AgriNER: An NER Dataset of Agricultural Entities for the Semantic Web

Sayan De[1], Debarshi Kumar Sanyal[2], and Imon Mukherjee[1]

[1] Indian Institute of Information Technology Kalyani, 741235, India
[2] Indian Association for the Cultivation of Science, Jadavpur 700032, India
sayan_jrf22@iiitkalyani.ac.in, debarshi.sanyal@iacs.res.in, imon@iiitkalyani.ac.in

Abstract. An immense amount of data relevant to agriculture is generated from the vast scholarly literature. To get as much relevant information as possible from the data, we need to extract the context and meaning from them. Semantic web technology can provide context and meaning to the data. Named entity recognition (NER) systems can help to extract the named entities and the relations between the entities. In addition to that, these entities and relations can be used to build a knowledge graph (KG) which can be stored using a resource description framework (RDF) and queried with SPARQL. In this paper, we propose an NER dataset that contains a total of thirty-six types of entities and nine types of relations, which can be used to build a KG.

Keywords: Agricultural Dataset · Named Entity Recognition · Relation Extraction · Knowledge Graph

1 Introduction

Research papers on agriculture contain information about the latest advances in the field, yet it is not always easily accessible to practitioners including scientists and farmers, for a variety of reasons including that the number of papers is huge and that the information is not available in a structured format. Agricultural industries, researchers, food processing companies, and many organizations need to extract entities such as crop names, pesticides, factors that affect plant growth, etc., and their relationships to make useful and strategic decisions. A named entity recognition (NER) system helps to extract knowledge entities from unstructured sources [2]. There are a few works on NER in agriculture, some of which like [1], [4] apply deep learning. A few datasets are available to train NER systems. Malarkodi et al. [7] have proposed nineteen entity types in the agriculture domain, but it does not cover many important aspects of agriculture. In addition to that, their corpus is not publicly available. Lun et al. [6] focus on four entity types, namely, Crop, Disease, Pest, and Drug, however, limiting to only Chinese agricultural websites. Gangadharan et al. [3] have worked with only three types of entities, namely, Disease, Soil, and Fertiliser, using only Indian agricultural websites. Liu et al. [5] have worked with six types of entities, namely Organism, Trait,

Method/Equipment, Chemical, Gene, Environment, and Miscellaneous using article abstracts of ten typical horticultural journals. In contrast to the above works, our corpus is an annotated collection of abstracts from agriculture research papers, and our set of entity types and relations is significantly larger. In this paper, we propose thirty six entity types and nine relations between the entities. Our contributions to this paper are as follows:

1. We introduce a fine-grained tag set comprising 36 useful entities in the agricultural domain.
2. We introduce 9 relations between the entities, including symmetric and asymmetric relations.
3. We introduce a publicly available fully annotated corpus with the above tags.

The corpus is publicly available on GitHub[3]. The rest of our paper is organized as follows. In Section 2 we propose a taxonomy for the entities and relations. We provide dataset statistics in Section 3. In Section 4, we apply a machine learning model for NER on this dataset. We conclude in Section 5.

2 Proposed Taxonomy

Our dataset is built from abstracts of research papers in agriculture. After analyzing the abstracts, we have developed a list of entity types and relations to cover most of the important knowledge aspects of the papers. The proposed tag set contains thirty-six named entities that we believe can help in research in the agriculture domain. The named entity types are Agri_Pollution, Agri_Process, Agri_Waste, Agri_Method, Chemical, Citation, Crop, Date_and_Time, Disease, Duration, Event, Field_Area, Food_Item, Fruit, Humidity, Location, ML_Model, Money, Natural_Disaster, Natural_Resource, Nutrient, Organism, Organization, Other, Other_Quantity, Person, Policy, Quantity, Rainfall, Season, Soil, Technology, Temp, Treatment, Vegetable, Weather. The terms are self-explanatory.

We have extracted nine relations to form meaningful connections between the entities. We define three symmetric relation types Coreference_Of, Conjunction, Synonym_Of, and six asymmetric relation types Caused_By, Helps_In, Includes, Originated_From, Used_For, Seasonal. A detailed description of the entity types and relations is available in our GitHub repository.

3 Dataset Statistics

The quality of the dataset influences the knowledge graph constructed and the machine learning models trained on it. We have hand-picked the abstracts of 180 papers from several reputed agricultural journals, such as Asian Journal of Agricultural and Food Sciences (AJAFS)[4], The Indian Journal of Agricultural

[3] `https://github.com/Tec4Tric/AgriNER`
[4] `https://www.ajouronline.com/index.php/AJAFS/index`

Sciences[5], and a few journals from IEEE and Springer Nature. We have analyzed the abstracts of these papers and recent trends in agriculture like [8], [9], and then we have decided on thirty six entities and nine types of relationships among the entities. Table 1 displays a summary of the number of occurrences of each annotated entity in the proposed dataset in percentage.

Table 1. Entities with their occurrences in AgriNER dataset.

Entity	Frequency	Entity	Frequency	Entity	Frequency
Agri_Method	4%	Other_Quantity	4%	Fruit	3%
Agri_Process	6%	Policy	<1%	Location	15%
Chemical	4%	Rainfall	1%	Money	<1%
Crop	13%	Soil	1%	Natural_Resource	2%
Disease	1%	Temp	1%	Organism	4%
Event	<1%	Vegetable	<1%	Other	<1%
Food_Item	2%	Agri_Ploution	1%	Person	3%
Humidity	<1%	Agri_Waste	2%	Quantity	2%
ML_Model	4	Citation	1%	Season	3%
Natural_Disaster	5	Date_and_Time	2%	Technology	2%
Nutrient	5%	Duration	2%	Treatment	1%
Organization	3%	Field_Area	1%	Weather	<1%

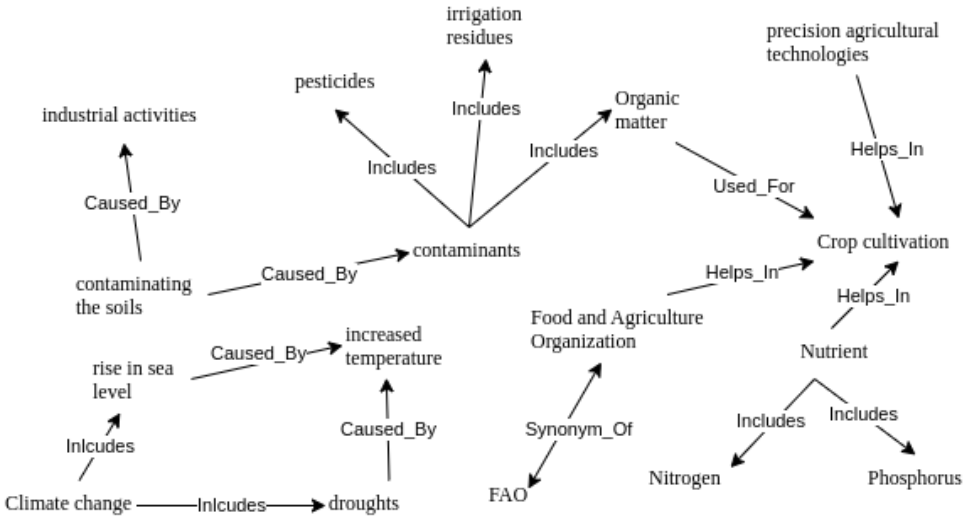

Fig. 1. Some parts of the knowledge graph using the dataset.

We have used the freely available `brat` tool[6] for annotation. One of the challenges was the entity class imbalance. To solve this problem, we have first

[5] `https://epubs.icar.org.in/index.php/IJAgS`
[6] `https://brat.nlplab.org`

counted the occurrences of the mentions of each entity type. Then, we added more data to the corpus to increase the count of the least frequent entity type. In total, we have 14,307 word-tokens and 1348 entity mentions. We have partitioned the dataset in a 70:30 ratio, with 70% data for training, and 30% data for testing.

4 Machine Learning-based Extraction of Named Entities

To provide a baseline for an automatic NER system for the dataset, we have trained `spaCy`[7] with the entities we have labeled. `spaCy` is a free open-source library for natural language processing in Python. `spaCy v3.0` provides a transformer-based pipeline, where we can train the model with our custom data. We first initialize the `spaCy` pipeline with `tok2vec` and `ner` models and then trained the model for several epochs with our custom entities. This model can recognize entities in unstructured data from the agricultural domain.

Table 2. A sample of the classification report.

	Precision	Recall	F1-Score	Support
Agri_Method	1.00	0.80	0.88	5
Chemical	0.98	0.87	0.91	10
Crop	0.93	0.77	0.84	18
Duration	0.83	0.75	0.80	5
Location	0.93	0.95	0.98	19
ML_Model	0.99	0.85	0.92	7
Organism	0.90	0.97	0.94	9
Other_Quantity	0.85	0.97	0.92	6
Season	0.98	0.97	0.98	6

Table 2 displays the classification metrics and the results. For simplicity, we have restricted to two digits after the decimal point. We have excluded the results for some of the entity types due to their very low occurrence in the test data. In Table 2, the support for some of the entities is low because of their low occurrence in the test dataset. Due to the size of the entity class, it is conceivable that not all of the entities were observed while predicting on the test dataset. Figure 1 displays some parts of the knowledge graph built using the proposed dataset.

5 Conclusion

In this paper, we have introduced a total of thirty six entities, and three symmetric and six asymmetric relations extracted from several agricultural research papers. The NER dataset is organized into a knowledge graph. In the future, we intend to use semantic web technologies to make the graph semantically richer by linking it to other relevant knowledge graphs. We hope better ML models will be built to improve the classification performance, and that our dataset will inform and motivate further research on the construction and application of agricultural knowledge graphs.

[7] `https://spacy.io`

6 Acknowledgement

This work is implemented as part of the "*Extraction, Organization and Query of Scholarly Information*", sponsored by the Science & Engineering Research Board, Govt. of India.

References

1. Devi, M., Dua, M.: Adans: An agriculture domain question answering system using ontologies. In: Proceedings of the 2017 International Conference on Computing, Communication and Automation (ICCCA). pp. 122–127. IEEE (2017)
2. Drury, B., Fernandes, R., Moura, M.F., de Andrade Lopes, A.: A survey of semantic web technology for agriculture. Information Processing in Agriculture **6**(4), 487–501 (2019)
3. Gangadharan, V., Gupta, D.: Recognizing named entities in agriculture documents using LDA based topic modelling techniques. Procedia Computer Science **171**, 1337–1345 (2020)
4. Li, W., Chen, P., Wang, B., Xie, C.: Automatic localization and count of agricultural crop pests based on an improved deep learning pipeline. Scientific Reports **9**(1), 7024 (2019)
5. Liu, Z., Luo, M., Yang, H., Liu, X.: Named entity recognition for the horticultural domain. In: Journal of Physics: Conference Series. vol. 1631, p. 012016. IOP Publishing (2020)
6. Lun, Z., Hui, Z., et al.: Research on agricultural named entity recognition based on pre train BERT. Academic Journal of Engineering and Technology Science **5**(4), 34–42 (2022)
7. Malarkodi, C., Lex, E., Devi, S.L.: Named entity recognition for the agricultural domain. Research in Computing Science **117**(1), 121–132 (2016)
8. Sinha, B.B., Dhanalakshmi, R.: Recent advancements and challenges of internet of things in smart agriculture: A survey. Future Generation Computer Systems **126**, 169–184 (2022)
9. Verma, K.K., Song, X.P., Joshi, A., Tian, D.D., Rajput, V.D., Singh, M., Arora, J., Minkina, T., Li, Y.R.: Recent trends in nano-fertilizers for sustainable agriculture under climate change for global food security. Nanomaterials **12**(1), 173 (2022)