DAGRI Subtask 1: 複雑な図表を含む農業文書から情報抽出・構造化するタスクの提案

森岡 幹 [1]　熊倉 梨央 [2]　木村 泰知 [2]　石原 潤一 [3]
大友 将宏 [3]　小林 暁雄 [3]　馬場 研太 [3]　桂樹 哲雄 [3]
[1]AIREV 株式会社 [2] 小樽商科大学 [3] 農研機構 農業情報研究センター
tmorioka@airev.co.jp　g1202212907@edu.otaru-uc.ac.jp　kimura@res.otaru-uc.ac.jp
{ishihara.junichi964, ohtomo.masahiro841}@naro.go.jp
{akio.kobayashi, baba.kenta285, katsuragi.tetsuo}@naro.go.jp

概要

農作物の栽培に関する技術文書には，作業手順やスケジュール，使用する機材や資材，および営農に際する収支構造など，農業に関する知識が体系的に記載されているが，ガントチャートやセル結合されている表など複雑な図表が含まれる，技術文書を作成する都道府県ごとにフォーマットやファイル形式が異なる等，ＡＩ活用の障壁となっている．この状況を大規模言語モデルをはじめ技術的に解決できるか調査・検証するため，複雑な図表を含む技術文書から必要な情報を認識・抽出し，統一フォーマットに変換・構造化するタスクを提案し，タスク概要，入出力データの仕様について解説する．

1　はじめに

大規模言語モデル (LLM) が台頭し，非構造化データ，半構造化データに対する解析・活用への障壁が下がっている．農業分野においても活用可能性がある非構造データが存在しており，特に，我々は従事者向けのガイドラインである標準農業技術文書に注目している．この技術文書には，作物栽培の作業手順やスケジュール，使用する機材・資材，栽培によって生じる収益および支出の構造，作物と栽培規模を仮定したときの収支といった経営モデルなど，営農・栽培における知識が包括的に記されており，知識継承や提言の補助としての活用が期待できる．一方で活用に至るには解決すべき課題が残っている．まず標準農業技術文書の作成にあたり規格化されたフォーマットとして FAPS-DB 形式があるが，細かに設定できる反面，営農に関する深い理解が求められる等，普及には一定の障壁があり，都道府県など自治体や農業協同組合がそれぞれ独自に技術文書を作成しているという現状がある [1]．記載する内容も共通する項目は多いものの，各自治体や協同組合によって記載内容にばらつきが生じたり，同じような内容でも表現方法やフォーマットが異なる．そこで，自治体によらない営農における重要項目を抽出し，FAPS-DB 形式より簡素な標準農業技術文書における統一フォーマットを定めた [2]．各自治体の技術基準を画一的に扱うには，標準農業技術文書を読み込んで統一フォーマットに変換する必要があるが，人手で変換作業を行うには膨大なコストがかかることは想像に難くない．機械的に変換できると好ましい一方，以下のような技術的課題が想定される．

- 標準農業技術文書には，文章のほか，ガントチャートやセル結合が多用されている複雑な図表が含まれている．
- 標準農業技術文書に記載される内容・トピック (作物の栽培手順，使用機材・資材，スケジュール，収益・支出構造など) は概ね共通している一方，表現・フォーマットは都道府県ごとに異なる．
- PDF，Excel，紙面をスキャンした画像等，様々なファイル形式で公開されている．また，ファイルやページの構成も作成者によって異なる．都道府県ごとに適切な方法でデータを読み込む必要がある．

長期文脈や画像などマルチモーダルを扱える等の技術進展がある一方で，上記の課題を，どこまでどのように解決できるのか，今後解決する見込みがあ

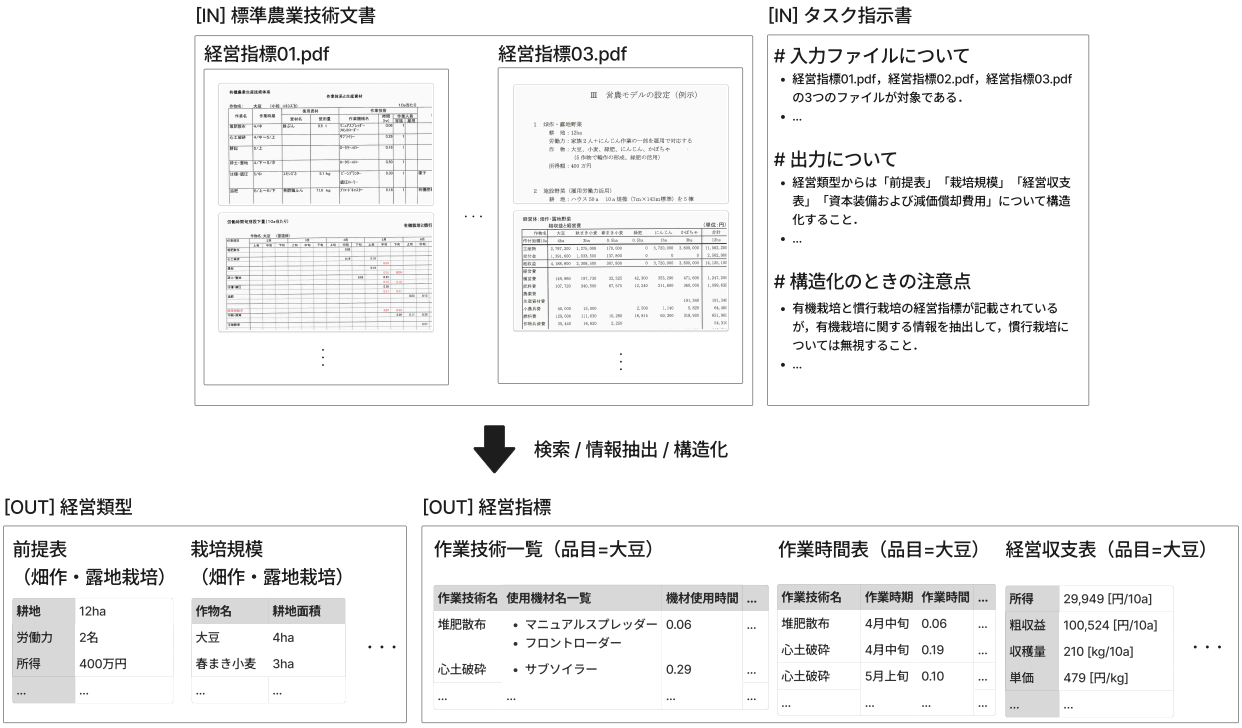

図 1　Subtask 1: Table IE のタスク像．都道府県ごとに形式・フォーマットが異なる標準農業技術文書とタスク指示を入力として，作型ごとの経営類型と作物・品種ごとに定めた経営指標をすべて抽出・構造化する．

るか，等を調査し明らかにするために，自治体ごとにフォーマット・表現形式が異なり，複雑な図表を含む標準農業技術文書から必要な情報を抽出し適切に構造化することを目標にしたタスク: Table Information Extraction (Table IE) を提案する．Table IE は，NTCIR-19 DAGRI (Data Analytics for aGRicultural Information) のサブタスクとしてコンペティション形式で実施される．本稿ではタスクの概要およびデータの入出力の仕様について述べる．

2　関連研究

広い分野で，非構造化データ・半構造化データに対し LLM を適用して情報抽出する，構造化する研究は行われており [3]，農業分野においても標準農業技術文書をはじめその他の非構造データに対する分析・解析に LLM を適用する研究が進められている．[4] では，農研機構が公表しているスマート農業実証プロジェクトの実証成果をまとめたポータルサイトのコンテンツから農業技術・農作業・導入効果を表形式に構造化する検証を実施した．実証課題名，目標，導入技術，達成目標と状況，導入技術の効果などが HTML 形式でまとめられている実証成果文書を LLM で構造化した場合，出力結果の特徴がどのようなものか，出力の精度はどの程度であるか評価し，LLM で構造化するときの効果と性能限界を分析した．検証・分析の対象にした実証成果のコンテンツは一定のフォーマットに従っており，全てのセクションにはテキスト形式で要旨が記載され，表組や画像については対象外にしていた．一方で，標準農業技術文書は自治体や協同組合ごとに独自のフォーマットを定めており表現方法が多様である，テキストや表組，図表が混在している，紙面・印刷物をスキャンして PDF 化したデータである等、統一フォーマットに変換するために適切な前処理が求められる．表組，図表の解析を目指して，[5] では長崎県の技術文書を対象に，文書中の表組の認識及び CSV 等のデータ形式への再構成のためのベンチマークを設計した．また，[6] では，表構造認識の様々な手法の性能をこのベンチマークで評価した．表の構造認識に加え，提案タスクでは文書構造を解析して関係性を解決する，表の意味理解が求められる．

3　タスク: Table IE の仕様

北海道における標準農業技術文書を例に，提案するタスクの概要を図 1 に示す．提案タスクの参加者は，本章で説明するデータを入力とし，[2] で定め

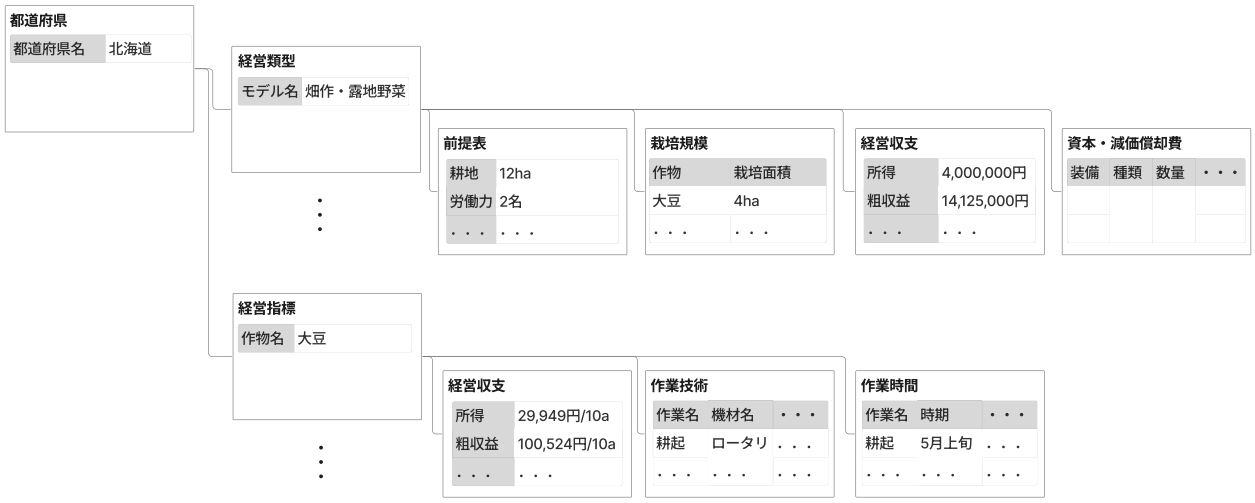

図 2　Subtask 1 における統一フォーマットのデータ概念図．各都道府県と，経営類型の作型および作物・品種の経営指標は一対多の関係にあり，一部項目は記載が無いこともある．

た統一フォーマットに従って構造化した結果を出力するシステム・アルゴリズムを開発する．

3.1　入力データ

提案タスクでは，以下を入力とする．

標準農業技術文書　各自治体が作成・公開している標準農業技術文書．ファイルの形式やファイル数が自治体によって異なるが，このような違いを吸収して処理する枠組があるか，等を検証するため，配布の際にファイルの形式変換や結合・分割といった加工は原則実施しない．

タスク指示書　標準農業技術文書のファイル形式・記載されている内容に関する説明や，出力したい項目，構造化を行うときの注意点をまとめたガイドラインで，都道府県ごとに定める．

3.2　出力データ

標準農業技術文書には，栽培に必要となる労働力や土地，機械設備，資材などの資本，栽培のスケジュールを作物ごとに策定した経営指標と，経営指標に基づき栽培する作物とその規模，栽培にあたり生じる収支目標などを定めた経営類型が記載されている．統一フォーマットにおける経営指標に該当する項目は「経営収支」「作業技術」「作業時間」であり，経営類型に該当する項目は「前提表」「栽培規模」「経営収支」「資本設備と減価償却費」である．経営指標と経営類型，構造化する各項目の関係を表したデータの概念図を図 2 に示す．

3.2.1　経営指標: 経営収支

作物ごとの，適当な耕地面積単位 (ほとんどのケースで 10a を単位としている) での収益・支出を表す．目標として設定した「所得」のもと，収益パートでは「粗収益」「販売収入」，副産物や交付金などの「そのほか収入」とともに，販売における「単価」と「収穫量」を定めている．支出パートでは「経営費計」，内訳として「変動費計」「固定費計」を定め，さらに変動費の内訳として「種苗費」「肥料費」「農薬費」「諸材料費」「動力光熱費」「農具費」「共済費」「荷物運賃手数料」を，固定費の内訳として「農機減価償却費」「機械修理費」「施設減価償却費」「施設修理費」を定めている．また従業員の「雇用費」のほか「土地改良及び水利費」，「生産費」についても支出の項目として定めている．

3.2.2　経営指標: 作業技術

作物・品種ごとの，栽培時に使用する技術，機材，資材および適当な耕地面積あたりでの作業時間や労働力を表す．表 1 に作業技術として構造化する項目一覧を示す．作物・品種ごとに作業は複数生じるものであり，係る作業を全て列挙する．また記載があった場合に限り，使用機材の機材の使用時間，使用資材の数量もあわせて構造化の対象に含める．

3.2.3　経営指標: 作業時間

作物・品種ごとの，栽培の各作業の年間スケジュールおよび，適当な耕地面積あたりで生じる月別の作業時間を表す．表 2 に作業時間として構造化

表 1　作業技術で構造化する項目

作業名	作業技術の名称
作業概要	作業の具体内容
使用機材	作業する上で使用する機械類一覧
使用資材	作業する上で使用する資材一覧
労働力	作業するのに必要な人員
総作業時間	作業するのにかかる時間の総計
費用	作業するのに生じるコスト
備考	作業技術の特記事項

表 2　作業時間で構造化する項目

作業名	作業技術の名称
時期	作業を実施する月 (上中下旬)
作業時間	単位耕地面積で生じる作業時間

する項目の一覧を示す．都道府県によって，月内を上中下旬の 3 区分，あるいは上下旬の 2 区分にわけてスケジュールを策定している．統一フォーマットで構造化するときにスケジュールの区分けが 2 区分・3 区分のいずれであるかの識別を同時に行う．

3.2.4　経営類型: 前提表

栽培・営農する作物一覧 (作型と呼ぶ) と，目標となる「所得額」，一時雇用を含めどのくらいの人数・時間で栽培するかを定めた「労働力」と「労働時間」，栽培する「耕地面積」を表している．記載があれば耕地の内訳（「借入地の耕地面積」と「自作地の耕地面積」）も構造化の対象とする．

3.2.5　経営類型: 栽培規模

定めた作型に対する作物・品種ごとの「栽培面積」を表す．

3.2.6　経営類型: 経営収支

設定した目標「所得額」の算出根拠となる，収益・支出を構造を表す．収益パート，支出パートの各項目は経営指標の経営収支の項目と同様である．定めた作型と対応する経営指標の経営収支の合計に相当する．

3.2.7　経営類型: 資本設備と減価償却費

定めた作型に対する栽培に必要な機材・施設などの資本の「種類」と「数量」，「規格・仕様」，「取得費用」，「耐用年数」，「年間償却額」を表している．

3.3　評価方針・指標

提案タスクは，タスク指示書の内容に基づき，標準農業技術文書の文書構造を正しく理解し，経営指標および経営類型が記載されているページ，章，表を正しく認識できるか，複雑な表形式から値を正しく読み取れるか，等を調査・検証することを目的としている．このため，単位やデータの型の変換は原則行わず，入力データの表記通りに該当項目を抽出できたか，という点にフォーカスする．経営指標・経営類型の各項目のスキーマを定めているため，各項目の値が適切に抽出・構造化できたかを適合率・再現率で評価できる．都道府県 $p \in \mathcal{P}$ の標準農業技術文書から抽出した経営類型・経営指標の項目集合 $\mathbf{Y}_p = \{Y_{p,i} \,|\, i \in \mathcal{I}_p\}$，正解となる経営類型・経営指標の項目集合 $\mathbf{T}_p = \{T_{p,i} \,|\, i \in \mathcal{I}_p\}$ としたとき，適合率，再現率は以下に従って求める

$$\text{Precision} = \frac{\sum_{p\in\mathcal{P}} \sum_{i\in\mathcal{I}_p} |T_{p,i} \cap Y_{p,i}|}{\sum_{p\in\mathcal{P}} \sum_{i\in\mathcal{I}_p} |Y_{p,i}|} \tag{1}$$

$$\text{Recall} = \frac{\sum_{p\in\mathcal{P}} \sum_{i\in\mathcal{I}_p} |T_{p,i} \cap Y_{p,i}|}{\sum_{p\in\mathcal{P}} \sum_{i\in\mathcal{I}_p} |T_{p,i}|}. \tag{2}$$

ただし，経営類型・経営指標の要素 $T_{p,i}, Y_{p,i}$ は文字列あるいは文字列の集合である．$|T_{p,i}|$ は項目があった場合，$T_{p,i}$ が文字列であれば 1，$T_{p,i}$ が文字列の集合であれば集合の大きさを返す．予測した項目の要素 $Y_{p,i}$ に対しても同様である．また $|T_{p,i} \cap Y_{p,i}|$ は正解と予測の類似度を表し，要素が文字列であれば完全一致あるいは文字列類似度を返し，要素が文字列の集合であれば共通集合の大きさを返す．

4　まとめ

複雑な図表を含む農業文書として標準農業技術文書を対象に，標準農業技術文書を読み込み必要な情報を認識・抽出し，[2] の統一フォーマットに変換・構造化するタスク Table IE を提案し，タスクの概要，入出力のデータ仕様について説明した．経営類型，経営指標のスキーマは，北海道と長崎県の標準農業技術文書を確認して設計したが，他の都道府県，協同組合の技術文書についても同様の確認を行い，提案したスキーマの表現力が十分か検証を継続する．またベースライン手法での実験と評価，課題の整理も今後の課題である．

謝辞

本研究は，内閣府研究開発と Society5.0 との橋渡しプログラム（BRIDGE）「AI 農業社会実装プロジェクト」JP23836805 の補助，および，JST，RISTEX, JPMJRS25L2 の支援を受けて行った．

参考文献

[1] 佐藤正衛. 農業技術体系データの作成・利活用の現状と課題. 農作業研究, Vol. 56, No. 3, pp. 197–203, 2021.

[2] 小林暁雄, 大友将宏, 石原潤一, 馬場研太, 桂樹哲雄, 森岡幹, 坂地泰紀, 木村泰知. DAGRI Data format: 農業経営データ解析タスクのための統一フォーマット構築. 言語処理学会 第 32 回年次大会 発表論文集, 2026.

[3] John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. **Nat. Commun.**, Vol. 15, No. 1, p. 1418, February 2024.

[4] 馬場研太, 佐藤正衛. 農業経営指標策定に向けた大規模言語モデル活用の可能性―農業文書の表形式への自動構造化実験―. 農業情報研究, Vol. 34, No. 2, pp. 51–67, 2025.

[5] 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 発表論文集, 2025.

[6] 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. 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. 言語処理学会 第 31 回年次大会 発表論文集, 2025.