

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

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
(小樽商科大学)

背景・目的

農業の技術継承にも用いられ、表形式データを数多く含む「長崎県農林業基準技術」の文書を基に、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、および評価ベンチマークである Technical Obtainment and Interpretation of Tables in Agricultural document (TOITA) を構築した

農業における技術継承の課題に対する取り組みとして、RAG (Retrieval-Augmented Generation) を用いた方法が検討されている。RAGにおいて「表」を含む文書が、正しく表構造認識されずに利用されることで、ハルシネーションを引き起こす要因となる。表の形式で記述されることが多い農業分野の文書においてはRAGへの適用が困難であるため、文書に含まれる表構造を正確に認識・変換する処理が重要となる。

本研究の目的

本研究では、農業分野のPDF文書をRAGに適用するため、PDF文書内の表構造を認識および変換可能なツールから最適なものを選定することを目指している。そのために、表構造認識ツールの性能を評価するベンチマークを構築することを目的とする。

表構造を評価する難しさ

横枠無し

経営類型	家族労働力	品目	技術の特徴
ひまわり	2 人	ひまわり	単棟ハウス
経営目標	1 総収入 1,597 千円 2 経営費 1,178 千円		

横枠無し

セル結合

非スカラー値

横枠が無い、セル結合がある、非スカラー値が含まれるセルといった問題があり正しく表構造認識されにくい

表構造認識を行うと仮定した場合

正解データの例

経営類型	家族労働力	品目	技術の特徴
ひまわり	2 人	ひまわり	単棟ハウス
経営目標	1 総収入 1,597 千円 2 経営費 1,178 千円		

予測データの例

経営類型	家族労働力	品目	技術の特徴
ひまわり	2 人	ひまわり	単棟ハウス
経営目標	1 総収入 1,597 千円 2 経営費 1,178 千円		

予測した際に欠落したデータ

- 農林業基準技術の文書に含まれる70種類の表 に対して、人手による**正解データ**を作成した
- 長崎県農林業基準技術の文書を基に、表構造理解のデータセットを構築した
- 表構造認識の評価手法GriTSを用いた**ベンチマークTOITA**を構築した

表構造認識ベンチマークTOITA

使用データ

- 令和6年度「長崎県農林業基準技術」のPDF
 - 表の内容やフォーマットを考慮し、以下の**4ファイル**を使用。
- | ファイル名 | テーブル数 | ページ数 |
|---------------|-------|------|
| 経営類型・単一品目 | 5 | 3 |
| 品目別施肥基準 | 35 | 21 |
| 長崎県農林業技術の考え方 | 21 | 10 |
| 農耕地土壌の概要と土壌管理 | 9 | 13 |
- 対象となる全ての表に対して、**人手アノテーション**により正解データを作成。

表構造認識の入出力

入力：PDFファイル

ぶどう
I 巨峰
1. 年間施肥量
1) 幼木・成木・収量別施肥量

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

2. 施肥時期及び割合

施肥時期	窒素	リン酸	加里
基肥 11月	60	100	60
追肥 5月下～6月上旬	20	—	20
礼肥 9月上～中旬	20	—	20

3. 摘要

1) 土壌の通気性や保水性を高めるための土づくり年次計画を立てて、確実に実行していく必要がある。

出力：ツールの予測結果 (HTML)

区分	None	収量 (kg/10a)	ぶどう (巨峰)
None	None	None	窒素 リン酸 加里
幼木	未結果樹	0	2 2 2
	結果樹	500	6 6 6
成木	None	1,200	2008 8 8

施肥時期

施肥時期	窒素	リン酸	加里
基肥 11月	60	100	60
追肥 5月下～6月上旬	20	—	20
礼肥 9月上～中旬	20	—	20

出力形式は、セル結合を反映できるHTML形式を想定している。

評価

GriTS

- 表構造認識で抽出した表に対し、行列として直接的に評価するフレームワーク
- セルの位置認識、**構造認識**、**内容認識**を同一のフレームワークで評価可能
- TEDS (Tree-Edit-Distance-based-Similarity) と比較して、行と列を同等に評価できる点で優れる

GriTS Topology (GriTS Top)

- セルの構造認識を評価する指標

GriTS Content (GriTS Con)

- セルの内容認識を評価する指標

評価の流れ

①ツールの予測結果を用意

区分	None	収量 (kg/10a)	ぶどう (巨峰)
None	None	None	窒素 リン酸 加里
幼木	未結果樹	0	2 2 2
	結果樹	500	6 6 6
成木	None	1,200	2008 8 8

②対応する正解データを紐付け

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 8

Aを正解行列、Bを予測行列とすると、GriTSは以下の式で算出可能である

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

$\mathbf{A}^\wedge, \mathbf{B}^\wedge$ 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

$$\begin{aligned} 2\text{D-MSS}(\mathbf{A}, \mathbf{B}) &= \arg \max_{\mathbf{A}^\wedge | \mathbf{A}, \mathbf{B}^\wedge | \mathbf{B}} \sum_{i,j} f(\mathbf{A}^\wedge_{i,j}, \mathbf{B}^\wedge_{i,j}) \\ &= \hat{\mathbf{A}}, \hat{\mathbf{B}} \end{aligned}$$

③GriTS Top、Conを算出

④F1スコアを算出し、最終的なスコアを確定