E-013

Twitter での画像情報を利用した日本語入力手法の提案
Proposal of Japanese Input Method using Image Information on Twitter

菅原 太一† 松原 雅文† Goutam Chakraborty† 馬淵 浩司†
Taichi Sugawara Masafumi Matsuhara Goutam Chakraborty Hiroshi Mabuchi

1. はじめに

現在の日本において，携帯端末は日々の生活に欠かすことのできない存在となっている．携帯電話・PHS の普及率は 116.8% （平成 25 年度末）[1] となっており，全国民が 1 台以上の携帯端末を所持・利用していることになる．更に，最近では iPhone や iPad などを筆頭としたスマートフォン・タブレットの普及，市場規模の拡大がめざましい．これらスマートフォン等においては，従来の携帯電話と比べ，多様なアプリケーションや Web サービスが容易に利用できる環境となっており，Facebook や Twitter，LINE などに代表されるソーシャルネットワーキングサービス（以下，SNS と表記）の普及の一翼を担っている．

SNS の普及により，「音声」から「文字・画像」へ，「書き溜めて発信」から「即時発信」へ，コミュニーションスタイルが変化してきている．現在，コミュニケーション系メディアの平均利用時間 [2] は，平日 1 日当たりの音声通話利用（携帯・固定・ネット）が 8.5 分なのに対し，文字利用（ソーシャルメディア・メール）が 41.5 分となっており，文字は音声通話の約 5 倍の時間利用されている．今や，文字でのコミュニケーションは人の生活に必要不可欠である．

このように，携帯端末上で文字列を入力する機会と必要性は増大している．携帯端末上で高速に精度よく入力を行うためには，入力方式などを工夫する必要があり，多くの研究がなされている[1]-[3]．

最近では，Twitter における「写真つきツイート」など，複数の情報を同時に発信することが可能となってきている．例えば，スイーツを食べるとき，スイーツに関する文字と画像を同時に発信するということなどである．このとき，発信者は文字と画像に何らかの関係性を持たせ発信しているものと考えられる．

そこで，本研究では，この画像情報を利用した日本語入力手法を提案する．複数の文字と画像の組において，文字どうしの類似度が高ければ，画像どうしの類似度も高くなるだろうという考えに基づいている．これにより，日本語入力の精度向上を目指す．

本稿では，本手法の概要を示し，実際のデータを用いて実験した結果から，本手法の有効性を述べる．

2. 提案手法

本手法では，ユーザが入力した文字列のかな漢字変換の際，変換候補の重み付けに画像情報を用いる．本手法における処理の流れを図 1 に示す．

はじめに，ユーザがかな文字・画像を入力する．Twitter における「写真つきツイート」を行うイメージである．その後，通常のかな漢字変換と同様，かな文字の辞書照合を行い，候補を漢字へと変換する．

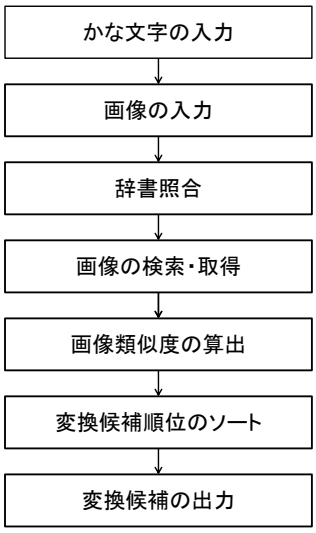

図 1：提案手法の流れ

次に，変換候補に出現した文字列をキーワードとし，Twitter 上で画像検索を行い，画像を取得する．そして，入力した画像と，取得した画像の類似度を算出する．算出された画像類似度について，変換候補それぞれに対し，その候補が正解だった場合について評価を行う．そして，その評価を加味して変換候補をソートする．このソート順に，変換候補を出力する．

最終的に，表示された変換候補一覧の中から，ユーザが正解である単語を選択することで，処理が完了となる．これにより，効率の良い日本語入力が期待できる．

3. 評価実験

本手法の有効性を確認するため，同音異表記語と Twitter の画像情報を用いた評価実験を行った．

3.1 概要

はじめに，同音異表記語と画像がセットになったデータを検索・取得し，それらの画像の類似度を算出する．その後，同音異表記語それぞれに対し，その語が正解だった場合の平均適合率を算出することで，同音異表記語と画像情報の関連の強さを評価する．

正解だった場合の同音異表記語の平均適合率が高ければ高いほど，ユーザが入力を意図していた可能性が高く，重み付けに有効であると判断できる．これを評価することにより，本手法の有効性を示すことができると考えられる．

† 岩手県立大学 Iwate Prefectural University

1 総務省, “情報通信統計データベース 携帯電話・PHS の加入契約数の推移”, http://www.soumu.go.jp/johotsusintokei/field/data/gt01020101.xls

2 総務省情報通信政策研究所, “平成 25 年 情報通信メディアの利用時間と情報行動に関する調査”, http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku-since2011.html

3.2 同音異表記語・画像の取得

はじめに，同音異表記語と画像がセットになったデータを検索・取得する．今回は，Twitter より，「写真つきツイート」のデータを以下の条件に基づき取得した．

- キーワード：「かんしん」（感心・関心）
- 取得方法：「キーワード＋filter:images」
- 利用ツール
  - Twitter API v1.1, python-twitter
- 最新上位 11 件のデータをそれぞれ収集
  - 内 1 件をテストデータとして利用
  - 2014/5/21 23:50 頃取得
- 重複ツイート（リツイートなど）は対象外
- 画像が複数存在した場合は，最初の 1 枚のみ取得

3.3 画像の類似度算出

取得した画像から，画像どうしの類似度を算出する．今回は，先述したテストデータをキーにし，それぞれのデータに対し，10(件)×2(単語) = 20(件) の画像の類似度を算出した．

今回，画像の類似度を算出するため，Simg [3] というソフトを利用している．Simg は，複数の画像を比較し，類似画像を検索できるソフトであり，2 つの画像を入力すると，類似度を 0 から 1 の間の値で返す．

このソフトで利用できるアルゴリズムの中で，Average Hash アルゴリズムを採択した．これは，画像を所定のサイズにリサイズしたあと，グレー化・2 値化を行い，bit 列のハッシュ値とするアルゴリズムである．生成されたハッシュ値は，bit ごとに比較可能となる．画像の特徴を端的に比較できるアルゴリズムといえる．

3.4 結果の評価

それぞれのキーワードに対し，画像の類似度が高い順に結果をソートし，キーワードが正解かどうかを順にチェックする．その後，適合率，平均適合率といった尺度[4]を用いて，実験結果に対する評価を行う．

- 適合率（Precision）

  チェックされたデータの中のうち，正解のデータの割合を示す．ノイズの少なさを示す尺度である．以下の計算式で求められる．

$$\text{適合率} = \frac{\text{チェックされたデータ中の正解データ数}}{\text{チェックされたデータ数}}$$

- 平均適合率（Average Precision）

  正解データがチェックされた時点での適合率の平均値である．この値が高ければ高いほど，ユーザが意図していた可能性が高い同音異表記語であると判断できる．

3.5 実験結果および考察

実験結果を表 1 に示す．今回は，正解データがすべてチェックされた時点での平均適合率を利用している．また，キーワードを 2 つ用いているため，本手法を用いない場合（デフォルト）の平均適合率は，0.50 となる．キーワード 2 つのどちらの場合でも，デフォルトより高い平均適合率が得られた．特に，キーワード「感心」においては，デフォルト比 1.36 倍の平均適合率が得られた．よって，本手法は有効であると考えられる．

表 1：実験結果

	キーワード	平均適合率
提案手法	感心	0.68
	関心	0.55
デフォルト		0.50

しかし，本実験で用いたデータ数は合計で 22 件，キーワード数は 2 件であり，データ数・キーワード数が少ない．そのため，結果が不安定である可能性が考えられる．よって，データ数・キーワード数を増加させ，実験を行っていく必要がある．

今回，画像の類似度を算出するため，Average Hash アルゴリズムを採択した．ピクセル同士が似通っていればいるほど高い類似度を表すものであり，画像の特徴を端的に比較できた．「関心」における類似度算出の際，「机の上に本が置いてある画像」という，人間の目で見ても似通っている画像のペアが，最も類似度が高く出力された．

しかし，今回採択したアルゴリズム以外でも画像の類似度を評価することは可能である．よって，画像類似度を算出するソフト・アルゴリズムに関して調査を進め，複数のソフト・アルゴリズムを用いて実験し，結果の評価を行っていく．

4. おわりに

本稿では，日本語入力の精度向上に向け，画像情報を利用した日本語入力手法を提案し，実験によってその性能評価を行った．本手法では，画像の類似度を考慮し，変換候補順位を変化させ，変換候補を出力する．

同音異表記語と Twitter の画像情報の関連の強さを評価する実験を行った．デフォルト比最大 1.36 倍の平均適合率が得られ，本手法の有効性が示された．

今後は，より多くのデータ，多くのキーワードにおいて実験を行う．また，画像類似度を算出するソフト・アルゴリズムに関して調査を進め，本手法の更なる精度向上を目指す．

参考文献

[1] 松原 雅文, 荒木 健治, 桃内 佳雄, 栃内 香次, “文字情報縮退方式を用いた帰納的学習によるべた書き文の数字漢字変換手法の有効性について”, 電子情報通信学会論文誌 D-II, J83-D-II, No.2, pp. 690-702 (2000).

[2] 田中 久美子, 犬塚 祐介, 武市 正人, “少数キーを用いた日本語入力”, 情報処理学会論文誌, Vol.44, No.2, pp.433-442 (2003).

[3] 菊地 直樹, 松原 雅文, Goutam Chakraborty, 馬淵 浩司, “携帯電話における入力誤り自動訂正手法の日常的な文章に対する有効性について”, FIT2011 第 10 回情報科学技術フォーラム講演論文集, E-056, pp.349-350 (2011).

[4] 北 研二, 津田 和彦, 獅々堀 正幹, “情報検索アルゴリズム”, 共立出版 (2002).

[3] namu, “Simg ユーザガイド”, http://www.vector.co.jp/soft/winnt/art/se492890.html