Expert Systems With Applications 277 (2025) 127126

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Fine-tuned encoder models with data augmentation beat ChatGPT in agricultural named entity recognition and relation extraction

Sayan De [a],*, Debarshi Kumar Sanyal [b], Imon Mukherjee [a]

[a] Indian Institute of Information Technology Kalyani, 741235, India
[b] Indian Association for the Cultivation of Science, Jadavpur, 700032, India

ARTICLE INFO

Dataset link: https://github.com/Tec4Tric/ESWA-NER-RE

Keywords:
Named entity recognition
Relation extraction
Knowledge graph
Text data augmentation

ABSTRACT

Agricultural research produces vast amounts of unstructured textual data, which remains largely underutilized due to the lack of robust tools for automated processing. If effectively processed, this underutilized data can provide critical insights to advance agricultural practices, decision-making, and sustainability. This work focuses on applying Named Entity Recognition (NER) and Relation Extraction (RE) to convert unstructured data into structured formats, addressing the challenges of domain-specific terminology and limited annotated datasets. This scarcity is primarily due to the domain-specific terminology, contextual complexity, and lack of annotated data in the agricultural domain. This study addresses these challenges by proposing sophisticated data augmentation techniques, validated using large language models and human reviewers, to enhance training data. We introduce AgNER-BERTa and AgRE-BERTa, two encoder-based models tailored for agricultural NER and RE tasks, and compare them with state-of-the-art (SOTA) baselines, including SciBERT, SpanBERT, and generative decoder models like ChatGPT. Our experiments demonstrate superior performance, achieving 98% accuracy for NER and 97% for RE outperforming SOTA models. The extracted entities and relations are used to construct the Agricultural Knowledge Graph (AgKG), providing structured, queryable insights to support precision agriculture, policy-making, and sustainable farming practices.

1. Introduction

In recent years, the thrust of research in agriculture has generated a large volume of information but in most cases, this information is unstructured (Fountas et al., 2024). This makes it extremely difficult and inefficient to extract and use the key information. It is practically impossible to analyze such volumes of data manually, which is very time-consuming and prone to human errors. To address this challenge, Natural Language Processing (NLP) techniques, particularly NER and RE, are increasingly being used to automate the identification of important entities and their relationships (Qiao et al., 2022). This process allows for the conversion of unstructured text into structured, machine-readable formats, facilitating more effective data utilization. However, a major complication in implementing deep learning models for NER and RE in the agricultural domain is the lack of large annotated datasets that cover a large variety of entities. Textual data augmentation can address this challenge (Shorten, Khoshgoftaar, & Furht, 2021), but its implementation is challenging due to the inherent complexity of natural language and the presence of domain-specific terminology.

1.1. Motivation

With the increasing availability of agricultural research data which is mostly unstructured, there is a growing need for automated methods to extract meaningful insights efficiently. However, extracting meaningful information from unstructured text is a complex task due to domain-specific terminology, contextual ambiguity, and the scarcity of annotated datasets. The **AgriNER** dataset, earlier proposed by us De, Sanyal, and Mukherjee (2023), contains thirty-six types of agricultural named entities and nine types of relations including symmetric and asymmetric relations, and sentences, extracted from agriculture research papers. The labeled dataset, though rich with diverse entity and relation types, is quite limited in size. In this paper, we aim to enlarge it automatically using Large Language Models (LLMs). Deep learning models, which are increasingly being adopted with great success for various NLP tasks, require large amounts of training data. The scarcity of annotated datasets in agriculture often poses a significant challenge to applying deep neural models like LLMs in this domain. We believe that our augmented dataset will help mitigate this issue to a great extent.

* Corresponding author.
E-mail addresses: sayan_jrf22@iiitkalyani.ac.in (S. De), debarshi.sanyal@iacs.res.in (D.K. Sanyal), imon@iiitkalyani.ac.in (I. Mukherjee).

https://doi.org/10.1016/j.eswa.2025.127126
Received 30 November 2024; Received in revised form 6 February 2025; Accepted 1 March 2025
Available online 10 March 2025