Vol.2014-NL-216 No.3
Vol.2014-SLP-101 No.3
2014/5/22

盛り上がり時間帯におけるツイートの言語的特性の解析

藤沼 祥成[1,2,a)]　横野 光[2,b)]　Pascual Martínez-Gómez[2,3,c)]　相澤 彰子[1,2,d)]

概要：あるイベントの盛り上がりに対して，それに関するツイートにも変化が現れその変化に着目することで盛り上がりを検出することが可能であると考えられる．本研究ではこの盛り上がり時間帯中のツイートに用いられている表現の特性を解析することを試みる．はじめに各時間帯のツイート集合とツイートより構築した言語モデルの関係をクロスエントロピーで算出した．実験結果より複数のハッシュタグ間における一部の盛り上がり時間帯のツイートはツイートより構築した n-gram 言語モデルに従うことを示す．また，盛り上がっている時間帯とそうでない時間帯において，クロスエントロピーにおいて統計的に有意差があることを示した ($p < 0.02$)．また，n-gram 言語モデルでは捉えられない素性も検討するため，Support Vector Machine (SVM) と Random Forest により各ツイートを盛り上がり時間帯の二値分類を行い，盛り上がり時間帯の特徴として漢字数が少ないことが明らかになった．

1. はじめに

近年マイクロブログの発展により，ソーシャルメディア上でユーザが，実際に遭遇したイベントなどに関してリアルタイムに言及することが増えてきた．2013 年現在では日に 5 億ものツイートが投稿され [*1]，これらのツイートの中にはイベント中の盛り上がりによって発言されるツイートも含まれる．このようなイベントの盛り上がりが Twitter ではツイート数の急激な増加によって現れることも多く，2010 年のワールドカップにおける日本対カメルーン戦においては，日本のゴールの瞬間に一秒あたり 2,940 ツイートという当時の一秒あたりの最高ツイート数を記録した [*2]．**図 1** に日本対カメルーン戦における一分あたりのツイート数の変化とツイート数が最も多くみられた上位 3 つの盛り上がり時間帯に何が起きたかを示す．

このようなイベントにおける Twitter 上の盛り上がりが検出できれば，実世界での重要な瞬間を捜す手がかりになる．それを利用した研究として，例えば Lanagan ら [11] は，スポーツの試合に関する要約の自動生成において，Twitter 上の盛り上がり時間帯を利用して重要な瞬間を捉える手法を提案している．

スポーツに限らず，Twitter 上のイベントにおける盛り上がりを検出する上で時間当たりのツイート数に着目した盛り上がりの指標が提案されている [8], [10], [19]．また，バースト検出 [7]，Topic Detection and Tracking[12], [15] を含む Twitter 上のタスクにおいては単位時間当たりのツイート数に焦点が当てられている．

ここで，従来の研究では盛り上がっている時間帯の言語的特性は注目されていない．しかし，**表 1** にあるように，盛り上がっているときのツイートと平常時でのツイートには，表現的な違いが見られる．これは盛り上がっているときにはユーザは興奮していることが多く，その感情がツイートにも反映されているためと考えられる．ツイート特有の言語的特性に関しては，Brody ら [2] が Twitter 上で文字の繰り返しが起こる単語は極性辞書に含まれる単語である割合が高いことを示しているが，実世界とのイベントとは関連付けられていない．

そこで，本稿では盛り上がっている時間帯における日本語のツイートの言語的特性を解析する．具体的には日本語ツイート全体よりサンプリングしたツイートをもとに言語モデルを構築し，盛り上がり時間帯のツイートと比較する．以下 2 章にて関連研究を，3 章にて使用した 3 つのデータセットについて述べる．4 章にてツイート言語モデルとクロスエントロピーに関して述べ，5 章では各ツイートにおける盛り上がり時間帯のツイートかどうかの二値分類を行い，その特徴量の解析を行う．6 章で結論を述べる．

[1] 東京大学
東京都文京区本郷 7–3–1
[2] 国立情報学研究所
東京都千代田区一ツ橋 2–1–2
[3] お茶の水女子大学
東京都文京区大塚 2–1–1
[a)] y_fujinuma@nii.ac.jp
[b)] yokono@nii.ac.jp
[c)] pascual@nii.ac.jp
[d)] aizawa@nii.ac.jp

[*1] https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

[*2] https://blog.twitter.com/2010/big-goals-big-game-big-records

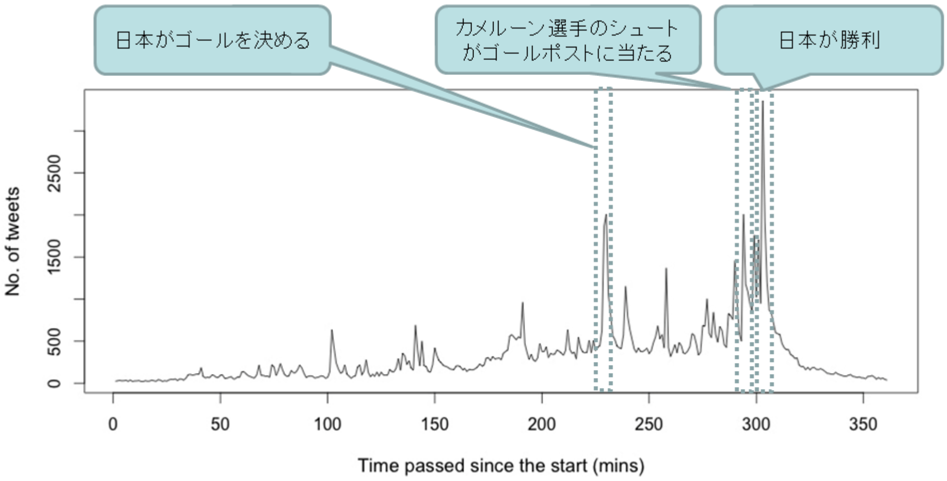

図 1　ワールドカップ 2010 における日本対カメルーン戦時のワールドカップに関連するハッシュタグ（#worldcup や#2010wc など）を含むツイート数の変化.

Fig. 1　The number of tweets per minute during Japan vs. Cameroon for hashtags related to World Cup 2010.

表 1　盛り上がり時間帯とその他の時間帯におけるツイート例

Table 1　Example of tweets from spikes and non-spikes.

	イベントの瞬間	ツイート例
盛り上がり時間帯	日本が勝った瞬間	やったあああああああああああああああああああああ #jpn #worldcup #2010wc
	日本がゴールを決めた瞬間	ゴーーーーーーーーーーーーーーーーーーーーール！！！！！！！！！ #2010wc
その他の時間帯	試合終了 50 分後	興奮してると見せかけて感動しすぎてずっと泣いてました．いや興奮はしてるけど信じてたから割と冷静でいられる #2010wc

2. 関連研究

本稿のタスクはイベント抽出，イベント要約，言語モデルの構築，そしてバースト検出を含む Twitter 上の様々なタスクと関連がある．

イベント検出は，地震の検出，株式市場の予測，スポーツの要約等の様々な分野において研究されている [1], [16]. Shen ら [18] はユーザに着目し，スポーツの要約生成に取り組んだ．また，Thelwall ら [19] は分野横断的にイベントの盛り上がり時間帯の極性を解析し，盛り上がりには負の極性が付随することを示した．

分野横断的にイベント検出やトピック検出を扱うには Twitter-LDA[20] 等のトピックモデルを利用し，各トピックを追跡する必要がある．しかしツイート中にハッシュタグが含まれる場合，ハッシュタグを用いてトピックを推定することができる．

また，近年ウェブやツイートより構築した言語モデルを利用することにより様々な性質を捉えられている．Danescu-Niculescu-Mizil ら [6] はオンライン・コミュニティにおけるユーザの利用年数とコミュニティ全体より構築した言語モデルとの間に相関があることを示した．Neubig ら [14] は日本語を含む 26 の言語においてリツイートされるツイートの情報量が多いことを示した．

盛り上がり検出では Kleinberg ら [10] のバースト検出等の先行手法が挙げられるが，このアルゴリズムはドキュメントの内容まで考慮していない．Diao ら [7] は TimeUserLDA を用いて各トピックのバーストを検出したが，文字列的特性に関しては考慮していない．

本稿ではイベントの盛り上がり時における言語的特性の解析に焦点を当てており，ハッシュタグを用いることによりトピックは検出済みであることとした．また，単一の言語モデルを構築することにより，イベントにおける一部の盛り上がり時間帯の共通の文字列的特性を持つことを示す．また言語モデルより算出したクロスエントロピーがなぜそのような性質を持つかを解析するため，ツイートの盛り上がり時間帯かその他の時間帯かの二値分類タスクにおける特徴量を解析した．

3. データセット

本稿で用いたツイートデータは Gnip*3 により収集されたものであり，以下の 5 つの区間に分かれている, 1) 2012-02-09〜2012-02-17, 2) 2012-03-21〜2012-03-22, 3) 2012-04-20〜2012-04-21, 4) 2012-05-18〜2012-05-19

本稿で対象としているツイートはイベントに対するユー

*3　http://gnip.com/

表 2　6 つのハッシュタグの統計とそのターゲット区間．時刻は全て UTC +0 である．
Table 2　Statistics of six hashtags and its target interval. All time are UTC +0.

ハッシュタグ	ジャンル	日時	ツイート数
#aibou	テレビドラマ	2012-03-21T10:31 - 2012-03-21T14:06	20,681
#hanshin	野球	2012-04-20T08:39 - 2012-04-20T12:54	6,176
#ACV	オンラインゲーム	2012-02-13T12:08 - 2012-02-13T15:41	1,562
#agqr	ラジオ番組	2012-02-15T11:39 - 2012-02-15T14:08	13,434
#figureskate	フィギュアスケート	2012-04-20T10:00 - 2012-04-20T12:20	1,410
#momoclo	音楽	2012-02-11T15:36 - 2012-02-11T17:21	1,823

ザからの反応である．そのため用いたツイートデータに関しては前処理としてボットからのツイートとリツイートを意味する文字列 'RT' を含むツイートを削除した．ボットのフィルタリングでは，Twiiter クライアントの情報を利用した．ツイート数が上位のクライアントにはボットが含まれにくいため，サンプリングした 3 日間におけるツイートに対して，クライアントから人の発言で用いられているものを人手で上位 43 クライアントを選択し，それらからの発言のみを利用した．このクライアントからのツイートはサンプリングした 3 日間全体のツイートの 90%以上を含んでいる．

以上のフィルタリングを施したデータより言語モデル構築用，クロスエントロピー算出用，特徴量解析用の 3 つのサブデータセットを構築した．

3.1　言語モデル構築用データセット

日本語ツイートの言語モデル (以下，ツイート言語モデルと呼ぶ) の構築に 2012 年 2 月 12 日中の全日本語ツイート，合計 413,008,939 ツイートよりサンプリングした 50,000 ツイートを用いた．

またツイート言語モデルとは別にハッシュタグ#aibou を含むツイートからなる言語モデル (以下，#aibou 言語モデルと呼ぶ) を 100 ツイートを超える時間帯のみからの計 15,663 ツイートより構築した．

3.2　クロスエントロピー算出用データセット

ハッシュタグのジャンルに関わらない，盛り上がり時間帯における共通の性質を解析するため，様々なジャンルのハッシュタグのツイートを収集した．ハッシュタグにおいて盛り上がりは実世界のイベントが起きている最中に起こりやすいため，その時間帯をハッシュタグにおけるターゲット区間とする．ターゲット区間は中央値の 10 倍を目安に設定し，ツイートを抽出した．また，盛り上がり時間帯以外の時間帯のツイートも含めるため，ターゲット区間の前後 20 分の時間におけるツイートもデータセットに含めた．**表 2** に実験に用いたデータセットの詳細を示す．

3.3　特徴量解析用データセット

各ツイートが盛り上がり時間帯のものか，その他の時間帯のものかどうかを判定する二値分類問題を考え，これに有効な特徴量から盛り上がり時の言語的特性を解析する．この実験のために 150 ツイートずつ盛り上がっている時間帯とそうでない時間帯より抽出した，偏りのないデータセットを構築した．具体的な抽出手順を以下に示す．

(1) 手動で設定したターゲット区間を 3 分割する．
(2) 各分割においてツイート数が上位の時間帯より，ツイートをテストデータとしてツイート数の合計が 50 件となるまで取得していく．
(3) 最後の時間帯中のツイートを合わせて，サンプリングしたツイートが 50 件を超える場合は 50 件になるようランダムサンプリングする．

4.　言語モデルによる解析

はじめにツイート言語モデルを構築し，各時間帯に対応するツイート集合とのクロスエントロピーを算出する．

4.1　言語モデルの構築

文字数 l のツイート t，ツイート中の文字 t_i に対し，t の確率を以下のように定義する．

$$P(t) = \prod_{i=1}^{l} P(t_i|t_{i-1}, ..., t_{i-n+1}).$$

CMU-Cambridge statistical language model toolkit [5] を用いて言語モデルを構築し，Katz バックオフスムージング [9] を用いた．

また，Neubig ら [14] が用いた手法であり，かつウェブ上の日本語において形態素解析の精度が十分でない [17], [21] ため，文字 7-gram モデルを採用した．

4.2　クロスエントロピー

Danescu-Niculescu-Mizil ら [6] の研究に準じて，あるツイート集合 T に対する言語モデルからの距離を数値化するため，クロスエントロピーを採用した．

クロスエントロピー H を以下に示す．

$$H(T) = -\frac{1}{N}\sum_{i}^{N} \log P(T_i)$$

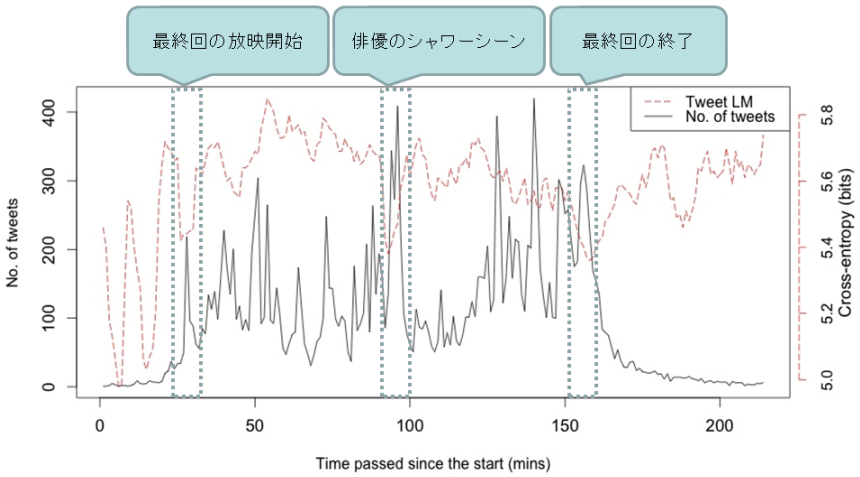

図 2　#aibou におけるツイート数の変化とツイート言語モデルを用いて算出したクロスエントロピーの変化.

Fig. 2　Cross-entropy between the general tweet language model and #aibou data.

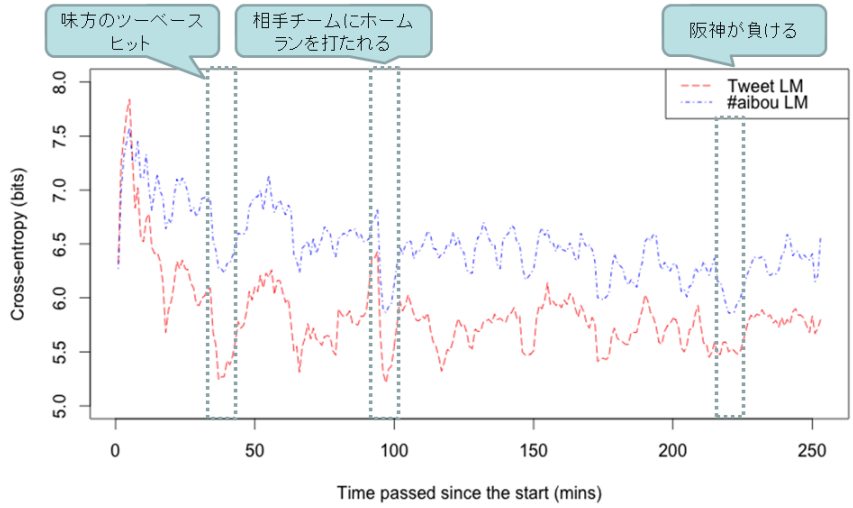

図 3　#hanshin データに対し，ツイート言語モデルと#aibou 言語モデルをそれぞれ用いてクロスエントロピーを算出した結果.

Fig. 3　Cross-entropy of #hanshin data compared with the general tweet language model and the #aibou language model.

なおクロスエントロピー算出時に前後 2 分間のツイートを含めた．これはクロスエントロピーを算出する際サンプリングエラーに弱いことが知られており [3]，サンプリングエラーに対して頑健にするためである．

4.3　クロスエントロピーの算出結果

各データに対するクロスエントロピーの算出結果を**図 2**，**図 3**，**図 4**，**図 5** に示す．

イベントの始まりと終わりの時間帯はクロスエントロピーが低くなることがわかる．図 2 では#aibou において「はじまた」等，イベントの始まりを示す多くツイートが多く見られた．

より詳細に解析するため，ハッシュタグ#aibou の盛り上がり時間帯のツイートだけからなる#aibou 言語モデルを構築した．図 3 は#hanshin に#aibou 言語モデルとツイート言語モデルをそれぞれ適用した結果を示す．#aibou 言語モデルとツイート言語モデル共にクロスエントロピーの変動が似たような傾向を示している．#aibou 言語モデルを適用した際にクロスエントロピーが最も低下した瞬間は 1) 相手チームがホームランを打った瞬間，2) 味方チームが負けた瞬間，である．一方ツイート言語モデルでは 1) 相手チームがホームランを打った瞬間，2) 味方チームがツーベースヒットを打った瞬間，である．#hanshin においては#aibou 言語モデルはより重要な場面においてクロスエントロピーが低下している．

しかし#aibou 言語モデルより算出したクロスエントロピーではツイート言語モデルで算出したクロスエントロピーと比較した結果，ツイート言語モデルのクロスエント

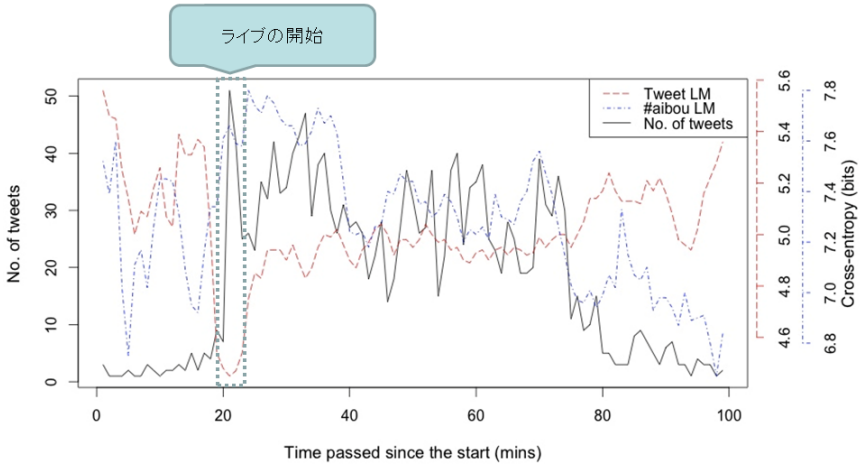

図 4 #momoclo データに対し，ツイート言語モデルと#aibou 言語モデルをそれぞれ用いてクロスエントロピーを算出した結果.

Fig. 4 The occurrence of #momoclo data shown together with the cross-entropy between the general tweet language model and the #aibou language model.

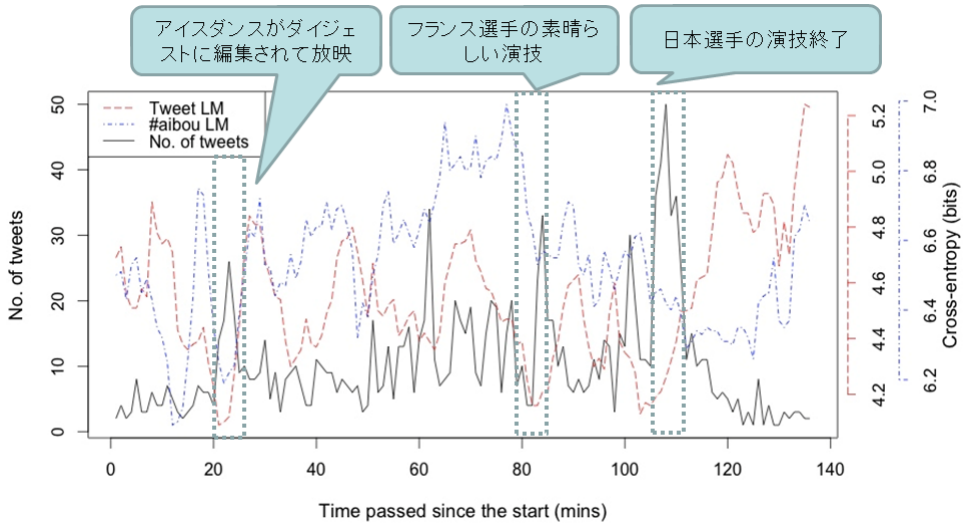

図 5 #figureskate データに対し，ツイート言語モデルと#aibou 言語モデルをそれぞれ用いてクロスエントロピーを算出した結果.

Fig. 5 The occurrence of #figureskate data shown together with the cross-entropy between the general tweet language model and the #aibou language model.

ロピーが低くなった盛り上がり時間帯の中で，低くならなかった時間帯がある．図 4 のイベントの始まった瞬間にツイート言語モデルではクロスエントロピーが低下したが，#aibou 言語モデルを用いた際には低下しなかった．このことから，#aibou 言語モデルはジャンルの異なるハッシュタグに対して盛り上がり時間帯を捉えることに向いていないことがわかる．

そのため，各ハッシュタグにおいて盛り上がり時間帯とその他の時間帯よりそれぞれ 150 ツイート，合計 300 ツイートをサンプリングした．サンプリングしたツイート集合に対し，ツイート言語モデルを用いて計算したクロスエントロピーを**表 3** に示す．表 3 の 4 列目に示した値は盛り上がり時間帯とその他の時間帯との二つのツイート集合に対して，Wilcoxon の順位和検定に算出した p 値である．また表 3 に示した 6 つのハッシュタグに対して，盛り上がり時間帯とその他の時間帯におけるクロスエントロピーの差に対し Wilcoxon の符号順位検定を適用した結果，統計的に有意差が存在した ($p < 0.02$)．n-gram 以外の特徴量も影響していると推測し，5 章でより詳細に解析した．

5. n-gram 以外の特徴の検討

n-gram 以外の特徴を考慮するために，教師ありデータを用いて入力ツイートが盛り上がり時間帯におけるものか否かの二値分類を行うモデルを構築した．採用した学習モデルは SVM と Random Forest であり，実装にはそれぞれ LIBSVM パッケージ [4] と RandomForest パッケージ [13]

表 3 各ハッシュタグの盛り上がり時間帯とその他の時間帯よりサンプリングしたツイート集合とツイート言語モデルを用いて算出したクロスエントロピー．
Table 3 Cross-entropy of sampled set of tweets from spikes and non-spikes computed using the general tweet language model.

ハッシュタグ	H(盛り上がり時間帯)	H(その他の時間帯)	H(その他の時間帯) - H(盛り上がり時間帯)	p 値
#aibou	5.45	5.58	0.13	$p < 0.3$
#hanshin	5.54	6.15	0.61	$p < 0.0004$
#ACV	5.44	5.61	0.17	$p < 0.06$
#agqr	4.22	5.11	0.89	$p < 10^{-9}$
#figureskate	4.18	4.78	0.60	$p < 10^{-11}$
#momoclo	4.73	5.17	0.44	$p < 0.00004$

を用いた．

5.1 使用した素性

使用した素性は以下の通りである．

(1) 漢字の数
(2) ひらがなの数
(3) カタカナの数
(4) ツイート中の合計文字数
(5) 各文字ユニグラムの数
(6) 同一文字の 3 文字繰り返しの数

Brody ら [2] の研究に基づき，盛り上がり時間帯において極性を持つ単語が現れやすいと推測したため，3 文字繰り返しの数を素性として入れた．また，学習データが 300 件と少ないので，$n \geq 2$ の n-gram 素性は扱わなかった．

5.2 結果と考察

二値分類において正しく盛り上がり時間帯のツイートとして分類した数と誤って盛り上がり時間帯のツイートとして分類した数を考慮するため評価指標として Accuracy (Acc.) を用いた．Accuracy の算出式を以下に示す：

$$\text{Acc.} = \frac{\text{正しく分類した数} + \text{誤って分類した数}}{\text{Total No. of tweets}}.$$

表 4 に SVM によるツイートの分類結果を示す．この実験では 1 ハッシュタグからサンプリングされたデータを学習データとして用い，残りの 5 ハッシュタグのデータをテストデータとした．

表 5 に各ハッシュタグに対して Random Forest を適用した結果算出された重要な特徴量の上位 5 件を示す．各特徴量に対して，Wilcoxon の順位和検定を適用した結果，統計的有意差を示した ($p < 0.01$)．Random Forest では n-gram 言語モデルでは捉えられなかった特徴量である漢字数が重要な特徴量であると捉えられている．5 ハッシュタグにおいて漢字数が上位 3 位以内にある．なお#agqr において漢字数は上位 5 位に含まれていないが，7 番目であった．また，文字ユニグラムは各ハッシュタグに依存していることがわかる．例を挙げると#ACV において上位 1 位，2 位にある文字ユニグラムはサーバが落ちたことによってツイート数が急激に増加したことに原因があるため，「落」と「ち」が上位に位置している．

表 4 より，#aibou データにおいてクロスエントロピーが低いにも関わらず，SVM による分類では高い性能を記録している．このことからクロスエントロピーのみでは盛り上がっている時間帯におけるツイートの性質を捉えきれていない，もしくはデータセットを構築する際のサンプリング手法においてノイズが含まれていると考えられる．

表 5 から盛り上がっている時間帯のツイートの性質として漢字数が少ないことが明らかになった．工藤ら [22] が述べているように，ユーザはツイートする時，漢字を入力するために Input Method Editor (IME) を利用しているため，ひらがなを入力する際より打鍵コストが多くかかる．そのため，漢字数が少ないという特徴は打鍵コストが関係しているのではないかと考えられる．

また，漢字の種類数はひらがな，カタカナと比較すると種類が多いことから一般的に高い情報量を保持している．しかし Web 上に用いられる言葉は特定の単語においてはひらがなが用いられやすいことが知られている [22]．例えば，「なく」という単語は「泣く」「無く」「鳴く」等の同音異義語が存在するが，即時性が要求される場面である盛り上がり時間帯では漢字が用いられにくい．以上のことから，盛り上がり時間帯におけるツイートは Twitter 全体からサンプリングしたツイートデータをもとに構築したツイート言語モデルに比べ，漢字数が少なく，クロスエントロピーが低い，と推測できる．

6. 結論

本稿では盛り上がっている時間帯のツイートはツイート言語モデルによって算出したクロスエントロピーが低いことを示した．SVM による分類結果より分類性能はハッシュタグに依存することがわかる．また，Random Forest において重要な特徴量の解析を行った結果，盛り上がっている時間帯のツイートにおいて漢字数が少ないことも示した．このことから盛り上がり時間帯におけるツイートは打鍵コストが関係しているのではないかと推測できる．

今後の課題として，1) クロスエントロピーが低いこと，

表 4　各ツイートが盛り上がり時間帯もしくはその他の時間帯から抽出されたかの二値分類の accuracy

Table 4　The accuracy of classifying whether a tweet is during spikes or non-spikes for each hashtag using SVM.

		Training data					
	hashtag	**#aibou**	**#hanshin**	**#ACV**	**#figureskate**	**#agqr**	**#momoclo**
Test data	#aibou	-	0.726	0.713	0.75	0.636	0.756
	#hanshin	0.680	-	0.663	0.713	0.606	0.713
	#ACV	0.680	0.633	-	0.663	0.633	0.610
	#agqr	0.626	0.620	0.636	-	0.620	0.633
	#figureskate	0.496	0.506	0.516	0.516	-	0.513
	#momoclo	0.653	0.550	0.686	0.686	0.550	-

表 5　Random Forest を各ハッシュタグに適用した際の上位 5 つの重要な特徴量.

Table 5　Variable importance of Random Forest classifier for each hashtag data.

Rank	**#aibou**	**#hanshin**	**#ACV**	**#figureskate**	**#agqr**	**#momoclo**
1	漢字数	見	落	漢字数	#	漢字数
2	ツイート中の全文字数	漢字数	ち	ぇ	q	機
3	space	く	漢字数	季	詞	ひらがな数
4	全	い	ひらがな数	ジ	ひ	来
5	機	ツイート中の全文字数	発	ス	か	ツイート中の全文字数

ならびに漢字数が少ないことが Twitter 上の他のタスクに有用か，2) 打鍵コストが影響し得る略語等の頻度がツイートの文字列的性質に影響しているか，の 2 点に関して検証することが挙げられる.

謝辞　本研究は JSPS 科研費 22240007 の助成を受けたものです.

参考文献

[1] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8 (2011).

[2] Brody, S. and Diakopoulos, N.: Cooooooooooooooollllllllllll!!!!!!!!!!!!!! Using word lengthening to detect sentiment in microblogs, *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., Association for Computational Linguistics, pp. 562–570 (online), available from ⟨http://www.aclweb.org/anthology/D11-1052⟩ (2011).

[3] Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D. and Lai, J. C.: Class-based n-gram models of natural language, *Computational linguistics*, Vol. 18, No. 4, pp. 467–479 (1992).

[4] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27 (online), available from ⟨http://www.csie.ntu.edu.tw/ cjlin/libsvm⟩ (2011).

[5] Clarkson, P. and Rosenfeld, R.: Statistical language modeling using the CMU-cambridge toolkit., *Eurospeech*, Vol. 97, pp. 2707–2710 (1997).

[6] Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J. and Potts, C.: No country for old members: User lifecycle and linguistic change in online communities, *Proc. of the 22nd International Conference on World Wide Web*, ACM, pp. 307–318 (2013).

[7] Diao, Q., Jiang, J., Zhu, F. and Lim, E.-P.: Finding bursty topics from microblogs, *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jeju Island, Korea, Association for Computational Linguistics, pp. 536–544 (online), available from ⟨http://www.aclweb.org/anthology/P12-1056⟩ (2012).

[8] Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A.: Information diffusion through blogspace, *Proc. of the 13th international conference on World Wide Web*, ACM, pp. 491–501 (2004).

[9] Katz, S. M.: Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 400–401 (1987).

[10] Kleinberg, J.: Bursty and hierarchical structure in streams, *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397 (2003).

[11] Lanagan, J. and Smeaton, A. F.: Using twitter to detect and tag important events in live sports, *Artificial Intelligence*, pp. 542–545 (2011).

[12] Lau, J. H., Collier, N. and Baldwin, T.: On-line trend analysis with topic models: #twitter trends detection topic model online, *Proc. of COLING 2012*, Mumbai, India, The COLING 2012 Organizing Committee, pp. 1519–1534 (online), available from ⟨http://www.aclweb.org/anthology/C12-1093⟩ (2012).

[13] Liaw, A. and Wiener, M.: Classification and regression by randomForest, *R News*, Vol. 2, No. 3, pp. 18–22 (online), available from ⟨http://CRAN.R-project.org/doc/Rnews/⟩ (2002).

[14] Neubig, G. and Duh, K.: How much is said in a tweet? A multilingual, information-theoretic perspective, *AAAI Spring Symposium on Analyzing Microtext* (2013).

[15] Petrović, S., Osborne, M. and Lavrenko, V.: Streaming First Story Detection with application to Twitter, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter*

of the Association for Computational Linguistics, Los Angeles, California, Association for Computational Linguistics, pp. 181–189 (online), available from ⟨http://www.aclweb.org/anthology/N10-1021⟩ (2010).

[16] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: Real-time event detection by social sensors, *Proc. of the 19th International Conference on World Wide Web*, ACM, pp. 851–860 (2010).

[17] Sasano, R., Kurohashi, S. and Okumura, M.: A simple approach to unknown word processing in Japanese morphological analysis, *Proc. of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, Asian Federation of Natural Language Processing, pp. 162–170 (online), available from ⟨http://www.aclweb.org/anthology/I13-1019⟩ (2013).

[18] Shen, C., Liu, F., Weng, F. and Li, T.: A participant-based approach for event summarization using Twitter streams, *Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, Association for Computational Linguistics, pp. 1152–1162 (online), available from ⟨http://www.aclweb.org/anthology/N13-1135⟩ (2013).

[19] Thelwall, M., Buckley, K. and Paltoglou, G.: Sentiment in Twitter events, *Journal of the American Society for Information Science and Technology*, Vol. 62, No. 2, pp. 406–418 (2011).

[20] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X.: Comparing Twitter and traditional media using topic models, *Proc. of the 33rd European Conference on Advances in Information Retrieval*, Berlin, Heidelberg, Springer-Verlag, pp. 338–349 (online), available from ⟨http://dl.acm.org/citation.cfm?id=1996889.1996934⟩ (2011).

[21] 斉藤いつみ，貞光九月，浅野久子，松尾義博：正規-崩れ表記のアライメントに基づく表記崩れパタンの抽出と形態素解析への導入，情報処理学会研究報告 NL214，pp. 1–9 (2013).

[22] 工藤　拓，市川　宙，Talbot, D.，賀沢秀人：Web 上のひらがな交じり文に頑健な形態素解析，言語処理学会第 18 回年次大会，pp. 1272–1275 (2012).