3Yin-A-17

DAGRI SubTask-1
農業PDF文書を対象とした情報抽出データセットの拡張

熊倉 梨央[1]、木村 泰知[1]、森岡 幹[2]、小林 暁雄[3]、大友 将宏[3]、石原 潤一[3]、馬場 研太[3]、桂樹 哲雄[3]
(1. 小樽商科大学、2. AIREV株式会社、3. 国立研究開発法人　農業・食品産業技術総合研究機構)

背景・目的

*https://sites.google.com/view/dagri/home-ja
2026年3月 - 2026年6月: プレ・コンペティション/ Dry Run
2026年7月: コンペティション/ Formal Run
2026年12月8日 - 2026年12月10日: NTCIR-19 カンファレンス（NII, 東京）

1.既存研究と課題

・文書を統一フォーマットに変換する
Table Information Extraction (Table IE) が提案され、NTCIR-19DAGRI*のサブタスクコンペティションが実施されている
➡**対象自治体が北海道と長崎県に限定されており、手法の汎用性評価が十分ではない**

2.産業分野の課題

・担い手の高齢化と減少が深刻
・基幹的農業従事者の平均年齢は約69歳に達し20年間で半減
➡**知識・経験の継承が急務**

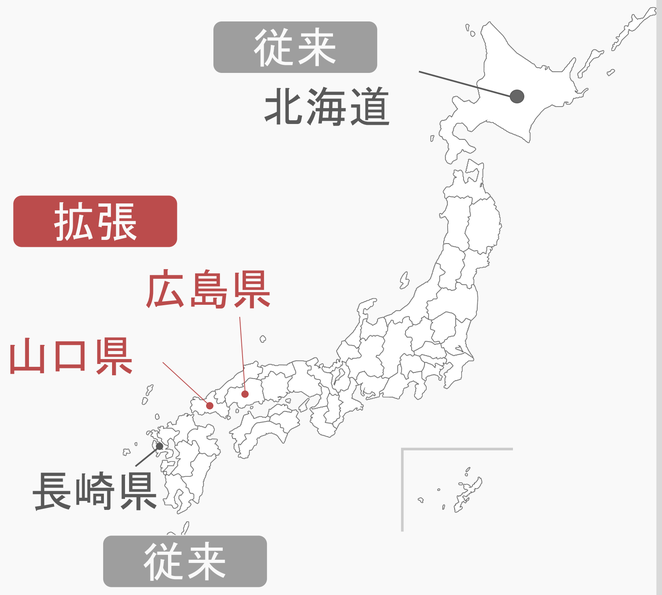
3.標準農業技術文書の重要性と課題

・PDF形式で図・写真・表が混在する非構造データ
・ガントチャートやセル結合を含む複雑名図表
・自治体ごとに文書構成や記載形式が異なる
➡**LLMによる機械的な読解や活用が困難**

本研究の目的

多様なPDF形式や記載構造に対する情報抽出手法の汎化性能を評価するために対象自治体を拡張したデータセットを構築する

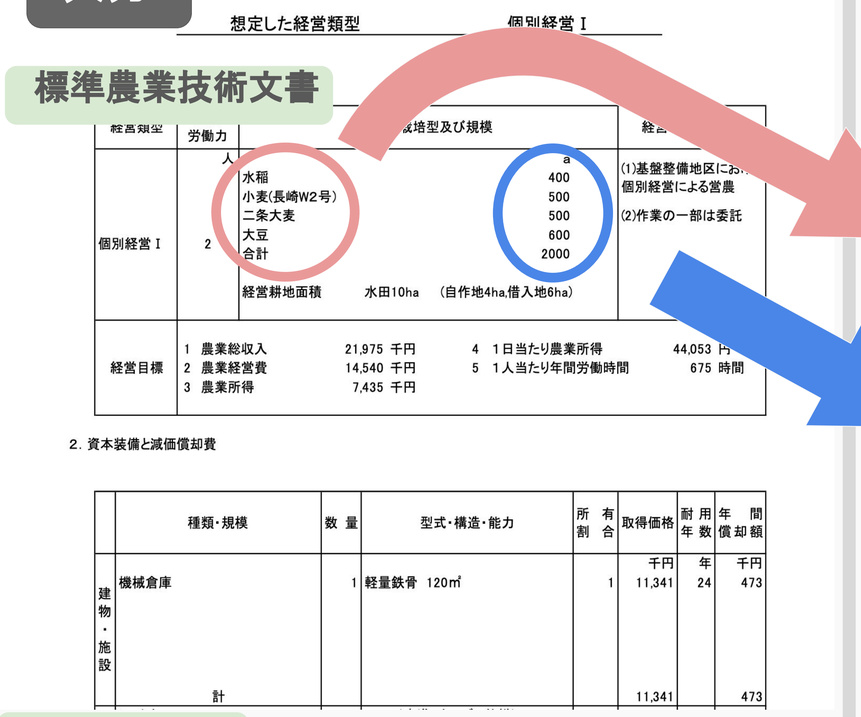
データセットの拡張



＊2自治体 → 4自治体へ拡張

入力

標準農業技術文書



想定した経営類型　個別経営Ⅰ

経営類型	労働力	栽培型及び規模		経営
	人		a	
個別経営Ⅰ	2	水稲	400	(1)基盤整備地区にお… 個別経営による営農
		小麦(長崎W2号)	500	
		二条大麦	500	(2)作業の一部は委託
		大豆	600	
		合計	2000	
		経営耕地面積	水田10ha （自作地4ha,借入地6ha）	

経営目標				
1	農業総収入	21,975 千円	4 1日当たり農業所得	44,053 円
2	農業経営費	14,540 千円	5 1人当たり年間労働時間	675 時間
3	農業所得	7,435 千円		

2. 資本装備と減価償却費

	種類・規模	数量	型式・構造・能力	所有割合	取得価格	耐用年数	年間償却額
					千円	年	千円
建物・施設	機械倉庫	1	軽量鉄骨 120㎡	1	11,341	24	473
	計				11,341		473

タスク指示書

例）**入力について**：経営指標01～03（PDF）
出力について：経営類型の各項目を構造化
構造化するときの注意点：有機栽培のみ抽出
（慣行栽培は除外）

出力

```
{
  # target prefecture
  "prefecture_name": "長崎県",
  # a list of crops to cultivate in this premise
  "crop_names": ["水稲", "麦類", "秋大豆"],
  # areas of cultivated land
  "cultivated_land": 1000,
  "cultivated_land_unit": "a",
  "borrowed_cultivated_land": 600,
  "owned_cultivated_land": 400,
```

検索・情報抽出・構造化

前提表

項目	値
耕地	2000a
労働力	2名
所得	7435万円

栽培規模

作物名	耕地面積
水稲	400a
麦類	1000a
秋大豆	600a

主な課題と対応方法

① 情報量の差異 ➡ **共通項目を基準に必要な情報を整理**

公開者	経営指標				経営類型			その他
	作業技術	作業時間	経営収支	資本装備	前提・栽培規模	経営収支	資本装備	印刷スキャン
北海道	○	○	○	×	○	○	×	○
長崎県	○	○	○	×	○	○	×	×
広島県	○	○	○	○	○	○	○	○
山口県	○	○	○	○	×	×	×	×

② 単位の不統一 ➡ **作物単位を10a基準に統一**

	山口県	広島県
作物	水稲	キャベツ
面積	10a	1ha
粗収益	101,250円/10a	3,670,000円/1ha
経営費	90,009円/10a	3,487,391円/1ha

※10a当たりの数値へと変換が必要

③ 作業技術一覧の様式 ➡ **項目ごとに再構成し統一フォーマットに変換**
＊同じ「作業技術一覧」でも、県によって記載の粒度や情報構造が異なる

山口県の特徴	広島県の特徴
・各作業が一行ごとに**独立**して記述 ・上から順に実行できる**手順重視**の構造 ・情報は**簡潔**で作業時間や労働力などの**定量情報は含まれない**	・各工程ごとに**複数**の情報項目を付与 ・**多層的**な構造 ・情報力が多く、労働投入量など**定量的な判断が可能**

自治体数の拡張により、さまざまな文書構造やフォーマットのデータが追加された