AKAZE と文字バイグラム確率を用いた分割表記文字の処理について

情報科学科　山中 健生　　指導教員：山村 毅

1　はじめに

現在、私たちの生活でインターネットや SNS を利用する機会が増えおり、誰でも情報発信ができるようになっている。そのため、私たちの周りには非常に多くの情報が溢れており、文章を分析する自然言語処理技術に対する需要が高まっている。しかし、SNS では正書法に従っていない文章もあるため、伝統的な自然言語処理では十分には対応できない。本研究では 「死ね」→「歹匕ね」のように１つの文字を複数に分割した表記について研究を行う。このような表記を本研究では分割表記文字と呼ぶこととする。先行研究である乾ら[1]の手法では分割表記文字の対象となる画像を OCR(画像データを文字データに変換する技術)を用いて文字認識することで、分割表記文字の処理を行なっているが、用いた OCR のライブラリが文字認識結果を一つしか出力できなかったため、文字認識の段階で誤ってしまうと正しく分割表記文字を処理できなかった。

そこで本研究では OCR の結果を複数出力させ、それらから適切なものを選ぶことでこの問題に対処する。

2　先行研究

乾ら[1]は、bi-gram 辞書を利用して文中から分割表記文字の部分を探し出し、これらを画像に変換したあと、OCR にて文字認識処理を行い、その文字を文の対応する位置に戻す方法を提案した。この時用いた OCR の Tessaract では結果が一つしか出力されず、出力結果が失敗ならうまく処理できないという問題があった。

3　提案手法

文中から分割表記文字の部分を抽出し、これを画像に変換したあと、AKAZE を利用して文字認識処理を行い複数の候補を出力できるようにし、文字の bi-gram を用いてその中から適切なものを選出する。具体的な手順を以下に示す。

1. 文中から分割表記文字の候補の部分を探し出す
2. 取り出した文字をそれぞれ画像にして、サイズ調整を行い一つの画像にする
3. AKAZE を用いて文字認識処理の候補を複数出力する
4. bi-gram 辞書を用いて複数の候補の中から適切なものを選出する
5. 選出した文字を入力文の対応する位置に戻す

4　評価実験

毎日新聞の文章から 100 個の分割表記文字を抜粋し、作成したシステムを用いて評価実験を行った。また、表記による違いを知るため、「明朝体」「丸ゴシック体」「角ゴシック体」の 3 つのフォントごとで実験を行った。さらに、比較のために Tessaract を用いた時の成功数も求めた。なお、問題を簡単にするため、分割表記文字の位置はあらかじめ与えておいた。

5　結果

成功結果を以下の表 1、失敗の内訳を以下の表 2 に示す。

表 1 評価実験結果・成功数

(データ数 100)	明朝体	丸ゴシ	角ゴシ
Tessaract	9	11	16
AKAZE の第一候補	22	32	26
AKAZE+バイグラム	66	68	62

表 2 評価実験結果・失敗内訳

(データ数 100)	明朝体	丸ゴシ	角ゴシ
AKAZE での出力失敗	31	30	37
バイグラムで選択失敗	3	2	1

失敗内訳において、「AKAZE での出力失敗」とあるのは、AKAZE で 10 個の候補を出したが、その候補の中に正しい変換の文字がなかったことによる失敗を指す。また、「バイグラムで選択失敗」とは、AKAZE の候補の中に正しい変換の文字はあったが、それ以外の文字を選んでしまったことによる失敗のことである。

6　考察

AKAZE は画像の特徴量を利用して複数候補を出しており、それゆえに簡単な単語は特徴が少なく、成功しにくい傾向があるのではないかと推測する。例えば、「イ白(伯)」、「シ工(江)」、「イ西(価)」は、3 つのフォントいずれにおいてもうまく文字認識できなかった。

バイグラムに数字や区切り記号が含まれている場合に、似たものはあるが全く同じものがないことがあるが挙げられる(例えば「『孫」はあったが、「【孫」はないなど)。バイグラム辞書を利用する際、文字そのものではなく、文字の種数(例えば「１人」でなく「<数>人」)を利用することなどの工夫が必要である。

7　まとめ

本研究では AKAZE を用いることで、OCR での文字認識の失敗を減らす手法を提案した。AKAZE とバイグラムを用いることでかなり精度が上昇したが、簡単な単語は特徴が少なく AKAZE の候補に出にくい傾向があることや、バイグラムの利用時に分割表記文字の前後に数字や符号がある時の処理の方法に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、今回扱わなかった分割表記文字の位置を特定すること、日常で使う話し言葉は新聞記事で取り扱っておらず、そういったものをどう取り扱っていくのかについて研究を取り組んでいきたい。

参考文献

[1] 乾　亮, 山村　毅: "視覚的「読み」を用いた分割表記文字の処理", 言語処理学会第 25 回年次大会発表論文集, pp.434-437(2019)