

# AI技術を用いた古典籍OCRテキストの 解析・校正支援手法

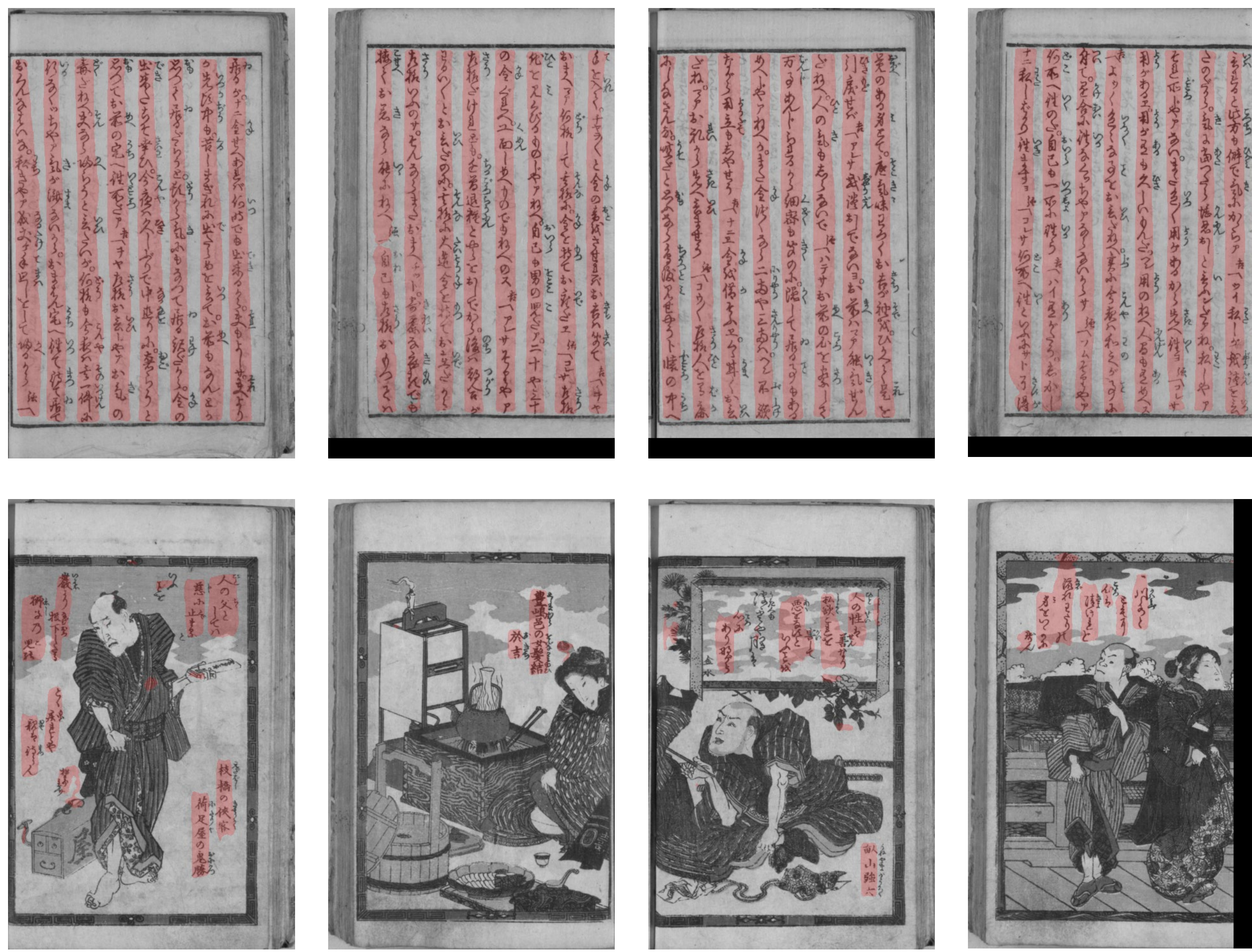
法政大学 藤田 悟、国文学研究資料館 大山敬三  
問い合わせ先: fujita\_s@hosei.ac.jp

## 1. 研究課題

古典籍に現れる崩し字の解析や校正を深層学習を用いて実現し、翻刻を支援する技術の研究を行う。特に、Transformer を一行分の文字列認識に適用して、崩し字の文字領域認識と文字コード認識を同時に進める技術について実用レベルでの認識精度を目指す。

## 2. 1行の文字領域の抽出

縦書き文字領域の抽出には、U-Net/CNN と呼ばれる深層学習を用いたセマンティックセグメンテーション技術が利用できる。学習データとして、文書や挿絵中の文字領域データを与えたところ、図1に示すような文字領域が抽出できた。



出展: 比翼連理花廻志満台 四編上

図1 文書や、挿絵の中の文字領域の抽出結果

## 3. 縦書き文字領域の認識

縦書き文字領域は、再び、U-Net/CNN を用いてパターン認識した後、1ピクセルごとに横長の細い短冊状のデータの集合としてデータ化する。このデータから Transformer により、縦方向の関係性を抽出して、個別文字領域の抽出と、文字コードを同定する。

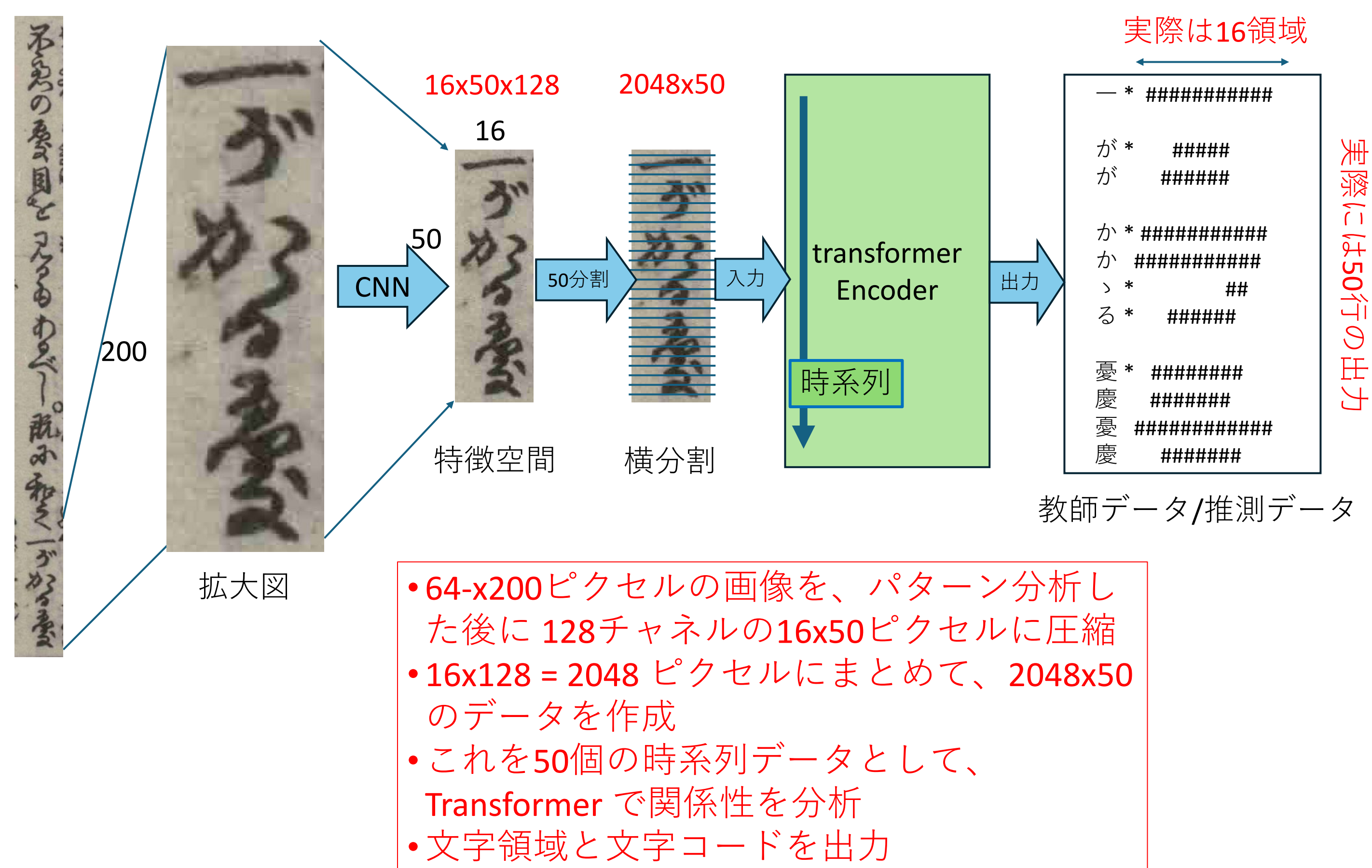


図2 縦につながる崩し字の認識

## 4. 翻刻の作成

図2の出力をさらに Transformer Decoder で分析し、最終的な古典籍の翻刻を生成する。。

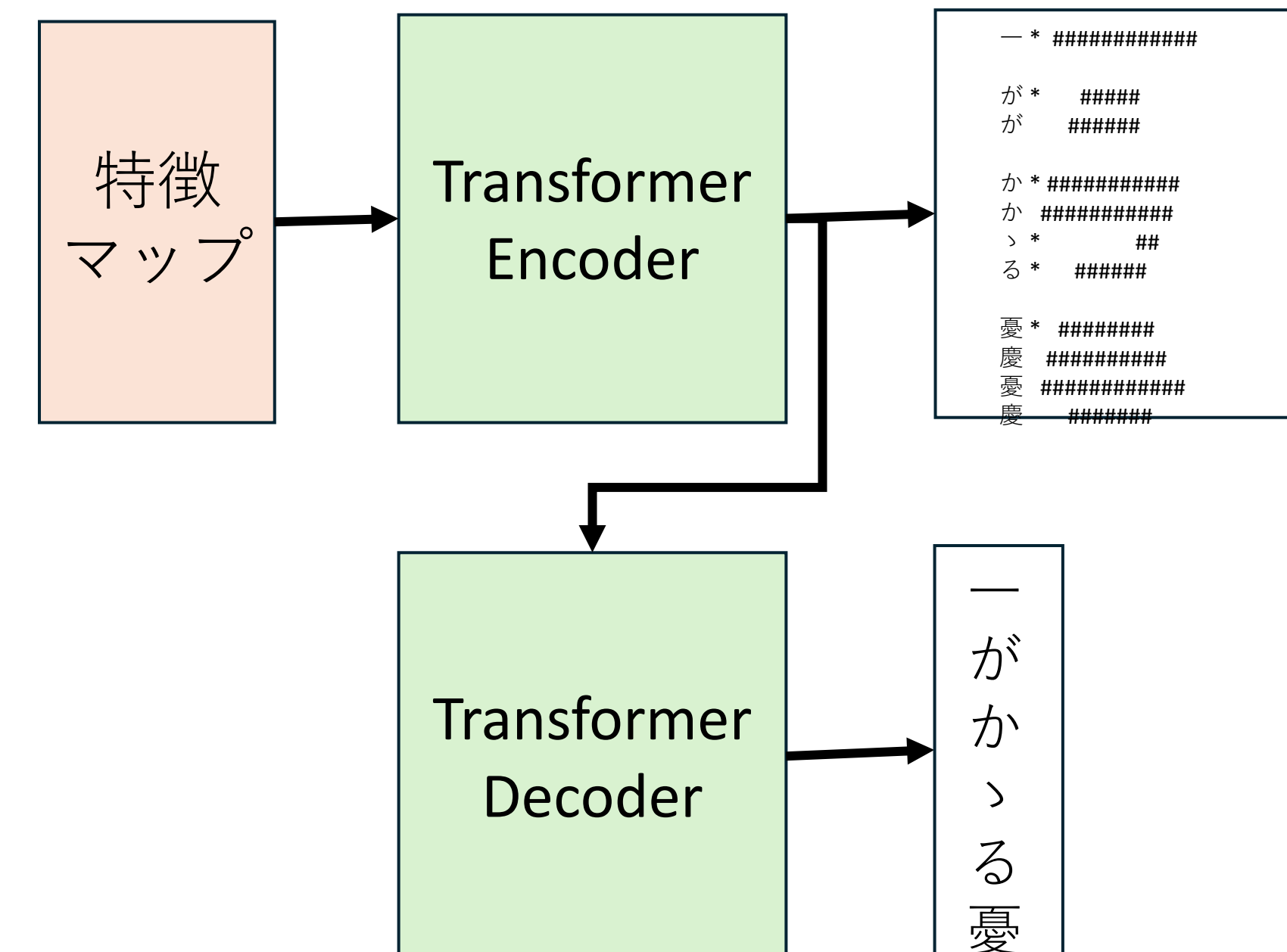
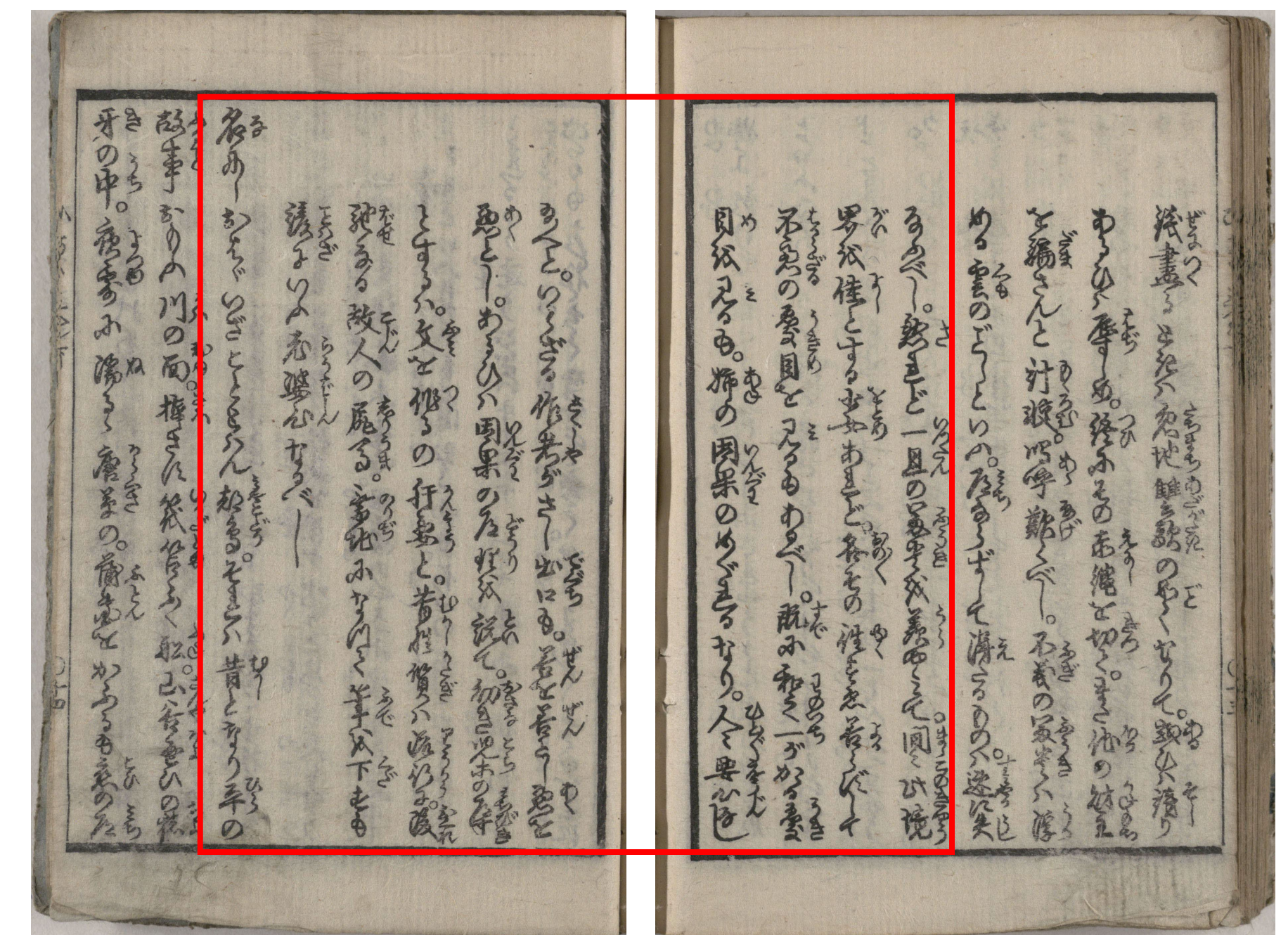


図3 最終翻刻を出力する Encoder Decoder の構成

## 5. 認識結果

Kaggle kuzushiji Recognition\* データセットを用いて、学習と認識精度評価を行う。

\* <https://www.kaggle.com/competitions/kuzushiji-recognition>



出展: 比翼連理花廻志満台 三編下

図4 認識実験に用いた古典籍の例

### 認識結果

なふべし然れど一旦の富生を美やみて間々此境  
果を倅とする女女あれど各その往す善らずして  
不悪の厚目を見るもあるべし既に和之一がかる慶  
目を見るも姉の因果のめれるなり人々異公しし  
給へといらざる作者がさし出口も善を善とし悪を  
悪としあるひは因果の道理を説き児等の節  
とするは文を作るの肝要善れ質は跡行に後  
馳なる談人の尻尊書地になつて筆を下すも  
諺にいふ老婆心なるべし  
名にしおはづいざことはん都鳥それは昔となり平の

※ 赤字は誤認識、緑字は文字不足

表1 認識精度

| モデル    | 適合率   | 再現率   | F値    |
|--------|-------|-------|-------|
| 図2のモデル | 0.896 | 0.900 | 0.893 |
| 図3のモデル | 0.865 | 0.864 | 0.865 |

※ 図3のモデルは、最終翻刻作成に課題を残す

## 6. 今後の課題

出現頻度の低い漢字認識の精度向上のため、ふりがな認識結果の活用などの研究を進める。

### 発表文献

- [1] Fujita, S. and Oyama, K., Learning a Sequence of Cursive-Style Japanese Characters in Classical Literary Works, Asia Pacific Signal and Information Processing Association, Annual Summit and Conference, 2024.