SNS 上の絵文字位置パターンの分析と Levenshtein 距離を用いたサンプル抽出

山崎由佳 [1]　西村綾夏 [2]

[1]京都大学大学院 [2]フリー
{yk.ymzaki+ct, ayaka.nishimura.jp}<@>gmail.com

概要

本研究は，SNS 上の投稿における絵文字の位置のパターンに着目する．SNS 上の投稿では，複数の絵文字が規則的に配されることを通じて，情報伝達を助ける機能が発揮されることがある．このような事象を探るにおいては，投稿中の絵文字単体のみでなく全体の位置を取り扱うことが有効である．

本研究は，特徴的なパターンを持つ投稿データを効率よく抽出するために，絵文字位置ラベルの列に対する Levenshtein 距離を用いたサンプル抽出（および分類）の方法を提案する．実際に，X（旧 Twitter）の投稿データを対象としてこれを実施し，得られた特徴的なパターンについて説明する．

1　研究背景

絵文字の位置については，自然言語処理の観点からも，言語学の観点からも触れられてきた．自然言語処理的な処理を行う研究の例に，絵文字自動挿入を目的として挿入位置を計算する [1] や，共起語を利用した絵文字の意味推定手法を提案する [2] がある．ただし [2] は「単語の後ろや文末に付くことで単語や文を装飾している絵文字は，装飾している対象の意味と同じ意味を持つので,」と述べ，位置と意味の関係を前提のように扱っている．

一方で言語学的な観点からは，絵文字の役割や用法について分析する上で，絵文字の出現位置についても考察や分類がなされてきた．ケータイメールの絵文字を題材とした [3] は，絵文字について「句読点の置かれる場所に付されている」とも述べ，前後のコンテキストとともに絵文字を取り上げて絵文字の役割分類を探索的に行っている．[4] は，選択体系機能言語学 (SFL) の枠組みを用い，(i) 経験的意味に関係する ideational な機能，(ii) 関係性の制定に関わる interpersonal な機能，(iii) 意味を首尾一貫したテキストに組織化するための資源を記述する textual な機能という 3 つのメタ機能のうち，(iii) にあたる “PERIODICITY system” の層で，絵文字位置（言語資源の前か後か間か）に対応した分類を提示している．[5] は文体パロディにおける絵文字の分析を行い，特定の絵文字が登場する位置を文体の特徴の一部として記述している．

ただし，コンテンツ中に存在する絵文字の位置がどのようなパターンをなしているのかという点，並びに，それと絵文字の役割・機能との関係については，踏み込む余地があるだろう．

実際に，絵文字の位置の反復性への着目を通じて語用論的な効果が見出せる例として，Instagram 上のある企業投稿を取り上げる（図 1）．

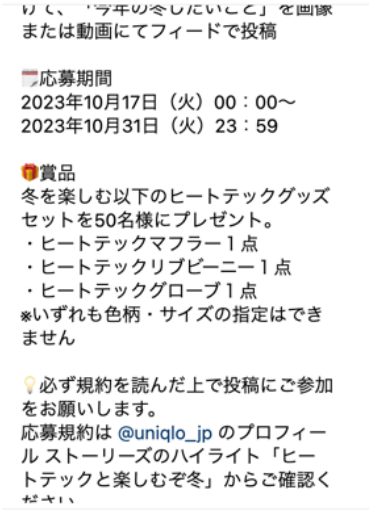

図 1　@uniqlo_jp. 2023. https://www.instagram.com/p/CyfoP3ju48O/, accessed on 2023/12/28.

分析のため，まず，[3] が提示した絵文字の役割の 7 分類を用いる．「①事物そのものを表す」「②事象の表現に前置き／後置き」「③身体動作」「④プロ

ソディー」「⑤前接文に合わせた感情を示す」「⑥前接文とは異なる感情を示す」「⑦装飾・雰囲気，リズムとり，句読点，トピック転換」である．図 1 における「🗓️応募期間」の「🗓️」は，「応募期間」という事象に対応しており，分類は②が最も適切だろう[1]．ただ，[3] は②の説明で「②も同様にアイコン的役割をもつが，その前後にすでに同じ意味を示す文字がある。にもかかわらず絵文字も使うのは，効率という観点からは無駄なのだが、メール全体の雰囲気を楽しくしたり和らげたりする効果を担っていると考えられる。」と述べている．しかし，図 1 全体を見れば，「🗓️」は数行後の「🎁」や「💡」とともに，テキスト中のまとまりの先頭に置かれていることがわかる．個々のまとまりは形式的に空行で区切られるとともに，話題としてのまとまりもなしている．このことから，「🗓️」「🎁」「💡」には投稿中の話題整理の役割も担うのではないか，という推測が導かれる．絵文字が視覚的な際立ちを生じさせ，閲覧者における情報の構造化を助ける効果を与えるのである．このような使用は「無駄」ではないはずだ[2]．留意すべきは，この効果は絵文字単体や絵文字を含む一行だけから生じるのではない，ということである．このため，コンテンツ全域（あるいは，閲覧者にある瞬間認識・記憶される範囲）における絵文字出現位置パターンを取り上げる意義があるといえよう．

そこで本研究は，SNS 上の投稿データを使用して絵文字の位置と機能を改めて分析するとともに，投稿データから特徴的な絵文字の出現位置パターンを抽出するための簡便な手法を提案する．

なお，先行研究における絵文字の位置の捉え方には，「テキスト文字列中の何番目か」「ある要素（たとえば絵文字の内容を表す語）との関係において前か後か」「一行あるいは一文において，先頭か末尾かそれ以外か」といったものが見受けられる．本研究は，ウェブ上，特に SNS 上のテキストを題材とし，「行」[3]における位置に着目する．

2　分析対象データの概要

ファッション小売チェーン「アダストリア」系列のブランドの，X（旧 Twitter）上の投稿を対象とする．アダストリアは 2021 年〜2022 年において，ファーストリテイリング，しまむらに次ぐ業界 3 位である．系列 20 ブランドの 2023 年 9 月 16 日における最新 200 投稿ずつを Twitter API を用いて取得し，その中で絵文字を含む 2093 件を対象とする．

3　分析手法：Levenshtein 距離を用いたサンプル抽出と分類

特徴的な絵文字位置登場パターンを抽出するために，Levenshtein 距離（編集距離）を用いて，サンプル抽出および分類を実行した．

大まかな流れは次の通りである．(i) 投稿中の各絵文字に対して位置ラベルを付与し，各ラベルをそれぞれ 1 文字として扱い，ラベルを連結し，得られた文字列に圧縮処置を施す．(ii)　Levenshtein 距離を用いた計算を行い，データ中で最も中間的なパターン 1 つ，それから最も「離れた」パターン 1 つ，……等を抽出していき，これらのサンプルへの距離を用いた分類を行う．

詳細を以下の小節に記す．

3.1　ラベルの付与と文字列化

対象データ中の各投稿に対して，以下の処理を実行し，文字列を得る．

1. 投稿に含まれる各絵文字に対し，1 文字ずつのラベルを付与する．用いるのは，行中での位置を表す 6 種類の位置ラベルである（「W:行全体」「A: 行頭」「Z: 行末」「B: 行頭絵文字から連なる絵文字の塊に入る」「Y: 行末絵文字に連なる絵文字の塊に入る」「M: それらのいずれでもない」）．なお，複数に該当する場合（W かつ A かつ Z など）は先頭のラベルを適用する．[4]
2. ラベルを連結して文字列を作成する．
3. 一部の位置ラベル（「B, Y, M」）が連続する際には，1 文字分に圧縮（短縮）する処理を行う．

例：(@repipi_armario の投稿 [5] より) “📣ウェブストア📣 [改行] 秋の新商品が色々入ってきてます🚚🚚💨ウェブの撮り方も秋に合わせてパープルにしてみたの☝🏻かわいいでしょ😆ウェブストアチェッ

1）　なお，⑦については，2 文の間にある例や行末の例が挙げられ，「⑦になると、絵文字と前接する文との関連性はほとんどなく，関連づけることさえ求められていない。」とある．
2）　もっとも，[3] はケータイメールを対象とした研究であるのに対し，本研究では企業の投稿を扱うため，テキストの目的が異なる場合も多いだろう．
3）　改行コードで区切られる範囲とする．
4）　なお，試行過程においては，4 種類の位置ラベル（「w：行全体」「f：行頭」「l：行末」「m：それらのいずれでもない」）や，9 種類の絵文字カテゴリラベル（ウェブサイト Emoji Terra https://emojiterra.com/ のカテゴリ分けをもととして絵文字に割り当てた意味的なカテゴリに対応）も試した．
5）　https://twitter.com/repipi_armario/status/903215107974488064, accessed on 2023/12/11.

クしてね💕➡️https://t.co/J8DE56Lk4e https://t.co/epq85pqsz3” には，“AZMMMMMMMM” という文字列を経由して，“AZM” が付与される．

3.2 サンプル抽出と分類

先述の処理で得た文字列をもとに，サンプル抽出と分類を行う[6]．

3.2.1 Levenshtein 距離

2 つの文字列間の「離れ具合」を計算する方法の一つに，Levenshtein 距離がある．文字列に対し，1 文字ごとの「挿入」「置換」「削除」といった 3 種の編集操作を許容する．ある文字列を別の文字列に一致させるための最小の操作回数が，両者間の Levenshtein 距離である[7]．たとえば，「A」と「B」の Levenshtein 距離は 1，「AB」と「BA」の Levenshtein 距離は 2 となる．

今回この方法を選定する理由として，(i) 文字の順番が結果に影響すること，(ii) 計算過程がクリアで再現性があること，を挙げる．

3.2.2 サンプル抽出

まず，サンプル抽出の手法を大まかに述べる．Levenshtein 距離に関して全体の中央値の 1 つ[8]にあたるラベル文字列を初めの（0 番目の）サンプルとする．以降の探索範囲を初めのサンプルからある程度離れた文字列集合とし，初めのサンプルから最も離れた文字列の 1 つを次のサンプルとする．回を重ねるごとに少しずつ探索範囲を狭めていくような形で，探索範囲からのサンプル取得および次の探索範囲の決定を行う．

より具体的には以下の通りである（例外処理を省いて述べる）．

パラメータ $R \in \mathbb{N}$（最初の探索範囲に確保したい個数）と $W \in \mathbb{N}$（最大距離からの許容幅）を与えておく．対象データ[9]中の各投稿を，添字を用いて d_i と記す．I で添字の集合を表す．各投稿に付与したラベル文字列を x_i で表し，L をこれらをまとめた配列 $(x_i)_{i\in I}$ とする．$I_0 := I$ と置く．Levenshtein 距離に関する L の median を 1 つ求め[10]，s_0 で表す．$x_i = s_0$ となる添字 $i \in I_0$ のうち最小のものを j_0 とする．

次のサンプル s_1 の探索範囲を以下のように作る．

$$\begin{aligned} u_0 &:= \max\left\{u \in \mathbb{Z}_{\geq 0} \middle| n(\{i \in I_0 | d(x_i, s_0) \geq u\}) \geq R\right\} \\ I_1 &:= \{i \in I_0 | d(x_i, s_0) \geq u_0\},\ L_1 := (x_i)_{i\in I_1} \end{aligned} \tag{1}$$

ただし，$d(\cdot,\cdot)$ で Levenshtein 距離を表し，$n($ 集合 $)$ で集合の要素の個数を表す．

ここから (2) のようにしてサンプル s_1 を抽出し，また次の探索範囲（I_2 および L_2）を定める．

$$\begin{aligned} d_{\max_1} &:= \max_{x\in L_1} d(x, s_0) \\ j_1 := \min\{i \in I_1 | d(x_i, s_0) = d_{\max_1}\},\ s_1 &:= x_{j_1} \\ u_1 &:= \max\{d_{\max_1} - W, 1\} \\ I_2 := \{i \in I_1 | d(x_i, s_1) \geq u_1\},\ L_2 &:= (x_i)_{i\in I_2} \end{aligned} \tag{2}$$

$d_{\max_2}$ 以降は，I_{k+1} が空集合になるまで (2) 同様のプロセスを繰り返す[11]．

3.2.3 分類

得たサンプル列をもとに，各データとサンプルの距離を用いた分類を実行する．ただし，パラメータとして，作成するクラスタ個数の上限 $(\in \mathbb{N})$ を与え，N と記す．分類番号は以下のように付与する．

- 文字列 x に対しては，サンプルらの先頭 N 個までの中で x との距離が最小になるようなサンプルのうち，最も添字が小さいものの添字を与える．（サンプルの添字の集合 J を $J := \{j_k\}$ で定め，J' を $J := \{j_k | k \leq N\}$ で定めると，x に対して分類を行う関数 $\mathrm{cl}(x)$ は 次のように表せる）

$$\begin{aligned} \mu &: x \mapsto \min_{j_k \in J'} d(x, j_k) \\ \mathrm{cl} &: x \mapsto \min\{j_k \in J' | \mu(x) = d(x, j_k)\} \end{aligned} \tag{3}$$

- 添字 i や投稿 d_i に対しては，x_i を通して同様に分類番号が付与できる．これを実行する．

4 分析結果と考察

実際に抽出されたサンプルを通して，見出された特徴的なパターンについて述べる．

6） 今回の手法は，各投稿が大体 10 行以下程度であるような状況を踏まえて作ったものである．もし対象の行数や含まれる絵文字の個数の最大値が非常に大きいような状況を扱うのならば，適宜変更が求められるだろう．
7） 拡張として，3 種の操作に対して異なる距離の加え方を与えるような距離を考えることもできる．
8） 一般には中央値は一意とは限らない．
9） 投稿数及び各投稿中の文字列長は有限であると仮定する．
10） 今回は，実装において，Levenshtein 距離の計算と median の算出には，python の Levenshtein ライブラリを用いた．
11） なお，$I_k \supsetneq \emptyset$ の場合，$j_k \in I_k$ となることと，$u_k \geq 1$ となることから，$I_k \supsetneq I_{k+1}$ となるといえる．

はじめに，パラメータとして $R = 20, W = 10$ を指定して得られたサンプル列[12]と，さらにクラスタ個数上限 $N = 10$ を指定して得られた分類について述べる．まず，得られたサンプル列の長さは 39 で，クラスタのサイズは，サンプル s_0（ラベル列「Z」）に紐づく分類番号 0 から順に，1838, 4, 17, 5, 46, 1, 100, 27, 1, 54（合計 2093）であった．各クラスタに入るサンプル個数は順に，1, 3, 10, 1, 6, 1, 7, 7, 1, 2 であった．以下，特徴的なパターンを数点，抽出されたサンプルに紐づく投稿例の説明とともに述べる．

パターン例 (A)：「AZ」の反復・頻出

(A-1)：一行の先頭と末尾に同じ絵文字

(1) 🎵UNI9UE PARK’23🎵[改行][改行]📣10 月 8 日出演📣[改行] 🕯️アカリトライブ🕯️[改行] \山本彩さんからコメント💌✨ / [略][13]

例 (1)（画面上表示例を付録に収録）は，サンプル s_1 に紐づき分類番号は 1．圧縮後の位置ラベル列は「AZAZAZMZAM」であり，3 個の「AZ」にあたる部分には，テキストを同じ絵文字で挟み目立たせるような効果が見出される．一方で個々の絵文字では，告知であることと「📣」，言及内容がライブであることと「🎵」，「アカリ」と「🕯️」の間に関連があるといえよう．形式（構造）と個々の絵文字の内容という二点で情報伝達に寄与するものだろう．

(A-2)：ブロック先頭に絵文字→中の行末に絵文字

(2) RT @up_6ch: こんにちは☀️[改行] きょうは…[改行][改行] 🐈#モリコロパーク に誕生✨[改行]　#猫の城遊具 を中継リポート📹 [改行]　#猫の恩返し　#ジブリ [改行][改行] 👑#藤井七冠 [改行]　王座への道は開くか🔥[改行]　八冠への挑戦権は…[改行][改行] ⚾#大谷翔平 が４０号！大台到達👏 [改行][改行] 💕#近藤千尋（@chipi1215…[14)]

例 (2)（s_{17}，分類番号 1）は，「RT」であり，元の投稿では “（@chipi1215” には “）さん大注目👀[改行]　暑い夏の🆕必需品☝️[改行]　#ペットボトルホルダー” と続いた．この例では，ニュースを述べるブロックが，空行を挟んで繰り返されている．各先頭行の頭にある絵文字は，図 1 同様，情報整理を助ける textual な働きを果たすものだろう．その 4 つの絵文字「🐈」「👑」「⚾」「💕」のうち，はじめ 3 つは，テキストの内容（「猫」「王座」）やトピック（野球）と結びつく ideational な機能をも兼ねると考えられるが，最後の「💕」の選定は異なる性質があると思われる．また，一部の行末には，ニュースへの評価や態度（語り手がどう受け止めているか，どう評価するかなど）を示し interpersonal な機能を持つと考えられる絵文字「✨」「🔥」「👏」がある．

なお，各ブロックの二行目以降は字下げされており，これらの行頭に絵文字はない．こういったことも情報整理を支えると考えられる．

パターン例 (B)：「AA」の反復

(3) 🌟人気ランキング🌟 [改行] 🥇ダブルルーズジャケット [改行] ▶️https://t.co/YZYCz190UI[略：「AA」パターン][15)]

例 (3)（s_3，分類番号 3）の省略部分では，前 2 行と同じような構造「（メダルを表す絵文字）アイテム名 [改行] ▶️(URL)」が空行を挟み 3 回登場していた（「🥇」の部分には「🥈」「🥉」「🏅」の順に登場）．

上述の例 (2) と例 (3) では，位置形式の反復自体が，投稿の構造化という点で textual な機能を果たしていると考えられる．ただし，例 (3) では絵文字の選択幅が狭く（メダルを表すものと「▶️」），メダル絵文字中の数字がブロックの順番と関わる一方，例 (2) 中では行頭にあたる絵文字は多様であった（なお「💕」「👑」の行頭用例は 25% 未満）．今後，一般に，反復されるもののあり方に着目し，反復的構成の多層性を分析することが可能だろう．

5 おわりに

本研究は，絵文字単体ではなく位置のパターンに着目することにより，情報整理に役立つ絵文字の使い方が見出せることを提示した．また，位置のパターンに基づく投稿サンプル抽出方法を考え，実際に興味深い例を抽出した．

本手法の発展としては，位置ラベル文字列の処理方法を，言語学的分析の実感を踏まえて調整することが考えられる．また，今回は複数ユーザの投稿で構成されるデータを対象とした分析を提示したが，特定のユーザを対象として，投稿のバリエーションを探ることもできるだろう．

12） 実際のところ，$d_{\max_k}$ の最大が 10 であったため，W をこれより大きくしても同じ結果となる．

13） @nikoand. https://twitter.com/nikoand/status/1701060180732260362, accessed on 2023/12/27.

14） @LAKOLE_official. https://twitter.com/LAKOLE_official/status/1687365843028721664, accessed on 2023/12/26.

15） @lowrystwit. https://twitter.com/LOWRYSTWIT/status/1439932330190131205, accessed on 2023/12/26.

謝辞

This work was supported by JST SPRING, Grant Number JPMJSP2110.

参考文献

[1] 橋本泰一. Twitter への絵文字自動挿入システム. 言語処理学会第 17 回年次大会発表論文集, pp. 1151–1154, 2011.

[2] 高本健太, 町田翔, 延澤志保. 用法に着目した文中の絵文字の意味推定. 情報処理学会第 80 回全国大会講演論文集, Vol. 2018, No. 1, pp. 313–314, 03 2018.

[3] 三宅和子. ケータイの絵文字 : ヴィジュアル志向と対人配慮. 日本語学, Vol. 31, No. 2, pp. 14–24, 02 2012.

[4] Lorenzo Logi and Michele Zappavigna. A social semiotic perspective on emoji: How emoji and language interact to make meaning in digital messages. **New Media & Society**, Vol. 25, pp. 3222 – 3246, 2021.

[5] 西村綾夏. 文体における視覚的文体素の役割 : Twitter ユーザを対象とした文体パロディ中の絵文字を例に. 言語科学論集, Vol. 23, pp. 19–38, 12 2017.

[6] Michael Halliday, et al. **Language as social semiotic**. Edward Arnold London, 1978.

[7] 小林一郎. 意味へのアプローチ：ハリデー言語学の観点から. 認知科学, Vol. 24, No. 1, pp. 8–15, 2017.

[8] Michele Zappavigna and Lorenzo Logi. Emoji in social media discourse about working from home. **Discourse, Context & Media**, Vol. 44, p. 100543, 2021.

付録：結果補足資料

サンプル一覧

$R = 20, W = 10, N = 10$ を指定して得られたサンプル列と分類番号を提示する．

表 1　サンプル列（$R = 20, W = 10$）と分類番号（$N = 10$）

サンプル	圧縮後位置ラベル列	分類番号
s_0	Z	0
s_1	AZAZAZMZAM	1
s_2	MYZZZYZYZ	2
s_3	AZAAAAAAAA	3
s_4	AMZZYZZZ	4
s_5	AZAZMZMZAM	5
s_6	ZMYZAZZ	6
s_7	AZAZMZAM	7
s_8	AZYZAZZYZ	8
s_9	ZAMYZAM	9
s_{10}	YZZZAZZ	2
s_{11}	AAYZAAAM	7
s_{12}	MZAZZYZ	2
s_{13}	YZZMAZZ	2
s_{14}	MZYZYZYZ	2
s_{15}	AZAAAMA	1
s_{16}	MZMZYZZ	2
s_{17}	ZAZZAZAZA	1
s_{18}	MZYZMYZ	2
s_{19}	AZAAAZA	7
s_{20}	ABYZZMZ	4
s_{21}	MZMZYZZA	4
s_{22}	ZYZZZAZZ	2
s_{23}	AZAZMZA	7
s_{24}	MZYZAAZ	6
s_{25}	ABYZZZA	4
s_{26}	YZMZAZZ	6
s_{27}	ZYZYZAB	9
s_{28}	AZYZAZZ	6
s_{29}	YZYZZYZA	2
s_{30}	AZMZZAB	7
s_{31}	YZYZAZZ	6
s_{32}	AZMZZAM	7
s_{33}	ZZYZAZZ	6
s_{34}	AYZZYZZ	4
s_{35}	MZYZAZZ	6
s_{36}	AZZZZZA	4
s_{37}	YZYZYZA	2
s_{38}	AZYZMZM	7

投稿例のキャプチャ

図 3　例 (1)

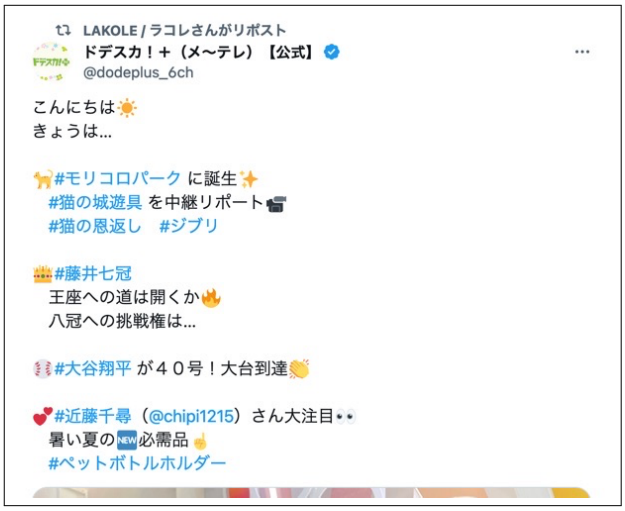

図 4　例 (2)

図 5　例 (3)