二段階抽出法を用いた実生活Tweetのマルチラベル分類

山本 修平[1,a)]　佐藤 哲司[2]

概要：身近な出来事や感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 Tweet が数多く投稿されている．本研究では，これらのような有益な Tweet を抽出するために，二段階抽出法を用いたマルチラベリングを提案する．第一段階では，大量の Tweet に対して教師データを必要としない LDA を用いてトピックを抽出し，第二段階では，ラベル付けされた少量の Tweet を用いてトピックと局面の対応関係を構築する．未知の Tweet に対して局面毎にスコアを算出し，スコアの分布の平均と標準偏差を用いて閾値を決定し，スコアが閾値を超えた複数の局面を動的に付与する．プロトタイプシステムを実装・評価を行い，未知の Tweet に対して複数の局面を柔軟に付与できることを明らかにした．

キーワード：Twitter，実生活，LDA，二段階抽出法，マルチラベリング

Multi-labeling Classification for Real Life Tweets using Two Phase Extraction Method

Yamamoto Shuhei[1,a)]　Satoh Tetsuji[2]

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．Twitter[*1] は，最も広く普及しているマイクロブログであり，最大 140 文字の短い文章からなる膨大な記事が日々投稿されている．Twitter では，多くのユーザがリアルタイムに，自分の経験や意見，また日常生活の中のイベントなど，身近な「今」を投稿しているため，最新かつ有益な記事が多い．例えば，電車の遅延情報やスーパーの特売情報といった，地域性が高く新鮮な記事がある．著者らは，このような記事を「実生活 Tweet」と呼び．大量の Tweet の中から実生活 Tweet を抽出することを試みてきている [11][12].

実生活 Tweet が実際にユーザの生活を支援した例として，2011 年 3 月に起きた東日本大震災が知られている [10]. 地震が起きた直後，被災地では断水や食料供給の不足，電車の運行中止など，大きな混乱が生じた．その際，給水や食料配布が行われる場所，電車の運行情報などについて書かれた有益な Tweet が数多く投稿され，被災地のユーザを支援したと報告されている．

このように，ユーザにとって有益な実生活 Tweet は，Twitter に数多く投稿されるようになってきている．一方で，実生活 Tweet 以外の Tweet も少なくない．特に，「ありがとう」や「なるほど」のような，誰かの投稿に対する相槌や共感といった，ユーザの生活を直接支援しない Tweet が多い．このような Tweet は，実生活 Tweet の発見を妨げる原因となっている．

実生活 Tweet は，生活の様々な局面に対応している．例えば，「電車が来ない」という Tweet は「交通」の局面が付与され，電車に乗ろうとしているユーザを支援できる．「今日は全商品半額です！！」という Tweet は「消費」の局面に付与され，買物に行こうとしているユーザを支援できる．Wikipedia の「地域コミュニティ」[*2] と「生活」[*3] を参考に，実生活を表 1 に示す 14 の局面に整理し，いずれかの局面を Tweet に付与する手法を提案しているが，Tweet

[1] 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
[2] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Ibaraki, 305-8550, Japan
[a)] yamahei@ce.slis.tsukuba.ac.jp

[*1] http://twitter.com
[*2] http://ja.wikipedia.org/wiki/地域コミュニティ
[*3] http://ja.wikipedia.org/wiki/生活

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 ...

によっては複数の局面を付与する方が適切な場合もある．例えば，「今日は昼間から暑くなるので，水分補給をしっかりしましょう」のような Tweet に対しては，「気象」と「健康」の二つの局面を付与する方が適切であると考えられる．

著者らが提案した二段階抽出法は，局面毎にトピックとの関連度を算出し，関連度が閾値を超えたトピックと局面に対応関係を構築する．トピック中に含まれる単語の生起確率と，閾値を超えたトピックと局面の関連度を用いて，未知の Tweet に局面を付与する手法であることから，実生活 Tweet に複数の局面を付与するマルチラベル法に容易に拡張することができる．

本論文で提案するマルチラベリング法は，未知の Tweet に対して局面毎に計算したスコアの分布から，標準偏差を用いて動的に閾値を決定する．スコアが閾値を超えた局面集合を，未知の Tweet に対して付与する．この結果，特定の局面のみスコアが高い Tweet には一つの局面が付与され，複数の局面のスコアが高い Tweet には複数の局面が付与される．

本論文の構成を以下に示す．第 2 章は関連研究について述べる．第 3 章はマルチラベリングに拡張した二段階抽出法について説明する．第 4 章は提案手法を用いたときの適合率と再現率について評価する．第 5 章で考察を行う．最後に，第 6 章でまとめと今後の課題を述べる．

2. 関連研究

実生活 Tweet は，ユーザ個人の経験や知識，あるいは地域に特有の情報であると考えられる．文書から経験情報を抽出する，「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 [6] は，人間の経験を｛状況，行動，主観｝からなる情報と捉え，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 [5] 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．これらの経験マイニングに関する研究は，ブログなどの長い文書に対して効果的であるが，Twitter に投稿される記事のような，非常に短い文書に対しては有効に機能しないと考えられる．Twitter に投稿される記事は，頻繁に主語や目的語が省略され，経験マイニングをより難しくしている．

Twitter に関する研究は数多く行われている．Ramage ら [8] は，ハッシュタグなどのラベルを教師情報として利用できるように，LDA を拡張した Labeled LDA を用いることで，推薦の性能が向上することを示している．Bollen ら [2] は，Twee をの 6 次元の mood（tension，depression，anger，vigor，fatigue，confusion）について分析した結果，それらは株価など実世界の出来事と相関があることを明らかにしている．Sakaki ら [9] は，Twitter ユーザをセンサーとみなし，地震などの現実世界でリアルタイムに起きるイベントを発見する手法を明らかにしている．Zhao ら [13] は，一つの Tweet は一つのトピックの内容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，Tweet 集合をトピック毎に分類した後で，トピックの内容を表すキーワードやフレーズを抽出している．Mathioudakis ら [7] は，収集した Tweet からバーストキーワードを見つけ出し，キーワードの共起を用いてグルーピングすることで，リアルタイムに変動するトレンドを発見しようとしている．

本論文では，人々の生活に有益である実生活 Tweet を抽出するための手法を提案する．実生活 Tweet は，ユーザの経験だけでなく，ユーザの知識に基づく情報も対象としているため，従来の研究とは大きく異なる．

3. 実生活 Tweet の二段階抽出法

実生活 Tweet は表 1 に提示したように，様々な局面を含んでおり，全ての局面に関連するキーワードを列挙することは困難である．また，経験マイニングで多用さ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省略されることが多いことから，十分な精度を得ることは難しいと考えられる．

本論文では，LDA を用いた二段階抽出法 [12] を拡張して，未知の Tweet に複数の局面を柔軟に付与するマルチラベル法を提案する．提案法の核となる二段階抽出法の概要を図 1 に示す．第一段階では，LDA を用いて大量の Tweet 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，教師無し学習モデルであるため，大量の教師（正解）データを必要としない特徴がある [1]．第二段階では，局面ラベルが付与された少量の Tweet を用いて，トピックと局面の対応関係を構築する．未知の Tweet から抽出した単語から，トピックと局面の対応関係を用い

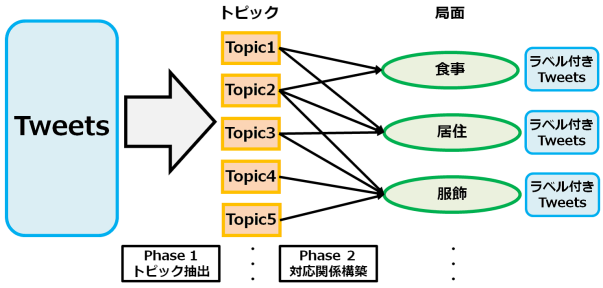

図 1 二段階抽出法

Fig. 1 Two phase extraction method

てスコアを算出する．スコアの分布から Tweet 毎に閾値を決定し，スコアが閾値を超えた局面集合を未知の Tweet に付与する．

以下．第 3.1 節ではトピックと局面の対応関係を構築する手法を，第 3.2 節では未知の Tweet に対する局面の付与方法について詳述する．

3.1 トピックと局面の対応関係構築

トピックと局面の対応関係を構築するため，少数のラベル付けされた Tweet を正解データとして用意する．局面 a としてラベル付けされた Tweet を形態素解析し，得られた単語の集合 [*4] を W_a とする．ここで，各局面を特徴付ける単語の重みとして情報利得を用いることで，特徴選択において良い特徴であるかを表現することができる．単語 w の情報利得 $IG(w)$ は，以下の式で与えられる．

$$IG(w) = H(A) - (P(w)H(A|w) + P(\bar{w})H(A|\bar{w})) \quad (1)$$

ここで，A は全ての局面を意味する．$P(w)$ は全ての Tweet の中で単語 w が出現する確率，$P(\bar{w})$ は全ての Tweet の中で単語 w が出現しない確率である．$H(A|w)$ は単語 w が出現する時の，全局面 A における条件付きエントロピー，$H(A|\bar{w})$ は単語 w が出現しない時の，全局面 A における条件付きエントロピーである．$IG(w)$ の値が高い時，単語 w は良い特徴であることを意味する．

トピック t と局面 a の関連度 $R(a,t)$ は，

$$R(a,t) = \frac{1}{|W_a|} \sum_{w \in W_a} IG(w) * p(w,t) \quad (2)$$

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．$|W_a|$ は単語集合 W_a の大きさである．この式は，単語の生起確率と情報利得を用いてトピックと局面の関連度を算出する式となっている．

0 から 1 の範囲にするため，正規化をする．ここでは，以下の式 (3) に示す，各局面で正規化した関連度 $\hat{Ra}(a,t)$ と，各トピックで正規化した関連度 $\hat{Rt}(a,t)$ を用意する．

[*4] 実際には，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

$$\hat{Ra}(a,t) = \frac{R(a,t)}{\sum_{t \in T} R(a,t)} \qquad \hat{Rt}(a,t) = \frac{R(a,t)}{\sum_{a \in A} R(a,t)} \quad (3)$$

ここで，T は第一段階の LDA で抽出した全てのトピック，A は全ての局面である．$\hat{Ra}$ は，局面がどのトピックによって表現されるかを示す指標であり，$\hat{Rt}$ は，トピックがどの局面を支持しているかを示す指標である．

正規化された関連度 $\hat{Ra}(a,t)$ が，各局面 a における閾値を超えた時，トピックと局面の対応関係を構築する．局面 a と対応関係が構築されたトピックの集合 T_a は，

$$T_a = \{t | \hat{Ra}(a,t) > E(\hat{Ra}(a,T)) + \sigma(\hat{Ra}(a,T)) * d\} \quad (4)$$

とする．ここで，$E(\hat{Ra}(a,T))$ は，局面 a における全トピック T の関連度の平均．$\sigma(\hat{Ra}(a,T))$ は，局面 a における全トピック T の関連度の標準偏差である．d を小さくすることでより多くのトピックが局面に関連付けらる．トピックが局面に関連付く度合いは局面毎に異なることから，標準偏差の d 倍をパラメータとし，第 4 章では d を変化させて，適合率と再現率を評価する．

3.2 未知の Tweet に対する局面の付与

未知の Tweet に局面を付与するため，第 3.1 節で述べたトピックと局面の対応関係を用いる．未知の Tweet tw と各局面 a のスコア $S(tw,a)$ は，以下の式で算出する．

$$S(tw,a) = \sum_{t \in T_a} \sum_{w \in W_{tw}} p(w,t) * \hat{Ra}(a,t) * \sigma(\hat{Rt}(A,t)) \quad (5)$$

ここで，W_{tw} は未知の Tweet から抽出した単語の集合，$\sigma(\hat{Rt}(A,t))$ は，トピック t における全局面 A の関連度の標準偏差である．$\sigma(\hat{Rt}(A,t))$ が高いとき，そのトピックは特定の局面を強く支持しており，局面にとって有用なトピックであることを表している．

スコア $S(tw,a)$ の高さは，tweet tw に局面 a の付与されやすさを表す値であるここで，スコアがより高い局面を未知の Tweet に付与する方法が求められるが，最もスコアが高い局面のみ付与する方法 [12] では，複数の局面が付与されるべき Tweet に対応できていなかった．予め与えられた K 件の局面を付与する方法でも，ある一つの局面だけが尤もらしい Tweet に K 個の局面を付与することになるため，未知の Tweet に対する局面付与数を動的に決定する方法が望まれる．

提案法では，Tweet に応じて付与すべき局面集合を動的に決定するため，スコアの分布を用いる．スコア間の値のばらつきが大きければ，標準偏差は大きくなり，ばらつきが小さければ，標準偏差は小さくなる．そこで，スコアの平均値に標準偏差を加算した閾値より高いスコアを持つ局面は，全ての局面の中でも未知の Tweet に付与されやすい局面であるとする．

未知の Tweet に付与される局面の集合は，以下の式で求

められる．

$$A_{tw} = \{a | S(tw, a) > E(S(tw, A)) + \sigma(S(tw, A))\} \quad (6)$$

ここで，$E(S(tw, A))$ は，未知の Tweet tw における全ての局面 A のスコアの平均であり，$\sigma(S(tw, A))$ は，未知の Tweet tw における全局面 A のスコアの標準偏差である．

4. 評価実験

第 3 章で提案した二段階抽出法による，局面を付与する適合率と再現率を評価する．実験では，つくば市で投稿された大量の Tweet から，LDA を用いてトピックを抽出する．正解データ作成のため，複数人による人手判定実験によって適切な局面を付与する．

以下，第 4.1 節では，評価実験に用いたデータセットとパラメータについて述べ，第 4.2 節では評価方法について説明する．第 4.3 節では実験結果について議論する．

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：トピック抽出のための Tweet

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 Tweet を使用する．その中から，つくば市で投稿された Tweet を抽出する．抽出条件は，各 Tweet のロケーション情報に「つくば」，あるいは「Tsukuba」と入力されている Tweet とした．以上の条件により収集した Tweet 数は，1,966,746 件となった．

4.1.2 データセット：実生活 Tweet

LDA で抽出したトピックと，生活の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 Tweet を用意する．1,500 件の Tweet に対して，第一著者（実験者 A）と他 2 名（実験者 B 及び C）の合計 3 名の実験者で人手判定を行った．実験者にはガイドラインとして，表 1 に示す局面に含まれる典型的な単語と，その局面に分類される例文（各局面 1 文ずつ）と，それが分類された理由を提示した．人手判定では，各 Tweet に対して最も適切な局面を一つだけ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，「非実生活」を付与することとした．なお，1,500 件の Tweet はいずれもロケーション情報に「つくば」あるいは「Tsukuba」と表記されたものであり，3 名の実験者はいずれも「つくば市」在住の大学生である．

人手判定によって分類が一致した Tweet 数を表 2 に示す．実験者間の κ 値 [3] は，実験者 A と実験者 B の κ 値が 0.609，実験者 A と実験者 C の κ 値が 0.645，実験者 B と実験者 C の κ 値が 0.664 となった．κ 値の平均は 0.639 であり，高い一致（substantial）が得られた．

トピックと局面の対応関係の構築，及び提案手法の評価では，実験者 3 名のうち，2 名以上の判定が一致した 1,382 件の Tweet を用いる．非実生活も一つの局面としてトピックとの関係を構築し，評価に含める．

表 2 人手分類による Tweet 数
Table 2 The number of tweets by humans classification

局面	二人以上一致	三人一致
服飾	97	73
交流	94	37
災害	97	69
食事	120	82
行事	78	11
消費	74	14
健康	98	68
趣味	105	64
居住	96	66
地域	56	20
学校	107	80
交通	104	72
気象	105	71
労働	79	45
非実	72	31
合計	1,382	803

4.1.3 パラメータ設定

LDA は，いくつかのパラメータを設定する必要がある．関連研究 [4] を参考にハイパーパラメータである α は $50/T$, また β は 0.1 とした．LDA の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，トピック数は 20, 50, 100, 200, 500, 1,000 と変化させた．

4.2 評価方法

4.2.1 トピック数

最適なトピック数を決定するため，各局面間の JS Divergence を用いて，ある一つの局面と他の局面との類似度を計算する．二つの局面の確率分布が同じであるとき，JS Divergence は 0 となる．本論文の場合は，局面間の確率分布 $\hat{Ra}$ が異なっている方が望ましい．そのため，各局面間の JS Divergence の合計値が最大であるとき，最適なトピック数であるとした．JS Divergence の合計値 JS_{sum} は，以下の式で求められる．

$$JS_{sum} = \sum_{(\forall P, \forall Q) \in A} JS(P||Q) \quad (7)$$

$$JS(P||Q) = \frac{1}{2}\left(\sum_x P(x) \log \frac{P(x)}{R(x)} + \sum_x Q(x) \log \frac{Q(x)}{R(x)}\right)$$

ここで，P と Q は確率分布を，A は全ての局面である．R は確率分布 P と Q の平均であり，$R = \frac{P+Q}{2}$ である．

4.2.2 比較手法

提案手法の有効性を確認するため，算出したスコアの上位 K 件の局面を付与する TOP@K 法を，比較手法として用いる．局面付与数である K を決定するため，提案手法

を用いたときの閾値 d にともなう局面付与数の平均を算出する．

4.2.3 適合率・再現率・F 値

未知の Tweet に対して局面を付与するときの適合率と再現率は，10 分割交差検定によって評価する．10 分割されたうちの 9 割の Tweet で局面とトピックの対応関係を構築し，残りの 1 割の Tweet で評価する．以上の操作を 10 回繰り返し，適合率と再現率，F 値の平均を算出する．

4.3 実験結果

4.3.1 トピック数の決定

最適なトピック数を決定するため，式 (7) に示す JS_{sum} を算出した．トピック数を変化させたときの JS_{sum} の値を表 3 に示す．この表から，最大の JS_{sum} となるトピック数 500 を用いて，以降の評価を行うこととした．

表 3 各トピック数のときの JS_{sum}
Table 3 JS_{sum} in each the number of topics

トピック数	20	50	100	200	**500**	1,000
JS_{sum}	11.74	16.67	20.10	21.63	**22.91**	22.31

4.3.2 比較手法の局面付与数の決定

TOP@K 法による局面付与数を決定するため，提案法を用いたときの閾値 d にともなう局面付与数の平均を，図 2 に示す．横軸はトピックとの対応関係を決定する閾値 d，縦軸は局面付与数である．閾値 d の変化に対して，局面付与数の平均は最大でも 3 以下となっている．このことから，各 Tweet に対する局面付与数として，TOP@1，TOP@2，TOP@3 を用いて以降の評価を行うこととした．

4.3.3 適合率・再現率・F 値

閾値 d にともなう適合率の変化について，提案手法を図 3 と図 5，TOP@2 を図 4 と図 6 に示す．いずれの図においても，横軸はトピックと局面の対応関係を決定する閾値 d，縦軸は適合率である．服飾や労働の局面は，d の変化に関わらず，TOP@2 に比べ提案手法の適合率が高く，災害の局面は TOP@2 の適合率が高くなった．また，閾値 d の増加にともなって，服飾や災害，交通の局面で適合率が高くなっている．

閾値 d にともなう再現率の変化について，提案手法を図 7 と図 9，TOP@2 を図 8 と図 10 に示す．横軸は閾値 d，縦軸は適合率である．労働の局面は，TOP@2 に比べ提案手法の再現率が高く，消費の局面は TOP@2 の再現率が高くなった．また，閾値 d の増加にともなって，交流や災害，消費の局面で再現率が低くなっている

$d = 0$ における各局面の適合率と再現率，F 値を表 4 に示す．服飾と交流，行事の局面では，提案手法と TOP@2 で再現率が同じ値を示しているが，適合率は提案手法の方

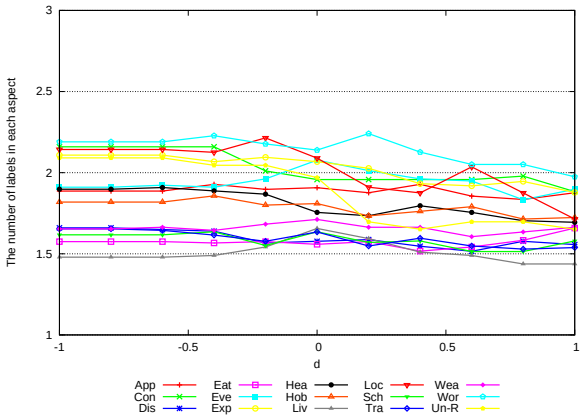

図 2 局面付与数
Fig. 2 The number of labeles in each aspect

が高い．食事と居住，交通，気象，労働の局面では，提案手法が適合率と再現率共に TOP@2 より高い値を示している．適合率はいずれの局面においても TOP@1 が最も高くなり，再現率はいずれの局面においても TOP@3 が最も高くなった．F 値については，服飾や交流，食事などの局面では提案手法が最も高くなったが．F 値の平均では TOP@1 が最も高くなった．

5. 考察

図 3 と図 7 から，交流の局面では閾値 d の増加にともなって，適合率は大きな変化をせず，再現率は低くなっていることが分かる．閾値 d はトピックと局面の対応関係を決定するパラメータであり，閾値 d の増加にともなって，局面と関連を持つトピックは少なくなる．交流の局面では，閾値 d の増加にともなって，局面を表現するために必要であったトピックとの関連が無くなったため，再現率が低くなったと考えられる．交通の局面では閾値 d の増加にともなって，再現率は大きな変化をせず，適合率は高くなっていることが分かる．交通の局面では，閾値 d の増加にともなって，局面を表現しないノイズとなるトピックとの関連が無くなったため，適合率が高くなったと考えられる．

労働の局面では，TOP@2 に比べ提案手法の適合率と再現率共に高いことが分かる．これは，スコア算出の際に労働以外の局面のスコアが高くなり，スコアが上位 2 つの局面を付与するだけでは十分でなかったことが原因と考えられる．提案手法では，スコアの分布に応じて付与する局面集合を決定するため，労働が他の局面にに比べて十分に高いスコアを持っている場合に，適切に局面を付与できていると考えられる．また，表 4 から，TOP@2 に比べて多くの局面で同等かそれ以上の再現率を示しながら，適合率が高くなっていることが分かる．これは，提案手法によって動的に閾値を決定し，正解となる局面のスコアが非常に高くなっている Tweet には，一つの局面のみを付与したためであると考えられる．

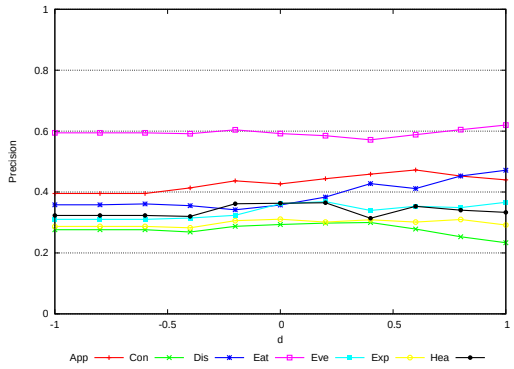

図 3 提案手法の適合率: 服飾 - 健康

Fig. 3 Precision of proposed method: Appearance - Health

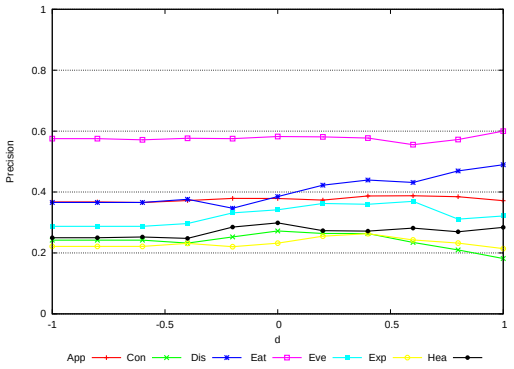

図 4 TOP@2 の適合率: 服飾 - 健康

Fig. 4 Precision of TOP@2: Appearance - Health

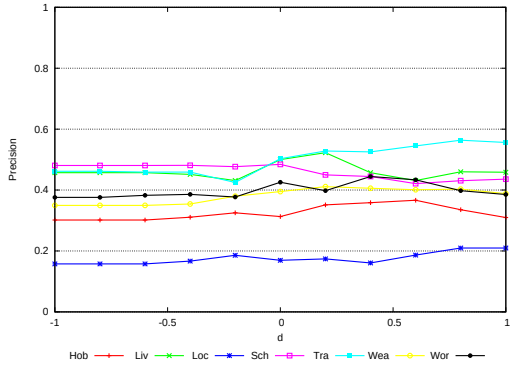

図 5 提案手法の適合率: 趣味 - 労働

Fig. 5 Precision of proposed method: Hobby - Working

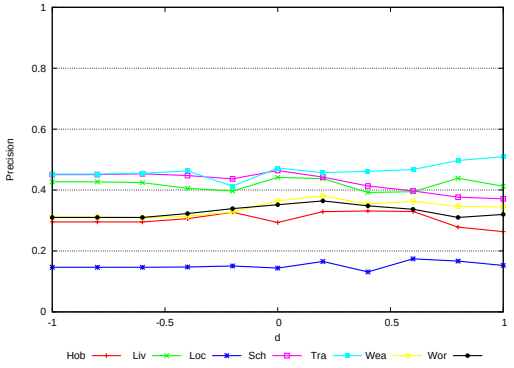

図 6 TOP@2 の適合率: 趣味 - 労働

Fig. 6 Precision of TOP@2: Hobby - Working

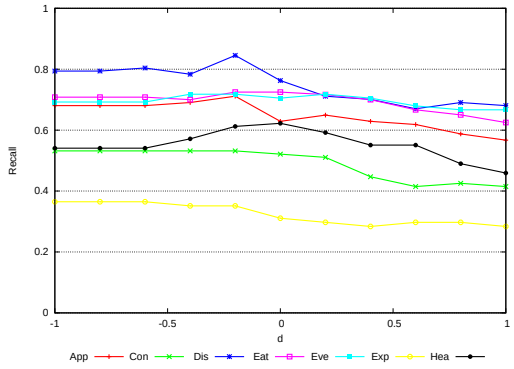

図 7 提案手法の再現率: 服飾 - 健康

Fig. 7 Recall of proposed method: Appearance - Health

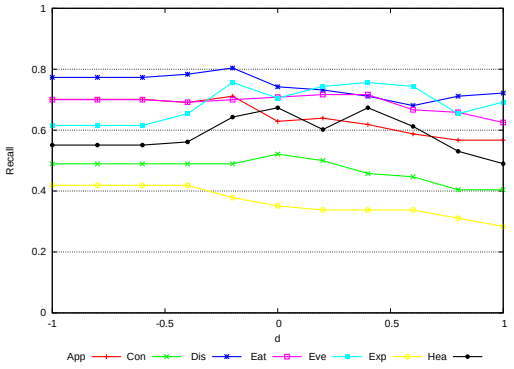

図 8 TOP@2 の再現率: 服飾 - 健康

Fig. 8 Recall of TOP@2: Appearance - Health

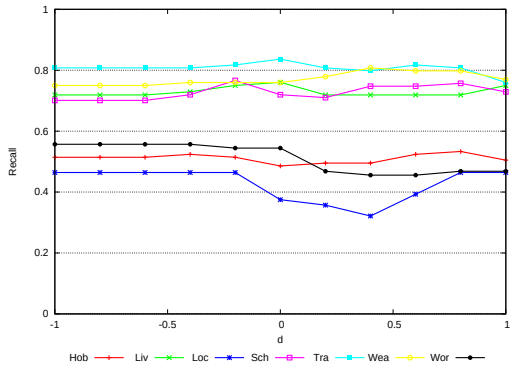

図 9 提案手法の再現率: 趣味 - 労働

Fig. 9 Reacall of propose method: Hobby - Working

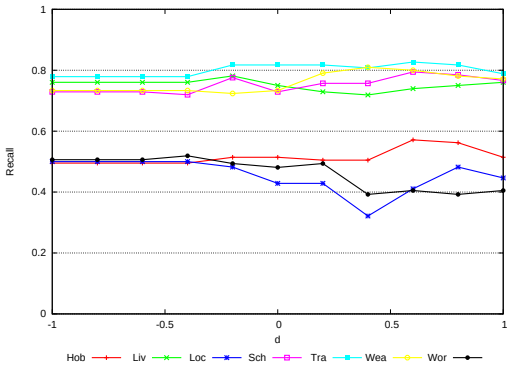

図 10 TOP@2 の再現率: 趣味 - 労働

Fig. 10 Recall of TOP@2: Hobby - Working

表 4 各手法における，閾値 $d = 0$ のときの適合率・再現率・F 値
Table 4 Precision, Recall, and F-measure of each method in $d = 0$

	提案手法			TOP@1			TOP@2			TOP@3		
局面	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
服飾	0.427	0.629	0.508	0.725	0.381	0.500	0.379	0.629	0.473	0.295	0.763	0.425
交流	0.293	0.521	0.375	0.375	0.223	0.280	0.272	0.521	0.358	0.196	0.628	0.299
災害	0.357	0.763	0.487	0.600	0.619	0.609	0.385	0.742	0.507	0.292	0.835	0.433
食事	0.592	0.725	0.652	0.611	0.642	0.626	0.582	0.708	0.639	0.470	0.775	0.585
行事	0.362	0.705	0.478	0.462	0.462	0.462	0.342	0.705	0.460	0.255	0.808	0.388
消費	0.311	0.311	0.311	0.375	0.284	0.323	0.232	0.351	0.280	0.207	0.419	0.277
健康	0.363	0.622	0.459	0.500	0.398	0.443	0.299	0.673	0.414	0.195	0.765	0.311
趣味	0.313	0.486	0.381	0.398	0.371	0.384	0.293	0.514	0.374	0.237	0.667	0.350
居住	0.500	0.760	0.603	0.588	0.625	0.606	0.442	0.750	0.556	0.326	0.792	0.462
地域	0.169	0.375	0.233	0.216	0.286	0.246	0.144	0.429	0.215	0.120	0.464	0.190
学校	0.484	0.720	0.579	0.684	0.626	0.654	0.464	0.729	0.567	0.311	0.832	0.453
交通	0.503	0.837	0.628	0.658	0.702	0.679	0.472	0.817	0.599	0.332	0.885	0.483
気象	0.395	0.760	0.520	0.444	0.571	0.500	0.365	0.733	0.487	0.297	0.810	0.435
労働	0.426	0.544	0.478	0.511	0.291	0.371	0.352	0.481	0.406	0.274	0.620	0.380
非実	0.080	0.424	0.134	0.106	0.265	0.151	0.074	0.441	0.126	0.078	0.632	0.139
平均	0.372	0.612	0.463	0.484	0.450	0.466	0.340	0.615	0.438	0.259	0.713	0.380

表 5 提案手法による混合行列
Table 5 Confusion Matrix

		局面付与結果														
		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
正解	服飾	**62**	7	4	2	2	18	5	16	10	3	6	3	12	1	33	184
	交流	1	**50**	8	11	24	1	17	11	1	4	7	8	8	13	21	185
	災害	1	5	**75**	0	2	0	7	7	4	15	2	7	7	1	17	150
	食事	8	9	3	**84**	4	10	16	5	3	3	2	6	10	3	21	187
	行事	6	14	18	3	**57**	1	2	5	4	16	7	3	6	4	16	162
	消費	23	8	7	13	12	**25**	5	15	4	10	7	4	4	1	10	148
	健康	4	16	3	9	5	2	**62**	1	7	1	6	4	17	8	29	174
	趣味	13	14	5	4	3	6	8	**55**	6	3	9	7	8	6	50	197
	居住	5	4	4	4	4	4	5	7	**71**	2	2	7	10	3	15	147
	地域	1	7	20	1	4	1	2	6	3	**20**	3	13	8	2	17	108
	学校	4	5	6	6	2	1	6	14	2	3	**77**	2	12	6	29	175
	交通	1	3	16	1	4	1	3	3	8	19	3	**85**	9	4	7	167
	気象	3	3	16	1	2	0	5	6	8	17	3	10	**81**	2	13	170
	労働	4	16	10	4	22	2	9	3	7	8	11	5	7	**45**	22	175
	非実	9	11	5	2	4	2	9	11	4	2	5	3	3	6	**21**	97
	合計	145	172	200	145	151	74	161	165	142	126	150	167	202	105	321	2426

TOP@1 と比較すると，服飾や交流，行事，健康の局面について，提案手法の再現率が非常に高くなっていることが分かる．このような局面では，最大となるスコアを持つ局面が正解の他に存在していたため，TOP@1 では適切な局面を付与することができていなかったと考えられる．しかし，提案手法を用いることにより，正解の局面が十分に高いスコアを持っている場合に正解の局面も付与できたため，再現率が高くなったと考えられる．

TOP@3 ではスコアが上位 3 つの局面を Tweet に付与するため，いずれの局面についても再現率が最も高くなっている．しかし，消費と地域の局面については再現率が 0.5 以下であることが分かる．これらの局面はスコアが適切に算出できていないことが考えられる．消費や地域の局面について，提案手法によって，どのような局面に混合されやすか，また混合しやすいかを分析するため，Cofusion Matrix を作成した．Consuion Matrix を表 5 に示す．行が正解データを，列が正解に対してどのような局面を付与したかを表している．地域の局面は災害や交通と，消費の局面は服飾や食事，行事，趣味など多くの局面と混合されている傾向が見られる．また，地域の局面は災害や行事，交

通，気象と混合している傾向が見られる．消費の局面は消費や食事と混合しやすい傾向が見られるが，消費の局面が付与された総数は 74 件であり，他の局面に比べて極めて少ないことが分かる．以上のことから，地域の局面は他の局面から混合されやすく，混合もしやすいため，十分に高いスコアを算出できているものの，より高いスコアが他の複数局面に現れやすいことが考えられる．消費の局面は他の局面から混合されやすいが，混合しにくいため，スコアが適切に算出できていないことが考えられる．

6. 結論

本論文では，実生活 Tweet を抽出するための二段階抽出法を用いたマルチラベリング法を提案した．第一段階では，LDA を用いて大量の Tweet からトピックを抽出する．第二段階では，少数のラベル付き Tweet を用いて．トピックと局面の関連度を計算し，関連度が閾値を超えたトピックについて，局面との対応関係を構築する．未知の Tweet から抽出した単語を用いて，局面とトピックの対応関係を用いて，局面毎にスコアを算出する．スコアの分布から平均値と標準偏差を用いて閾値を算出し，閾値を超えた局面集合を Tweet に付与する．

評価実験の結果，未知の Tweet から複数の局面を抽出できることが明らかになった．TOP@K 法との比較により，正解となる局面が最大スコアを持たないときでも，提案手法を用いることで適切な局面を付与できたため，高い再現率を示した．. また，正解となる局面のスコアが非常に高い Tweet には，一つの局面のみを付与したため，高い適合率を示した．消費や地域の局面は，他の局面に比べて適合率と再現率が低くなった．Confusion Matrix を用いた分析により，地域の局面は他の局面と混合されやすく混合しやすいことが，消費の局面は他の局面から混合されやすいことが明らかになった．

今後の課題は，トピックと局面の関係構築手法及びスコア算出法を洗練し，他の局面との混合が見られる局面を適切に分離することと，提案手法を実装したシステムの構築があげられる．

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．

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] Bollen, J., Pepe, A. and Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, *In Proceedings of the 2010 World Wide Web*, pp. 450–453 (2010).

[3] Cohen, J.: A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46 (1960).

[4] Griffiths, T. L. and Steyvers, M.: Finding scientific topics, *Proceedings of the National Academy of Science*, Vol. 101, pp. 5228–5235 (2004).

[5] Inui, K., Abe, S., Morita, H., Eguchi, M., Sumida, A., Sao, C., Hara, K., Murakami, K. and Matsuyoshi, S.: Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents, *In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321 (2008).

[6] Kurashima, T., Tezuka, T. and Tanaka, K.: Extracting and Geographically Mapping Visitor Experiences from Urban Blogs, *The 6th International Conference on Web Information Systems Engineering*, pp. 496–503 (2005).

[7] Mathioudakis, M. and Koudas, N.: Twittermonitor: trend detection over the twitter stream, *In Proceedings of the 2010 International Conference on Management of Data*, pp. 1155–1158 (2010).

[8] Ramage, D., Dumais, S. and Liebling, D.: Characterizing microblogs with topic models, *In Proceedings of the 4th Int'l AAAI Conference on Weblogs and Social Media*, pp. 130–137 (2010).

[9] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes twitter users: Real-time event detection by social sensors, *In Proceedings of 18th International World Wide Web Conference*, pp. 851–860 (2010).

[10] Yamamoto, M., Ogasawara, H., Suzuki, I. and Furukawa, M.: Tourism Informatics:9. Information Propagation Network for 2012 Tohoku Earthquake and Tsunami on Twitter, *IPSJ Magazine*, Vol. 53, No. 11, pp. 1184–1191 (2012 (in Japanese)).

[11] Yamamoto, S. and Satoh, T.: Real Life Information Extraction Method from Twitter, *The 4th Forum on Data Engineering and Information Management* (2012 (in Japanese)).

[12] Yamamoto, S. and Satoh, T.: Two Phase Extraction Method for Extracting Real Life Tweets using LDA, *The 15th Asia-Pacific Web Conference*, pp. 340–347 (2013).

[13] Zhao, X., Jiang, J., He, J., Song, Y., Achananuparp, P., LIM, E. P. and Li, X.: Topical key phrase extraction from Twitter, *The 49th Annual Meeting of the Association for Computational Linguistics*, pp. 379–388 (2011).