Twitterへの絵文字自動挿入システム

橋本 泰一
東京工業大学 総合プロジェクト支援センター
hashimoto.t.ab@m.titech.ac.jp

1　はじめに

近年，インターネットの普及とウェブサービスの発展にともない，ウェブが情報発信の場という側面に加え，コミュニケーションの場という側面を持ち始めた．特にブログやSNSといったウェブサービスはコミュニケーションの可能性を広げ，それまでつながることが困難であった人たち同士を結び付けることを可能にした．

最近注目されているコミュニケーションのためのウェブサービスの一つにTwitter [1] がある．Twitterはマイクロブログの一種で，投稿されたメッセージはツイートと呼ばれ，個人の他愛もないつぶやき（Tweet）をインターネットに公開するというウェブサービスである．一回に投稿できるツイートは140文字以内という制限があることが特徴的である．

また，Twitter内の他のユーザをブックマーク（フォロー）することができ，フォローしたユーザのツイートは自動的に自分のツイートの履歴（タイムライン）へマージされ表示される．一般にSNSではユーザへリンクを貼るには承認が必要であるが，Twitterのフォローは概念がブックマークに近く，一方的にリンクを貼ることができる．そのため，SNSよりもユーザのつながりを作り易いという利点を持つ．

Twitterが大きな人気を得た要因の一つは，早い段階でのWeb APIの一般公開，それにともなってスマートフォンなどの携帯電話におけるクライントが作成されたことである．一般に携帯電話では文字入力が困難であるが，140文字という短さとテキストのみという手軽さが携帯電話と相性がよく，いつでもどこでも簡単に投稿できるというメリットがある．さらに，Web APIが公開されることにより，短縮URL，Togetter，Twitopicなどの関連サービスも多数生まれている．

ユーザのつながりを作ることが容易であるために，Twitter上での情報の伝達がブログやSNSに比べ非常に速いという特徴を持つ．その情報伝達の速さを利用し，企業や商店が広告用のアカウントを作成し自身の製品情報やセール情報などを伝える手段として用いられることも多い．

一方で，日本では携帯電と使ったコミュニケーションとして，メールが早い段階で整備されてきた．そして，絵文字という日本独特の文字が生まれ，さらに携帯メールのコミュニケーションを豊かにしてきた．しかし，絵文字という文化は日本のみで，あまり諸外国には受け入れられていない．しかし，文字だけなく，絵もコミュニケーションの一部として利用することは全世界に通用する方法であると考えられる．

本研究では，Twitterに投稿されたツイートの表現をより豊かにするために，自動的に絵文字を挿入するシステムについて述べる．まず，ツイートをいくつかの部分文字列へ分割し，それぞれの文字列と類似した絵文字入りの文脈を検索する．そして，絵文字と類似文脈の統計値をもとに，絵文字の挿入位置と優先度を計算し，挿入する絵文字を決定する．

2　Twitterへの絵文字自動挿入システム「勝手にデコツィっ」

本研究では，Twitterのツィートに対して自動的に絵文字を挿入するシステム「勝手にデコツィっ」[1]について述べる．システムの概要を図1に示す．

2.1　対象となるツイート

対象となるツイートは公開タイムライン[2]，キーワードタイムライン[3]，ユーザタイムライン[4]である．公開タイムラインは，Twitter Streaming APIで取得可能な日本語のツイートの一部である．すべてのツィートを対象にしないのは，Twitter Streaming APIで取得できるツイートが膨大であり，システムのハード的な

[1] http://riverstone.star.titech.ac.jp/deco/
[2] http://riverstone.star.titech.ac.jp/deco/
[3] http://riverstone.star.titech.ac.jp/deco/#キーワード
[4] http://riverstone.star.titech.ac.jp/deco/@ユーザ名

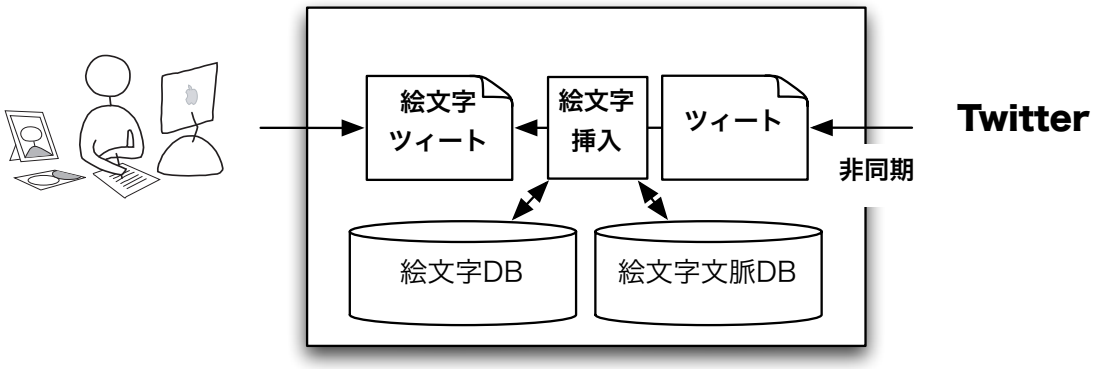

図 1: Twitter への絵文字自動挿入システム「勝手にデコツィっ」概要図

制約ですべてのツイートを処理することが困難であったためである．キーワードタイムラインは，指定されたキーワードを含む日本語のツイートである．ユーザタイムラインは，ある特定のユーザのツイートである．

2.2 絵文字挿入処理

Twitter から取得したツイートに対して，絵文字を挿入する処理について述べる．そのアルゴリズムの概要は以下のとおりである．

1. ツイートからの品詞 3 グラムの抽出
2. 絵文字を含む類似文脈の検索
3. 絵文字候補の選出
4. 絵文字候補のスコアの計算
5. 絵文字候補の選択

2.3 ツィートからの品詞 3 グラムの抽出

まず，半角空白を区切りとして，ツィートを分割する．そして，分割した文字列が「RT」,「@」「#」で始まる文字列，URL である場合には，絵文字挿入の対象としない．なぜならば,「RT」は，Twitter においてそれ以後の文字列は他のユーザのツイートの引用（リツイート，ReTweet）を表す特別な文字列,「@」で始まる文字列はユーザ名,「#」で始まる文字列は「ハッシュタグ」と呼ばれるツイートを分類するタグを表すためである．絵文字挿入の対象となった文字列は，MeCab と IPADic によって形態素解析を行い，品詞 3 グラムへ分割する．

例えば，「台風くんのか…お盆なのに厄介だな。#taifu」というツイートの場合,「#taifu」が処理対象外となり,「台風くんのか…お盆なのに厄介だな。」が形態素解析され，単語 3 グラムへ分割される．

例: 台風くんのか…お盆なのに厄介だな。 #taifu

↓

台風	くん	の
くん	の	か
の	か	…
か	…	お盆
…	お盆	な

…

2.4 絵文字を含む類似文脈の検索

前節で抽出した単語 3 グラムを用いて，絵文字を含む類似文脈を検索する．類似文脈の検索エンジンには SimString [2] を用い，絵文字を含む文脈情報は Baidu 絵文字入りモバイルウェブコーパス [3] を用いた．

Baidu 絵文字入りモバイルウェブコーパスは，絵文字を含む単語 1 グラムから 5 グラムのコーパスである．このコーパスより絵文字を含む単語 N グラムを抽出し，絵文字を除いた単語列を SimString を用いて，インデックス化する．(絵文字文脈 DB)

ツイートから抽出した単語 3 グラムと類似した文脈を SimString を使って Baidu 絵文字入りモバイルウェブコーパスから検索する．

図 2: 「勝手にデコツィっ」スクリーンショット

例: 台風くんのか…お盆なのに厄介だな。 #taifu

↓

台風　くん　の

　　類似文脈: 台風の

くん　の　か

　　類似文脈: んので

　　類似文脈: くもんの

　　類似文脈: くれんの

の　か　…

　　類似文脈: のか

…

2.5 絵文字候補の選出

類似文脈検索で検索された文脈に挿入されていた絵文字を挿入絵文字候補として選出する．Baidu 絵文字入りモバイルウェブコーパスを用いて，絵文字の統計データのデータベースを構築する．(絵文字 DB) 絵文字 DB は，以下の 6 項目についてデータベース化する．

- 類似文脈： c'
- 絵文字の種類： e
- 絵文字位置： $Posi(c')$
- 類似文脈の文字列長： $Len(c')$
- 絵文字頻度： $Freq(e)$
- 類似文脈の頻度： $Freq(c')$

類似文脈として検索された文字列を使って，絵文字 DB の文脈 c を検索し，絵文字候補を選出する．

2.6 絵文字候補のスコアの計算

類似文脈 c' 挿入する絵文字の候補 e とから実際の文脈 c における絵文字の挿入位置 $Posi(c', e)$ と優先度 $Prio(c', e)$ を計算する．

$$Posi(c', e) = \frac{Posi(c')}{Len(c')} * 3 + 2 \quad (1)$$

$$Prio(c', e) = \frac{\log Freq(e)}{Freq(c') + 1} \quad (2)$$

$$(3)$$

まず，類似文脈 c' と絵文字候補 e から文脈 c のおける絵文字の挿入位置 $Posi(c', e)$ を式 (1) を用いて計算し，絵文字のスコア $Prio(c', e)$ を計算する．

例: 台風くんのか…お盆なのに厄介だな。 #taifu
↓
台風　くん　の
類似文脈: 台風の
候補: E005,1,3,10,1480
くん　の　か
類似文脈: んので
候補: EB5B,2,3,70,7262
類似文脈: くもんの
候補: E546,0,3,18,13744
類似文脈: くれんの
候補: EB05,3,4,12,5058
の　か　…
類似文脈: のか
候補: EB5A,2,3,20,1566
候補: EB5B,2,3,65,7262
候補: EB5C,2,3,22,1969
…
…

例えば，文脈「台風くんの」において，類似文脈「台風の」絵文字候補「E005」の場合，絵文字 DB より以下の値が取り出せ，

$$\begin{aligned} Posi(c') &= 1 \\ Len(c') &= 3 \\ Freq(e) &= 10 \\ Freq(c') &= 1480 \end{aligned}$$

挿入位置と絵文字候補の優先度は，

$$\begin{aligned} Posi(c', e) &= 3 \\ Prio(c', e) &= 0.0015 \end{aligned}$$

と計算される．つまり，「台風くんの¡E005¿」という表現のスコアが 0.0015 であると計算される．

2.7　絵文字候補の選択

先の絵文字候補の優先度をもとに，文脈 c におけるある位置 p の絵文字候補 e のスコア $Score(c, e, p)$ を計算する．文脈 c の各位置で，その位置に挿入されると予想された絵文字候補のスコアの和を，その位置におけるその絵文字候補の優先度とする．

$$\begin{aligned} Score(c, e, p) &= \sum_{c' in C} Prio(c', e) \qquad (4) \\ C &= \{c' | Posi(c', e) = p\} \end{aligned}$$

各位置での最もスコアが高かった絵文字を挿入する．ただし，ある特定の品詞列の場合，絵文字は挿入しない．その規則を下記に示す．

- 「助詞」と「助詞」の間
- 「名詞, 非自立」の後

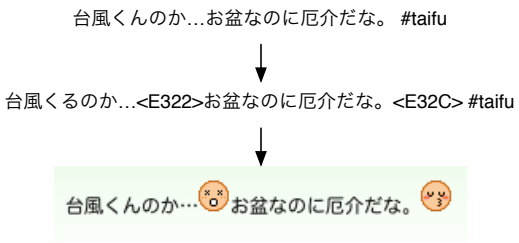

図 3: 絵文字が挿入されたツイートの例

3　まとめ

本研究では，マイクロブログ Twitter のツイートに対して絵文字を挿入するシステムについて述べた．Twitter では，日本の携帯電話のメールや携帯サイトで用いられている絵文字をツイート内で使用することができない．本システムは，Twitter へ投稿されたツイートに対して，絵文字を自動的に挿入する．まず，ツイートをいくつかの部分文字列へ分割し，それぞれの文字列と類似した絵文字入りの文脈を検索する．そして，絵文字と類似文脈の統計値をもとに，絵文字の挿入位置と優先度を計算し，挿入する絵文字を決定する．

本研究では，絵文字を挿入したツイートに対する評価を行っていない．どの絵文字を挿入することが正しいのか決めることは非常に困難であるが，今後，評価できるように検討したい．また，絵文字自身が持つ意味を活用した応用研究についても検討していきたい．

参考文献

[1] Twitter. http://twitter.com/.

[2] 岡崎直観, 辻井潤一. 高速な類似文字列検索アルゴリズム. 情報処理学会創立 50 周年記念全国大会, pp. 1C–1, 2010.

[3] 萩原正人, 大原一輝, 水野貴明, 橋本泰一, 荒牧英治, 竹迫良範. 「不自然言語処理コンテスト」第 1 回開催報告. 言語処理学会第 17 回年次大会, 2011.