

# トピックと局面の対応関係に基づく 実生活ツイートのマルチラベル分類

山本 修平<sup>1,a)</sup> 佐藤 哲司<sup>1,b)</sup>

受付日 2013年12月22日, 採録日 2014年4月7日

**概要:** 身近な出来事や関心事を投稿し共有する Twitter 上には, 食事や交通, 災害, 気象など, 様々な生活の局面で有益なツイートが数多く投稿されている. 本論文では, これらの有益なツイートに適切な局面を付与する, 階層的推定法を用いたマルチラベリングを提案する. 第1段階では, 大量のツイートに対して LDA を用いてトピックを抽出し, 第2段階では, ラベル付けされた少量のツイートを用いてトピックと局面の対応関係を構築する. 構築した対応関係を用いて未知のツイートに局面ごとのスコアを算出し, 閾値を超えたスコアを持つ局面を動的に付与する. 収集した大量のツイートを用いた評価実験を行い, 未知のツイートに対して適切な複数の局面を推定できることを明らかにした.

キーワード: Twitter, 実生活, LDA, 階層的推定法, マルチラベル分類

## Multi-label Classification for Real Life Tweets Based on Association between Topics and Aspects

SHUHEI YAMAMOTO<sup>1,a)</sup> TETSUJI SATOH<sup>1,b)</sup>

Received: December 22, 2013, Accepted: April 7, 2014

**Abstract:** Recently, lots of users share their current events and opinions by using the Twitter. Hence, some of these tweets are beneficial in several aspects of user's real life, i.e., eating, traffic, disaster, weather, and so on. In this paper, we propose multi-label classification using hierarchical estimation method for estimating appropriate aspects. In the first phase, many topics are extracted from a sea of tweets using Latent Dirichlet Allocation (LDA). In the second phase, associations between many topics and fewer aspects are built using a small set of labeled tweets. Our sophisticated evaluations demonstrate that the proposed method can estimate several appropriate aspects of each unknown tweets.

**Keywords:** Twitter, real life, LDA, hierarchical estimation method, multi-label classification

### 1. はじめに

現在, 知識共有コミュニティサイトやブログ, マイクロブログなど, 多くの情報共有サービスが存在している. ツイートと呼ばれる短文を投稿する Twitter<sup>\*1</sup>は, 最も広く普及しているマイクロブログの1つであり, 2013年末に2億4,100万人の月間アクティブユーザ数を記録している<sup>\*2</sup>.

ユーザは自らの経験や意見, また日常生活でのイベントなど, 身近な「今」を投稿している. このため, 他のユーザにとっても最新かつ有益なツイートが多く, たとえば電車の遅延情報は交通機関を利用するユーザに役立ち, 近所のスーパーマーケットの特売情報は買物に出かけようとしているユーザを支援できる. これらのような地域性が高く新鮮かつ, 他のユーザに有益なツイートを, 著者らは「実生活ツイート」と呼び, 大量のツイートの中から実生活ツイートを抽出することを試みてきている [16], [20].

<sup>1</sup> 筑波大学大学院図書館情報メディア研究科  
Graduate School of Library, Information and Media Studies,  
University of Tsukuba, Tsukuba, Ibaraki 305-8550, Japan

a) yamahei@ce.slis.tsukuba.ac.jp

b) satoh@ce.slis.tsukuba.ac.jp

<sup>\*1</sup> <https://twitter.com/>

<sup>\*2</sup> Twitter, Inc., <https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=823321>

ユーザにとって有益な実生活ツイートは、Twitter に数多く投稿されている。一方で、実生活ツイート以外のツイートも少なくない。特に、「ありがとう」や「なるほど」といった、誰かの投稿に対する相槌や共感などの、ユーザの生活を直接支援しないツイートが多い。このようなツイートの混在は、実生活ツイートの発見を妨げる原因となっている。

実生活ツイートは、生活の様々な局面に対応している。たとえば、「電車が来ない」というツイートは生活の中の「交通」の局面に対応し、これから電車に乗ろうとしているユーザを支援できる。「雨が降ってきた」というツイートは「気象」の局面に対応し、これから外出する人や、洗濯しようとしている人など、幅広いユーザを支援できる。著者らは、Wikipedia の「地域コミュニティ」\*3 と「生活」\*4 を参考に、人々の生活を表 1 に示す 14 の局面に整理し、いずれかの局面を未知のツイートに付与する手法を提案している [16]。しかし、ツイートによっては複数の局面を付与するのが適切な場合もある。たとえば、「激しい雨のため、電車が遅延しています」というツイートは、「激しい雨が降っている」と、「電車が遅延している」とについて言及している。ツイートの主題は「電車の遅延」であるが、このツイートは屋外の状況を確認できないユーザに対して「雨が降っている」ことを伝えることもできる。ユーザは、現在の活動状況に適した局面を指定して、必要とする情報を選択して入手するであろうから、このツイートには「交通」だけでなく、「気象」も付与しておくことが望ましい。「気象」を付与しておくことで、電車で外出する生活者だけでなく、洗濯物を外に干している生活者にも有益な実生活ツイートとなる。

本論文では、分類の手がかりとなる単語が少ない短文の

ツイートに対して、大量のツイートを用いてトピックを抽出する第 1 段階と、少量のラベル付きツイートを用いてトピックと局面の対応関係を学習する第 2 段階からなる、階層的推定法によるマルチラベル分類を提案する。提案法では、トピック中に含まれる単語の生起確率と、トピックと局面の関連度を用いて、未知のツイートに対して局面ごとにスコアを算出する。閾値を超えたスコアを持つ局面を、未知のツイートに対して推定することによって、マルチラベル分類を実現する。

本論文の構成を以下に示す。2 章は関連研究について述べる。3 章は、少量のラベル付きツイートでマルチラベル分類を実現する階層的推定法について提案する。4 章は、大量に収集した実際のツイートを用いて提案手法の有効性について評価する。5 章で考察を行い、6 章で結論と今後の課題を述べる。

## 2. 関連研究

### 2.1 経験マイニングに関する研究

実生活ツイートは、ユーザの経験や知識、あるいは地域に特有の情報を含んでいる。文書から経験情報を抽出する方法として、経験マイニングに関する研究がいくつか行われている。Kurashima ら [10] は、人間の経験を {状況, 行動, 主観} からなる情報ととらえ、文章中から {時間, 空間, 動作, 対象, 感情} を自動抽出する手法を述べている。Inui ら [7] は、人間の経験を {時間, 極性, 話者態度} の観点から、{トピック, 経験主, 事態表現, 事態タイプ, 事実性} の各項目に索引付けする枠組みを提案している。これらの文章の構造に着目した経験マイニングに関する研究は、ブログなどの長い文章に対して効果的であるが、Twitter に投稿される短い文章では、有効に機能しないと考えられる。

### 2.2 Twitter からの情報抽出に関する研究

Twitter から有益な情報を抽出する研究は、数多く行われている。Sakaki ら [13] は、Twitter ユーザをセンサと見なし、地震などの現実世界で起きるイベントを発見する手法を明らかにしている。Mathioudakis ら [11] は、収集したツイートからバーストキーワードを見つけ出し、キーワードの共起でクラスタリングを行い、リアルタイムに変動するトレンドを発見することを目指している。Zhao ら [19] は、Twitter に投稿された情報要求に関するツイートを Support Vector Machine (SVM) [4] によって抽出し、ユーザの情報要求から現実世界のイベントやトレンドを発見する手法を提案している。本論文は未知のツイートに対して実生活の局面を推定することを目的としているため、これらの Twitter に関する研究とは異なる。

### 2.3 トピックモデルを利用した研究

トピックモデルに関する研究では、Blei ら [1] によって提

表 1 実生活の局面

Table 1 Aspects of real life.

局面	典型的な単語
服飾	衣服, 服装, 着る, 装飾, 化粧, 理髪, 衣装 ...
交流	約束, 出会い, 招待, 友人, 誘い, 勧誘, 飲み会 ...
災害	洪水, 竜巻, 地震, 火事, 津波, 二次災害 ...
食事	料理, 外食, 食べ物, レストラン, ジャンクフード ...
行事	祭り, 冠婚葬祭, 日程, 開催日, 学園祭, 文化祭 ...
消費	購入, 買う, 注文, 安売り, 特売, ショッピング ...
健康	風邪, 体調, 怪我, 痛み, 健康法, 病気予防 ...
趣味	余暇, 娯楽, おもちゃ, 音楽, テレビ, ゲーム ...
居住	掃除, 家具, 洗濯, 住まい, 隣人, アパート ...
地域	観光, 地域情報, 地理情報 ...
学校	勉強, 宿題, 課題, 試験, テスト, 資格, 研究 ...
交通	電車, バス, 飛行機, 時刻表, 渋滞, 混雑, 遅延 ...
気象	天気, 気温, 湿度, 風, 花粉, 雨量, 空模様 ...
労働	アルバイト, 研修, 就職活動, 営業, 仕事 ...

\*3 <http://ja.wikipedia.org/wiki/地域コミュニティ>

\*4 <http://ja.wikipedia.org/wiki/生活>

案された、潜在的ディリクレ配分法 (LDA) が広く知られている。LDA とは、1つの文書に複数のトピックが存在すると仮定した確率的トピックモデルであり、それぞれのトピックがある確率を持って文書上に生起するという考えのもと、各トピックの確率分布を導出する教師なし学習モデルである。Zhao ら [18] は、1つのツイートは1つのトピックを表すという仮説に基づいて、Twitter-LDA と呼ばれるモデルを提案している。Zhang ら [17] は、LDA を用いて、アーティストを推薦する手法を提案している。ユーザの選好アーティストを特徴量として生成したトピック集合と、アーティストのコミュニティに所属するユーザを特徴量として生成したトピック集合を用いて、アーティスト間の類似度、ユーザ間の類似度を算出し、精度だけでなく意外性のあるアイテムの推薦も行えるとしている。Weng ら [15] は、Twitter 上で影響力のあるユーザの発見のために、LDA を用いてユーザのトピック同定をしている。トピックごとにユーザのネットワークを構築し、各ネットワークに対して PageRank を拡張した TwitterRank を適用することで、ユーザをランキングする手法を提案している。本論文は、LDA を用いてトピック抽出をしたうえで、トピックと局面对の対応関係を構築し、トピックと局面の関連度とトピック中の単語の生起確率を用いて、未知のツイートに対してマルチラベル分類をすることに特徴がある。

### 3. 局面の階層的推定法

#### 3.1 階層的推定法の概要

実生活ツイートは表 1 に提示したように、様々な局面を含んでおり、局面に関連するすべてのキーワードを列挙することは困難である。また、経験マイニングで用いられているルールベースの解析手法は、Twitter に投稿される記事が短く単語の省略も多いことから、ツイートへ局面を付与することは難しいと考えられる。

本論文では、教師ありの機械学習によるアプローチにより、未知のツイートに対して適切な複数の局面を推定することを試みる。ここでの課題は以下に示す 3 項目である。

##### 課題 1

ツイートは平均 45 文字 [21] と短いことから、少ない手がかり語からツイートの言及している局面を推定する必要があること。

##### 課題 2

推定対象が生活に密着した局面であるため、1つのツイートが複数の局面に言及していることを考慮し、マルチラベリングをする必要があること。

##### 課題 3

与えられた局面にツイートを分類するには、局面ごとに正解データを作成する必要があるが、人々の生活は時間とともに変化していくことから、なるべく最新に投稿されたツイートを訓練データとすることが望まし

く、できる限り少量の訓練データで高い推定精度が得られること。

従来の教師あり機械学習による手法では、Naive Bayes 分類器 [5] や SVM [4] を用いたマルチクラス分類が広く知られている。両手法ともマルチラベリングへ拡張されており [8], [14], またトピックモデルの 1 つである Labeled LDA [12] もマルチラベリングを目的に提案されている。いずれの手法も十分な訓練データを用いることで、ブログやニュース、Web ページなどの比較的長い文書を分類することを目的とし、高い推定精度を示している。しかし、本論文で課題とする、短文である場合や訓練データが少ない場合には、考慮できる手がかり語が少なくなるため、十分な性能を得られないと考えられる。

本論文で提案する階層的推定法の概要を図 1 に示す。階層的推定法は、第 1 段階で LDA を用いて大量のツイートからトピックを抽出する。LDA は大量の文書集合をクラスタリングするための、教師なし学習モデルであり、訓練データを必要としないという特徴がある [1]。第 2 段階では、局面を特徴付けるトピックを抽出するため、局面ラベルが付与された少量のツイートをを用いて、トピックと局面对の対応関係を構築する。

入力されたツイートから抽出した単語のトピック中の生起確率と、トピックと局面の関連度を用いてスコアを算出し、スコアが閾値を超えた局面を未知のツイートに対して推定する。

従来の教師あり機械学習手法は、訓練データから直接クラスラベルに対する単語の尤度を学習しているが、提案する階層的推定法は、局面とトピックの関連度を算出し、関連度に基づいて対応関係を構築する。提案手法の特徴は、ツイートに出現する単語をトピックに展開し、ツイートが言及している話題をトピックという単位で確率的に推定した後に、少量の訓練データであらかじめ学習したトピックと局面の関連度からツイートに局面を推定することにある。

このように、トピックを中間層として 2 段階の推定を行うことで、訓練データとなるツイートに出現しない単語であっても、トピック内で訓練データに含まれる単語と共起していれば、局面を推定することができる。すなわち、局

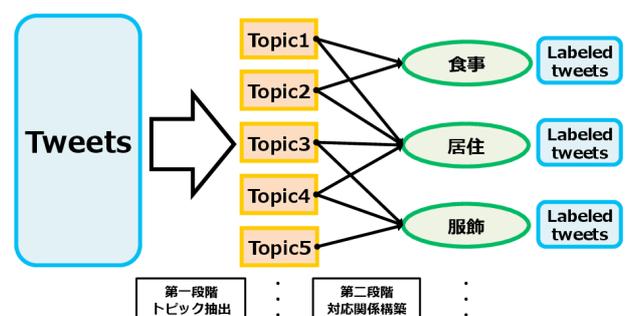


図 1 階層的推定法

Fig. 1 Hierarchical estimation method.

面を推定しようとするツイートに出現する単語を、トピックを使って確率的に拡張していることに特徴がある。しかし、むやみに拡張するとノイズとなる単語によって推定精度が低下することから、LDAによって大量のツイートからトピックに属する単語の出現確率をあらかじめ学習しておく。

以下、3.2節ではトピックと局面の対応関係を構築する手法を、3.3節では未知のツイートに対して複数の局面を推定する手法について詳述する。3.4節では対応関係を構築する際に最適なトピック数の決定方法を説明する。

### 3.2 トピックと局面の対応関係の構築方法

大量のツイートをLDAの入力とし、あらかじめトピックを生成しておく[16]。抽出するトピック数は3.4節で最適化方法を示すが、数百トピックで最適となる。トピックと局面の対応関係を構築するため、局面がラベル付けされた少量のツイートを正解データとして用意する。日本語形態素解析器MeCab[9]を用いてツイートを形態素解析し<sup>\*5</sup>、得られた語彙集合を $W$ とする。

各局面 $a$ における単語 $w$ の重要度として、単語の生起確率 $p(a, w)$ を算出する。本論文では、マルチラベルが付与されているデータから、単語の生起確率を算出できるLabeled LDA[12]を用いる。Labeled LDA(L-LDA)は、RamageらによってLDAを教師あり学習へ拡張したトピックモデルである。L-LDAは文書にあらかじめ付与されているラベルを、その文書の内容を表すものととらえることで、潜在トピックの抽出における教師ラベルとして利用することを考えたモデルである。

トピックと局面の関連の強さを表す関連度を算出する。トピック $t$ と局面 $a$ の関連度 $R(a, t)$ は、

$$R(a, t) = \sum_{w \in W} p(a, w) * p(w, t)^g \quad (1)$$

で算出する。ここで、 $p(w, t)$ は、LDAを用いて抽出したトピック $t$ における単語 $w$ の生起確率である。この式は、単語 $w$ の局面 $a$ における生起確率と、単語 $w$ のトピック $t$ における生起確率を用いて、局面 $a$ とトピック $t$ の関連度を算出する。 $g$ は、トピック中の単語の生起確率の広がり調整するためのパラメータである。

関連度を0から1の範囲とするために正規化する。ここでは、式(2)に示す、各局面で正規化した関連度 $\hat{R}a(a, t)$ と、各トピックで正規化した関連度 $\hat{R}t(a, t)$ を用意する。

$$\hat{R}a(a, t) = \frac{R(a, t)}{\sum_{t \in T} R(a, t)}, \quad \hat{R}t(a, t) = \frac{R(a, t)}{\sum_{a \in A} R(a, t)} \quad (2)$$

ここで、 $T$ はLDAで抽出したすべてのトピック、 $A$ はす

<sup>\*5</sup> 形態素解析の結果に基づいて、名詞と動詞、形容詞のみを使用する。

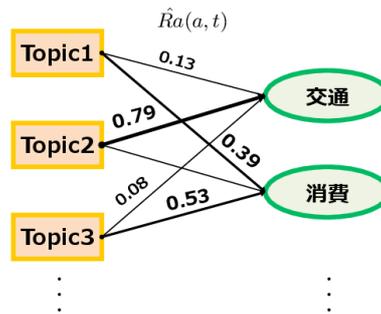


図2 対応関係の例

Fig. 2 Association examples.

べての局面である。 $\hat{R}a(a, t)$ は、局面 $a$ がどのトピックによって支持されているかを表す指標であり、 $\hat{R}t(a, t)$ は、トピック $t$ がどの局面を支持しているかを表す指標である。

関連度 $\hat{R}a(a, t)$ を用いてトピックと局面の対応関係を構築する。局面とトピックの対応関係の例を図2に示す。この図は、交通の局面はトピック2から0.79の高い関連度を持っており、消費の局面はトピック1とトピック3から0.39と0.53と中程度の関連度を持っている。このような場合、交通の局面はトピック2と対応関係を構築し、消費の局面はトピック1と3の2つのトピックと対応関係を構築する。以上のような対応関係をすべてのトピックとすべての局面の間で構築するために、局面ごとに閾値を決定し、関連度 $\hat{R}a(a, t)$ が閾値を超えたトピックと対応関係を構築する。局面 $a$ と対応関係を構築するトピック集合 $T_a$ は、媒介変数 $d$ を用いて、

$$T_a = \{t | \hat{R}a(a, t) > \max_{t \in T} (\hat{R}a(a, t)) - \sigma(\hat{R}a(a, T)) * d\} \quad (3)$$

とする。ここで、 $\sigma(\hat{R}a(a, T))$ は、局面 $a$ の全トピック $T$ に対する関連度の標準偏差である。 $d$ はトピックとの対応関係を調整するためのパラメータであり、 $d$ を大きくすることで、より多くのトピックが局面に関連付けられる。

### 3.3 局面の推定方法

未知のツイートの局面を推定するため、3.2節で構築したトピックと局面の対応関係を用いる。局面の推定方法の概要を図3に示す。ツイートから単語を抽出し、トピック中の各単語の生起確率と、トピックと局面の関連度を用いて、局面ごとにスコアを算出する。未知のツイート $tw$ と各局面 $a$ とのスコア $S(tw, a)$ は、以下の式で算出する。

$$S(tw, a) = \sum_{t \in T_a} \sum_{w \in W_{tw}} p(w, t)^g * \hat{R}a(a, t) * \hat{R}t(a, t) \quad (4)$$

ここで、 $W_{tw}$ は未知のツイート $tw$ から抽出した単語集合、 $p(w, t)$ はトピック $t$ における単語 $w$ の生起確率である。

$\hat{R}a(a, t)$ は局面にとって重要となるトピックに対して高い関連度を与えるが、複数の局面が同じトピックに対して高い関連度を持つ場合もある。たとえば、動詞は多くの局

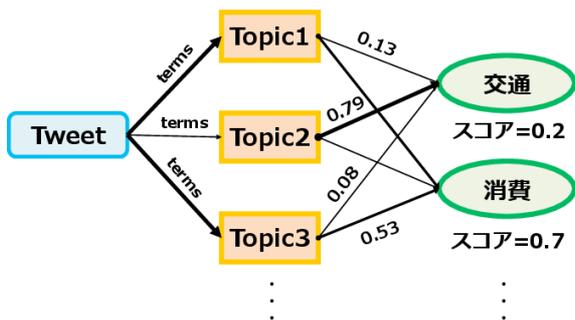


図 3 局面の推定方法  
Fig. 3 Aspects estimation method.

面において特徴的な語になりうるため、動詞が生起確率上位に集中しているトピックは、 $\hat{R}a(a, t)$  の値が複数の局面から高い関連度として与えられる。このようなトピックの存在は、局面の推定精度を低下させる要因となりうる。

一方、 $\hat{R}t(a, t)$  は局面にとって特徴的なトピックに対して高い関連度を与えるため、局面間で共有するトピックには低い関連度を与える。しかし、特定の局面間で共有されるべきトピックに対しても低い関連度を与えてしまう。たとえば、洪水は大雨が原因で発生することから、同じ文書で言及されやすいため、洪水と大雨に関連する単語は同じトピックに集まることが考えられる。表 1 より洪水は災害の局面に含まれ、大雨は気象の局面に含まれるために、洪水と大雨に関するトピックは両局面で共有されるべきトピックであるが、 $\hat{R}t(a, t)$  は両局面に対して低い関連度となる。

スコア算出式 (4) では、 $\hat{R}a(a, t)$  と  $\hat{R}t(a, t)$  の関連度を考慮したスコアとするために、両者の積をとることとした。

未知のツイートに対して算出されたスコアが高い局面ほど、推定すべき局面であり、動的に決定する上位  $k$  局面をツイートに付与する。本論文では、算出されたスコアの平均値と標準偏差を用いて各スコアを正規化し、閾値  $r$  を超えたスコアを持つ局面を、未知のツイートに対して推定する。未知のツイート  $tw$  に対して推定する局面集合  $A_{tw}$  は、

$$A_{tw} = \left\{ a \mid \frac{S(tw, a) - E(S(tw, A))}{\sigma(S(tw, A))} > r \right\} \quad (5)$$

とする。ここで、 $E(S(tw, A))$  と  $\sigma(S(tw, A))$  は、スコア  $S(tw, a)$  の全局面  $A$  における平均値と標準偏差である。各スコアについて、平均値との差を標準偏差で除することで、平均を 0、標準偏差を 1 とした値に正規化する。スコアが閾値  $r$  よりも大きければ、そのスコアを持つ局面は未知のツイートに対する局面として推定する。

### 3.4 トピック数の最適化方法

LDA では、生成するトピック数をパラメータとして与える。3.2 節で説明したように、関連度に基づいてトピックと局面の対応関係が構築されることから、生成するトピック数によって、局面と結びつくトピックが変動する。

最適なトピック数を決定するため、JS Divergence を用いて、ある 1 つの局面と他の局面との類似度を計算する。2 つの局面の確率分布が同じである場合、JS Divergence は 0 となり、より異なっている場合、JS Divergence は 1 に近づく。本論文の場合は、局面間の確率分布  $\hat{R}a(a, t)$  が異なっている方が望ましい。そのため、各局面間の JS Divergence の合計値を最大とするトピック数を最適であるとした。JS Divergence の合計値  $JS_{sum}$  は、以下の式で求められる。

$$JS_{sum} = \sum_{(\forall p, \forall q) \in A} D_{JS}(\hat{R}a(p, *) || \hat{R}a(q, *)) \quad (6)$$

$$D_{JS}(P || Q) = \frac{1}{2} \left( \sum_{t \in T} P(t) \log \frac{P(t)}{R(t)} + \sum_{t \in T} Q(t) \log \frac{Q(t)}{R(t)} \right)$$

ここで、 $R(t)$  は確率分布  $P(t)$  と  $Q(t)$  の平均であり、 $R(t) = \frac{P(t)+Q(t)}{2}$  で与えられる。

## 4. 評価実験

提案手法の有効性を明らかにするために、適合率、再現率、F 値を評価する。比較対象として、Labeled LDA, SVM, および Naive Bayes 分類器をマルチラベリングへ拡張した NBML を用いる。また、トピックと局面の対応関係について分析することで、提案手法が有効に機能する局面と機能しない局面の違いを明らかにする。

### 4.1 データセットとパラメータ設定

#### 4.1.1 データセット：トピック抽出に用いるツイート

提案手法の第 1 段階における LDA を用いたトピック抽出のため、2012 年 4 月 15 日から 2012 年 8 月 14 日の間に、日本語で Twitter に投稿されたツイートを Search API \*6 を用いて収集した。その中から、ツイートのロケーション情報に「京都」あるいは「Kyoto」と入力されているツイートを使用した。以上の条件により評価に使用するツイートは、2,390,553 件となった。

#### 4.1.2 データセット：実生活ツイート

トピックと局面の対応関係を構築するため、人手によってツイートに局面をラベル付けした。1,500 件のツイートに対して、第 1 著者（実験参加者 A）と他 2 名（実験参加者 B および C）の合計 3 名で、各ツイートに対して適切な局面を付与する人手判定を行った。実験参加者にはガイドラインとして、表 1 に示す各局面に含まれる典型的な単語と、その局面に分類されるツイートの例（各局面 1 件ずつ）と、それが分類された理由を提示した。人手判定では、各ツイートに対して適切な局面として第 1、第 2、第 3 候補まで付与することとした。いずれの局面にも適さないと判断した場合は、「非実生活」を付与することとした。いずれの候補にも該当しなかった局面はすべて第 4 候補とした。なお、用意した 1,500 件のツイートは、いずれもロケーショ

\*6 <https://dev.twitter.com/docs/api/1/get/search>

表 2 人手判定の結果, 正解ラベルとして付与された局面の数

Table 2 The number of labeling aspects as correct.

局面	服飾	交流	災害	食事	行事	消費	健康	趣味	居住	地域	学校	交通	気象	労働	非実	合計
ラベル数	181	379	86	287	311	435	177	348	213	432	195	169	226	262	1,391	5,092

ン情報に「京都」あるいは「Kyoto」と表記されたものであり, 上記のトピック抽出に使用するツイートと重複はない. また, 3名の実験参加者はいずれも「つくば市」在住の大学生であり, 京都で生活したことはない.

人手判定の結果, 第1候補に分類された局面について, 実験参加者間の一致度を  $\kappa$  値 [3] によって評価した. 実験参加者 A と実験参加者 B の  $\kappa$  値は 0.687, 実験参加者 A と実験参加者 C の  $\kappa$  値は 0.595, 実験参加者 B と実験参加者 C の  $\kappa$  値は 0.576 となった.  $\kappa$  値の平均は 0.619 であり, 高い一致 (substantial) であった.

各ツイートに対して複数の局面をラベル付けするため, 3名の手判定の結果を用いる. ツイート  $tw$  に対して正解となる局面集合  $AC_{tw}$  は,

$$AC_{tw} = \{a | Uscore(tw, a) \leq 10\} \quad (7)$$

とする. ここで,  $Uscore(tw, a)$  は, 実験参加者がツイート  $tw$  に対して, 局面  $a$  を第何候補に選択したかを合計した値であり, 以下の式で求められる.

$$Uscore(tw, a) = \sum_{u \in U} candidate(tw, a, u) \quad (8)$$

ここで,  $U$  はすべての実験参加者を表し,  $candidate(tw, a, u)$  は, ツイート  $tw$  に対して実験参加者  $u$  が局面  $a$  を分類した候補番号である. 実験参加者がいずれの候補にも入れなかった局面は, 値を 4 として計算する.  $Uscore(tw, a)$  の最小値は, 実験参加者 3 名が同じ局面を第1候補に選択した場合であり,  $candidate(tw, a, u) = 1$  となるため,  $Uscore(tw, a) = 3$  となる. 最大値は, 実験参加者が特定の局面をいずれの候補にも選択しなかった場合, すなわち,  $candidate(tw, a, u) = 4$  の場合で, そのとき  $Uscore(tw, a) = 12$  となる.

以上の処理によって, 人手判定した 1,500 件のツイートに対して, ラベル付けされた局面数を集計した結果を表 2 に示す. 服飾の局面は, 1,500 件のツイートの中で合計で 181 件ラベル付けされている. 1,500 件のツイートに対するすべてのラベル数は 5,092 件となっており, 1つのツイートに対して平均 3.39 件のラベルが付与されている. 評価実験では, いずれの局面にも属さない「非実生活」についても 1つのクラスとしてトピックと対応関係を構築し, 非実生活を推定できるか否かを評価する.

正解データとして使用する 1,500 件のツイートについて, 各ツイートに付与された局面の数を集計した結果を表 3 に示す. 最も多いラベル数は 3 で, 820 件のツイートが存在する. ラベル数が 6 あるツイートは, 11 件存在する.

表 3 ラベル数別のツイート数

Table 3 The number of labeling by examinee.

ラベル数	1	2	3	4	5	6	合計
ツイート数	1	111	820	442	115	11	1,500

表 4 各トピック数における  $JS_{sum}$  値

Table 4  $JS_{sum}$  in each the number of topics.

トピック数	50	100	200	500	1,000
$JS_{sum}$	294.17	386.34	497.51	<b>510.03</b>	454.21

#### 4.1.3 パラメータ設定

LDA は, 事前にいくつかのパラメータを設定する必要がある. 関連研究 [6] を参考に, ハイパーパラメータである  $\alpha$  は  $\frac{50}{|T|}$ ,  $\beta$  は 0.1 とした.  $|T|$  は LDA で生成するトピック数である. イテレーション回数は, 予備実験の結果から安定した値が得られる 100 とした. また, 提案手法の関連度算出式 (1) やスコア算出式 (4) のパラメータ  $g$  は, 予備実験の結果から  $g = 0.3$  とした.

LDA で生成するトピック数は, 3.4 節で説明した  $JS_{sum}$  の値が最大値となるトピック数を選択する. トピック数を 50, 100, 200, 500, 1,000 と変化させ, 各トピック数で  $JS_{sum}$  を算出した結果を表 4 に示す.  $JS_{sum}$  が最大となったトピック数 500 を最適なトピック数とし, トピックと局面の対応関係を構築した.

#### 4.2 評価尺度

提案手法の有効性を議論するには, 推定した局面がどれだけ正解しているかという正確性と, すべての正解のうちどれだけ提案手法で局面を推定できたかという網羅性の, 2つの観点からの評価が必要となる. 本論文では, 正確性を適合率 (Precision), 網羅性を再現率 (Recall), 適合率と再現率の調和平均である F 値 ( $F$ -measure) によって提案手法の推定精度を評価する. 実験では, 1,500 件の正解データについて, 10 分割交差検定によってトピックと局面の対応関係を構築, および推定精度の評価を行う. 用意した 1,500 件の正解データのうち, 訓練データとして 9 割 (1,350 件) を使用して Labeled LDA を学習し, 残りの 1 割 (150 件) をテストデータとして推定精度を評価する.

#### 4.3 比較手法

##### 4.3.1 L-LDA

Labeled LDA は, マルチラベルが付与された訓練データを用いて, 各クラスにおける単語の生起確率を算出でき

る。また、単語の生起確率を用いてテストデータに対するマルチラベリングが可能であり、Web ページの分類においては、SVM を上回る性能を示している [12]。提案手法は Labeled LDA で算出した単語の生起確率を用いていることから、Labeled LDA は提案手法のベースライン手法と考えることができる。Labeled LDA のイテレーション回数は、いずれの訓練データ集合に対しても安定した値が得られる 100 とした。ハイパーパラメータである  $\alpha$  と  $\beta$  は、ともに 0.1 とした。

### 4.3.2 SVM

SVM (Support Vector Machine) は、教師あり学習を用いる識別手法の 1 つであり、現在知られている多くの手法の中でも、高い分類性能と汎化能力を有している [4]。本論文では、マルチラベル分類\*7 も可能である SVM 用のライブラリの 1 つである LIBSVM を用いる [2]。

カーネルは線形カーネルを選択し、パラメータは、LIBSVM のツールでグリッドサーチ\*8 を実行し、 $C = 1.0$  とした。SVM に与える素性は提案手法と同様に、ツイートを形態素解析した結果得られた、名詞、動詞、形容詞の品詞に該当する単語の Bag of Words とした。

### 4.3.3 NBML

比較対象の 1 つとして、Naive Bayes Multi-label Classification (NBML) [14] を用いる。NBML は Wei らによって、Naive Bayes 分類器 [5] をマルチラベル分類へ拡張された手法である。Naive Bayes 分類器は、テキスト中に含まれる単語が互いに独立に発生したものであるという仮定をおき、それらの単語が出現したときの文書のクラスへの所属確率をベイズの定理により求め、所属確率が最も高いクラスへ文書を分類する手法である。

NBML は、クラス別の所属確率から平均値を求め、平均値を超える所属確率を持つクラスへ文書を分類することで、マルチラベリングを実現している。NBML は単語の生起確率を算出する際に、各文書に対して 1 つのラベルが付与されていることを前提としているため、4.1.2 項で説明したような複数ラベルが付与された文書に対応していない。そこで、本論文では文書に付与されているラベルを 1 つにするため、ラベルごとに同じ文書を複製する。つまり、ラベルの数だけ文書を複製するため、表 2 に示しているように、すべてのラベルを合計した 5,092 件のツイートによって訓練する。

## 4.4 実験結果

### 4.4.1 各局面に結びつくトピック数の遷移

すべての局面に結びつくトピック数が最大の 500 となる

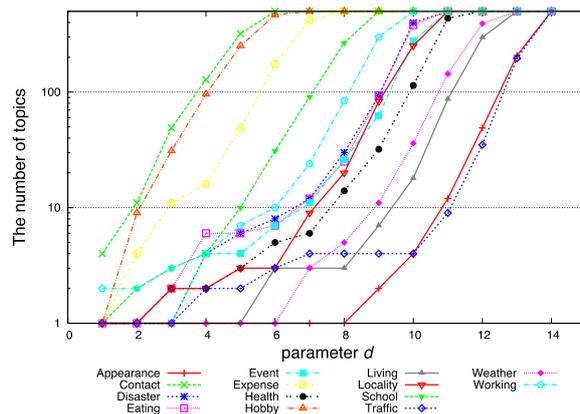


図 4 各局面に結びつくトピック数の遷移  
Fig. 4 Connectivity among topics and aspects.

まで、パラメータ  $d$  を 1 ずつ変化させた。各局面に対して結びつくトピック数の遷移を図 4 に示す。すべての局面で、パラメータ  $d$  の増加にともなって、結びつくトピック数が増加している。服飾の局面は  $d \leq 8$  まで結びついているトピック数が 1 つである。  $d = 14$  となったとき、すべての局面で結びついているトピック数が 500 となり、トピックと局面の対応関係は完全二部グラフとなっている。

### 4.4.2 トピックと局面の対応関係

トピックと局面の対応関係の詳細を表 5 に示す。表では、局面ごとに  $\hat{R}a(a, t)$  の高い上位 4 トピックを抽出し、そのトピック番号と関連度  $\hat{R}a(a, t)$  と  $\hat{R}t(a, t)$  をそれぞれ示している。服飾 (App.) の局面に対して最も強く結びつくトピックは、Topic119 であり、 $\hat{R}a(App., Topic119) = 0.023$ ,  $\hat{R}t(App., Topic119) = 0.667$  である。

太字のトピック番号は、他の局面とも上位 4 トピック中で結びついていることを示し、Topic60 は地域と交通の局面に、Topic178 は災害と気象の局面に、Topic281 は交流と非実生活の局面に強く結びついていた。

### 4.4.3 提案手法の最適なパラメータ

提案手法は式 (5) に示すように、パラメータ  $r$  によってツイートに付与する局面を決定する。最適なパラメータ  $r$  を選択するため、0.0, 0.5, 1.0 と  $r$  を切り替え、それぞれの適合率、再現率、F 値を評価した。10 分割交差検定をした平均を算出した後、F 値が最大となるパラメータ  $d$  を選択した。結果を表 6 に示す。適合率では  $r = 1.0$ 、再現率と F 値では  $r = 0.0$  が最大となった。

### 4.4.4 提案手法の評価値の遷移

提案手法で  $r = 0.0$  としたときの、トピック数の遷移にともなう適合率、再現率、F 値を評価した。学校と交通の局面の評価結果を、それぞれ図 5 と図 6 に示す。いずれの図でも、横軸はパラメータ  $d$ 、左の縦軸は評価値、右の縦軸はトピック数である。学校の局面では、 $d$  の増加にともなって再現率は上昇し、 $d \geq 8$  から急激に下降に転じている。適合率と F 値は、 $d = 1$  で最大となっている。交通

\*7 LIBSVM Tools: Multi-label classification, <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/multilabel/>

\*8 LIBSVM Tools: Grid parameter search for regression, [http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#grid\\_parameter\\_search\\_for\\_regression](http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#grid_parameter_search_for_regression)

表 5 各局面に対して高い関連度  $\hat{R}a$  で結びつくトピック

Table 5 Relevance of high  $\hat{R}a$  topics.

局面	$\hat{R}a1$ 位			$\hat{R}a2$ 位			$\hat{R}a3$ 位			$\hat{R}a4$ 位		
	topic	$\hat{R}a$	$\hat{R}t$									
服飾	#119	0.023	0.667	#474	0.009	0.399	#240	0.009	0.385	#454	0.007	0.170
交流	#49	0.007	0.246	#429	0.007	0.167	#281	0.007	0.133	#247	0.007	0.317
災害	#178	0.020	0.234	#469	0.014	0.205	#277	0.013	0.404	#380	0.013	0.253
食事	#341	0.024	0.795	#484	0.018	0.526	#352	0.017	0.574	#197	0.016	0.546
行事	#345	0.023	0.303	#314	0.016	0.214	#190	0.015	0.297	#18	0.014	0.307
消費	#35	0.011	0.269	#132	0.010	0.191	#437	0.009	0.261	#419	0.008	0.182
健康	#393	0.023	0.445	#22	0.019	0.582	#193	0.014	0.395	#348	0.012	0.212
趣味	#412	0.008	0.490	#273	0.007	0.166	#430	0.007	0.359	#75	0.007	0.186
居住	#290	0.019	0.563	#230	0.010	0.284	#133	0.010	0.344	#301	0.007	0.343
地域	#60	0.019	0.279	#314	0.015	0.258	#44	0.010	0.182	#486	0.008	0.147
学校	#111	0.014	0.643	#3	0.012	0.403	#418	0.011	0.338	#118	0.009	0.538
交通	#60	0.032	0.369	#201	0.025	0.582	#149	0.020	0.455	#42	0.016	0.440
気象	#23	0.020	0.480	#451	0.010	0.207	#178	0.010	0.250	#230	0.008	0.299
労働	#321	0.017	0.468	#436	0.016	0.286	#253	0.011	0.208	#334	0.011	0.221
非実	#281	0.005	0.142	#432	0.005	0.184	#216	0.005	0.185	#273	0.005	0.189

表 6  $r$  を変化させたときの適合率, 再現率, F 値

Table 6 Precision, Recall and F-measure are evaluated by  $r$  varying.

$r$	Precision	Recall	F-measure
0.0	0.57	0.66	*0.59
0.5	0.61	0.57	0.58
1.0	0.68	0.54	0.58

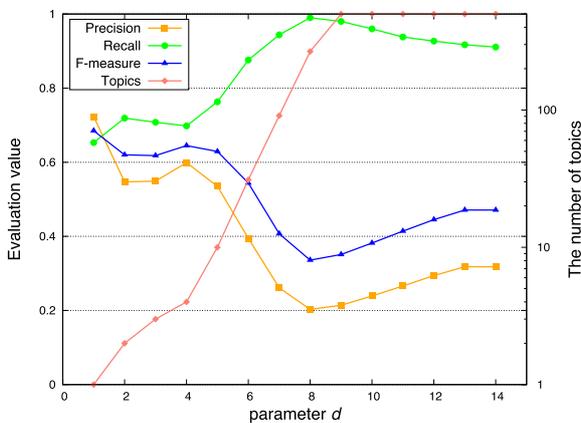


図 5 学校の適合率, 再現率, F 値

Fig. 5 Precision, Recall and F-measure of School.

の局面では, いずれの評価値についても  $d = 4$  で上昇しており, そのときのトピック数は 2 である. 適合率と F 値は,  $d = 10$  で最大となっており, 再現率は  $d = 14$  で最大となっている.

#### 4.4.5 各手法の推定精度

各手法の適合率, 再現率, F 値を表 7 に示す. 10 分割交差検定をし, それぞれの局面で平均を算出した. 提案手法については,  $r = 0.0$  とし F 値を最大とするときのパラメータ  $d$  を選択した. なお, 本来はパラメータ決定と性能評価

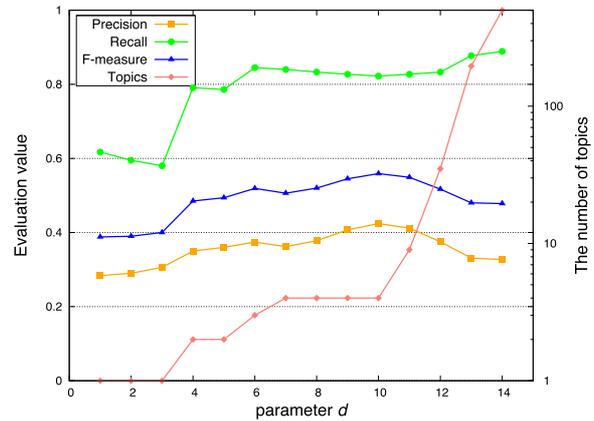


図 6 交通の適合率, 再現率, F 値

Fig. 6 Precision, Recall and F-measure of Traffic.

では異なるデータセットを用いるべきであるが, 表 6 から明らかな差がないように, F 値は  $r$  の値にほとんど影響を受けないことから, ここで得られる  $r$  の最適値  $r = 0.0$  を用いて以降の評価を行うこととした. また,  $d$  は値は暫定値であるが, 適切な  $d$  を設定することで達成できる推定精度であると考えられる. 各局面に対応する  $d$  とトピック数も, 表の項目として示している. 左に「\*」のある数値は, すべての手法で比較したときの最大値である. 服飾の局面では, 適合率の最大値は 0.83 で NBML, 再現率の最大値は 0.69 で L-LDA, F 値の最大値は 0.63 で提案手法となっており,  $d = 12$  で服飾の局面に結びつくトピック数は 49 である. すべての局面についてマクロ平均をとった結果, 各手法で最大値を示したのは, 適合率では NBML, 再現率と F 値で提案手法であった.

提案手法と比較手法について, 各ツイートに対して推定したラベル数を集計した結果を表 8 に示す. ラベル数

表 7 各手法の適合率, 再現率, F 値  
Table 7 Precision, Recall and F-measure of each method.

局面	Precision				Recall				F-measure				<i>d</i>	topics
	提案法	L-LDA	SVM	NBML	提案法	L-LDA	SVM	NBML	提案法	L-LDA	SVM	NBML		
服飾	0.72	0.43	0.64	<b>*0.83</b>	0.57	<b>*0.69</b>	0.28	0.37	<b>*0.63</b>	0.52	0.38	0.51	12	49
交流	0.37	0.43	0.41	<b>*0.53</b>	<b>*0.68</b>	<b>*0.68</b>	0.35	0.54	0.48	<b>*0.53</b>	0.37	<b>*0.53</b>	1	4
災害	0.34	0.67	0.44	<b>*0.76</b>	<b>*0.59</b>	0.49	0.44	0.21	0.43	<b>*0.54</b>	0.44	0.33	3	1
食事	<b>*0.77</b>	0.41	0.51	0.73	0.75	<b>*0.77</b>	0.64	0.51	<b>*0.76</b>	0.53	0.57	0.60	6	7
行事	0.42	0.47	<b>*0.56</b>	<b>*0.56</b>	<b>*0.57</b>	0.51	0.20	0.45	<b>*0.49</b>	0.48	0.29	<b>*0.49</b>	8	26
消費	0.52	<b>*0.64</b>	0.43	0.52	<b>*0.55</b>	0.40	0.45	0.46	<b>*0.53</b>	0.49	0.43	0.49	5	49
健康	0.37	0.43	0.48	<b>*0.76</b>	<b>*0.63</b>	0.55	0.28	0.38	0.47	0.48	0.35	<b>*0.50</b>	14	500
趣味	0.38	0.44	0.43	<b>*0.57</b>	<b>*0.77</b>	0.62	0.54	0.44	<b>*0.51</b>	<b>*0.51</b>	0.48	0.49	13	500
居住	0.58	0.38	0.64	<b>*0.71</b>	0.57	<b>*0.62</b>	0.34	0.41	<b>*0.57</b>	0.46	0.44	0.51	11	87
地域	<b>*0.81</b>	0.62	0.62	0.62	0.58	<b>*0.73</b>	0.54	0.65	<b>*0.67</b>	<b>*0.67</b>	0.57	0.63	13	500
学校	0.73	0.37	<b>*0.88</b>	0.81	0.65	<b>*0.81</b>	0.36	0.52	<b>*0.68</b>	0.51	0.49	0.63	1	1
交通	0.43	0.33	0.71	<b>*0.82</b>	<b>*0.82</b>	<b>*0.82</b>	0.44	0.50	0.56	0.47	0.54	<b>*0.62</b>	10	4
気象	0.69	0.25	0.47	<b>*0.81</b>	0.70	<b>*0.84</b>	0.63	0.58	<b>*0.70</b>	0.38	0.53	0.67	8	5
労働	0.44	<b>*0.64</b>	0.52	0.56	<b>*0.60</b>	0.50	0.19	0.35	0.51	<b>*0.55</b>	0.28	0.43	8	84
非実	0.93	<b>*0.94</b>	0.93	0.93	0.93	0.51	<b>*0.99</b>	0.93	0.93	0.66	<b>*0.96</b>	0.93	4	388
平均	0.57	0.50	0.58	<b>*0.70</b>	<b>*0.66</b>	0.63	0.44	0.49	<b>*0.59</b>	0.52	0.47	0.56		

表 8 各手法のラベル数別のツイート数  
Table 8 The number of labeling by each method.

ラベル数	提案法	L-LDA	SVM	NBML	人手判定
1	20	101	0	165	1
2	23	154	137	531	111
3	83	259	1,250	442	820
4	277	369	80	243	442
5	504	307	33	90	115
6	425	182	0	23	11
7	143	80	0	6	0
8	25	37	0	0	0
9	0	9	0	0	0
10	0	2	0	0	0
平均	5.13	4.16	3.00	2.75	3.39

の最頻値は提案手法では5, L-LDAでは4, SVMでは3, NBMLでは2であった。ツイートあたりの平均ラベル付与数は, 提案手法では5.13, L-LDAでは4.16, SVMでは3.00, NBMLは2.75となり, 人手判定の結果に最も近い平均ラベル付与数を示したのはSVMであった。

#### 4.4.6 訓練データが少ない場合の推定精度

各手法について, 訓練データを減じた場合の推定精度を評価した。10分割した正解データから, 1個(150件)を正解データとし, 余った9個から訓練データとして1, 3, 5, 7個をランダムに抽出し, 正解データを変えながら10回繰り返して平均をとる。それぞれの評価値について算出した結果を図7に示す。提案手法のパラメータ  $r$  と  $d$  は4.4.5項と同様にして決定した。また, 提案手法は訓練データの変更にもとない, 3.4節に示す最適なトピック数を  $JS_{sum}$  を再計算する必要がある。実際には, 訓練データ

の割合を変えても  $JS_{sum}$  は500トピックのときに最も高くなったため, いずれの割合でもLDAで抽出するトピック数は500として評価をした。

提案手法とNBMLの適合率は, 訓練データが9個のときは大きな差があるが, 割合を下げていくとともに, 2つの手法の適合率の差は小さくなり, 訓練データ数が少ない場合でも提案手法の適合率はほとんど低下しないという結果となった。再現率では, L-LDAとSVMでは訓練データを減じていくとともに大きく下降しているが, 提案手法とNBMLではほとんど変化しなかった。F値では, 提案手法は訓練データが3個までは大きな変化はなく, 1個にしたときにやや下降したが, 他の手法と比較してつねに最大となっていた。SVMは訓練データを減じていくとともに大きく下降した。

## 5. 考察

### 5.1 トピック局面の対応関係に関する考察

表7より, 提案手法が他の手法に比べ高いF値を示した局面は, 服飾, 食事, 行事, 消費, 趣味, 居住, 地域, 学校, 気象である。この中でも特に服飾, 食事のF値は他の手法に比べて大きな差があることが分かる。この2つの局面について, 高い関連度で結びついているトピックに着目すると, 表5より, 服飾の  $\hat{R}a1$  位はTopic119, 食事の  $\hat{R}a1$  位はTopic341である。それぞれのトピックの上位語を10語ずつ取り出したものを表9に示す。Topic119, Topic341のいずれについても, それぞれの局面に典型的な語が存在することが分かる。このように, 局面に特徴的な語が上位に集中しているトピックは, 関連度  $\hat{R}t$  が他のトピックに比べて高くなっている。趣味や居住, 学校, 気象の局面に

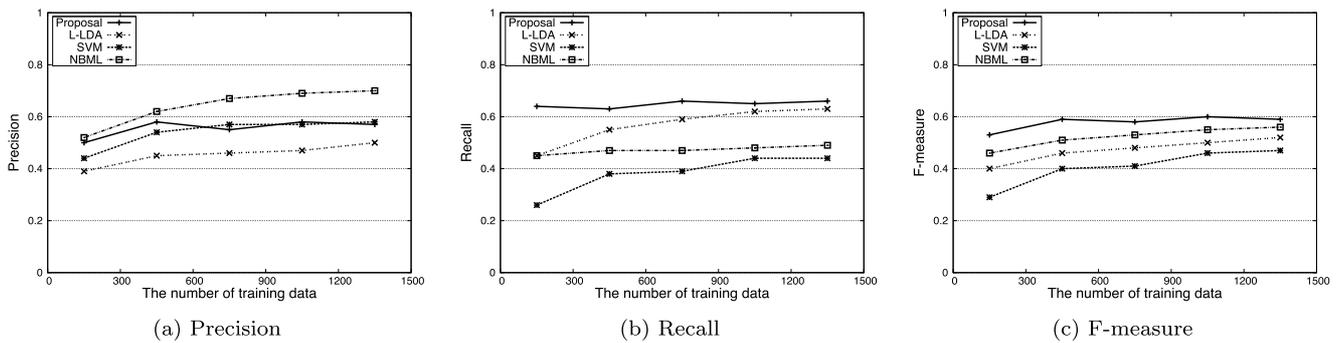


図 7 訓練データの数を变化させた評価値

Fig. 7 Precision, Recall and F-measure are evaluated by varying the number of training data.

表 9 高い関連度で結びつくトピックの上位語

Table 9 High occurrence probability terms.

トピック	上位語
Topic119 服飾 $\hat{Ra}1$ 位	着る, 浴衣, 似合う, T シャツ, デート, スーツ, ピンク, 衣装, シャツ, ワンピース
Topic341 食事 $\hat{Ra}1$ 位	カレー, 野菜, トマト, 美味しい, 作る, スープ, サラダ, ソース, 食べる, おいしい
Topic67 交通 $\hat{Ra}2$ 位	電車, 乗る, 阪急, 降りる, 遅れる, 特急, 快速, 車両, 乗車, 列車

表 10 多くの局面に対して結びつくトピックの上位語

Table 10 High occurrence probability terms.

トピック	上位語
Topic60 地域, 交通	京都, 交通, 河原町, 四条, 三条, 地下鉄, 烏丸, 便利, 嵐山, 案内
Topic178 災害, 気象	台風, 京都, 警報, 大雨, 降る, 地方, 暴風, 影響, 発表, 近畿
Topic281 交流, 非実	今日, 明日, 行く, 終わる, 仕事, がんばる, 済ませる, 早い, 用事, 夕方

についても、服飾や食事の局面と同様の理由で、推定精度が高くなったと考えられる。

地域の局面も高い F 値を示しているが、 $\hat{Ra}1$  位で結びつく Topic60 の  $\hat{R}t$  は、Topic119 や Topic341 に比べ低く、交通の局面に対しても  $\hat{Ra}1$  位で結びついている。表 10 より、Topic60 の上位語を見ると、「京都」や「河原町」、「四条」などの地名に関する語と、「交通」や「地下鉄」、「案内」などの交通に関する語がともに存在している。したがって、Topic60 は「京都特有の交通」に関するトピックであると考えられる。このようなトピックを関連度  $\hat{Ra}$  によって地域と交通の局面で共有できたために、地域の局面で推定精度が高くなったと考えられる。行事や消費の局面でも、他の局面とトピックを共有することで高い推定精度が得られたと考えられる。

図 6 より、交通の局面では  $d = 4$  で結びつくトピック数が 2 つになり、すべての評価値が上昇している。表 5 より、トピック数が 2 つのときは Topic60 と Topic67 が結び

つく。Topic60 は先ほど確認したとおり、「京都特有の交通」に関するトピックであるが、Topic67 は表 9 より、交通の局面に特徴的な語が上位に集中していることが確認できる。交通の局面において、Topic60 は局面を表現するうえで必要であるが、Topic67 の方がより広範かつ特徴的に交通の局面を表現できているために、すべての評価値が上昇したと考えられる。

災害や交流の局面は、他の手法に比べて推定精度が低い結果となった。災害の局面では、Topic178 が  $\hat{Ra}1$  位で結びついており、Topic178 は気象の局面に対しても  $\hat{Ra}3$  位で結びついている。表 10 より、Topic178 は「台風」に関連するトピックであることが分かる。台風は洪水や家屋の崩壊など、二次被害を及ぼしやすいことから、災害の局面について言及されやすく、また気象情報とともに言及されやすいことから、気象の局面についても言及されやすいトピックである。災害と関連付けられた  $\hat{Ra}2$  位以下のトピックはいずれも関連度が低く、このトピックのみでは災害に関する話題を適切に表現できなかったために、災害の局面は推定精度が低くなったと考えられる。

## 5.2 各手法の推定精度に関する考察

表 7 より、各手法の推定精度を比較すると、適合率では NBML、再現率では提案手法と L-LDA、F 値では提案手法と NBML が他の手法に比べ高い値を示した。表 8 より、提案手法は L-LDA に比べると多くのラベルを付与するが、L-LDA に比べ適合率と再現率が高い値を示した。提案手法は L-LDA に比べ多くの正解を推定できていることから、局面に対するスコアも適切に計算できていると考えられる。人手判定の平均ラベル付与数に最も近い手法は SVM であるが、SVM は他の手法に比べて F 値が低い。これは、SVM が他の手法に比べ局面に対するスコアを適切に計算できていないことが原因と考えられる。NBML は人手判定の平均ラベル付与数に比べ少ないラベルを付与するために、適合率は高いが再現率が低くなった。

訓練データが少ない場合の推定精度について調査した図 7 より、提案手法は他の手法に比べて、訓練データの数

表 11 提案手法が有効に機能したツイート例

Table 11 Our proposed method effectively performed tweets.

正解局面 (人手)	気象, 健康
NBML が推定した局面	気象
提案手法が推定した局面	気象, 健康
ツイート本文	明日も関西は晴天で黄砂も飛んでこない様子. 最高気温も 28°C くらいまで上がるよう. 外出時には UV ケアもしっかりとして直射日光に気をつけなければ! 熱中症にも気をつけないとね. 水分補給を忘れずに ...

表 12 提案手法が有効に機能したツイート例

Table 12 Our proposed method effectively performed tweets.

正解局面 (人手)	学校, 行事
NBML が推定した局面	学校, 行事
提案手法が推定した局面	学校, 行事, 地域
ツイート本文	【拡散希望】5月31日18時から桂集会を開催します. 開催場所は桂川の近くです. 大学院試験のための勉強会やります. 気軽に連絡ください.

が減っても推定精度の下降が小さいことが確認できる. 特に再現率については, 訓練データの個数が9個では提案法とL-LDAはほとんど差はないが, 訓練データの数が減っていくのにもなって, L-LDAは大きく下降していくが, 提案手法はほとんど下降しない. 提案手法以外の手法では, 訓練データの数が減ると手がかりとできる語が少なくなり, 再現率が下降することが考えられる. しかし, 提案手法は訓練データの数が減っても, トピックという単位で局面と対応関係を構築するために, 手がかり語の数は変化しない. 以上のことから, 提案手法は訓練データの数が減っても再現率がほとんど下がらずに, F値でも他の手法に比べ上回ることができたと考えられる.

提案手法が正確に正解ラベルを推定した例を, 表 11 に示す. 表では, 提案手法と最も近いF値を示したNBMLが推定した局面も示している. ツイートに正解ラベルとして付与されている局面は気象と健康であり, NBMLでは気象の局面のみ推定しているが, 提案手法はこの2つの局面を正確に推定することができている. ツイート文に着目すると, 翌日の天気予報について文章の前半で言及し, 文章の後半では天候によって体調を悪くしないための対策について言及している. NBMLでは, 「晴天」や「黄砂」, 「気温」の単語が気象の局面に対して高い生起確率であったために, 気象に対する所属確率が他の局面に比べて高くなり, 1つの局面のみを推定したと考えられる. 対して, 提案手法では「熱中症」や「水分」といった単語を高い生起確率として持つトピックが, 健康の局面に対して高い関連度を持って結びついていたために, 気象に加えて健康の局面も推定することができたと考えられる.

提案手法が正解ラベルに加えて, 別の局面も推定した例を表 12 に示す. ツイートに正解ラベルとして付与されている局面は学校と行事であり, NBMLではこの2つの局面を正確に推定することができている. 一方, 提案手法はこ

の2つの局面に加えて地域の局面を推定した. ツイート文に着目すると, 「桂川」という単語が存在することが分かる. 「桂川」は京都市内を流れる川の名前であるため, このツイートは, 「桂川」の近くで勉強会を開催するという内容であることが分かる. したがって, このツイートは勉強会を開催する内容であるが, その勉強会が特定の地域に限定したイベントでもあるため, 地域の局面は必ずしも間違ったラベルではないと考えられる. 以上のような短い文章からも, 提案手法はトピックという単位で話題を推定した後, 局面を推定できるために, 様々な局面を想起して推定できているといえる.

## 6. 結論

本論文では, より少ない訓練データで未知のツイートに対して適切な複数の局面を推定する, マルチラベル分類に適した階層的推定法を提案した. 提案手法は, 教師なし学習と教師あり学習を組み合わせているところに特徴がある. 第1段階では, 教師なし学習で文書をクラスタリングできるLDAを用いて, 大量のツイートからトピックを抽出する. 第2段階では, 少量のラベル付きツイートを用いて, トピックと局面の対応関係を構築する. 未知のツイートに対して適切な局面を推定するため, トピックと局面の対応関係と, ツイートから抽出した単語のトピック中の生起確率を用いて, 局面ごとにスコアを算出する. 閾値を超えたスコアを持つ局面を, 他の局面よりもふさわしい局面として, 未知のツイートに対して推定する.

提案手法の有効性を評価するため, 京都市内で投稿された日本語ツイートを用いて評価実験を行った結果, 未知のツイートに対して複数の適切な局面を推定できることを明らかにした. トピックと局面の対応関係について分析した結果, 推定精度が高い局面では, その局面を表現するために必要な単語が集まったトピックが高い関連度で結びつい

ていた。L-LDA, SVM, NBML を比較対象として推定精度を比較評価をした結果、再現率と F 値で提案手法が他の手法に比べて高い値を示した。訓練データの数を減らした場合に、他の手法の推定精度が下降していく中で、提案手法はほとんど下降しないという特徴が明らかになった。以上の結果から、ツイートのような短文に対して、より少ない訓練データでマルチラベリングをする手法として、提案手法が有効であるといえる。

今後の課題は、トピックと局面の対応関係を決定するパラメータ  $d$  を自動的に最適化する手法を明らかにすること、京都以外の地域においても同様の結果が得られることを示すこと、最適なトピック数を決定する  $JS_{sum}$  の値と推定精度の相関を明らかにすることなどである。

謝辞 本研究の一部は、JSPS 科研費 25280110 の助成を受けたものです。ここに記して謝意を示します。

## 参考文献

- [1] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
- [2] Chang, C. and Lin, C.: LIBSVM: A Library for Support Vector Machines, *ACM Trans. Intell. Syst. Technol.*, Vol.2, No.3, pp.1–27 (2011).
- [3] Cohen, J.: A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, Vol.20, No.1, pp.37–46 (1960).
- [4] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Mach. Learn.*, Vol.20, No.3, pp.273–297 (1995).
- [5] Domingos, P. and Pazzani, M.: On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss, *The Journal of Machine Learning Research*, Vol.29, No.2-3, pp.103–130 (1997).
- [6] Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *The National Academy of Science*, Vol.101, pp.5228–5235 (2004).
- [7] Inui, K., Abe, S., Hara, K., Morita, H., Sao, C., Eguchi, M., Sumida, A., Murakami, K. and Matsuyoshi, S.: Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents, *The 2008 International Conference on Web Intelligence*, pp.314–321 (2008).
- [8] Kazawa, H., Izumitani, T., Taira, H. and Maeda, E.: Maximal margin labeling for multi-topic text categorization, *Advances in Neural Information Processing Systems 17*, pp.649–656, MIT Press (2005).
- [9] Kudo, T.: Yet Another Part-of-Speech and Morphological Analyzer (2005), available from <http://mecab.sourceforge.net/>.
- [10] Kurashima, T., Tezuka, T. and Tanaka, K.: Blog Map of Experiences: Extracting and Geographically Mapping Visitor Experiences from Urban Blogs, *Proc. 6th International Conference on Web Information Systems Engineering, WISE'05*, pp.496–503 (2005).
- [11] Mathioudakis, M. and Koudas, N.: TwitterMonitor: Trend Detection over the Twitter Stream, *Proc. 2010 ACM SIGMOD International Conference on Management of Data, SIGMOD '10*, pp.1155–1158 (2010).
- [12] Ramage, D., Hall, D., Nallapati, R. and Manning, C.D.: Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09*, pp.248–256 (2009).
- [13] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proc. 19th International Conference on World Wide Web, WWW '10*, pp.851–860 (2010).
- [14] Wei, Z., Zhang, H., Zhang, Z., Li, W. and Miao, D.: A Naive Bayesian Multi-label Classification Algorithm with Application to Visualize Text Search Results, *International Journal of Advanced Intelligence*, Vol.3, No.2, pp.173–188 (2011).
- [15] Weng, J., Lim, E.-P., Jiang, J. and He, Q.: TwitterRank: Finding Topic-sensitive Influential Twitterers, *Proc. 3rd ACM International Conference on Web Search and Data Mining, WSDM '10*, pp.261–270 (2010).
- [16] Yamamoto, S. and Satoh, T.: Two Phase Extraction Method for Extracting Real Life Tweets using LDA, *The 15th Asia-Pacific Web Conference, APWeb '13*, pp.340–347 (2013).
- [17] Zhang, Y.C., Séaghdha, D.O., Quercia, D. and Jambor, T.: Auralist: Introducing Serendipity into Music Recommendation, *Proc. 5th ACM International Conference on Web Search and Data Mining, WSDM '12*, pp.13–22 (2012).
- [18] Zhao, W.X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P. and Li, X.: Topical Keyphrase Extraction from Twitter, *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11*, pp.379–388 (2011).
- [19] Zhao, Z. and Mei, Q.: Questions About Questions: An Empirical Analysis of Information Needs on Twitter, *Proc. 22nd International Conference on World Wide Web, WWW '13*, pp.1545–1556 (2013).
- [20] 山本修平, 佐藤哲司: Twitter からの実生活情報の抽出法の提案, 第 4 回データ工学と情報マネジメントに関するフォーラム DEIM2013 論文集, F3-4 (2012).
- [21] 水沼友宏, 池内 淳, 山本修平, 山口裕太郎, 佐藤哲司, 島田 諭: Twitter におけるバーストの生起要因と類型化に関する分析, 情報社会学会誌, Vol.6, No.2, pp.69–84 (2013).



山本 修平

2012年筑波大学情報学群知識情報・図書館学類卒業。2014年筑波大学大学院図書館情報メディア研究科博士前期課程修了。修士(情報学)。現在、同研究科博士後期課程に在籍。情報検索、データマイニング等に興味を持つ。日本データベース学会学生会員。



佐藤 哲司 (正会員)

1980年山梨大学工学部電子工学科卒業。同年日本電信電話公社(現NTT)武蔵野電気通信研究所に入所。以来、データベースマシン、マルチメディアデータベース、情報検索・情報共有の高次化等に関する研究・開発に従事。

1994年工学博士(大阪大学)取得。2007年4月より筑波大学大学院図書館情報メディア研究科教授、現在に至る。情報検索・知識発見、社会ネットワーク分析、社会インタラクションに興味を持つ。電子情報通信学会、日本データベース学会各会員。

(担当編集委員 河合 英紀)