農林業基準技術文書を対象とした PDF 解析ツールの表構造認識の性能評価

中村彩乃 [1] 杉山陽菜乃 [1] 阿部瑞稀 [1] 前多陸玖 [1] 坂口遥哉 [1] 佐藤栄作 [1] 木村泰知 [1]
[1] 小樽商科大学
{g1202219307,g1202200907,g1202228307,g2021344,g2021163}@edu.otaru-uc.ac.jp
g1202453006@edu.otaru-uc.ac.jp kimura@res.otaru-uc.ac.jp

概要

本研究では，PDF 文書から表データを正確に抽出することを目的とし，主に農業分野で利用される技術文書を対象に主要企業の表抽出ツールを評価する．既存研究が指摘する固定レイアウト構造の課題を考慮し，複雑な表レイアウトや多言語への対応を踏まえて，表構造認識ベンチマーク TOITA を用いて性能を比較することで，その有用性を明らかにする．

1 はじめに

PDF 文書は，企業の統計資料や財務報告書のみならず，教育や農業，交通などの行政関連分野においても，情報共有や保存の標準フォーマットとして広く利用されている．特に農業分野では，生産計画や収支報告，技術継承など，多岐にわたる情報提供の手段として表形式データを含む PDF 文書が活用される場面が多く，これらのデータを効率的かつ正確に抽出することは，組織や業務のデジタル化をはじめとする新たな技術導入による効率向上においても不可欠な基盤である [1, 2]．また，PDF 文書からの情報抽出技術に関する研究が進められている [3, 4]．

しかしながら，PDF は作成者が意図したレイアウトを保持するように設計されたフォーマットであり，「表」のような構造化データの抽出が困難な場合もある [5]．特に，セル結合や非スカラー値などの複雑なレイアウト表現が用いられている場合，その抽出はさらに困難となる [6, 7]．一方，近年の AI 技術の急速な進歩により，機械学習や深層学習を活用した表抽出手法が，従来のルールベースや OCR ベース手法では困難とされてきた複雑なレイアウトや多言語への対応においても有効性を示しつつある．

そこで，本研究では，PDF に対応した表構造認識ツールを対象として，それらの表構造認識の精度を評価する．具体的には，PDF 形式の農業基準技術文書に含まれる表を対象とした表構造認識ベンチマークである Technical Obtainment and Interpretation of Tables in Agricultural document (TOITA)[8] を用いて各ツールの精度を評価することで，その有用性を明らかにする．

2 対象ツール

本研究は，表構造認識ツールの性能を評価することを目的としている．表構造認識とは，PDF や画像など固定フォーマット文書を入力とし，表情報を再構築して別フォーマットで出力する機能のことを指す．このタスクは，「表の検出」（table detection），「表構造の認識」（table structure recognition），および「表データの抽出」（table data extraction）の三つの段階に分けられる [9][10]．

- 表の検出：文章内における表の領域を特定する
- 表構造の認識：検出された表内のヘッダー，列，行，セルの関係性を特定する
- 表データの抽出：認識した構造に基づき，データを任意の形式に再構成し出力する

本研究では，この 3 つのプロセスを統合的に実行可能なアーキテクチャ，すなわち End-to-End 型のツールを採用している．これは，API 等を用いてこれらのプロセスをスクリプトで表現・実行可能なものも含んでいる．また，対象とするデータが日本語で記述されているため，日本語非対応のツールは選定対象から除外した．

表 1 PDF からの表構造解析が可能なツールの一覧

ツール名	有料 / 無料	インターフェース	ライセンス	採用 / 不採用
camelot [1]	無料	ライブラリ	MIT license	採用
Docling [2]	無料	ライブラリ	MIT license	採用
img2table [3]	無料	ライブラリ	MIT license	採用
pdfplumber [4]	無料	ライブラリ	MIT license	採用
PyMuPDF [5]	無料	ライブラリ	AGPL-3.0 license	採用
tabula-py [6]	無料	ライブラリ	MIT license	採用
YomiToku [7]	無料	ライブラリ	CC BY-NC 4.0	採用
Azure AI Document Intelligence [8]	有料	API	未確認	採用
Google Cloud Document AI [9]	有料	API	未確認	採用
Nanonets [10]	有料	API	未確認	採用
UPDF AI [11]	有料	GUI	未確認	採用
PDFium [12]	無料	ライブラリ	BSD-3-Clause license	不採用
Unstructured [13]	無料	ライブラリ	Apache-2.0 license	不採用
ABBYY [14]	有料	GUI	未確認	不採用
AlgoDocs [15]	有料	API, GUI	未確認	不採用
Asprise [16]	有料	API	未確認	不採用
Docparser [17]	有料	GUI	未確認	不採用
Docsumo [18]	有料	API	未確認	不採用
Eden AI [19]	有料	API	未確認	不採用
IBM Cloud Watson Discovery [20]	有料	API	未確認	不採用
OCR.space [21]	有料	API	未確認	不採用
OpenText Core Capture Services [22]	有料	API	未確認	不採用
PDF.co [23]	有料	API, GUI	未確認	不採用
PDF Element [24]	有料	GUI	未確認	不採用
Rossum [25]	有料	API	未確認	不採用
Tableau prep [26]	有料	GUI	未確認	不採用
UiPath Document Understanding [27]	有料	GUI	未確認	不採用

表 1 は，PDF や画像などのドキュメントを対象にテキスト抽出や内容解析を行うための各種ツールとサービスを整理した一覧である．本表では，各ツールの名称，料金形態，インターフェース，ライセンス，評価対象としての選定結果を比較可能な形式で提示している．具体的には，左列よりツール名，「無料」もしくは「有料」といった料金形態，「ライブラリ」や「API」「GUI」といったインターフェース，「MIT license」「AGPL-3.0 licens」などのライセンス情報，並びに本研究において評価の対象としたかどうかを「採用」「不採用」という形で表記した．不採用となったツールの背景としては，「End-to-End

1) https://github.com/atlanhq/camelot
2) https://github.com/DS4SD/docling
3) https://github.com/xavctn/img2table
4) https://github.com/jsvine/pdfplumber
5) https://github.com/pymupdf/PyMuPDF
6) https://github.com/chezou/tabula-py
7) https://github.com/kotaro-kinoshita/yomitoku
8) https://azure.microsoft.com/ja-jp/products/ai-services/ai-document-intelligence
9) https://cloud.google.com/document-ai
10) https://nanonets.com/
11) https://updf.com/jp/
12) https://github.com/PDFium/PDFium
13) https://github.com/Unstructured-IO/unstructured
14) https://pdf.abbyy.com/ja/
15) https://www.algodocs.com/
16) https://github.com/Asprise/
17) https://docparser.com/
18) https://github.com/docsumo/docsumo_python_client
19) https://github.com/edenai
20) https://github.com/IBM/watson-discovery-ui
21) https://github.com/topics/ocr-space
22) https://github.com/orgs/opentext/repositories
23) https://pdf.co/
24) https://github.com/streetturtle/pdf-element
25) https://github.com/RossumAI/
26) https://www.tableau.com/ja-jp/products/prep
27) https://www.uipath.com/

型のツールではない」「企業向けであり個人利用が対象外である」「表の範囲選択をユーザーに依存している」等の理由が挙げられる．

これらのツールは，無料で利用可能なオープンソースのライブラリから，有料のクラウド API サービスや商用製品にまで多岐にわたる．無料ツールでは，主に PDF ファイルを入力とし，ライブラリとして動作し，結果を CSV 形式で出力するものが多く存在する．また，GUI ツールやクラウドベースの解析サービスを提供する有料製品も含まれている．

3 TOITA ベンチマーク

TOITA ベンチマークとは，農業の技術継承にも用いられ表形式を多く含んでいる「長崎県農林業基準技術」の文書をベースに，各ツールの表構造認識精度，とりわけ構造認識精度とデータの抽出精度を評価するための正解データセットと評価ベンチマークである [8]．それぞれタイプの異なる表を含んだ 4 つの PDF ファイルと，そこに含まれる計 70 個の表を対象としている．

3.1 評価方法

TOITA ベンチマークの評価方法に倣い，表構造認識の評価には GriTS[11] を使用する．入力は HTML で記述された表であり，属性はセル結合を示す rowspan 属性と colspan 属性のみが考慮される．データセットに含まれる表それぞれについて GriTS スコアを算出し，それらを用いて算出した F1 スコアを最終的な評価値とする．

3.2 GriTS

以下に GriTS の算出方法を記すが，詳細は引用元の論文を参照されたい．$\mathbf{A}$ を抽出対象である表の行列，$\mathbf{B}$ を抽出した表の行列とすると，GriTS は以下の式で算出される．

$$\mathrm{GriTS}_f(\mathbf{A},\mathbf{B}) = \frac{2\sum_{i,j} f(\hat{\mathbf{A}}_{i,j},\hat{\mathbf{B}}_{i,j})}{|\mathbf{A}|+|\mathbf{B}|}$$

ここで，$\hat{\mathbf{A}}$，$\hat{\mathbf{B}}$ はそれぞれ 2 次元最大類似部分構造 (2D-MSS) であり，以下の式で算出される．

$$\begin{aligned}\text{2D-MSS}_f(\mathbf{A},\mathbf{B}) &= \arg\max_{\mathbf{A}'|\mathbf{A},\mathbf{B}'|\mathbf{B}} \sum_{i,j} f(A'_{i,j},B'_{i,j}) \\ &= \tilde{\mathbf{A}},\tilde{\mathbf{B}},\end{aligned}$$

GriTS は，入力する行列の形態を変更することで，構造評価（GriTS Toporogy）と内容評価（GriTS Content）を同一のアルゴリズムで評価することが可能である．

3.2.1 GriTS Topology

GriTS Topology（$\mathrm{GriTS}_{\mathrm{Top}}$）は，セルの構造を評価するための指標である．正解データと予測データの表をグリッド化した際の，セルの「サイズ」と「相対位置」を示す情報の行列を入力として，$\mathrm{GriTS}_{\mathrm{Top}}$ を計算する．

3.2.2 GriTS Content

GriTS Content（$\mathrm{GriTS}_{\mathrm{Con}}$）は，セルの内容を評価するための指標である．正解データと予測データの表をグリッド化した際の，各セルに含まれるテキストの行列を入力として，$\mathrm{GriTS}_{\mathrm{Con}}$ を計算する．文字列間の類似度の計算には，正規化された最長共通部分列（LCS）を用いる．

4 表構造解析とその評価

4.1 実験手順

TOITA ベンチマークのデータセットに含まれる 4 つの PDF を対象として，表構造解析ツールを適用した．出力内容を確認し，データセットの正解データとの対応付けを人手で行なった．出力が CSV 形式であるものに対しては，HTML 形式に変換して予測データを作成した．正解データと予測データを用いて GriTS スコアを求め，最終的な評価値である $\mathrm{GriTS}_{\mathrm{Top}}$ と $\mathrm{GriTS}_{\mathrm{Con}}$ のそれぞれの F1 スコアを算出した．

4.2 評価結果

各表構造認識ツールの評価結果について，$\mathrm{GriTS}_{\mathrm{Con}}$ の F1 スコアについて降順でソートし，表 2 に示す．6 つ全ての指標において，YomiToku が最良のスコアとなった．その理由として，YomiToku は日本語文書に特化したツールであることや，出力がセル結合の情報を含んだ HTML 形式であるため，セル結合の情報が GriTS のスコアに反映されたということが考えられる．次ぐ Azure Docuent Intelligence と Nanonets では，構造認識（$\mathrm{GriTS}_{\mathrm{Top}}$）についての差はほとんど見られないものの，内容認識（$\mathrm{GriTS}_{\mathrm{Con}}$）については 3 ポイントほどの差が見られる．

表 2 評価結果

tool	top_recall	top_precision	top_F1	con_recall	con_precision	con_F1
YomiToku	**0.8943**	**0.8824**	**0.8883**	**0.8430**	**0.8318**	**0.8373**
Azure AI Document Intelligence	0.8098	0.7990	0.8044	0.7915	0.7810	0.7862
Nanonets	0.8099	0.7991	0.8045	0.7634	0.7532	0.7583
Docling	0.7260	0.7163	0.7211	0.6852	0.6760	0.6806
Google Cloud Document AI	0.6554	0.6466	0.6510	0.6080	0.5999	0.6039
UPDF	0.7305	0.7208	0.7256	0.6027	0.5947	0.5987
pdfplumber	0.6496	0.6409	0.6452	0.5437	0.5365	0.5401
PyMuPDF	0.6186	0.6104	0.6144	0.4902	0.4837	0.4870
tabula	0.5231	0.5162	0.5196	0.4172	0.4116	0.4144
camelot	0.5106	0.5038	0.5072	0.3837	0.3786	0.3811
img2table	0.7619	0.7517	0.7568	0.3278	0.3234	0.3256

大半のツールにおいて，$\text{GriTS}_{\text{Top}}$ と $\text{GriTS}_{\text{Con}}$ の間で半分以上の差はついていないものの，img2table では半分以上の差がついている．img2table は OCR の処理を複数のライブラリから選択して適用することができる（本研究では tesseract-ocr を使用）ため，本ツールの $\text{GriTS}_{\text{Con}}$ は OCR ライブラリの性能に依存する点に留意すべきである．

5 おわりに

本稿では，PDF に対応した表構造認識ツールを対象として，TOITA ベンチマークを用いてそれらの表構造認識の精度を評価した．6 つの項目全てで YomiToku が最良のスコアを計測し，Azure AI Document Intelligence と Nanonets が後に続く結果となった．

謝辞

本研究は，JSPS 科研費 21H03769 および，内閣府「研究開発と Society 5.0 との橋渡しプログラム（BRIDGE）」における農林水産省実施施策「AI 農業社会実装プロジェクトの助成を受けて実施された．また，協力頂いた合同会社 Nita Limo に感謝します．

参考文献

[1] 小林暁雄, 坂地泰紀, 桂樹哲雄, 森翔太郎, 橋本祥, 鈴木雅弘, 川村隆浩. 普及指導員の知識を回答可能な生成 ai のための農産物市場価値を表現するデータセットの構築. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 3M5OS12b01–3M5OS12b01, 2024.

[2] 杉山陽菜乃, 阿部瑞稀, 中村彩乃, 前多陸玖, 坂口遥哉, 佐藤栄作, 木村泰知, 小林暁雄, 大友将宏, 石原潤一, 桂樹哲雄, 川村隆浩. 農林業基準技術に含まれる表を対象とした pdf から csv へ変換する際の課題分析. 言語処理学会第 31 回年次大会, 2025.

[3] Lei Sheng and Shuai-Shuai Xu. Pdftable: A unified toolkit for deep learning-based table extraction, 2024.

[4] Burcu Yildiz, Katharina Kaiser, and Silvia Miksch. pdf2table: A method to extract table information from pdf files. In **Indian International Conference on Artificial Intelligence**, 2005.

[5] Andrey Mikhailov and Alexey Shigarov. Page layout analysis for refining table extraction from pdf documents. In **2021 Ivannikov Ispras Open Conference (ISPRAS)**, pp. 114–119, 2021.

[6] 奥山和樹, 木村泰知. 有価証券報告書を対象とした機械判読が困難な表構造の分析. 言語処理学会第 30 回年次大会 (NLP2024), pp. P3–20–, 3 2024.

[7] 前多陸玖, 奥山和樹, 佐藤栄作, 木村泰知. 有価証券報告書を対象とした機械判読が困難な表のセル分類に向けて. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 3M5OS12b03–3M5OS12b03, 2024.

[8] 阿部瑞稀, 杉山陽菜乃, 中村彩乃, 前多陸玖, 坂口遥哉, 佐藤栄作, 木村泰知. Pdf 形式の農業技術文書を用いた表構造認識ベンチマーク toita. 言語処理学会第 31 回年次大会, 2025.

[9] Max Goebel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. A methodology for evaluating algorithms for table understanding in pdf documents. pp. 45–48, 09 2012.

[10] Tapio Elomaa. Anssi nurminen algorithmic extraction of data in tables in pdf documents. 2013.

[11] Brandon Smock, Rohith Pesala, and Robin Abraham. Grits: Grid table similarity metric for table structure recognition, 2023.