E-036

OCR を利用した崩れた表記の自動修正手法の性能評価

Performance Evaluation of Automatic Correction Method for Informal Words using OCR

秋山大五郎 † 松原雅文 ‡
Daigoro Akiyama Masafumi Matsuhara

1. はじめに

近年，SNS は多くのユーザが積極的に情報を発信する場となっている．多種多様なユーザが発信するデータは膨大であり，それらのデータを利用する動きも活発になっている．しかし，SNS 上のショートメッセージには崩れた表記の単語が含まれるため，データを活用しにくい課題がある．そこで我々は，OCR を利用した崩れた表記の自動修正手法を提案している [1]．しかし，提案手法では，評価に使用したデータ数が少ないといった課題があった．

そこで，本稿では，崩れた表記に変換する辞書を用いて自作したデータにより性能評価を行った結果を報告する．また，提案手法の画像生成処理などを変更し提案手法の精度向上を目指す．

2. 関連研究

崩れた表記の単語を正しい単語に修正する手法として，崩れた表記が少ない文書から修正候補を取得する手法 [2] がある．文章中の崩れた表記の単語において左右の単語とマッチする文章を崩れた表現の少ない文書から検索し，修正候補を取得する．取得した修正候補に対して修正候補の出現頻度，修正前後の文字列間における編集距離，修正前の形態素解析コスト値に基づいたスコアリングを行い，総合スコアが一定以上のものを修正ルールとする手法である．

この手法では，本来修正する必要がない未知語に対しても修正処理を行ってしまう問題点がある．これに対し，フィルター処理を行うことで修正が不必要な未知語を修正対象から除外する手法 [3] がある．未知語として判断された単語に対し，TF-IDF(Term Frequency-Inverse Document Frequency) 処理を行い，TF-IDF 値をもとに検出された未知語が一般的かの判断を行う．一般的な未知語を処理対象から除外することによって崩れた表記の単語のみを処理する手法である．

これらの手法は文脈情報から修正候補を取得しているため，文章全体が崩れた表記の場合，修正候補を正しく取得することが難しい．そのため，我々は崩れた表記が含まれる文章に対し OCR を用いることで，文字の形から修正候補を取得可能とする手法を提案した．しかし，提案した手法では，評価に使用したデータ数が少ない課題があった．

そこで，本稿では，崩れた表記に変換する辞書を用いて自作した崩れた表記のデータにより性能評価を行った結果を報告する．また，提案手法の画像生成処理などを変更し提案手法の精度向上を目指す．

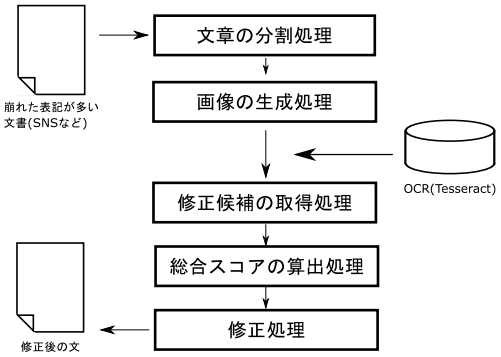

図 1: 提案手法の OCR 部分

3. 提案手法

3.1. 提案手法の概要

提案手法における OCR を用いた修正候補取得部分の処理を図 1 に示す．提案手法では，崩れた表記を多く含む SNS などの文章を入力とし，文章を 2 文字区切りで分割する．次に分割した文字列が描かれている画像を生成する．そして生成した画像から OCR を用いて修正候補を取得する．最後に取得した修正候補をスコアリングし，最適な修正候補を取得する．

3.2. 文章の分割処理

崩れた表記が多く含まれた文章を 2 文字区切りで分割する．これは，「ネ申」(神) や「弓長」(張) などの崩れた表記の多くは 2 文字以下で構成されているためである．

我々が過去に提案した手法では，形態素解析で未知語と判定された文字列に対し修正を行っていたが，崩れた表記の組み合わせによって未知語と判定されないものや，組み合わせて 1 つの文字を表している崩れた表記の構成要素が別々の未知語として判定されるという問題点があったため，一定の文字数で文章を区切り分割することとした．

3.3. 画像の生成処理

分割した文字列が描かれている画像を生成する．崩れた表記は 2 文字で正規表記 1 文字を表していることが多いため，崩れた表記の 2 文字が描かれている画像を生成し，画像をリサイズすることで，崩れた表記の文字列に対応する正規表記文字を OCR が認識できる可能性を向上させる．

3.4. 修正候補文字列の検索

生成された画像から OCR を用いて修正候補文字列を取得する．今回は OCR エンジンに Tesseract [1] を用いた．

†岩手県立大学大学院ソフトウェア情報学研究科
‡岩手県立大学ソフトウェア情報学部

[1] https://tesseract-ocr.github.io/

表 1: 画像生成設定

設定名称	値	設定名称	値
リサイズ比率 (縦)	1.9	リサイズ比率 (横)	0.9
フォントサイズ	20	画像サイズ (縦)	60
画像サイズ (横)	80	書き始め座標 (縦)	-3
書き始め座標 (横)	7		

表 2: 実験結果

文字種	総数	合致数	正解率 (%)
ひらがな	35	9	25.7
カタカナ	33	1	3.0
漢字 1	1,075	334	31.1
漢字 2	148	15	10.1
合計	1,291	359	27.8

3.5. 総合スコアの算出

スコアリングには，修正後文章の形態素解析器の連接コスト，修正前後の文字列間における編集距離を用い，これらをもとに総合スコアを算出する．その後，修正候補文字列の総合スコアが最も高いものを修正文字列として出力する．

4. 実験

4.1. 実験条件

提案手法における OCR を用いた修正候補文字列の出力について評価実験を行う．

崩れた表記辞書は，正規表記 1 文字とそれに対応する 1 ～ 3 文字の崩れた表記で構成されている．この崩れた表記辞書の中で，正規表記 1 文字に対して 2 文字以上の崩れた表記の文字列が登録されている約 1,300 個の組み合わせを実験データとして使用した．

データに対し，最適な画像生成設定を探索するため Beam Search を用いた．各組み合せのうち上位 3 つの組み合わせを保持するよう設定した．各組み合わせの評価には，正解率を使用した．

予備実験により決定した画像生成設定を表 1 に示す．この設定で実験を行った．

4.2. 実験結果

崩れた表記から生成した画像に対して OCR を用いて修正候補文字列を出力した結果を表 2 に示す．文字種は崩れた表記の文字列に対応する正規表記の文字の種類を示している．「漢字 1」は対応する正規表記の漢字が「へん」と「つくり」に分解できるものを表し，「漢字 2」は「へん」と「つくり」に分解できない漢字と 3 文字の崩れた表記になるものを表している．

4.3. 考察

実験結果から，「へん」と「つくり」に分けることができる漢字 1 の正解率が最も高くなった．これはデータにおいて，漢字の占める割合が多く，画像生成の設定が漢字に最適となっていたことによるものであると考えられる．

ひらがなはデータにおいて占める割合が少ないが，漢字 1 の次に正解率が高かった．これは「ﾚけ」(け) や「ｶゝ」(か) などのひらがなの崩れた表記の一部が，正規表記文字の形容に似ていたことが要因だと考えられる．しかし，認識できなかったひらがなの崩れた表記 (む：£ゝ，ぞ：ξ″ など) は目視での判断も難しいため OCR と他の手法を組み合わせるなどの対策が必要であると考えられる．

「へん」と「つくり」に分けることができない漢字に関しては，構成されている要素に「にょう」が含まれている場合に，認識が難しいことが分かった．これは，「にょう」は 1 文字の領域のうち，左側から右下までを占めているのに対して，「にょう」の崩れた表記では 2 文字に分割されるため，右下の領域が使用できないことが原因だと考えられる．しかし，一部の構成要素に「にょう」を持つ漢字 (超：走召，題：是頁など) は正しく認識できていたため，「しんにょう」のような対応する漢字が無い「にょう」への対策が必要である．

なお，生成画像のリサイズを行わずに，正規の表記である漢字が描かれている画像を生成し，OCR で認識した場合，精度は約 64%だった．そのため，一部の漢字に関しては OCR が対応していない可能性があり，OCR を再学習させるなどの対策が必要であると考えられる．

5. おわりに

本稿では，OCR を用いて崩れた表記の単語の形容から正規の表記である単語を認識する手法の性能評価を行った．今回の実験において，OCR を用いて 2 文字で構成された崩れた表記文字列を正規表記文字としてある程度認識可能であることが確認された．

今後は，OCR のみでは認識することが難しい崩れた表記に対応するため，OCR と他の手法を組み合わせる方法を検討する．また，今回の実験では，修正候補文字列に対しスコアリングを行っていないため，修正候補文字列を複数個取得し，スコアリングを行い，その性能を評価する実験を行う予定である．

謝辞

本研究の一部は JSPS 科研費 21K12611 の助成を受けたものである．

参考文献

[1] 秋山 大五郎, 松原 雅文: OCR を利用した崩れた表記の自動修正手法の提案, 情報処理学会第 84 回全国大会,5V-07, 愛媛大学城北キャンパス, ハイブリッド開催,March 2022.

[2] 池田 和史, 柳原 正, 松本 一則, 滝嶋 康弘：くだけた表現を高精度に解析するための正規化ルール自動生成手法，情報処理学会論文誌データベース（TOD）,Vol.3,No.3,pp.68-77,2010

[3] 星野 恵以子, 寺田 篤史, 村上 久, 秋吉 政徳：ソーシャルメディアに現れるくだけた表現を含む口語的表現の自動修正方式，第 79 回全国大会講演論文集,Vol.2017,No.1,pp.595-596,2017