OCR を利用した RAG における PDF 文書内のタイトルや表の利用

蒲原悠登 [1] 竹内孔一 [2]

[1] 岡山大学工学部
[2] 岡山大学大学院 環境生命自然科学学域
p3z19g5j@s.okayama-u.ac.jp takeuc-k@okayama-u.ac.jp

概要

本論文では，大量かつ複雑な構造を持った PDF 文書を扱う際に，Retrieval- Augmented Generation (RAG) の利用において，OCR ライブラリ surya[1)] を利用したチャンキング手法を提案する。surya を用いてタイトルや表を自動判定し，それぞれの構造に応じた処理を適用することで精度の向上を目指す．学内事務手続き PDF を用いた RAG による質問応答実験では，単純なテキスト分割と比べて BLEU および ROUGE 各種指標で高い性能を示した．その際特に表領域の処理の効果が確認された．

1 はじめに

現代において，解説書やマニュアルなどの複雑な資料を理解することは，多大な労力を要する．これは，大規模なプロジェクトや財務資料の積み上げなど，機械的に解釈するのが難しい場面では特に明確である．

これを解決する為の手段として，Retrieval-Augmented Generation (RAG) は有力な選択肢である [1]．RAG は資料をベースに，意味のある回答を生成することが可能であり，複雑な資料の理解を支援することが期待される．

文書内に対して LLM を利用した RAG による質問応答が研究されており [2]，RAG の性能向上には，vector 検索をベースにさまざまな改善の手段がある．例えば，chunk の作成，メタデータの利用，ranking，LLM のプロンプトの改善，クエリ拡張などの手法が挙げられる．本論文では，chunk の作成方法に注目する．

表や段落などのどういう構成かがデータの中に書いてある場合，文書内の表や段落それぞれに対して処理を適用することが言われている [3]．しかしながら本稿での実験で扱うような，段落構造としてのメタデータが存在しないため PDF データでは利用することが出来ない．

そこで OCR ライブラリである surya を使用し，レイアウトを判断することで，表と文本とで chunk の作成方法を変えたところ，学内の事務手続き PDF に関する検索で本文に関する質問については BLEU スコアで約 0.03 ポイント，表に関する質問については BLEU スコアで約 0.12 ポイントの上昇が見られたので報告する．

2 タイトルおよび表解析手法

本論文で提案する OCR を使用したレイアウトを意識したチャンキングは，下記のようなフローによって実現される．

2.1 surya を用いたレイアウト解析

初めに，PDF ファイルを JPEG 形式の画像に変換する．その後 JPEG に変換された画像は，OCR ライブラリ surya を用いてレイアウト解析を行う．surya は図 1 のようにテキストを始め，タイトルや表など複数のレイアウトタイプを判定することが可能である．本論文では，特に「Title」「Section-header」をタイトルとして判定し，タイトルと判断されたレイアウトに対してはタイトルの処理を，「Table」と判断されたレイアウトに対しては表の処理を，それ以外に対して本文の処理を行う．

1） https://github.com/VikParuchuri/surya

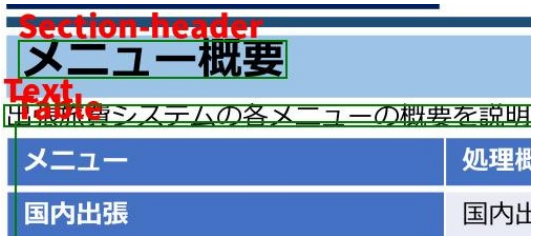

図 1　レイアウトの判定例

2.2　テキスト抽出

PyMuPDF[2] というライブラリを使用してレイアウト解析の結果を活用し，個別のレイアウト型に対して適切な処理を行う．具体的な処理は下記に記述する。

2.2.1　タイトルの処理

surya によって「Title」や「Section-header」と判断された範囲内のテキストに対しては，PyMuPDF の get_text を使用してテキストを取得した後タイトル情報として保存しておき，そのタイトルに属する文章が現れたときにその文章とともに保存する．これにより，文章の背景情報を明確化することが可能となる．

2.2.2　表の処理

「Table」と判断された範囲内のテキストに対しては，PyMuPDF の find_tables 機能を用い，表情報を Markdown 形式で抽出する．こうすることにより表の構造の情報を失うことなく表をテキストに変換した上で保存することが出来る．また検索用に表を一行毎取り出し，Markdown の表情報と一緒に保存する．

2.3　本文の処理

「Section-header」「Table」「Title」と判断されなかった範囲のテキストは，PyMuPDF の get_text を使用して収集した後，ノイズ低減のため同タイトルのテキストと 64 文字以上になるように連結させ，現在保存されているタイトルと一緒に保存する．

2）　https://pymupdf.readthedocs.io/ja/latest/

3　実験と結果

3.1　チャンクの構築方法

チャンクの構築方法の比較を行うために実験では，PDF からページ毎に取り出した文字列を以下の方法で切り出す．

- テキストを全てつなげた後 64 文字毎に切り出す手法 (文字数)
- テキストを全てつなげた後「。」で区切る手法 (文章)
- 上記 2 節で紹介した方法 (レイアウト)

3.2　実験の設定

3.2.1　使用したモデル

RAG の LLM のモデルには日本語の処理が期待できる Llama-3-ELYZA-JP-8B を使用する．

3.2.2　RAG の設定

チャンクの文章をベクトルにする際のモデルには多言語対応の埋め込みモデルである multilingual-e5-large を使用し，リトリーブする際の検索方法として，ベクトル検索を使用する．また LLM に渡すテキストにはランキング上位 4 つを選ぶようにした．

3.2.3　評価指標

評価は生成された回答と事前に作成された質問に対する回答との類似度を以下の評価指標を使用することで行った．

- BLEU
- ROUGE-1
- ROUGE-2
- ROUGE-L

3.2.4　データ

実験のデータには 74 ページからなる学内の事務手続き PDF を対象とする．また RAG の性能を評価するために質問と回答のセットをその PDF のうち本文部分から 67 個，表部分から 43 個それぞれ作成した．

以下に本文部分からの質問と回答の例を示す．

- 質問: 前泊・後泊等を行う際で、移動のみの日が休日の場合は？

表 1　本文の質問を使用した実験結果

	BLUE	ROUGE-1	ROUGE-2	ROUGE-L
文字数	0.1604	0.4259	0.2452	0.3462
文章	0.1901	0.4528	0.2708	0.3720
レイアウト	**0.2213**	**0.4674**	**0.3131**	**0.3946**

表 2　表の質問を使用した実験結果

	BLUE	ROUGE-1	ROUGE-2	ROUGE-L
文字数	0.1694	0.4340	0.2687	0.3861
文章	0.1831	0.4699	0.2881	0.4120
レイアウト	**0.3001**	**0.5514**	**0.4209**	**0.5250**

回答: 振替の対象にはなりません。

- 質問: 宿泊を伴う場合は？（宿泊費不要も含む）
  回答: 出張後、［出張報告入力］で宿泊先を入力してください。宿泊費不要の場合は［旅行者メモ］にも記入（例：兄の家に宿泊）し、ホテル以外の宿泊先にチェックを入れてください。
- 質問: 参照作成メニューで作成されるデータの伝票区分は概算払いですか?
  回答: いいえ、精算払いです。

続いて以下に表部分からの質問と回答の例を示す.

- 質問: メニューの国内出張はどのような機能ですか?
  回答: 国内出張データを新規作成するという機能です。
- 質問: 出張計画画面の伝票区分はどのような入力項目ですか
  回答: 精算払か概算払の選択を行う項目です。
  ※精算払の場合は選択不要
- 質問: 一時保存後の画面の伝票番号はどのような入力項目ですか?
  回答: 自動で番号が割り振られる項目です。

3.3　実験結果と考察

本文からの質問を使用した実験を表 1 に示す．また表からの質問を使用した実験を表 2 に示す.

本研究では，従来の単純なテキスト分割に加え，OCR レイアウト解析の結果を用いて文書内のタイトルや表を意識したチャンキング手法を導入した．表 1 および表 2 の結果から，以下の点が明らかとなった.

まず，表 1 より本文に対する質問であっても「レイアウト」アプローチがすべての評価指標（BLUE, ROUGE-1, ROUGE-2, ROUGE-L）において最も高いスコアを示している．これはタイトルや段落構造を考慮したチャンキングが，LLM の回答生成に必要な情報のまとまりをより最適化し，RAG による再現性や正確性を向上させているためと考えられる．また文字数単位で分割した場合や文章単位で分割した場合に比べ，タイトルという文書構造を無視しないことで，質問に必要な情報がより取得できていることが確認できる.

続いて，表 2 では表領域を含む質問に対してレイアウトを使用したチャンキング手法が顕著な効果を示した．BLUE や ROUGE-1, ROUGE-2, ROUGE-L のすべてにおいて数値が大幅に向上していることから，OCR レイアウト解析による表領域の抽出と Markdown 形式での保存が，表に対する質問に対して非常に有効であることが分かる．単に文字として結合する手法では，表の構造や見出しとの関連情報が失われやすいが，レイアウト情報を保持することで，回答時に必要な文脈を正しく活用できているものと考えられる.

一方で，今後の課題として今回扱った PDF は学内の事務手続き用ということもあり，レイアウト構造が比較的定型化されているという特徴がある．すなわち，タイトルや表の配置パターンがある程度一定であり，surya による OCR レイアウト解析が円滑に機能しやすい環境であったと言える．今後，企業の内部文書や他分野の報告書など，レイアウトが大きく変化しうる非定型的な PDF や，斜め配置といった複雑なレイアウトを含む文書においても，この手法が同程度の有効性を示すかが今後の課題と言える.

4　まとめ

本研究では，PDF 文書を扱う際に，surya の OCR を用いたレイアウト解析に基づくチャンキング手法が RAG の回答精度向上に有効であることを示した．具体的には，surya によるレイアウト解析を活用してタイトルや表の領域を判定し，タイトル情報と本文，表情報を個別にテキスト抽出・保存することで，従来の単純な文字数ベースや文章ベースの切り出し手法よりも優れた性能が得られることが実験結果から確認された.

学内の事務手続き PDF（74 ページ）を対象に Q&A ペア（本文と表の質問各々を複数用意）を作成し，RAG で回答を生成する実験を行ったところ，BLEU および ROUGE 各種（ROUGE-1, ROUGE-2, ROUGE-L）の指標において，レイアウトに基づくチャンキングが最も高いスコアを示した．特に表領

域に対する質問においては，表構造を Markdown 形式で保持することで文脈情報を失わずに参照できるため，回答生成が大きく向上することが示唆された．また，タイトルの情報を活用してテキストを再編成することで LLM が必要な背景知識を効率よく取り込めることも，回答の再現性や正確性を高める要因となっている．

一方で，本研究の対象文書は学内事務手続きという比較的定型化されたレイアウトを持つ PDF であり，企業の内部資料や斜め配置などを含む非定型的なドキュメントへの適用可能性は今後の検証が必要である．

またテキストや図をベクトル化して扱う先行研究 [4] と比較する必要がある．

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. https://arxiv.org/abs/2005.11401.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. https://arxiv.org/abs/2312.10997.

[3] Demiao Lin. Revolutionizing retrieval-augmented generation with enhanced pdf structure recognition, 2024. https://arxiv.org/abs/2401.12599.

[4] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding, 2024. https://arxiv.org/abs/2411.04952.