Adapting BERT and AgriBERT for Agroecology: A Small-Corpus Pretraining Approach

Oussama Mechhour[1,2*], Sandrine Auzoux[1], Clément Jonquet[3], Mathieu Roche[2]

[1*]AIDA, Univ. of Montpellier, CIRAD, Reunion Island, France.
[2]TETIS, Univ. of Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.
[3]MISTEA, Univ. of Montpellier, INRAE, Institut Agro, Montpellier, France.

*Corresponding author(s). E-mail(s): oussama.mechhour@cirad.fr;
Contributing authors: sandrine.auzoux@cirad.fr;
clement.jonquet@inrae.fr; mathieu.roche@cirad.fr;

Abstract

Source variables, or observable properties, used to describe agroecological experiments are often heterogeneous, non-standardized, and multilingual, making them challenging to understand, explain, and utilize in cropping system modeling and multicriteria evaluations of agroecological system performance. A potential solution is data annotation via a controlled vocabulary, known as candidate variables, from the Agroecological Global Information System (AEGIS). However, matching source and candidate variables via their textual descriptions remains a challenging task in agroecology. Domain-general language models, such as BERT, often struggle with domain-specific tasks due to their general-purpose training data. In the literature, these models are adapted to specialized domains through further pretraining, pretraining from scratch, and/or fine-tuning on downstream tasks. However, pretraining a domain-general model on a domain-specific corpus is resource-intensive, requiring substantial time, energy, and computational resources. To the best of our knowledge, no study has further pretrained a domain-general model on a small corpus (less than 100 MB) to adapt it to a domain-specific task and evaluated it on downstream tasks without fine-tuning. To address these shortcomings, this paper proposes further pretraining BERT and AgriBERT on a small agroecology-related corpus. This approach is designed to be both time- and resource-efficient while enhancing domain adaptation. We evaluate the pretrained models on the task of matching source and candidate

variable descriptions without fine-tuning. Our results show that our further pretrained AgriBERT (+ Experts + Core) model outperforms all others by more than 8% from P@1 to P@10. These findings showed that small-scale pretraining can significantly improve performance on domain-specific tasks without requiring fine-tuning.

Keywords: Semantic matching, Word embeddings, Masked Language Modeling (MLM), Pretrained Language Models (PLMs), Observable properties

1 Introduction

Source variables or observable properties[1] used to describe agroecological experiments are often described using homonyms, synonyms, acronyms, multiple languages, and nonstandard terms, making them challenging to interpret and explain. Each researcher may describe these source variables differently, which complicates their use in cropping system modeling and multi-criteria evaluations of agroecological system performance.

To address this issue, the French Agricultural Research Centre for International Development (CIRAD) developed the Agroecological Global Information System (AEGIS) [1]. AEGIS integrates a harmonized data acquisition and processing chain that utilizes a set of observable properties referred to in this paper as candidate variables. These candidate variables combine semantic terms from reference ontologies, such as the Plant Ontology [2], Crop Ontology [3], Environment Ontology [4], and Agronomy Ontology [5], alongside expert agroecological knowledge. Table 1 presents examples of correct matching between source and candidate variable descriptions.

Table 1 Samples of source and candidate variable descriptions. In total, we have 84 source descriptions and 170 candidate descriptions.

Source descriptions	Candidate descriptions
Cane yield (in fresh machinable stem)	measurement of fresh stem biomass at plot level
Cover plant 2 growth in height	Height of the apex of the sugar stem sample
Plant height up to the apex	Height of the apex of the sugar stem sample
...	...

For instance, source variable descriptions in lines 2–3 differ lexically but refer to the same candidate variable description because they share the same meaning. The goal of this work is to match these two types of descriptions using Natural Language Processing (NLP).

To achieve this, it is essential to understand the semantic meaning of descriptions in order to match them correctly. This challenge can be addressed using semantic matching approaches from the literature, such as word embeddings. Semantic matching involves determining whether two words or phrases have the same meaning, even if they do not share similar lexical structures.

[1] An observable property is the description of something observed or derived.

One family of semantic matching techniques is word embeddings, which represent each word as a vector capturing its features and meaning. Word embeddings can be divided into two categories: (i) context-independent and (ii) context-dependent. Context-independent methods, such as Word2Vec [6], FastText (Fast Text Representation Learning) [7], and GloVe (Global Vectors for Word Representation) [8], represent each word as a unique vector regardless of its context. For example, two synonyms will be represented differently by these models. To overcome this limitation, context-dependent embedding techniques emerged, such as ELMo (Embeddings from Language Models) [9] and BERT (Bidirectional Encoder Representations from Transformers) [10].

ELMo represents each word using both its left and right context through a bidirectional LSTM (Long Short-Term Memory) [11]. However, the left and right contexts are processed sequentially, not simultaneously. Transformers-based models, such as BERT, address this issue. BERT has proven efficient across multiple tasks, such as question answering [12], text classification [13], machine translation [14], and more. BERT represents each token in its context, ensuring that the same word has different representations based on its context. BERT is bidirectional, meaning it considers both the left and right context simultaneously, enabling richer contextual representations.

BERT was pretrained on a large corpus from BookCorpus and English Wikipedia using self-supervised learning, which eliminates the need for annotated data. It was trained on two tasks: (i) Masked Language Modeling (MLM), where 15% of input tokens are randomly masked, and the model predicts them, and (ii) Next Sentence Prediction (NSP), where the model determines whether one sentence follows another. While BERT is pretrained on general data to learn the English language, it struggles with domain-specific data, such as agroecology.

Another context-dependent word representation model, AgriBERT (Agricultural Bidirectional Encoder Representations from Transformers) [15], was pretrained on an agricultural corpus using the same architecture as BERT. While AgriBERT is more relevant to agroecology than BERT, it is still too general for agroecology-specific tasks. To adapt these general models to the agroecology domain, pretraining and fine-tuning on domain-specific data are required [16, 17]. Pretraining such models, however, is resource-intensive, requiring large datasets, extensive computational resources, and significant time.

To address these constraints, we propose to answer the following research question (RQ): *Can further pretraining a general model (i.e., BERT and AgriBERT) on a small agroecology-related corpus outperform these general models on agroecology tasks without fine-tuning?*

The contributions of this work are as follows:

- We collected a small corpus (less than 100 MB) to further pretrain BERT and AgriBERT.

- Our pretrained models on AgriBERT outperformed BERT, AgriBERT, and all our pretrained models on BERT in the task of matching source and candidate variable

descriptions.

- The pretrained models on BERT were competitive with BERT for the same task.
- The datasets, code, and pretrained models can be accessed at https://github.com/OussamaMECHHOUR/MAEVa-v1.5.

The remainder of this study is structured as follows: we reference pertinent literature, elucidate their shortcomings, and present our proposal in the subsequent section. Section 3 delineates our proposed methodology, Section 4 presents the experimental configurations and findings, and Section 5 concludes with a summary of our work, its limitations, and prospective avenues for further research.

2 Related Work

In the literature, domain-general models have been adapted to domain-specific tasks through further pretraining, pretraining from scratch, and/or fine-tuning. Further pretraining involves initializing the weights of the domain-specific model with those of domain-general models, followed by additional pretraining on domain-specific corpora.

For instance, **Legal-BERT** [18] is a family of BERT models designed to support legal NLP research, computational law, and legal technology applications. It explores various approaches for adapting BERT to the legal domain through pretraining and fine-tuning. The authors pretrained BERT on the random MLM task using 12 GB of diverse English legal text. They demonstrated that fine-tuning the pretrained Legal-BERT from scratch or further pretraining it outperforms BERT and the fine-tuned BERT on domain-specific legal tasks. Notably, the two pretraining strategies followed by fine-tuning showed comparable performance across three legal datasets. Additionally, the study highlighted that a smaller model pretrained from scratch and fine-tuned on legal data achieves performance comparable to larger models while being significantly more efficient. This smaller model operates approximately four times faster, requires fewer computational resources, and is three times smaller in size, making it a cost-effective and environmentally friendly solution.

BioBERT [19] is another domain-specific adaptation of BERT, focusing on biomedical tasks. It was further pretrained on the random MLM task using biomedical corpora, including PubMed abstracts (4.5 billion words) and PMC full-text articles (13.5 billion words). BioBERT was fine-tuned on three downstream biomedical tasks: (i) Named Entity Recognition (NER), (ii) Relation Extraction (RE), and (iii) Question Answering (QA). For both NER and QA tasks, BioBERT consistently achieved higher scores than BERT in all datasets.

SciBERT [20] pretrained BERT from scratch using the random MLM task on a corpus of 1.14 million full-text papers from Semantic Scholar [21]. The corpus comprises 18% computer science papers and 82% biomedical papers. SciBERT was fine-tuned on several downstream tasks, including Named Entity Recognition (NER), Relation Classification (REL), and Text Classification (TC). The results showed that SciBERT outperforms BERT (both fine-tuned and without fine-tuning) across biomedical, computer science, and multidomain tasks. Furthermore, SciBERT demonstrated better performance than BioBERT on some NER and RE datasets.

AgriBERT [15] pretrained BERT from scratch on the random MLM task, utilizing a large corpus of 46,446 food- and agriculture-related journal articles from 26 journals (4 GB). This domain-specific pretrained model was subsequently fine-tuned for an agriculture-specific downstream task involving the mapping between food descriptions and nutrition data. The results showed that fine-tuning AgriBERT outperformed BERT and the fine-tuned BERT for this specific task.

One shortcoming of these works is the use of large datasets for pretraining domain-specific models, which is time- and resource-intensive. Additionally, none of these works pretrained a general model on a small domain-specific dataset to determine whether it can enhance performance or not. To the best of our knowledge, no study has further pretrained a domain-general model on a small corpus (less than 100 MB) to adapt it to domain-specific and evaluated it on downstream tasks without fine-tuning.

To address these shortcomings, and motivated by the non-existence of agroecology-specific models, we propose to further pretrain BERT and AgriBERT on a small corpus and evaluate the performance of our pretrained models on matching source and candidate variable descriptions.

3 Proposed Approach

BERT is a general language representation model pretrained on English Wikipedia and BooksCorpus, while AgriBERT is pretrained on agricultural corpus. However, both models remain general for agroecology domain-specific tasks, making it challenging to effectively represent specialized vocabularies, particularly in agroecology. For example, terms such as "Intercropping," "Biological pest control," and "Soil organic matter" may not be well understood by BERT and/or AgriBERT due to the general-purpose nature of their training data.

For this reason, it is necessary to pretrain BERT and AgriBERT on an agroecology-related corpus to enhance their ability to represent such vocabularies, thereby improving the performance of downstream tasks. Our proposed approach begins with data preparation, as detailed in Subsection 3.1, followed by the pretraining process described in Subsection 3.2.

3.1 Data Preparation

Fig. 1 illustrates the complete process of collecting our small corpora for further pretraining BERT and AgriBERT. The process begins with the collection of 78 full-text articles provided by three experts in agroecology. In addition, we utilized two APIs to retrieve articles related to agroecology using predefined queries[2]. Through the Core API [22], we retrieved 1,666 full-text articles, while the Europe PMC API[3] provided 2,831 full-text articles. The queries were constructed based on keywords extracted from our variable names and description content, as well as keywords manually defined by agroecology experts. This means that our corpora do not cover all key agroecological terms—given the vastness of the domain [23]—but are instead specific to our task.

[2] https://github.com/OussamaMECHHOUR/MAEVa-v1.5/blob/main/datasets/keywords/default_keywords.txt

[3] https://europepmc.org/RestfulWebService#!/Europe32PMC32Articles32RESTful32API/search

After collecting the PDF articles, we used GROBID[4] (GeneRation Of BIbliographic Data) to extract, parse, and restructure the raw PDFs into XML/TEI-encoded documents. XML/TEI files are well-organized and facilitate the extraction of important information by segmenting the content into structured sections such as introductions, conclusions, and main text. Unnecessary information, such as author affiliations, corresponding author details, page numbers, and inline references, was removed to ensure clean and unbiased content.

For each PDF article, we constructed a corresponding TXT file containing only the essential content, excluding tables, figures, and metadata. Subsequently, the TXT files were merged into three distinct corpora:

- **Corpus (Experts):** Comprising 78 TXT files from expert-gathered articles.
- **Corpus (Core):** Comprising 1,666 TXT files retrieved via the Core API.
- **Corpus (PMC):** Comprising 2,831 TXT files retrieved via the Europe PMC API.

These merged corpora form the foundation for further pretraining, providing a small focused dataset for adapting BERT and AgriBERT to agroecology-specific tasks. Table 2 illustrates basic statistics of these three corpora.

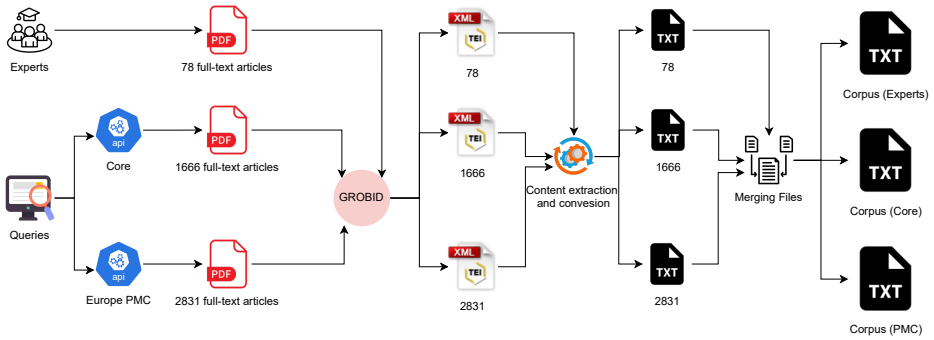

Fig. 1 Full process of data preparation used for the further pretraining of BERT and AgriBERT.

Table 2 Basic statistics of our datasets.

Datasets	Articles	Words	Size (MB)
Corpus (Experts)	78	281,520	1.66
Corpus (Core)	1,666	5,653,653	33.66
Corpus (PMC)	2,831	8,562,323	53.6

[4] https://github.com/kermitt2/grobid

3.2 Further Pretraining: BERT and AgriBERT

The random MLM task is applied in BERT and AgriBERT, where 15% of input tokens are randomly masked, and the model attempts to predict them correctly based on bidirectional context. This task has been widely used in the literature to pretrain domain-specific models. As described in Subsection 3.1, we have three small corpora: (i) Corpus (Experts), (ii) Corpus (Core), and (iii) Corpus (PMC).

As shown in Table 2, Corpus (Experts) is relatively small in size (1.66 MB), which limits its potential for further pretraining of a general model and increases the risk of overfitting. Given that prior research has demonstrated the benefits of combining corpora for domain adaptation, we adopt a similar strategy. For instance, **BioBERT** [19] achieved superior performance on biomedical named entity recognition and question answering tasks by merging corpora from PubMed and PMC, compared to using individual corpora. Similarly, **SciBERT** [20], which was pretrained on a corpus composed of 18% computer science papers and 82% biomedical papers, demonstrated strong results across multiple benchmarks.

Motivated by these findings, we propose the following steps:

- Merge **Corpus (Experts)** and **Corpus (Core)** into a single corpus named **Corpus (Experts + Core)**, increasing the total size and making it more suitable for pretraining.
- Merge all three corpora into a unified corpus named **Corpus (Experts + Core + PMC)** to leverage the strengths of each source.

The final corpora used for further pretraining BERT and AgriBERT are summarized in Table 3.

Table 3 Basic statistics of our final corpora used to further pretrain BERT and AgriBERT.

Datasets	Articles	Words	Size (MB)
Corpus (Experts + Core)	1,744	5,935,173	35.3
Corpus (PMC)	2,831	8,562,323	53.6
Corpus (Experts + Core + PMC)	4,575	14,497,498	89

4 Experiments

4.1 Experimental Setup

Data Preparation Details To retrieve PDF articles using the Core API, it took approximately 5 hours, followed by about 1 hour to extract the Corpus (Core) from those PDF articles, using the detailed steps described in Subsection 3.1. For the Europe PMC API, the retrieval process took approximately 9 hours, and generating the Corpus (PMC) required an additional 1 hour and 15 minutes. The Corpus (Experts) was generated in just 5 minutes.

Our corpora are significantly smaller compared to those used in the literature. For instance, AgriBERT pretrained $\text{BERT}_{\text{BASE}}$ from scratch using 4 GB of agricultural

corpus. $\text{BERT}_{\text{BASE}}$ itself was pretrained on 11,000 full books from BooksCorpus (0.8 billion words) and English Wikipedia articles (2.5 billion words). Similarly, LegalBERT was pretrained from scratch on 12 GB of legal data. In contrast, our largest corpus is just 89 MB.

Further Pretraining Details $\text{BERT}_{\text{BASE}}$ was originally pretrained on BooksCorpus and English Wikipedia for 1M steps (approximately 40 epochs). We refer to this version as $\text{BERT}_{\text{BASE}}$ (BooksCorpus + English Wikipedia). Given our three corpora—Corpus (Experts + Core), Corpus (PMC), and Corpus (Experts + Core + PMC)—we further pretrained $\text{BERT}_{\text{BASE}}$ on each, resulting in the following pretrained models:

- $\text{BERT}_{\text{BASE}}$ (+ Experts + Core),
- $\text{BERT}_{\text{BASE}}$ (+ PMC),
- $\text{BERT}_{\text{BASE}}$ (+ Experts + Core + PMC).

The further pretraining process for $\text{BERT}_{\text{BASE}}$ was identical across all the above corpora. We pretrained $\text{BERT}_{\text{BASE}}$, i.e., the checkpoint "bert-base-uncased" from Hugging Face[5], on the random MLM task for 20 epochs with a batch size of 64, a sequence length of 512, and the same hyperparameters as those used for the original $\text{BERT}_{\text{BASE}}$.

For AgriBERT, we used the checkpoint "agriculture-bert-uncased" from Hugging Face[6], which was further pretrained from the checkpoint of the SciBERT model. We refer to the pretrained AgriBERT on agriculture corpus as AgriBERT (agricultural corpus). For further pretraining of AgriBERT, we used the same methodology as for $\text{BERT}_{\text{BASE}}$, resulting in the following pretrained models:

- AgriBERT (+ Experts + Core),
- AgriBERT (+ PMC),
- AgriBERT (+ Experts + Core + PMC).

All the pretraining tasks were executed on an A100 GPU with 80 GB of memory. The entire pretraining process took 3 days to complete.

4.2 Results and discussion

4.2.1 Results

To evaluate our pretrained models and compare them to the original BERT and AgriBERT models, we selected the matching between source and candidate variable descriptions as a downstream task. The dataset consists of 84 source descriptions and 170 candidate descriptions. Most of these candidate variable descriptions are highly similar, making the matching task challenging and non-trivial. Without fine-tuning, for each source variable description, we calculate its similarity with all candidate variable descriptions and rank the returned results from highest to lowest similarity. Given that each source variable description has exactly one correct candidate variable

[5] https://huggingface.co/google-bert/bert-base-uncased
[6] https://huggingface.co/recobo/agriculture-bert-uncased

description, we use precision at position K (P@K) as our evaluation metric.

Evaluation Metric Precision at position K (P@K) evaluates the model's ability to rank the correct candidate variable description within the top K positions of the similarity-ranked list for each source variable description. Formally, P@K is defined as in Equation 1:

$$P@K = \frac{N_{\text{correct@}K}}{N_{\text{total}}} \tag{1}$$

Where:

- $N_{\text{correct@}K}$: Number of source variable descriptions where the correct candidate is ranked in the top K.
- N_{total}: Total number of source variable descriptions (84).

Table 4 illustrates the results obtained by $\text{BERT}_{\text{BASE}}$ and our further pretrained models on $\text{BERT}_{\text{BASE}}$, while Table 5 illustrates the results obtained by AgriBERT and our further pretrained models on AgriBERT.

Table 4 Results of matching source and candidate variable descriptions using the default $\text{BERT}_{\text{BASE}}$ (row 1) and our three further pretrained models (rows 2-4).

$\text{BERT}_{\text{BASE}}$	P@1	P@3	P@5	P@10
(BooksCorpus + English Wikipedia)	**11.90%**	**15.47%**	**20.23%**	22.61%
(+ PMC)	3.57%	13.09%	**20.23%**	25%
(+ Experts + Core)	2.38%	9.52%	17.85%	**27.38%**
(+ Experts + Core + PMC)	1.19%	2.38%	9.52%	**27.38%**

Note: The best score(s) for each $P@k$ are in bold.

Table 5 Results of matching source and candidate variable descriptions using the default AgriBERT (row 1) and our three further pretrained models (rows 2-4).

AgriBERT	P@1	P@3	P@5	P@10
(agricultural corpus)	28.57%	40.47%	45.23%	51.19%
(+ Experts + Core)	**41.66%**	**50.00%**	52.38%	**60.71%**
(+ PMC)	38.09%	48.80%	**54.76%**	**59.52%**
(+ Experts + Core + PMC)	22.61%	38.09%	46.42%	57.14%

Note: The best score(s) for each $P@k$ are in bold.

4.2.2 Discussion

Table 4 shows that $\text{BERT}_{\text{BASE}}$ outperforms our three pretrained models on P@1 (+8%) and P@3 (+2%). For P@5, our pretrained $\text{BERT}_{\text{BASE}}$ (+ PMC) is equivalent to $\text{BERT}_{\text{BASE}}$, while for P@10, all our pretrained models outperform $\text{BERT}_{\text{BASE}}$ (+5%).

Additionally, to compare the impact of using our three corpora on performance, we observe from Table 4 that the corpus containing articles collected from the Europe PMC API (2,831 articles) outperforms the other pretrained models in P@1, P@3, and P@5 scores. This suggests that the PMC corpus may contain more representative articles for the task compared to those collected by experts and via the Core API. However, when merging all articles (a total of 4,575 articles), the results decrease. The decline in performance after merging all corpora indicates that adding more data does not always improve model performance, possibly due to domain noise or variations in data quality.

Table 5 shows that our pretrained models outperform AgriBERT by 13% on P@1 and P@3, and by 9% on P@5 and P@10. It is not surprising that further pretraining AgriBERT is more efficient than further pretraining $\text{BERT}_{\text{BASE}}$, as agriculture is more closely related to the agroecology field. Moreover, to compare the impact of using our three corpora on performance, we observe from Table 5 that the corpus containing articles collected from experts and the Core API (1,744 articles) outperforms the PMC corpus (2,831 articles) on P@1 and P@3. However, when merging all corpora (4,575 articles), the results decrease across all P@K values. This further confirms that adding more data does not always enhance model performance, likely due to domain noise or inconsistencies in data quality.

From both tables, we conclude that our further pretrained AgriBERT models outperform AgriBERT, $\text{BERT}_{\text{BASE}}$, and all our pretrained $\text{BERT}_{\text{BASE}}$ models on our task. With a very small dataset, short training time, and without fine-tuning, we achieved enhanced results, which positively answers our research question (RQ). Pretraining on a small dataset can improve results, especially when the base model is significantly different from the domain-specific task. However, when there is a close relationship between the general model used as the base and the domain-specific task, the results improve significantly while requiring less time, fewer resources, shorter pretraining durations, and no need for fine-tuning.

Error analysis Even though our pretrained models achieved strong performance, they still exhibit some errors. To analyze these errors, we selected our best-performing models—**$\text{BERT}_{\text{BASE}}$ (+ PMC)** and **AgriBERT (+ Experts + Core)**—and focused on the $P@10$ metric. This means we considered all source descriptions for which the correct candidate description was not retrieved within the top 10 proposed candidates.

We observed that the proposed candidate descriptions were semantically distant and/or semantically close to the source description. In all cases, the proposed candidates could share the same unit of measurement as the source description or not.

- **Row 1** in Table 6 illustrates a case where the proposed candidate is semantically distant and does not share the same unit of measurement with the source description.
- **Row 2** shows a proposed candidate that is also semantically distant but shares the same unit of measurement with the source description.

- **Row 3** presents a proposed candidate that is semantically close[7] but does not share the same unit of measurement with the source description.
- **Row 4** shows a proposed candidate that is semantically close and shares the same unit of measurement with the source description.
- **Rows 1–2 and 3–4** highlight that, within the top 10 proposed candidates, both types of errors—semantic and unit mismatch—can occur simultaneously for a single source description.

These observations highlight the importance of considering units of measurement alongside textual descriptions when matching variable descriptions. One potential solution for future work is to use a unit-of-measurement-based filter, comparing source descriptions only with candidate descriptions that share the same unit dimension. This approach could simplify and improve the matching process and increase the likelihood of correct matches. However, it also presents a major limitation: two variables may share the same unit of measurement dimension but still be semantically unrelated.

To address this in future work, we plan to explore a hybrid approach that combines unit of measurement matching with semantic similarity based on descriptions.

Table 6 Samples illustrating all types of errors made by our models.

Source Variables		**Candidate Variables**	
Description	**Unit**	**Description**	**Unit**
Weed coverage index over the trial	%	Height of the apex of the sugar stem sample	cm
Weed coverage index over the trial	%	soluble sugar concentration of the leaves	%
Cane yield (in fresh machinable stem)	$t.ha^{-1}$	total fresh biomass	kg
Cane yield (in fresh machinable stem)	$t.ha^{-1}$	aerial biomass of straws at the harvest of the associated plant per unit area measured on the plot	$t.ha^{-1}$
...	...	...	...

Further analysis of errors made by **$\text{BERT}_{\text{BASE}}$ (+ PMC)** and **AgriBERT (+ Experts + Core)** also revealed these models do not make the same mistakes:

1. Both models correctly matched 17 of the same source descriptions.
2. 48 source descriptions were correctly matched by **AgriBERT (+ Experts + Core)** but not by **$\text{BERT}_{\text{BASE}}$ (+ PMC)**.
3. 19 source descriptions were correctly matched by **$\text{BERT}_{\text{BASE}}$ (+ PMC)** but not by **AgriBERT (+ Experts + Core)**.

These results suggest that the two models are complementary. Therefore, we plan to investigate strategies for combining them in future work.

[7] These variables are semantically close because cane yield is a specific part of the total fresh biomass, both measured in fresh weight and referring to sugarcane productivity.

5 Conclusion

In this paper, we sought to address the following research question (RQ): *Can further pretraining a general model (i.e., BERT and AgriBERT) on a small agroecology-related corpus outperform these general models on agroecology tasks without fine-tuning?* To answer this RQ, we collected a small dataset (less than 100 MB) from agroecology-related full-text articles and further pretrained $\text{BERT}_{\text{BASE}}$ and AgriBERT models on the random Masked Language Modeling (MLM) task. We evaluated these models on the task of matching source and candidate variable descriptions, a challenging task in the agroecology field, and demonstrated that:

- Our further pretrained AgriBERT models outperformed $\text{BERT}_{\text{BASE}}$, AgriBERT, and all our pretrained $\text{BERT}_{\text{BASE}}$ models on this task.

- $\text{BERT}_{\text{BASE}}$ and our pretrained $\text{BERT}_{\text{BASE}}$ models are competitive.

Our study shows that it is possible to further pretrain domain-general models into domain-specific models and achieve efficient results using a small dataset, minimal pretraining time, limited computational resources, and no need for fine-tuning. Although these results are encouraging, our work has certain limitations. One limitation lies in the use of the random MLM task, which masks 15% of input tokens randomly and attempts to predict them. This approach may not effectively target important tokens, which restricts the model's ability to learn domain-specific terminology. A known alternative in the literature is the use of selective masking, where tokens to be masked are chosen based on their importance. For example, EntityBERT [24] selectively masks detected clinical entities, which has shown promising results. Another limitation of our work is the absence of agroecology-related ontologies, which are rich in semantic information such as synonyms, labels, definitions, and relationships between entities. Incorporating such ontologies could significantly enhance the model's ability to understand domain-specific knowledge. In this work, we applied two models: one is too general (BERT), and the other is more closely related to our domain (AgriBERT).

As future work, we plan to compare the performance of random MLM and selective masking tasks during pretraining, leverage agroecology-related ontologies rather than relying solely on article-based corpora, and apply more general models to compare them. We also aim to compare our innovative approach with approaches such as fine-tuning.

Acknowledgements. This work was parpartially funded by the French National Research Agency (ANR) as part of the France 2030 program, under the reference ANR-16-CONV-0004 (#DigitAg), and by the Horizon Europe research and innovation program under grant agreement 101081973 (IntercropVALUES). This research received support from the Regional Council of La Réunion, the French Ministry of Agriculture and Food, and the European Union (Feder program, grant AG/974/DAAF/2016-00096 and Feder program, grant GURTDI 20151501-0000735).

References

[1] Auzoux, S., Christina, M., Goebel, F.-R., Mansuy, A., Marion, D.: A dictionary

of variables to harmonize data from agro-ecological experiments on sugarcane. (2018). ISSCT

[2] Walls, R., Cooper, L., Elser, J., Gandolfo, M., Mungall, C., Stevenson, D., Jaiswal, P.: The plant ontology facilitates comparisons of plant development stages across species. Frontiers in Plant Science **10** (2019) https://doi.org/10.3389/fpls.2019.00631

[3] Arnaud, E., Matteis, L., Laporte, M.-A., Espinosa, H., Hyman, G., Shrestha, R., Portugal, A., Chibon, P., Devare, M., Akintunde, A., White, J., Wilkinson, M., Caracciolo, C., Celli, F., Mclaren, G.: The crop ontology, a resource for enabling access to breeders'data. (2014)

[4] Buttigieg, P.L., Morrison, N., Mungall, C., Lewis, S.: The environment ontology: contextualising biological and biomedical entities. Journal of biomedical semantics **4**, 43 (2013) https://doi.org/10.1186/2041-1480-4-43

[5] Aubert, C., Buttigieg, P.L., Laporte, M.-A., Devare, M., Arnaud, E.: CGIAR Agronomy Ontology. Licensed under CC BY 4.0 (2017). http://purl.obolibrary.org/obo/agro.owl

[6] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[7] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Fasttext: Efficient learning of word representations and sentence classification. arXiv preprint arXiv:1607.01759 (2016)

[8] Pennington, J., *et al.*: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[9] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR **abs/1802.05365** (2018) 1802.05365

[10] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019). https://api.semanticscholar.org/CorpusID:52967399

[11] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**, 1735–1780 (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[12] Yin, J.: Research on question answering system based on bert model. In: 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL &

ICCEA), pp. 68–71 (2022). https://doi.org/10.1109/CVIDLICCEA56201.2022.9824408

[13] Yu, S., Su, J., Luo, D.: Improving bert-based text classification with auxiliary sentence and domain knowledge. IEEE Access **7**, 176600–176612 (2019) https://doi.org/10.1109/ACCESS.2019.2953990

[14] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., Liu, T.: Incorporating BERT into neural machine translation. CoRR **abs/2002.06823** (2020) 2002.06823

[15] Rezayi, S., Liu, Z.-L., Wu, Z., Dhakal, C., Ge, B., Zhen, C., Liu, T., Li, S.: Agribert: Knowledge-infused agricultural language models for matching food and nutrition. In: International Joint Conference on Artificial Intelligence (2022). https://api.semanticscholar.org/CorpusID:250635911

[16] Araci, D.: Finbert: Financial sentiment analysis with pre-trained language models. CoRR **abs/1908.10063** (2019) 1908.10063

[17] Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. CoRR **abs/1904.05342** (2019) 1904.05342

[18] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: the muppets straight out of law school. CoRR **abs/2010.02559** (2020) 2010.02559

[19] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. CoRR **abs/1901.08746** (2019) 1901.08746

[20] Beltagy, I., Cohan, A., Lo, K.: Scibert: Pretrained contextualized embeddings for scientific text. CoRR **abs/1903.10676** (2019) 1903.10676

[21] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H., Peters, M.E., Power, J., Skjonsberg, S., Wang, L.L., Wilhelm, C., Yuan, Z., Zuylen, M., Etzioni, O.: Construction of the literature graph in semantic scholar. CoRR **abs/1805.02262** (2018) 1805.02262

[22] Knoth, P., Novotny, J., Zdrahal, Z.: Automatic generation of inter-passage links based on semantic similarity. In: Computational Linguistics (COLING 2010), pp. 590–598 (2010). http://oro.open.ac.uk/22933/

[23] Helmer, T., Roche, M., Martin, P., Enten, F., Reynaud, L., Lebret, M.-C., Bienabe, E., Blanchard, M., Ehrensperger, A., Hernandez, R., Priour, G.: ASSET Theoretical Lexicon: An Agroecology Lexicon. https://doi.org/10.18167/DVN1/TVN3AC . https://doi.org/10.18167/DVN1/TVN3AC

[24] Lin, C., Miller, T., Dligach, D., Bethard, S., Savova, G.: EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In: Demner-Fushman, D., Cohen, K.B., Ananiadou, S., Tsujii, J. (eds.) Proceedings of the 20th Workshop on Biomedical Language Processing, pp. 191–201. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.bionlp-1.21 . https://aclanthology.org/2021.bionlp-1.21/