The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

農業ナレッジグラフを用いた営農記事からの農作物の関係の発見

A Method for Judging the Semantic Similarity between Crops from Farm Management Articles based on Agricultural Knowledge Graph

朱 成敏 *1 Sungmin JOO　武田 英明 *1*2 Hideaki TAKEDA　竹崎 あかね *3 Akane TAKEZAKI　吉田 智一 *3 Tomokazu YOSHIDA

*1国立情報学研究所 National Institute of Informatics

*2総合研究大学院大学 SOKENDAI University(The Graduate University for Advanced Studies)

*3農業・食品産業技術総合研究機構 National Agriculture and Food Research Organization

Terminology in the agricultural field are difficult to process data automatically since they are often used in various expressions. It is especially difficult for crops whose names vary depending on the agriculture activity, edible parts or cultivation method in which various names are used. This study suggests semantic analysis using agricultural knowledge graph to solve these problems. This study also confirmed that the use of knowledge graph's semantic structure not only enabled the extraction of agriculture activity and the crop name clearly, but also enabled agriculture-specific analysis.

1. はじめに

農業分野の用語には同じ意味にも拘らず，様々な表記が存在している．例えば，農作業の一つである「荒代」は「かじり」や「荒代かき」のように習慣によって異なる名称が使われる場合や農業 ICT システムでは「あらしろ」や「荒しろ」など異なる表記で処理される場合もある．農作物名は地域や利用部位，栽培方法などの基準によって名称が異なる場合や，品種名や食品名が農作物の名称として使われる場合もある．ICT システムのデータや関連文書にはこのような場合，異なる用語として処理される可能性がある．こういった農業分野の語彙が持つ表記の多様性は農業情報の利活用において妨げとなる．

一方，様々な表記が収録されていて，かつその意味関係まで記述されている知識体系があればそれを参照することによって上記の問題は解決できる．例えば,「荒代」の場合は同義語として「荒代かき」と「かじり」を持ち,「あらしろ」や「荒しろ」などで表記される基準があれば，ICT システムはこれらの名称を一括して「荒代」として処理することができる．また，農作物の場合も「レッドクイーン」や「シャインマスカット」のような品種名と一般的な名称である「ブドウ」との関連性が定義された基準があれば，システムは「レッドクイーン」と「シャインマスカット」を「ブドウ」の一種として扱うことができる．このように名称に対する基準があれば農業データやコンテンツに対する正確な情報の抽出や分析など高度な利用が可能となる．

そこで，本研究では農業語彙が持つ表記の多様性に対応し，農業コンテンツの分析に農業ナレッジグラフを用いることを提案する．農業ナレッジグラフは農業分野を対象に構築された知識体系であり，農業における概念と概念間の関係性について定義をしている．それぞれの概念は表記を持っており，表記は見出し語である代表表記以外にも同義語も収録されている．また， 関連情報体系と連携されており，様々な情報を活用することもできる．本研究では農業関連の文書を対象とし，農業ナレッジグラフを用いて農作業と農作物の名称を抽出する．そして，抽出された農作業と農作物の共起を用いて農作業と農作物の関連性を発見する．発見された関連性から農作物同士の関連性を推測する．

連絡先: 朱成敏，国立情報学研究所，〒 101-8430 東京都千代田区一ッ橋 2-1-2, joo@nii.ac.jp

2. 農業ナレッジグラフ

筆者らは農業 ICT システムのデータ連携のために農業分野のナレッジグラフを構築してきた [朱 18]．本章では農作業を体系化した農作業基本オントロジー (AAO，Agriculture Activity Ontology) と農作物の語彙を整理し，Linked Data 化した農作物語彙体系 (CVO，Crop Vocabulary) について概観し，それぞれの意味構造について述べる．

2.1 農作業基本オントロジー

農業 ICT システム間の相互運用性を確保するために農作業の標準語彙として筆者らは農作業基本オントロジーを開発し，推進してきた [朱 16]．農作業基本オントロジーは農作業概念を定義するために目的，行為，対象，副対象，場所，手段，機資材，作物，時期，作業条件の 10 項目を属性を用い，それぞれの属性が持つ値の包含関係から農作業概念を体系化した．また，記述論理による設計を行い，矛盾や重複のない論理性も確保した．最新版である Ver2.01 には 475 語の農作業名称が収録されており，それぞれの農作業名称は固有の名前空間 (URI) を持つ．共通農業基盤 *1 では農作業基本オントロジーを用いた語彙変換 API，用語集などのサービスを提供しており，Turtle/RDF と CSV 形式の関連データも公開している．農作業基本オントロジーでは農作業に対し代表表記，同義語などの別名，英名の３つの表記を与えている．また，これらの表記は共通農業基盤にて提供している語彙変換 API を用いて容易に変換することができる．図 1 は「荒代」の定義と表記を表した例である．

2.2 農作物語彙体系

筆者らは農作物名称の標準語彙として農作物語彙体系を構築した [竹崎 17]．農作物語彙体系は植物学的分類に基づいて様々な農作物名を分類し，それぞれの農作物名は同義語，英名，学名を基本情報として収録している．また，既存語彙である「農薬登録における適用作物名」の作物名,「農産物等の食品分類表」の食品名,「日本食品標準成分表」の食品名，情報

*1 CAVOC, http://cavoc.org

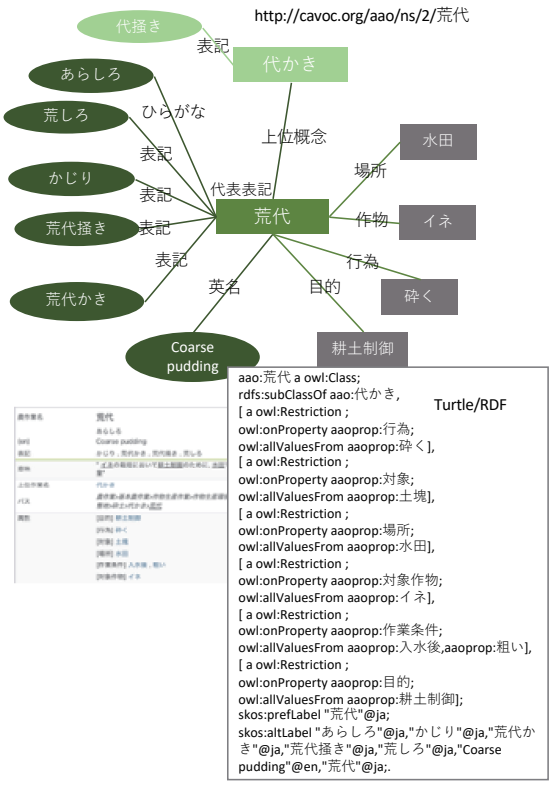

図 1: 「荒代」における情報の意味関係.

体系である NCBI の Taxonomy ID，Wikipedia の項目名との対応関係を調査し，関連情報として収録した．最新版である Ver1.52 では 1,249 語が収録されており，共通農業基盤にて公開している．それぞれの農作物名は固有の名前空間を持ち，収録情報を閲覧することが可能である．これらの情報は CSV と Turtle/RDF 形式でも公開されており，機械可読性も確保している．農作物語彙体系では農作物名に対し，英名や科名，学名のような植物学的情報，別名，一般的な名称にあたる総称を上位概念として持つ．また，既存の語彙体系に収録されている名称も収録されており，様々な語彙の意味関係を把握することができる．図 2 は「シャインマスカット」に関連する意味関係を表す例である．

3. 農作物間の関係性発見

本章では実際の新聞記事のデータから発見された農作業名称と農作物名称の関係性について考察し，作物の生産過程の類似性から農作物同士の関係性を判定する手法について述べる．

3.1 営農記事の分析

本研究では農業コンテンツとして営農に関連する記事を用いる．営農関連記事は研究目的での利用許可を得た 2014 年 4 月から 2017 年 3 月までの日本農業新聞 *2 の営農面の記事 3,479 件である．オープンソース形態素解析エンジンである Mecab[MeCab 13] を用いて本文の形態素分析を行い，品詞が名詞の場合に農作業基本オントロジーと農作物語彙体系の名称を抽出した．抽出対象となる名称は代表表記と同義語を含む全ての関連語彙である．実行結果，99 件の農作業名，292 件の農作物名がそれぞれ 1,371 件，1,265 件の記事から発見された．

*2 日本農業新聞，https://www.agrinews.co.jp/

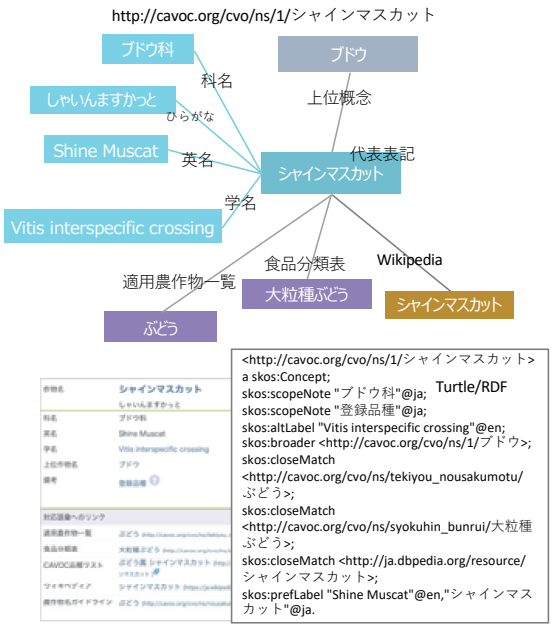

図 2: 「シャインマスカット」における情報の意味関係.

3.2 表記の調整

そして，出現頻度と共起頻度を確認するために発見された農作業と農作物の名称に対して表記の整理を行った．同じ作業にも関わらず異なる表記で書かれており，別の作業として処理される場合に備えて，代表表記に変換する処理を農作業基本オントロジーの語彙変換 API を用いて行った．例えば，同じ農作業である「播種」と「種まき」がそれぞれ異なる名称として含まれていた記事が一部あり，農作業基本オントロジーの代表表記である「は種」にまとめた．農作物の場合は品種や用途，利用部位によって異なる名称になる場合が多く，代表表記に統一するために総称にあたる上位概念の名称を用いてまとめた．この処理によって「シャインマスカット」や「レッドクイーン」など品種名は総称である「ブドウ」として処理された．

3.3 関連農作業の抽出

次は農作物名と農作業名の共起を確認した．農作物単位で共起した農作業名を抽出し，関連農作業のリストを作成した．最も多くの農作業と共起した農作物は「イネ」であり，52 件であった．一方，共起した農作業が 5 件以下の農作物も 79 件があり，農作物 1 件にあたり農作業の平均共起数は 12.49 件であった．

3.4 農作物間の類似性判別

本研究では関連農作業名のリストを用いて農作物同士の類似性を判別するために Jaccard 係数を用いる．共起した農作業の集合 $Activities_a$ を持つ農作物 $crop_a$ と集合 $Activities_b$ と共起した農作物 $crop_b$ の Jaccard 係数は次のように計算できる．

$$Jaccard(crop_a, crop_b) = \frac{|Activities_a \cap Activities_b|}{|Activities_a \cup Activities_b|} \quad (1)$$

本実験では共起した農作業名称が 5 件以上の農作物を対象にし，Jaccard 係数を求めた．表 1 はコムギ，タマネギ，トマトと関連度がある上位 5 件の農作物の順位である．比較のた

め農作物間の共起数による順位も表示した．実験の結果，コムギと最も類似性を持つ農作物はオオムギであり，タマネギはネギ，トマトの場合はナスに判明された．表 2 はコムギ，タマネギ，トマトと最も関連があると判明した農作物の関連農作業のリストである．

表 1: 農作物間の類似判別.

(a) コムギ

#	Jaccard 係数		共起数	
1	オオムギ	0.6563	イネ	70 回
2	ダイズ	0.6364	ダイズ	47 回
3	トウモロコシ	0.5938	ジャガイモ	17 回
4	イネ	0.5789	テンサイ	15 回
5	キャベツ	0.5428	ソバ	14 回

(b) タマネギ

#	Jaccard 係数		共起数	
1	ネギ	0.7274	イネ	11 回
2	ホウレンソウ	0.6957	トマト	11 回
3	ジャガイモ	0.6400	ネギ	9 回
4	レタス	0.6087	ニンジン	9 回
5	キュウリ	0.5770	ダイズ	6 回

(c) トマト

#	Jaccard 係数		共起数	
1	ナス	0.7083	イチゴ	41 回
2	ピーマン	0.6207	キュウリ	40 回
3	イチゴ	0.6061	イネ	32 回
4	ウンシュウミカン	0.5833	ナス	29 回
5	CO	0.5517	チャ	19 回

コムギは 31 件の農作業と，オオムギの 22 件の農作業と共起しており，22 件全部コムギの農作業と一致した．一方，イネと共起した 52 件の農作業の中でコムギと一致する農作業は 25 件だった．一致した農作業名はオオムギより 2 件多かったが，27 件の農作業が一致していないことが判明したので，農作業の類似性はオオムギとコムギの方が強い類似性を持つと考えられる．採種，追肥，融雪の 3 つの農作業はコムギのみ共起していたことがわかった．

タマネギは 21 件の農作業名との共起が確認された．ネギの 17 件の共起農作業の中で排水作業を除く 16 件が一致した．一方，イネは 20 件の共起農作業の一致が確認されたが，関連性がない農作業が 32 件発見された．今回の実験ではコムギとタマネギの場合，最も共起した農作物はイネであることがわかった．イネは農作業名称が発見された 2,505 件の記事の中で 674 件の記事に出現しており，共起頻度が農作物の中で最も多かったと考えられる．

トマトは 28 件の農作業と共起しており，選別作業，意見交換，土寄せの 4 件以外はナスとイチゴの農作業と一致した．ナスは 22 件農作業の中で 20 件が，イチゴは 23 件の中で 19 件がトマトの農作業と一致している．実際，イチゴは Jaccard 係数による順位でも 3 番目である．

3.5 農作物間の類似性による関連記事の提案

前節で求めた農作物間の類似性を用い，類似性の高い農作物に関する記事を関連記事として提案した．Jaccard 係数の上位 3 件までの農作物を選択し，該当する農作物の出現頻度が高い記事を掲載日が近い順で提示した．図 3 は「タマネギ」に関する記事の例である．1 の (b) の結果より「ネギ」と「ホウレンソウ」,「ジャガイモ」に関する記事が「タマネギ」に関する

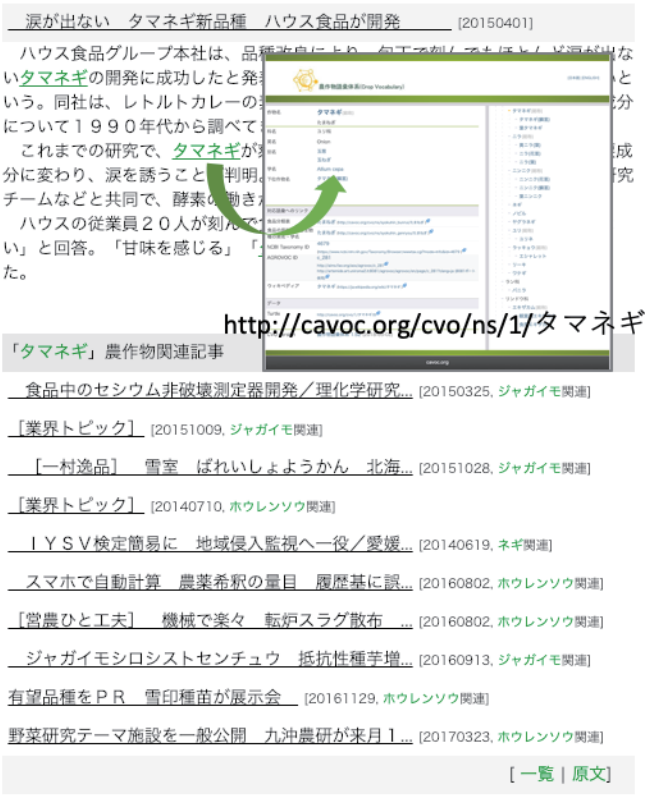

図 3: 農作物間の類似性による関連記事リスト.

記事と共に関連記事として表示されている．

4. 考察と今後の課題

今回，農作物名と共起した農作業の集合を Jaccard 係数を用いて類似性を確認した．この場合，どちらかの農作業の数が少なく，さらに出荷作業や収穫作業のような栽培において基本となる農作業名の場合は強い類似性を見せた．そのため比較的に関連農作業の数が少ない農作物が上位になることが多かった．一方，麦踏みのように特定農作物のための農作業やハウス栽培の農作業など農作物の栽培特徴が反映されたことも確認できた．

営農記事は栽培や経営など幅広い情報が書かれているため農作業と農作物の関連性が一般的な記事より多く含まれていると考えられる．この場合は共起頻度も関連性を判別する重要な因子となる．今回は Jaccard 係数を用いて類似性を判別したが，共起頻度を重み付けとして取り込む場合，改善の可能性がある．今後の課題として取り組む予定である．

今回は農作業基本オントロジーの語彙変換を用いて同義語を代表表記に整理し，また，農作物語彙体系の階層構造を用いて利用部位や栽培方法などによる別名と品種名を総称に変換して共起を確認した．この調整作業により抽出する単語の数や集計による手間が軽減されたと考えられる．

農作業名称の中では収穫や運搬のように農業以外にも使われる単語が多く含まれている．こういう単語の共起は本研究の提案手法の性能に影響を与える可能性が高い．形態素解析方法の改善や文脈情報を用いて作業名を特定する必要がある．今回は農作業に注目して農作物の類似性判別したが，今後は農機具や資材など様々な農業関連語彙を用いて類似性の判別を行いたい．また，そのために農業ナレッジグラフの拡張する予定して

表 2: 農作物名と共起した農作業名のリスト.

農作物名	共起した農作業名
コムギ	収穫作業, 評価作業, 受粉, 施肥, 雑草抑制作業, 出荷作業, は種, 移植作業, 定植, 育苗, 出荷調製作業, 排水作業, 耕耘, 選別作業, 稲刈り, 採種, 生物制御作業, 基肥, せん定, 追肥, 出芽, 中耕, 運搬作業, 土寄せ, 刈取り, 田植え, 麦踏み, 融雪, 誘引, 意見交換, 砕土
オオムギ	移植作業, 育苗, 出荷作業, 収穫作業, 出荷調製作業, 耕耘, 施肥, 排水作業, 雑草抑制作業, 選別作業, 評価作業, は種, 出芽, 生物制御作業, 基肥, 麦踏み, 中耕, 運搬作業, 定植, 刈取り, 稲刈り, 田植え
イネ	追肥, 生物制御作業, 収穫作業, 意見交換, 評価作業, せん定, 受粉, 摘果, 施肥, 雑草抑制作業, 出荷作業, は種, 出芽, 中耕, 土寄せ, 定植, 育苗, 耕耘, 鎮圧, 排水作業, 給水作業, 出荷調製作業, 移植作業, プラスチックマルチング, 基肥, かん水, 物理的雑草抑制作業, 砕土, 運搬作業, 選別作業, モニタリング作業, 稲刈り, 間伐, 誘引, 包装, 代かき, 摘粒, 清掃作業, 草姿調整作業, 田植え, 催芽, 保水作業, 刈取り, 暖房, ハロー作業, 風制御作業, 遮光, 脱穀, 精米, 中干し, 冷房, 手取除草
タマネギ	収穫作業, は種, 中耕, 土寄せ, 施肥, 基肥, 追肥, 生物制御作業, 移植作業, 定植, 育苗, 出荷作業, 選別作業, プラスチックマルチング, 運搬作業, 評価作業, 給水作業, かん水, 出荷調製作業, 遮光
ネギ	は種, 定植, 排水作業, 出荷調製作業, 移植作業, 育苗, 出荷作業, 遮光, 収穫作業, 生物制御作業, 選別作業, 土寄せ, 給水作業, かん水, 評価作業, 施肥
イネ	追肥, 生物制御作業, 収穫作業, 意見交換, 評価作業, せん定, 受粉, 摘果, 施肥, 雑草抑制作業, 出荷作業, は種, 出芽, 中耕, 土寄せ, 定植, 育苗, 耕耘, 鎮圧, 排水作業, 給水作業, 出荷調製作業, 移植作業, プラスチックマルチング, 基肥, かん水, 物理的雑草抑制作業, 砕土, 運搬作業, 選別作業, モニタリング作業, 稲刈り, 間伐, 誘引, 包装, 代かき, 摘粒, 清掃作業, 草姿調整作業, 田植え, 催芽, 保水作業, 刈取り, 暖房, ハロー作業, 風制御作業, 遮光, 脱穀, 精米, 中干し, 冷房, 手取除草
トマト	接ぎ木, 移植作業, 定植, 生物制御作業, 収穫作業, 施肥, かん水, 出荷作業, は種, 排水作業, 育苗, 評価作業, 暖房, 出荷調製作業, 換気, 選別作業, 遮光, 包装, 意見交換, モニタリング作業, プラスチックマルチング, 送風, 誘引, 受粉, 保温, 土寄せ, 草姿調整作業, 冷房
ナス	接ぎ木, 生物制御作業, 収穫作業, 出荷作業, 定植, せん定, 施肥, 保温, プラスチックマルチング, 育苗, 誘引, 受粉, 出荷調製作業, 暖房, 換気, かん水, 雑草抑制作業, 評価作業, 移植作業, 意見交換, 草姿調整作業, モニタリング作業
イチゴ	収穫作業, 定植, 生物制御作業, 育苗, 出荷作業, 出荷調製作業, プラスチックマルチング, 暖房, 評価作業, 換気, 予冷, かん水, 草姿調整作業, せん定, 施肥, 送風, 受粉, 冷房, 包装, は種, 遮光, 追肥, 摘果

いる.

5. おわりに

農作物の栽培には特定の農作業，または複数農作業の組み合わせが用いられており，農作物の特徴として用いることができる．そこで，本研究では農作業に注目して農作物間の類似性の判別を提案した．営農記事から農作業名称と農作物名称を抽出し，共起を基準にその関連性を定義した．農作業と農作物の抽出は農業ナレッジグラフを用いて行い，同義語や意味関係を考慮した名称の抽出ができた．今後は共起頻度による重み付けや農作業と農作物以外の農業関連語彙を用い，本研究の提案手法を改善していきたい.

謝辞

本研究で利用した記事データは株式会社日本農業新聞により提供されたものである.

本研究（の一部）は，総合科学技術・イノベーション会議の SIP（戦 略的イノベーション創造プログラム）「次世代農林水産業創造技術」（管理法人：農研機構生物系特定産業技術研究支援センター）の支援を受けて行った.

参考文献

[朱 18] 朱成敏, 武田英明, 竹崎あかね, 吉田智一： 農業データ連携のためのナレッジグラフに基づく標準語彙の運用, 2018 年度人工知能学会全国大会 (第 32 回), No. 2G2-OS-10b-01, 2018.

[朱 16] 朱 成敏, 小出 誠二, 武田 英明, 法隆 大輔, 竹崎 あかね, 吉田 智一：記述論理に基づく農作業オントロジーの設計と応用, 第 38 回人工知能学会セマンティックウェブとオントロジー研究会（SIGSWO）, 06, 2016.

[竹崎 17] 竹崎あかね, 朱成敏, 武田英明, 吉田智一： 農業 IT システム間のデータ連携を促進する農作物語彙体系の構築, 電子情報通信学会技術研究報告，知的環境とセンサネットワーク研究会, vol. 117, no. 310, ASN2017-J27, pp. 98-99, (2017).

[MeCab 13] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, <http://taku910.github.io/mecab/> 2019 年 2 月 1 日参照.