令和3年度
OCR処理プログラム研究開発作業
報告資料

資料作成：株式会社モルフォAIソリューションズ

事業の概要

本事業の目的・実施概要

本事業の目的

国会図書館デジタルアーカイブ画像のOCR処理を活用したテキストデータの作成による、本文検索サービスや視覚障害者等へのデジタル化資料の提供の利便性向上

実施概要

貴館の所蔵するデジタル化資料の画像を利用して、所定の性能を満たすOCR処理プログラムの研究開発として以下二点を実施

① OCR学習用データセット及びレイアウト情報データセット作成

- 1870年代～1960年代　理系図書・文系図書
- 1870年代～1990年代　雑誌
  計17,375件を作成

② OCR処理プログラムの開発

- 本文テキスト領域の文字の認識精度向上及び、画像のテキスト化処理速度の向上

本事業のアプローチ

■課題

貴館の所蔵資料は膨大であるため、開発するOCR処理プログラムには高い認識精度と高速な処理速度の両立が求められる。一方で従来手法には実用面の課題がある。例えば商用の日本語OCR処理プログラムは現代の資料を対象としているため、近代書籍に特有の複雑なレイアウトや文字配置、字形、旧字旧仮名遣いに十分に対応しておらず精度面で課題がある。また、近年、OCR処理において深層学習等の先端的な機械学習技術を取り入れた事例が注目を高めている。ただし、これらの多くは研究段階や試行段階のものであるため、精度が高い場合であっても処理速度が十分でない等の傾向がみられる。

上記を踏まえ、当社及び当社再委託先は本事業の目的に適い実用化に資するOCR処理プログラム実現のために以下のアプローチを採用する。

① OCR学習用データセット及びレイアウト情報データセットの作成

良質な学習データの構築

精度の高いOCR処理プログラム構築のためには良質な学習用データセットの構築が不可欠である。学習用データセットの構築を担う凸版印刷は、明治期以降の近代書籍に関するOCR処理の実績、及び字形データセット構築の実績を豊富に有しています。そのため過去の知見を活かし、貴館が保有する書籍画像データの特徴を踏まえ、複雑なレイアウトや特殊な文字配置、字形といった近代書籍の特徴を網羅した良質なデータセットの構築を目指す。

② OCR処理プログラムの開発

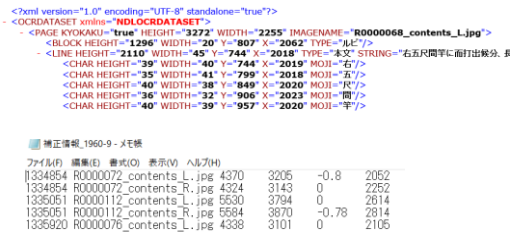
精度面と処理速度の面を両立した機械学習アルゴリズムの構築

一般に深層学習をはじめとする機械学習のアルゴリズムは、GPU等の豊富な計算リソースを前提に処理速度よりも精度面での工夫を行うものが多い傾向にある。一方で、モルフォ及びモルフォAIソリューションズでは、これまでスマートフォンや車載カメラといった計算資源の限られた端末での機械学習・画像処理プログラムの開発を多数行ってきた。したがって、モルフォが保有する技術・ノウハウを活かすことにより精度のみならず処理速度でも実用に足るOCR処理プログラムの開発を目指す。

実施内容と成果

OCR学習用データセットの実施概要

	元画像データ抽出・選別 画像角度補正	要件定義	OCR学習用データセット構築
実施内容	• 貴館貸与画像データ約250万資料の集計 • データセット用の画像抽出・選別（約250万資料→2.5万件）作業 • 抽出・選別した画像の補正（複製/拡張子変更/傾き補正/トリミング等）	• 納品用データセットの詳細仕様の検討（データ構成/文字種/行矩形要素/ブロック要素/インライン要素/欧文/数式・化学式/出力形式/矩形情報の並び順）	• 要件定義で固めたOCR学習データセット仕様書に基づき、納品用のデータセットを作成 • 7/29初回納品、以降9/27までに計6回、17,375件を納品（納品済みのOCR学習用データセットについて修正対応含む） • レイアウト情報データセットは、12/21初回納品、以降22/2/7までに計4回、27,878件を納品
アウトプット	• OCR学習用データセット画像抽出手順書 • OCR学習用データセット作成用画像_画像補正手順書 • データセット構築用画像（2.5万件）	• OCR学習用データセット作成仕様書	• 納品データ一式（アノテーションデータ/xmlデータ/（画像）補正情報



OCR処理プログラムの実施概要

OCR処理プログラムの実施フロー

【Input】 貴館保有の デジタルコレクション 収載画像	• 貴館保有の画像をOCR処理に最適な形式に変換
画像の前処理	• 色修正、ノイズ除去、画像の統合、傾き補正等を実施
レイアウト抽出・ 行矩形抽出・ 文字領域抽出	• 様々なレイアウトの画像から文章の領域を抽出した上、行の矩形や読み取り順、文字領域の抽出を実施
文字認識（OCR）	• 取得した文字の識別のため以下を実施 • 正規化・特徴量抽出、テンプレートマッチングを実施
出力結果の後処理	• 文字認識で得られた候補から文字の誤読を検出・修正
【Output】 テキスト	• 認識した文字を所定のフォーマットで出力

実施項目

- OCR処理プログラムに実装する手法の検討・検証
  -分割
  -傾き補正
  -レイアウト/行認識
- （上記の検討・検証を踏まえ）実装手法の課題整理・対応
- プログラム設計（内部/外部設計）・開発※

中間報告

- 精度・速度改善
- 見出し認識・著者名対応

※　プログラム開発は10月以降も継続

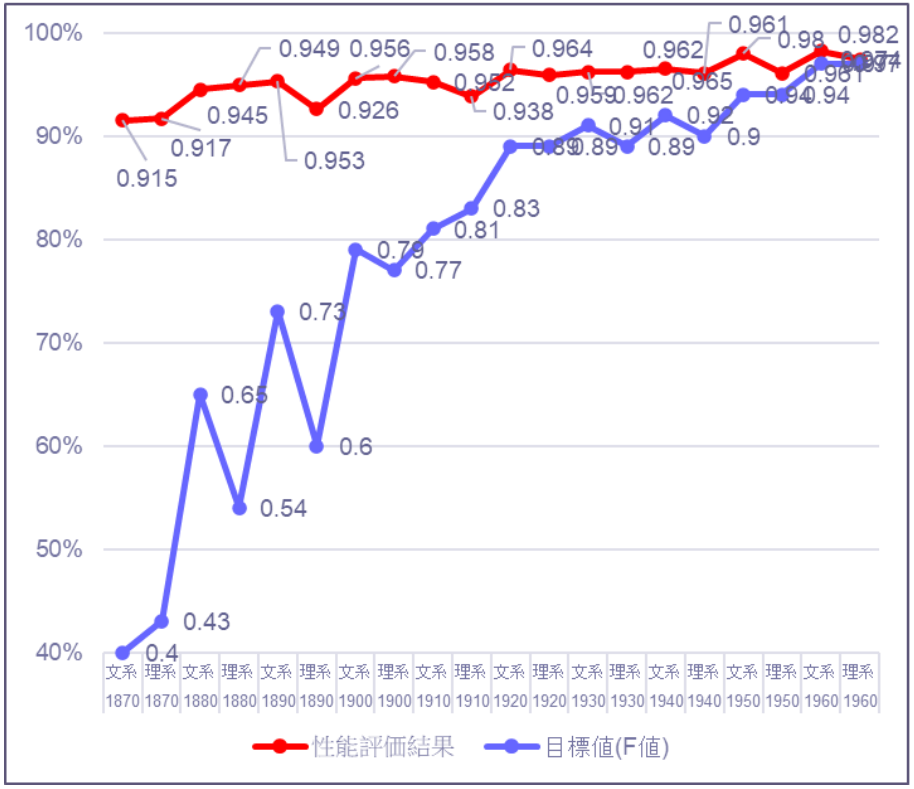
中間報告～最終報告

OCR処理プログラム認識性能の達成状況

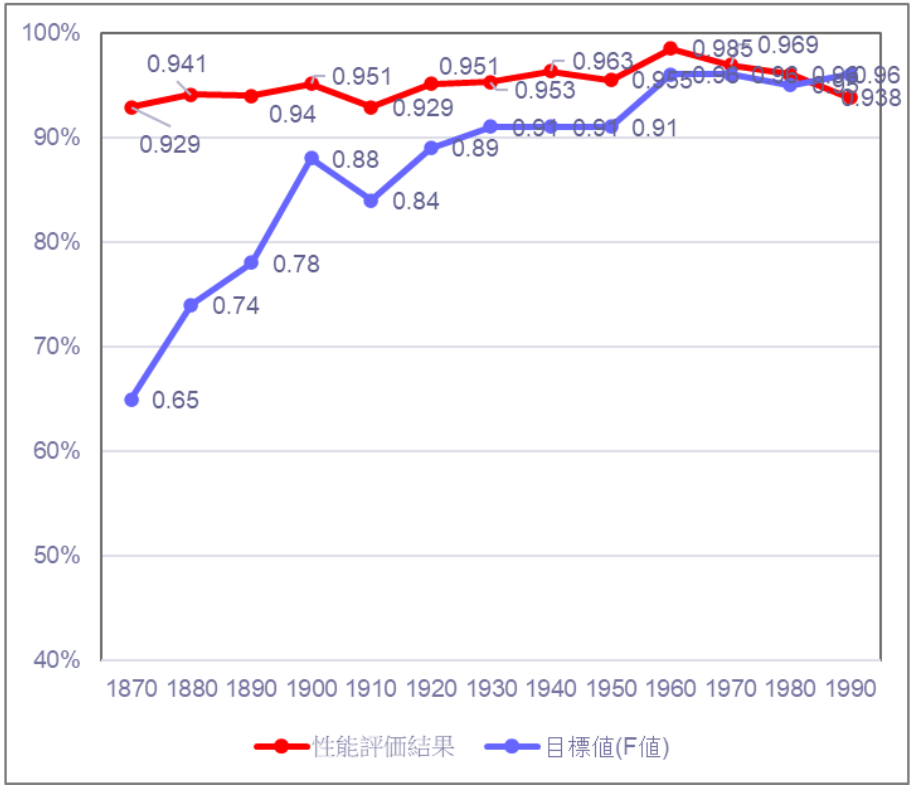
- 全納品データ17,375件のOCR処理プログラムに実装する全手法（①見開き分割、②傾き補正、③レイアウト認識、④行認識）を統合した際の認識性能は以下の通り、30個（目標値達成）/30個（評価対象、参考値3個は除く）となり、仕様書上の認識精度基準は達成することができた。
  ※各年代ごとに10件のスコアを算出した中の中央値で判定（全330件）

認識精度基準と認識性能評価結果まとめ

書籍種別・年代別の精度評価

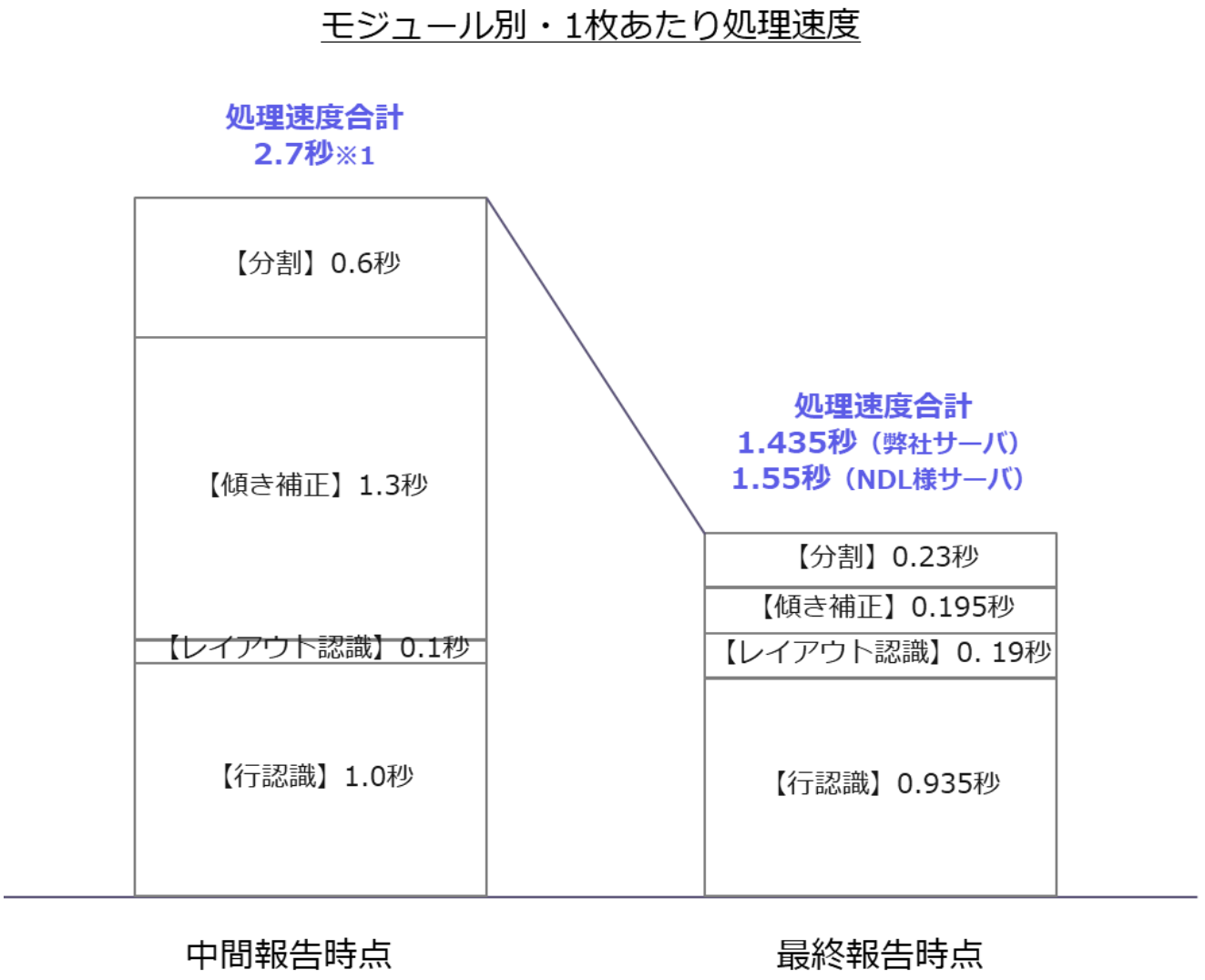


雑誌の年代別の精度評価
(1970-1990年代は参考値)



OCR処理プログラム速度性能の達成状況

- 中間報告時点から更に43%処理時間を削減し約1.55秒/枚の処理時間を達成、目標の2秒/枚を大きく上回る速度性能を達成



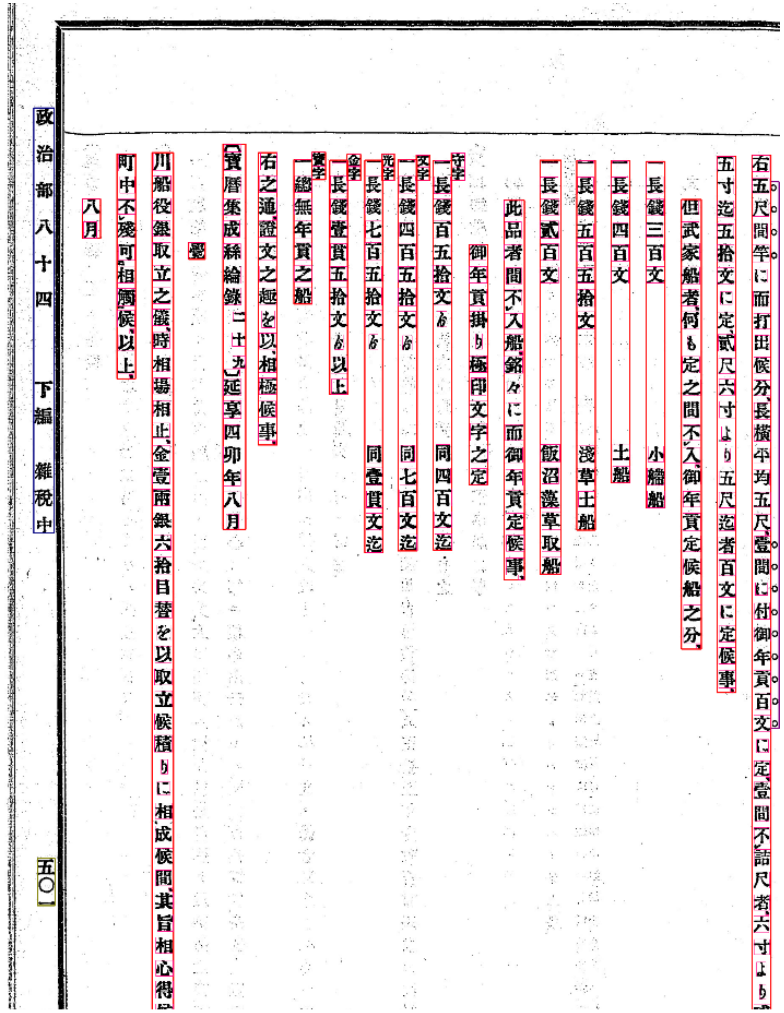
※1　貴館と弊社で動作環境で異なること、分割と傾きは入出力を含めた値

OCR学習用データセット

※データセットの作成作業は凸版印刷株式会社が担当

【ご参考】OCR学習用データセット納品物のサンプル（1910年代図書）

imgファイル



本文テキスト内外の座標情報が記載されたxmlファイル


```
<?xml version="1.0" encoding="UTF-8" standalone="true"?>
- <OCRDATASET xmlns="NDLOCRDATASET">
  - <PAGE KYOKAKU="true" HEIGHT="3272" WIDTH="2255" IMAGENAME="R0000068_contents_L.jpg">
      <BLOCK HEIGHT="1296" WIDTH="20" Y="807" X="2062" TYPE="ルビ"/>
    - <LINE HEIGHT="2110" WIDTH="45" Y="744" X="2018" TYPE="本文" STRING="右五尺間竿に而打出候分、長横平均五尺、壹間に付御年貢百文に定、壹間不詰尺者、六寸より貳尺" DIRECTION="縦"
        <CHAR HEIGHT="39" WIDTH="40" Y="744" X="2019" MOJI="右"/>
        <CHAR HEIGHT="35" WIDTH="41" Y="799" X="2018" MOJI="五"/>
        <CHAR HEIGHT="40" WIDTH="38" Y="849" X="2020" MOJI="尺"/>
        <CHAR HEIGHT="36" WIDTH="36" Y="906" X="2023" MOJI="間"/>
        <CHAR HEIGHT="40" WIDTH="39" Y="957" X="2020" MOJI="竿"/>
        <CHAR HEIGHT="31" WIDTH="29" Y="1015" X="2025" MOJI="に"/>
        <CHAR HEIGHT="37" WIDTH="37" Y="1066" X="2021" MOJI="而"/>
        <CHAR HEIGHT="37" WIDTH="39" Y="1119" X="2022" MOJI="打"/>
        <CHAR HEIGHT="35" WIDTH="34" Y="1171" X="2025" MOJI="出"/>
        <CHAR HEIGHT="38" WIDTH="39" Y="1223" X="2022" MOJI="候"/>
        <CHAR HEIGHT="37" WIDTH="41" Y="1277" X="2022" MOJI="分"/>
        <CHAR HEIGHT="9" WIDTH="8" Y="1318" X="2053" MOJI="、"/>
        <CHAR HEIGHT="39" WIDTH="38" Y="1331" X="2023" MOJI="長"/>
        <CHAR HEIGHT="40" WIDTH="40" Y="1382" X="2023" MOJI="横"/>
        <CHAR HEIGHT="34" WIDTH="37" Y="1439" X="2022" MOJI="平"/>
        <CHAR HEIGHT="37" WIDTH="38" Y="1490" X="2022" MOJI="均"/>
        <CHAR HEIGHT="35" WIDTH="40" Y="1543" X="2022" MOJI="五"/>
        <CHAR HEIGHT="38" WIDTH="40" Y="1596" X="2022" MOJI="尺"/>
        <CHAR HEIGHT="9" WIDTH="9" Y="1636" X="2049" MOJI="、"/>
        <CHAR HEIGHT="40" WIDTH="37" Y="1647" X="2024" MOJI="壹"/>
        <CHAR HEIGHT="38" WIDTH="34" Y="1701" X="2025" MOJI="間"/>
        <CHAR HEIGHT="30" WIDTH="29" Y="1757" X="2026" MOJI="に"/>
        <CHAR HEIGHT="38" WIDTH="36" Y="1806" X="2024" MOJI="付"/>
        <CHAR HEIGHT="34" WIDTH="39" Y="1863" X="2023" MOJI="御"/>
        <CHAR HEIGHT="39" WIDTH="39" Y="1915" X="2022" MOJI="年"/>
        <CHAR HEIGHT="40" WIDTH="40" Y="1966" X="2022" MOJI="貢"/>
        <CHAR HEIGHT="39" WIDTH="37" Y="2019" X="2025" MOJI="百"/>
        <CHAR HEIGHT="38" WIDTH="40" Y="2072" X="2022" MOJI="文"/>
        <CHAR HEIGHT="31" WIDTH="28" Y="2129" X="2029" MOJI="に"/>
        <CHAR HEIGHT="36" WIDTH="37" Y="2180" X="2023" MOJI="定"/>
        <CHAR HEIGHT="9" WIDTH="9" Y="2219" X="2051" MOJI="、"/>
```


納品対象画像の補正情報

補正情報_1910-0 - メモ帳

ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

PID	画像名	回転補正後横幅	回転補正後後高さ		補正角度	トリミング矩形X座標
1766303	R0000068_contents_L.jpg	5850	3996	-0.19	2863	
1766303	R0000068_contents_R.jpg	5856	4004	-0.28	3063	
1905974	R0000065_contents_L.jpg	5116	3786	-0.21	2387	
1905974	R0000065_contents_R.jpg	5110	3780	0.13	2587	
897690	R0000068_contents_L.jpg	4350	3272	0.11	2055	
897690	R0000068_contents_R.jpg	4358	3282	0.23	2255	

画像選別基準

画像選別については、貴館の内容確認の上、以下の抽出及び除外対象とした。

[抽出対象画像]

- テキストと図版入り画像
- テキストのみ画像
- 図版のみ画像※キャプション付き
- 表組入り画像
- 全面表組
- 版本

[除外画像]

- 表紙、扉
- 画質の悪い画像（解像度、極端な傾き、ピンボケ、折り目、汚れ、虫食い、ノドの開き不足）
- 図版のみの画像※キャプションなし
- 全面広告
- グラビア（＋グラビア内テキスト）画像
- ブランクページ
- 片観音、両観音、Z折り、特殊折り
- 全面外国語
- 楽譜
- マンガの吹き出し（雑誌・図書内のマンガ記事を含む）
- 新聞の縮刷版
- 目次

OCR学習用データセットの納品実績

貴館と合意した分割単位で均等に画像を抽出、企画提案時に考慮していなかった理系/文系が不明な図書は、全データを考慮して330資料を抽出するという方針に基づき、各年代の納品数量バランスを取りつつ、最終的に17,375件の納品を9/27までに完了させた。

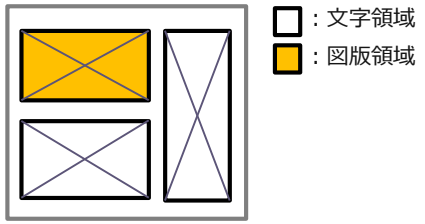
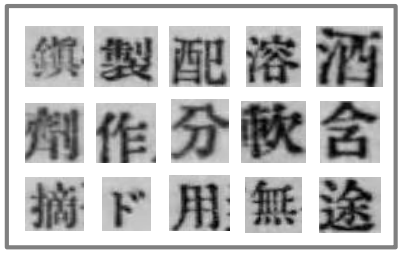
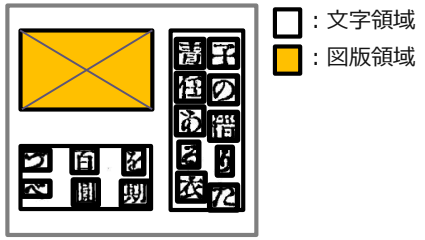
OCR学習用データセットの納品数量及び内訳（全体サマリー）

No	年代	7/29納品	7/30納品	8/16納品	8/30納品	9/13納品	9/27納品	合計
		実績	実績	実績	実績	実績	実績	実績
1	新字（1960年以降の理系図書・文系図書・雑誌）	521	614	1,007	1,173	913	1,570	5,798
2	旧字① （1910～1959年の理系図書・文系図書・雑誌）	534	520	1,112	971	786	2,166	6,089
3	旧字② （1909年以前の理系図書・文系図書・雑誌）	516	546	926	1,047	1,314	1,139	5,488
合計		1,571	1,680	3,045	3,191	3,013	4,875	17,375

OCR学習用データセット構成（１/2）

本事業におけるOCR学習用データセットは、様々なAI-OCRの実装に対応するため、領域に関する情報と文字に関する情報はまとめて1データでOCR学習用データセットを作成する。

■領域情報データセット、文字情報データセット、および両者を統合したデータセットとしては、それぞれ以下の例があげられる。

領域情報データセット	• NDL-DocLデータセット (https://github.com/ndl-lab/layout-dataset) • PRImA (https://www.primaresearch.org/dataset/)		従来のOCR学習用データセット
文字情報データセット	• MNIST (手書き数字データセット) (http://yann.lecun.com/exdb/mnist/) • スタンフォードOCR (手書き英数字データセット) (http://ai.stanford.edu/~btaskar/ocr/) • ETL文字データベース (手書き・活字 和文英数字データセット) (http://etlcdb.db.aist.go.jp/)		
領域情報・文字情報統合データセット	• MTHv2 (中国漢字データセット) (https://github.com/HCIILAB/MTHv2_Datasets_Release)		領域情報と文字情報を統合したOCR学習用データセット

本事業でのOCR学習用データセット構成

OCR学習用データセット構成（2/2）

OCR学習用データセットに含まれる情報の一覧を示す。本文テキスト内の行矩形情報、文字矩形情報に加えて、本文外の画像情報、基本的な文字種以外のブロック要素（図版、欧文、数式、化学式）についても以下の通り定める。

OCR学習用データセットに含まれる情報		
本文テキスト内	（本文内）行矩形情報	• 座標（X,Y） • 高さ（H） • 横幅（W） • 書字方向（縦、横、右から左） • 行内文字列情報（テキスト）
本文テキスト内	（本文内）文字矩形情報	• 座標（X,Y） • 高さ（H） • 横幅（W） • 文字情報（文字コード）
本文テキスト外	画像情報	• 画像名 • 画像高さ（H） • 画像横幅（W） • 匡郭有無 • 傾き修正情報（回転方向・角度）※ • トリミング情報（左右・切断位置）※ • その他
本文テキスト外	文字以外の領域情報	• 座標（X,Y） • 高さ（H） • 横幅（W） • 種類（図版、欧文、数式、化学式など）

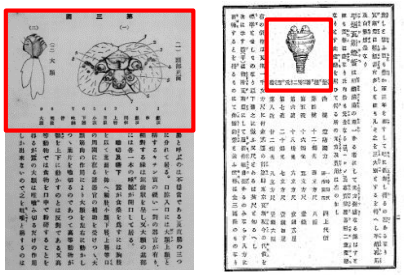
※画像補正作業で作成するTSVファイルに記載

文字以外の領域情報

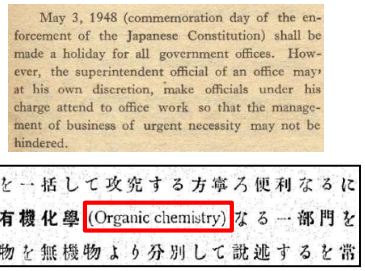
- 図版、数式、化学式は、本事業の目的である本文検索用テキスト作成においてテキスト認識の阻害要因であり、精度向上のためには文字認識の対象外として排除することが望ましい。
  本件OCR学習用データセットでは、これらの要素については文字以外の領域情報として作成し、将来的にOCR処理プログラムの機能として排除処理を実装する際の学習用データとしての活用を想定する。
- 欧文は領域情報として、ブロック要素およびインライン要素でOCR学習用データセットに文字以外の領域情報として作成する。

※欧文は既存OCRソフトの認識精度が高いため、本事業の目的を鑑みて、本件OCR処理プログラムでは欧文領域の認識のみ実行し、別途既存OCRソフトで文字認識を実施する、という実装を想定する。
欧文のOCR学習用データセットは既存のオープンデータが存在するため、本提案では新規に作成せず、既存オープンデータの使用を想定する。

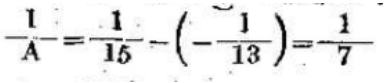
[図版]（ブロック要素）



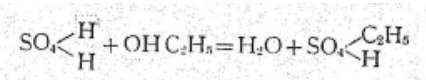
[欧文]（ブロック要素・インライン要素）



[数式]（ブロック要素・インライン要素）



[化学式]（ブロック要素・インライン要素）

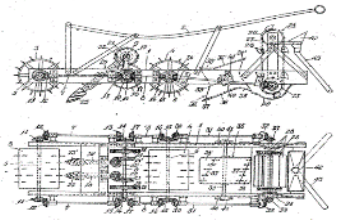


OCR学習用データセット レイアウト情報ーブロック要素

- BLOCKの種類 (TYPE) は「図版」「表組」「柱」「ノンブル」「ルビ」「系統図」「数式」「化学式」「欧文ブロック」

BLOCK

図版



表組

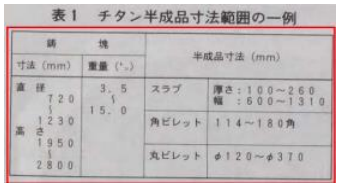


表1　チタン半成品寸法範囲の一例

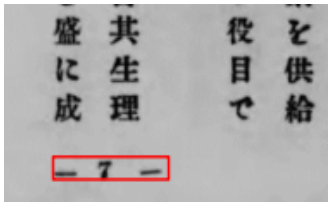
鋳塊		半成品寸法 (mm)	
寸法 (mm)	重量 (t)		
直径 720 / 1230	3.5 / 15.0	スラブ	厚さ：100～260 幅：600～1310
高さ 1950 / 2800		角ビレット	114～180角
		丸ビレット	φ120～φ370

柱

数式

$$\frac{1}{A}=\frac{1}{15}-\left(-\frac{1}{13}\right)=\frac{1}{7}$$

ノンブル



ルビ

系統図

化学式

$$SO_4\langle\begin{matrix}H\\H\end{matrix}+OHC_2H_5=H_2O+SO_4\langle\begin{matrix}C_2H_5\\H\end{matrix}$$

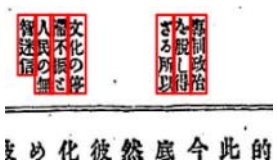
OCR学習用データセット レイアウト情報ーライン要素・インライン要素

- LINEの種類（TYPE）は、「キャプション」「頭注」「割注」「見出し」「著者名」「本文」
- INLINEの種類（TYPE）、「手書き」「縦中横」 「数式」「化学式」「欧文」「色付文字」「回転欧文」

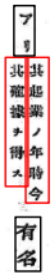
LINE

キャプション

頭注

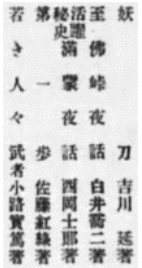


割注

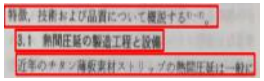


見出し

著者名



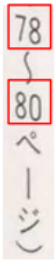
本文



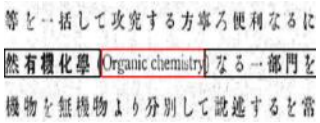
INLINE

手書き

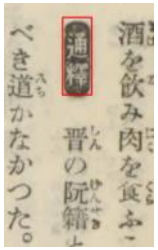
縦中横



欧文



色付き文字



回転欧文

数式

$$\frac{1}{A}=\frac{1}{15}-\left(-\frac{1}{13}\right)=\frac{1}{7}$$

化学式

$$SO_4<\begin{matrix}H\\H\end{matrix}+OHC_2H_5=H_2O+SO_4<\begin{matrix}C_2H_5\\H\end{matrix}$$

OCR学習用データセット 文字種

OCR学習用データセットの文字種

基本的な文字種

- ひらがな
- カタカナ
- 数字
- JIS第一水準漢字・JIS第二水準漢字
- 下記の記号
  - 半角記号　,.- /()
  - 句読点　、。
  - 括弧　() 〔〕 『』 【】 「」
  - 丸付き文字　①②③④⑤⑥⑦⑧⑨⑩⑪⑫⑬⑭
㊀㊁㊂㊃㊄㊅㊆㊇㊈㊉
❶❷❸❹❺❻❼❽❾
  - 丸付き漢字　㊞
  - 括弧付き文字　㈱㈲㈾㈲
  - 繰り返し記号　ヽヾ〻ゝゞヽヾ
  - 図形　○●□◆▲△▽
  - その他の記号　― ・ ※↓ →↑←⇩⇨⇧⇦ ？～ =〆…

凸版印刷による事前調査で出現率3,000位までのJIS第一水準漢字・JIS第二水準漢字に包摂が可能なJIS第二水準外の漢字

欧文・ギリシア文字

- 本文中に出現する3文字程度までの欧文・ギリシア文字
  - 半角アルファベット52文字（U+0041-U+005A、U+0061-U+007A）
  - ギリシア文字48文字（U+0391-U+03A9、U+03B1-U+03C9）

留意事項

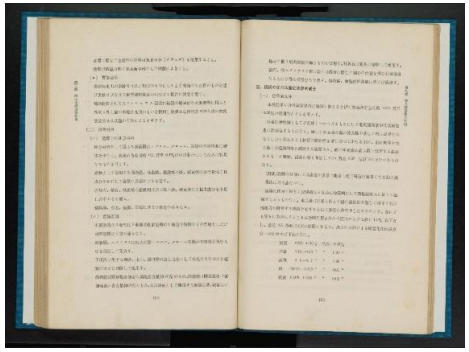
- 欧文・ギリシア文字は、十分な領域がありレイアウト認識ができるものについては、矩形情報のみのOCR学習用データを作成し、文字情報は入力対象外とする。
- 半角文字が存在する(ASCIIコード内)の英数字記号は、別紙2「性能評価対象とする資料の内訳及び判定基準」の仕様を踏まえ、すべて半角のコードポイントを割り当てる。
- 合略仮名や結合文字、特殊な丸付文字など、コードポイントがないため１文字で入力不可能な字形は本件OCR学習用データ作成の対象外とする。
- 訓点、図版内文字、表組内文字は文字種に関わらずOCR学習用データ作成の対象外とする。
- 変体仮名は対応する現代仮名遣いのコードポイントを割り当てる。
漢字と変体仮名どちらとも取れる文字は漢字のコードポイントを割り当てる。
- 行頭字下げ、文中スペースは学習用データ作成の対象外とする。
- 包摂不可能なJIS第二水準外文字、および汚れやかすれなどで文字が判読不能な文字は「**二**」を割り当てる。

OCR処理プログラム

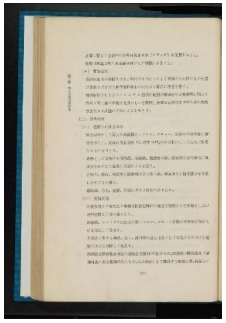
実装した手法全体の処理フロー概要

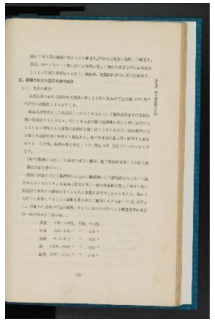
実装した手法全体の処理フローは以下の通りである。

入力画像

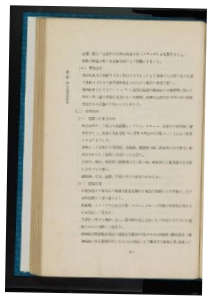


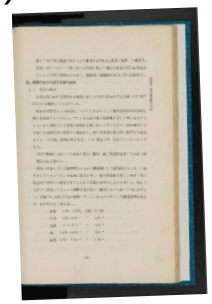
①見開き分割(分割モジュール)



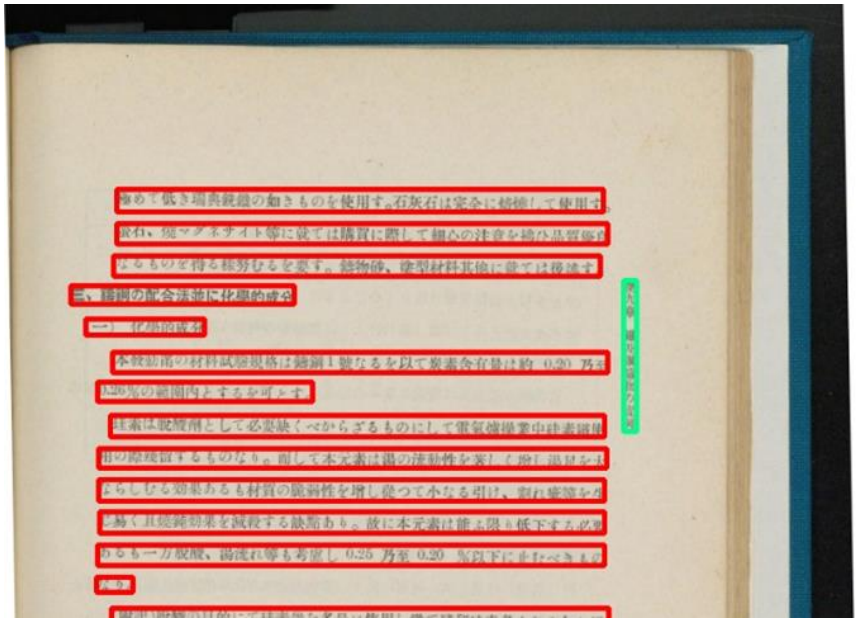


②傾き補正(傾き補正モジュール)

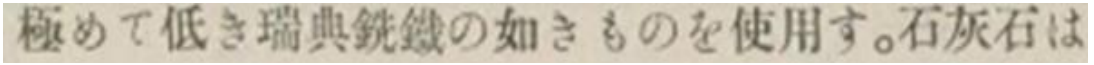




③レイアウト認識(レイアウト認識モジュール)



④行認識(行認識モジュール)



極めて低き瑞典銑鐵の如きものを使用す。石灰石は…

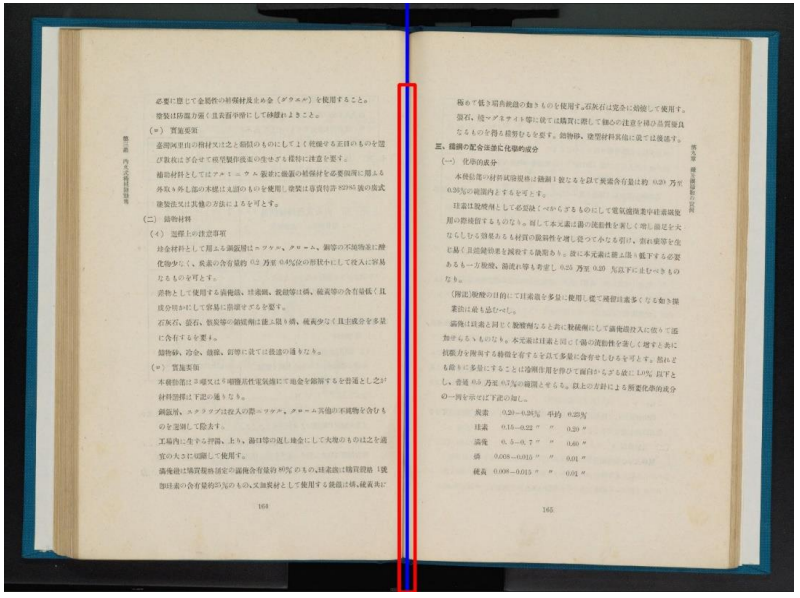
実装した各手法について
中間報告までの振り返り

見開き分割

アプローチ

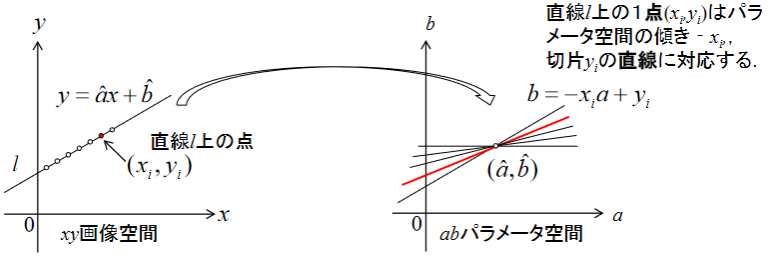
- NDLラボで開発されたノド元分割プログラムを採用
  (https://github.com/ndl-lab/ssd_keras)
- SSDを用いたノド元のx座標検出器
- 貴館よりモデルを提供いただき利用

手法のイメージ



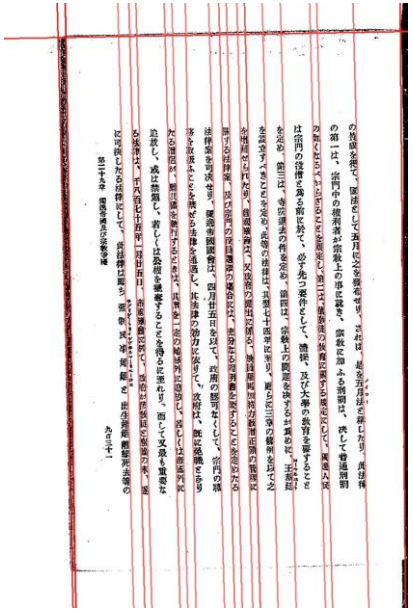
傾き補正

- Hough変換を用いた画像内の直線要素の検出
- 検出された直線の角度のうち最頻値をページの傾きとして採用
- Hough変換のイメージ

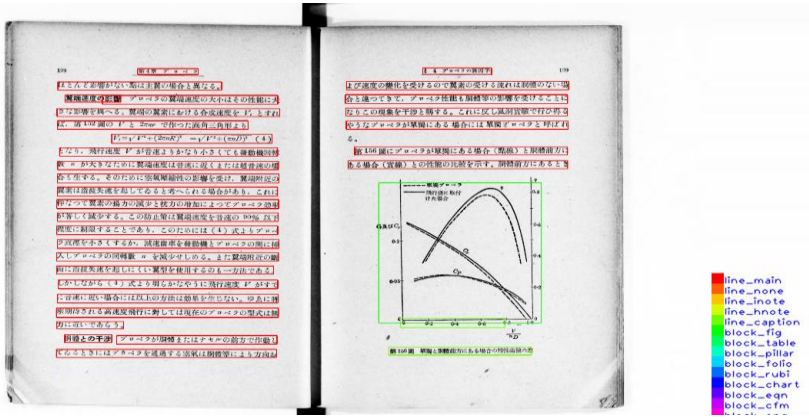
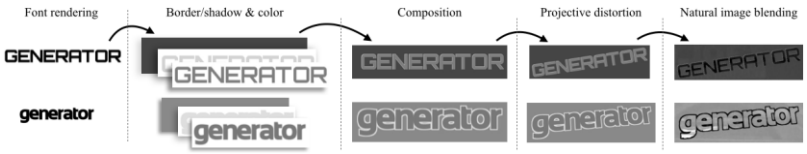


直線l上の点のabパラメータ空間への写像

http://www.cfme.chiba-u.jp/~haneishi/class/digitalgazo/6HoughTransform.pdf



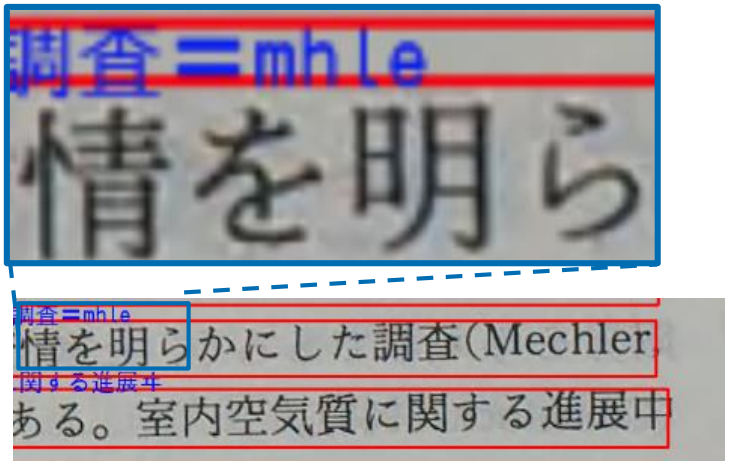
実装した各手法について
中間報告までの振り返り

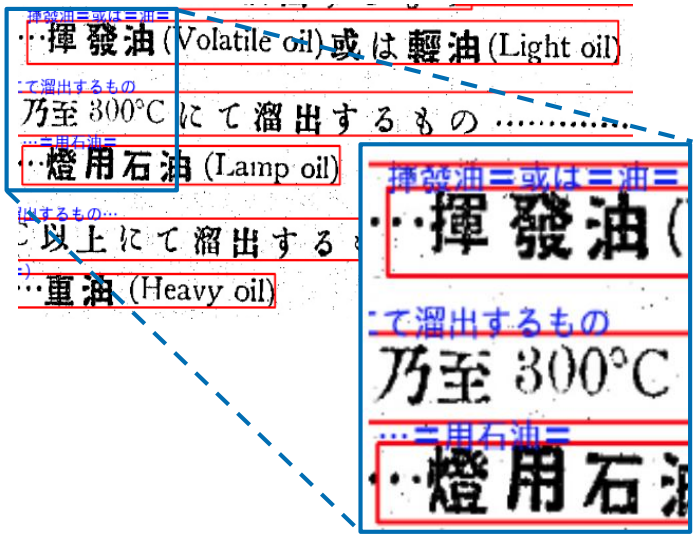
	アプローチ	手法のイメージ
レイアウト認識	• Cascade Mask RCNNのDetection部分を採用 以下のアプローチを参照 - インドの古文書行認識 (Indiscapes) (https://arxiv.org/abs/1912.07025) - 2020年SIGNATEコンペ (3rd, 4th) (https://signate.jp/competitions/218) • レイアウト認識特有のチューニング - 画像に対して対象が小さい → 入力解像度を大きく - 対象の数が多い → Region Proposal(途中段階の候補領域提示)の上限増 - アスペクト比が極端 (縦長、横長) → アンカーボックスのサイズ調整	
行認識	• シーンテキスト認識手法Rosettaを採用 (https://arxiv.org/abs/1910.05085) 以下のシーンテキスト認識手法の比較研究を参照 (https://arxiv.org/abs/1904.01906) • backboneにResNet, decoderにCTCを利用 • 合成画像英単語データセットMJを利用 • 学習時の各種Augmentation アスペクト比変更 パディング追加 角度ずれ	MJサンプル  

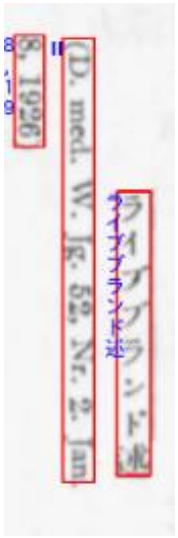
中間報告時点での課題

1．特に理系文書における欧文、数値、記号などの認識性能改善

英文字、数値の認識欠落のサンプル

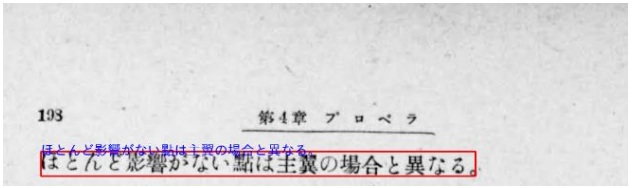


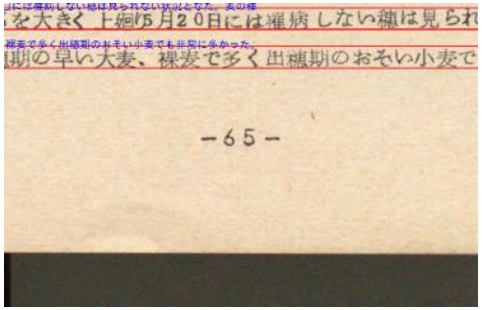


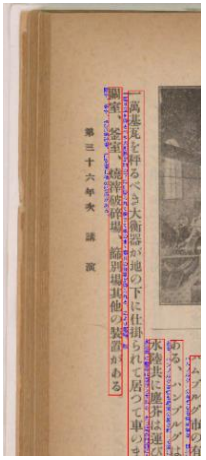


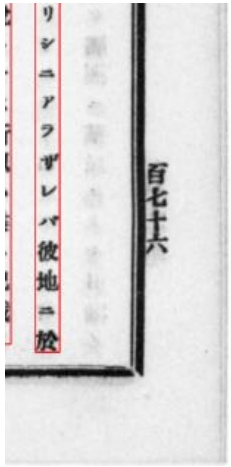
2．レイアウト、文字間違いの認識性能改善

タイトル、ページ数の認識欠落のサンプル（本文ではない柱やノンブルは認識対象外としていた）





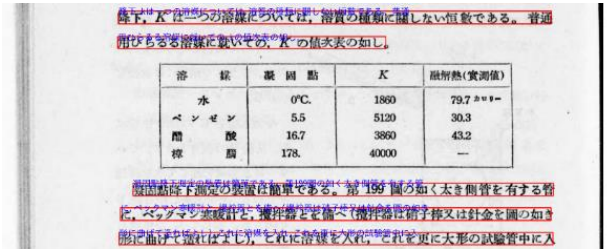




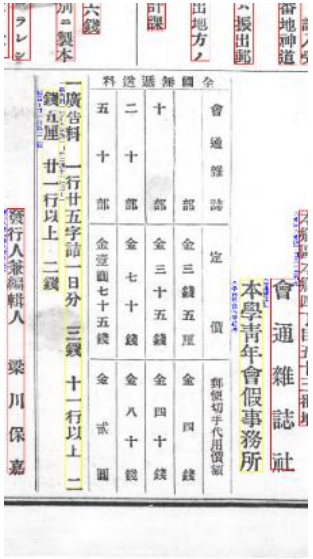
中間報告時点での課題

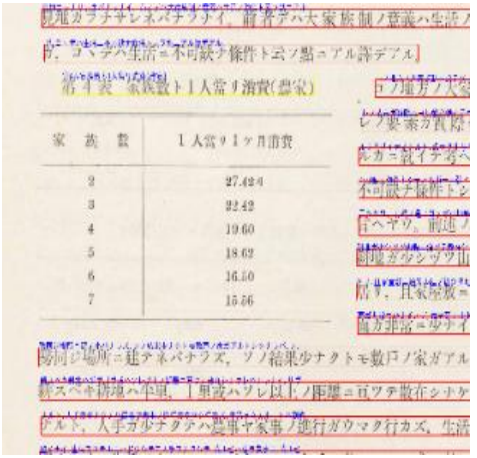
2．レイアウト、文字間違いの認識性能改善（前ページの続き）

表内文字の認識欠落のサンプル（表内文字は認識対象外としている）



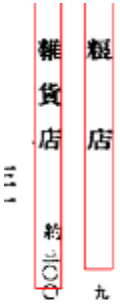
溶媒	凝固點	K	融解熱(實測値)
水	0°C.	1860	79.7 カロリー
ベンゼン	5.5	5120	30.3
醋酸	16.7	3860	43.2
樟腦	178.	40000	—

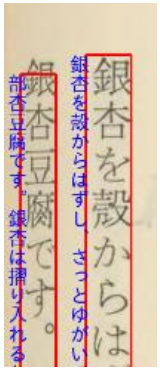


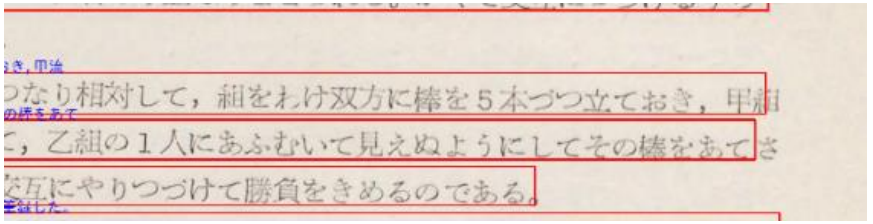


家族數	1人當り1ヶ月消費
2	27.42
3	22.42
4	19.60
5	18.62
6	16.10
7	15.56

枠の認識欠落（枠が足りていない）のサンプル

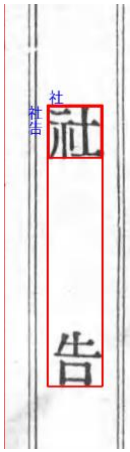




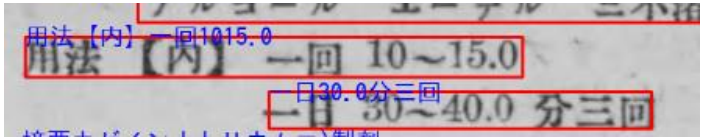


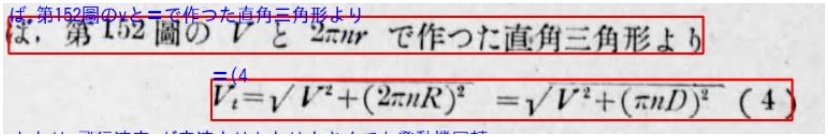
中間報告時点での課題

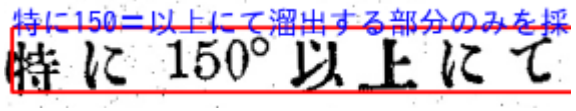
枠の重複認識のサンプル

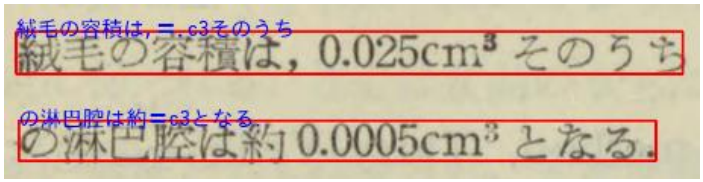


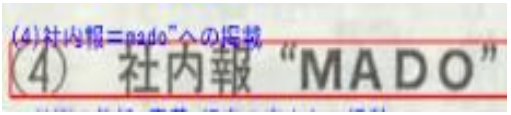
数字・記号間違いのサンプル











中間報告時点の課題に対する取り組み

1. 性能改善

1）特に理系文書での欧文、数値、記号などの認識性能改善

→ アノテーションで〓になっているインライン欧文に高精度英語OCRでアノテーション

2）レイアウトの縦書き横書き混在の誤認、行の重複の誤認、行の長さの不足

→ ルールベースの後処理で効果を検証、(学習データの増加でも改善)

3）行認識での〓の取り扱い

→ 頻出するアノテーションが〓になる文字があればデータセットの修正を検討する想定だったが今回は実施せず

→ 行認識で〓が第一候補として認識された場合に第二候補以降を採用するようなオプションを追加

4）柱・ノンブル対応

→ 柱とノンブルがテキスト化の対象外だったが、オプションで対象/非対象を切り替えられるよう対応

2. 速度改善

1）個別の手法ごとに現状分割0.6s、傾き補正2.6s、レイアウト認識0.2s、行認識約2.0s、合計5.4s (1ページ辺り2.7s)

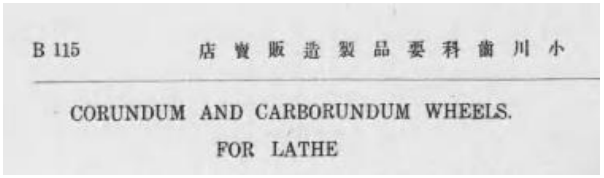
速度要件は見開き左右の1ページで2s以内

→ 特に処理時間のかかっている傾き補正の改善を実施

1．性能改善
1）理系文書における欧文、数値、記号などの認識性能改善

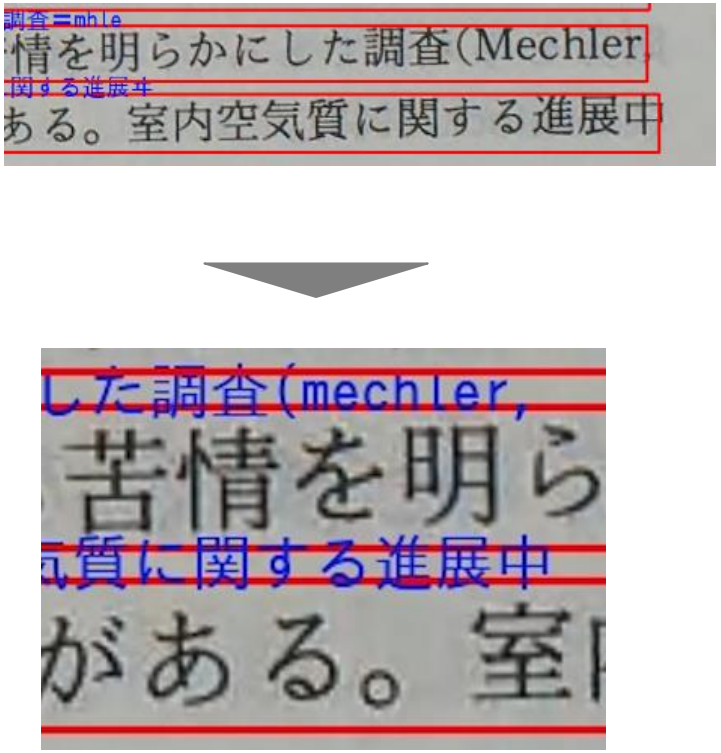
■ アノテーションで〓になっているインライン欧文に対して高性能な英語OCRでアノテーション実施
TrOCR：Transformerを利用した高精度なOCRモデル
精度変化：0.818 → 0.825（欧文が含まれる行に限定した場合 0.730 → 0.972）

欧文アノテーションサンプル



(before)
<LINE DIRECTION="横" TYPE="キャプション" STRING="〓" X="527" Y="485" WIDTH="804" HEIGHT="42">
(after)
<LINE DIRECTION="横" TYPE="キャプション" STRING="CORUNDUM AND CARBORUNDUM WHEELS." X="527" Y="485" WIDTH="804" HEIGHT="42">
(before)
<LINE DIRECTION="横" TYPE="キャプション" STRING="〓" X="788" Y="550" WIDTH="235" HEIGHT="38">
(after)
<LINE DIRECTION="横" TYPE="キャプション" STRING="FOR LATHE" X="788" Y="550" WIDTH="235" HEIGHT="38">

改善例



1．性能改善
2）レイアウトの縦書き横書きの混在、行の重複の誤認、行の長さの不足

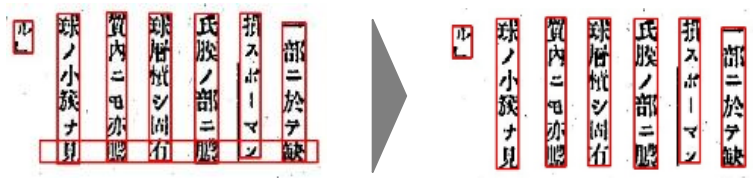
■ ルールベースの後処理を追加して効果を検証

縦横混在：横(縦)行に対して複数の縦(横)行が重なる場合は削除

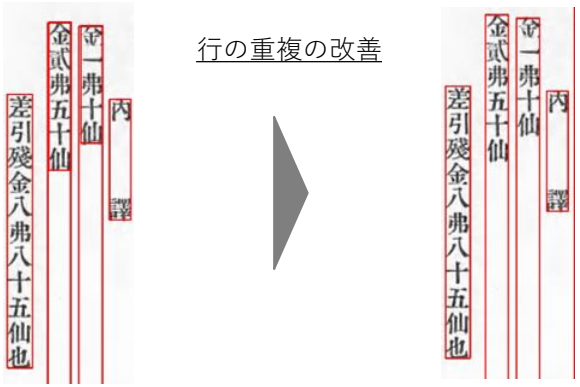
重複：すべての矩形に対して内包されているかをチェックし、内包されている矩形を削除(多少のはみ出しも許容)

長さ不足：学習データの増加で改善

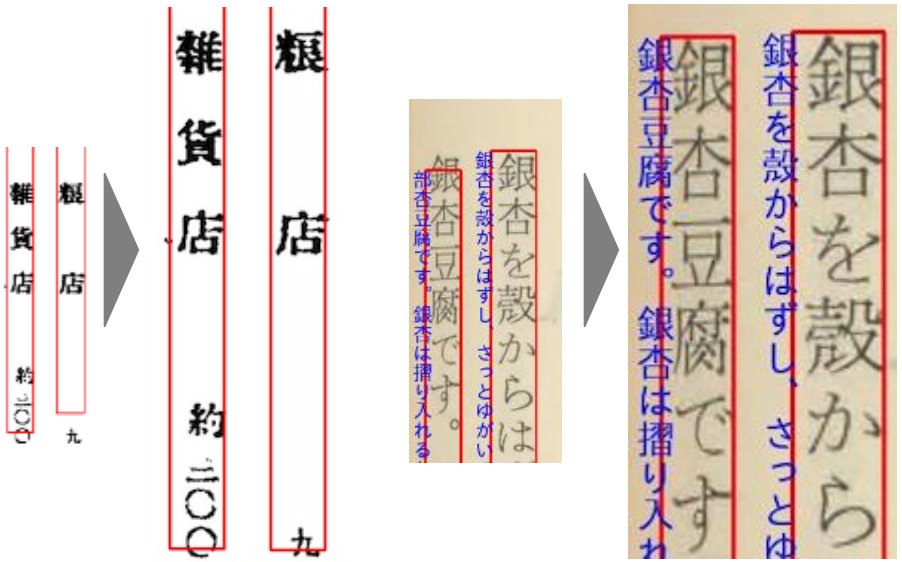
縦横混在の改善



行の重複の改善



長さ不足の改善



2．速度改善
1）傾き補正と統合プログラムでの速度改善

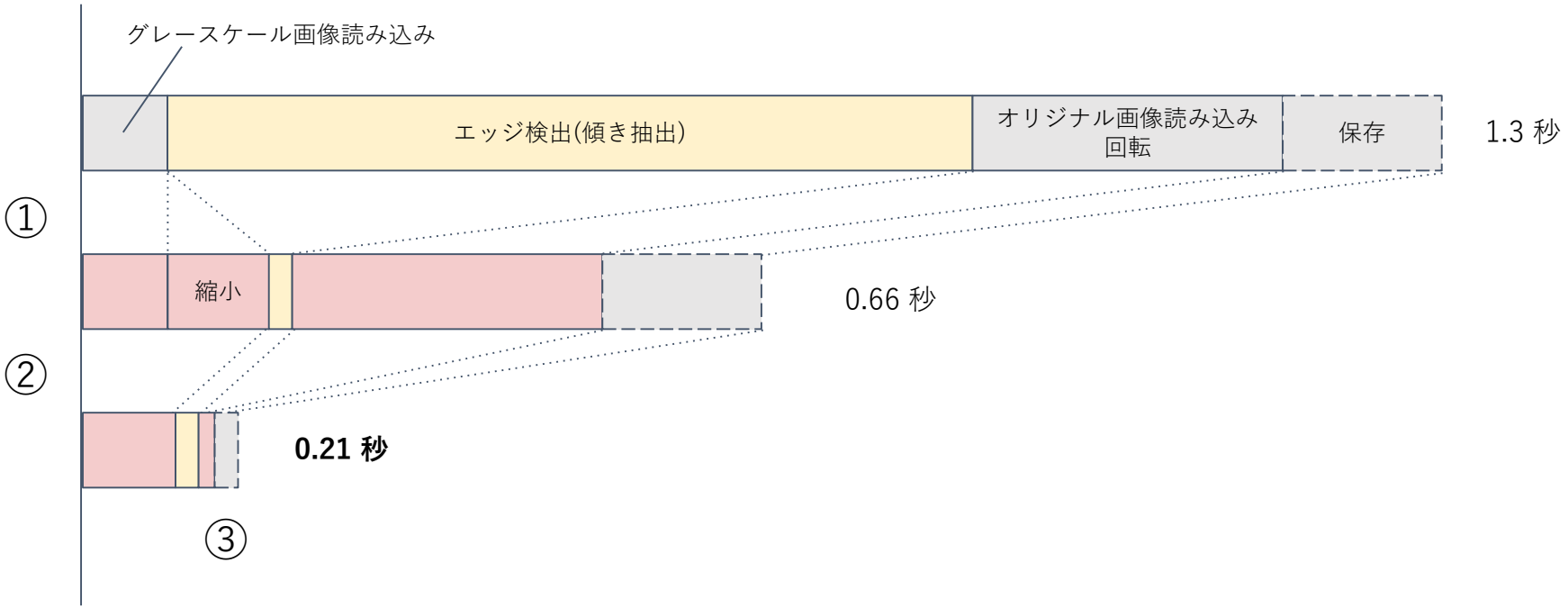
■ 傾き補正の速度改善

① 処理時間の半分以上を占めるエッジ検出(傾き抽出)の時間を入力画像を縮小することで短縮
　画像縮小による精度劣化を低減するためにパラメータを調整

② 画像の読み込み、縮小や回転などに使用していた画像処理ライブラリをscikit-imageからOpenCVに変更

精度変化：評価画像99枚 x 2(左右) で平均誤差 0.259度 → 0.265度

③ 統合プログラムで画像の保存を介さずメモリ上での画像の受け渡しを実行



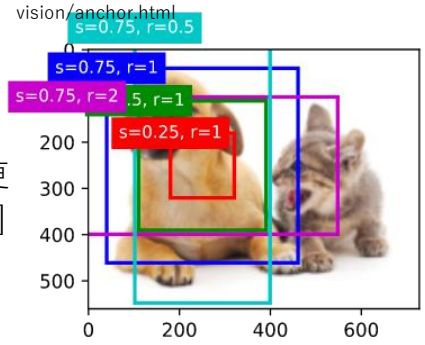
その他の改善検討

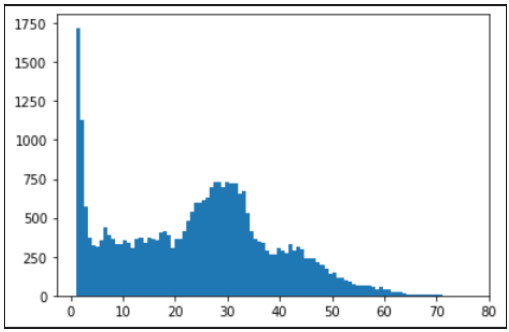
■ レイアウト

アンカーボックスサイズ最適化

アスペクト比の構成を全データの矩形情報から以下の通り変更
[1/24, 1/16, 1/8, 1, 8, 16, 24] → [1/32, 1/16, 1/4, 1, 4, 16, 32]

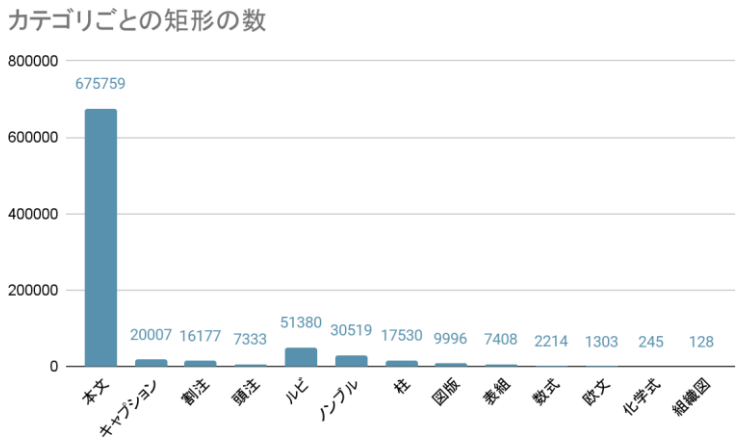
https://d2l.ai/chapter_computer-vision/anchor.html





Equalization Loss

全データの8割を占める本文矩形に偏った学習となるのを回避
少数カテゴリでは一定割合で多数カテゴリデータを無視する



■ 行認識

合成漢字データの利用

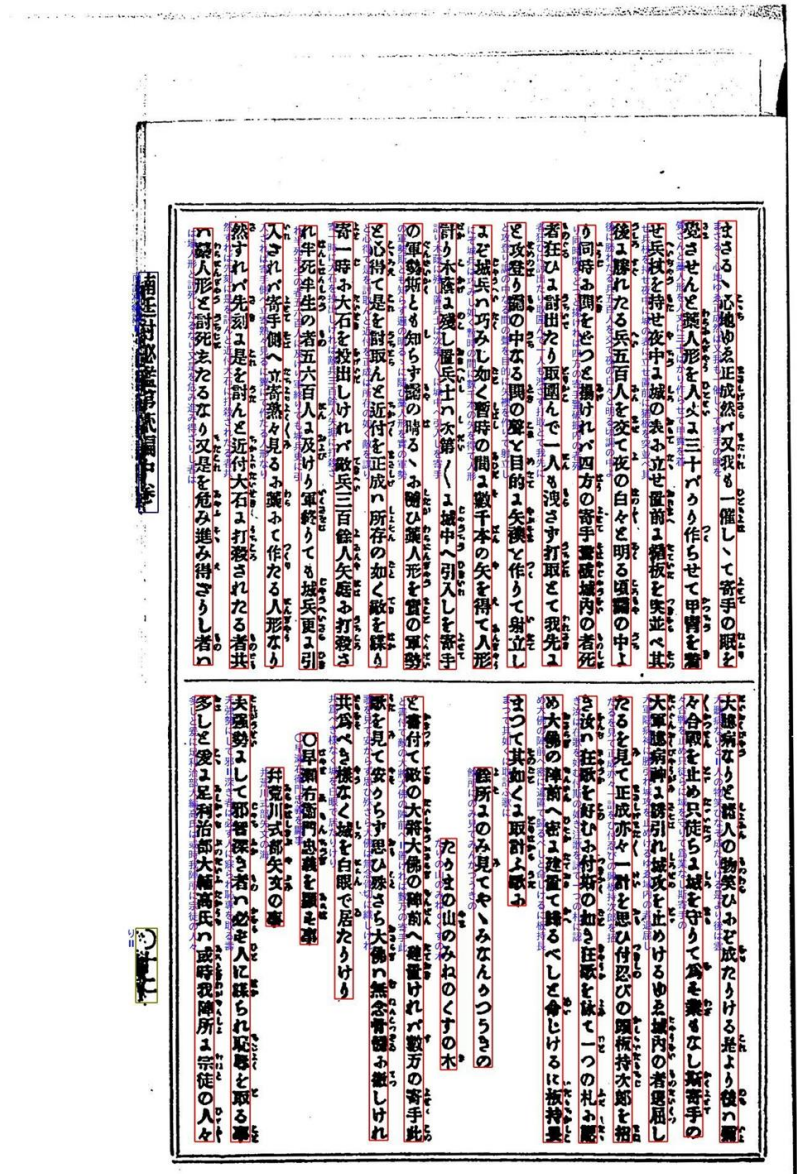
認識対象文字7083字に対して784字はデータセットに出現しない

Image: 退津頽端燿暨ｃ滔夸他象苫巧犂綫耀董伜幢熙瀬旋杼薹：串鵞燒美海癜歹彡煉聢路包哩墻稼
True: 退津頽端燿暨ｃ滔夸他象苫巧犂綫耀董伜幢熙瀬旋杼薹：串鵞燒美海癜歹彡煉聢路包哩墻稼
Pred: 退津頽端燿暨c滔夸他象苫巧犂綫耀董伜幢熙瀬旋杼墓:串鵞燒美海歷ダ彡煉聢路包哩墻稼

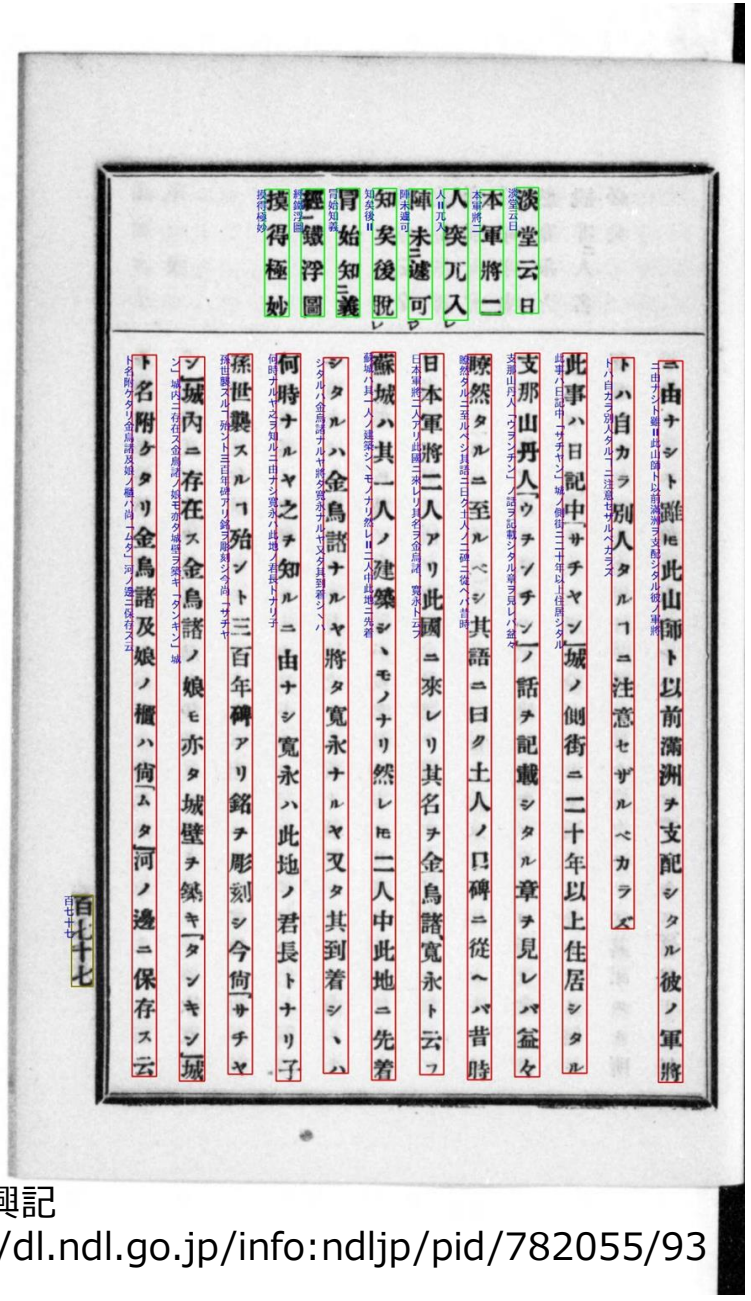
精度変化：0.788 → 0.800 (出現数50以下の文字でF1 0.267→0.405)

処理結果サンプル

旧字②（1909年以前の文系図書）



楠廷尉秘鑑：今古実録 第２編　中巻
https://dl.ndl.go.jp/info:ndljp/pid/880917/13



義経再興記
https://dl.ndl.go.jp/info:ndljp/pid/782055/93