別紙３

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】

本研究では令和３年度において、構造判定機械学習モデルの開発、テキスト抽出エンジンの開発、クラウドサービスのプロトタイプ版の開発の３点について取り組んだ。

１．構造判定機械学習モデルの開発

開発内容

PDF の版面を画像化し、Google Cloud Vision に解析させた情報を取得。得られたそれぞれの段落に構造情報についてのラベリングを行った。当初は段落よりも大きな単位であるブロックに対してラベリングをする想定であったが、ブロック自体には十分な情報が含まれていなかったため、段落を対象とした。一段落あたり 30 の特徴量を持ち、主要なものは版面上の位置情報、面積、書字方向、文字数である。ラベルについては本文、見出し、ノンブル、柱、ルビ、その他の６種類とした。当初は、より細かい粒度の構造情報（複数のレベルの見出し、図表、注記など）の判定を期待していたが、Google Cloud Vision の抽出単位には適していなかったため初年度は簡潔なものに留まった。

対象

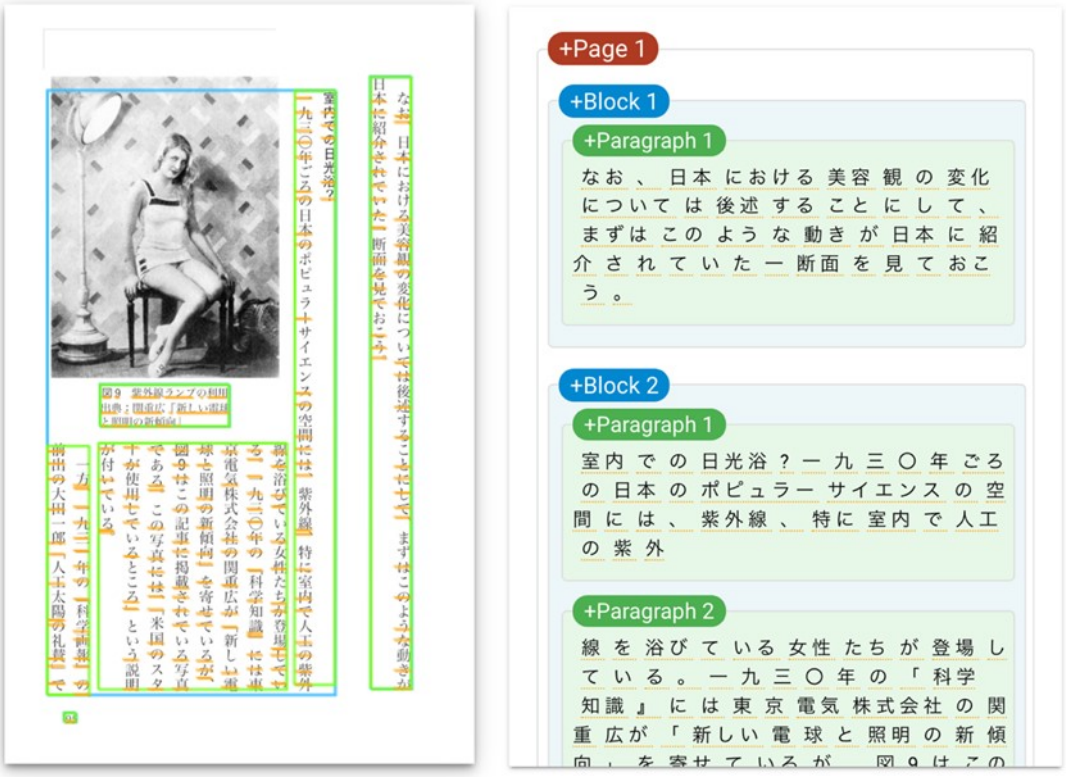
簡潔なレイアウトを持つ一般書、新書、論文 60 点を対象とした。段落数は一般書、新書で 4000 程度、論文で 500 程度である。

評価

ノンブル、柱、ルビについては高い精度で判定できたものの、見出しについては本文と区別できないケースが多く見られた。これは Google Cloud Vision の情報だけでは構造を判定するための特徴が不足していたためと考えられる。

また、Google Cloud Vision のレイアウト解析自体にも課題が見られた。見出しと本文を区別せず同じブロックにしてしまう。図や表を一塊のブロックとして認識せず、中の文字だけを段落とし

て拾ってしまうなど、期待する解析結果が得られなかった。



Cloud Visionのレイアウト解析見出しと本文が同一パラグラフと見なされている。図版はキャプションのテキストのみを拾っているが図版も含めて一塊のブロックをして扱われるのが望ましい。

今後の課題と方向性

１　レイアウト解析の問題

オープンソースのLayoutParserなど、Google Cloud Vision以外のレイアウト解析モデルの調査及び開発。また将来的に多様なレイアウトの出版物に対応するため、出版物の種類ごとに特化した複数のレイアウト解析モデルを使い分けるようにしてゆきたい。

２　構造判定の精度向上

OCR機能による段落の情報を手がかりとした構造判定には限界があり、２つの改善方法が考えられる。

一つは、段落の特徴にPDF2MD（弊社PDF解析プログラム）のフォント情報を含めて学習させる方法である。もともとPDF2MDにはあらかじめ見出しや本文のフォントを指定することで構造化された文書を出力する機能を有していたが、初年度はGoogle Cloud Visionの情報で構造を判定する想定でいたため活用できていなかった。PDF2MDに機械学習に適したデータを出力する機能を設け、学習させることで本来の機能が発揮できるようにしてゆきたい。

もう一つは、画像認識技術による構造判定である。画像の中の図や表をブロックとして扱い意味を機械的に判定するモデルを作成する。こちらもPDF2MDではある程度、判定のアルゴリズムが実装されていた（特定の長さ以上のパスが含まれていれば表と見なす、など）がそれほど精度の高

いものではなかった。機械学習を活用することで精度を向上させることが期待される。
上記２つの方法のうち、令和四４年度は前者のフォント情報を手がかりにした手法を試行すし、画像認識技術の活用は次善の手段とする。

２．テキスト抽出エンジンの開発

開発内容
PDF2MDを改修し、PDFファイルに対してページ番号と矩形の座標を指定することで、範囲内のテキストを出力できるようにした。

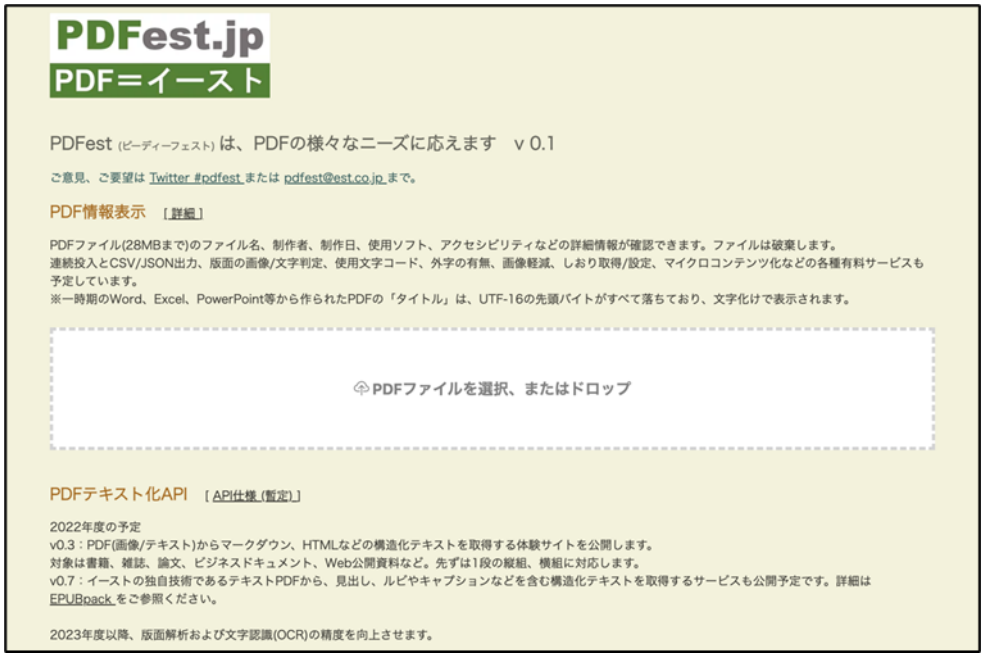
評価
予定していた開発内容は完了し、Google Cloud Visionのブロックや座標の段落に対しても対応するテキストを取り出せるようになった。
ファイルが持つテキストそのものを出力しているためGoogle Cloud VisionのOCRよりも高い精度のテキストを得ることができている。両者のテキストを比較したところ、特に記号類、小書きの仮名、画数の多い漢字などでは、OCRが誤認識したテキストを回避できていた。反対に稀ではあるが、印刷会社が独自に作成したフォントなどで、見た目と異なるコードポイントが割り当てられていた文字については、OCRのほうが正確というケースもあった。

今後の課題と方向性
前述したOCRのほうが正確なテキストとなるケースは、仕様として対処しない。
他の課題としては段落の改行問題がある。ページの終端で段落が折り返している場合、途中で改行されひと繋がりの文章とならない事象である。原因のひとつにはGoogle Cloud Visionにおける段落に相当するparagraph という単位には実際には段落ではなく行が出力されていた点である。またそれぞれのparagraphには終端が文章の途中であるか否かを示すパラメタがあるが、日本語の文章では適切に機能していなかった。PDF2MD単体では、行を跨ぐ段落を繋げるロジックがあるが、初年度は段落の単位の判定をGoogle Cloud Vision側に委ねていたため活用できていないのが現状であった。これについては、PDF2MDの持つ段落整形ロジックを活用するか、機械学習でより良い判定ができるのかを見極めた上で対処する。
また、構造判定機械学習モデルの項で述べたとおり、フォント情報を含む学習データを作成するために、ページのある全ての文字の位置とフォントの種類、サイズをJSON形式で出力するAPIを実装する。
この他に、PDFファイルにあらかじめ設定されたしおりの抽出や、それに基づく章単位でのコンテンツの切り出し機能についても研究を進める。

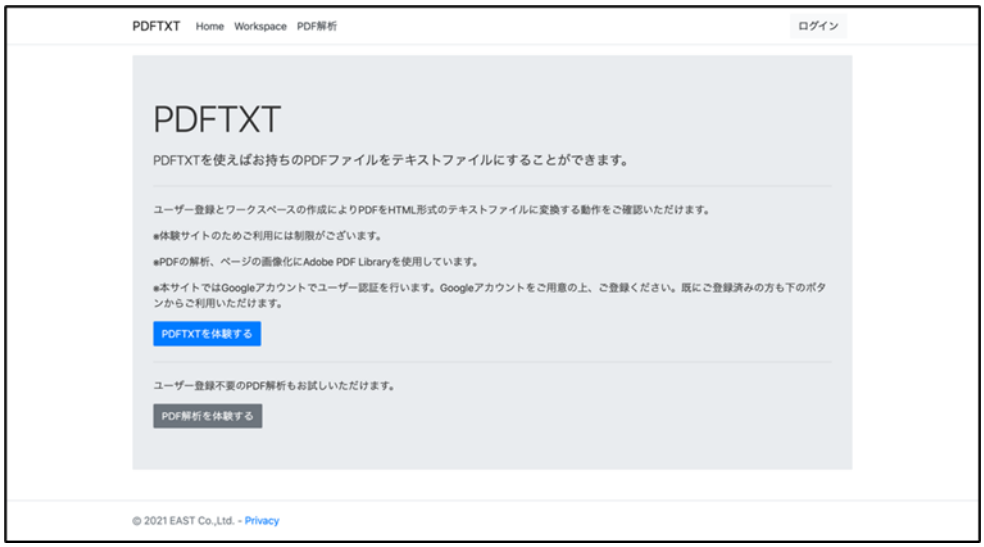
３．クラウドサービスのプロトタイプ版

開発内容
本研究を含む弊社のPDFソリューションデモサイトとして pdfest.jp を公開した。このサイトで

は本研究の現時点での機能も同サイトで利用できる他、PDFファイルが持つ様々な属性なども手軽に確認することができる。



PDFest.jpトップページ



クラウドサービス　トップページ

本研究のプロトタイプとしての実装状況は次のとおりである。

機能	内容と開発状況
テキスト抽出機能	版面の各構成要素に含まれるテキストを文字単位取得し成形する。 → 実装済み
構造判定機能	機械学習モデルにより、版面の構成要素の構造を判定する。

	→実装済み
出力機能	プレーンテキスト、構造化テキスト（HTML・DaisyXML）、音声による出力。 →プレーンテキストとHTMLの出力を実装済み。
その他サービス機能	ユーザー認証、アップロード、ダウンロード、変換キュー制御、プレビュー。 → プレビュー機能を除いて実装済み

補足

その他サービス機能

エンドユーザーは自身のアカウントを作成しサイトにログインする。個々のユーザーは「マイスペース」を持ち、自身がPDFの管理（アップロード、変換状況の確認、削除）や変換されたアクセシブルなテキストのダウンロードができる。

Web API

すべての機能はWeb APIを介して利用できるようにした。これによって他のアプリケーションや外部のシステムからでも本サイトの機能を活用することができる。認証にはアカウント作成時に個々のユーザーに自動で割り当てられるAPIキーを仕様する。

機能	メソッド	URL	説明
PDF アップロード	POST	/api/books	PDF ファイルをアップロードして変換処理を開始する。
処理状態取得	GET	/api/books/{fileId}	指定した PDF ファイルの変換処理の状態を表示する。
版面画像取得	GET	/api/books/{fileId}/{pageNum}.png	指定した PDF ファイルの指定したページを画像として表示する。（レイアウト解析に利用するためのもの）
テキスト取得	GET	/api/books/{fileId}/{pageNum}.txt	指定した PDF ファイルの指定したページのテキストを取得する。パラメタに矩形の座標を指定することで特定のブロックのみを主力することも可能。
処理結果受信	PUT	/api/books/{fileId}	指定した PDF ファイルの変換結果（txt、HTML）を受信する。
PDF 削除	DELETE	/api/books/{fileId}	指定した PDF ファイルをマイスペースから削除する。

今後の課題と方向性

クラウドサービスについて、本年度は出力機能にDaisyXMLを追加する。公開したデモサイトについて、広くユーザーの反応を取り入れる。またデザインやUIの向上を図る。追加が必要となった機能は随時Web APIとしても切り出し、ドキュメントも整備する。

以上