5V-07

OCRを利用した崩れた表記の自動修正手法の提案

秋山 大五郎 † 松原 雅文 ‡

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. はじめに

近年, ソーシャル・ネットワーキング・サービス (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, 特にSNS上のショートメッセージには「ぉはよう」,「ｷｮｩ」のような特有の表現や「ネ申」,「イ為」のような文字の構成要素を組み合わせた表現などの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，本稿では，崩れた表記の単語を正しい単語に修正する手法を提案する．まず，先行研究などで提案されている文脈情報に加えてOCRを用いることで修正候補を取得する．取得した修正候補文字列をスコアリングし，最もスコアが高いものを正しい単語として修正する．

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 [1] 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 [2] 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．

Automatic Correction Method for Informal Words using OCR
Daigoro Akiyama†, Masafumi Matsuhara‡
†Graduate School of Software and Information Science, Iwate Prefectural University, ‡Faculty of Software and Information Science, Iwate Prefectural University

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，崩れた表記の単語のみに対しては処理が難しい．

そこで，本稿では，崩れた表記の単語に対し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する．

3. 提案手法

3.1. 提案手法の概要

提案手法の全体像を図1に示す．提案手法では，崩れた表記を多く含むSNSなどの文書を入力とし，形態素解析を行い未知語を抽出する．抽出した未知語の修正候補を，崩れた表記が少ない文書から得られる文脈情報とOCRを用いて取得する．取得した修正候補をスコアリングし，最適な修正候補を選択する．

3.2. 未知語の抽出処理

崩れた表記が多く含まれた文書から未知語を抽出する．崩れた表記の単語の多くは正しい表記の単語とは異なり形態素解析の辞書に登録されていないため，形態素解析の際に未知語として検出される．また，連続で検出された複数の未知語は1つの未知語として扱う．本研究では形態素解析器にMeCab [1] を用いた．

3.3. 修正候補文字列の検索

3.3.1 崩れた表記が少ない文書を用いた手法

関連研究と同様に検出された未知語に対して，未知語において左右の単語を崩れた表記の少ない文書で検索し，修正候補文字列を取得する．

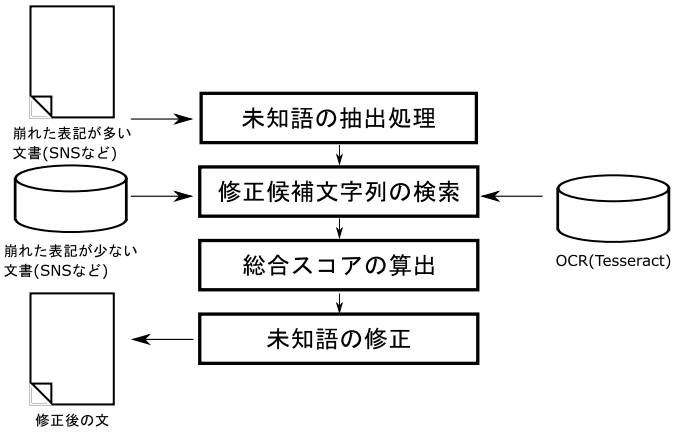

図 1: 提案手法の全体像

[1] http://taku910.github.io/mecab/

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

3.3.2 OCR を用いた手法

検出された未知語について，この画像を生成し OCR を用いて修正候補文字列を取得する．本研究では OCR エンジンに Tesseract [2] を用いた．

3.4. 総合スコアの算出

関連研究と同様，スコアリングに，検索結果における修正文字列の出現頻度，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．検出された未知語に対して文脈情報を用いずに OCR のみを用いて修正候補文字列を取得し，取得した修正候補文字列に対応した正しい表記の単語が含まれているかを評価する．

文字列から生成した画像は，サイズを 150*10 ピクセル，背景色を白，文字の色を黒，文字の大きさを 10 ピクセル，フォントをメイリオに設定した．

4.2. 実験結果

崩れた表記の単語に対して OCR を用いた修正結果を表 1 に示す．各ラベルはぞれぞれ崩れた表記の単語，形態素解析の際に連続して未知語と認識された文字列，未知語として認識された文字列画像の OCR 認識結果，認識結果を置き換えた単語，最終結果が正しい単語かの判断を表している．

表 1: 実験結果

単語	未知語	OCR 結果	最終結果	評価
ｹむし	ｹ	け	けむし	○
さほｏーと	ｏー		さほと	×
ﾚ) んご	ﾚ)	り	りんご	○
にﾚまん	ﾚ	し	にしまん	×

4.3. 考察

OCR を用いた場合，半角 2 文字が未知語として検出される崩れた表記の単語は正しい表記の単語と認識された．しかし，「さほｏーと」(正しい表記はさぽーと) のように半角以外の未知語が検出される崩れた表記の単語や「にﾚまん」(正しい表記はにほん) のように半角 1 文字が未知語として検出される崩れた表記の単語は正しい表記の単語と認識されなかった．これは OCR が半角 2 文字を全角 1 文字として認識した結果であると考えられる．

「さほｏーと」という単語全体に対して OCR を用いた場合，正しい表記の単語ではないが「さほｏ一と」として認識する．しかし，形態素解析器を用いて取得した未知語の「ｏー」から生成した画像の OCR 結果は「」(空) だった．また，「さほｏーと」の崩れた表記である「ほｏ」から生成した画像に OCR を用いなければ，対応する正しい表記の単語である「さぽーと」と認識できない．そのため，OCR を適用する文字列の範囲についても，検討が必要である．

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しい表記の単語を認識する手法を提案した．今回の実験において，半角 2 文字の未知語として検出された崩れた表記の単語については，OCR を用いて正しい表記の単語と認識可能であることが確認された．それ以外の崩れた表記の単語に対応するため，OCR を適用する文字列の範囲についても検討する必要がある．

今後は，OCR から得られる修正候補文字列を増やすため，未知語に対し複数の画像生成を行う．次に，今回の実験では文脈情報を用いていないため，文脈情報を用いた実験を行う．同様に今回の実験では OCR を用いて取得した修正候補に対して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) 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

2) 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

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

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.