Vol.2015-DPS-163 No.34
Vol.2015-MBL-75 No.34
2015/5/29

# 天気・時期コンテキストを考慮したトピックモデル

伊藤拓[†1]　朱丹丹[†1]　深澤佑介[†2]　太田順[†1]

本研究では，天気コンテキストと時期コンテキストに着目し，天気・時期コンテキストと Twitter の投稿内容の関係性を表すトピックモデルを提案した．天気を考慮したモデル，時期のみを考慮したモデル，両方を考慮したモデルを文書における単語の予測精度(Perplexity)の観点で比較評価し，天気と時期の両方を考慮することによる性能向上を示した．

# Topic Models Considering both Weather and Seasonal Contexts

TAKU ITO[†1]　DANDAN ZHU[†1]　YUSUKE FUKAZAWA[†2]　JUN OTA[†1]

In this research, we focused on weather and seasonal contexts, and proposed topic models that represented the relationship between weather/season contexts and posted sentences on Twitter. We compared a model considering only weather, a model considering only season, a model considering both weather and season from view point of how models were able to predict posted words in documents(Perplexity). Then, we discovered that the evaluation improved because of considering both weather and seasonal contexts.

## 1. 序論

近年，インターネット上での交流を通じて社会的なネットワークを構築するソーシャル・ネットワーキング・サービス(SNS)が流行しており，その中でも特にユーザが気軽に短文を投稿できる Twitter は多くのユーザが登録している．Twitter の投稿内容はユーザ個人の趣味や嗜好，ユーザ全体のトレンドなどが反映されており，逆に Twitter の投稿内容を見ることでそうした情報を知ることができる．ユーザの趣味嗜好，トレンドを知ることで，商品の需要予測やユーザへのレコメンデーションシステムに役立てることができる．

Twitter からユーザの趣味嗜好，トレンドを知るうえで，Twitter の投稿内容が何によって影響を受けているのかを知ることは重要である．Twitter の投稿内容は，ユーザが文書を投稿する際の文脈情報（コンテキスト）によって大きく影響されると考えられる．中でも，天気というコンテキストは，Twitter の投稿内容に大きな影響を与えるコンテキストであると言える．ただし，同じ天気であっても時期によってユーザの天気の感じ方は異なり，Twitter の発話内容も変わると考えられる．したがって、天気コンテキストと時期コンテキストを両方考慮し，どのように組み合わさって Twitter の発話内容に影響を与えているかを知ることは重要である．

コンテキストと Twitter の発話内容の関係を調べる手法として，トピックモデルという手法が存在する．トピックモデルとは，ユーザが文書を投稿する際の文書生成過程の仮説を表したものであり，状況に応じていくつかの話題から 1 つの話題が選択され，その話題に応じていくつかの単語から 1 つの単語が選択されるという考えに基づいている．この過程を単語ごとに繰り返すことで文書が生成されるというのがトピックモデルの考え方である．

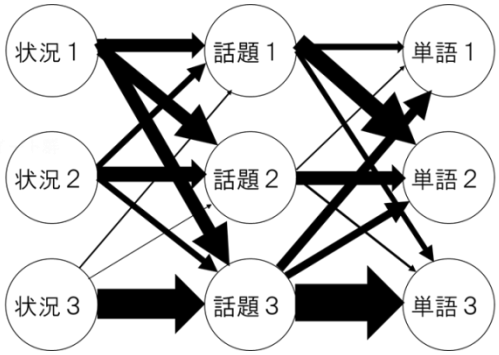


図 1　トピックモデルにおける文書生成過程の仮説
Figure 1　The Hypothesis about the process of generating documents in topic models

それを図示したものが図 1 である．状況によってどのような話題が選ばれるかという確率は変化し，話題によってどの単語が選ばれるかという確率が変化する．図 1 では，矢印の太さが確率の大きさを表しており，状況に応じて話題を選ぶ確率，話題に応じて単語を選ぶ確率が変わることを意味している．たとえば場所について考えてみると，スタジアムにいるという状況（状況 1）でサッカーに関する話題（話題 1）は出やすいが，野球場にいるという状況（状況 2）でサッカーに関する話題（話題 1）は出にくい．また，どのような単語が発話されるかは選択された話題によっても異なる．サッカーに関する話題（話題 1）を選んでいる

†1 東京大学
The University of Tokyo
†2 (株)NTT Docomo
NTT Docomo Ltd.

Vol.2015-DPS-163 No.34
Vol.2015-MBL-75 No.34
2015/5/29

ときは，ゴールという単語（単語 2）は発話されやすいが，ホームランという単語（単語 1）は発話されにくい．状況と話題，話題と単語の関係性を学習することにより，ある状況でどのような単語が発話されやすいかを知ることができる．

コンテキストを考慮したトピックモデルとして，場所[1], 時間[2], 同行者[3]を考慮したトピックモデルが提案されている．しかし，天気をコンテキストとしたトピックモデルは提案されていないため，天気・時期コンテキストがどのように組み合わさって Twitter の投稿内容に影響を与えているかという問題が解決されていない．

そこで本研究の目的を，「天気・時期コンテキストを考慮したトピックモデルを提案し，2 つのコンテキストがどのように組み合わさって Twitter の発話内容に影響を与えているかを調べる」こととする．目的の達成にあたって，2 つ課題がある．1 つ目は，天気コンテキストには気温，湿度，降水量など複数の要素があるがどの要素をコンテキストとして採用するかという点，2 つ目は天気と時期という 2 個のコンテキストを，どのように組み合わせてモデル化するかという点である．

1 つ目の課題には，線形分類器を利用し，天気要素のツイート分離性能を評価することで，Twitter の発話内容への影響度を評価し，影響が大きい要素を絞り込む方法をとる．2 つ目の課題には，各ツイートを天気と時期によってクラス分類し，その 2 つのクラスの組を 1 つのクラスとみなすという方法をとる．

## 2. 問題設定

本研究ではトピックモデルを提案することを目的としているが，課題で述べたように採用する天気要素を絞る必要があるため，天気要素を絞る問題とトピックモデルを提案する問題の 2 つに分ける．

### 2.1 天気要素の絞り込み

はじめに，天気要素の絞り込みを行う．線形分類器を用いたツイートの天気要素による分類を行い，天気要素のツイート分離性能の高さを，天気要素のツイート内容への影響の大きさの評価値として採用し，天気要素同士の比較を行う．ツイート内容への影響度の大きさは，その天気要素を閾値としてツイートを 2 グループに分け，そこに登場する名詞がグループ間でどの程度異なるかによって評価する．これは，天気要素がツイート内容に影響を与えているならば，その天気要素において天気条件が異なれば発話される名詞も変わるだろうとの考えに基づくものである．具体的な評価関数としては，第 3 章で述べる F 値という関数を用いる．比較する天気要素は，気温，湿度，降水量，風速，雲量，日照時間，天気概況である．本問題では，線形分類器による評価が高かった天気要素が出力となる．

### 2.2 トピックモデルの提案

次に，トピックモデルの提案について述べる．コンテキストを考慮しないモデル，天気のみを考慮したモデル，時期のみを考慮したモデル，天気と時期両方を考慮したモデルについて，Twitter の発話内容とコンテキストの関係性を示しているかの性能比較を行う．トピックモデルは文書の生成過程を確率的に表現したモデルであり，各単語について出現確率が存在する．たとえば，ツイート $d$ の $i$ 番目の単語が「アイス」である確率 $P(w_{di}=\text{アイス})=0.0056$ などと与えられる．実際に出力された単語の尤度が高ければ，そのモデルは Twitter の文書生成過程を正しく表せていると言える．したがって，単語の尤度の高さをトピックモデルの評価値とする．具体的な評価関数としては，第 4 章で述べる Perplexity という関数を用いる．

なお，本実験では 2011 年 5 月から 12 月までに東京都で投稿された日本語の位置情報付きツイート 928051 件を分析対象としている．

## 3. 提案アルゴリズム

### 3.1 天気要素の絞り込み

本稿では，著者らが[4]において天気要素の絞り込みを行った結果を利用する．本章ではその手法と結果について概要を説明する．

(1)天気データとツイートデータの紐づけ

はじめに，気象庁[5]からダウンロードした天気データと，ツイートデータを日付によって紐づける．

(2)編集距離によるフィルタリング

次に，編集距離によるフィルタリングを行う．編集距離とは 2 つの文章の近さを表す指標であり，ツイート投稿時間の直近 1000 件以内のツイートとの編集距離が 30 未満であった場合，そのツイートをデータセットから除去する．これにより，bot とよばれる自動生成によるツイートを排除することができる．

(3)天気要素と閾値によるツイートのグループ分類

続いて，天気要素と閾値を選択し，ツイートデータの 2 グループ分類を行う．天気要素がツイートに影響を与えているならば，その天気要素が異なる天気のときに Twitter の発話内容も変わると考えられるので，2 つに分けられたツイートの内容は大きく変わるはずである．

(4)天気要素と閾値のツイート分離性能に関する評価

最後に，形態素解析を用いてツイートから名詞のみを抽出し，2 つのグループに分けられたツイートに登場する名詞がどれだけ明確に分けられたかを評価する．まず全ツイートの 8 割を学習用データ，残りを試験用データに分ける．線形分類器として Stanford Classifier[6]を用いて，学習用データからグループの特徴を学習する．学習した特徴をもと

Vol.2015-DPS-163 No.34
Vol.2015-MBL-75 No.34
2015/5/29

に，Stanford Classifier は試験データに含まれるツイートのグループを予測する．この正答率を表すのが F 値という値であり，天気要素の評価値となる．

以上の評価実験の結果について，天気要素と F 値の最大値，それと閾値の関係を表した表を表 1 に載せる．

表 1　天気要素と F 値の最大値，閾値の関係
Table 1　Max F-value for Each Weather Condition

| 天気要素 | *F* 値(最大値) | 閾値 |
|---|---|---|
| 気温 | 0.648 | 8℃ |
| 湿度 | 0.634 | 42% |
| 降水量 | 0.599 | 15mm |
| 雲量 | 0.599 | 5 |
| 風速 | 0.587 | 7m/s |
| 天気概況 | 0.562 | 晴れ |
| 日照時間 | 0.555 | 8h |

表 1 について説明する．たとえば，気温という天気要素について調べたとき，複数の閾値について F 値を計算しているが，その中でも 8℃が閾値のときにもっとも F 値が高く，0.648 であったことを意味する．

以上の結果から，F 値の比較により，ツイート内容により影響を与えている天気要素は気温と湿度であることが分かった．よって，気温と湿度をトピックモデルの天気コンテキストとして組み込む．

### 3.2 トピックモデルの提案

前節での結果をもとに，天気コンテキストとして気温と湿度，時期コンテキストとして日付をトピックモデルに組みこむ．今回比較を行う，3 つのトピックモデルについてそれぞれの文書生成過程の仮説を述べる．

(a)　天気コンテキスト考慮モデル

このモデルでは気温と湿度をパラメータとしてツイートが天気クラスに分類され，クラスに応じてトピックが選ばれる．選ばれたトピックからそのトピックごとの確率分布に従って単語が生成される．これにより，気温と湿度の違いによるトピックの違いを表現できる．このモデルにおけるグラフィカルモデルを図 2 に載せる．

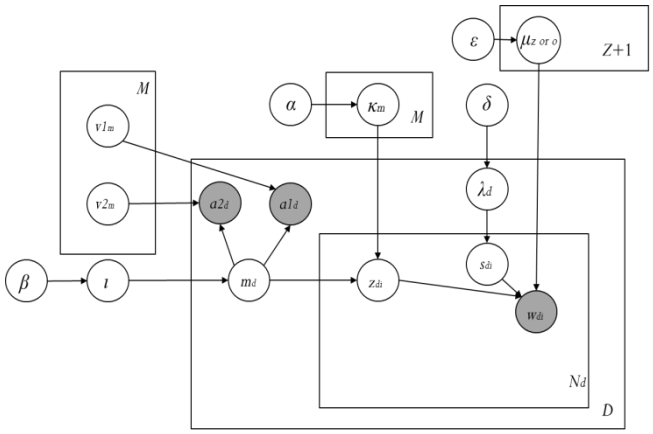


図 2　天気コンテキスト考慮モデルのグラフィカルモデル
Figure 2　The graphical model of the topic model considering weather context

図 2 における変数の定義を表 2 に載せる．この定義は，後述する全モデルで共通である．図 2 において，$\kappa_m$ は天気クラスごとに存在するトピックの確率分布を表している．このモデルにおいては，各ツイートにつき天気クラス($m_d$)が割り当てられ，天気クラスに応じて気温や湿度を選ぶというモデルになっている．本来天気は入力であるが，このモデルでは天気に基づく潜在的な欲求が存在し，その欲求に応じて発話する天気を選択し，その天気のときに発話をするという仮説になっている．また，モデル(a)においてはスイッチ変数($s_{di}$)を導入している．これは，「私」などどのトピックでも共通して現れる単語を，背景トピックとして排除するための機構である．

(ア) トピック固有の単語($s_{di}$ =0)
(イ) トピックで共通する単語($s_{di}$ =1)

スイッチ変数は，モデルが自動的に学習を行う．$s_{di}=0$ のとき，単語 $w_{di}$ のトピックは $z$ から選ばれる．$s_{di}=1$ のとき，単語 $w_{di}$ のトピックは背景トピック $o$ から選ばれる．すなわち，単語の出現確率分布 $\mu$ は，トピック数 $Z$ と背景トピック 1 個を足し合わせた $Z$+1 個存在する．

Vol.2015-DPS-163 No.34
Vol.2015-MBL-75 No.34
2015/5/29

表 2　変数の定義

Table 2　The definition of valuables

| Variable | Meaning |
|---|---|
| $M$ | 天気クラスの数 |
| $T$ | 時期クラスの数 |
| $Z$ | トピック数 |
| $D$ | ツイート数 |
| $V$ | ユニークな単語数 |
| $N_d$ | ツイート$d$の単語数 |
| $m_d$ | ツイート$d$の天気クラス |
| $t_d$ | ツイート$d$の時期クラス |
| $z_{di}$ | ツイート$d$の$i$番目の単語のトピック |
| $k_{di}$ | |
| $s_{di}$ | ツイート$d$の$i$番目の単語のスイッチ変数 |
| $w_{di}$ | ツイート$d$の$i$番目の単語 |
| $\iota$ | あるツイートの天気クラスの分布 |
| $\tau$ | あるツイートの時期クラスの分布 |
| $\nu_{1m}$ | 天気クラス$m$の気温への正規分布 |
| $\nu_{2m}$ | 天気クラス$m$の湿度への正規分布 |
| $a_{1d}$ | ツイート$d$の気温 |
| $a_{2d}$ | ツイート$d$の湿度 |
| $y_d$ | ツイート$d$の日付 |
| $\omega_t$ | 時期クラス$t$の日付への正規分布 |
| $\mu_{zorb}$ | 単語のトピック$z$もしくは背景トピック$b$の単語への分布 |
| $\lambda_d$ | ツイート$d$のスイッチ変数への分布 |
| $\kappa_m$ | 天気クラス$m$のトピックへの分布 |
| $\theta t$ | 時期クラス$t$のトピックへの分布 |
| $\beta$ | $\iota$のディリクレ事前分布のパラメータ |
| $\gamma$ | $\tau$のディリクレ事前分布のパラメータ |
| $\delta$ | $\lambda$dのディリクレ事前分布のパラメータ |
| $\alpha$ | $\kappa_m$のディリクレ事前分布のパラメータ |
| $\rho$ | $\theta_t$のディリクレ事前分布のパラメータ |
| $\varepsilon$ | $\mu_{zorb}$のディリクレ事前分布のパラメータ |

(b)　時期コンテキスト考慮モデル

このモデルでは日付をパラメータとしてツイートが時期クラスに分類され，クラスに応じてトピックが選ばれる確率が変化する．これにより，日付の違いによるトピックの違いを表現できる．このモデルにおけるグラフィカルモデルを図 3 に載せる．

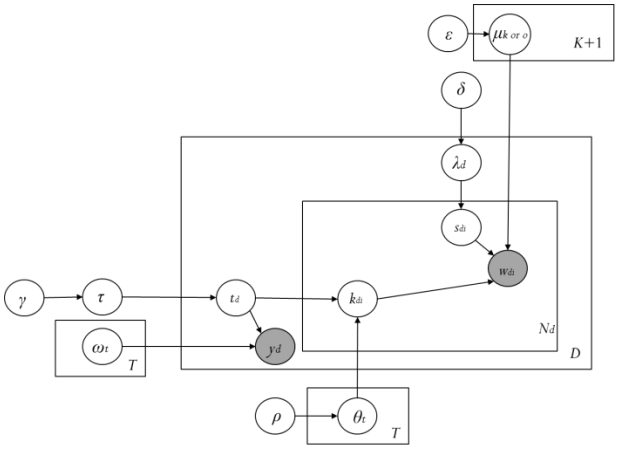


図 3　時期コンテキスト考慮モデルのグラフィカルモデル

Figure 3　The graphical model of the topic model considering season context

このモデルも天気コンテキスト考慮モデルと同様に，潜在的な時期に基づく欲求が存在し，その欲求に応じて適切な時期を選択し，その時期の時に発話をするという仮説になっている．

(c)　天気＋時期コンテキスト考慮モデル

このモデルでは天気コンテキスト考慮モデルや時期コンテキスト考慮モデルと同様に，ツイートが天気クラス，時期クラスに分類される．天気クラスに応じて天気のトピックが時期クラスに応じて時期のトピックが選ばれる．その後，天気トピックか時期トピックのどちらかが選ばれ，そのトピックにおける確率分布に従って単語が選ばれる．これにより，気温，湿度と日付の違いによるトピックの違いを表現できる．このモデルにおけるグラフィカルモデルを図 4 に載せる．

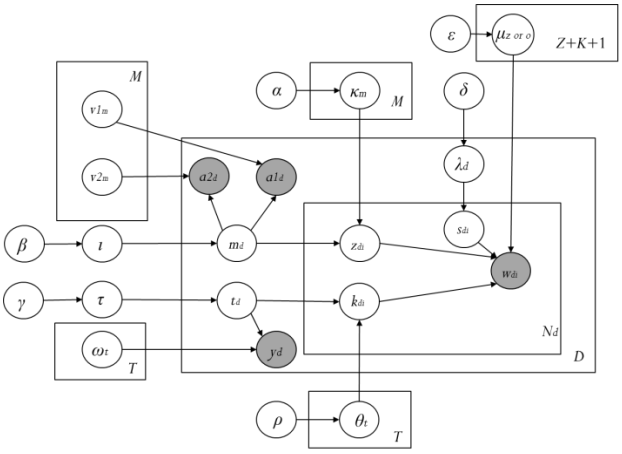


図 4　天気＋時期コンテキスト考慮モデルの
グラフィカルモデル

Figure 4　The graphical model of the topic model considering both weather and season context

## 4.　トピックモデルの量的評価

定量的な評価として，Perplexity による評価を行う．Perplexity は，トピックモデルにおいて文書に登場する単語の予測精度を表したもので，値が低いほど予測精度が高いことを示す．式 1 に定義を載せる．

$$\text{Perplexity} = \exp\left(-\frac{1}{\sum_{d=1}^{D} N_d} * \sum_{d=1}^{D}\sum_{i=1}^{N_d} \log(w_{di})\right) \quad (1)$$

上の式において，$D$ は全ツイート数，$N_d$ はツイート $d$ の単語数，$w_{di}$ はツイート $d$ の $i$ 番目の単語の尤度を表している．単語の尤度が高いほど，正しく単語の出現確率を学習しているということであり，モデルが正しく文書生成過程を表していると言える．各モデル間での Perplexity の比較を行った結果を図 5 に載せる．

Vol.2015-DPS-163 No.34
Vol.2015-MBL-75 No.34
2015/5/29

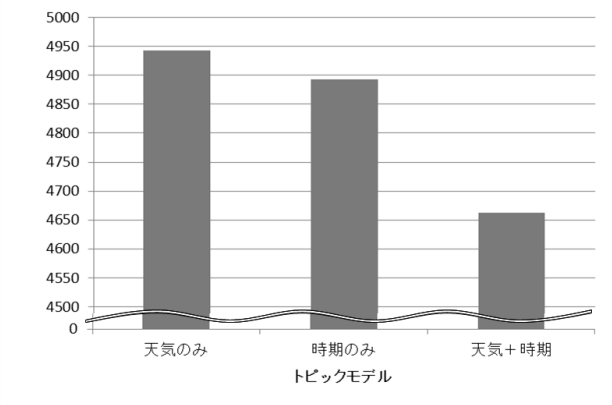


図 5　Perplexity の比較
Figure 5　Comparison of Perplexiy

縦軸が Perplexity，横軸がトピックモデルである．この結果を見ると，天気＋時期コンテキスト考慮モデルの Perplexity が，その他 2 つのトピックモデルにおける Perplexity よりも低くなっていることが分かり，天気と時期を両方考慮することによって性能が向上していることが分かる．

## 5. トピックモデルの質的評価

天気とツイートとの関係を学習した結果を載せる．表 3 は，天気コンテキスト考慮モデルによって学習した結果であり，表 4 は天気＋時期コンテキスト考慮モデルによって学習した結果である．なおここでは，ユーザの嗜好を推定するという研究の背景に基づき，「食べ物」にデータセットを限定して単語を抽出した．方法としては，「○○を食べる」または「○○食べる」という形で投稿された名詞のみを学習対象としている．

表 3　天気コンテキスト考慮モデルによる
天気クラスの単語分布
Table 3　Results of distributions of words associated with each weather class by the topic model considering weather context

| 天気クラス | | 2 | 4 | 7 |
|---|---|---|---|---|
| 気温平均 | | 31.28±1.29℃ | 9.29±0.61℃ | 23.45±3.72℃ |
| 湿度平均 | | 55.28±3.64% | 20.70±4.34% | 48.59±12.02% |
| 上位トピック | 1番目 | 3 | 33 | 3 |
| | 2番目 | 26 | 36 | 31 |
| | 3番目 | 79 | 49 | 26 |
| 出力された語 | 1番目のトピック | カレー | そば | カレー |
| | | バーガー | 寿司 | バーガー |
| | | おにぎり | コロッケ | おにぎり |
| | 2番目のトピック | 料理 | ケーキ | ラーメン |
| | | タイ | パフェ | 中華 |
| | | クレープ | カルビ | オムライス |
| | 3番目のトピック | うどん | 年越し | 料理 |
| | | チーズ | ポン酢 | タイ |
| | | カツ | アラモード | クレープ |

表 4　天気＋時期コンテキスト考慮モデルによる
天気クラスの単語分布
Table 4　Results of distributions of words associated with each weather class by the topic model considering both weather and season context

| 天気クラス | | 2 | 3 | 7 |
|---|---|---|---|---|
| 気温平均 | | 33.69±1.11℃ | 9.09±1.21℃ | 24.63±1.02℃ |
| 湿度平均 | | 46.63±5.21% | 44.98±9.84% | 23.86±2.27% |
| 上位トピック | 1番目 | 9 | 5 | 8 |
| | 2番目 | 5 | 34 | 19 |
| | 3番目 | 14 | 1 | 9 |
| 出力された語 | 1番目のトピック | アイス | カレー | カレー |
| | | 納豆 | うどん | ライス |
| | | お茶 | クレープ | たこやき |
| | 2番目のトピック | カレー | お菓子 | 野菜 |
| | | うどん | まんじゅう | チキン |
| | | クレープ | 屋台 | 天ぷら |
| | 3番目のトピック | かき氷 | お好み焼き | アイス |
| | | そば | ラーメン | 納豆 |
| | | スイカ | モンブラン | お茶 |

表では，各天気クラスとそれに対応するトピックとの対応関係が記されている．天気クラスには，そのクラスに属するツイートが投稿された際の平均気温，平均湿度が載っている．また，天気クラスに属するツイートに含まれる単語が，どのトピックに多く分類されたかという情報と，そのトピックに含まれる単語のうち，そのトピックに登場する回数の多い単語が表示されている．たとえば，表 3 において天気クラス 2 に分類されたツイートが投稿されたときの気温平均は 31.28℃であるということを意味し，天気クラス 2 に含まれるツイートに登場する名詞が分類されたトピックを登場回数の多い順に並べると，トピック 3，トピック 26，トピック 79 になる．さらに，トピックに含まれる単語を見てみると，トピック 3 に含まれる登場回数上位の単語として「カレー」「バーガー」「おにぎり」が含まれているということを表している．

気温が低いときの単語について考察する．表 3 において，天気クラス 4 は気温平均 9.29℃と寒いときのクラスであるが，天気クラス 4 の単語を見てみると，「そば」「ケーキ」「年越し」などの単語が含まれている．この「年越し」という単語は食べ物ではないが，形態素解析を行う際に「年越しそば」という単語を「年越し＋そば」と分解してしまったために登場していると推測できる．天気クラス 4 は，平均気温が 9℃と寒い時期に多く投稿されたツイートであるが，上記単語は寒さという天気的な要素によるものよりは，クリスマスや年末といった特定の季節に強く結びついた単語であると考えられるため，天気クラスに属する単語の学習結果としては適切でないと考えられる．一方，表 4 における寒いときのクラスは天気クラス 3 であるが，天気クラス 3 では，「そば」「ケーキ」など特定の季節に依存した食品は除かれており，寒いときに食べる傾向の高い食品が抽出されている．このような，時期コンテキストによるトピックを排除することにより，より天気コンテキストと

の関連度が高いトピックを抽出することができたと考えられる．

次に，気温が高いときの単語について考察する．表 3 で気温が高いクラスは天気クラス 2 であり，表 4 で気温が高いクラスは天気クラス 2 である．表 4 の天気クラス 2 では「アイス」「かき氷」「スイカ」など，気温が高いときに好まれる食品が学習されていることが分かる．

また，「カレー」「うどん」は表 3，表 4 ともに複数の天気クラスで登場しているが，これは登場回数の多い単語が各クラスでそのまま特徴語として登場しているためであり，こうして全時期に共通して登場する名詞を，特徴語から除外する機構を改良する必要があると考えられる．

## 6. 結論

トピックモデルを用いて天気・時期コンテキストと Twitter の発話内容の関係を調べることを目的とし，天気コンテキスト・時期コンテキストを組み込んだ複数のトピックモデルを提案し，比較を行った．

天気コンテキストとしては，気温・湿度・降水量など複数の要素が考えられるため，線形分類器による評価を行うことによって，気温と湿度がより Twitter の発話内容に影響を与える要素であることを見つけ，天気コンテキストとして採用した．

提案した複数のトピックモデルについて，量的な評価として，文書に登場する名詞の予測精度を表す Perplexity という指標を用いて評価を行い，天気と時期のコンテキストを組み合わせることによる予測精度の向上を示した．また，質的な評価として，実際に学習によって得られた天気クラスの単語分布を見比べることによって，天気と時期を組み合わせたモデルでは，天気とは直接関係のない時期のコンテキストによって影響を受けたと考えられる単語を排除することができ，より天気コンテキストと関連度の高いトピックを抽出することができた．

今後は，ユーザごとのコンテキストから受ける影響度の大きさの違いを考慮したモデルを構築したい．

## 謝辞

本研究成果の一部はグローバルプレナーズ（株）と東京大学の共同研究によって得られたものである．

## 参考文献

1) J. Eisenstein, B. O’Connor, N. A. Smith, and E. P. Xing: “A latent variable model for geographic lexical variation,” *Proc. of Conference on Empirical Methods on Natural Language Processing*, pp. 1277-1287, 2010.
2) D. Blei and J. Lafferty: “Dynamic topic models,” Proc. of International Conference on Machine Learning, Vol.23, pp. 113-120, 2006.
3) 深澤 佑介, 太田 順: “同行者に応じたトピックモデル,” 情報処理学会論文誌, Vol. 55, No. 1, pp. 413–424, 2014
4) 伊藤 拓, 深澤 佑介, 朱 丹丹, 太田 順, Tweet 内容に影響を与える気象条件と特徴語の抽出, 情報処理学会, 2014-MBL-73, No.1, 2014.
5) 気象庁: http://www.jma.go.jp/jma/
6) Stanford Classifier:http://nlp.stanford.edu/software/classifier.shtml