7ZK-05

日本近代公文書画像における文脈を考慮した文字検出手法
A Context-Aware Text Detection Method for Modern Official Japanese Document Images

宮川 裕貴† 山田 雅之† 中 貴俊† 兼松 篤子†
Yuki Miyagawa Masashi Yamada Takatoshi Naka Atsuko Kanematsu
宮崎 慎也† 長谷川 純一‡
Shinya Miyazaki Junichi Hasegawa

1. はじめに

各行政機関が保管する公文書は政治決定の背景や日本国内外の情勢を知ることができるというように史料的価値があるが，戦前期の文書の多くは近世古文書の流れを汲む手書き文字による文書であるため，解読には専門的な知識が必要である．したがって一般行政職員が解読することは容易ではなく，また，解読の知識を持つ研究者も少ないため公文書史料が活用できていない．そこで，我々の研究グループは台湾総督府文書を題材として自動解読システムとそのためのデータセットの開発を進めている．本稿では Transformer[1] をベースとした文脈情報の考慮が可能な文字認識モデルを提案する．そして，認識精度の検証を通じて本モデルの有効性を検証する．

2. 関連研究

歴史的文書に対する文書画像認識技術として Lamb らが古典籍を対象として KuroNet と呼ばれる U-Net ベースのモデルを提案している[2]．Le らは Attention と LSTM を組み合わせたモデルを提案し，Kindai-OCR V1.0 として公開している．また，Kindai-OCR V2.0 では Transformer ベースのモデルを採用しており，本研究とは題材としている書籍の時代やモデル構造で強く関連していると言える[3]．

3. 近代公文書データセット

我々が開発を進めている近代公文書データセットは，近代の日本の行政文書の中で唯一体系的に保管されている台湾総督府文書を題材とし，現時点で 5,002 ページ分の文書画像と画像内の各文字に対応するアノテーションデータで構成される．図 1 に示すように旧字体と新字体の混在や略字，くずし字が確認できる他，文字の大きさや間隔が不揃いであることが特徴である．

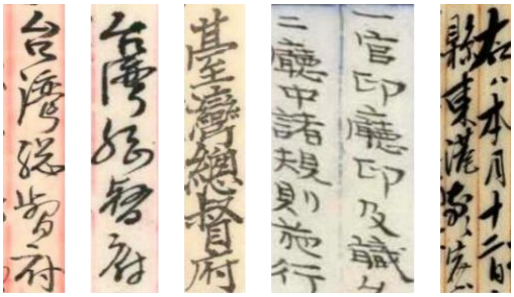

図 1：様々な字体

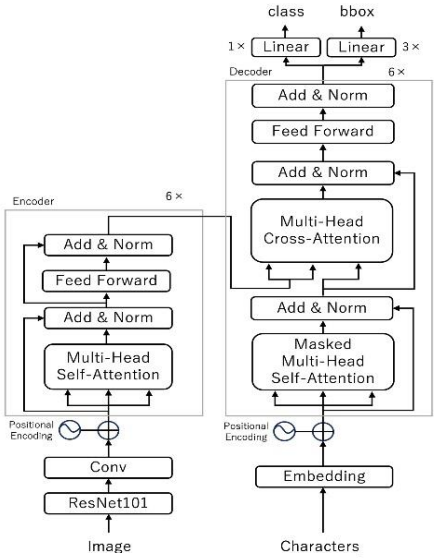

図 2：提案モデル

4. モデル構造

提案モデルの構造を図 2 に示す．本モデルの

† 中京大学 Chukyo University
‡ 中京大学人工知能高等研究所 Institute for Advanced Studies in Artificial Intelligence

表 1：適合率・再現率

	適合率	再現率
1 行画像	0.8806	0.8776
複数行画像	0.1897	0.1814

Encoder は画像を，Decoder は文字列の埋め込みベクトルを入力とする．これにより，文脈情報を考慮した推論が可能になる．モデルの学習時は n 文字目までの情報から $n+1$ 文字目の予測が行えるように学習を進める．推論の際には，はじめに 1 文字目を予測し，その結果を Decoder ブロックに入力することで次の 1 文字を予測する．これを指定した上限文字数に達するか終わりを示す特殊トークンが出力されるまで繰り返す．

5. 認識実験

本実験では近代公文書データセットを元とした 99,637 枚の 1 行画像によるデータセットと，34,467 枚の 3 行までの複数行画像によるデータセットを作成した．両データセットは訓練:検証:テスト用に 8:1:1 に分割し，50 エポックの訓練を実施した．訓練時のハイパーパラメータは初期学習率 2×10^{-5}，最適化関数 AdamW，Transformer の特徴量次元 256，FFN 次元 2048 とした．損失は $L = L_{Class} + \alpha L_{L1} + \beta L_{GIoU}$ とする．L_{Class} はクラス分類のクロスエントロピー誤差，L_{L1} は平均絶対誤差による矩形の回帰誤差，L_{GIoU} は Generalized IoU[4] によって与えられる値である．テストデータに対する適合率，再現率を表 1 に示す．1 行画像では適合率が 0.8806，再現率が 0.8776 の一方，複数行画像では適合率が 0.1897，再現率が 0.1814 と大きく悪化している．図 3 に示すモデルの出力例において，1 行画像では全ての文字領域とクラスラベルが正しく検出されている．複数行画像に対しては文字領域が正しく検出されていないものがあり，また，クラスラベルの検出が誤っている場合が多い．複数行画像で精度が悪くなっている理由としては，1 行画像と比べて学習データ数が減少していることや，改行の特徴が学習できていないなどが考えられる．

図 3：モデルの出力例

6. おわりに

本稿では文書画像認識のためのモデルとして Transformer をベースとした文脈情報を考慮可能なモデルを提案した．認識実験では 1 行画像では高い精度を得られたが，複数行画像では精度に課題が残る．今後はモデルパラメータの増加や学習の際に改行を示す特殊トークンを挿入するなどを通じて精度向上を目指す．

謝辞

本研究は JSPS 科研費 JP20J01304，および中京大学戦略的研究「デジタル・ヒューマニティーズプロジェクト：日本近代公文書自動解読システムの開発」の助成を受けた．

参考文献

[1] shish Vaswani, Noam Shazeer, et al. “Attention is All you Need”, NIPS, pp.5998-6008, 2017
[2] Alex Lamb, Tarin Clauwat et al. “KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition”, SN Computer Science, Vol.1, No.177, pp.1-15,
[3] Anh Duc Le, Daichi Mochihasi et al. “Recognition of Japanese historical text lines by an attention-based encoder-decode and text line generation”, https://doi.org/.10.1145/3352631.3352641
[4] Hamid Rezatofighi, Nathan Tsoi et al. “Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression”, CVPR, pp.658-666, 2019