ISS-P-022

古文書翻刻ツールにおける パターン辞書を用いた認識候補文字の一意性

片山歩希 †　　松尾賢一 † †
Ayuki Katayama　　Ken'ichi Matsuo

奈良工業高等専門学校 システム創成工学専攻 †　　奈良工業高等専門学校 情報工学科 † †

1　まえがき

歴史上の出来事を理解する上で古文書は文献資料の一種としての大きな役割を果たす．しかし，江戸時代以前に書かれた古文書は書体や体裁により簡単に理解することができない．そのため，日本で古文書に対して現代語訳まで翻刻（解読）する者は日本人口の 0.01%未満といわれている [1]．現在では，翻刻者を増加させるために初めて翻刻を学ぶ初学者に向けて様々なツールが開発されている．本稿では，ツールの中でも自動くずし字認識ツール [2] に着目した．着目したツールでは，一文字画像を認識ルーチンにかけ，複数の候補文字が出現する．しかし, 初学者は複数の候補文字から文や単語に沿った文字を選択するために難易度が高い．

そこで本研究では，一文中の連続した一文字画像を複数枚入力することで，初学者が文や単語に沿った文字を選択することなく使用できるツールを開発，評価する．

2　提案する古文書翻刻ツール

2.1　古文書翻刻ツールの流れ

本研究で開発するツールは，図 1 のように動作する．28x28 ピクセルに正規化した各カラー画像を単一くずし字認識ルーチンに入力する．ルーチンでは，入力画像に類似した文字を候補文字として複数出力する．次に，候補文字からすべてを組み合わせ，文字列を生成する．最後に，生成した文字列と予め用意したひらがな単語辞書 DB と比較することで，実際に存在する文字列のみを出力する．

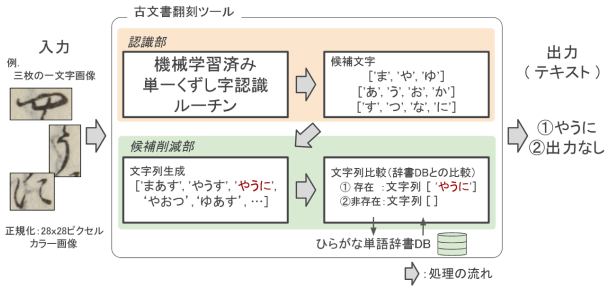

図 1: 古文書翻刻ツール 全体図

2.2　認識部について

認識部では，畳み込みニューラルネットワークによる教師あり学習を用いて，一文字画像を認識する．学習用のデータは CODH が作成したデータセット（訓練用：232,365 枚 テスト用：38,547 枚）を利用する [3]．

2.3　候補削減部について

候補削減部では，まず認識部で出力された候補文字をすべての組み合わせることで，文字列を生成する．組み合わせた文字列とくずし字用例辞典をもとにしたひらがな単語辞書 DB と比較し，DB に存在する文字列と総和スコアを出力する．

3　実験及び結果

本ツールの評価として，源氏物語 総角（あけまき）から 1 文字ずつ画像が分かれている 2-7 文字の文字画像を 115 セット分用意する．次に，用意したデータセットをツールに入力し，出力された結果からツールを評価する．

本ツールにおける評価項目は，出力回数，正解を含む出力回数，一意に決定した出力回数である．さらに，本ツールによって削減された文字列についても評価とする．

表 1: 源氏物語総角 115 セットに対しての出力回数

出力回数	正解を含む出力回数	一意に決定した出力回数
101	88	28

表 2: 源氏物語総角 115 セットに対しての削減数

生成した文字列数	出力された文字列数
6665025	3801

4　あとがき

開発したツールの性能評価として，結果より 1 入力あたりの出力数が平均して 36 となっている．これは，入力が二文字である単語の場合のときに生成される文字列が一意に定まらず，複数の二文字の単語候補を挙げてしまうこと．パターン辞書中の 3 文字や 4 文字なども候補となるために二文字単語以上の候補が出力されてしまうことが原因である．今後の課題としては，出力を一位にするために，一文での評価をする必要があると考える．

参考文献

[1] 橋本雄太．Ai 文字認識とクラウドソーシングを組み合わせた歴史資料の大規模テキスト化．人工知能, Vol. 35, No. 6, pp. 754–760, 2020.

[2] Tarin Clanuwat, Alex Lamb, and Asanobu Kitamoto. Kuronet: Pre-modern japanese kuzushiji character recognition with deep learning. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 607–614, 2019.

[3] CODH. 日本古典籍くずし字データセット．