邦字新聞に対応したOCRシステムの開発

人間文化総合科学研究科　情報衣環境学専攻　生活情報通信科学コース

城研究室　博士前期課程　2年

22720045　熊谷もも

目次

1 研究背景

邦字新聞とは

- 明治維新以前の時代に刊行
- 当時のアメリカ大陸，アジアにて日本人移民により刊行
- コミュニティ，政治，軍事プロパガンダ等

邦字新聞デジタル・コレクション

- スタンフォード大学フーバー研究所が公開
- 邦字新聞画像データに誰でもアクセス可能

研究への活用

全文検索機能の実現

全文自動テキスト化が求められている

『奥州日報』1918/05/01 page.01

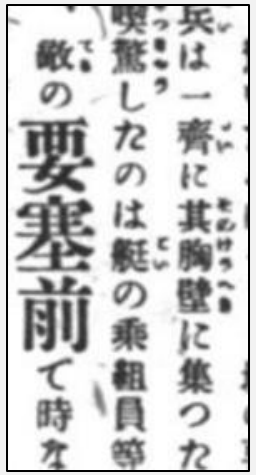
邦字新聞画像データのテキスト化

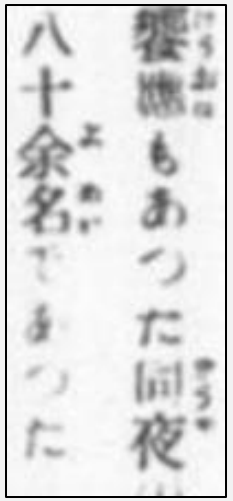
邦字新聞デジタル・コレクション
『Dōhō, 1941.02.15』より抜粋

露∞規に繋顳が

テキスト化の成果は不十分

- 図や広告を含む複雑な構成
- 活版印刷による不統一な文字サイズ
- 損傷等による新聞の質の低下





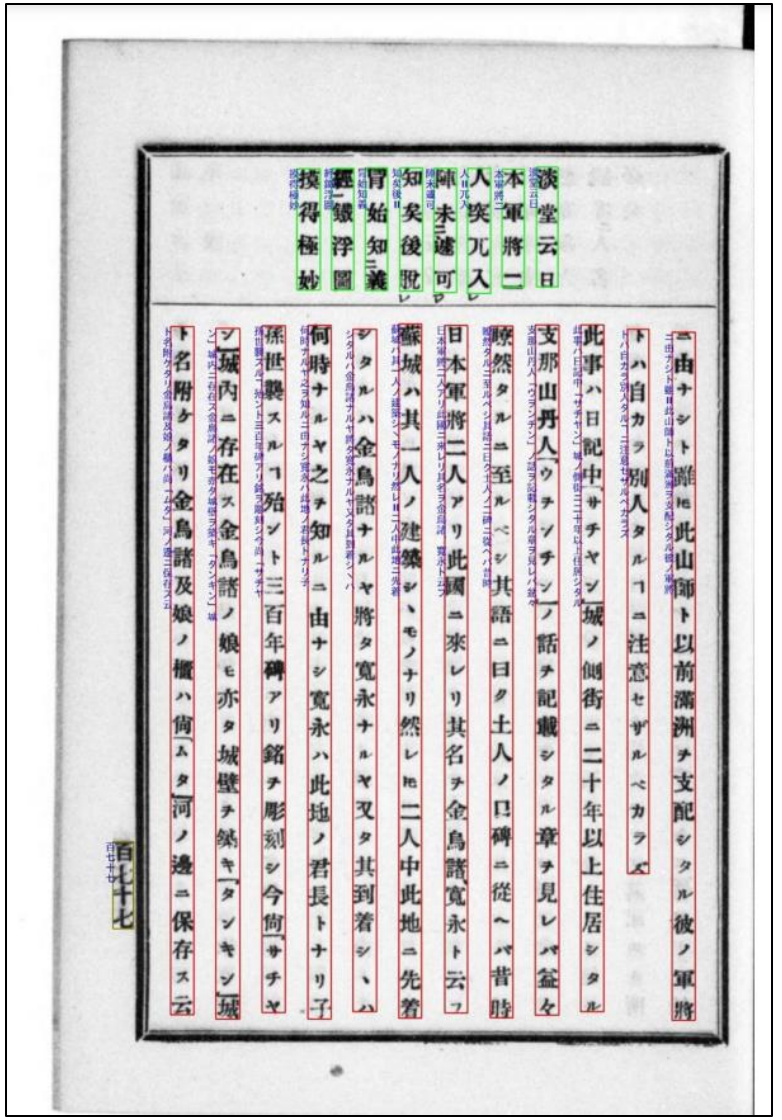
邦字新聞に対応したOCRシステムが求められている

2 NDLOCRについて

NDLOCRとは

- 日本語文書に対応したOCR
- 国立国会図書館の所蔵する約262万タイトルの資料
- 幅広い年代の資料を利用
- 日本語文書に用いられるほとんどの文字種に対応

様々な日本語文書に対して高精度なOCRが可能



『義経再興記』
https://dl.ndl.go.jp/pid/782055/1/93

NDLOCR（ver 2.1） 処理の流れ

前処理

- 見開き分割
- 傾き補正

レイアウト解析

- RCNN
- 行領域抽出

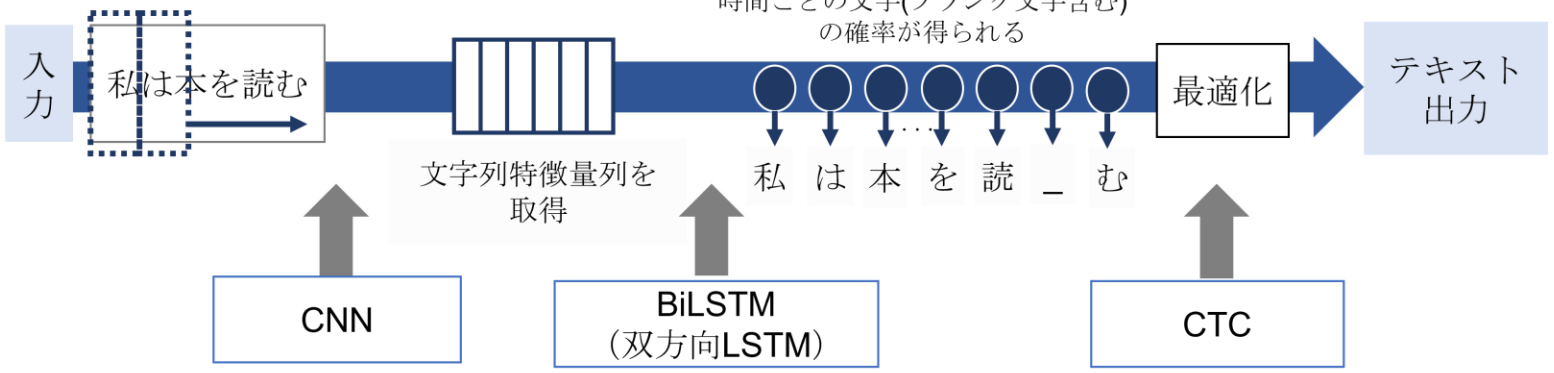
文字認識

- **シーケンス認識**

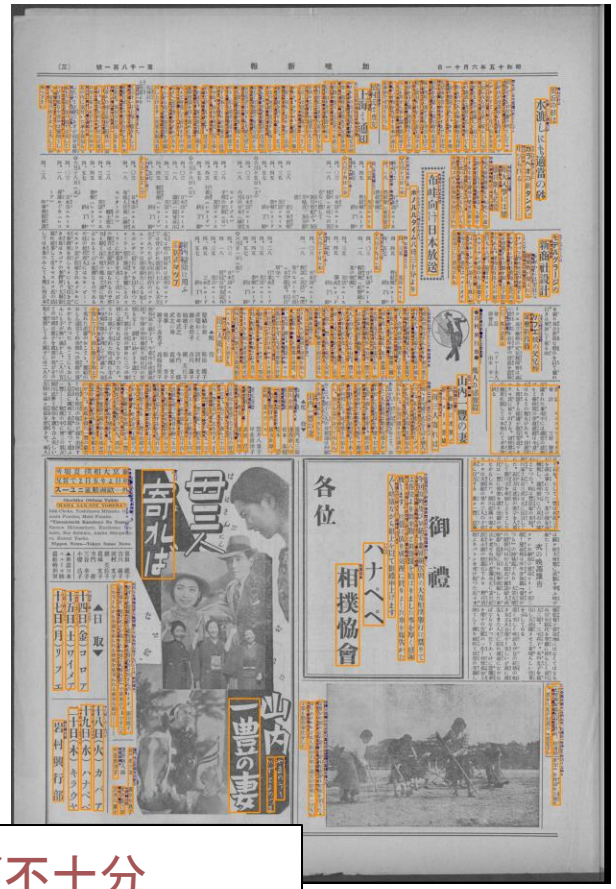
読み上げ機能用処理

- XYCutによる読み順の推定
- 見出し
  著者推定
- 漢字の
  読み推定

シーケンス認識



邦字新聞に対するNDLOCRの精度



認識精度が不十分

邦字新聞に対応するOCRを目指す

一部状態の良いもの
解像度が高いものは高精度に認識可能

邦字新聞に対するNDLOCRの課題点

1 レイアウト解析精度の低さ

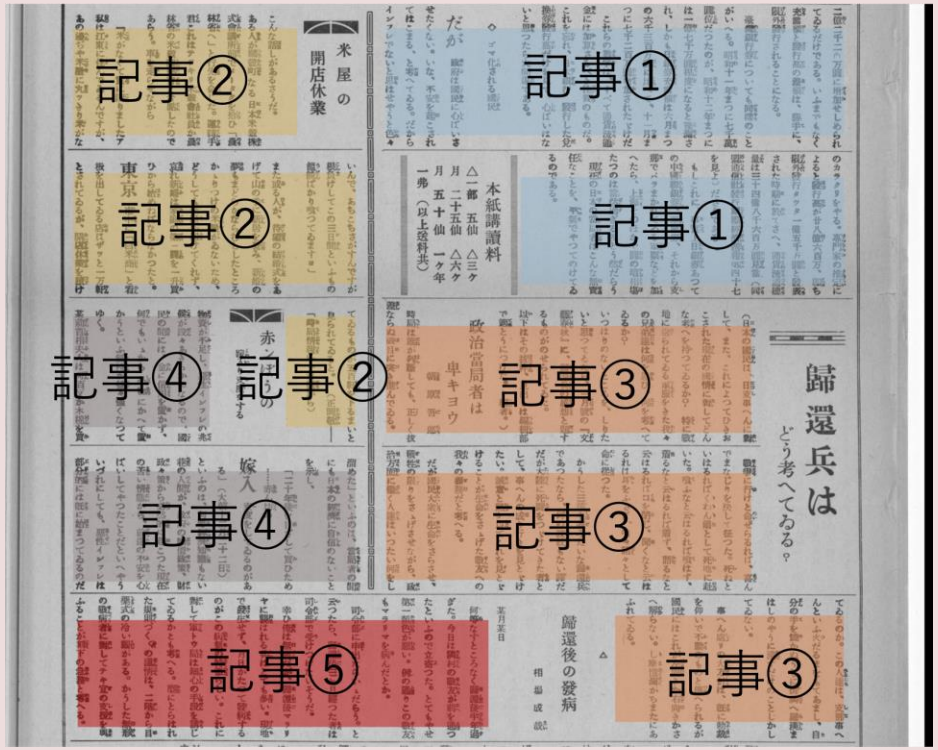
- 多段組み・複雑なレイアウト構成に対応する手法を検討

2 ルビが除去できない

- ルビがノイズとなり文字認識率低下
- 規格化されていない文字・位置に対応する手法を検討

3 読み順検出精度の低さ

- 記事が複数含まれ，構造が複雑化
- 簡単なルールベースでは決定できない



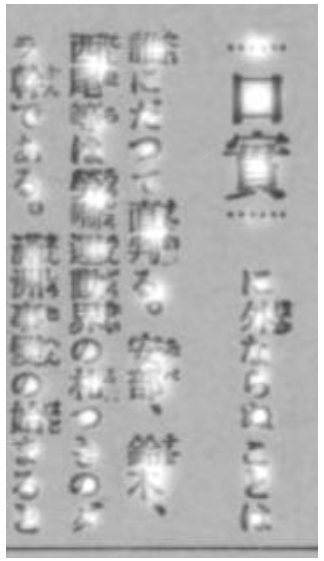
3　邦字新聞に対応したＯＣＲ

① レイアウト解析手法の改善

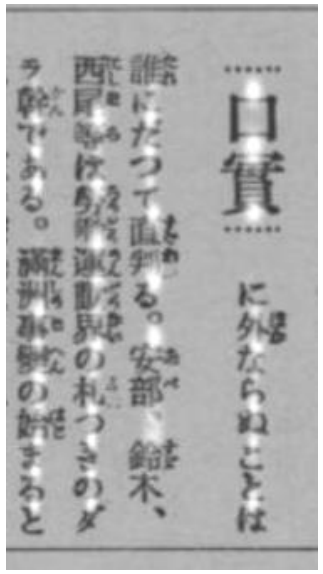
■ CRAFT

- テキスト領域検出手法
- CNNで2つのヒートマップを出力

Region Score



Affinity Score



■ 解像度ピラミッドを適用したCRAFT

CRAFT ＋ 解像度ピラミッド

- 計算資源を削減
- 膨大，質の高いデータによる学習モデル構築
- 多段組みレイアウトの近代書籍に対応

行領域抽出処理の流れ

解像度ピラミッドを適用した CRAFT

行領域矩形推定

拡大処理

行領域出力

■ 行領域を推定

Region Score

Affinity Score

2値化

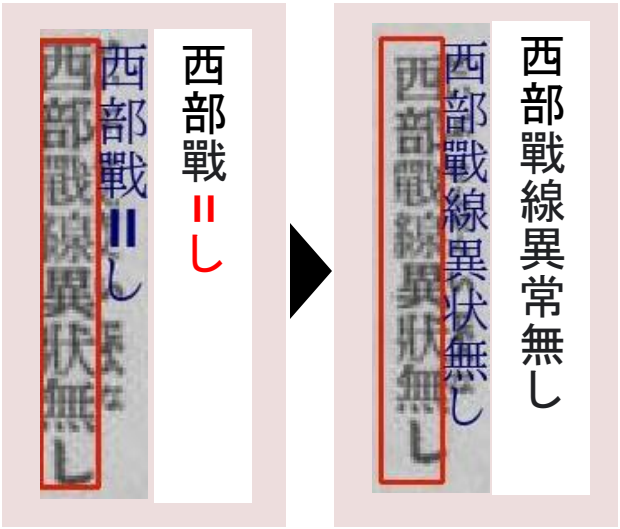
加算

連結成分ラベリング

外接する長方形領域を取得

■ 行領域を拡大処理

- 文字の端々が切れる問題を解決
- 上下左方向の行領域拡大
- 文字認識率向上につながる



② ルビ除去処理の追加

行領域抽出

ルビ除去前

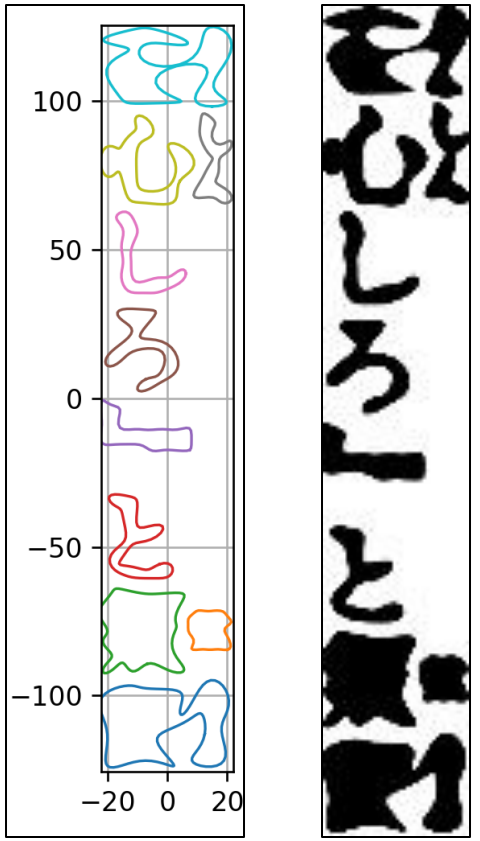
前処理

- Canny法
- 輪郭点取得

■ フーリエ記述子による輪郭線記述

楕円フーリエ記述子

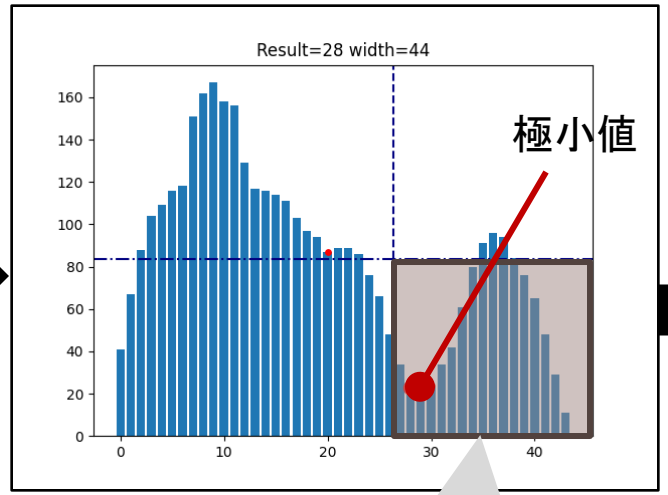
- 座標情報を周期関数として捉える（閉曲線）
- フーリエ級数展開により得られるフーリエ係数から形状を近似する
- 展開次数が大きいほど輪郭が微細に



② ルビ除去処理の追加

輪郭概形

■ 濃度ヒストグラムによるルビ有無判定



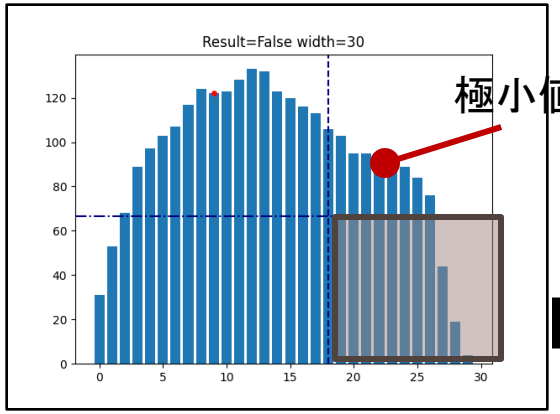
- 縦方向に濃度ヒストグラムをとる
- ヒストグラム谷となる位置で
  ルビ有無判定

判定範囲

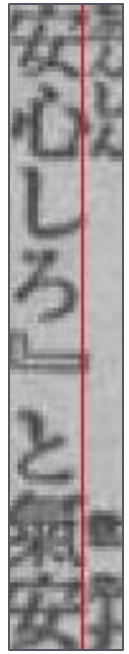
最大濃度値 1/2 以下

右側 2/5 の範囲内

ルビ無しと判定されるもの



ルビ除去



ルビ除去なし

③ 読み順検出手法の改善

■ 邦字新聞のレイアウト構成

- 複数の記事により構成
- 記事ごとにレイアウトが異なる

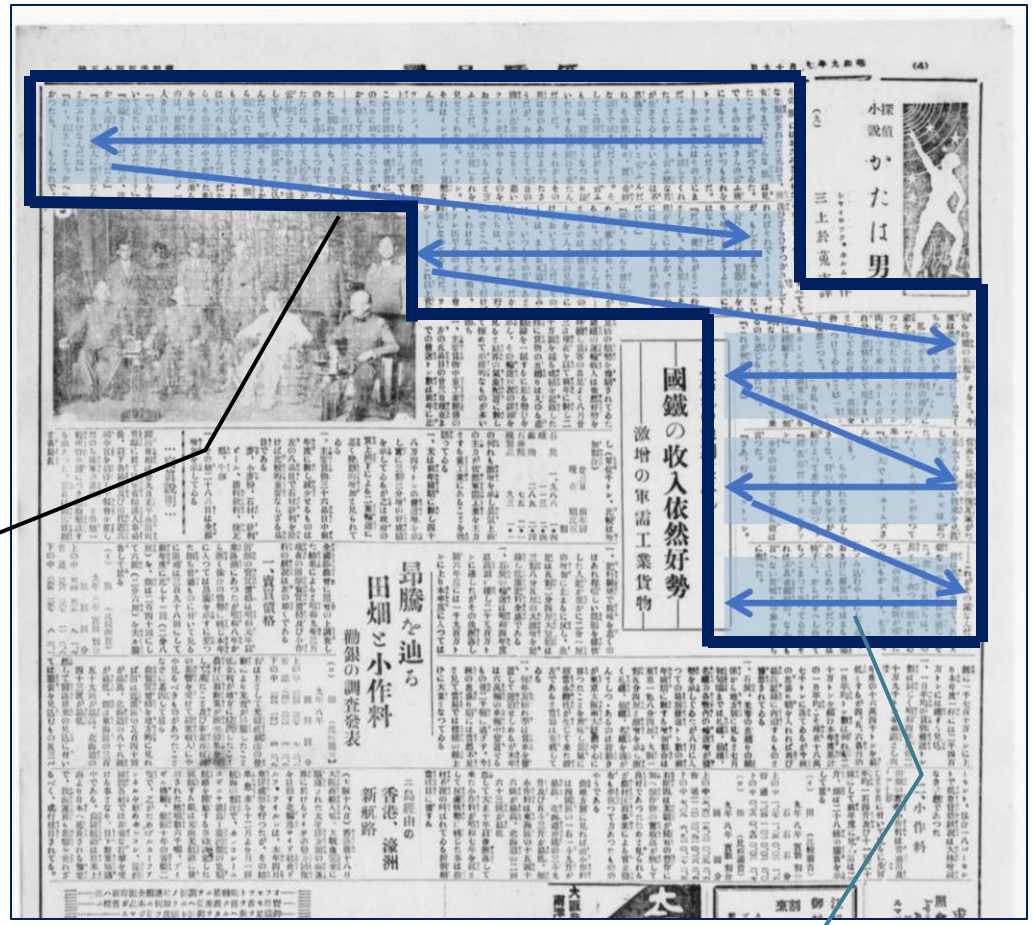
主題に関連した文章

文章ブロック ①

文章ブロック ②

…

同一記事



文章ブロック

文章類似度計算による同一記事推定

同一記事の推定

■ 文書同士の類似度計算手法

単語の重要度計算

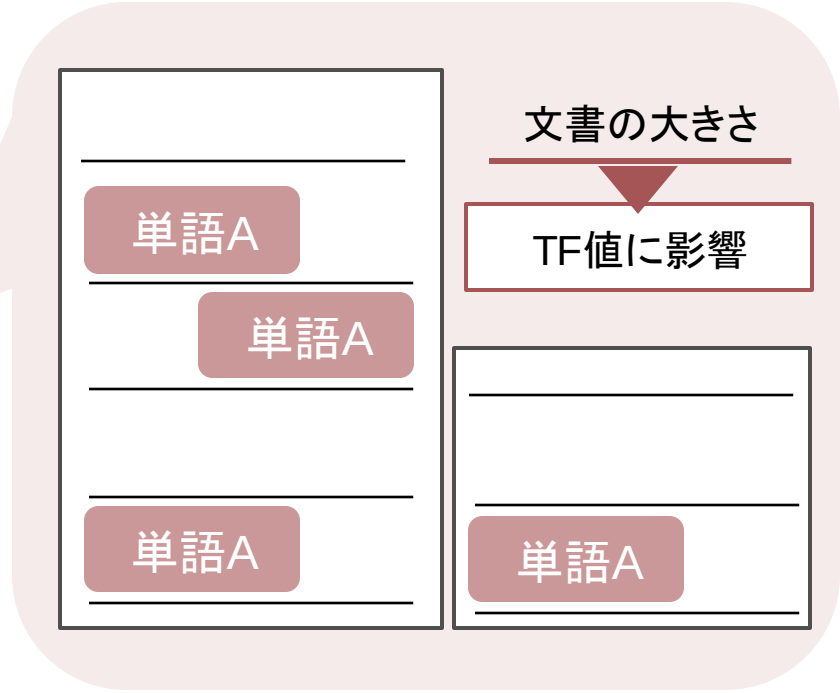
- TF-IDF

TF値	×	IDF値
文書内の単語出現頻度		文書集合中の単語出現頻度

- **Okapi-BM25**

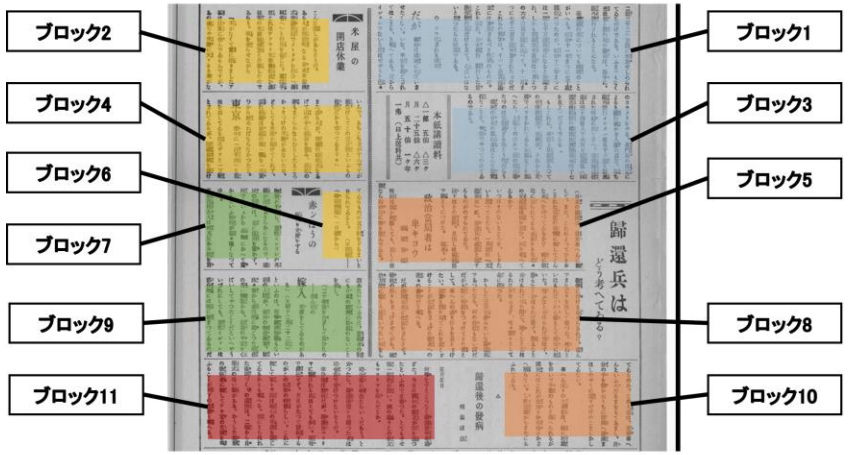
TF値	＋	IDF値	＋	DL値
				文書内の総単語数

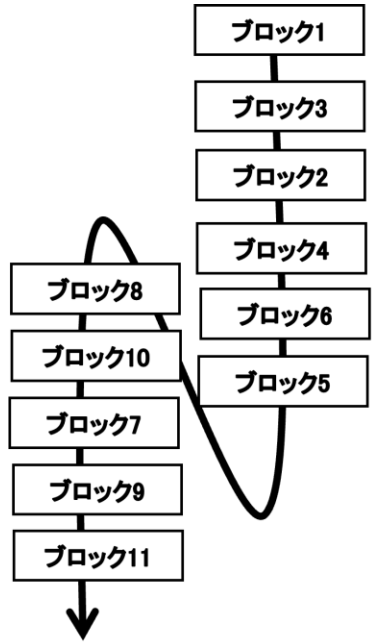
  - TF-IDFの改善手法
  - DL 値の平均値を用いて文書ごとの単語数の差による影響を軽減



■ コサイン類似度 → 文章ブロック同士の類似度を算出

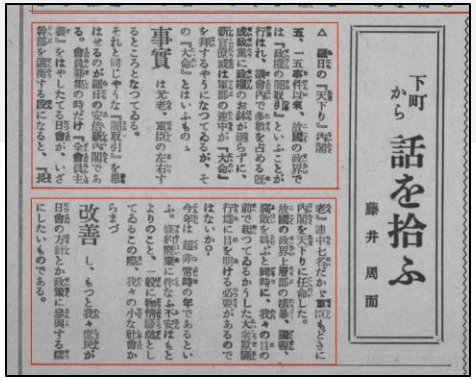
読み順検出処理の流れ





文章ブロックの分割 → 形態素解析 → 同一記事の推定 → 後処理

- Affinity Scoreの拡大処理



- MeCabの利用
- 名詞を抽出

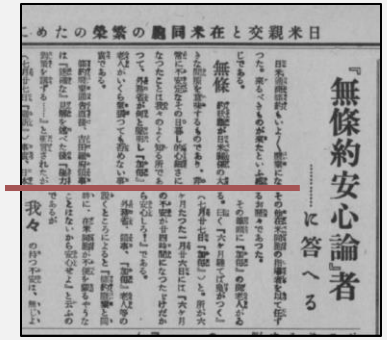
- Okapi-BM25とコサイン類似度
- 類似度0.2以上

- 並び替え

提案手法　処理の流れ

前処理 → レイアウト解析 → 文字認識 → 読み順検出

- 2値化
- 横線の除去



解像度ピラミッドを適用したCRAFT

ルビ除去

- フーリエ記述子
- 濃度ヒストグラム

- シーケンス認識

同一記事の推定

- 文章ブロック分割
- Okapi-BM25

実行環境の構築 → **Docker**

- コンテナ型仮想化ソフトウェア
- 実行環境のライブラリなどをまとめて分離

移植性　再現性

4 検証と考察

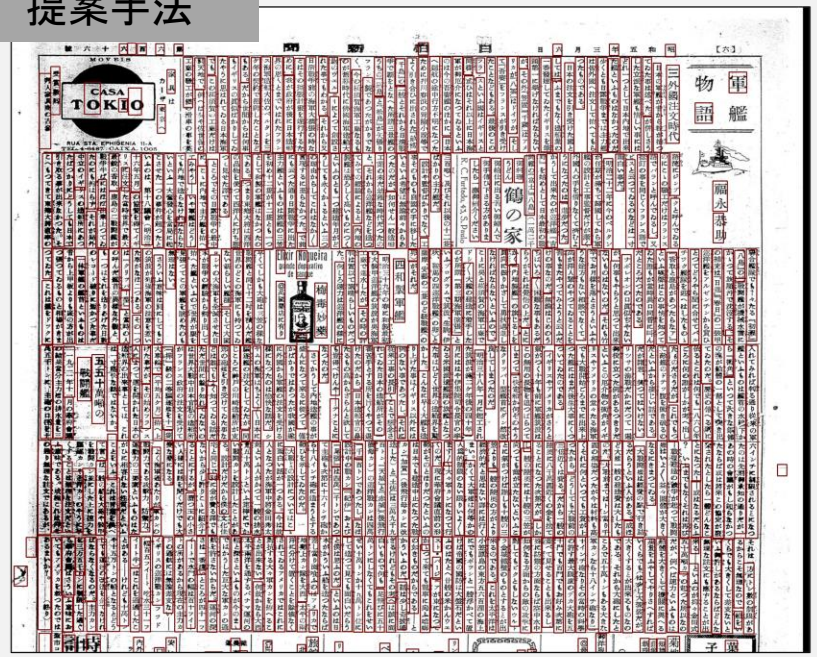
レイアウト解析手法　検証

■ 提案手法とNDLOCRのレイアウト解析精度比較による検証

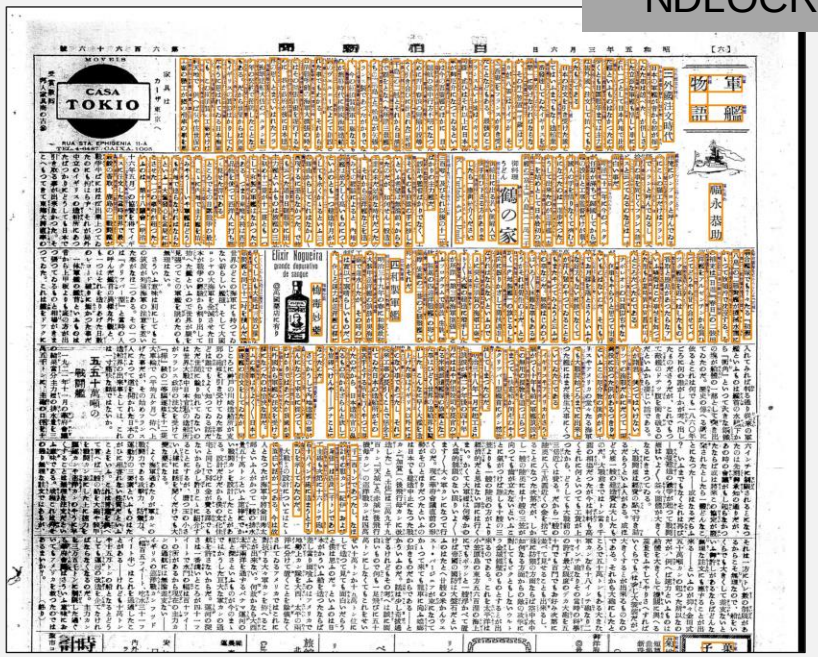
	画像A	画像B	画像C	画像D	画像E	画像F	画像G	平均
提案手法	95.81%	88.26%	92.52%	87.32%	79.89%	97.69%	71.73%	87.60%
NDLOCR	96.34%	89.39%	71.02%	49.26%	57.22%	98.46%	29.62%	70.19%

大幅な精度向上

提案手法



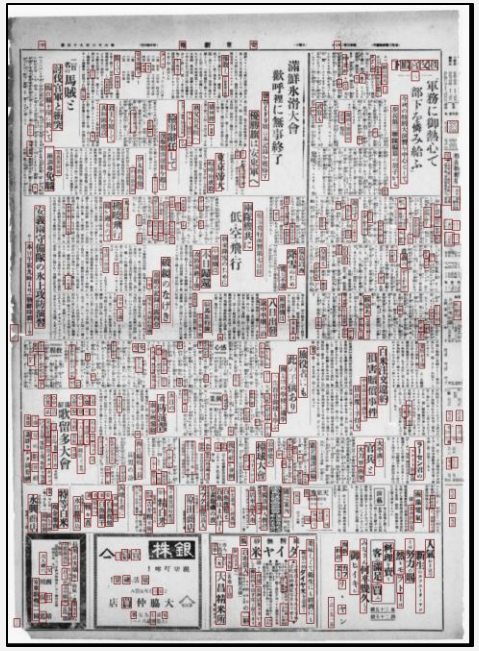
NDLOCR



レイアウト解析手法　考察

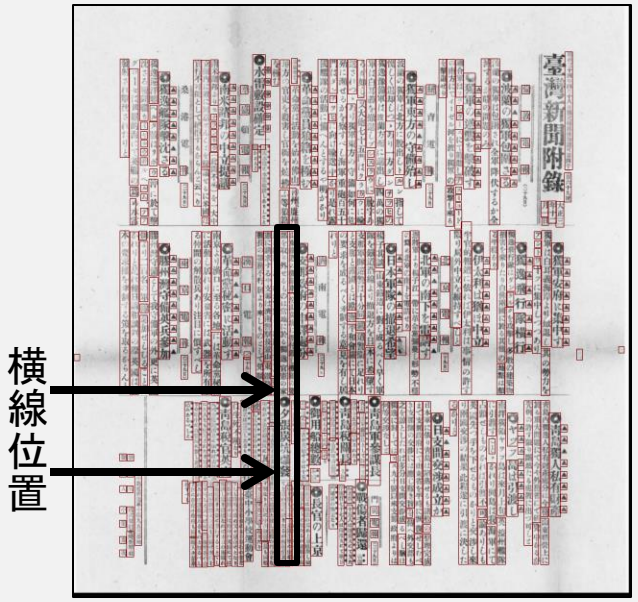
認識誤りについて

解像度の低い画像



全体の認識精度が低い

段組みをまたいで1行認識



横線が除去できておらず
横線を文字として認識してしまう

改善点

- 解像度の低い画像への対応
- CRAFTの後処理手法の改善
- 段組みを区切る横線除去方法

ゆがみ，かすれによる途切れのある線に対応する

ルビ除去手法　検証

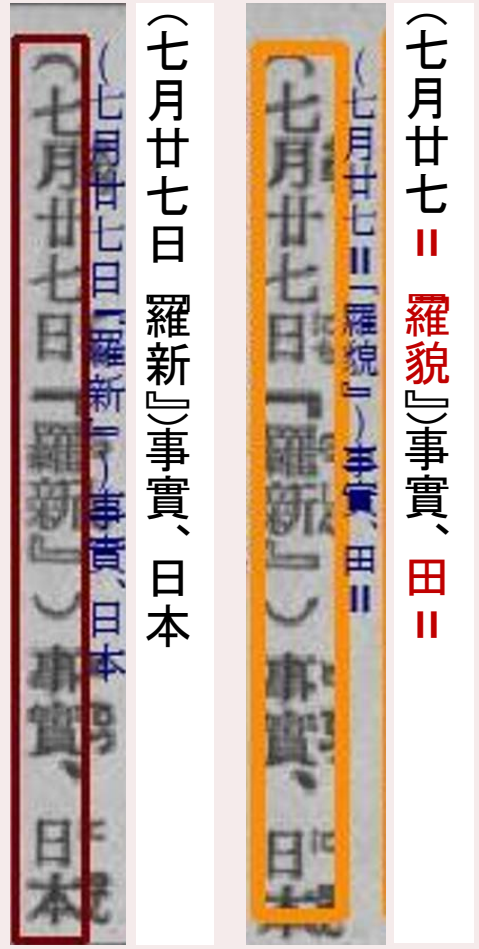
■ 文字認識精度の比較による検証

- レイアウト解析精度が高精度の画像を利用

	画像A	画像B	画像F
提案手法	60.10%	45.57%	73.23%
NDLOCR	42.30%	26.25%	34.77%

いずれの画像も文字認識精度が向上

ルビの除去により漢字の認識誤りが改善



提案手法　NDLOCR

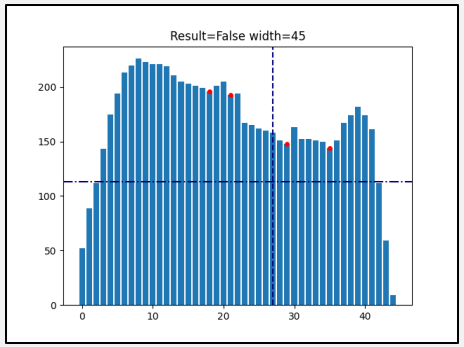
ルビ除去手法　考察

ルビ除去　失敗例

分離する要素を持つ ひらがな行

漢字を多く含む行

濃度の差が出にくい



濃度ヒストグラム

改善点

ルビの有無判定誤り

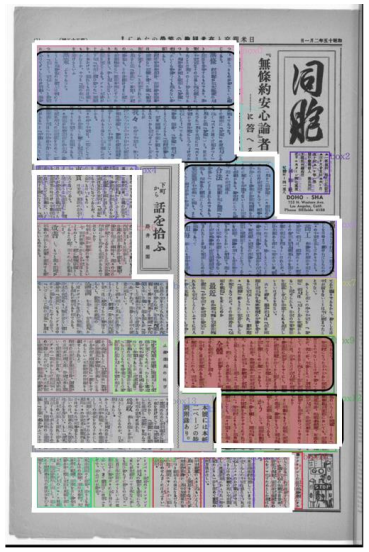
- ルビの無い行に対する誤りは，文字認識精度低下の原因となる
- ルビ有無判定手法の改善が必要

濃度ヒストグラム以外の判定方法を検討

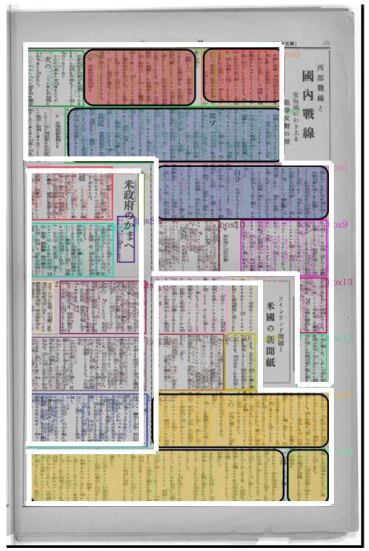
読み順検出手法　検証

■ 同一記事推定精度による検証

- レイアウト解析，文字認識の出力を利用する
- ノイズの少ない鮮明な画像を用意



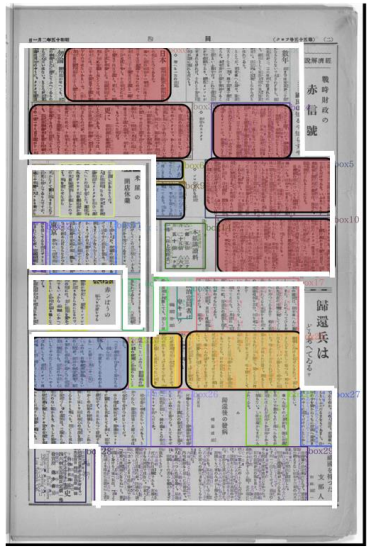
画像H



画像I

画像J

画像K



画像L

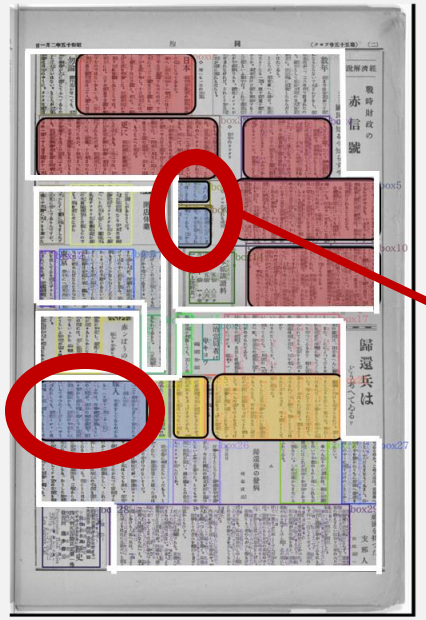
結果

- 記事の一部分を抽出可能
- 記事すべてを網羅することは不可
- 画像J, K, Lでは誤りが見られた

読み順検出手法　考察

同一記事推定　誤推定例

DL値　少ない文章量に含まれる単語の影響力が大きくなる



小さい文章ブロックが影響

形態素解析が原因のもの

「ハツキリ」

- 近代文語体:「ハツキリ」
- 現代口語体:「はっきり」(副詞)

MeCab(形態素解析ソフト)

近代文語体に対応していない

名詞を抽出

ハツ　キリ

改善点

推定精度が不足

他手法との組み合わせを検討

テキストセグメンテーション

3 つの分類層を用いたマルチタスク学習技術

“Topic Segmentation Model Focusing on Local Context”. (https://arxiv.org/abs/2301.01935)

5 今後の課題

邦字新聞に対応したOCRシステム

実用に向けて 99.9% 以上の精度を目指す

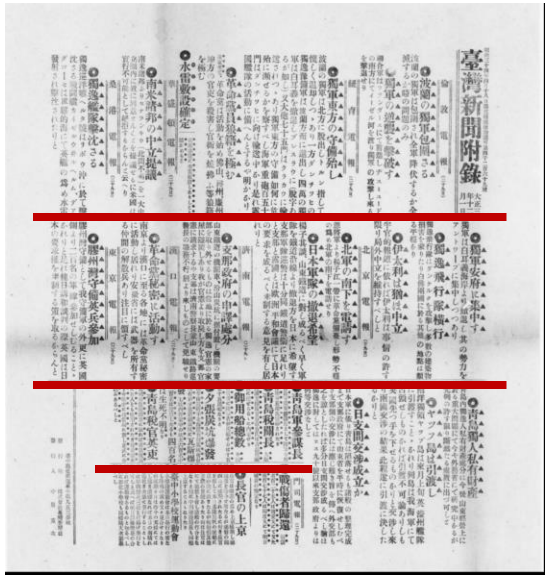
レイアウト解析手法の精度向上

- OCRシステムの全体精度に関わる
- 99.9%以上の精度を目指す

前処理手法の改善

段組みを区切る横線の検出方法

ルビ除去：ルビ有無判定の改善



読み順検出の改善

- 近代文語体ー現代口語体
  相互翻訳機能の実現

読み順通りの出力が必須

同一記事推定の改善

他手法の検討

研究業績

国際研究集会（口頭発表のみ）

- 熊谷 もも，邦字新聞OCRの概要と設置，スタンフォード大学上田研究室セミナー，スタンフォード大学フーバー研究所，2023年10月．

国内学会（口頭・査読無）

- 熊谷 もも，古磯 則江，髙田 雅美，上田 薫，城 和貴，多段組みで構成される近代書籍の読み順推定手法の検討，研究報告数理モデル化と問題解決（MPS），Vol.2023-MPS-143，No.21，pp.1-6，2023年6月．