令和4年度
視覚障害者等用データ作成のための
OCR処理プログラム追加研究開発作業
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

令和3年度にNDLが開発したOCR処理プログラム（NDLOCR）に対して生成するテキストデータを視覚障害者用の読み上げ用途にも活用できる更なる機能改善に取り組む。

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

- 学習用データセット：計10,803件
- 評価用データセット：計3,013件

② OCR処理プログラムの開発

- OCR処理プログラムの追加開発
  ・読み上げ順序の調整機能の開発
  ・レイアウト情報の自動付与及び読み上げ上不要な要素の削除機能の開発
  ・漢字の読み情報の自動付与機能の開発

- OCR処理プログラムのテキスト化性能改善
  ・1960年代以降の現代書籍を主な対象に、読み上げ順序も考慮したテキスト化認識精度の改善

- 読み上げ順序及び文字認識性能の評価方法の設計及び評価
  ・開発したOCR処理プログラムを適切に評価するための評価指標・評価方法の設計
  ・評価の実施

本事業のアプローチ

■受託者のアプローチ

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処理プログラムの多くは帳票を対象としているため、書籍・雑誌に特有の複雑なレイアウトや、それに基づいた読み順の推定に対応しておらず精度面で課題がある。また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、書籍・雑誌を含む様々な日本語資料のOCR処理の実績、及び字形データセット構築の実績を有する。そのため過去の知見を活かし、多段組や回り込みといった複雑なレイアウト、多様な文字デザインといった現代活字資料の特徴を網羅した良質なデータセットの構築が可能。

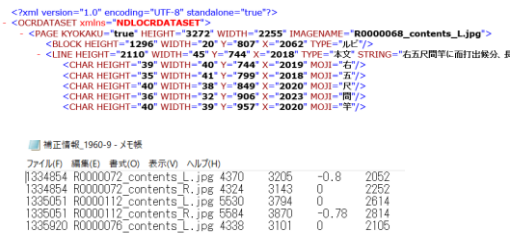
② OCR処理プログラムの開発

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262万資料の集計 • データセット用の画像抽出・選別作業 • 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
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 

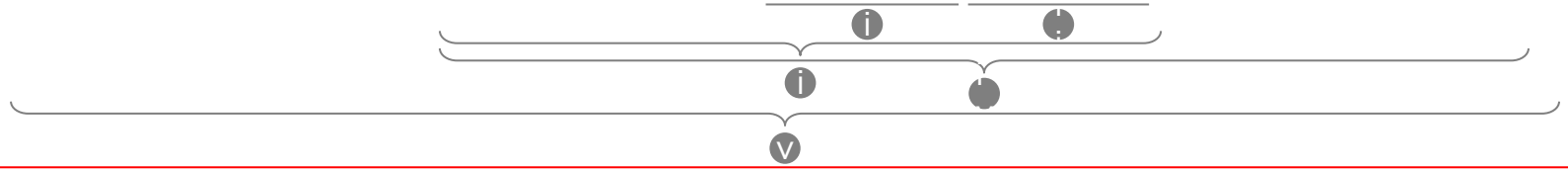
読み上げ順序及び文字認識性能の評価方法
本文行内の読み順の正しさを評価するための指標

読み順を考慮した評価指標として「①本文行内の読み順の正しさを評価するための指標」と「②本文行間の読み順の正しさを評価するための指標」の２つを定義し、開発したプログラムの性能改善に生かす。

①本文行内の読み順を評価するための指標

評価指標は正解文字列と推論結果の文字列の編集距離(Levenshtein distance)を用いる。
各行の正解文字列数を文字列の長さで正規化し、正規化編集距離として定義する。

評価指標： *Average{k=1~n} (Levenshtein distance(line_true_k, line_pred_k) / Length(line_true_k))*



評価例

		ⓘ正解テキスト	ⓙ推論テキスト	ⓚ編集距離	ⓛ正規化編集距離	ⓥ平均値
		line_true_k	line_pred_k	Levenshtein distance	Levenshtein distance / Length(true_line_k)	Average{k=1~n} (Levenstein distance /Length(true_line_k))
k番目	1	ひさかたの	ひさかた	1 (追加)	0.2 (1/5)	1ページに含まれる各行内の文字の読み上げ順序の評価指標 平均値 0.12 (0.6/5)
	2	光のどけき	二のどけき	1 (置き換え)	0.2 (1/5)	
	3	春の日に	春の日に	0 (編集不要)	0	
	4	しづ心なく	づ心なく	1 (追加)	0.2 (1/5)	
	5	花の散るらむ	花の散るらむ	0 (編集不要)	0	

*Average=平均値　*n=正解文字列の数、k=各文字列に割りあてられた番号（1~nを附番）
*Levenshtein distance =文字列に変形するのに必要な手順の最小回数として定義、各行の編集距離を算出
*Length　文字列の長さ
*line_true_k＝k番目の本文行の正解文字列、line_pred_k＝k番目の推論結果の文字列（該当する推論結果がない場合は空の文字列）

読み上げ順序及び文字認識性能の評価方法
本文行間の読み順の正しさを評価するための指標

続いて、「②本文行間の読み順の正しさを評価するための指標」として以下の定義を提案する。

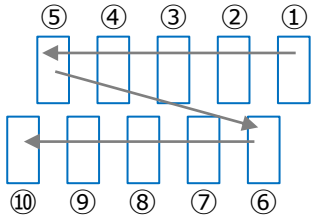
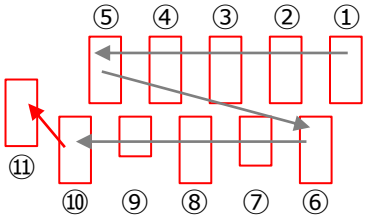
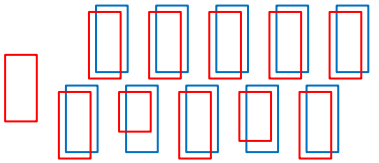
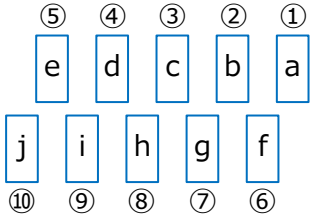
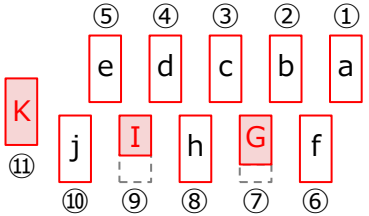
②本文行間の読み順の正しさを評価するための指標

②-1.本文行の順序を文字列とみなし、正解の本文行の順序と推論結果の本文行の順序の正規化編集距離を算出する。

評価指標： $\frac{Levenshtein\ distance\ (line\_order\_true,\ line\_order\_pred)}{n}$ = $\frac{a}{d}$

評価例

<凡例>
数字=読み順
英字=ユニークな記号

	正解行レイアウト・読み順	推論行レイアウト・読み順	判定/割り当て/算出結果
IoUで行の一致を判定			IoU=0.8以上を閾値に一致判定  （左記例） IoU=0.89 ↓ 0.8以上の為 一致判定
ユニークな記号を割り当て			<記号割り当て結果> b ***line_order_true***: abcdefghij （正解） c ***line_order_pred***: abcdef**G**h**I**j**K**（推論）
正規化編集距離算出	正解行読み順：abcdefghij	<編集距離の計算> 2置換、1文字削除 推論行読み順：abcdef**G**h**I**j**K** 正解行読み順：abcdefghij	<正規化編集距離算出> a 編集距離(***Levenshtein distance***)＝3 d 正解行読み順数(***n***)＝10 **正規化編集距離＝ a / d ＝0.3**

*IoU=2つのボックスの積集合の面積 / 2つのボックスの和集合の面積
*line_order_true = 正解行読み順（正解データの行の読み順を文字列化した値）
*line_order_pred = 推論行読み順（推論結果の行の読み順を文字列化した値）　*n = 正解の行数

想定される紙面のバリエーションと評価結果

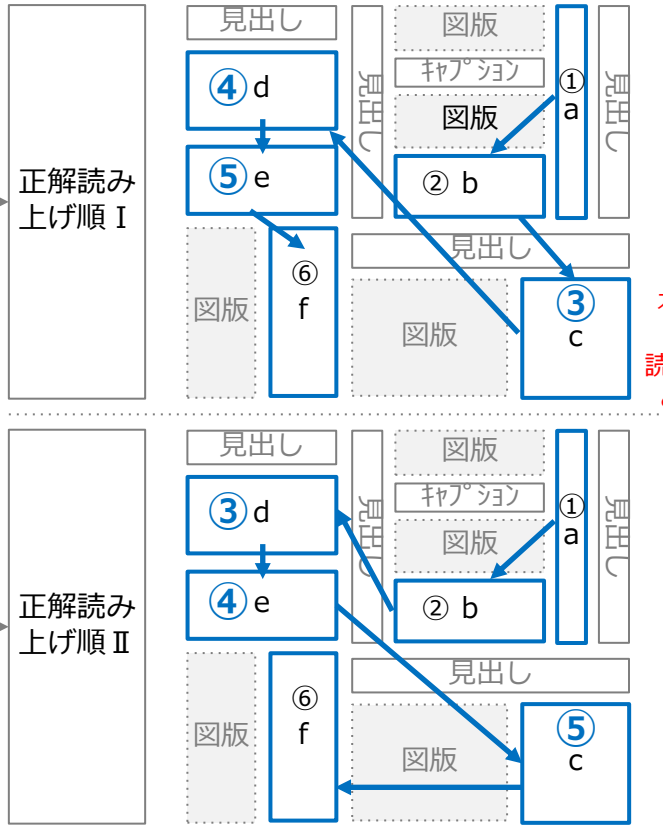
必ずしも一意に読み上げ順序の定まらない資料に対する評価例を以下に示す
　読み上げ順が一意にの定まらない資料に対しては「複数の正解行読み上げ順序を用意」し、
　推論行読み上げ順序との正規化編集距離の値が低いものを正解行読み上げ順序として評価する。

紙面のバリエーション

読み順が一意に定まらない資料

正解レイアウト・読み上げ順序

正解の読み上げ順序を複数用意

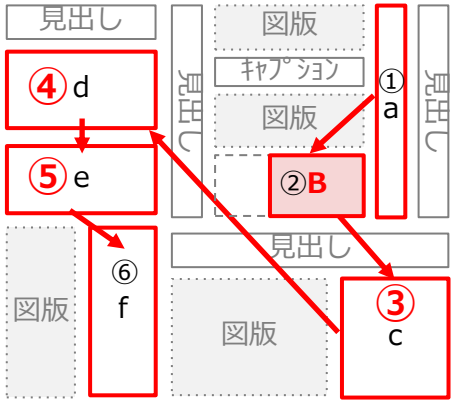


本ケースは
Ⅰを正解
読み上げ順序
として評価

推論レイアウト・読み上げ順序

**正規化編集距離が小さい
読み上げ順序を正解とする**

正解行読み上げ順序＝ abcdef
推論行読み上げ順序＝ aBcdef
編集距離　　＝ 1
正規化編集距離　　＝ **0.166**
(行内文字編集距離は記載略)



正解行読み上げ順序＝ abdecf
推論行読み上げ順序＝ aBcdef
編集距離　　＝ 4
正規化編集距離　　＝ **0.666**
(行内文字編集距離は記載略)

- **データセット全体としてのスコア算出方法**

  左右のページ単位で算出された評価スコア全体の中央値

- **行内（行認識）評価対象カテゴリ**

  本文、広告文字

  ※本文がすべてインライン要素で正解文字列が＝のものは対象外

  ※本文の一部に存在するインライン要素については該当部分を黒塗りした画像で評価

- **行間（読み順）評価対象カテゴリ**

  本文

- **行内・行間スコア目標値**

  レベル4については中間報告段階スコアからの改善で判断し明示的に設定はしない

	行内（行認識）スコア	行間（読み順）スコア
レベル1	0.01	0.0
レベル2	0.01	0.0
レベル3	0.015	0.1
レベル4	(0.142)	(0.374)

OCR処理プログラム認識性能の達成状況

- ページごとに算出された編集距離ベースの評価スコア（※１）の中央値で評価を実施
- 中間報告後に設定された仕様に基づいて最終スコアを取得し目標スコアの達成を確認

※１スコア算出方法については、「読み上げ順序及び文字認識性能の評価方法」（スライド７～10）を参照
※最終スコアの算出ではアノテーションに基づいてインライン要素を黒塗りするため、分割傾き補正された評価画像を利用（レイアウト認識から処理開始し黒塗りなし画像を利用、その結果と黒塗り画像で行認識以降の処理実施）
※参考スコアは上記の黒塗り画像を利用せず、分割処理からNDLOCRで処理を行った際のスコア
※各スコアは０から１の範囲を取り、０に近づくほど正確であることを表す

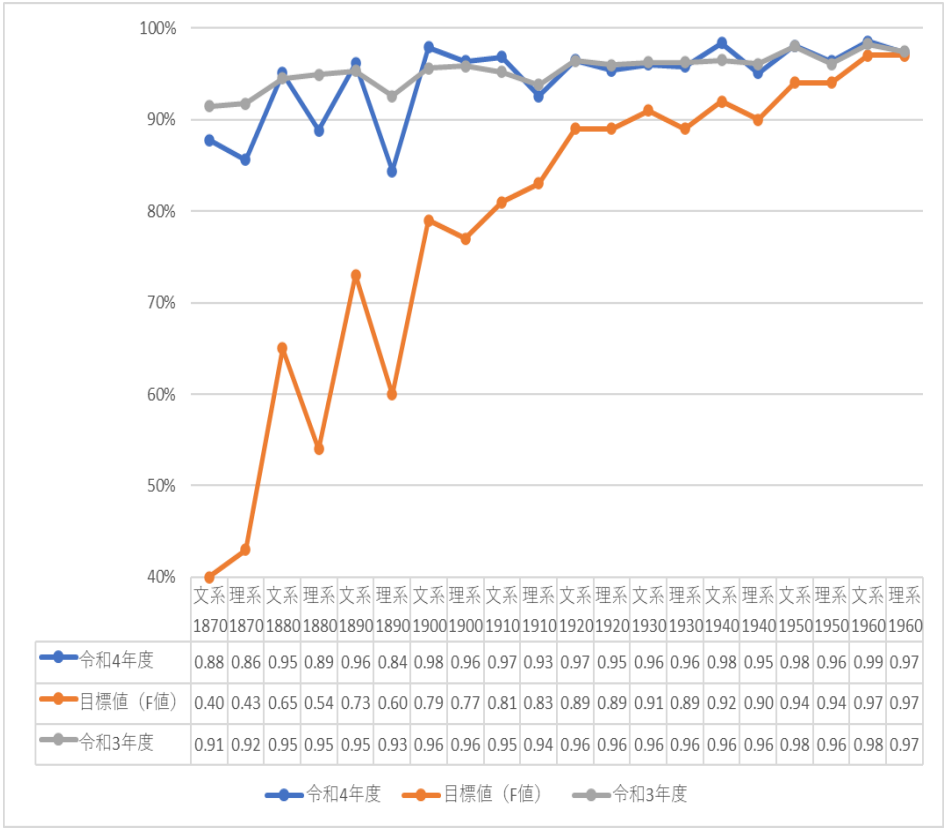
		目標スコア	最終スコア （調整あり）	参考スコア （調整なし）
行内スコア （行認識）	レベル１	0.01	0.0024	0.0031
	レベル２	0.01	0.0034	0.0108
	レベル３	0.015	0.0150	0.0419
	レベル４	(0.142)	0.0725	0.1703
行間スコア （読み順）	レベル１	0.0	0.0	0.0
	レベル２	0.0	0.0	0.0256
	レベル３	0.1	0.0556	0.125
	レベル４	(0.374)	0.2111	0.4286

OCR処理プログラム認識性能の達成状況（参考値：前年度評価指標での評価）

- 令和３年度のテキスト化認識性能評価で採用されたF値での評価を令和４年度モデルでも評価実施
- 1870年代から1960年代までは令和３年度NDLOCRの認識性能が高い
- 1960年代以降（雑誌のみ）は令和４年度NDLOCRの認識性能が高い
- レイアウト認識はデータ仕様変更で令和４年度データのみ、行認識は令和３年分も学習していたため全体的に性能劣化少ない
  ※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）

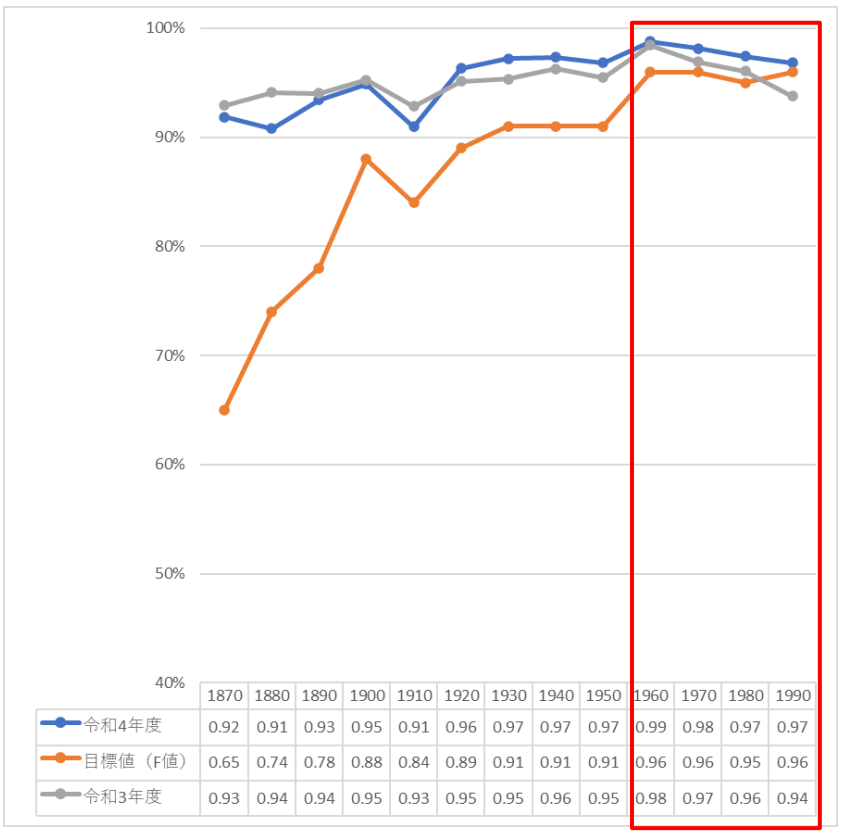
前年度認識性能基準と認識性能評価結果まとめ

書籍種別・年代別の評価



	文系1870	理系1870	文系1880	理系1880	文系1890	理系1890	文系1900	理系1900	文系1910	理系1910	文系1920	理系1920	文系1930	理系1930	文系1940	理系1940	文系1950	理系1950	文系1960	理系1960
令和4年度	0.88	0.86	0.95	0.89	0.96	0.84	0.98	0.96	0.97	0.93	0.97	0.95	0.96	0.96	0.98	0.95	0.98	0.96	0.99	0.97
目標値（F値）	0.40	0.43	0.65	0.54	0.73	0.60	0.79	0.77	0.81	0.83	0.89	0.89	0.91	0.89	0.92	0.90	0.94	0.94	0.97	0.97
令和3年度	0.91	0.92	0.95	0.95	0.95	0.93	0.96	0.96	0.95	0.94	0.96	0.96	0.96	0.96	0.96	0.96	0.98	0.96	0.98	0.97

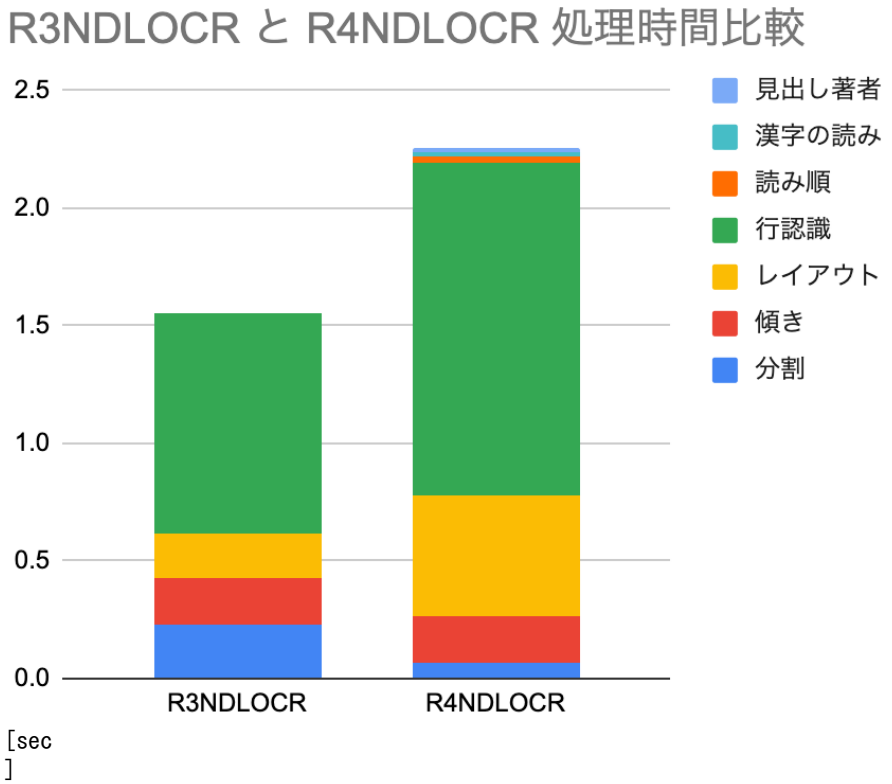
雑誌の年代別の評価

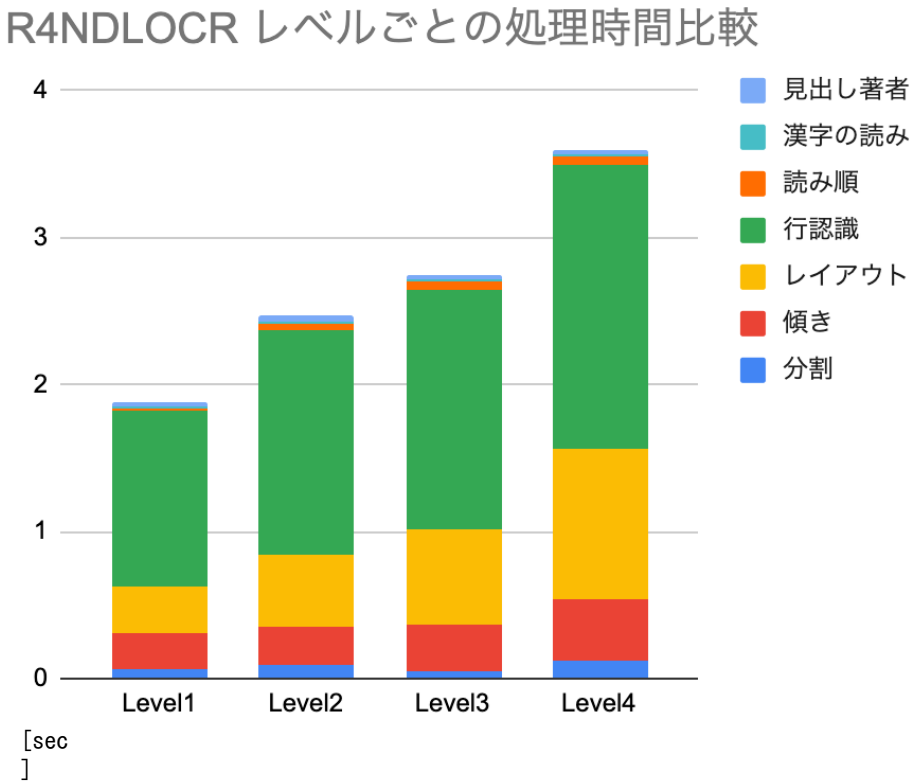


	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
令和4年度	0.92	0.91	0.93	0.95	0.91	0.96	0.97	0.97	0.97	0.99	0.98	0.97	0.97
目標値（F値）	0.65	0.74	0.78	0.88	0.84	0.89	0.91	0.91	0.91	0.96	0.96	0.95	0.96
令和3年度	0.93	0.94	0.94	0.95	0.93	0.95	0.95	0.96	0.95	0.98	0.97	0.96	0.94

OCR処理プログラム速度性能の状況

- 令和３年度1.55秒だった処理時間は今年度2.26秒となった
- 要因としては行認識（0.935秒→1.420秒）とレイアウト認識（0.19秒→0.508秒）の増加が大きい
- 分割は改善（0.23秒→0.066秒）が見られたが全体のインパクトとしては小さかった
- その他の追加機能に関してはほぼ無視できる程度での増加となった
- レベルごとに処理時間の増加傾向が見られた（ページの複雑化による行数の増加の影響）
  ※レイアウト認識のメモリ不足問題の2つの対応（mmdetectionの変更とPyTorchの設定）で速度劣化発生





OCR学習用データセット

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
- 全面広告
- 目次
- グラビア（＋グラビア内テキスト）画像
- 全面外国語

[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
- ブランクページ
- 片観音、両観音、Z折り、特殊折り
- 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

OCR評価/学習用データセットの納品実績

貴館と合意した定義に基づき、抽出した雑誌、書籍のデータをレベルごとに仕分けし、それぞれ定められた数量を納品。また後半期開始にあたってはレベルごとの数量調整を行った。最終的に評価データ3,013件、学習データ10,803件の納品を1/26までに完了させた。

OCR評価/学習用データセットの納品数量及び内訳（全体サマリー）

レベル	内容		7/22納品	8/05納品	8/19納品	9/02納品	9/16納品	9/30納品	11/25納品	12/07納品	12/22納品	01/01納品	01/26納品	合計
			実績	実績	実績	実績	実績	実績	実績	実績	実績	実績	実績	実績
1	一段組みの簡単なレイアウト	評価	340	321	385		124	30						1,200
		学習	100	100	55		753	410		408	409	403	407	3,045
2	単純な多段組レイアウト	評価	87	133			24	42	65					351
		学習	133	87			194	178		101	151	154	153	1,151
3	少し複雑な多段組レイアウト	評価		350	411	500								1,261
		学習		108	446	498	622	629	424	815	706	651	506	5,405
4	複雑なレイアウト	評価		201※										201
		学習				100	500	602						1,202
合計		評価	427	1, 005	796	500	148	72	65					3, 013
		学習	233	295	501	598	2, 069	1, 819	424	1, 324	1, 266	1, 208	1, 066	10,803

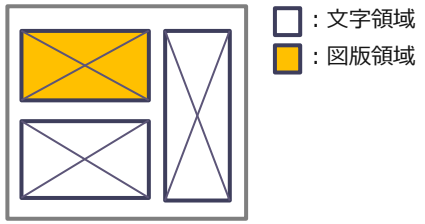
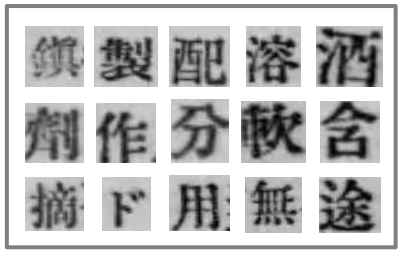
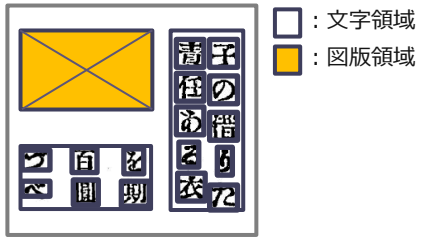
※1件、読み方がふた通りあり

OCR学習用データセット構成（1/2）

※令和3年度開発と同じ

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、広告、数式、化学式など）についても以下の通り定める。

データセットに含まれる情報

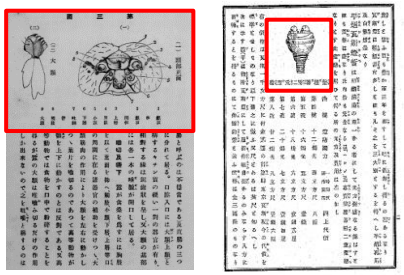
本文テキスト内	テキストブロック情報	• 多角形の頂点（ポリゴン）（X_1,Y_1,X_2,Y_2,X_3,Y_4…）
	行矩形情報	• 座標（X,Y） • 高さ（px） • 横幅（px） • 書字方向（縦、横、右から左） • 90度回転情報（右、左） • 行内文字列情報（文字コード） • 見出し情報（TRUE/FALSE） • 著者名情報（TRUE/FALSE） • 種類（本文、キャプション、割注、頭注、広告文字）
	文字矩形情報	• 座標（X,Y） • 高さ（px） • 横幅（px） • 文字情報（文字コード） • 種類（欧文、回転欧文、色付文字、縦中横、手書き、数式、化学式）
本文テキスト外	画像情報	• 画像名 • 画像高さ（px） • 画像横幅（px） • 匡郭有無（TRUE/FALSE） • 傾き修正情報（角度）※ • トリミング情報（左右・切断位置）※
	文字以外の領域情報	• 座標（X,Y） • 高さ（px） • 横幅（px） • 種類（図版、広告、表組、数式、化学式、系統図、柱、ノンブル、ルビ）

※画像補正作業で作成するTSVファイルに記載

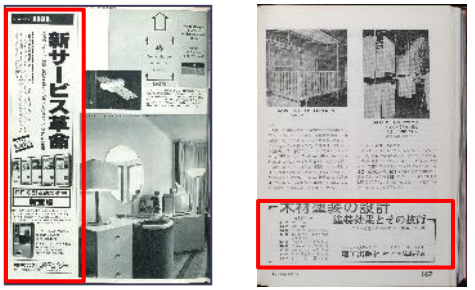
文字以外の領域情報

- 図版、広告、表組、数式、化学式、系統図、柱、ノンブル、ルビは、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、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。
  ただし、広告内の文字については文字情報の活用を想定し、行矩形情報「広告文字」としてアノテーションを行う。

[図版]（ブロック要素）



[広告]（ブロック要素）



[数式]（ブロック要素・インライン要素）

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

[化学式]（ブロック要素・インライン要素）

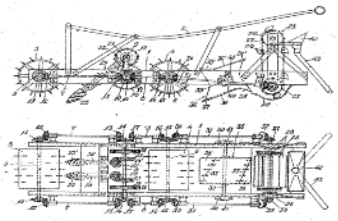
$$SO_4{<}^{H}_{H}+OHC_2H_5=H_2O+SO_4{<}^{C_2H_5}_{H}$$

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」「広告」
  ※「広告」内の文字は別途LINE要素の「広告文字」としてアノテーション

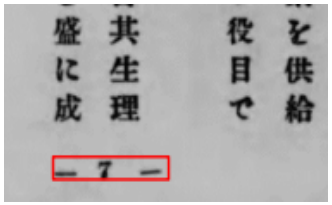
BLOCK

図版



表組

ノンブル



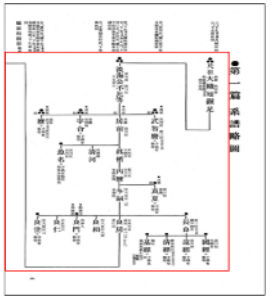
数式

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

化学式

$$SO_4\langle^{H}_{H}+OHC_2H_5=H_2O+SO_4\langle^{C_2H_5}_{H}$$

系統図



柱

ルビ

広告

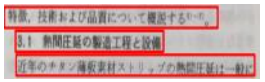
OCR学習用データセット レイアウト情報ーライン要素・インライン要素

- LINEの種類（TYPE）は、 「本文」 「キャプション」「頭注」「割注」「広告文字」。

  また、LINE内に「見出し」「著者名」「90度回転」有無の情報を含む。

LINE

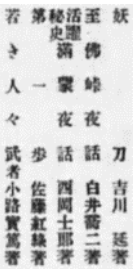
本文



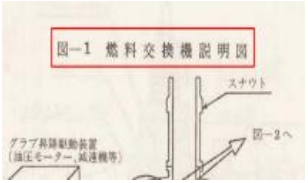
頭注

広告文字

著者名



キャプション



割注

見出し

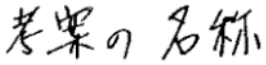
90度回転

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

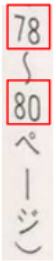
- INLINEの種類（TYPE）、「手書き」 「数式」「化学式」 「縦中横」 「色付文字」 「欧文」「回転欧文」

INLINE

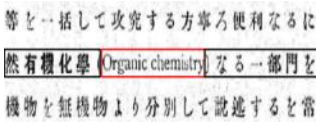
手書き



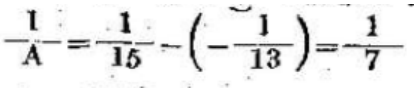
縦中横



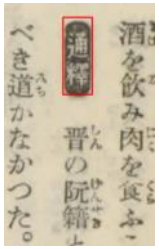
欧文



数式

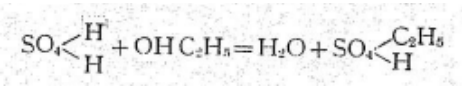


色付文字



回転欧文

化学式



OCR学習用データセット 文字種

※令和3年度開発と同じ

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「〓」を割り当てる。

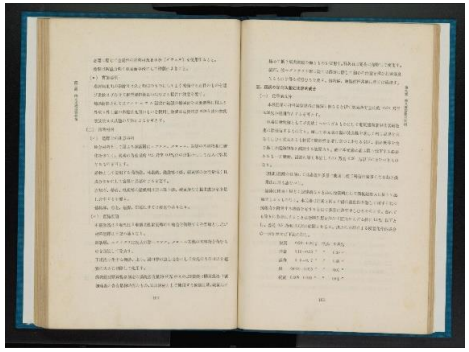
OCR処理プログラム

実装した手法全体の処理フロー概要

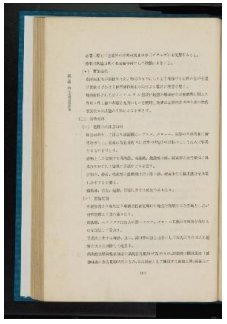
画像の出典：
海軍艦政本部 編『鋳造作業標準』,日本鋳物協会,昭和7. 国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/1234548/1/87

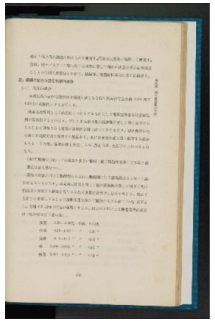
実装した手法全体の処理フローは以下の通りである。

入力画像

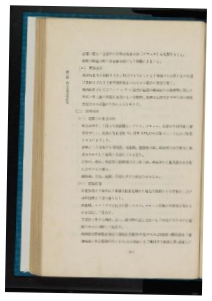


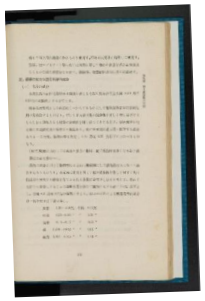
①見開き分割（分割モジュール）



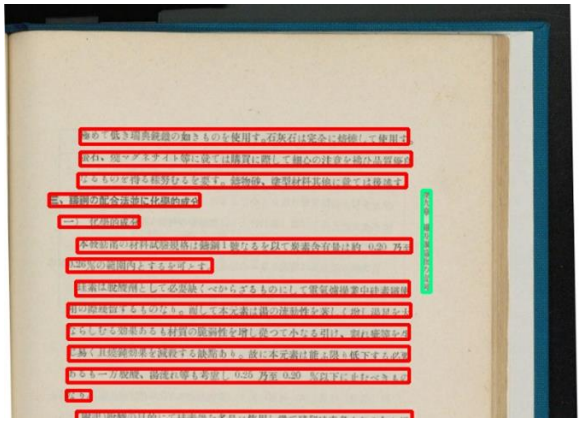


②傾き補正（傾き補正モジュール）

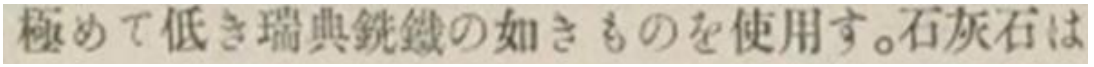




③レイアウト認識（レイアウト認識モジュール）

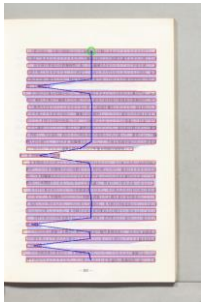


④行認識（行認識モジュール）



極めて低き瑞典銑鐵の如きものを使用す。石灰石は
…

⑤読み順推定（読み順推定モジュール）



⑥漢字の読み推定
（漢字の読み推定モジュール）

⑦見出し著者認識
（行属性推定モジュール）

実装した各手法について
アプローチと昨年度分からのアップデート概要

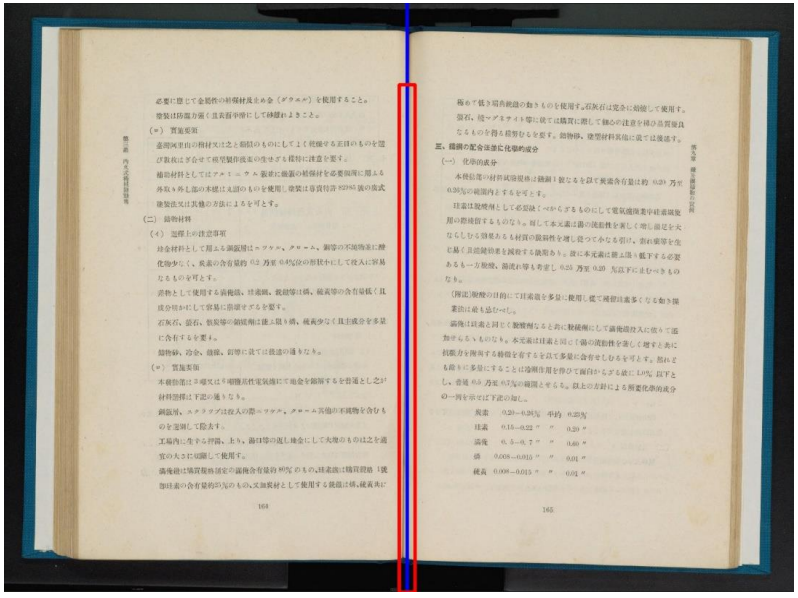
上の画像の出典：前ページと同様
下の画像の出典：ミュルレル 著 ほか『欧洲新政史』下,八尾書店,明27.10.
国立国会図書館デジタルコレクション https://dl.ndl.go.jp/pid/776486/177/1

見開き分割

アプローチ

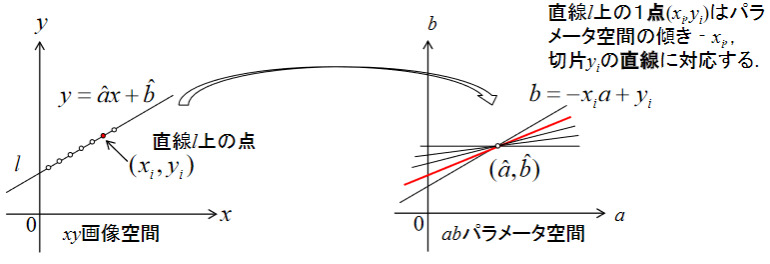
- アプローチとしてはノド元のx座標検出器で変更なし
- モデルをMobileNet SSDから**Cascade RCNNを利用**する形に変更
- **ノド元が中央にないケースに対応**するため、昨年度ノドが中央にあった画像から横方向のクロップを行った人工データで学習実施

手法のイメージ



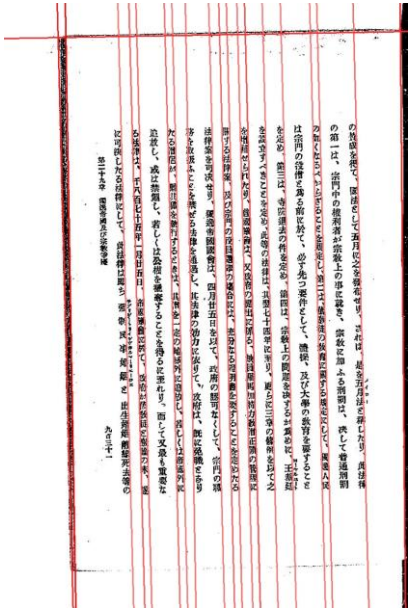
傾き補正

- **今年度変更なし**
- 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
アプローチと昨年度分からのアップデート概要

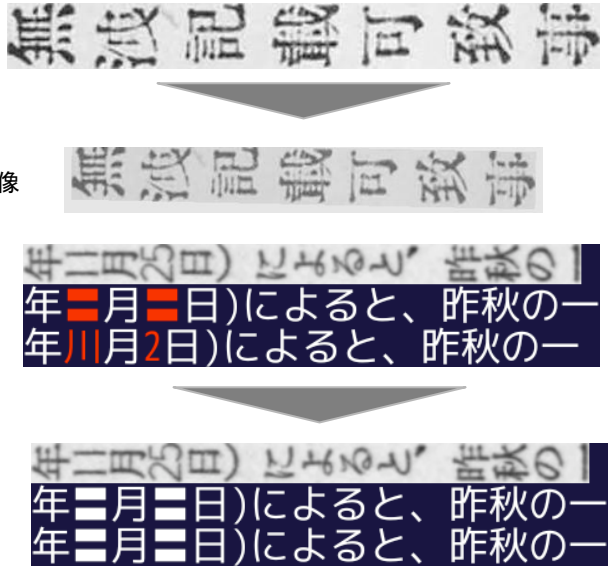
アプローチ

レイアウト認識

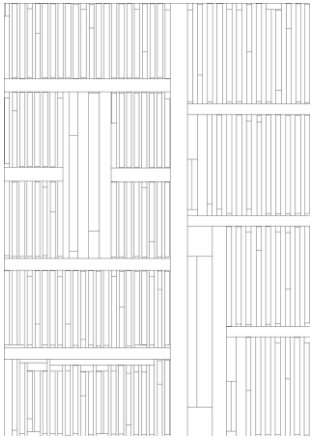
- **本文ブロックのマスク推定**を行うためCascade Mask RCNNを継続採用（昨年度**検出部分のみだったがセグメンテーションまで行う形に変更**）
- Resnetバックボーンから速度劣化を抑えて精度向上が期待できる**ConvNeXtバックボーンに変更**
- レイアウト認識特有のチューニング
  - 画像に対して対象が小さい → 入力解像度を大きく
  - 対象の数が多い → Region Proposal(途中段階の候補領域提示)の上限増
  - アスペクト比が極端（縦長、横長）→ アンカーボックスのサイズ調整
- **矩形の検出過不足に対して後処理を実装して改善（後述）**

行認識

- シーンテキスト認識手法Rosettaを採用（**変更なし**）
  (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
- ”川”と”三”など縦行を横向きで認識する際に問題になる文字の対応のため、**STRIDE（ネットワーク内部の縦/横行推定モジュール）を導入**
- **レイアウト認識結果に基づき不要要素（柱・ノンブル）をテキスト化対象/対象外として選択可能に**



実装した各手法について
アプローチと昨年度分からのアップデート概要

	アプローチ	手法のイメージ
読み順推定	• Recursive XY Cutを採用 ページ内の文字の集まりでない部分の余白を大きな順に縦方向か横方向で切るという処理を行い、切られた部分領域に対しても同様の処理を繰り返してページ内でブロックを作成 上下と左右に切断する際にどちらを先に読むかのルールを設定することでブロックを作成しながら読み順を構成していく • レイアウト認識の本文ブロックの情報も利用して、Recursive XY Cutで対応できないケースに対する補正を実施	 
行属性付与 見出し 著者認識	• Random Forestを採用 見出しは行内の文字ではなく位置やサイズの情報だけで判定できるのではという仮説 これに対して利用した情報がX座標、Y座標、文字数、行の幅と高さ、画像内の全矩形の幅と高さの平均	
漢字の読み推定	• 形態素解析器KyTeaを採用 • 行ごとの処理で行末で単語が途切れるのを避けるため、1文ごとの処理実施	

中間報告時点の課題に対する取り組み

改善アプローチ

１）レイアウト認識：本文行、本文ブロックの欠落

レイアウト認識のモデル構造の変更
→ 特徴抽出を主に行うバックボーンをResNetからConvNeXtに
→ モデル変更に伴って矩形の重複検知や本文ブロックマスクはあるが本文行矩形はないケースなどが発生したため、別途レイアウト認識の後処理でこれらの問題に対応

２）読み順推定：縦書き横書きの混在、段抜き見出し、割注など複雑なレイアウトにおける読み順の混乱

縦書き横書きの混在、段抜き見出し対応
→ ListNetやLightGBMなどの特徴ベースの学習手法からRecursice XY Cutに手法変更

割注対応
→ ルールベースで割注を膨張処理で1つにまとめ順序を決定、割注内は縦書きか横書きかで右→左か上→下を決定する

３）その他

学習データ増加
→ 直接的な課題対応ではないが中間報告時点では学習データが半分程度しかない状態であった
すべてのモジュールにおいて学習データが倍になることでの精度向上が見込まれる

漢字の読み推定及び見出し・著者推定の機能実装

■ 漢字の読み推定

形態素解析器のKyTeaを利用して本機能を実装

KyTea採用の理由：

処理速度、読み付与不可数の少なさ、精度のバランス

青空文庫・サピエ点字振り仮名データセット1944件で

正規化編集距離が0.0672となった

■ 見出し・著者推定

Random ForestとBERTの2手法を実装

高速なRandom Forestとスコアの高いBERTという結果に

手法	スコア	読み付与失敗漢字数	速度（1000文字平均）
kakasi	0.905806	130	0.0415
mecab	0.950673	1412	0.0986
sudachi	0.911929	3118	0.0568
kytea	**0.911424**	**124**	**0.0127**
vaporetto	0.927852	13150	0.0022

Random Forest

TITLE	LEVEL1	LEVEL2	LEVEL3	LEVEL4
F1	0.4974	0.3499	0.5156	0.589
precision	0.4942	0.4497	0.6109	0.5523
recall	0.5007	0.2863	0.4461	0.6309
AUTHOR	LEVEL1	LEVEL2	LEVEL3	LEVEL4
F1	0.2993	0.0793	0.4014	0.1635
precision	0.2857	0.0811	0.5009	0.4483
recall	0.3143	0.0776	0.3349	0.1

BERT

TITLE	LEVEL1	LEVEL2	LEVEL3	LEVEL4
F1	0.6087	0.5577	0.5857	0.5050
precision	0.5149	0.4895	0.5092	0.4255
recall	0.7445	0.6479	0.6893	0.6212
AUTHOR	LEVEL1	LEVEL2	LEVEL3	LEVEL4
F1	0.3113	0.2768	0.5401	0.574
precision	0.2026	0.1689	0.4281	0.4694
recall	0.6714	0.7672	0.7316	0.7385