文字認識を用いた分割表記文字の処理

情報科学科　乾　亮　　指導教員：山村　毅

1　はじめに

現代社会では，様々な場所に文章が存在し，一般的に新聞や本などでは，正書法に則った正しい文章が使われている．しかし近年では，LINE や Twitter などの SNS が使われるようになり，そこでは，通常は検閲や校正などは行われないので，様々な正規でない表現が含まれる文章が使われることがある．

本研究では，インターネットの文章で見られるような分割表記の文字を含む日本語文章を解析する方法を提案する．

2　分割表記の文字とは

分割表記の文字とは例えば「死」という文字を「歹」と「匕」に分けて，「おまえ歹匕ねよ」のように, 文字を複数の文字に分割して表記するというものである．これは Twitter などで暴言を書き込むときによく使われ, 監視に引っかからないように文を偽装する一つの方法である．このような表記を，本研究では分割表記文字と呼ぶことにする．

3　提案手法

分割表記文字の部分は，コンピュータの側からは，「誤り」と同じであるので，これまでの自然言語処理システムで用いられてきた誤り検出の方法を用いれば，特定することができる (例えば，林ら [1] は，bi-gram 辞書を利用して漢字変換誤り検出を行う方法を提案している)．

本研究では，文字の bi-gram を利用して，文中から分割表記文字部分を抽出し，これを画像に変換したあと，文字認識処理を行なって，分割表記文字を元の文字に変換する．具体的な処理手順は以下の通りである．

1. 文字の bi-gram 辞書をコーパスから計算して用意しておく
2. 入力文の bi-gram を先頭から順に取り出して並べる
3. bi-gram 辞書を用いて入力文の bi-gram で低頻度または存在しないもの (ゼロ bi-gram) を検出する
4. 検出されたゼロ bi-gram の位置から分割表記文字候補の 2 文字を取り出す
5. 取り出した文字をそれぞれ画像にし，サイズや間隔を調整して一つの画像にする
6. 画像を文字認識し，入力文の対応する位置に戻す

システムの流れを図 1 に示す．この図の例ではゼロ bi-gram を検出した結果，「整王」,「王里」,「里す」の 3 つの並んだ部分が検出されている．次にこれらの結果から，「整王」と「里す」で挟まれた「王里」を分割表記文字として取り出し，これを画像に変換，文字認識を行う (「理」となる)．最終的には，入力文は「物を整理する」となった．

4　実験方法

Twitter の文を抜粋して作った 100 文を対象に，システムで評価実験を行った．なお，本研究では 1 文につき分割表記文字が 1 つの文を対象としている．

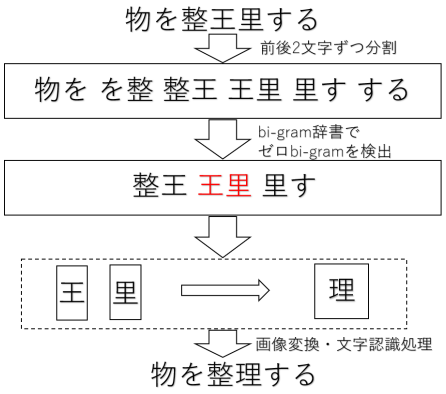

図 1　システムの流れ

5　実験結果と考察

結果を表 1, 成功した例を以下に示す．

表 1　評価実験結果

成功した文	失敗した文	合計
69	31	100

早く歹ヒね → 早く死ね
糸エ白に出る → 紅白に出る

失敗の内訳を表 2, その説明を以下に示す．

表 2　失敗の内訳

内訳	文数
文字認識の失敗	19
ゼロ bi-gram 未検出	5
ゼロ bi-gram の過剰検出	7

文字認識の失敗

文字認識で, 誤った結果が出力されたことによる失敗

ゼロ bi-gram 未検出

検出で, ゼロ bi-gram が見つからなかったことによる失敗

ゼロ bi-gram の過剰検出

検出で, 分割表記文字でない文字列が誤ってゼロ bi-gram として検出されたことによる失敗

6　おわりに

本研究では，文字認識を取り入れることで分割表記文字を含む日本語文章を解析する方法を提案した．文字認識を使うことは有効ではあったが，検出の仕方，文字認識結果の複数候補の出力など課題があることがわかった．

今後は，本研究で得られた課題や，1 文につき分割表記文字が 2 つ以上の文の場合について研究を取り組みたい．

参考文献

[1] 林 秀治, 山本 和英: “漏れのない漢字変換誤り検出と誤り可能性によるレベル分け”, 言語処理学会第 22 回年次大会発表論文集, pp.1145–1148, 2016