「開催場所」「開催期間」を特定する手法について検討する。次に得られたデータから各イベントのカテゴリの推定を教師ありと半教師ありのトピックモデルを用いて行う。これにより将来的に、店舗運営にとって重要な情報となりうるイベントの年齢層や男女比の推測を行う上で、それらの予測がしやすい形での情報提供などを目指す。

2 予告型地域イベント収集

予告型地域イベントとは開催日よりも前に、「イベント三要素」[2] が告知されているイベントのことと定義する。イベント三要素とは「開催日時」、「開催場所」、「イベント名」のことを指す。本節ではこのイベント三要素のうち前者二つを手掛かりに予告型イベントを収集する手法を検討する。予告型地域イベントの収集にはイベント情報サイトから収集と Twitter などの SNS からの収集を行う。

2.1 イベント情報サイトからの予告型地域イベント収集

まず WEB 上で公開されているイベント情報サイトからイベント情報を収集する。イベント情報サイトとは、旅行会社や地方自治体などより提供されるイベント告知のためのウェブサイトを意味する。具体的には「るるぶ.com[1]」「地域情報サイト ZAQ[2]」「ことさが[3]」「Walkerplus[4]」「Let ' s enjoy Tokyo[5]」などがある。これらのサイトはそれぞれサイトのフォーマットが統一されているため、各サイトに合わせたパーサを作ることで、イベント三要素を容易に抽出することが可能である。

2.2 ソーシャルメディアからの予告型地域イベント収集

Twitter をはじめとするマイクロブログでは、イベント主催者やイベントに興味を持つ人が Twitter 上でイベント情報を多く発信している。

しかしながら、このようなソーシャルメディア上でのイベント予告は定められたフォーマットがないため、情報の抽出が困難であり、自然言語処理の技術を用いて柔軟に抽出することが必要となる。榊ら [2] の先行研究では Tweet の本文について行われているが、本研究では、本文中に含まれる url の先のテキストも分析対象に含める。次節では恣意的なパターンマッチング及び、辞書を用いて各要素を抽出手法を述べる。

[1] http://www.rurubu.com/event/
[2] http://zaq.ne.jp/
[3] http://cotosaga.com/
[4] http://www.walkerplus.com/
[5] http://www.enjoytokyo.jp/

3 イベント開催地域、開催期間抽出

3.1 データ収集

イベントの開催地域、開催期間の情報を Twitter から収集する手法について検討する。データの収集には、手がかり語「開催」という単語を用いてイベント予告候補ツイートを収集する。それらのデータのうち、url が付いているツイートに対し、リンク先の本文を抽出し、分析を行う。

3.2 地名抽出部

地名抽出部では、収集されたイベント予告候補ツイートの URL が含まれるものについて、そのリンク先について分析を行い、イベント開催地域の抽出を行う。以下正規表現、地名辞書の利用、CRF(条件付き確率場) の 3 つの手法の提案を述べる。

3.3 正規表現の利用

まず住所の正規表現を用いた抽出を行う。正規表現を用い「都道府県/市区町村/それ以降」の抽出を行う。使用した正規表現を下に示す[6]。

(…??[都道府県])((?:旭川｜伊達｜石狩｜盛岡｜奥州｜田村｜南相馬｜那須塩原｜東村山｜武蔵村山｜羽村｜十日町｜上越｜富山｜大町｜蒲郡｜姫路｜大和郡山｜下松｜岩国｜田川｜大村｜四日市｜廿日市｜野々市) 市｜.+?郡 (?:玉村｜大町｜.+?)[町村]｜.+?市.+?区｜.+?[市区町村])(.+)

市と区が並んでいるとき「市区町村」に区まで含む場合と、市のみとする場合がある。前者は「政令指定都市の行政区」であり、後者は「地域自治区・合併特例区」または「町名の一部の最後が区になっている」場合である。基本的に市と区が並んでいれば区まで市区町村に含めるようにするが、上の正規表現の「旭川」から「大村」の部分は市と区が並んでいるが、市のみとする例外に対応するためである。上の正規表現内の「四日市市」、「廿日市」、「野々市市」は市の中に「市」という文字が含まれているものに対応するためである。また上の正規表現内の「玉村」、「大町」は町村内に「町」や「村」が含まれている場合に対応するためである。これによりイベントの開催場所が住所表現で書かれている場合に「都道府県/市区町村/それ以降」の抽出を行うことが可能となった。

3.4 地名辞書の利用

あらかじめ用意した地域や施設名の含んだ辞書（GSK 地名施設名辞書）と参照し、地名と一致する名詞がリンク先に含まれるかを検証する。本研究では言語資源協会 (GSK) で販売されている GSK2012-C GSK 地名施設名辞書第 2 版[7]を使用した。GSK 施設名辞書には、全国の施設名 1,000 件，全国の住所 117,075 件，それらの住所の表記揺れ，読み仮名，緯度経度などが登録されている。これを用いることにより、住所表現がなく、会場名のみの表現による開催場所の情報の取得が可能になった。

3.5 CRF の利用

最後に CRF(条件付き確率場) を用いてイベント名称を抽出する。CRF は J.Lafferty らによって提案された系列ラベリング問題を解く手法であり、系列に対してそれに対応するラベルを推定する [3]。本稿ではイベントの説明をしているテキストの形態素の並びを入力系列とし、抽出したい情報を IOB2 タグで表し、これを出力ラベルとする。この形態素に対して特長を表す素性関数を設定しその素性関数の重要度を表す重みを CRF は学習する。学習に用いた IOB2 タグはは 8 種類（ART 固有物名、LOC 地名、ORG 組織、PSN 人名、DAT 日付、TIM 時間、MNY 金額、PNT 割合）である。本稿では、CRF の実装として CRFsuite [8]を用いた。

3.6 CRF の適用具体例

実験データとして学習に用いたコーパスは、ウィキニュース日本語版 500 文を MeCab で形態素解析して IOB2 タグでタグ付けしたものを用いる。以下抽出例を示す。

- 抽出成功例：下北沢「ザ・新年会!」でホリエ、NCIS 村松、KeishiTanaka が弾き語るナタリー 1 月 26 日に東京・下北沢 CLUB251 にてライブイベント「ザ・新年会！」が開催される。

  抽出結果　：東京・下北沢 CLUB251

一方抽出に失敗した例を以下に載せる。このように開催場所が@の前後に書かれている場合など抽出に失敗した。

- 抽出失敗例：「ビックカメラ水戸駅店@フォトコン開催中」や「10/29・30HALLOWEEN FESTIVAL@湘南 T-SITE」

CRF を用いることで、上記の例のように完全な住所表現の書かれ方をしておらず、施設名に登録していない施設でも、前後の単語の並びより抽出が可能になった。しかし、この手法の場合、日本や韓国といった、地名と判断されるが実際は検索対象ではない地名も検索に引っかかる可能性がある。対策としてはこのような分析目

[6]なるべく短い正規表現で住所を「都道府県/市区町村/それ以降」に分けるエクストリームスポーツ,http://qiita.com/zakuroishikuro/items/066421bce820e3c73ce9, (最終アクセス 2017 年 5 月 9 日)

[7]http://www.gsk.or.jp/catalog/gsk2012-c/

[8]CRFsuite, http://www.chokkan.org/software/crfsuite/.

的とは異なるワードについては NG ワードとして省くか、実際のイベント情報サイトを元にコーパスを作成し学習させるなどの手法があると考えられる。また CRF を用いた手法では名前だけしか取れないので、緯度経度を推定するために GeoNLP[9],Google Place API[10] または Google Geocoding[11] を利用することが必要だと考えられる。

3.7 精度評価

2016 年 10 月 24 日〜10 月 31 について「ハロウィン」「開催」を手掛かりとして取集したツイートのうち、100 個について人の目で精度確認を行った。正規表現を用いた結果 71 %、GSK 地名辞書を用いた結果 98 %、CRF を用いた結果 58 %の精度となった。GSK 地名辞書を用いた場合は精度が高いものの、載っている地名には限りがあり、抽出できたデータが少なかった。一方 CRF は抽出される数は多い一方、精度が他に比べて劣る結果となった。

4 開催期間の抽出

イベントの開催期間の書き方は、全てが 1/2〜1/5 といった特定のフォーマットに従って書かれているわけではない。例えば、本日から三日間や、10/2 から 4 日間といった開催日が単一の日付ではなく、期間となっている場合がある。このような期間にも対応するように抽出した日付のまわりに「より」「から」「、」といった手がかり語が入っているかを調べ、パターンと合致する場合については開催期間として抽出する。開催期間についても、開催期間の表現は限られているため、地名抽出手法と同様に正規表現の手法を適用する。使用した正規表現を表 1 に示す。範囲を示す手掛かり語には「〜」「、」「と」「から」「d 日間」を用いる。

4.1 精度評価

3.5.2 節と同様,100 個のツイートについて、人の目で精度評価を行った結果、全てのツイートで正しく日付が取得されていることが確認された.

5 教師ありトピックモデルを用いたカテゴリ推定手法

5.1 概要

カテゴリとはそのイベントを特徴づけるクラスのことで、例えば「花火」「イルミネーション」などが挙げられる。本章では、抽出したイベントの予告文から、カテゴリを自動推定する手法について述べる。推定には教師ありのトピックモデル、Labeled LDA によって、教師あり多クラス分類問題として解く。

[9]https://geonlp.ex.nii.ac.jp/
[10]https://developers.google.com/places/
[11]https://developers.google.com/maps/documentation/

表 1 使用した正規表現

正規表現	取得できる日付例
mm 月 dd 日	11 月 30 日
mm/dd	10/23
mm.dd	9.12
dd 日	24 日
dd 日	6 日
dd（月｜火｜水｜木｜金｜土｜日）	3 金
dd+((｜()+（月｜火｜水｜木｜金｜土｜日）+()｜）)	3（金）

5.2 Labeled LDA によるカテゴリ推定

Labeled LDA(LLDA) は Ramage らによって提案された人によって文書にあらかじめつけられたタグをその文書の意味を表現するものと捉え、潜在トピックの抽出の際に教師信号として利用することを考えた、教師ありのトピックモデルである [5]。図 4.2 に LLDA のグラフィカルモデルを示す。L-LDA と LDA との違いは，ラベル（文書に与えられているタグ）の情報 Λ が，θ を推定する際に影響を与えているという点である。具体的には，ラベルの有無の情報を 1 または 0 の 2 値によって与えることによって、文書ごとに射影行列を生成し、α を制限した新しいパラメータを生成し、トピック分布を求める。他の過程は，LDA と同様である。

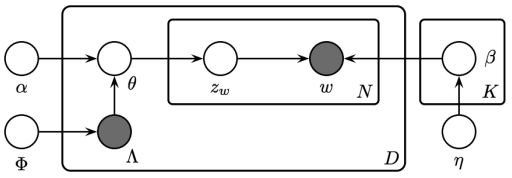

図 1 システム構成

5.3 実験

Walkerplus のデータ 2786 件を用いて、LLDA によるカテゴリ推定の精度を確認する。ここでは、各イベントから、当該イベントに関する説明文を抽出し、それらに対し MeCab による形態素解析を行い、名詞、動詞、形容詞、形容動詞のみを抽出して LLDA への入力とした。

各イベントデータには、Walkerplus によって大カテゴリ 6 種類と小カテゴリ 31 種類があらかじめタグとして付与されている。本実験では、小カテゴリを学習に用いて実験を行った。学習データを元に得られたトピック内の上位単語の一部のうち、5 トピックについて抜粋してを表 2 に示す。LLDA により、一つ一つのイベントは

音楽やグルメや物産展といった、ラベルの混合トピックとして表現することが可能となる。

表 2 LDA を用いて得られたトピックの一部

カテゴリ名	内容
トピック 0	体験　定員　開催 参加　自分
トピック 2	イベント　紹介　体験　開催　本展
トピック 28	イルミネーション　温泉　健康　お餅　今年

5.3.1　LLDA を用いた具体的なトピック分布例

以下下記に示すイベント記事の説明文について LLDA でトピック分布を調べた結果、各カテゴリへの所属度がどの程度であったかを図 2 に示す。縦軸にカテゴリを、横軸にカテゴリへの所属度を示す。Walkerplus でもともと付与されていた正解ラベルは「物産展」であった。一方出力結果は「物産展」が 27.8 %、「フェア」が 19.0 %、「グルメ」が 13.2 %となった。正解カテゴリの「物産展」が一番上位になった他、文章内には食べ物に関する記述もあり、正解ラベルのようにただ一つのカテゴリで表現するのではなく、混合トピックを用いて表したほうが、より良くカテゴリを表現できていると言える。

新春から東西の寿司の名店、行列店に話題のスイーツなど全 60 店舗が大集合。東から銀座 久兵衛の太巻き、西から初登場のアウームは目と舌で堪能できる新しい感覚の彩り寿司で西武池袋本店限定品を提供。出来たてを楽しめるイートインには大田原牛超の大田原牛ローストビーフうにのっけ丼や五ノ神製作所の海老トマトつけ麺、福岡の行列店アイボリッシュの華やかなベリーデラックスなど、今食べたい日本の美味づくしの物産展。

正解ラベル：物産展

5.3.2　Labeled LDA を用いた精度評価

Labeled LDA で得られたカテゴリの分布を特徴ベクトルとし、二番目以降の所属度のカテゴリも学習に用いるため、RandomForest を用いてカテゴリ予測して精度検証を行った。また比較手法として教師無しトピックモデルである LDA を用いて次元圧縮した後、RandomForest を用いてカテゴリを予測したものと DBN を用いて次元圧縮した後、RandomForest を用いてカテゴリを予測したもの用いる。LDA での仮定するトピック数は LLDA でのトピック数（正解カテゴリ数）と同じ 31 とした。

以上の実験の結果を図 3 に示す。横軸は推定したカテゴリの上位 n 番目以内を示し、縦軸はその n 番目以内に正解があった場合正解カテゴリがある場合を正解とした時の精度を示す。LLDA+RandomForest を用いた結果は LDA+RandomForest の結果に比べ約 35 %、DBN+RandomForest の結果に比べ 25 %改良され、約 77 %の確率で推定上位 3 番目以内に推定し、一般的な教師なしの特徴抽出の手法よりも、精度良く特徴の抽出ができていることが確認された。

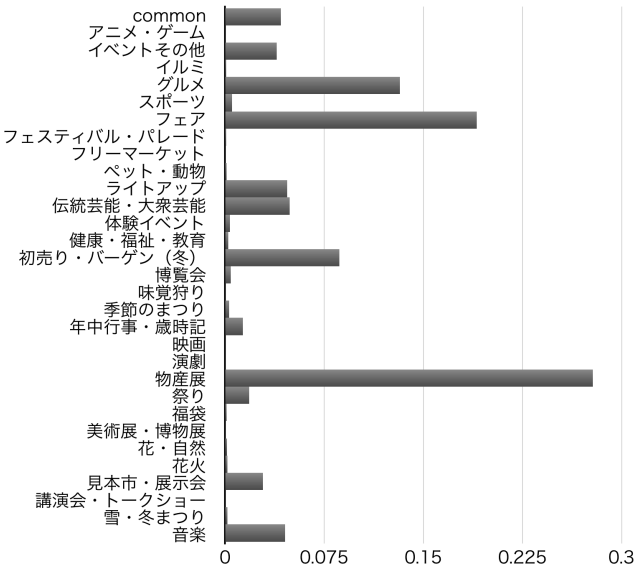

図 2 トピック分布例結果

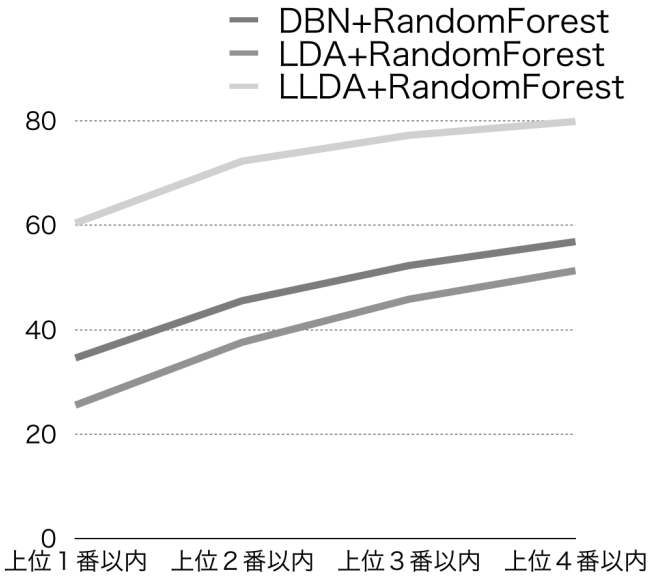

図 3 トピック分布例結果

6 半教師ありトピックモデルを用いたカテゴリ推定

6.1 半教師ありトピックモデル

大量のイベント情報のデータを集めることは可能になったとしても、大量の正解ラベルの取得はコストの観点からも難しい場合が多い。このような場合、半教師あり学習を用いることでラベルの付いたデータのみを用いるよりも、ラベルなしのデータも学習に用いることで精度の向上が期待される。また半教師ありトピックモデルは、ラベル付きデータとラベル無しデータの両方を入力に用い、人があらかじめ一部の文書や単語にラベルを与えることで、人手で与えたラベル以外に未知のラベルを自動的に検出し付与することも可能である。本研究では、半教師ありトピックモデルとして semi-supervised LDA(以降 ssLDA)[6] を用いた。ssLDA の利点としては LLDA と異なりトピック数を自由に決められる点があげられる。それにより、学習に用いたラベルを代表するトピックと任意の数の新しいラベル無しのトピックが得られる。

以下では、ssLDA で未知のラベルを検出する精度を確認するために、既知のラベルの一部をあえて付与せずに学習させ、ssLDA を用いたトピック抽出を行う。それにより得られた未知のトピックに、学習に用いていないカテゴリが再現されているかを検証する。また結果について、LLDA と比較した精度検証を行う。

6.2 ssLDA を用いた大カテゴリ分類精度評価

本節での分類実験では　学習で用いたイベント記事にもともと付与されている大カテゴリ 6 分類で行う。ssLDA でトピック分解を行う際、6 つのうち 1 つのラベルを削除し，残った 5 つのラベルを用い、かつ学習させるトピック数を 6 に設定した上で、各イベントをトピック分解した。その結果、得られたトピックのうち、もっとも所属度が高いものをカテゴリの推定結果とする。ここで、精度は未知のラベルに所属するイベントが削除したラベルに所属している割合とする。例えば、「趣味・生活」を削除して ssLDA による学習を行った場合、得られた未知のカテゴリに所属したイベントに本来「趣味・生活」ラベルが付与されていた割合が、精度となる。一方、再現率は削除したラベルが付与されていたイベントのうち未知のラベルが付与されたイベントの割合となる。なお「その他」ラベルは色々なトピックの集合であることを考え、今回の ssLDA の学習から抜くラベルには用いなかった。

結果を表 3 に載せる。比較手法として 6 ラベル全てを用いて学習を行う LLDA を用いた結果を表 4 に載せる。結果として、LLDA よりは精度や再現率はトピックによるが若干落ちるものの、ラベルなしのトピックとして元々のラベルを推定が出来ていることが分かる。

表 3 ssLDA を用いた 6 カテゴリ分類評価

抜いたカテゴリ	精度	再現率	f 値
お祭り	0.2	0.62	0.30
趣味・生活	0.76	0.47	0.58
食べる・買う	0.63	0.65	0.58
季節のイベント	0.40	0.37	0.39
文化・芸術・スポーツ	0.84	0.47	0.60

表 4 LLDA を用いた 6 カテゴリ分類結果

カテゴリ	精度	再現率	f 値
お祭り	0.15	0.88	0.25
趣味・生活	0.82	0.58	0.68
食べる・買う	0.75	0.77	0.76
季節のイベント	0.59	0.52	0.64
文化・芸術・スポーツ	0.84	0.55	0.68

6.3 ssLDA で得られたラベルなしトピックの中身

前節で得られた、学習に用いなかったカテゴリの推定結果のうち、元々付与されていたラベルがそれぞれどれほど含まれていたかを表 5 に示す。縦軸は推定されたラベルであり、横軸は正解ラベルである。結果として、カテゴリによって推定のしやすさが異なることが分かった。特に「お祭り」カテゴリは「食べる・買う」を除く全カテゴリと類似傾向にあり、「季節のイベント」は「趣味・生活」「文化・芸術・スポーツ」「その他」と類似傾向にあることが分かった。「文化・芸術・スポーツ」を学習に用いなかった場合、比較的良い精度が出ていることから、「文化・芸術・スポーツ」は他とはトピックとして独立したカテゴリであることが考えられる。ssLDA を用いる場合、類似したカテゴリは学習に用いたラベルに吸収されやすく、独立したカテゴリなら再現されることが分かった。

以下具体例を示す。以下の例では、「お祭り」を除き学習を行った。正解ラベルは「お祭り」である一方、推定結果は「食べる・買う」であった。実際に文章中には食べ物の話題が含まれており、この推定結果は妥当と言える。このように学習に用いなかったカテゴリの文章のうち、似たカテゴリを含む場合は、そのカテゴリに吸収されやすいことが観測された。

全国各地の食や祭りの魅力を余すことなく発信。9度目の開催となる今回も一度は見てみたい全国各地の祭りが東京ドームに集結「青森ねぶた祭」や「弘前ねぷたまつり」など、全国各地に伝わる巨大な山車や躍動感あふれる演舞がステージを彩る。また、大人気コーナー「全国ご当地どんぶり選手権」や「ご当地スイーツマルシェ」、「絶品！逸品！ちょいのせ市場」、「イケ麺スタンプラリー」など食の企画も内容盛りだくさん。

正解ラベル：　「お祭り」
学習に用いなかったラベル：「お祭り」
推定結果：「食べる・買う」

表 5 推定されたトピックの Confusion Matrix

	お祭り	趣味･生活	食べる買う	季節のイベント	文化･芸術･スポーツ	その他
お祭り(推定)	11	11	0	9	7	9
趣味･生活(推定)	1	30	1	0	9	7
食べる買う(推定)	1	7	26	0	0	5
季節のイベント(推定)	2	8	1	10	8	7
文化･芸術･スポーツ(推定)	0	4	0	0	26	2

7　結論と課題

本研究では、イベント情報サイトと Twitter からイベントの情報を開催前に抽出を行い、会場周辺のコンビニの仕入れの手助けとなるように、「開催地」、「開催期間」、「イベントのカテゴリ」を推定し、マーケティングへの利用を容易にするための研究を行なった。Twitter のイベント告知にはフォーマットがないため、自然言語処理を行い、柔軟な抽出を行なった。開催地の抽出には正規表現、地名辞書、CRF を用いて抽出し、開催期間は正規表現を元に抽出した。

カテゴリ推定においては、Labled LDA を用いた教師ありの手法を検討した。通常 LDA から得られるトピックは意味づけが難しく、イベントのカテゴリを推定するのは難しかった。しかし LLDA を用いた結果、正解カテゴリと感覚的に一致する単語の分布となった。LLDA を用いた多クラス分類として扱った場合、元の正解カテゴリを約 77 %の確率で上位 3 番以内に推定することができた。

しかし一般的に正解データの収集は困難であり、全てのカテゴリを手動で全て用意することは難しく、思わぬイベントカテゴリが含まれて入る可能性もある。半教師ありのトピックモデル ssLDA を使った場合、学習に用いたのとは別の新しいトピックが得られることがわかり、感覚的にも納得しやすい結果が得られた。これらのことから膨大なイベント予告の文章から例えばキーワードベースで対象の文書にあらかじめ用意したカテゴリを付与し、ラベルを付与しないものとともに半教師ありトピックモデルを用い、ラベルなしのトピックのイベントには新しいカテゴリがあることを示唆し、新たなカテゴリ発見の手助けにもなることが分かった。

今後の課題としては、正解データの存在する Web だけではなくイベント検出の手法を使った Twitter のデータを使ったイベント情報のカテゴリ推定を検証する。あらかじめ地域や特徴語を含むツイートに対し、ラベル付けを行い、教師データとして使用する。その後ラベルの付与されていないデータとともに ssLDA で学習を行い、ラベルなしのトピックの抽出を行い、新たなトピックの発見を目指す。

参考文献

[1] 石垣司，竹中毅，本村陽一“条件付層別差分モデルによる需要予測の高精度化，第 25 回人工知能学会全国大会,2011.

[2] 榊剛史, 松尾豊：ソーシャルメディアの予告型イベント及び参加条件の抽出手法,JSAI’2013.

[3] Lafferty, J., McCallum, A. and Pereira, F. : Conditional Random Fields : Probabilistic Models for Segmenting and labeling Sequence Data, In Proc. of 18th International Conference on Machine Learning, pp. 282-289 (2001).

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., Vol. 3, pp. 993〜1022, March 2003.

[5] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In Proc. ICWSM 2010. American Association for Artificial Intelligence, May 2010.

[6] Di Wang,Marcus Thint,Ahmad Al-Rubaie :Semi-Supervised Latent Dirichlet Allocation and Its Application for Document Classification. WI-IAT ’12 Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 03 Pages 306-310