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。

概要

PDF変換サービスの改良

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研究開発は、マルチモーダル大規模言語モデルを採用することによって大きく前進した。本事業で開発した、文書画像を手がかりにしてマルチモーダル大規模言語モデルに抽出済みテキストを再構成させる手法は精度が高く、当初想定していなかった雑誌のような複雑なレイアウトの誌面にも応用できることが判明した。

サイト内 PDF 探査ツールの開発

URLを起点としてサイト内に含まれるPDFファイルを一覧化して取得するツールの開発が完了し、ウェブサイトのデータ把握に活用している。

EPUB のアクセシビリティ向上

EPUB のアクセシビリティメタデータをチェック・生成するコマンドラインツールを開発したが、国内全体でも EPUB のアクセシビリティ向上の動きは鈍い。本年度開発したマルチモーダル大規模言語モデルを活用する手法に、省コスト化と高品質化の可能性を感じている。

PDF変換サービスの事業化

企業や団体に PDF 変換サービスのデモを行なったところ、以下の領域・用途に反響があった。

- 医学系論文の閲覧サービスにおける論文のデジタル化移行
- 製造業における、PDF 仕様書からのデータ抽出
- 教育機関における、オンライン学習に向けた印刷教材のデジタル変換
- 出版社における雑誌バックナンバーからのデータ抽出

各企業・団体からのデータ提供を受け調査やコンテンツ 制作を進めており、本格的な受託を目指している。

研究開発の経緯と詳細

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研究開発は、主に見出しを中心とした構造情報の取得と、ページ番号や柱など不要な情報の除去を中心として行ってきた。初年度は、PDFを解析して文字のフォントサイズと位置情報を手がかりとして見出しを判別するモデルを開発したが、十分な成果には繋がらなかった。次年度は、OCRによるレイアウト解析モデルを中心に据え、得られた情報にフォント情報を組み合わせることで、見出し判定の精度を向上させることができたが、要素の読み順や段落整形に大きな課題を残すこととなった。

本年度は、はじめにテキストと文書の画像をペアで学習させたマルチモーダルなモデルをレイアウト解析に活用することにした。Microsoft社が開発したLayoutLMv3はテキストの位置情報を画像の特徴量を学習した代表的なモデルであり、クラウドサービス Azure上で提供されているLayoutモデルは本モデルをベースにしていると推測される。

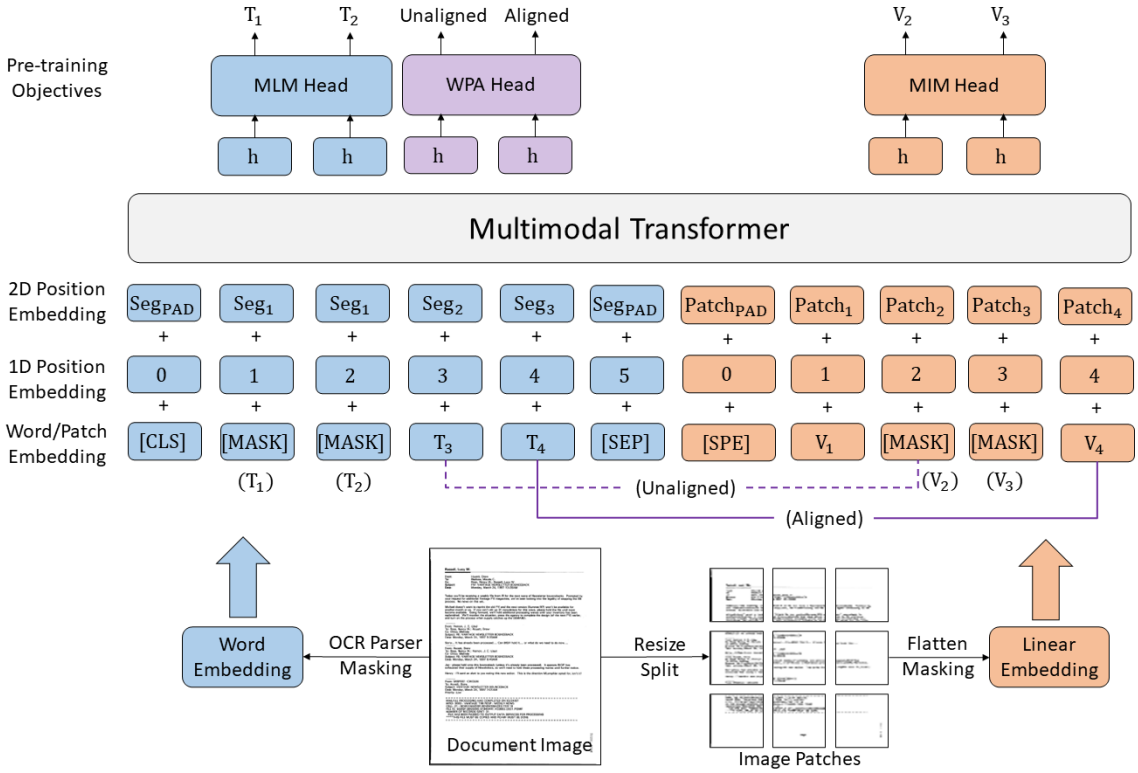


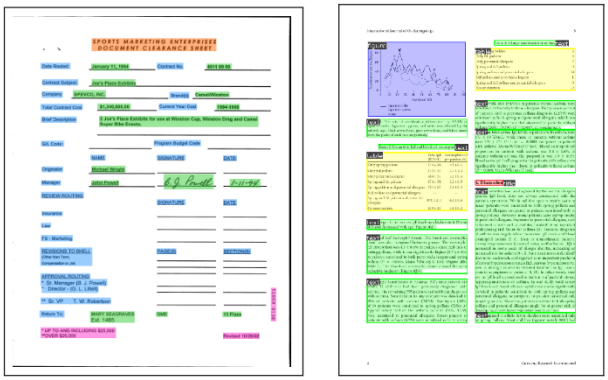
図 LayoutLMv3のアーキテクチャ（https://arxiv.org/abs/2204.08387）

このLayoutモデルは、タイトル、見出し、本文、フッター、ページ番号、図、表の判別とテキスト抽出が可能である。しかし縦書きの日本語文書を解析させたところ、行が読み順の逆になってしまう、一部の段落が、表として扱われてしまう等、テキスト抽出に問題を抱えていることが明らかになった。そこで、Layoutモデルについてはレイアウト解析・構造判定情報のみを利用し、テキスト抽出は従来弊社でテキスト抽出に使ってきたPDF2MDで行うことにした。PDF2MDは、事前に手作業で見出しや図表に注釈を加えることで、テキストに構造を与えながら抽出するソフトウェアである。この手作業の部分をLayoutモデルに代替

させる方式を開発し、自社クラウドサービスに実装した。

しかし、AzureのLayoutモデルは現時点では追加学習させるサービスの対象となっておらず、日本語のさまざまなレイアウトの誌面を学習させてチューニングする手段が提供されていなかった。

弊社では医学論文から構造化テキストを抽出するニーズがあり、LayoutLMv3をベースに英語の医学論文PubLayNetで学習したモデルに相当するものを必要としていたが、いずれのモデルもライセンス上商用利用が禁止されており、事業化に繋げることができない。そのため、性能の向上については、Azureでチューニング対象モデルが拡大されるか、Layoutモデルのアップデートを待つかの足踏み状態となってしまった。



(a) Text-centric form understanding on FUNSD **(b) Image-centric layout analysis on PubLayNet**

図 PublayNetの医学論文で学習したLayoutLMv3（https://arxiv.org/pdf/2204.08387）

転機となったのは、大規模言語モデルのマルチモーダル化である。2023年にOpenAI社のGPT-4で画像が扱えるようになり、2024年のGPT-4oは大幅に性能が向上した。

単純にGPT-4o単体で画像からのテキスト抽出・文書化のタスクを実行させると、元の文書に存在しない文章が生成されるハルシネーションや、勝手な要約、言い換え、文体の変化、など不正確な結果が多発した。しかし、事前にPDFのテキスト出力またはOCRでテキスト（構造化されないない、不正確な読み順、不自然な改行やスペースを含む）を抽出しておき、誌面の画像をヒントに文書として再構成させたところ、高い精度で元の文書を構造化テキスト（マークダウン）にすることができたため、プロンプトを改良しながらウェブアプリケーションとして実装した。

またこの手法は、従来難しかった不自然な改行やスペースの除去の問題も解消されており、ごく自然に段落が整形されている。さらに、当初想定していたシンプルな論文だけでなく、複雑なレイアウトを持つ雑誌の誌面でも有効に機能した。誌面の各所に配置されたコラムや図版などのブロックは、本文と混在することなく、独立して書き出されている。

課題としては、マークダウンの出力にムラがあり、見出しの階層や、強調表現、箇条書きの制御に改良の余地がある点が挙げられる。また表や画像の説明文の生成や、校正タスクを追加することにより、さらにアクセシビリティと品質を高めることができるだろう。

変換事例

（別紙参照）

【別紙】

変換事例：総務省2024年5月号 p2～3をOCR経由し大規模言語モデルで再構成

特集

独立行政法人シンポジウムを開催しました

～社会環境の変化に対応する独立行政法人の取組～

◆独立行政法人とは、国の政策を実現するため、法律に基づいて設置されるものです。
◆現在、87の独立行政法人があり、社会環境の変化に対応しつつ、政策実施機能を最大化し、国民生活および社会経済に貢献することが求められます。
◆このような取組の推進に当たり、①独立行政法人評価制度委員会が独立行政法人のマネジメントに期待することおよびその実現を支えるために委員会が果たす役割等についてお伝えするとともに、②独立行政法人の業務運営を支える人材の確保・育成について、先進的な取組事例を基に、理事長等によるリーダーシップやマネジメントの在り方について議論を深めるため、今年1月29日に「独立行政法人シンポジウム」を開催しました。

プログラム

○基調講演
「独立行政法人のマネジメントに期待すること」
澤田道隆独立行政法人評価制度委員会委員長（花王（株）取締役会長（当時））
○パネルディスカッション
「法人の使命を果たすための人材の確保・育成の取組」

開会挨拶を行う長谷川鉱務大臣政務官

当日は、会場とオンライン合わせて約400名の方に聞参加いただきました。

基調講演「独立行政法人のマネジメントに期待すること」

澤田道隆独立行政法人評価制度委員会委員長（花王（株）取締役会長（当時））

◆独立行政法人内のマネジメント・内部統制の在り方への期待
・主務大臣が示す使命・ミッションを踏まえた法人自身のビジョンの確立
・法人の長によるトップマネジメントと監事機能を活用した内部統制
⇒職員に対しては、各法人が持つ長期的ビジョンと併せてメッセージを伝え、「大きく社会の役に立つ仕事を担っている」という「ワクワク感」を持って業務を遂行いただく環境を作ることが、今の時代の経営層には求められるのではないか。
◆目標管理を中心とした主務大臣によるガバナンスの在り方への期待
・将来的な法人のあるべき姿である使命等の提示
・独立行政法人とのコミュニケーションを重視したガバナンス
◆独立行政法人評価制度委員会（法人の中期目標、業績評価等をチェックする機関）における調査審議に当たっての基本的視座
・府省・法人横断的に求められる対応の促進・支援
・主務大臣と法人の両者の意識の共有に立脚した効果的な PDCA サイクルの実現
・法人の長等によるマネジメント・内部統制の改善を促進

基調講演を行う澤田委員長

独立行政法人の政策実施機能の最大化の実現に向けて、独立行政法人評価制度委員会が、積極的にサポート
・各法人の目標見直しの機会を中心に、主務省・法人に対するヒアリングなどを通じて、法人の将来像等について認識を共有、委員会の考えをフィードバック
・各主務省・法人が具体的に業務改善の参考となるよう、取組事例を積極的に収集・展開
・独立行政法人の業務管理・内部管理の共通的な方向性を提示

パネルディスカッション

「法人の使命を果たすための人材の確保・育成の取組」

パネルディスカッションでは、国民生活および社会経済に貢献することが求められる独立行政法人の業務運営を支える人材の確保・育成について、まず、3つの法人からそれぞれの取組が紹介され、その後、パネリストによる活発な議論が交わされました。ここでは、その概要を紹介します。

【原田委員長代理】 私は独立行政法人評価に携わって約10年になります。この間、運用の大きな変化がいくつかありましたが、その1つに、目標期間の途中で各法人に次々と新しい目標が付与される、ということがあります。これは法人に対する信頼の高まりの表れと考えられますが、他方、現在のリソースでやれるのだろうか、と考えることがあります。

こうした問題意識は各法人共通ではないかと存じます。

モデレーター　原田委員長代理

本日は、「人材の確保・育成」の観点から、法人のリーダーのお三方に、取組とお考えを伺いました。委員の皆さまはどうお感じになったでしょうか。

【栗原委員】 独立行政法人評価制度委員会では、中長期計画を見させていただくわけですが、業務目標の議論とともに、経営基盤についての議論も意識的に行っていま す。その経営基盤の中でも特に重要なのは「人材」です。民間でも法人でも、「人材」は経営上の非常に重要なテーマで、人材戦略を開示していく動きは民間ではすでに始まっています。

しかし、人材確保やエンゲージメント向上、どう人材を育成していくか、ということなどはなかなか王道がありません。本日のお話には、様々な工夫と参考になる例が多分に含まれていたと思いまし た。ぜひ他の法人でも参考にしていただきたいと思います。

パネリスト　栗原委員

【各法人の取組の概要】

○森林研究・整備機構「ダイバーシティ推進の取組」
・理事長自らがダイバーシティ推進本部長を務め、多様な人材がそれぞれの能力を存分に発揮できる職場環境の充実に向けた取組を推進
・全国の研究教育機関をメンバーとする「ダイバーシティ・サポート・オフィス」を通して関係機関と連携

○製品評価技術基盤機構「デジタル人材の確保・育成の取組」
・デジタル人材育成を段階的に実施
・役職員の IT パスポート試験等情報処理技術者試験の取得率80%以上を達成
・その結果、業務効率化や新たな価値の創出が促進

○住宅金融支援機構「法人の使命の徹底、働きやすさやモチベーション向上に資する取組」
・全職員参加型でパーパスを決定し、内外に浸透
・専門人材等を確保・育成するとともに、働きやすい職場づくりを推進
・カイゼン活動（職員による業務の効率化・事務ミス防止・職場環境改善等の工夫）と優良事例の表彰を通じたモチベーションの向上

テキストデータ利用率　96.3% 53/55

与えられたテキスト群を完全に使用して文書を再構成できたかの割合。

写真のキャプションが2箇所漏れ。未使用テキストのチェック機構を追加予定。

誤字脱字　0.011% 23/1917

半角文字と全角文字の違いなど、OCR時点で発生したものが多い。他に括弧「」の抜け等

PDFのテキストデータを利用する場合はより低くなる。

例　赤文字が正解

~社会環境の変化に対応する独立行政法人の取組~
～社会環境の変化に対応する独立行政法人の取組～

独立行政法人とのコミュニケーションを重視したガバナンス
「独立行政法人」とのコミュニケーションを重視したガバナンス

ハルシネーション等発生率　0.018% 1/55

元の文章を変更してしまうケースが稀に発生。画像のキャプション付近のため、写真の説明テキストを生成してしまったか。

長谷川総務大臣政務官が開会の挨拶を行いました。
開会挨拶を行う長谷川総務大臣政務官

見出し検出率 77.78%
文字サイズもフォントも異なる見出しやコラムの見出しも正しく検出。
2つ余計に検出。パネルディスカッションの太字発言者名を見出しと誤認識。

変換記事全文
グレー背景は見出し。他はHTMLのデフォルトスタイル

特集

独立行政法人シンポジウムを開催しました

~社会環境の変化に対応する独立行政法人の取組~

- 独立行政法人とは、国の政策を実現するため、法律に基づいて設置されるものです。
- 現在、87 の独立行政法人があり、社会環境の変化に対応しつつ、政策実施機能を最大化し、国民生活および社会経済に貢献することが求められます。
- このような取組の推進に当たり，1. 独立行政法人評価制度委員会が独立行政法人のマネジメントに期待することおよびその実現を支えるために委員会が果たす役割等についてお伝えするとともに、2. 独立行政法人の業務運営を支える人材の確保･育成について、先進的な取組事例を基に、理事長等によるリーダーシップやマネジメントの在り方について議論を深めるため、今年 1 月 29 日に「独立行政法人シンポジウム」を開催しました。

プログラム

- **基調講演**
  - 「独立行政法人のマネジメントに期待すること」
  - 澤田道隆（独立行政法人評価制度委員会委員長、花王株取締役会長(当時)）
- **パネルディスカッション**
  - 「法人の使命を果たすための人材の確保･育成の取組」

開会挨拶

- **長谷川総務大臣政務官** が開会の挨拶を行いました。

当日は、会場とオンライン合わせて約 400 名の方に御参加いただきました。

基調講演「独立行政法人のマネジメントに期待すること」

澤田道隆（独立行政法人評価制度委員会委員長、花王株取締役会長(当時)）

- **独立行政法人内のマネジメント･内部統制の在り方への期待**
  - 主務大臣が示す使命･ミッションを踏まえた法人自身のビジョンの確立

- 法人の長によるトップマネジメントと監事機能を活用した内部統制
- 職員に対しては、各法人が持つ長期的ビジョンと併せてメッセージを伝え、「大きく社会の役に立つ仕事を担っている」という「ワクワク感」を持って業務を遂行いただく環境を作ることが、今の時代の経営層には求められるのではないか。

- **目標管理を中心とした主務大臣によるガバナンスの在り方への期待**
  - 将来的な法人のあるべき姿である使命等の提示
  - 独立行政法人とのコミュニケーションを重視したガバナンス
- **独立行政法人評価制度委員会における調査審議に当たっての基本的視座**
  - 府省･法人横断的に求められる対応の促進･支援
  - 主務大臣と法人の両者の意識の共有に立脚した効果的な PDCA サイクルの実現
  - 法人の長等によるマネジメント･内部統制の改善を促進

パネルディスカッション「法人の使命を果たすための人材の確保･育成の取組」

パネルディスカッションでは、国民生活および社会経済に貢献することが求められる独立行政法人の業務運営を支える人材の確保･育成について、まず、3 つの法人からそれぞれの取組が紹介され、その後、パネリストによる活発な議論が交わされました。ここでは、その概要を紹介します。

[原田委員長代理]

私は独立行政法人評価に携わって約 10 年になります。この間、運用の大きな変化がいくつかありましたが、その 1 つに、目標期間の途中で各法人に次々と新しい目標が付与される、ということがあります。これは法人に対する信頼の高まりの表れと考えられますが、他方、現在のリソースでやれるのだろうか、と考えることがあります。

こうした問題意識は各法人共通ではないかと存じます。

モデレーター: 原田委員長代理

本日は、「人材の確保･育成」の観点から、法人のリーダーのお三方に、取組とお考えを伺いました。委員の皆さまはどうお感じになったでしょうか。

[栗原委員]

独立行政法人評価制度委員会では、中長期計画を見させていただくわけですが、業務目標の議論とともに、経営基盤についての議論も意識的に行っています。その経営基盤の中でも特に重要なのは「人材」です。民間でも法人でも、「人材」は経営上の非常に重要なテーマで、人材戦略を開示していく動きは民間ではすでに始まっています。

しかし、人材確保やエンゲージメント向上、どう人材を育成していくか、ということなどはなかなか王道がありません。本日のお話には、様々な工夫と参考になる例が多分に含まれていたと思いました。ぜひ他の法人でも参考にしていただきたいと思います。

パネリスト: 栗原委員

各法人の取組の概要

- **森林研究･整備機構「ダイバーシティ推進の取組」**

- 理事長自らがダイバーシティ推進本部長を務め、多様な人材がそれぞれの能力を存分に発揮できる職場環境の充実に向けた取組を推進
- 全国の研究教育機関をメンバーとする「ダイバーシティ･サポート･オフィス」を通して関係機関と連携

- **製品評価技術基盤機構「デジタル人材の確保･育成の取組」**
  - デジタル人材育成を段階的に実施
  - 役職員の IT パスポート試験等情報処理技術者試験の取得率 80%以上を達成
  - 業務効率化や新たな価値の創出が促進
- **住宅金融支援機構「法人の使命の徹底、働きやすさやモチベーション向上に資する取組」**
  - 全職員参加型でパーパスを決定し、内外に浸透
  - 専門人材等を確保･育成するとともに、働きやすい職場づくりを推進
  - カイゼン活動（職員による業務の効率化･事務ミス防止･職場環境改善等の工夫）と優良事例の表彰を通じたモチベーションの向上

独立行政法人の政策実施機能の最大化の実現に向けて、独立行政法人評価制度委員会が、積極的にサポート

- 各法人の目標見直しの機会を中心に、主務省･法人に対するヒアリングなどを通じて、法人の将来像等について認識を共有、委員会の考えをフィードバック
- 各主務省･法人が具体的に業務改善の参考となるよう、取組事例を積極的に収集･展開
- 独立行政法人の業務管理･内部管理の共通的な方向性を提示