デジタル化資料のOCRテキスト化

改善結果報告書

資料作成：LINE株式会社

全文テキスト化
サンプル画像に対する性能検査結果

特化型の学習をさせていない弊社のベースラインモデルではF値の目標値を上回っているのは21/33区分でしたが、本件での改善計画を実施した結果、第１回性能検査では**31/33区分**、第２回性能調査*では**32/33区分**で目標値を上回りました。

性能検査の目標値達成件数

性能検査の区分	目標値達成件数
ベースラインモデルでの検査	**21**/33
第１回性能検査	**31**/33
第２回性能検査	**32**/33

全文テキスト化
環境準備、サンプリング手法の実装と提案手法の検証

計画の通り、環境準備と特徴量抽出・クラスタリング・サンプリングをするための実装をしました。
また、ご紹介いただいたデータセットを利用して、サンプリング手法の検証をしました。

計画と実績

凡例　**赤字**：主な変更点

	計画	実績
環境準備	• Step1:特徴量抽出～Step3:サンプリングを実行するための環境準備	• 計画と同様
サンプリング手法の実装	• CLOVA OCR検出モデルを用いたドキュメントイメージの特徴量抽出機の実装 • 特徴量データのクラスタリングプログラムの実装 • 機能テスト	
サンプリング手法の検証	• 一部のデータを利用して提案したサンプリング手法を検証	• ご紹介いただいた**NDL-DocLデータセット***を利用して提案したサンプリング手法を検証

* NDL-DocLデータセット(資料画像レイアウトデータセット) https://github.com/ndl-lab/layout-dataset

全文テキスト化
特徴量抽出・クラスタリング・サンプリング(Step1～3)

計画時は、Step1～3を順番に実施する予定でしたが、HDDから弊社ストレージへの転送に想定以上の時間を要したため、4回に分割して転送が完了した画像から順次処理を実施するという方針に変更させていただきました。

Step1～3を分割して実施した時期と処理枚数

	実施時期(2021年)	処理枚数
1回目	06/09-06/20	約1,000万枚
2回目	06/16-07/02	約2,000万枚
3回目	06/30-07/16	約2,000万枚
4回目	07/14-07/30	約2,000万枚
	全体	**約7,000万枚**

特徴量抽出・クラスタリング・サンプリングの手順

4分割した各回において、下記の手順をそれぞれ実施

Step	タスク名称	詳細
Step1	イメージ変換	処理の速度向上のため、同じサイズ(1280x1280)のjpeg(Quality:75%)へ変換
	特徴量抽出	入力イメージに対して Bottleneck Layerの特徴量を抽出
Step2	クラスタリング	FAISSを利用してk-means法で区分毎にクラスタリングを実施
Step3	サンプリング	各クラスターから中心に一番近いイメージを1枚選択
	サンプル確認	上記サンプリングで選択されたサンプルに偏りがないかなどを目視で確認

全文テキスト化
アノテーション(Step4)　対応実績のサマリ

学習用データセット・性能検査用テキストデータ作成のためのアノテーション実績は以下の通りです。予備も含め、当初の予定よりやや多く作成しました。

Step4での対応枚数に関する計画と実績

凡例　**赤字**：主な変更点

	計画	実績
学習用データセット	• 対象画像：約10,000枚 　• 各カテゴリ(全34区分)約300枚毎の対応 • 構造化テキストデータの作成 • 全ての画像に対してInspection(二次チェック)実施 　• 一次作業：約10,000枚 　• 二次作業：約10,000枚	• 対象画像：**10,450**枚 　• 各カテゴリ(全34区分)約300枚毎の対応 • 構造化テキストデータの作成 • 全ての画像に対してInspection(二次チェック)実施 　• 一次作業：**10,510**枚 　• 二次作業：**10,450**枚
性能検査用テキストデータ	• 対象画像：3,300枚 • 各カテゴリ(全33区分)100枚毎の対応 　• 非構造化テキストデータの作成 • 全ての画像に対してInspection(二次チェック)実施 　• 一次作業：3,300枚 　• 二次作業：3,300枚	• 計画と同様

区分毎の枚数の内訳は、「5.参考資料」の①②をご参照ください。

全文テキスト化
アノテーション(Step4)　文字コード

計画時は文字コードを指定させていただきましたが、実際のアノテーションでは、文字コードに固執せず、アノテーターが読み取れた文字は原則読み取れた形のまま入力しました。

Step4での文字コードに関する計画と実績

凡例　**赤字**：主な変更点

	計画	実績
文字コード	• 文字コード「JIS X 0213」の一部文字種除くすべての文字を入力の対象とします • JIS第1・第2水準の漢字を網羅する形でアノテーションを進行します	• 文字コードに固執せず、**アノテーターが読み取れた文字は、原則読み取れた形のまま入力**しました • 理由は下記の２点です 　• 当社OCR対象外文字については、学習した場合も推論時に出力されないことから、入力時に存在していても学習に大きく影響がないため 　• 作業時に文字種を確認して入力の対象外であるかどうかの判断に時間がかかってしまうため

全文テキスト化
アノテーション(Step4)　入力対象文字種

計画時から大きな変更はありませんでしたが、Unicodeでも定義されていない変体仮名や合字は入力不可能であるため入力しませんでした。

Step4での入力文字種に関する計画と実績

凡例　**赤字**：主な変更点

入力対象文字種	計画(例)	実績(例)
英数字	0123ABCabc	計画と同様
記号	！？。、〃゜ゝゞ／＼	
ひらがな	あぁいぃうぅゐゑ	あぁいぃうぅゐゑ **※Unicodeでも定義されていない変体仮名や合字は入力不可能であるため入力していません。**
カタカナ	アァイィウゥヰヱ	アァイィウゥヰヱ **※Unicodeでも定義されていない変体仮名や合字は入力不可能であるため入力していません。**
囲み英数字/ローマ数字	①②③ Ⅰ Ⅱ Ⅲ	計画と同様
単位記号	㎜㎝㎞㎡	
囲み文字	㊤㊥㊦㊧㊨	
外字	✓ ★ ⌫ ↑ ‹ ›	
省略文字/年号	No.㈱㍾㍽㍼㍻	
漢字	すべての漢字が対象	

全文テキスト化
アノテーション(Step4)　入力対象外文字

音楽記号や印影・意匠につきましては、アノテーションを進めていく過程で課題となり、貴館と協議の上で入力対象外としました。

Step4での入力対象外文字に関する計画と実績

凡例　**赤字**：主な変更点

入力対象外	計画(例)	実績(例)
ギリシャ文字	ΔΘΣΦΨΩ	ΔΘΣΦΨΩ **※ただし主に理系の書籍にて記号として使用されるものは文章中に単一で出現する可能性もあり、入力しているものもあります。**
ロシア文字	БГДЁЖЗЙ	計画と同様
罫線素片	─ ┌ ┌┐ ┘ └ ├┼┤	
数学記号	∑∫∮√∠∟⊿∴∩∪	
音楽記号	（定義なし）	𝄞♪♮**Andante** **※Unicodeでも定義されていないものも多く、五線譜上の配置や形が変化するものもあり、アノテーションでの表現が困難、また主にイタリア語で書かれる演奏記号等はアルファベットで構成されており改めて入力を行わずともOCR可能であるため音楽記号はすべて対象外**
印影・意匠	（定義なし）	**字体が特殊である場合もあり、図等との区別もつきにくいため**

全文テキスト化
アノテーション(Step4)　追加対応方針

以下４点につきましては、アノテーションの進捗にあわせて出てきた問題点等に関して貴館と協議の上、方針として追加させていただきました。

No	追加となった対応方針	例
1	書籍の特定のページが貴館のデジタル化作業時点で欠損しており画像にその旨が記載されているものはアノテーション対象外	欠 MISSING
2	書籍の外表紙・裏表紙は(本文ではないので)対象外	外表紙・裏表紙のタイトル等
3	主に書籍外の最初や最後に添えられている管理用の文字は対象外	始、終、管理番号、バーコード等
4	手書加筆の中で、明らかに出版者の記述したものではない落書きやメモ等は対象外	落書き、アンダーライン等

具体的な例は、「5.参考資料」の③をご参照ください。

計画時の1次改善を第１回性能検査に向けて、2次・3次改善を第２回性能検査に向けて実施しました。

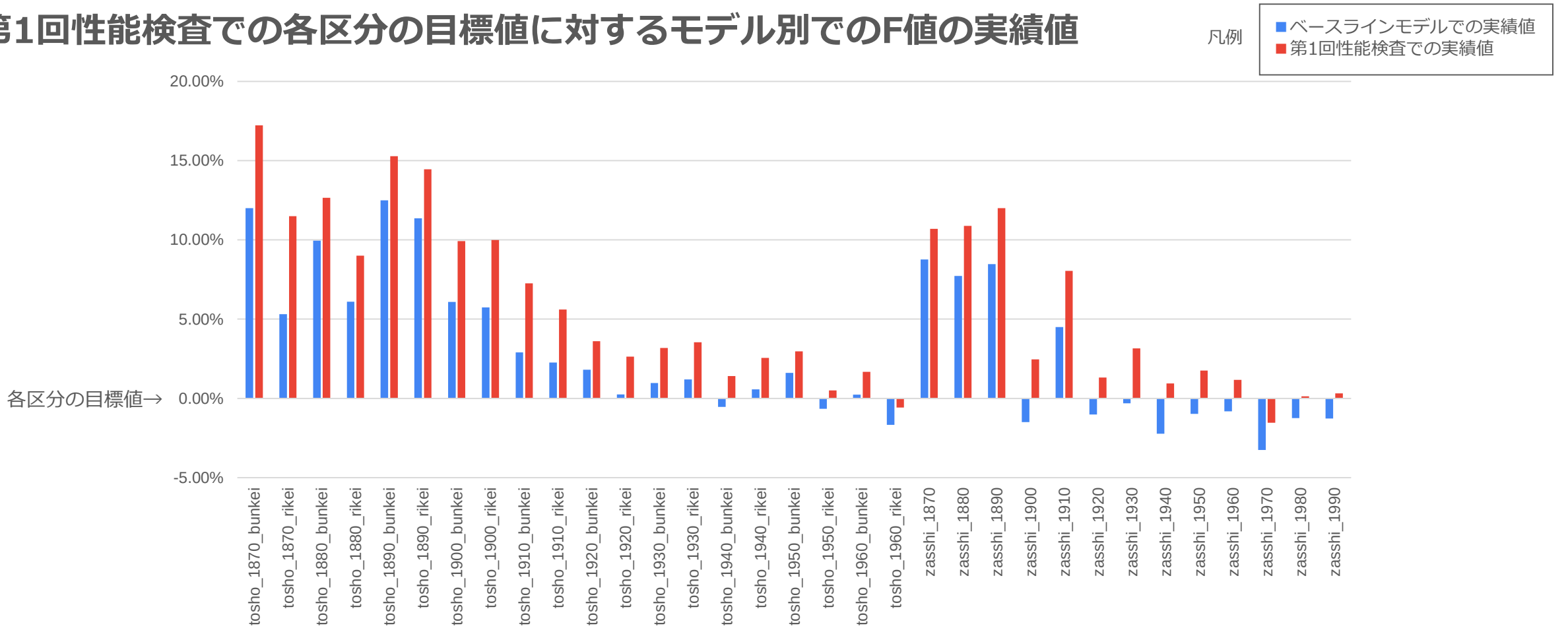
性能検査に向けて実施した改善策

凡例　取り消し線：不採用
※不採用の理由はp.24参照

		検出器	認識器
1次改善(7月)	**第1回性能検査**	• 特化モデル学習・評価 　• 約３千枚(30%)のアノテーションデータを追加で利用 • PoCで確認できた下記の問題点がどれほど解決できるのかを確認 　• 古い図書での縦・横書きの文字領域の検出精度が低い問題 　• 本文とルビ・割注が一つの文字領域として検出される問題 　• 図や写真内に存在する認識対象外の文字領域が検出される問題など • エラー分析を通じて精度向上ができる改善点を探索	~~• 古い図書への文字認識精度の向上のために使えるテキストデータの調査・収集~~ 　~~• 著作権問題がないデータ（例、青空文庫など）の活用~~ • 認識器学習用の合成データ生成 　• 文書データに合うスタイルの合成データ生成オプションの実験
２次改善(8月)	**第2回性能検査**	• ルビ・割注などを本文から分離する手法の実装 • １次改善段階で確認された改善策の適用 ~~• 特化モデルの学習・評価~~ 　~~• 約６千枚(60%)のアノテーションデータを追加で利用~~ • エラー分析を通じて精度向上ができる改善点を探索	~~• 特化モデルの学習・評価~~ 　~~• 合成データ及びリアルデータ（約６千枚のアノテーションデータから得られた分）を追加で利用~~ • エラー分析を通じて精度向上ができる改善点を探索
3次改善(9月)		• 2次改善段階で確認された改善策の適用 • 特化モデルの学習・評価 　• 約１万枚(100%)のアノテーションデータを追加で利用 　• Hyper-parameter チューニング適用 • Testデータへの性能検証 • 今回では対処できなかった問題点をレポート	• 2次改善段階で確認された改善策の適用 • 特化モデルの学習・評価 　• 合成データトラルデータ（100%）を追加で利用 　• Hyper-parameter チューニング適用 • Testデータへの性能検証 • 今回では対処できなかった問題点をレポート

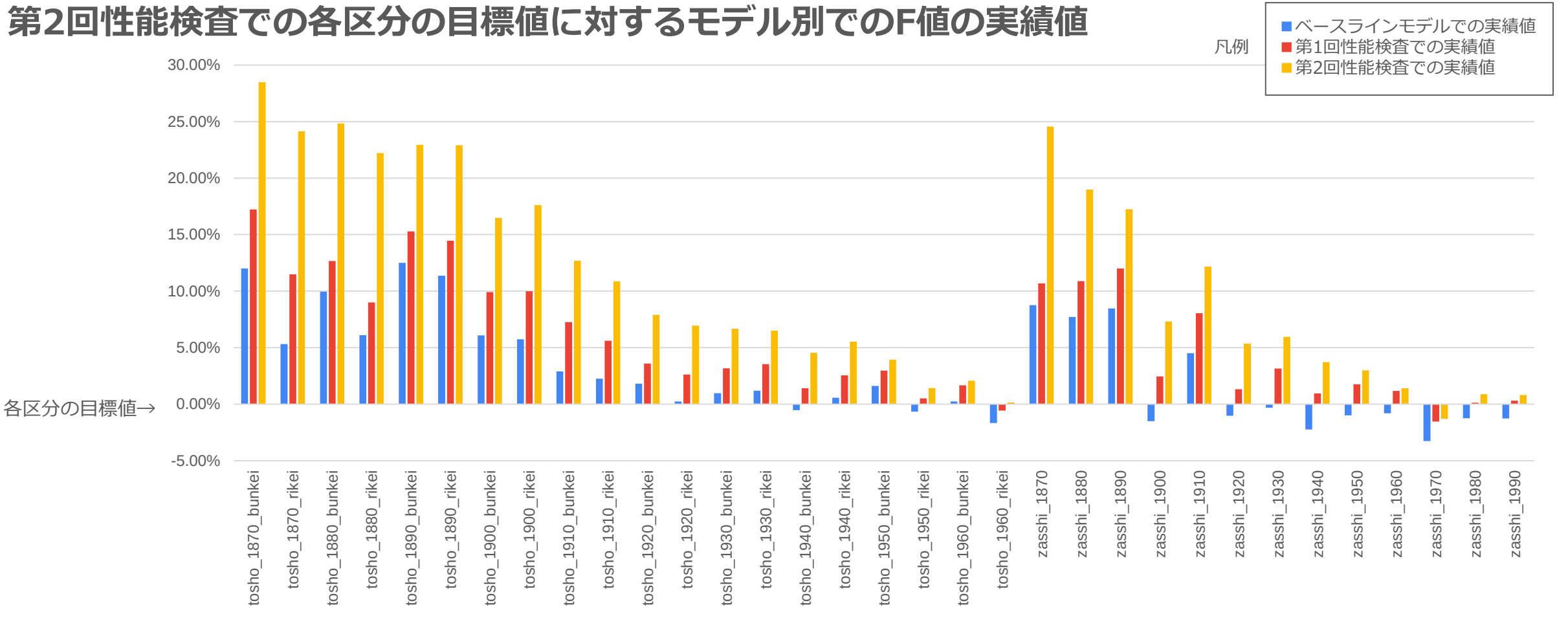
改善策の結果、第１回性能検査では31/33区分でF値の目標値を上回りました。
特に1910年以前の古い図書・雑誌については、概ね3%以上のF値の改善が見られました。

第1回性能検査での各区分の目標値に対するモデル別でのF値の実績値



改善策の結果、第2回性能検査では32/33区分でF値の目標値を上回りました。
特に1880年以前の古い図書・雑誌については、概ね10%以上の大きなF値の改善が見られました。
また、第1回では達成できなかった1960年代の理系図書についても、目標値を達成しました。

第2回性能検査での各区分の目標値に対するモデル別でのF値の実績値



ベースラインモデルでの誤検出・誤認識の傾向は主に下記３点でした。要因としては、**学習で利用したのは現代の書籍**であり、本件の対象である**幅広い年代や多様な資料を含んでいないため**と考えています。本章では、弊社のベースラインモデルと、本件で開発した特化モデルのOCR出力結果の違いを検出器・認識器それぞれで定性的に分析し、改善した部分をご説明します。

ベースラインモデルで誤検出・誤認識であった場合の画像の傾向

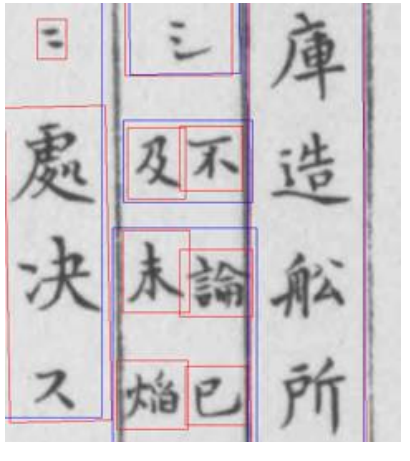
① 文字サイズ・字間・行間が現代の書籍と異なる

② 画像にノイズが多く含まれている

③ 言葉の使い方が違う

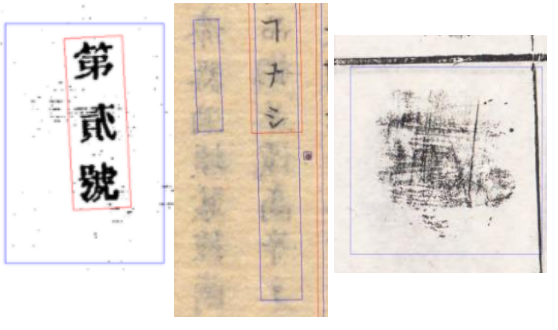
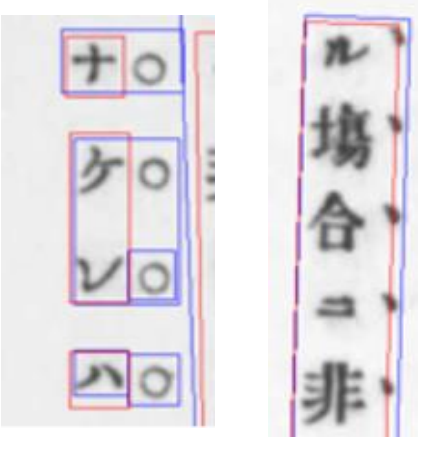
全文テキスト化
モデル学習、検出器・認識器それぞれのチューニング(Step5)　定性分析　検出器(1/2)

凡例
- ベースラインモデル
- 特化モデル

	検出結果	改善前(ベースラインモデル)	改善後(特化モデル)
① 文字サイズ・字間・行間が現代の書籍と異なる		**OCR結果の傾向** 左図の検出結果の例では、文字の半分以上が検出できておらず、検出できた文字列も横書きと誤検出していました。 **原因** ベースラインモデルの学習で利用したデータは現代の書籍・文書が多く、横書きの文書かつ行間も確保されているものがほとんどでした。しかし、古い年代の書籍は左図の様に行間がほぼない縦書きの文書なども多いため、誤検出が発生していたと考えています。	特化モデルの場合、行間がほぼない文書についてもきちんと縦書きと認識し、かつ文字列もほぼ100%正しく検出できています。
		**OCR結果の傾向** 古い年代の書籍は、左図のように１行の中に小さい文字で２行を書く割注が存在します。 ベースラインモデルは、これらの2行の文字列を多くの場合１行の文字列として誤検出していました。 **原因** 学習データにこのようなパターンのデータが少なかったため、小さい文字を適切に扱うための必要なチューニング(例:入力イメージの解像度など)がされてないため、と考えています。	特化モデルの場合、このようなケースでも1行にまとめてではなく2行でそれぞれ別々に正しく文字列を検出することができています。

全文テキスト化
モデル学習、検出器・認識器それぞれのチューニング(Step5)　定性分析　検出器(2/2)

凡例
- ベースラインモデル
- 特化モデル

	検出結果	改善前(ベースラインモデル)	改善後(特化モデル)
② 画像にノイズが多く含まれている		**OCR結果の傾向** 書籍内の汚れ・ノイズ・透かして見えている文字などを文字列の一部と誤検出していました。 **原因** 過去の書籍・文書は、現代の資料と比べて画像にノイズが含まれているケースが多いですが、ベースラインモデルの学習で利用している書籍・文書データでは、このようなノイズはほとんど含まれていなかったため、と考えています。	特化モデルの場合、これらのノイズを無視するように学習しているため、誤検出が改善されました。
③ 言葉の使い方が違う		**OCR結果の傾向** OCR対象ではない文字または記号が検出結果に混入していました。 **原因** 他の事例と同様、現代の書籍・文書にあまり含まれていない文字・記号であり、学習されていなかったため、と考えています。	特化モデルの場合、アノテーション時にこのような部分を除外しているため、検出ボックスからうまく除外できています。

全文テキスト化
モデル学習、検出器・認識器それぞれのチューニング(Step5)　定性分析　認識器

凡例

青字	誤認識
=	判読不可文字

③ 言葉の使い方が違う

認識結果

元画像	正解	ﾍﾞｰｽﾗｲﾝ	特化
	防ク┐ヲ得ヘシ二人以上二至テ八大概子防禦ノ力及ハ	防ク**コチ**得ヘシ二人以上**‖**至テ**六**大**概予**防禦ノ力及	防ク**＝**ヲ得ヘシ二人以上二至テ八大**＝**子防禦ノ力及ハ

『仏蘭西刑法講義. 第11-20巻』
https://dl.ndl.go.jp/info:ndljp/pid/794222/143

改善前(ベースラインモデル)

OCR結果の傾向
画像にノイズが少なく文字も綺麗な場合でも、認識精度で低い場合がありました。

原因
ベースラインモデルの学習に使われたデータ（現代言語のテキスト）ではほぼ出現しない、傾向が大きく異なる文章・文字の使い方のため、モデルの認識性能が低下していると考えています。

改善後(特化モデル)

古い年代の書籍を学習させることで、正しく認識できるようになりました。

全文テキスト化
モデル学習、検出器・認識器それぞれのチューニング(Step5)　本文と非本文の分離機能

本文と非本文の分類は各画像に対するOCR処理結果を統計的に処理して本文と思われる文字サイズを推定した後に実施しています。また、分類する際の閾値については、**本文を非本文に分類しない**よう、保守的な値を設定しています。

本文・非本文分類のアルゴリズム詳細

1. OCR結果のテキストボックスに対して文字数と文字サイズを推定
- 文字数　　：OCR結果で得られた文字数
- 文字サイズ　：長辺の長さ/文字数

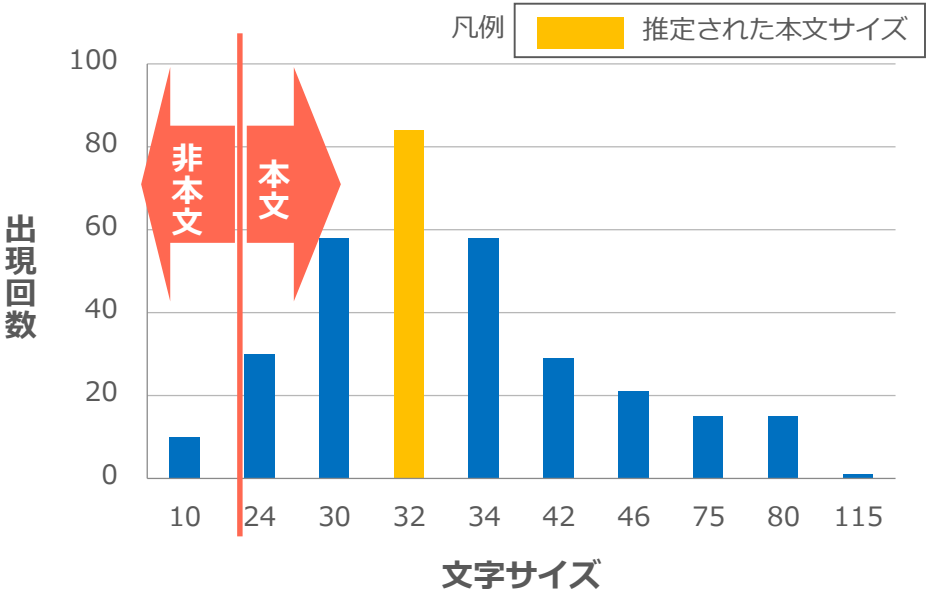
2. K-meansクラスタリングを実施
- 入力値　　：全てのテキストボックスから得られた文字サイズ(一つのテキストボックスに文字数分存在)と出現回数
- クラスタ数　：10(検証により決定)

3. 本文の文字サイズを推定
- 一番サンプルを多く含むクラスタの中心値(centroid)を利用

4. 本文・非本文を分類
- 本文の文字サイズのR%より小さい文字を含むテキストボックスを非本文として分類
- R%は、**本文を非本文に分類しない**ため、保守的な数値として60%を利用

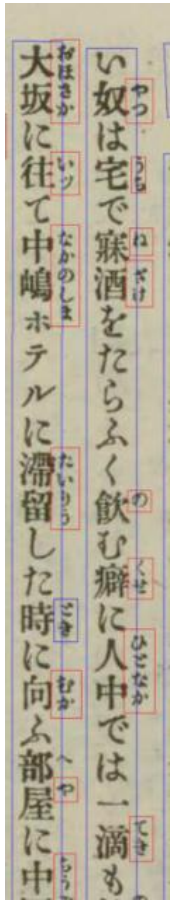
本文・非本文分類の例

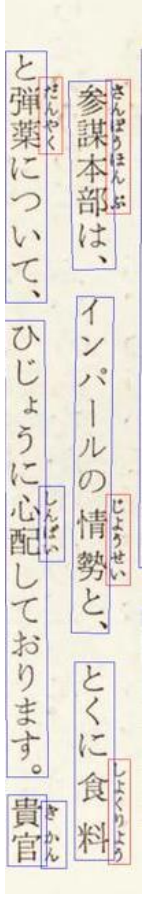
左記のNo.1~3を経て本文サイズが32と推定された場合、その60%である19.2を境界として本文・非本文を分類
- 19.2以上　→　本文
- 19.2未満　→　非本文



本文と非本文の分離の例を下記に示します。多くの非本文(ルビ)をきちんと特定できており、
かつ本文を誤って非本文と分類していないことがご確認いただけます。

本文・非本文分類の例





凡例
本文
非本文

全文テキスト化
モデル学習、検出器・認識器それぞれのチューニング(Step5)　採用しなかった改善策

改善計画時に予定していたが採用しなかった改善策は、主に下記の3つです。ただ、予定されたスケジュール通りに性能検査を合格できており、性能への影響はなかったと考えております。

No	分類	改善計画の方針	採用しなかった理由
1	検出器	約６千枚(60%)のアノテーションデータを利用しての特化モデルの学習・評価	本文・非本文分離器の開発とテストに予定より多くの工数を消費してしまい、内部的な学習・評価サイクルを全３回から全２回へ減らしたため。
2	認識器	約６千枚(60%)のアノテーションデータを利用しての特化モデルの学習・評価	
3		古い図書への文字認識精度の向上のために使えるテキストデータの調査・収集、著作権問題がないデータ（例、青空文庫など）の活用	明確に著作権問題がない古い図書を収集することが困難だったため。

全文テキスト化
右読みの文章への対応

計画書に記載の通り、アノテーション・OCR処理プログラムでのテキスト化までは左から読むそのままの順番通りで対応しました。その後、後処理の工程で右読みか左読みかどうかを行単位で判定するようにし、構造化データでは右読みである旨のタグ付け、非構造化データでは文字列を反転して結果を出力しました。

アノテーション → **テキスト化** → **後処理**

右読みの文章については、
左から読む順番通りに文字入力

左から読む順番通りに文字出力

横書きの文字列について
右読み・左読みのどちらが
適切かを判定(詳細は次項参照)

『少年保護事業と宗教』
https://dl.ndl.go.jp/info:ndljp/pid/1455587/29

```
{
    "id": 2,
    “boundingBox”: [～],
    “text”: “スピルカ",
    "confidence": 0.9862
},
```

◆構造化テキストデータ

```
{    ～
    “text”: “スピルカ",
    "isRTL": true,
},
```

◆非構造化テキストデータ

カルピス

全文テキスト化
右読みの文章への対応　具体的な判定方法

方針

- 文字列が短い場合には誤った確度となることが多いため、短い文字列での判定を避ける目的で、テキストボックス単位ではなく行単位での判定結果を用いて最終判定します。
- 同一行に縦書きが含まれる場合に意味の繋がる単位で判定をするため、行単位での判定では隣接する横書きのテキストボックス毎にテキストデータを結合して使用します。
- 右横書きを検知する方針とし、誤検知を避けることを目指します。
    - 文字列が短い場合には誤った確度となることが多いため、最終判定において行単位の結合した文字列が3文字以下の場合は確度によらず左横書きとして扱います。

判定方法

- 下表の判定方法に従い処理を行います。

判定単位/結果	Key	判定方法
テキスト ボックス単位	isBoxRTL	言語モデルを用いて各書字方向の判定確度を求め、確度の高い方向を判定結果とします。 確度が一致する場合は左横書きとして扱います。 （最終判定には使用しませんが、参考情報として出力します。）
行単位	isRowRTL	隣接する横書きのテキストボックスのテキストデータを結合し、判定を行います。 言語モデルを用いて各書字方向の判定確度を求め、確度の高い方向を判定結果とします。 確度が一致する場合は左横書きとして扱います。
最終判定結果	isRTL	行単位の判定結果を最終判定結果とします。 ただし、結合した文字列が3文字以下の場合は、確度によらず左横書きとして扱います。

レイアウト情報付与
改善結果 - 概要

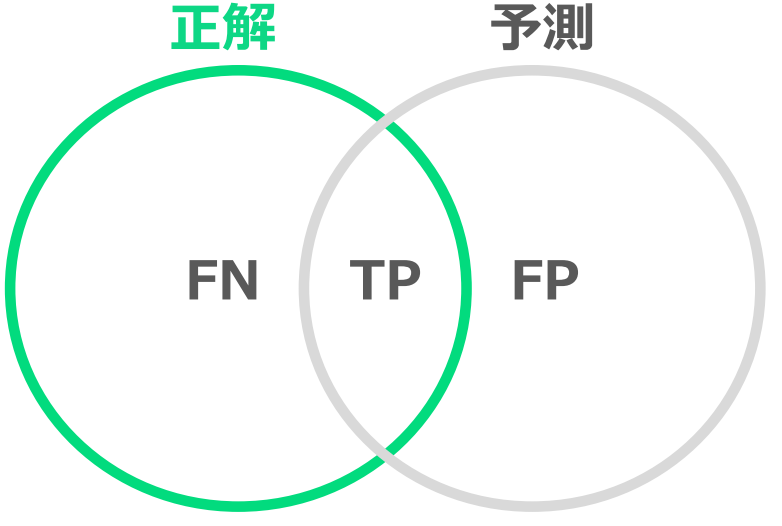
全文テキスト化モデルが検出したテキストにレイアウト情報付与できるよう、ドキュメント画像を入力するとPixel単位でレイアウト属性を推論するSemantic Segmentationモデルを開発しました。Semantic Segmentationの研究領域でデファクトスタンダードとなっていたDeepLab v3+をベースラインとして、**9point**のACC向上を達成しました。

評価データセットによるカテゴリ別のACC

項目	DeepLab v3+	開発したモデル
見出し	70.23	80.96
キャプション	48.43	54.59
注釈	43.00	60.92
ノンブル	70.78	95.04
柱	68.18	86.06
本文	90.66	98.47
画像・図など*	87.42	96.92
全項目の平均	71.95	**81.78**

* 主に画像・図・表・グラフ

ACC（Accuracy）= TP / FN+TP



レイアウト情報付与の定量的な評価方法は、
貴館と協議の上、ACCで測定しました

貴館より受領したデータを用いて、有効性の確認と並行して下記の改善策を実施しました。

実施した改善策

手法の選定	データ前処理と学習パラメタの調整

• 下記条件に基づいた有望手法の抽出
  • 実用性観点からパラメタ数と処理速度を考慮した手法であること
  • ソースコードが公開されていて再現性があること
  • ベンチマークデータセットでDeepLab v3+より優位な結果が報告されていること
• 比較検証による手法の選定
  • 比較対象：DeepLab v3+、SwinTransformer、SegFormer
• NDL様のデータセットで学習と評価を実施
  • 性能と速度の両面で優位性を示したSegFormerを採用
	• 背景領域を学習から除外することの有効性検証
	  • 除外する方針に決定
	• 損失関数"Cross Entropy Loss"と"Focal Loss"の比較検証
	  • "Focal Loss"を採用する方針に決定
	• 学習時にCROPによるデータ拡張の有効性検証
	  • CROPする方針に決定
	• 学習時にFLIPによるデータ拡張の有効性検証
	  • FLIPしない方針に決定

レイアウト情報付与
手法の選定

有望手法として抽出したDeepLab v3+、SwinTransformer、SegFormerを比較検証しました。
僅差ではありますが、優位性を示した**SegFormerを採用**しました。

候補として抽出された手法

DeepLab v3+
- atrousという畳み込みが特徴的なCNNベースのモデル
- 2018年にGoogleが提案
- https://arxiv.org/abs/1802.02611

SwinTransformer
- CNN層をTransformerに置き換える手法を提案したモデル
- 2021年にMicrosoft Research Asiaが提案
- https://arxiv.org/abs/2103.14030

SegFormer
- Semantic Segmentationに特化したTransformerモデル
- 2021年に香港大学やNVIDIAらが共同研究で発表
- https://arxiv.org/abs/2105.15203

比較実験の結果（ACC）

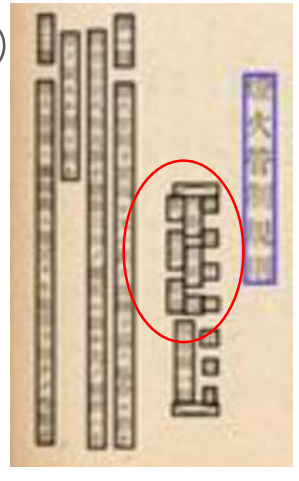
項目	DeepLab v3+ (type: r101) (param:63M)	SwinTransformer (type: Base) (param:121M)	SegFormer (size:B5) (param:85M)
見出し	75.76	82.62	77.19
キャプション	60.42	57.54	57.00
注釈	44.82	40.97	56.04
ノンブル	93.71	92.64	94.16
柱	88.06	85.82	87.71
本文	98.44	98.35	98.26
画像・図など	96.80	96.84	97.53
全項目の平均	79.72	79.25	**81.13**

レイアウト情報付与
背景領域を学習から除外することの有効性検証

全文テキスト化モデルでテキスト領域を検出済みという前提を利用して、背景領域を無視しながら前傾領域のみ学習させることで、見逃しリスクを軽減した上で、性能改善を施しました。

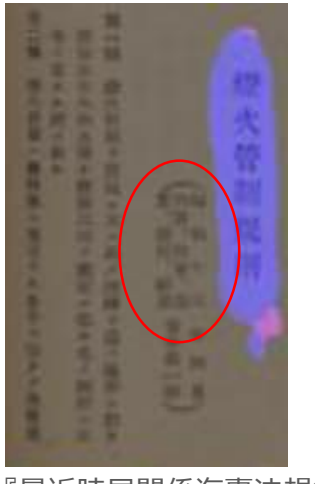
学習時に背景領域を考慮した場合としない場合の比較

①



②

③



『最近時局関係海事法規集』
https://dl.ndl.go.jp/info:ndljp/pid/1262569/119

左から順に①正解、②背景考慮した場合の推論結果、③背景無視した場合の推論結果。青が見出し、黒が本文、無色が背景領域。

テキスト領域を背景領域だと誤認識する"見逃しリスク"が減るため、テキスト領域を別途検出する前提ならば背景無視の学習の方が良い。

比較実験の結果（ACC）

項目	背景を考慮	背景を無視
見出し	66.83	77.19
キャプション	44.78	57.00
注釈	41.28	56.04
ノンブル	70.27	94.16
柱	68.58	87.71
本文	88.08	98.26
画像・図など	88.07	97.53
全項目の平均	70.70	**81.13**

背景領域を学習から除外した場合、8クラス分類が7クラス分類になるので、ACC改善は必然とも言えます。そこで、異なる評価指標でも比較実験をして、背景除外の有効性を確認しました。

マージ後の性能を比較した結果

項目	背景を考慮	背景を無視
見出し	64%	69%
キャプション	81%	83%
注釈	14%	21%
ノンブル	82%	95%
柱	96%	98%
本文	97%	99%
画像・図など	97%	98%
全項目の平均	93%	**96%**

表の値
全文テキスト化モデルの検出結果に、
レイアウト情報付与モデルの検出結果をマージして、
boxに正しいレイアウト情報を付与できた割合。

対象データ
評価データセットのうち「1960年代以降の図書」に該当する74枚を対象に評価実施。

レイアウト情報付与
損失関数"Cross Entropy Loss"と"Focal Loss"の比較検証

登場頻度が少ないサンプルを重点的に学習できる**Focal Lossを採用**した結果、
標準的に利用されるCross Entropy Lossと比べて、特にキャプションや注釈で性能改善されました。

Cross Entropy Loss

- 深層学習で標準的に用いられる損失関数です。
- モデルが推論した各レイアウトクラスに属する確率分布と、正解データとの差分を算出します。

Focal Lossとは？

- モデルにとって認識が難しいサンプルほど大きくなる重みパラメタを、Cross Entropy Lossに掛け合わせます。
- 今回の開発では、比較的容易に認識できる本文が、サンプルが圧倒的に多いため重点的に学習される傾向にありました。
- Focal Lossにより、サンプル数が少ないキャプションや注釈の認識性能向上が期待できます。

比較実験の結果（ACC）

項目	Cross Entropy Loss	Focal Loss
見出し	81.83	82.57
キャプション	58.15	**60.03**
注釈	41.16	**46.67**
ノンブル	92.25	93.45
柱	83.77	84.82
本文	98.42	98.51
画像・図など	96.60	96.60
全項目の平均	78.88	**80.38**

レイアウト情報付与
データ拡張の有効性検証

一般的なデータ拡張手法であるCROPとFLIPのドキュメント画像における有効性を検証しました。
僅差ですが、ACCが大きかった**CROP ON、FLIP OFFを採用**しました。

部分切り出しするCROPの有効性検証結果

項目	CROP ON	CROP OFF
見出し	81.96	79.98
キャプション	57.23	51.17
注釈	41.07	39.56
ノンブル	92.6	94.34
柱	83.36	89.23
本文	98.34	98.57
画像・図など	96.65	96.71
全項目の平均	**78.74**	78.51

上下左右ランダム反転するFLIPの有効性検証結果

項目	FLIP ON	FLIP OFF
見出し	81.96	81.83
キャプション	57.23	58.13
注釈	41.07	41.16
ノンブル	92.60	92.25
柱	83.36	83.77
本文	98.34	98.42
画像・図など	96.65	96.60
全項目の平均	78.74	**78.88**

レイアウト情報付与
アノテーション　対応実績のサマリ

レイアウト解析のモデル性能向上のための学習用データセット、モデルの性能評価用テキストデータの作成のためのアノテーション実績は以下の通りです。予備も含め、当初の予定よりやや多く作成しました。

アノテーションでの対応枚数に関する計画と実績

凡例　**赤字**：主な変更点

	計画	実績
数量	約10,000件 └作成済みの全文テキスト化用学習データを使用 34区分の各カテゴリあたり約300件対応予定	**10,307**件 └作成済みの全文テキスト化用学習データを使用 34区分の各カテゴリあたり約300件対応 完成したデータを、学習用データ９割、 評価用データ1割合で分割 ・学習用データ: 9,287件 ・評価用データ: 1,020件
作業工程	アノテーションに加えて全画像に対してインスペクションを実施予定 • アノテーション　約10,000件 • インスペクション　約10,000件	アノテーションに加えて全画像に対してインスペクションを実施 • アノテーション　**10,354**件 • インスペクション　**10,353**件 • 最終結果　**10,307**件
出力形式	JSONフォーマット	計画と同様

レイアウト情報付与
アノテーション　対応方針

レイアウト情報付与のアノテーションに関しましては、特に計画時から変更なく実施しました。

アノテーションの対応方針に関する計画と実績

凡例　**赤字**：主な変更点

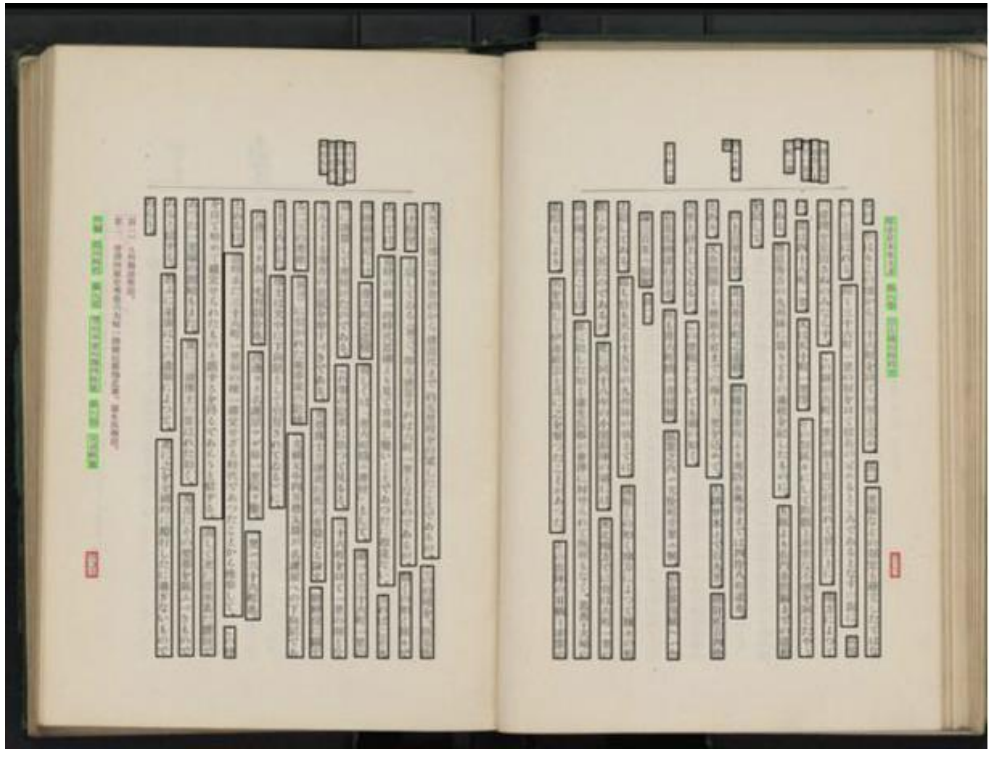
	計画	実績
対応方針	当初、対応方針として整理した項目* • 見出し・キャプションの取扱 • 注釈の取扱 • ノンブル・柱の取扱 • 本文の取扱 • 画像・図・表・グラフの取扱 • Holdの対象	計画と同様

レイアウト情報付与
正しくレイアウト情報付与できた例(1)

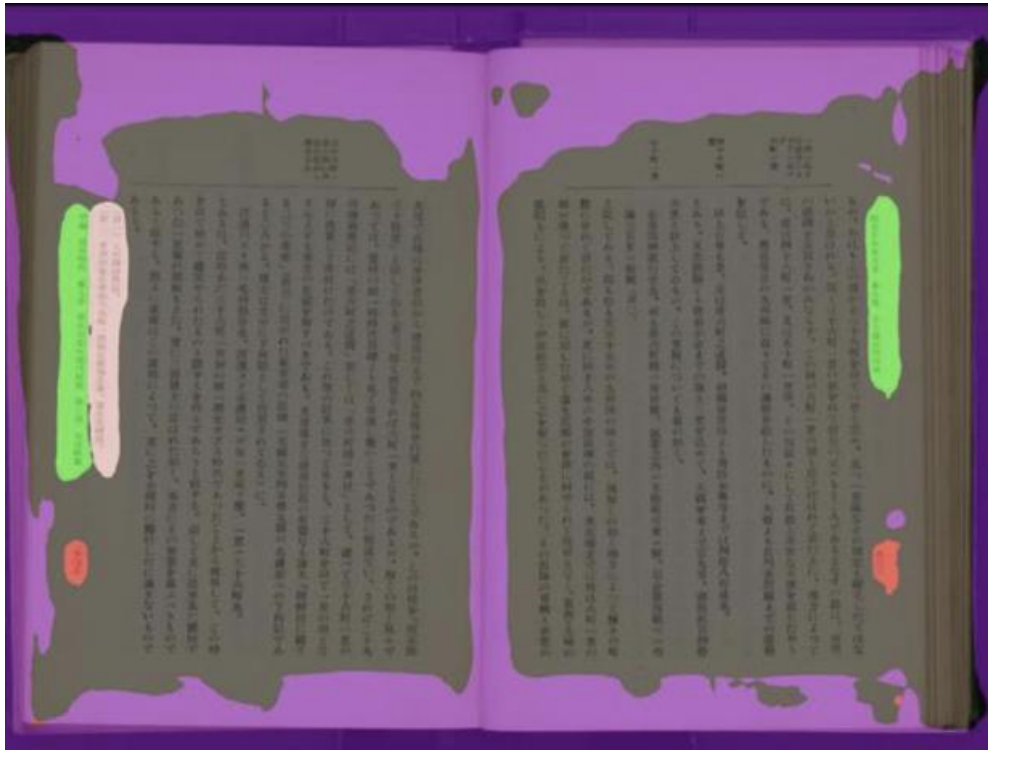
文字領域に対して、正解データと同じようにレイアウトラベルと推論できています。
余白にもいずれかのレイアウトラベルが必ず付与されるため、見た目は少々ノイジーになります。

凡例
- 見出し
- 柱
- 注釈
- 本文
- ノンブル
- 画像・図など

正解データ



レイアウト情報付与モデルの出力

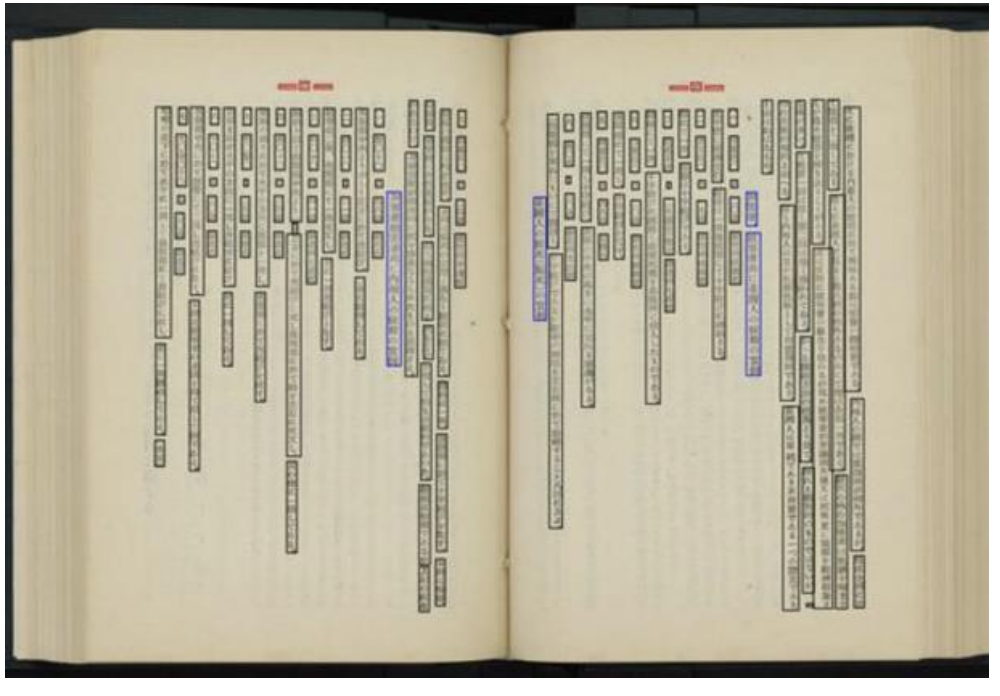


レイアウト情報付与
正しくレイアウト情報付与できた例(2)

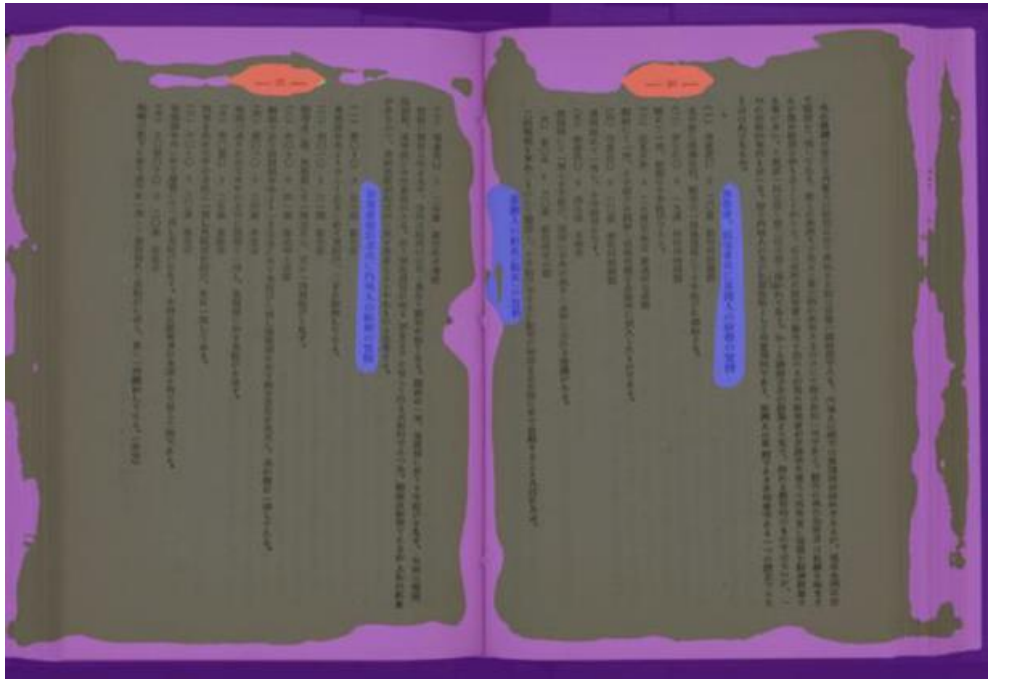
文字領域に対して、正解データと同じようにレイアウトラベルと推論できています。
余白にもいずれかのレイアウトラベルが必ず付与されるため、見た目は少々ノイジーになります。

凡例
- 見出し
- 柱
- 注釈
- 本文
- ノンブル
- 画像・図など

正解データ



レイアウト情報付与モデルの出力

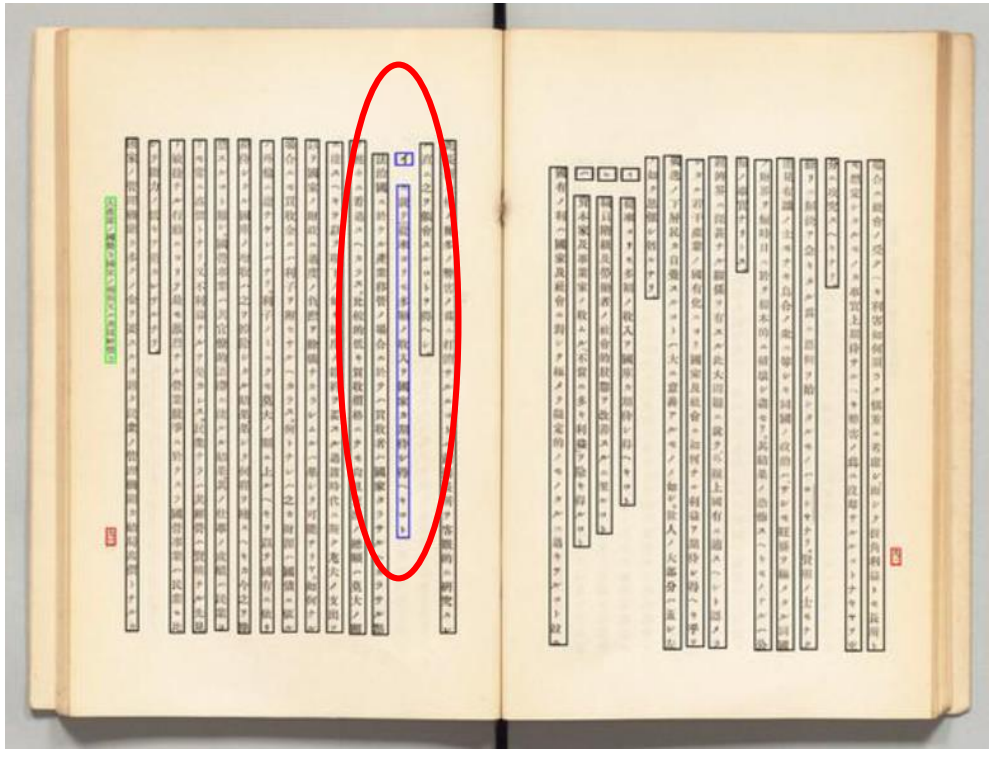


レイアウト情報付与
レイアウト情報付与を一部誤った例(1)

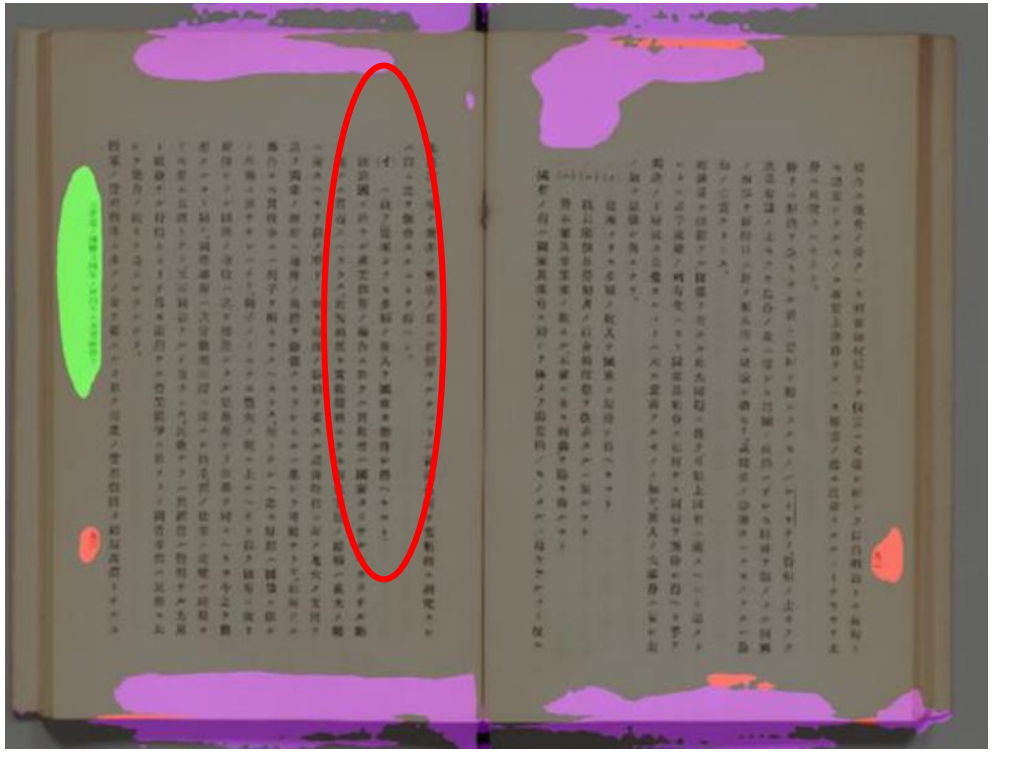
正解データでは見出し(青)としている箇所を、本文(黒)だと推論しています。

凡例
- 見出し
- 柱
- 注釈
- 本文
- ノンブル
- 画像・図など

正解データ



レイアウト情報付与モデルの出力

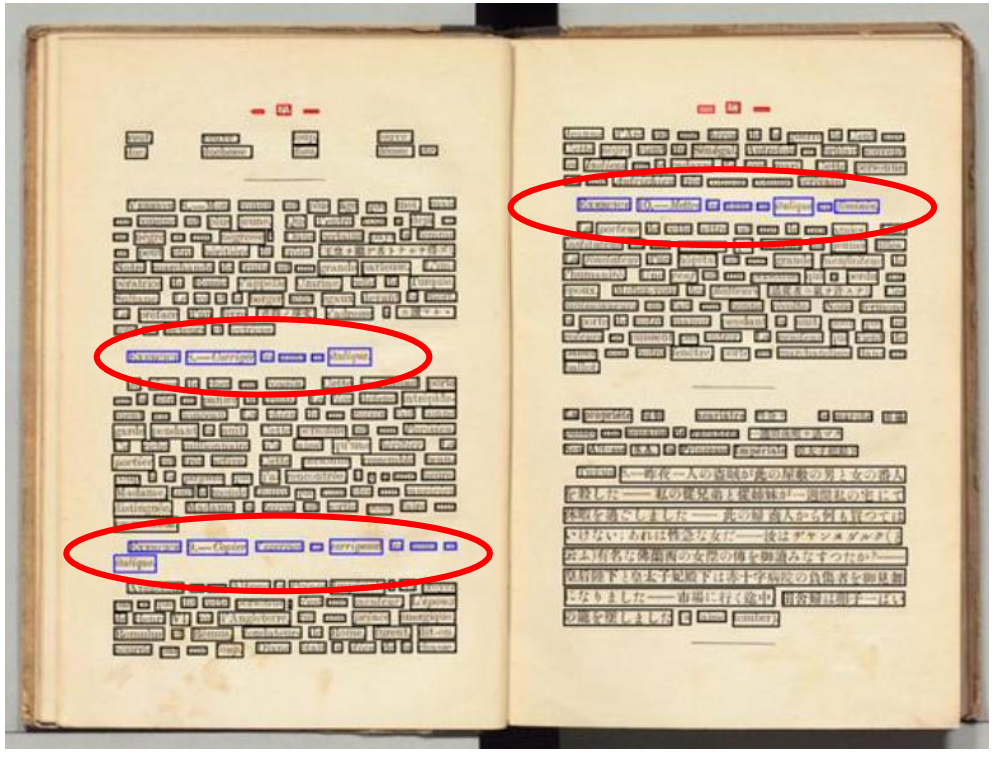


レイアウト情報付与
レイアウト情報付与を一部誤った例(2)

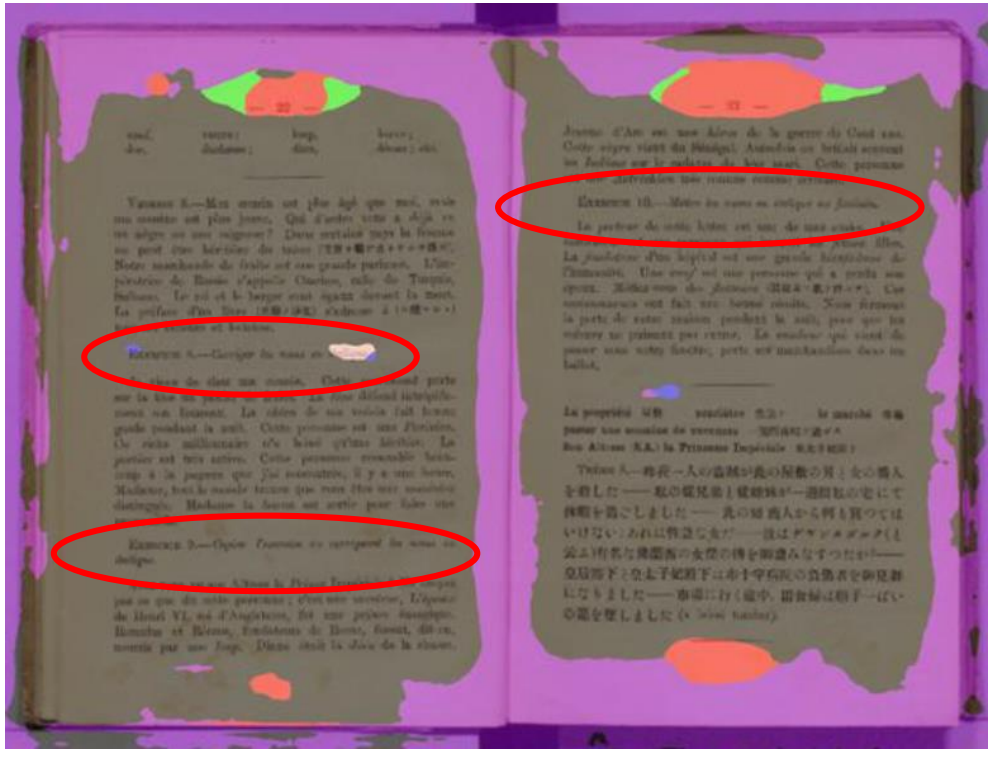
正解データでは見出し(青)としている3箇所を、いずれも本文(黒)だと推論しています。

凡例
- 見出し
- 柱
- 注釈
- 本文
- ノンブル
- 画像・図など

正解データ



レイアウト情報付与モデルの出力



アノテーション実績詳細(学習用データセット)

区分	Annotation	Inspection1	Inspection2
合計	10,510	10,510	**10,450**

#	区分	Annotation	Inspection1	Inspection2
1	other	308	308	301
2	tosho_1870_bunkei	317	317	317
3	tosho_1880_bunkei	316	316	311
4	tosho_1890_bunkei	326	326	323
5	tosho_1900_bunkei	302	302	301
6	tosho_1910_bunkei	321	321	318
7	tosho_1920_bunkei	312	312	312
8	tosho_1930_bunkei	301	301	301
9	tosho_1940_bunkei	300	300	300
10	tosho_1950_bunkei	305	305	300
11	tosho_1960_bunkei	304	304	304
12	tosho_1870_rikei	300	300	300
13	tosho_1880_rikei	301	301	301
14	tosho_1890_rikei	303	303	302
15	tosho_1900_rikei	322	322	319
16	tosho_1910_rikei	303	303	300
17	tosho_1920_rikei	305	305	300

#	区分	Annotation	Inspection1	Inspection2
18	tosho_1930_rikei	307	307	300
19	tosho_1940_rikei	317	317	314
20	tosho_1950_rikei	302	302	300
21	tosho_1960_rikei	315	315	312
22	zasshi_1870	305	305	305
23	zasshi_1880	308	308	305
24	zasshi_1890	320	320	318
25	zasshi_1900	322	322	320
26	zasshi_1910	307	307	307
27	zasshi_1920	311	311	310
28	zasshi_1930	303	303	303
29	zasshi_1940	314	314	314
30	zasshi_1950	314	314	314
31	zasshi_1960	307	307	306
32	zasshi_1970	303	303	303
33	zasshi_1980	304	304	304
34	zasshi_1990	305	305	305

アノテーション実績詳細(性能検査用テキストデータ)

区分	Annotation	Inspection1	Inspection2
合計	3,586	3,579	**3,300**

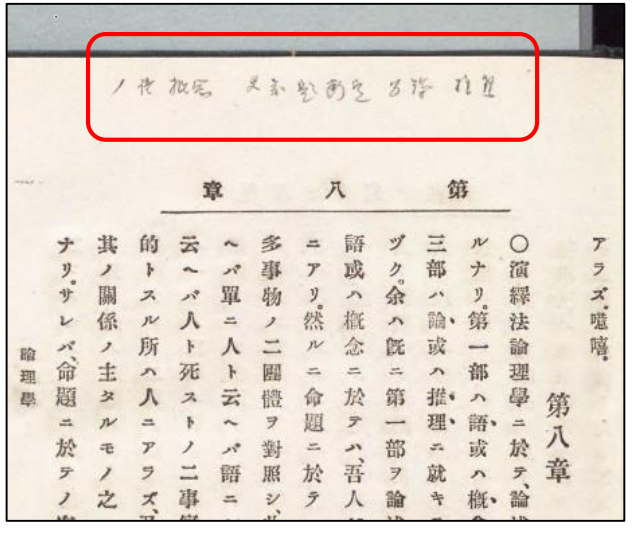
#	区分	Annotation	Inspection1	Inspection2
1	tosho_1870_bunkei	103	103	100
2	tosho_1880_bunkei	106	106	100
3	tosho_1890_bunkei	101	100	100
4	tosho_1900_bunkei	102	100	100
5	tosho_1910_bunkei	105	101	100
6	tosho_1920_bunkei	108	108	100
7	tosho_1930_bunkei	100	100	100
8	tosho_1940_bunkei	129	129	100
9	tosho_1950_bunkei	122	122	100
10	tosho_1960_bunkei	117	117	100
11	tosho_1870_rikei	130	130	100
12	tosho_1880_rikei	137	137	100
13	tosho_1890_rikei	116	116	100
14	tosho_1900_rikei	106	106	100
15	tosho_1910_rikei	100	100	100
16	tosho_1920_rikei	103	103	100
17	tosho_1930_rikei	109	109	100

#	区分	Annotation	Inspection1	Inspection2
18	tosho_1940_rikei	114	114	100
19	tosho_1950_rikei	102	102	100
20	tosho_1960_rikei	103	103	100
21	zasshi_1870	131	131	100
22	zasshi_1880	107	107	100
23	zasshi_1890	107	107	100
24	zasshi_1900	101	101	100
25	zasshi_1910	100	100	100
26	zasshi_1920	104	104	100
27	zasshi_1930	106	106	100
28	zasshi_1940	102	102	100
29	zasshi_1950	110	110	100
30	zasshi_1960	103	103	100
31	zasshi_1970	101	101	100
32	zasshi_1980	101	101	100
33	zasshi_1990	100	100	100

参考資料
③全文テキスト化　追加対応方針の具体例

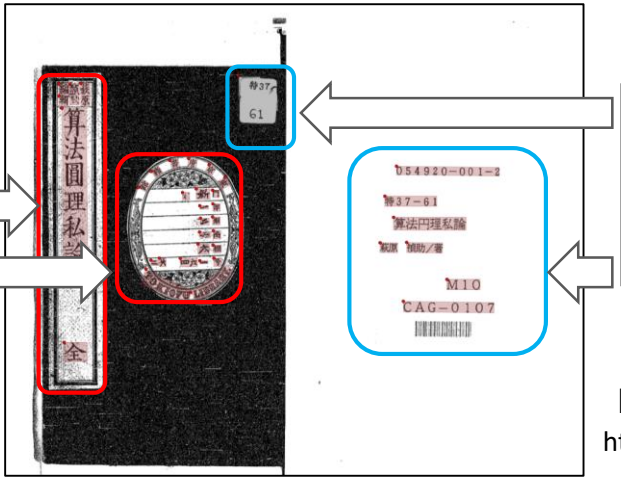
No.1
書籍・書籍画像が紛失しており画像にその旨が記載されているもの

No.4
手書加筆の中で、明らかに出版者の記述したものではない落書きやメモ等は対象外



『演繹歸納論理學 再版』
https://dl.ndl.go.jp/info:ndljp/pid/2937042

No.2
書籍の外表紙・裏表紙は(本文ではないので)対象外



No.3
主に書籍外の最初や最後に添えられている管理用の文字

『算法円理私論．〔本編〕』
https://dl.ndl.go.jp/info:ndljp/pid/829116

アノテーション実績詳細

区分	Annotation	Inspection1	Fix
合計	10,354	10,353	**10,307**

#	区分	Annotation	Inspection	Fix
1	other	298	298	298
2	tosho_1870_bunkei	313	313	311
3	tosho_1880_bunkei	307	307	307
4	tosho_1890_bunkei	320	320	319
5	tosho_1900_bunkei	298	298	298
6	tosho_1910_bunkei	318	318	317
7	tosho_1920_bunkei	312	312	311
8	tosho_1930_bunkei	301	301	299
9	tosho_1940_bunkei	299	299	298
10	tosho_1950_bunkei	296	296	294
11	tosho_1960_bunkei	300	300	300
12	tosho_1870_rikei	297	297	295
13	tosho_1880_rikei	296	296	296
14	tosho_1890_rikei	297	297	296
15	tosho_1900_rikei	314	314	312
16	tosho_1910_rikei	298	298	295
17	tosho_1920_rikei	298	298	297

#	区分	Annotation	Inspection	Fix
18	tosho_1930_rikei	300	300	300
19	tosho_1940_rikei	311	311	309
20	tosho_1950_rikei	297	297	294
21	tosho_1960_rikei	308	308	308
22	zasshi_1870	303	303	303
23	zasshi_1880	302	302	301
24	zasshi_1890	314	314	313
25	zasshi_1900	317	317	313
26	zasshi_1910	303	303	301
27	zasshi_1920	307	307	304
28	zasshi_1930	300	300	298
29	zasshi_1940	311	310	310
30	zasshi_1950	312	312	312
31	zasshi_1960	306	306	306
32	zasshi_1970	299	299	297
33	zasshi_1980	301	301	298
34	zasshi_1990	301	301	297

区分ごとの学習用データ・評価用データの件数

区分	学習用	評価用	総計
合計	**9,287**	**1,020**	**10,307**

#	区分	学習用	評価用	小計
1	other	268	30	**298**
2	tosho_1870_bunkei	281	30	**311**
3	tosho_1870_rikei	265	30	**295**
4	tosho_1880_bunkei	277	30	**307**
5	tosho_1880_rikei	266	30	**296**
6	tosho_1890_bunkei	289	30	**319**
7	tosho_1890_rikei	266	30	**296**
8	tosho_1900_bunkei	268	30	**298**
9	tosho_1900_rikei	282	30	**312**
10	tosho_1910_bunkei	287	30	**317**
11	tosho_1910_rikei	265	30	**295**
12	tosho_1920_bunkei	281	30	**311**
13	tosho_1920_rikei	267	30	**297**
14	tosho_1930_bunkei	269	30	**299**
15	tosho_1930_rikei	270	30	**300**
16	tosho_1940_bunkei	268	30	**298**
17	tosho_1940_rikei	279	30	**309**

#	区分	学習用	評価用	小計
18	tosho_1950_bunkei	264	30	**294**
19	tosho_1950_rikei	264	30	**294**
20	tosho_1960_bunkei	270	30	**300**
21	tosho_1960_rikei	278	30	**308**
22	zasshi_1870	273	30	**303**
23	zasshi_1880	271	30	**301**
24	zasshi_1890	283	30	**313**
25	zasshi_1900	283	30	**313**
26	zasshi_1910	271	30	**301**
27	zasshi_1920	274	30	**304**
28	zasshi_1930	268	30	**298**
29	zasshi_1940	280	30	**310**
30	zasshi_1950	282	30	**312**
31	zasshi_1960	276	30	**306**
32	zasshi_1970	267	30	**297**
33	zasshi_1980	268	30	**298**
34	zasshi_1990	267	30	**297**