系列ラベリングによる顔文字の自動抽出と顔文字辞書の構築

渡邉 謙一†　高橋 寛幸‡　但馬 康宏†　菊井 玄一郎†
†岡山県立大学　情報工学部
‡ＮＴＴレゾナント株式会社
{cb24031r, tajima, kikui} @cse.oka-pu.ac.jp, h-taka @nttr.co.jp

1. はじめに

ブログやマイクロブログの普及に伴い，これらに投稿されている膨大なテキスト（記事）を分析することで企業や社会全体の効率化に活用しようとする研究が活発化している．これらの殆どが形態素解析の利用を前提としていることから，形態素解析の精度は十分に高い必要がある．しかしながら，現在の多くの形態素解析処理は新聞記事を中心としたコーパスに対しては極めて高い精度(95%以上)で解析できるものの，言語的な性質の異なるブログやマイクロブログに適用した場合，精度が低下してしまう．この原因として，新語や略語，表記の大幅な揺れ（ひらがな化や小文字化，長音化など），顔文字などがある．新語についてはウエブからの自動獲得手法([1],[2],[5]など)，表記のひらがな化については表記のゆれを考慮した形態素解析手法（[3],[5]など）が提案されている．一方，顔文字については juman-7.0[1]において顔文字辞書のよる識別が行われているほかは，特段の処理が行われておらず，しばしば前後の単語と顔文字の一部が結合した誤まった解析結果が出力される．また，前者においても後述するようにマイクロブログに出現する顔文字すべてに対応しているわけではない．

一方，顔文字の抽出に特化した処理として，HMM を使った顔文字抽出が公開されているが[2], web 上のアプリケーションなので解析プログラムとして利用するのは難しい．

そこで，本稿では，テキスト解析前処理としての顔文字箇所の同定，および，形態素解析器における顔文字辞書の拡充を目的として，まず，系列ラベリングによる顔文字の自動抽出処理を構築し，これを用いて大量のマイクロブログテキストから顔文字辞書の構築を試みる．

2. 関連研究

顔文字を単語の一種と考えると，テキストからの顔文字抽出は特定カテゴリの語彙抽出と考えることができる．村脇ら[1]は周辺の文字や単語の頻度統計によって「ハブる」などのカタカナ用言や異表記語などを抽出する手法を提案している．一方，福島ら[3]はカタカナ用言を，後続文字列を素性とする SVM により自動獲得している．これらは各カテゴリの周辺に共起する語彙分布の特徴を利用しているが顔文字ではこのような特徴は存在しない．また浅原ら[4]らは未知語の付近で形態素解析の候補が特異な振る舞いをすることを利用して未知語の自動獲得を行っているが，顔文字では形態素の候補が殆んど存在しないためこの手法は利用できない．

3. 提案手法

そこで本研究では入力文字列から CRF による系列ラベリングを用いて顔文字の抽出を行う．具体的には入力の各文字に対して，1)FACE-B(顔文字の先頭文字), 2)FACE-I(顔文字の 2 文字目以降), 3) TEXT(顔文字以外のテキスト文字), 4) EOS(文末)の 4 つのラベルのいずれかを自動で付与し，FACE-B から始まって FACE-I が 0 文字以上連続している文字列を顔文字として抽出する．

CRF の素性としては，入力された文字そのものと表 2 に示す文字種の情報を用いる．素性と正解ラベルの例を図 1 に示す．素性関数としては以下に示すものを用いた([]内は図 1 の文字位置 2 に対する素性を示す).

1.　当該位置のラベルと以下のそれぞれの素性の組

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[2] http://www.nishiohirokazu.org/kaomoji/kaomoji.py

1-1) 当該位置の直前，その直後の文字および文字種
[っ, (, ´, CHARA, SYMBOL, SYMBOL]
1-2) 当該位置及び直前の文字の組，および，文字種の組
[っ/(, CHARA/SYMBOL]
1-3) 当該位置及び直後の文字の組，および，文字種の組
[(/´, SYMBOL/SYMBOL]
2. 当該ラベルと直前のラベルの組(バイグラム素性)

位置	文字	文字種	ラベル
	…	…	…
1	っ	CHARA	TEXT
2	(	SYMBOL	FACE-B
3	´	SYMBOL	FACE-I
4	Д	SYMBOL	FACE-I
5	`	SYMBOL	FACE-I
6	¥s	SYMBOL	FACE-I
7	)	SYMBOL	FACE-I
8	EOS	EOS	EOS
	…	…	…

図 1　与える素性とラベル

表 1　文字種の割り当て

素性値	文字種
CHARA	平仮名　片仮名 漢字　アルファベット
NUM	数字
EOS	EOS
SYMBOL	CHARA，NUM 以外の文字

表 2　ラベルの一覧

ラベル	ラベルの意味
FACE-B	顔文字の開始地点
FACE-I	顔文字の内側
TEXT	顔文字の外側
EOS	文の終端

4. 抽出実験

4.1. コーパスと評価尺度

Twitter 社の streaming API を使い，2011 年 10 月から 2012 年 10 月の間の public timeline から日本語文字を含む記事(tweet)を収集して「基本コーパス」とした．収集した総記事数は約 5,000 万(50,277,727)記事である．

これらの中から下記を満たす記事をそれぞれ 1,000 個ずつランダムに抽出し学習データとした．

1)　全角または半角の括弧で囲まれた文字列を含む記事
2)　記号類を全く含まない記事

前者に対しては顔文字を人手で抽出し，ラベルを付与した．後者には顔文字が全く含まれていないと考えて TEXT，EOS のみを自動で付与した． 1)のように括弧を含むツイートに絞ってサンプリングを行ったのは，顔文字の出現密度を高めて作業の効率化を図るためである．一方，顔文字でない部分のデータが相対的に減少した分を補償するために 2)のサンプルを加えた．

評価データは 2 種類用意した．一つ目は学習データを除く基本コーパスから学習データと同じ方法で収集した 1)，2)500 個ずつの合計 1000 個であり，これを評価データ A とする．二つ目は基本コーパスから，学習データ，評価データ A を除いてランダムに 500 個選出したものであり，これを評価データ B とする．

評価尺度として，以下の式で与えられる適合率(precision)，再現率(recall)，F 値(F-measure)を使用した．

$$precision = \frac{R}{N}$$

$$recall = \frac{R}{C}$$

$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

ここで R は出力された顔文字の正解数，N は出力された顔文字数，C は評価データ内に存在する顔文字数とする．なお，正解数とは顔文字の始点と終点が正解データと完

全に一致するものである．

4.2. 実験結果と考察

上記の教師データを使って CRF のパラメータを学習し，評価データに適用した．なお，CRF のツールは CRF++[3]を用いた．

評価データ A の中には 403 個の顔文字が存在し，本手法によって，396 個の顔文字を抽出した．このうち 362 個の顔文字が正しい顔文字であった．

評価データ B の中には 115 個の顔文字が存在し，同様に顔文字の抽出を行った結果，112 個の顔文字を抽出した．このうち 105 個の顔文字が正しい顔文字であった．

上記の結果より，適合率，再現率，F 値を算出したものを表 3 に示す．

表 3　顔文字の出力精度

	評価データ A	評価データ B
適合率(R/N)	0.911(362/396)	0.937(105/112)
再現率(R/C)	0.895(362/403)	0.913(105/115)
F 値	0.903	0.925

表 3 より評価データ A の F 値が 90.3%，評価データ B の F 値が 92.5%となり，高い精度で獲得できていることがわかる．また獲得できた顔文字の例を以下に示す．

＼(^o^)／　　(*^^*)　　(／ω・＼)

(ρ＿;)　　_(:3 」∠)_　　(´・ω・`)

．・゜・(ノД`)・゜・．　(　´　▽　`)ﾉ　　(-．-;

獲得した顔文字から顔以外に「手」がついてある顔文字，左右非対称な顔文字なども獲得できていることが分かる．

次に正しく抽出できていない特徴と顔文字例を以下に挙げる．以下では<FACE></FACE>内で囲まれている文字列が自動抽出した顔文字の範囲である．

A)　顔文字の一部が切れている
- m<FACE>(＿)m</FACE>
- <FACE>o(*⌒―⌒*)</FACE>o

B)　顔文字の前後にある文字の一部が顔文字に含まれている
- <FACE>(゜Д゜)』</FACE>
- <FACE>傷´；ω；)</FACE>

C)　顔文字でない文字列を顔文字としている
- <FACE>(　)(　)</FACE>
- <FACE>＾＾＾＾＾＾</FACE>

D)　複数の顔文字を一つの顔文字としている
- <FACE>(-_-)(*^^*)</FACE>
- <FACE>(O.O;)(oo;)</FACE>

E)　顔文字であるべき文字列を顔文字としていない
- `・ω・）b
- ∩＾ω＾∩

これらの抽出誤り例から顔文字の開始部分に括弧が存在しない顔文字や，顔文字の手を表す部分にアルファベットなどの文字を使っている顔文字の一部が正しく抽出できていないことが分かる．これは学習データとして括弧で囲まれた文字列を含む tweet を重点的に利用しており括弧で囲まれていない顔文字の学習例が少なかったためと考えられる．

5. 顔文字辞書の構築

提案手法を大規模な web コーパスに適用することにより，顔文字辞書の構築を行った．使用したデータは 4. で収集した基本コーパス全て(約 5,000 万記事)である．抽出した顔文字のうち一定の閾値以上の頻度のものを収集し，顔文字辞書とした．収集数は閾値 500，100，5 でそれぞれ異なり顔文字数が 1,487 個，5,013 個，56,523 個となった．

抽出した全ての顔文字のうち，出現頻度上位 1,000 個(1,000 個目の頻度は 812)を対象に適合率を算出すると 95.6%であった(956 個が正解)．このことから出現回数の多い顔文字は高い精度で獲得できていることがわかる．再現率については顔文字の全体が不明なので正確な評価は難しいが，参考として，juman-7.0 に含まれている顔文字辞書(972 個)のうちどの程度の顔文字が自動抽出結果の顔文字と一致したかを算出した．なお，juman-7.0 は全角文字のみを対象としているため，自動抽出した顔文字をすべて全角に統一して計算した．

F500)　出現回数が 500 回以上の顔文字 1302 種類

[3]http://crfpp.googlecode.com/svn/trunk/doc/index.html

F100） 出現回数が 100 回以上の顔文字 4199 種類
F005） 出現回数が 5 回以上の顔文字 46014 種類
表 4 に自動抽出した顔文字辞書の上記のパターンと JUMAN7.0 に実装されている顔文字辞書との一致顔文字数と一致率を示す.

表 4　juman-7.0 顔文字辞書に対する一致率

	一致数	一致率
F500)	101	0.103
F100)	160	0.164
F005)	289	0.297

表 4 の結果を見ると閾値を下げて 46000 個収集してもカバー率は 29.7%と非常に低い結果となった.

一致していない顔文字は細かいバリエーションによるものが多いが，特徴としては次に例を示すような西洋式の顔文字があげられる.

:）　　:－（　　:－P　　（:.）彡

このような西洋式の顔文字は juman の辞書に 75 個存在するが，本手法で作成した顔文字辞書にはほとんど含まれていなかった．これは，本手法の顔文字の収集対象が日本語の Twitter であること，学習コーパスとして括弧で囲まれているものを用いたことなどが理由と思われる. 一方， juman の辞書にない顔文字が作成した辞書に大量に存在する理由は些細なバリエーションによるものと思われる.

ここでテキスト解析の際の顔文字同定について考えてみよう．一つの方法は今回の抽出手法のようにモデルによって動的に顔文字区間を検出する方法であり，もう一つの方法は juman-7.0 のように形態素解析辞書で扱う方法である．後者の方が形態素解析に無理なく組み込める上に解析結果をコントロールしやすいため，本研究では後者の辞書を自動構築することを想定していた．しかしながら，顔文字のバリエーションが予想外に多いことから，辞書に顔文字全てを登録して入力との完全一致で解析するのは網羅性の点で疑問が残ることも分かった.

形態素解析と融合する一つの方策として森らの「点予測による自動単語分割」[7]との結合が考えられる．本研究で用いた素性は基本的に森らの用いた素性に含まれるのでこれらを同時に実行することは可能と思われる．但し，本研究で有効であったラベルのバイグラムが森らの手法では利用できないのが懸念点である.

6. おわりに

本稿では CRF を用いた系列ラベリングによる顔文字の自動抽出する方法を実装し，大量のマイクロブログから顔文字辞書の自動構築を行った．自動抽出の性能は適合率が 94%，F 値が 0.93 と，高い精度が達成できることが分かった．またマイクロブログ 5 千万記事から顔文字の自動抽出を行った結果，頻度 5 以上で 5 万 6 千種類もの大量の顔文字が収集できることが分かった．今後はさらに精度の向上を目指すほかに，抽出した顔文字の感情推定を行っていきたいと考える.

謝辞 : twitterAPI を公開されている Twitter 社に謝意を表します.

参考文献

[1] 村脇有吾, 黒橋禎夫. 日本語未知語のテキストからの自動取得. 信学技報, vol. 111, no. 119, NLC2011-8, pp. 37-42, 2011-7
[2] 福島健一, 鍛治伸裕, 喜連川優. 機械学習を用いたカタカナ用言の獲得. 言語処理学会　第 13 回年次大会　発表論文集. pp.815-818, 2007.3.
[3] 工藤拓,市川宙,David Talbot,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　発表論文集. pp.1276-1279, 2012.3.
[4] 浅原正幸,松本祐治. 形態素解析とチャンキングの組み合わせによる日本語テキスト中の未知語出現箇所同定. 情報処理学会研究報告. 自然言語処理研究会報告 2003(23), 47-54, 2003-03-06
[5] 柴田知秀, 村脇有吾, 黒橋禎夫, 河原大輔. 実テキスト解析をささえる語彙知識の自動獲得. 言語処理学会　第 18 回年次大会　発表論文集. pp.81-84, 2012.3.
[6] 篠山学, 松尾朋子. 顔文字を考慮した対話テキストの感情推定に関する研究. 香川高等専門学校研究紀要 1, 151-153, 2010-06
[7] 森信介, ニュービッグ　グラム, 坪井祐太.点予測による自動単語分割. 情報処理学会論文誌　Vol.52, No.10　pp.2944-2952, 2011.10.