補助事業成果報告書

補助事業の名称	機械学習を活用した非アクセシブルなPDF文書の構造化とテキスト抽出に関する研究開発
補助事業の概要	非アクセシブルなPDFから正しく成形された構造化テキストを抽出するために、レイアウト解析された版面の構成要素に機械学習モデルによって構造情報を付加した上で、独自のPDFテキスト抽出エンジンによって成形されたテキストを抽出する。これらの機能が統合されたクラウドサービスを開発する。

【研究開発の実施内容と成果】

令和４年度において本研究では主に、PDF の版面情報取得 API の改良、フォント DB の構築、レイアウト解析エンジンの切り替え、見出し判定モデルの作成、DaisyXML の出力対応、の４点について取り組んだ。

１.PDF の版面情報取得 API の改良

前年度は版面のレイアウト解析と画像版面のOCR 解析にGoogle Cloud Vision を利用し、解析された版面データから、PDF2MD によるテキスト抽出と機械学習モデルによる構造判定を行った。PDF2MD はCloud Vision から得たブロックの位置情報に基づいて、OCR よりも精度の高いテキストを得ることに成功した。一方、機械学習モデルによる構造判定はヘッダー、柱、ノンブルの判別ではある程度機能するものの、構造情報として重要な見出しの判定においては十分な精度を上げられなかった。これはGoogle Cloud Vision から得たパラメータだけでは判別に必要な特徴量が足りなかったためと考えられる。
そこで本年度は特徴量にフォント情報を追加するためにPDF2MD を改良し、版面のすべての文字の位置とフォントの種類およびサイズをページごとに出力するAPI を実装した。出力形式はGoogle Cloud Vision のJSONフォーマットを拡張する形式とした。また PDF ファイル全体のページ数とフォント情報を一括で取得するAPI も追加した。

冗長化を避けるため、フォント毎にidを付与し、それぞれの文字データは自身のフォントをid参照する形式とした。

各ページの解析データ

```
{
  "fullTextAnnotation": {
    "pages": [ // ページ
      {
        "width": 1414, // ページ幅
        "height": 2000, // ページ高さ
        "blocks": [ // ブロック
            "paragraphs": [ // パラグラフ
                "words": [ // 単語
                    "symbols": [ // 文字
                        "text": "あ", // 文字
                        "confidence": 1,
                        "font": { // フォント値
                          "id": 5, // フォント ID
                          "size": 10.56 // フォントサイズ(推定)
                        }
                    },
                    (略)
                ],
              }
            ],
          }
        ],
        "blockType": "TEXT",
      }
    ],
  }
],
"text": "あいうえお \n かきく..." // テキスト全文
"fonts": [ // フォント一覧
  {
    "id": 5, // フォント ID
    "name": "YuMincho-Regular", // 完全フォント名
    "isVertical": false, // 縦書フォントか
    "metadata": { // フォントメタデータ
```

```
                "name": "Yu Mincho Regular",
                "postscriptName": "YuMincho-Regular",
                "family": "Yu Mincho",
                "subfamily": "Regular",
                "format": "",
                "license": "Microsoft ...", //  ライセンス
                "copyright": "Copyright @ ..." //  コピーライト
              }
            },
            (略)
        ],
        "estimatedMainFont": { //  推定本文フォント
          "id": 5, //  フォント ID
          "size": 10.56 //  フォントサイズ(推定)
        }
    }
}
```


フォント DB の構築

フォントには OpenType や TrueType などさまざまなフォーマットがあり、含まれている情報の形式もばらばらであるため、そのままでは機械学習で扱いづらい。たとえば文字の太さやイタリック体などはフォントの重要な特徴であるが、統一的な表現形式があるわけではない。そこで正規化を行いデータベースに登録する機構を作成した。

以下はフォント名やサブファミリー名を手がかりとして、文字の太さを fontWeight という数値に正規化する変換ルールである。

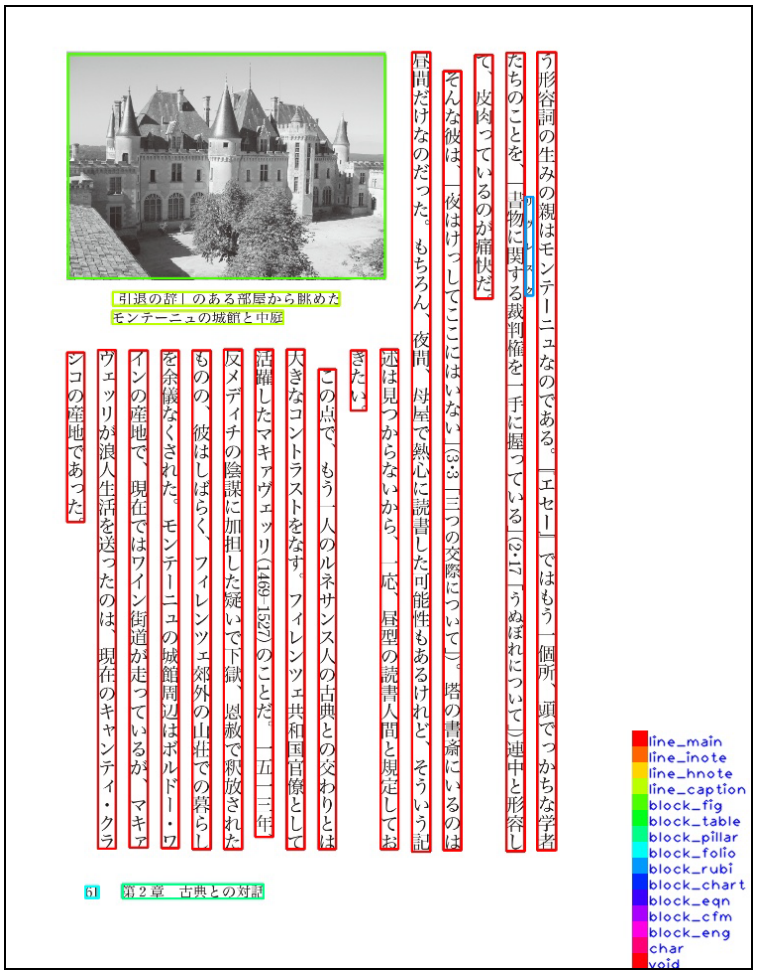
項目	値	fontWeight
サブファミリー名	Thin	100
サブファミリー名	ExtraLight	200
サブファミリー名	Light	300
サブファミリー名	Regular	400
サブファミリー名	Medium	500
サブファミリー名	SemiBold	600
サブファミリー名	Bold	700
サブファミリー名	ExtraBold	800

サブファミリー名	Black, Heavy	900
フォント名	-W0 を含む	100
フォント名	-W1 を含む	100
フォント名	-W2 を含む	200
フォント名	-W3 を含む	300
フォント名	-W4 を含む	400
フォント名	-W5 を含む	500
フォント名	-W6 を含む	600
フォント名	-W7 を含む	700
フォント名	-W8 を含む	800
フォント名	-W9 を含む	900

レイアウト解析エンジンの切り替え

レイアウト解析エンジンを Google Cloud Vision から NDL OCR に変更。そのため専用のレイアウト解析サーバーが必要となったためサーバー構築を行った。ブロック構造の推論精度が向上し、アクセシビリティの観点から不要な要素の除外が容易となった。（次図に例示）

図. NDLOCR によるレイアウト解析と構造判定



う形容詞の生みの親はモンテーニュなのである。『エセー』ではもう一個所、頭でっかちな学者たちのことを、「書物に関する裁判権を一手に握っている」(2・17「うぬぼれについて」)連中と形容し、皮肉っているのが痛快だ。

そんな彼は、一夜はけっしてここにはいない」(3・3「三つの交際について」)。塔の書斎にいるのは仕事だけなのだった。もちろん、夜間、小屋で熟心に読書した可能性もあるけれど、そういう記述は見つからないから、一応、昼型の読書人間と規定しておきたい。

この点で、もう一人のルネサンス人の古典との交わりとは大きなコントラストをなす。フィレンツェ共和国官僚として活躍したマキアヴェッリ(1469-1527)のことだ。一五一三年、反メディチの陰謀に加担した疑いで下獄、恩赦で釈放されたものの、彼はしばらく、フィレンツェ郊外の山荘での蟄居を余儀なくされた。モンテーニュの城館周辺はボルドー・ワインの産地で、現在ではワイン街道が走っているが、マキアヴェッリが浪人生活を送ったのは、現在のキャンティ・クラシコの産地であった。

<?xml version='1.0' encoding='utf-8'?>
<OCRDATASET><PAGE HEIGHT="2000" IMAGENAME="R0000001_pp.jpg" WIDTH="1236">
<LINE CONF="1.000" HEIGHT="982" TYPE="本文" WIDTH="36" X="431" Y="782" />
<LINE CONF="1.000" HEIGHT="977" TYPE="本文" WIDTH="34" X="308" Y="784" />
<LINE CONF="1.000" HEIGHT="332" TYPE="本文" WIDTH="33" X="122" Y="788" />
<LINE CONF="1.000" HEIGHT="982" TYPE="本文" WIDTH="34" X="555" Y="782" />
<LINE CONF="1.000" HEIGHT="942" TYPE="本文" WIDTH="33" X="618" Y="821" />
<LINE CONF="1.000" HEIGHT="958" TYPE="本文" WIDTH="34" X="493" Y="782" />
<LINE CONF="1.000" HEIGHT="478" TYPE="本文" WIDTH="35" X="926" Y="203" />
<LINE CONF="1.000" HEIGHT="1571" TYPE="本文" WIDTH="36" X="802" Y="198" />
<LINE CONF="1.000" HEIGHT="1530" TYPE="本文" WIDTH="35" X="864" Y="235" />
<LINE CONF="1.000" HEIGHT="1568" TYPE="本文" WIDTH="36" X="988" Y="199" />
<LINE CONF="1.000" HEIGHT="982" TYPE="本文" WIDTH="35" X="740" Y="782" />
<LINE CONF="0.999" HEIGHT="1567" TYPE="本文" WIDTH="36" X="1049" Y="198" />

```
<LINE CONF="1.000" HEIGHT="26" TYPE="キャプション" WIDTH="446" X="212" Y="669" />
<LINE CONF="1.000" HEIGHT="26" TYPE="キャプション" WIDTH="337" X="211" Y="706" />
<BLOCK CONF="1.000" HEIGHT="440" TYPE="図版" WIDTH="624" X="124" Y="202" />
<BLOCK CONF="0.992" HEIGHT="27" TYPE="柱" WIDTH="278" X="231" Y="1833" />
<BLOCK CONF="0.821" HEIGHT="23" TYPE="ノンブル" WIDTH="26" X="158" Y="1837" />
<BLOCK CONF="0.987" HEIGHT="194" TYPE="ルビ" WIDTH="15" X="1026" Y="482" />
```

ブロックの読み順が不規則に出力される問題があったため、位置情報から読み順を推論する機械学習モデルを作成してブロックの順序を並べ替える処理を行う必要があったが、並べ替え後でも Google Cloud Vision よりも読み順の精度は劣る結果となった。

見出し判定モデルの作成

レイアウト解析されたブロックの座標にフォント DB の情報を追加して、ブロックが見出しか本文のどちらかを判定する分類モデルを作成した。

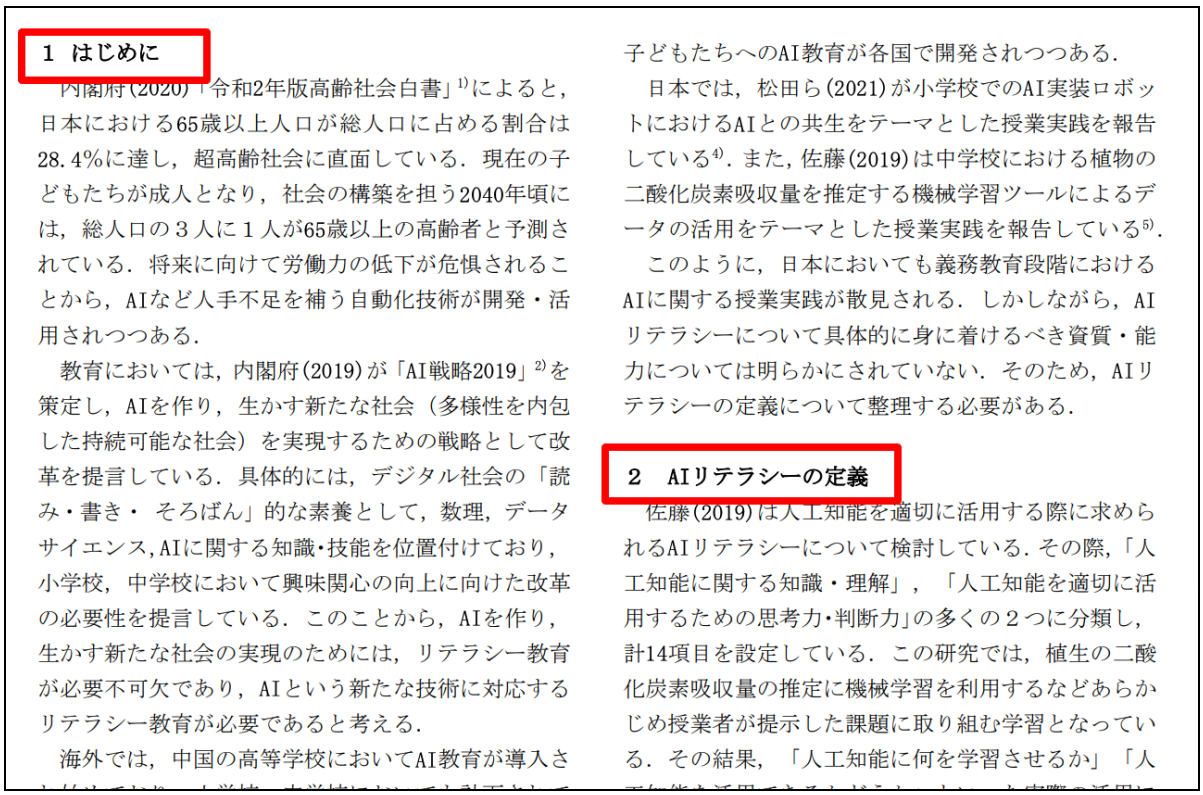
ブロックのデータ

パラメタ	説明
page_width	ページ幅
page_height	ページ高さ
block_type	ブロックの種類
block_left	ブロック左座標
block_top	ブロック上座標
block_right	ブロック右座標
block_bottom	ブロック下座標
block_width	ブロック幅
block_height	ブロック高さ
font_size	フォントサイズ
font_weight	フォントウェイト
font_isItalic	イタリック識別
main_font_size	本文フォントサイズ
main_font_weight	本文フォントウェイト
main_font_isItalic	本文イタリック識別

font_weightRatio	フォントウェイト比
font_scale	フォントサイズ比
label	本文／見出し

分類にはランダムフォレストのアルゴリズムを使用し、k-Fold 交差検証によって精度を検証した結果、83%程度の精度が得られた。これらの特徴量が、見出しの検出に有効であることが確認できたが、さらなる精度向上が必要である。

しかしこの手法では太字やイタリック体のグリフを持たないフォントを PDF の機能を利用して体裁を変えているケースを検出できない問題が判明した。以下の論文の版面は、見出しに本文よりも太いウェイトの文字が使われているように見える。しかしフォントが太字の字形（グリフ）を持っているのではなく、本文と同じ MS Mincho の文字を PDF の機能を利用して太く見せているにすぎない。今回の作成したモデルではこのような見出しを本文と区別することができなかった。解析の際に、こうしたケースを検出して PDF2MD の出力データに反映可能か調査・検討を行っている。



1 はじめに

内閣府(2020)「令和2年版高齢社会白書」[1]によると，日本における65歳以上人口が総人口に占める割合は28.4%に達し，超高齢社会に直面している．現在の子どもたちが成人となり，社会の構築を担う2040年頃には，総人口の３人に１人が65歳以上の高齢者と予測されている．将来に向けて労働力の低下が危惧されることから，AIなど人手不足を補う自動化技術が開発・活用されつつある．

教育においては，内閣府(2019)が「AI戦略2019」[2]を策定し，AIを作り，生かす新たな社会（多様性を内包した持続可能な社会）を実現するための戦略として改革を提言している．具体的には，デジタル社会の「読み・書き・ そろばん」的な素養として，数理，データサイエンス,AIに関する知識･技能を位置付けており，小学校，中学校において興味関心の向上に向けた改革の必要性を提言している．このことから，AIを作り，生かす新たな社会の実現のためには，リテラシー教育が必要不可欠であり，AIという新たな技術に対応するリテラシー教育が必要であると考える．

海外では，中国の高等学校においてAI教育が導入さ

子どもたちへのAI教育が各国で開発されつつある．

日本では，松田ら(2021)が小学校でのAI実装ロボットにおけるAIとの共生をテーマとした授業実践を報告している[4]．また，佐藤(2019)は中学校における植物の二酸化炭素吸収量を推定する機械学習ツールによるデータの活用をテーマとした授業実践を報告している[5]．

このように，日本においても義務教育段階におけるAIに関する授業実践が散見される．しかしながら，AIリテラシーについて具体的に身に着けるべき資質・能力については明らかにされていない．そのため，AIリテラシーの定義について整理する必要がある．

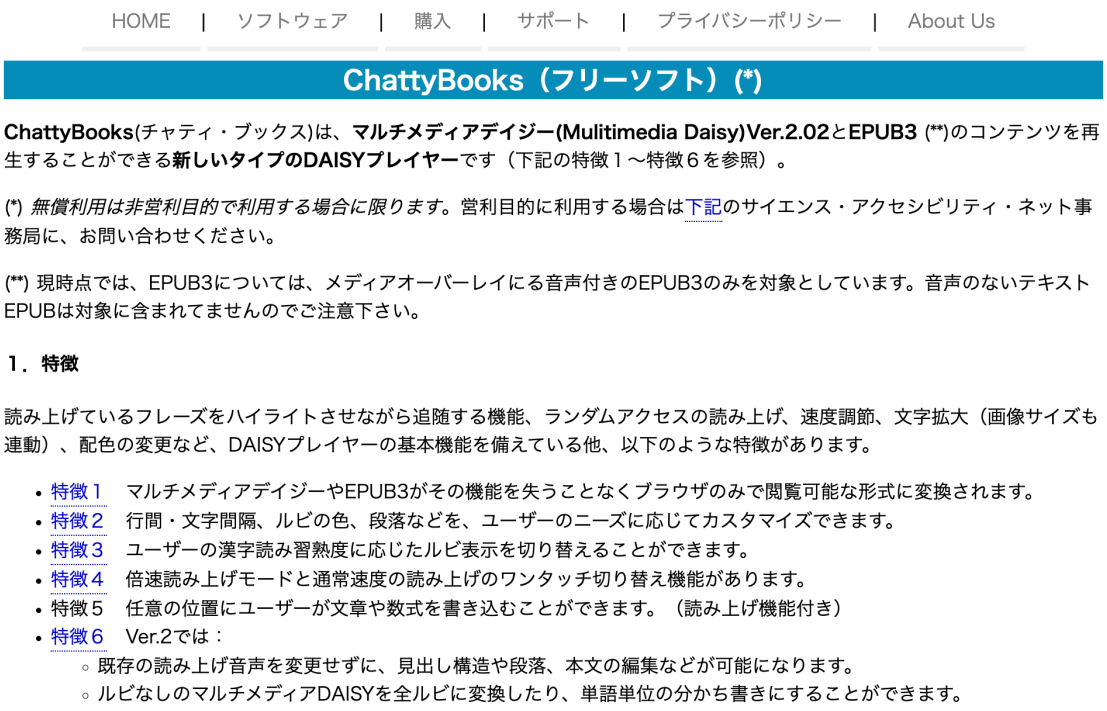
2 AIリテラシーの定義

佐藤(2019)は人工知能を適切に活用する際に求められるAIリテラシーについて検討している．その際，「人工知能に関する知識・理解」，「人工知能を適切に活用するための思考力･判断力」の多くの２つに分類し，計14項目を設定している．この研究では，植生の二酸化炭素吸収量の推定に機械学習を利用するなどあらかじめ授業者が提示した課題に取り組む学習となっている．その結果，「人工知能に何を学習させるか」「人

DaisyXML の出力対応

変換ライブラリとして Daisy Pipeline2 を選定したが、HTML から DAISY への変換には対応しておらず、HTML→EPUB 3→DAISY という 2 段階の変換を行う必要があることが判明した。 また Daisy Pipeline2 そのものの開発も停滞が見受けられた。

日本 DAISY コンソーシアム・日本障害者リハビリテーション協会を含む有識者へのヒアリングでは、「過去のフォーマットである DAISY よりも EPUB 3 を対象にすることが望ましい」という意見が得られた。 そのため出力フォーマットのターゲットを DAISY から EPUB 3 に変更することにしたが、現時点では EPUB 向けの入力情報に不足があるため今後の予定とする。 「WordTalker」の表示、発音ユーザインタフェースの再開発 努力目標として挙げていたが、主要な EPUB3/DaisyXML プレイヤーである ChattyBook の利用をユーザーに推奨することとした。



HOME | ソフトウェア | 購入 | サポート | プライバシーポリシー | About Us

ChattyBooks（フリーソフト）(*)

ChattyBooks(チャティ・ブックス)は、**マルチメディアデイジー(Mulitimedia Daisy)Ver.2.02**と**EPUB3** (**)のコンテンツを再生することができる**新しいタイプのDAISYプレイヤー**です（下記の特徴１～特徴６を参照）。

(*) *無償利用は非営利目的で利用する場合に限ります*。営利目的に利用する場合は下記のサイエンス・アクセシビリティ・ネット事務局に、お問い合わせください。

(**) 現時点では、EPUB3については、メディアオーバーレイにる音声付きのEPUB3のみを対象としています。音声のないテキストEPUBは対象に含まれてませんのでご注意下さい。

１．特徴

読み上げているフレーズをハイライトさせながら追随する機能、ランダムアクセスの読み上げ、速度調節、文字拡大（画像サイズも連動）、配色の変更など、DAISYプレイヤーの基本機能を備えている他、以下のような特徴があります。

- 特徴１　マルチメディアデイジーやEPUB3がその機能を失うことなくブラウザのみで閲覧可能な形式に変換されます。
- 特徴２　行間・文字間隔、ルビの色、段落などを、ユーザーのニーズに応じてカスタマイズできます。
- 特徴３　ユーザーの漢字読み習熟度に応じたルビ表示を切り替えることができます。
- 特徴４　倍速読み上げモードと通常速度の読み上げのワンタッチ切り替え機能があります。
- 特徴５　任意の位置にユーザーが文章や数式を書き込むことができます。（読み上げ機能付き）
- 特徴６　Ver.2では：
  - 既存の読み上げ音声を変更せずに、見出し構造や段落、本文の編集などが可能になります。
  - ルビなしのマルチメディアDAISYを全ルビに変換したり、単語単位の分かち書きにすることができます。

ChattyBooks https://www.sciaccess.net/jp/ChattyBooks/

その他

版面の解析精度を向上のために、版面画像の解析前処理（余白、見開き、傾き補正、回転処理）等の機能を開発した。

PDF のナビゲーション機能を活用しアクセシビリティを向上させるために、しおりデータの設定機能(SetBookmark)およびしおりデータの抽出機能（GetBookmark）を開発した。マイクロコンテンツ化については検討段階に留まっている。

京都大学美馬秀樹教授とオンライン会議を行い、過去の研究における構造判定のアドバイスを頂いた。

- 岩波書店『思想』について構造判定を行った際には、フォント情報、サイズ情報、位置情報が有効だった。
- ページの上下左右の余白も重要。
- 形態素解析も併用した。見出しには名詞が多く含まれる傾向にある。
- 日本語なのか英語なのか、ピリオドやスペースの扱いまで含めないと精度が出ない。
- 単純に句読点やピリオドで区切ってセンテンスにすればよい訳ではない。メールや URL の扱いにも注意が必要。
- 見出しをツリー構造にするのはレイアウト解析の分野かもしれない。

日本 DAISY コンソーシアム・日本障害者リハビリテーション協会等の有識者と会議を行い出版物のアクセシビリティに関する課題や意見をヒアリングした。

- 現在のスクリーンリーダーは未だにルビの親文字とルビ文字を二重に読み上げる問題を解消できていない。
- PDF のアプリケーションの読み上げは不便であるため、公文書等は PDF だけではなく Word や Excel ファイルなどの形式でも提供してもらいたい。
- ブラウザの JavaScript で実装されたビューワーもあるが DOM 構造がばらばらであるため、アクセシビリティ面では期待できない。
- デジタル庁等がアクセシビリティについてのガイドラインを設けていることは認識しているが、活用されているとは言い難い。
- 国立国会図書館以外でも文書資料のアーカイブを画像フォーマットで持っている企業や団体があるが、視覚障害者が利用するのは困難である。AI を役立てるソリューションができないだろうか。

次年度に向けて

- レイアウト解析、読み順、見出し判定モデルについては、開発を継続しさらなる精度向上を目指す。
- デモサイトの公開や企業・団体に PDF 提供を依頼し、より多様な PDF ファイルによる検証とフィードバックの獲得を目指す。
- ChatGPT に代表される大規模言語モデル(LLMs)の活用を調査・検討する。これまで対応を断念していた画像の説明テキスト生成や表画像のデータ化の活路となる可能性がある。
- 2023 年 5 月に発表された Adobe Sensei PDF Accessibility Auto-Tag API には、本研究の今後の課題であった見出しの階層化を実現できる可能性がある。また、より詳細な構造

判定やタグ付き PDF の生成にも対応している。現時点では日本語のサポートを表明していないが、本研究を事業化する上で役立つ機能を複数有している。弊社は Adobe 社の PDF ライブラリの日本販売代理店を務め、自社ソフト PDF2MD も同ライブラリをベースとしているなど深いつながりがあるため、日本導入支援などのパートナーシップを模索してゆきたい。