

農林業基準技術文書を対象としたPDF解析ツールの表構認識の性能評価

中村彩乃 杉山陽菜乃 阿部瑞稀 前多陸玖 坂口遥哉 佐藤栄作 木村泰知
(小樽商科大学)

背景・目的

農業の技術継承の課題解決のため、農業基準技術文書に含まれる表を PDF形式から機械判読可能な形式へ変換する。
①機械判読可能な形式に変換する際の課題を明らかにする。(Q7-20)
②各ツールの表構造認識の精度を評価するためのベンチマークTOITAを構築する。(Q7-14)
③ベンチマークTOITAを用いて各表抽出ツールの性能を比較・評価する。
農業分野では情報提供の手段として表形式を含むPDF文書が多く活用される。しかし複雑なレイアウト表現が用いられている場合、これらのデータを正確に抽出することは困難となる。一方、AI技術の進歩により複雑なレイアウトや多言語に対応する表構造認識ツールが開発されている。

本研究の目的 表構造認識ベンチマーク TOITAを用いて主要企業の表抽出ツールの性能を比較・評価する。

評価方法

表構造認識ベンチマーク TOITAを用いて各表抽出ツールの性能評価を行う。
セル構造とセル内容それぞれの抽出精度について、GriTSを用いて評価を行う。

入力: TOITAベンチマークに含まれるPDF

区分	収量 (kg/10a)	ぶどう (巨峰)		
		窒素	リン酸	加里
幼木	未結果樹	0	2	2
	結果樹	500	6	6
成木		1,200	8	8

正解: PDFから人手で作成された表

区分	収量 (kg/10a)	ぶどう (巨峰)		
		窒素	リン酸	加里
幼木	未結果樹	0	2	2
	結果樹	500	6	6
成木		1,200	8	8

出力: 表構造認識ツールによる抽出結果

比較

評価

TOITAベンチマーク

長崎県農業基準技術文書をベースとした各ツールの表構造認識の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。(詳細はQ7-14参照)

GriTS

表構造認識の評価手法として使用する。(詳細はQ7-14参照)

GriTS Topology(GriTS Top) : セル構造の評価指標。

GriTS Content (GriTS Con) : セル内容の評価指標。

対象ツール

以下の表構造認識ツールの性能評価を行う。

PDFからの表構造認識解析が可能なツール一覧

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

表構造認識

PDFや画像など固定フォーマット文書を入力として、表情報を再構築して別フォーマットで出力する機能である。

PDF → CSV

表の検出 → 表構造の認識 → 表データの抽出

対象ツール

採用

日本語対応である。

表構造認識の3つのプロセスを統合的に実行可能なEnd-to-End型ツールである。

不採用

End-to-End型ではない。

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。

表構造解析とその評価

評価結果

tool	構造認識の評価			内容認識の評価		
	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

Yomitokuが最良スコア

出力がセル結合の情報を含んだHTML形式である。

日本語文書に特化したツールである。

AzureDocuentIntelligence,Nanonets

構造認識の差はないが、内容認識に多少の差がある。

大半のツール

構造認識 > 内容認識 (半分以下の差)

img2table

構造認識 > 内容認識 (半分以上の差)

→OCRの処理を複数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GriTSConはOCRライブラリの性能に依存する。