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 [1] 杉山陽菜乃 [1] 中村彩乃 [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，農業の技術継承にも用いられる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検証するための正解データセットを作成し，さらに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．

1 はじめに

農業の技術継承は，高齢化が進む日本において，極めて重要な課題となっている [1]．技術継承の問題としては，高齢化や後継者不足に加え，地域特有の作物の栽培技術や土壌管理法など，経験に基づく技術が多く存在すること，デジタル技術との融合が進んでいないことなどが挙げられる．このような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 [2]．RAG とは大規模言語モデルにおける追加学習を伴わずに外部知識を活用する手法のことで，外部文書をベクトルに変換し，データベースに保存することで検索を可能にしている．これによりハルシネーションのリスクを軽減させ，信頼性の高い出力を期待できる．しかしながら，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可能性がある．特に，栽培技術や土壌管理・作付け体系など，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 [3]．そのため，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．ICDAR2013 の Table competition では，PDF からの表の位置検出と構造認識タスクが提案されたが，データセット含まれていない日本語データに対する認識精度は明らかでない [4]．現状では，日本語に対応した表構造認識を行うツールは数多く存在するものの，各ツールの性能を体系的に検証する手段は確立されていない．

そこで，本研究では，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．

本研究の貢献は，以下の 3 点である．

- 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

2 表構造認識の評価手法

表構造認識の精度を評価するためには，抽出対象の表と抽出した表の一致度を適切に測定する必要がある[1)]．一致度を測定する際の観点は，以下の 3 点が挙げられる．

1. セルの構造
2. セルの内容
3. セルの位置

本研究では，表構造認識を行った後の出力をデータソースとして利用することを前提としているため，データの構造や内容に影響がないと思われる 3. セルの位置については評価を行わず，1. セルの構造と 2. セルの内容についてのみ評価を行う．

1) https://nanonets.com/blog/the-ultimate-guide-to-assessing-table-extraction/

ただし，PDF 上に存在する表の多くは，見た目に重きが置かれており，表構造認識はもちろんその評価も複雑なものになる．図 1 は，長崎県農林業基準技術の PDF に存在する表を一部抽出し，簡略化したものである．ここに示された，横枠がない表に対して表の範囲を正確に推測できるかという点や，非スカラー値（一つのセルに複数の情報が含まれている状態）のセルを正確に分割できるかという点は，表構造認識の難しさであると同時に，適切に評価すべきポイントでもある．

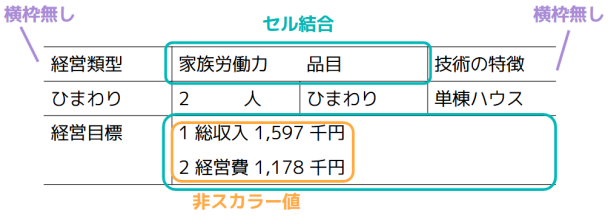


図 1　表構造を評価する際の難しさ

図 2 は，図 1 で示した表に対して表構造認識を行うと仮定した場合の，正解と予測データの例である．認識対象の表に枠線がないことから，予測データは両端の列が欠損してしまっている．加えて，「2 人」のセルが分割されてしまっている，文字を誤って認識してしまっているというミスも窺える．

図 2　表構造認識ツールをと正解の例

以降，図 2 で示した正解データと予測データを例に取り，セルの構造と内容の評価について説明を行う．

2.1　セルの構造の評価

セルの構造については，正解データの行数と列数，同様に予測データの行数，列数を用いることにで，行の正解率（rows_accuracy）と列の正解率（cols_accuracy）を求めることができる．それらの調和平均をとることで，セルの構造の正解率（cells_shape_accuracy）を算出し，評価値として用いることができる．

1. rows_accuracy

$$= 1 - \frac{|\text{rows}_{truth} - \text{rows}_{pred}|}{max(\text{rows}_{truth}, \text{rows}_{pred})}$$

2. cols_accuracy

$$= 1 - \frac{|\text{cols}_{truth} - \text{cols}_{pred}|}{max(\text{cols}_{truth}, \text{cols}_{pred})}$$

3. cells_shape_accuracy

$$= \frac{2 \times \text{rows\_accuracy} \times \text{cols\_accuracy}}{\text{rows\_accuracy} + \text{cols\_accuracy}}$$

正解データ，予測データについてこれらを計算すると，rows_accuracy は $\frac{3}{4}$，cols_accuracy は $\frac{3}{4}$，となり，cells_shape_accuracy は $\frac{3}{4}$ となる．しかし，この事例において rows_accuracy と cols_accuracy が同値になることの妥当性には疑問を抱かざるを得ない．事実，この手法ではセル結合を考慮することはできない．

2.2　セルの内容の評価

セルの内容については，セルを特定の順序で並べ，内容が一致するかどうかを測ることで評価を行うことができる．図 3 では，各データのセルの内容を左上から右下にかけて一列に並べたもので，完全一致したものをカウントしている．

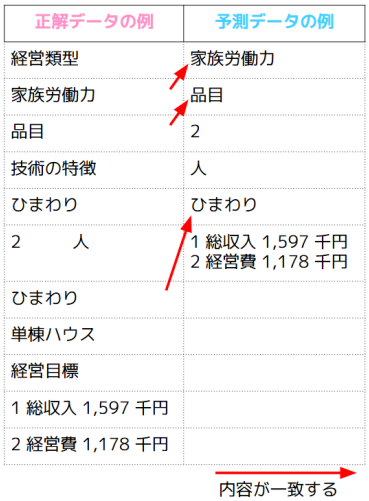


図 3　内容評価の例

- Precision

$$= \frac{\text{予測した中の正解セル数}}{\text{予測したセル数}}$$

- Recall

$$= \frac{\text{予測した中の正解セル数}}{\text{正解セル数}}$$

• F1 Score

$$= \frac{2 \times Precision \times Recall}{Precisiion + Recall}$$

上記に倣い，F1 スコアを求めることで，セルの内容の評価値として用いることができる．件のデータでは，Precision が $\frac{3}{6}$，Recall が $\frac{3}{11}$，F1 Score が $\frac{6}{17}$ となる．ただし，この手法は完全一致を前提とすることや，行や列の順番がシャッフルされていても完全一致をとれてしまうという問題を抱えている．

2.3 セルの構造と内容の評価

Tree-Edit-Distance-based-Similarity(TEDS)[5] は，HTML で表現した表について編集距離を求めることにより，セルの構造と内容の両方を考慮することが可能な指標である．HTML には，セル結合を示す属性として rowspan 属性と colspan 属性が含まれるため，セル結合を考慮することが可能である．また，HTML のツリー構造を利用することで行や列の順番を考慮できる点や，セルの内容を編集距離で評価することにより完全一致を前提としない点で優れている．TEDS は下記の式で算出することができる．

$$\mathrm{TEDS}(T_a, T_b) = 1 - \frac{\mathrm{EditDist}(T_a, T_b)}{\max(|T_a|, |T_b|)}$$

ただし，理想的な指標と思われる TEDS にも，行と列を同等に扱うことができないという問題点が存在する．というのも，欠落（追加）したセルの数が同じであっても，行が欠落した場合と列が欠落した場合とでスコアに差が生じてしまうのである．

これらの問題を解決した手法として，Grid table similarity(GriTS)[6] が挙げられる．GriTS は，行と列を同等に扱うほか，部分一致を許容し，セルの絶対位置に依存しないという点で優れた評価指標である．その詳細については，第 4.1 節で説明する．本研究で提案するベンチマークでは，表構造の評価指標として $\mathrm{GriTS}_{\mathrm{Top}}$ を，表の内容の評価指標として $\mathrm{GriTS}_{\mathrm{Con}}$ を採用する．

3 TOITA データセット

本研究で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である Technical Obtainment and Interpretation of Tables in Agricultural document (TOITA) を作成した．

3.1 目的

本研究における表構造ツールの評価ベンチマークの作成目的は，各表構造認識ツールが PDF 文書に現れる表構造を適切に認識・変換し，RAG への入力により適したものにすることであった．そのため，正解データセットの作成もそれに則している必要がある．正解データセットの作成においては，PDF 上の表を再現できるように作成するのではなく，正規化してデータベースに格納できるようにすることを目的とする．

3.2 対象となるデータ

表 1 対象となる PDF ファイルごとの統計情報

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

図 1 は，対象となる PDF ファイルごとの統計情報である．収集した令和 5 年度の長崎県農林業基準技術の PDF から，データセットとして使用するファイルを選定する．選定する上での留意点として，表フォーマットの偏りが出ないようにする必要がある．そのため，正解データセットの対象として PDF ファイルの「経営類型・単一品目」「品目別施肥基準」「長崎県農林業技術の策定に伴う基本的な考え方」「農耕地土壌の概要と改良基準，土壌管理」からそれぞれ PDF ファイルを 1 つずつ使用する．これにより各 PDF ファイルに含まれる表フォーマットに多様性を持たせ，より実用的なデータセットを作成した．

3.3 アノテーション

作成したデータセットに対して，Google スプレッドシート上で正解データを人手で作成する．作成時には，PDF 上の表を再現するのではなく，データベースに格納しやすい形式に変換する．図 2 の上の表は作成した正解データであり，下の表は PDF に含まれる表を表構造認識ツールを用いて CSV 変換したものである．

正解データの作成では，「セル結合の分割」と「非スカラー値の分割」に注意し作業を行った．「セル結合の分割」においては，正規化した後の検索の問題から，一つのセルではなく，分割してセル分けが必要だと判断した場合に行った．「非スカラー値の

分割」においては，header に対応せる内に複数の値が含まれている場合は，値をそれぞれを分割し一つのセルごとに格納した．これらの作業を行い，計 70 個のテーブルデータとそれに対応する正解データを作成した．

4　表構造認識の評価

4.1　GriTS

本ベンチマークでは，表構造認識の評価指標として GriTS[6] を使用する．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．

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

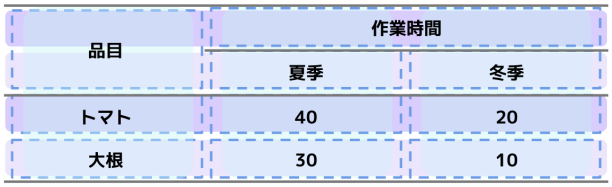
GriTS は，入力する行列の形態を変更することで，構造評価と内容評価を同一のアルゴリズムで評価することが可能である．例えば，図 4（A）で示す表を評価するにあたり，構造を評価したい場合は（B）で示す行列，内容を評価したい場合は（C）で示す行列に変換したものを入力とする．

4.1.1　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（図 4（B））を入力として，$\mathrm{GriTS}_{\mathrm{Top}}$ を計算する．

4.1.2　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（図 4（C））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．



品目	作業時間	
	夏季	冬季
トマト	40	20
大根	30	10

（A）GriTSによる評価で用いる表の例

[0, 0, 1, 2]	[0, 0, 2, 1]	[-1, 0, 1, 1]
[0, -1, 1, 1]	[0, 0, 1, 1]	[0, 0, 1, 1]
[0, 0, 1, 1]	[0, 0, 1, 1]	[0, 0, 1, 1]
[0, 0, 1, 1]	[0, 0, 1, 1]	[0, 0, 1, 1]

（B）$\mathrm{GriTS}_{\mathrm{Top}}$の入力

品目	作業時間	作業時間
品目	夏季	冬季
トマト	40	20
大根	30	10

（C）$\mathrm{GriTS}_{\mathrm{Con}}$の入力

図 4　GriTS の入力例

4.2　評価の流れ

評価は以下の流れで行われる．

1. 予測データを用意する
2. 正解データと内容の一致する予測データの組を人手で作成する
3. 正解データと予測データの組を用いて GriTS を算出する
4. 再現率と適合率の調和平均をとり，最終的な評価スコアを算出する

5　おわりに

本稿では，表構造認識ツールが PDF 文書に現れる表構造を適切に認識・変換し，RAG への入力により適したものにすることを目的として，TOITA ベンチマークを提案した．長崎県農林業基準技術の PDF ファイルから，表のフォーマットを考慮して 4 つのファイルを選定し，それに含まれる 70 の表に対して人手による正解データを作成した．評価手法には GriTS を採用し，入力を HTML 形式で記述された表形式データ（正解データと予測データ）として，$\mathrm{GriTS}_{\mathrm{Top}}$ と $\mathrm{GriTS}_{\mathrm{Con}}$ を出力とした．

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] Patrick S. H. 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. **CoRR**, Vol. abs/2005.11401, , 2020.

[3] 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.

[4] Max C. Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. Icdar 2013 table competition. In **2013 12th International Conference on Document Analysis and Recognition (ICDAR)**, pp. 1449–1453, Los Alamitos, CA, USA, aug 2013. IEEE Computer Society.

[5] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation, 2020.

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