

顔文字推薦のための感情辞書自動生成手法

鬼沢 和也[†] 田中 友樹[‡] 山名 早人[§]

[†] 早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1

[‡] 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

[§] 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-1

E-mail: {kazuya_o, ytanaka, yamana}@yama.info.waseda.ac.jp

あらまし 携帯電話やスマートフォンの普及に伴い、テキストによるコミュニケーションを行う機会が増え、顔文字がより頻繁に利用されている。しかし、顔文字の種類は多く、適切な顔文字を選ぶには手間がかかるという問題がある。そこで、入力文に対し適切な顔文字を推薦し、顔文字の入力を補助するという研究が行われている。既存研究では、入力文から使用する顔文字のカテゴリを推定することにより、顔文字を推薦する。使用する顔文字のカテゴリの推定は、極性（ポジティブ、ネガティブ、なしのいずれか）およびカテゴリを手で付与した学習データを基に行っている。しかし、人手で作成した学習データを用いるという方法には、学習データが少なく、学習データ中の単語が入力文に存在しない場合がある。そこで本論文では、人手ではなく、自動的に単語のカテゴリを評価することにより、対応可能な単語数を多くする手法を提案する。

キーワード 顔文字, 顔文字推薦, 感情推定

1. はじめに

近年、Twitter や facebook をはじめとするサービスの出現に伴い、テキストによるコミュニケーションを行う機会が増加している。しかし、テキストによるコミュニケーションでは、会話のように声のトーンや身振り手振り、表情等を伝えられない。そこで、テキストのみでは伝えづらい情報を伝えるために顔文字が利用されている。顔文字とは「(^*)」や「ε=ε=ο(*^▽)ο」のように、文字や記号を組み合わせて表情や動作を表したものである。

顔文字がインターネットを介して初めて使用されたのは 1982 年と言われている[1]。現在では携帯電話やスマートフォンを始めとする、コンピュータの入力システムに顔文字が組み込まれており、顔文字は一般的に使用されている。顔文字の入力方法としては、一文字ずつ手動で入力する方法、コピーアンドペーストで貼り付ける方法、入力システムに登録しておく方法、専用のアプリケーションを利用する方法がある。しかし、入力システムに登録しておく方法以外は顔文字の入力に手間がかかる。また、入力システムに登録する方法の場合は、どのキーワードが各顔文字と対応しているか記憶しておかなければならない。さらに、あらかじめ登録した顔文字しか使用できないという問題があり、こうした問題を解決するため、顔文字の入力を補助する研究がされている。

顔文字の入力を補助する研究として Suzuki ら[2]と、江村ら[3]の研究がある。Suzuki らは文章の感情を分析し、その感情に対応する顔文字を付与する手法を考案

した。文章の感情分析は、感情を表す語と、その語が表す感情をセットにした辞書を作成し、その辞書を基に文章が表す感情を分析する。江村らは、入力文から使用する顔文字のカテゴリを推定することにより、顔文字を推薦する方法を考案した。使用する顔文字のカテゴリの推定は、手がかり語辞書とコーパスを基に行っている。手がかり語辞書には 1,440 語の単語が含まれており、コーパスには 3,887 件の文が含まれている。各語と文には極性（ポジティブ、ネガティブ、なしのいずれか）およびカテゴリが設定されている。

しかし、Suzuki らと江村らの手法には、文章中にあらかじめ登録した語が存在しない場合がある。そこで本稿では、単語の評価を自動化し、既存研究より多くの単語に対応できる手法を提案する。

具体的には、初期データセットとして既知の顔文字のカテゴリとそのカテゴリの顔文字を登録しておく。そして、各カテゴリの顔文字を含むツイートを Twitter より収集し、そのツイートに含まれる単語を当該カテゴリに関連する語としてデータベースに登録する。登録にあたっては、単語情報として、当該顔文字と共に出現する確率も登録する。入力文が与えられると、データベースの情報を基に使用する顔文字のカテゴリを推定し、適切なカテゴリの顔文字を推薦する。既存手法では、学習データ中の単語が入力文中に存在せず、適切な顔文字を推薦できないという問題があったが、提案手法では、収集したツイートに含まれる全ての単語を自動的に評価できるため、学習データ中の単語が入力文中に存在しない場合に適切な顔文字を推薦でき

ないという問題を解決することができる。実験により、入力文に学習データ中の単語が入力文に存在する割合と推薦の正解率を既存手法と比較する。

本稿では以下の構成をとる。まず2節で関連研究について述べ、3節で提案手法について述べる。次に4節で実験と評価を行い、最後に5節でまとめを述べる。

2. 関連研究

本節では既存研究である、入力文に対し顔文字を付与、または推薦する研究について説明する。2.1項で入力文に対し顔文字を付与する Suzuki らの研究[2]、2.2項で入力文に対し顔文字を推薦する江村らの研究[3]について述べる。最後に、2.3項でまとめる。

2.1. Suzuki らの研究

Suzuki らは、入力文の感情を分析し、その感情に対応する顔文字を付与するという方法により、顔文字付与システムを実現した。定義されている感情は、Plutchik.R[4]の定義する基本的な感情であり、例えば、喜びや怖れ、驚きなどと、実際に収集した文章において顔文字がよく付与されている「疲れた」や、「眠い」といった感情である。入力文の感情分析は、形態素と感情をセットにした静的な辞書をあらかじめ人手で作成し、入力文に現れる形態素から感情を分析することにより行った。辞書に含まれる形態素は、名詞の一部と、形容詞の一部である。

2.2. 江村らの研究

江村らの研究では、まず、コーパスと手がかり語辞書を作成する。コーパスとは、収集したツイートにカテゴリを付与したものである。手がかり語辞書とは、入力文の属するカテゴリを推定する手がかりになる語を集めた辞書である。このコーパスと手がかり語辞書を用いて入力文のカテゴリ推定し、対応する顔文字を推薦する。推定に使用するカテゴリは、Twitter のつぶやき 1,722 件を分類することにより決定した、37 カテゴリである。入力文のカテゴリ推定は、次の手順で行なっている。

1. 手がかり語辞書に含まれる語の、入力文中における出現頻度を計算する。
2. 1 の結果から得られた手がかり語の出現頻度を求め、カテゴリ毎に素性ベクトルを構築する。
3. コーパス中の文章との類似度を計算する。
4. 類似度の高い上位 k 件が、最も多く属しているカテゴリを推定結果とする。

この手法がユーザの入力した文に対し、どの程度正確に顔文字を推薦できるか実験したところ、68.4%の文章に対し、適切な顔文字を推薦することができた。

2.3. 既存研究のまとめ

既存の研究[2][3]は、どちらもあらかじめ用意した辞書の情報を利用し、入力文に顔文字を付与または推薦している。しかし、人手で辞書を構築しているため、登録件数が少なく、辞書に含まれている形態素や語が入力文に存在しない場合がある。この問題を解決するためには、辞書に登録する形態素や語を増やすという方法がある。しかし、人手で登録する形態素や語を増やすには手間がかかるという問題がある。

3. 提案手法

本節では、既存手法では人手で作成していた辞書を、自動的に構築する手法について述べる。また、構築した辞書を用いて、入力文に対し顔文字を推薦する手法についても述べる。提案手法では、既存研究の人手でタグ付けをしていて語の数を増やすには手間がかかるという問題を、語の評価を自動化することにより解決した。

提案手法の流れを図 1 に示す。提案手法の処理は、使用するデータの用意、形態素データベースの構築、顔文字推薦処理の 3 つに大別される。3.1 項にて使用するデータの用意、3.2 項で形態素データベースの構築、3.3 項で顔文字推薦処理について説明する。なお、3.3 項の顔文字推薦処理は、江村らの手法と同様の手法である。

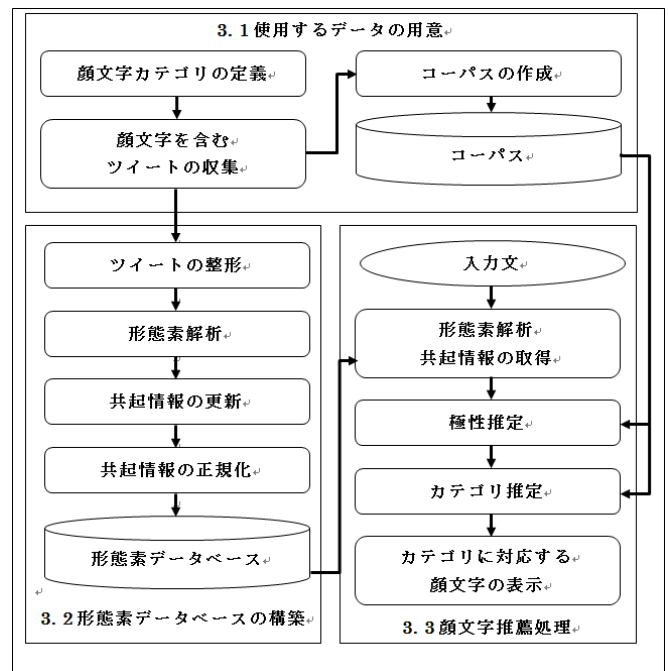


図 1 提案手法の流れ

3.1. 使用するデータの用意

使用するデータは、図 1 で示したように、顔文字カ

カテゴリの定義、顔文字を含むツイートの収集、コーパスの作成という、3つの処理から成っている。それぞれの処理について詳細を説明する。

顔文字カテゴリの定義

顔文字カテゴリの定義は、次の手順で行う。

- 1. 実際のツイートから顔文字を手で抽出する。
- 2. 抽出した顔文字を、人手により分類する。顔文字の分類は、顔文字の見た目をもとに人手で行う。
- 3. 全ての顔文字を分類した後、各分類に名前と極性をつけ、それを顔文字カテゴリとする。

今回は2012年7月24日の3,511件のツイートから、794個の顔文字を抽出した。その中には476種類の顔文字が存在した。これらの顔文字を分類し、顔文字カテゴリの定義と、各カテゴリへの分類基準の設定をした。また、顔文字と顔文字カテゴリをセットにし、顔文字辞書として保存した。収集した全ての顔文字は、「謝罪・お願い・感謝」「照れ」「泣き」「焦り」「興奮」「あいさつ」「驚き」「気合・怒り」「悲しみ」「ため息」「汗・困る」「安らぎ」「不快」「微笑」「喜び」「笑い」「無」「特殊」の18種類に分類することができた。このカテゴリ分類は、まず、出現回数の多い200個の顔文字を分類することにより、「謝罪・お願い・感謝」「照れ」「悲し泣き」「嬉し泣き」「焦り」「興奮」「あいさつ」「驚き」「気合・怒り」「悲しみ」「ため息」「汗」「困る」「安らぎ」「不快」「微笑」「喜び」「笑い」「無」「了解」「睡眠」「特殊」の計22カテゴリを作成する。その後、残りの276種類の顔文字を分類したところ、泣いている顔文字に、「悲し泣き」と「嬉し泣き」のどちらのカテゴリにも分類できない顔文字があったため、この2カテゴリを統合し、カテゴリ「汗」と「困る」についても同様に統合する。さらに、「了解」と「睡眠」カテゴリについては該当する顔文字が2つしか無かったため、この2カテゴリに属する顔文字は他のカテゴリに分類することにより、最終的に18カテゴリとした。各カテゴリの極性と、各カテゴリに分類された顔文字の数と例を表1に示す。各カテゴリへの顔文字の分類は、表2に示した基準で行った。この基準は、収集した顔文字を分類しながら作成した。各顔文字は、この基準を参考に、最も適切だと思われるカテゴリに著者が分類した。

表 1 各カテゴリに分類された顔文字の数と例

カテゴリ	極性	分類された顔文字の数	顔文字の例
謝罪・お願い・感謝	なし	6	m(__)m, <(_ _)>
照れ	POS	9	(//▽//), (//ω\)
泣き	NEG	50	(T_T), (;つД`)
焦り	NEG	8	><, (>_<)ノ
興奮	POS	13	(≥◇≤), (ε>▽<ε)
あいさつ	なし	25	(´▽`)/, (^_^)/

驚き	なし	15	(´O´), ** (Д_)/
気合・怒り	なし	15	(´・ω・´), (*`ω`)
悲しみ	NEG	23	(´・_・´), (´_`)
ため息	NEG	16	(´Д`), ㄣ (´Д`) ㄣ
汗・困る	NEG	32	(~::~), (@p@)
安らぎ	POS	50	(*´▽`), (*´///`*)
不快	NEG	11	(ε_ε), (ε-ε)
微笑	POS	5	(¯▽¯), (ε_ω_ε)
喜び	POS	61	\(^o^)/, (^q^)
笑い	POS	62	(*^^*), (●´艸`)
無	POS	42	(・ω・), (:3」∠)_
特殊	なし	33	((_ _))..zzzzZ, (๖´ڇ`๖)๖

表 2 各カテゴリの分類基準

カテゴリ	属する顔文字
謝罪・お願い・感謝	謝罪や感謝, お願いを表していると思われる顔文字。
照れ	照れていると思われる顔文字。具体的には、目や頬の部分が斜線で表現されている顔文字か、両目を手で隠している顔文字。
泣き	泣いていると思われる顔文字。
焦り	目の部分が「>」と「<」で表現されていて、苦しそうな顔文字。
興奮	目の部分が「≥」と「≤」や、「>」と「<」で表現されていて、楽しそうな顔文字。
あいさつ	あいさつをしていると思われる顔文字。例えば片手を上げている顔文字。
驚き	驚きを表現していると思われる顔文字。
気合・怒り	目や眉の部分が「`」と「´」で表現されている顔文字。
悲しみ	悲しみを表現していると思われる顔文字。例えば眉が下がっている顔文字。
ため息	目が「´」と「`」で表現されている顔文字のうち、ため息を吐いていると思われる顔文字。
汗・困る	困っていたり、汗をかいていたりする様子を表現していると思われる顔文字。
安らぎ	目が「´」と「`」で表現されている顔文字のうち、ため息カテゴリに属せず、口の部分が「ε」または「艸」、「v」でないもの。
不快	目がハイフン等、真ん中の横棒で表されている顔文字。
微笑	目がアップバー等、上側の横棒で表されている顔文字。
喜び	喜びを表している顔文字のうち、口が開いているもの。
笑い	喜びを表している顔文字のうち、口が開いていないもの。または、目が「´」と「`」で表現されている顔文字のうち、口の部分が「ε」、「艸」または「v」で表されているもの。
無	目の部分が点で表されている顔文字。
特殊	他のカテゴリに分類できない顔文字。

顔文字を含むツイートの収集

各カテゴリの顔文字を含むツイートは、Twitter のストリーミング API を用いて 2012 年 7 月 1 日から 2012 年 7 月 29 日に収集したデータより抽出した。なお、1 ツイートに複数カテゴリの顔文字が含まれている場合、含まれていた全てのカテゴリでカウントする。また、行頭から改行までで 1 ツイートとカウントしているため、ツイートの改行が含まれていて、改行前と後に顔文字が存在する場合は、1 ツイートが複数ツイートとカウントされる。

コーパスの作成

コーパスは、3,887 件の顔文字付きの文から、文と付与されていた顔文字のカテゴリをセットにして保存することにより作成した。顔文字付きの文は、2012 年 7 月 30 日にストリーミング API を用いて収集したツイートから人手で抽出した。

3.2. 形態素データベースの構築

本節では、形態素データベースの構築方法について詳細を述べる。形態素データベースの構築は、各ツイートに対し、整形、形態素解析、各形態素の共起情報の更新、共起情報の正規化の 4 手順から成る。

ツイートの整形

ツイートの整形では、取得したツイートから文と顔文字以外の情報の削除と、ツイートの分割を行う。ツイートから文と顔文字以外の情報の削除については、URL、誰宛のメッセージか示すための @ユーザ名、リツイートを示す RT、顔文字の部分以外のスペース、を削除する。例えば、「@dareka この動画めっちゃ面白いよ (≧▽≦) bit.ly/UPwews」というツイートを整形すると、「この動画めっちゃ面白いよ (≧▽≦)」となる。また、ツイートの分割については、1 文に対し、1 顔文字が文末に付与されている形になるよう分割する。例えば、「疲れた (´Д`) けど楽しかった \ (´○`) / また行こう！」というツイートの場合、「また行こう！」の部分は文末に顔文字が付与されていないため削除し、残りの部分を「疲れた (´Д`)」と「けど楽しかった \ (´○`) /」に分割する。

形態素解析

形態素解析は、lucene-gosen[5]を利用して行う。

共起情報の更新

各形態素について、表 3 のように各カテゴリの評価

値を持つ。これを共起情報と呼ぶことにする。各カテゴリの評価値の初期値は 0 であり、ツイートでの顔文字出現回数によって評価値を更新する。具体的には、顔文字を内包するツイートが得られた時、当該ツイートに含まれる各形態素に対して、当該顔文字の属するカテゴリの評価値を更新する。例えば、今、「喜び」を代表する顔文字があるツイートに含まれているとする。このツイートが形態素として「面白い」を内包していた場合、表 3 の「面白い」という形態素の「喜び」カテゴリについて更新（プラス 1）する。

表 3 共起情報のイメージ

形態素	喜び	汗	怒り	驚
面白い	4	3	2	1

形態素情報の正規化

収集した全ツイートに対して以上の処理をした後、式(1)により、各形態素の評価値を顔文字の出現回数で正規化する。式(1)において、 $value(m, c_i)$ は、形態素 $m \in M$ のカテゴリ c_i の評価値、 $num_kaomoji(m, c_i)$ は、カテゴリ c_i の顔文字が形態素 m と共にツイートに現れた回数である。ただし、 M は形態素の全体集合とする。表 3 を用いて説明すると各カラムで正規化することに等しい。

$$modified_value(m, c_i) = \frac{value(m, c_i)}{\sum_{m \in M} num_kaomoji(m, c_i)} \quad (1)$$

次に、式(2)により各形態素の補正後の評価値 $modified_value(m, c_i)$ の最大値を形態素毎に 1 とするために正規化を行う。正規化を終えた後、評価値を形態素データベースへと保存する。表 3 を用いて説明すると各行で正規化することに等しい。

$$normalized_value(m, c_i) = \frac{modified_value(m, c_i)}{\max_{c_i \in C} (modified_value(m, c_i))} \quad (2)$$

ただし全カテゴリの集合を C とする。

3.3. 顔文字推薦処理

顔文字推薦の手順は以下の通りである。

1. 入力文を形態素解析する。
2. 1. で得られた各形態素について、形態素データベースから共起情報を取得する。
3. 入力文の極性を推定する。
4. 3. で推定した極性の下位カテゴリとして、顔文字のカテゴリを推定する。
5. カテゴリに対応する顔文字を表示する。

入力文の形態素解析は形態素データベースを構築した際と同様の方法で行う。入力文の極性とカテゴリの推定については、以下で説明する。なお、入力文の極性とカテゴリの推定方法は、既存手法である江村らの手法と同様の方法である。カテゴリに対応する顔文字の表示は、顔文字辞書から、推定結果のカテゴリに

属す顔文字を取得し、使用回数の多い順に表示することにより行う。

入力文の極性の推定

入力文の極性は、以下の手順で推定する。

1. 入力文に含まれる各形態素の共起情報から、極性ごとの素性ベクトルを作成する。
2. コーパス中の各文との平方距離を計算する。
3. 平方距離の近い上位 k 件のうち、学習データにおいて最も多く付与されている極性を推定結果とする。

入力文のカテゴリの推定

入力文の推定された極性の下位カテゴリとして、カテゴリを以下の手順で推定する。

1. 極性の推定の場合と同様に、カテゴリごとの素性ベクトルを作成し、コーパス中の各文との平方距離を計算する。
2. 平方距離の近い上位 k 件のうち、学習データにおいて最も多く付与されているカテゴリを推定結果とする。

4. 実験

本節では、既存手法(Suzuki らの手法[2]と江村らの手法[3])と提案手法における、入力文に学習データ中の単語が存在する割合と、顔文字カテゴリの推薦精度を比較し、提案手法の有効性を示す。

4.1. テストデータの作成と既存手法の実装

本項では、テストデータの作成方法と既存手法の実装について述べる。

テストデータの作成

まず、2012年8月3日に、TwitterのストリーミングAPIを用いてツイートを収集した。次に、収集したツイートから、顔文字が付与されている文を手で200件抽出する。その後、各文に対し、付与されている顔文字のカテゴリを3.1項で述べた方法で判断する。最後に、各文と付与されていた顔文字、顔文字のカテゴリをセットで保存し、これをテストデータとする。

既存手法の実装

比較対象として実装した既存手法[2][3]について、本研究と同じ条件のもとで評価するために論文と異なる実装をした部分について説明する。

まず、Suzuki らの手法[2]の実装について説明する。Suzuki らの手法については、3つの変更を加えて実装した。1つ目は、顔文字カテゴリである。Suzuki らはPlutchikの考案した感情を拡張したものを顔文字カテゴリとして使用していたが、推薦精度を比較するためには顔文字カテゴリを揃える必要があるため、3.1

項で定義した顔文字カテゴリを使用した。2つ目は、文の感情部がどの感情を表しているかの判断方法である。Suzuki らはアンケートにより決めたが、今回は著者が決めた。最後の変更点は、各顔文字カテゴリの対となるカテゴリについてである。Suzuki らの手法では、“名詞、ナイ形容詞語幹”の直後に“助動詞、特殊・ナイ”があった場合、もともと判定していたカテゴリとは反対の意味を表すカテゴリを推定結果とする。しかし、今回のテストデータには“名詞、ナイ形容詞語幹”の直後に“助動詞、特殊・ナイ”が続くものが無かったため、各カテゴリの反対の意味を表すカテゴリについては考えていない。

次に、江村らの手法[3]の実装について説明する。江村らの手法は、カテゴリ、コーパスに変更を加え、実装した。まず、カテゴリの変更について説明する。江村らは、ツイートを37カテゴリに分類し、その分類をカテゴリとした。しかし、今回は、顔文字を分類することにより作成した、3.1項で説明したカテゴリを使用する。コーパスについては、江村らは、2011年5月と7月1日～16日のパブリックタイムライン上のツイートから、きざしラボの提供する顔文字と、アンケートを用いて決めた顔文字、全163種類を含むツイートを抽出し、コーパスを作成した。しかし、今回は、提案手法のコーパスと同じコーパスを利用した。

また、手がかり語の選択方法と、 k -NNの k の値については、江村らの論文に記載されていない。そのため、手がかり語はコーパス中の文から人手で選択した。 k -NNの k の値については、コーパスをテストデータとし、 k の値を1から100まで変化させ、式(3)の値が最大になる $k(k=28)$ を選択した。なお、 $precision@n$ は、推定結果の n 番目までに正解が含まれている確率である。

$$Score(k) = \sum_{n=1}^3 \frac{precision@n}{n} \quad (3)$$

4.2. 学習データ中の単語が存在する割合の測定

本項では、既存手法と提案手法の、学習データ中に存在する単語を含む入力文の割合を測定した。Suzuki らの手法に対しては、テストデータの文に文の感情部が含まれているかどうかによって測定した。江村らの手法に対しては、テストデータの文に手がかり語辞書に含まれている語が存在するかどうかによって測定した。提案手法に対しては、形態素データベースに含まれている形態素が含まれているかどうかによって測定した。

測定結果を表4に示す。この結果から、入力文に学習データ中の単語が存在する割合が既存手法よりも提案手法のほうが高いことがわかる。

表 4 各手法の対応している文章率

手法	入力文に学習データ中の単語が存在する割合
Suzuki らの手法	36.5%
江村らの手法	81.5%
提案手法	99.0%

4.3. 推薦精度の評価

本項では、推薦精度の評価実験について説明する。本実験では、学習データ中の単語が存在する文に対する、顔文字カテゴリの推薦精度を測定した。テストデータに含まれる文を入力文として入力し、推定結果が実際に使用されていた顔文字のカテゴリと一致するかどうかにより、推薦精度を測定した。提案手法 2 は、江村らの手法で使用している、手がかり語辞書中の語が入力文に含まれない場合、提案手法で自動構築した辞書を使用するように拡張した場合である。結果を表 5 に示す。P@N は、カテゴリ推定が可能だったデータのうち、カテゴリ推定結果の N 番目までに正解カテゴリが含まれていた割合である。また、Suzuki らの手法については 1 つのカテゴリしか推定できないため、P@2 と P@3 は計算していない。

表 5 推薦精度

手法	推薦精度			
	P@1	P@2	P@3	F 値
Suzuki らの手法	12.3%	—	—	0.184
江村らの手法	20.9%	41.1%	48.5%	0.333
提案手法	18.7%	36.9%	45.5%	0.315
提案手法 2	20.7%	39.4%	47.5%	0.342

江村らの手法と提案手法を比べると、提案手法の方が江村らの手法より推薦精度が低い。この結果より、人手で評価した単語を用いた方が、提案手法による自動的に評価した単語を用いるより、精度が高くなることがわかる。また、江村らの手法と提案手法 2 を比べると、こちらも提案手法 2 の方が江村らの手法より推薦精度が低い。これは、江村らの手法を、江村らの手法よりも推薦精度の低い、提案手法で拡張したためだと考えられる。しかし、提案手法 2 は、江村らの手法では推薦できない文に対し、提案手法で推薦を行うことにより、対応できる文章率が向上している。対応できる文章率が向上した事により、既存手法よりも多くの入力文に対し、顔文字を入力する手間を削減できると考えられる。また、対応できる文章率が向上した結果、提案手法 2 の F 値は、江村らの手法よりも 0.009

高い 0.342 となり、提案手法が既存手法よりも優れている事を示す事ができた。

5. まとめ

本稿では、自動的に単語のカテゴリを評価する、顔文字推薦のための、顔文字カテゴリ推定手法について述べた。従来の研究では、人手で単語とカテゴリを対応付けたデータをもとに、顔文字を付与または推薦しており、対応できる単語数に限りがあった。それに対し、提案手法では、自動的に単語のカテゴリを評価したデータをもとに顔文字カテゴリを推定するため、対応できる単語数が増え、その結果、より多くの文に対し顔文字を推薦できる。ツイッターにおいて実際に入力された文の内、学習データ中の単語が含まれている文の割合を測定する実験の結果、既存手法では 81.5% の文に学習データ中の単語が含まれていたが、提案手法では 99.0% の文に含まれていた。また、推薦精度を測定する実験では、既存手法の精度が 20.9%、既存手法を提案手法で拡張したものの推薦精度が 20.7% と、精度は低くなるが、F 値は 0.009 向上し、提案手法の有効性を示す事ができた。

今後の課題としては、何も入力されていない状態への対応と、推薦精度の向上が考えられる。現在の手法では、入力文から使用する顔文字を予測するので、何も入力されていない状態では顔文字を推薦することができない。何も入力されていない状態でも使用する顔文字を推定できるようになると、例えば悲しい感情を示す場合、悲しい顔文字を入力するだけで良い。実際に、顔文字しか入力されていないツイートも存在する。何も入力されていない状態への対応は、今まで入力したツイートや、相手から送られたツイートからある程度予測できると考えられる。推薦精度も、今まで入力したツイートや、他のユーザのツイートを利用することによって、向上できると考えられる。

参考文献

- [1] 鈴木晃二, "顔文字の歴史について", 明海日本語, 第 12 号, pp.91-96, 2007.
- [2] N. Suzuki and K. Tsuda, "Automatic Emoticon Generation Method for Web Community", IADIS Int'l Conf. Web Based Communities, pp. 331-334, 2006.
- [3] 江村優花, 関洋平, "テキストに現れる感情, コミュニケーション, 動作タイプの推定に基づく顔文字の推薦", IPSJ SIG Notes, 2012-DD-85(1), pp.1-7, 2012.
- [4] 福井康之, "感情の心理学", 川島書店, 1990.
- [5] "lucene-gosen", 入手先
<<http://code.google.com/p/lucene-gosen/>>