Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Special Track on AI for Good

# AgriBERT: Knowledge-Infused Agricultural Language Models for Matching Food and Nutrition

**Saed Rezayi**[1], **Zhengliang Liu**[1], **Zihao Wu**[1], **Chandra Dhakal**[1], **Bao Ge**[2], **Chen Zhen**[1*], **Tianming Liu**[1*] and **Sheng Li**[1*]
[1]University of Georgia
[2]Shanxi Normal University
{saedr,zl18864,zihao.wu1,chandra.dhakal25,czhen,tliu,sheng.li}@uga.edu, bob_ge@snnu.edu.cn

## Abstract

Pretraining domain-specific language models remains an important challenge which limits their applicability in various areas such as agriculture. This paper investigates the effectiveness of leveraging food related text corpora (e.g., food and agricultural literature) in pretraining transformer-based language models. We evaluate our trained language model, called AgriBERT, on the task of semantic matching, i.e., establishing mapping between food descriptions and nutrition data, which is a long-standing challenge in the agricultural domain. In particular, we formulate the task as an answer selection problem, fine-tune the trained language model with the help of an external source of knowledge (e.g., FoodOn ontology), and establish a baseline for this task. The experimental results reveal that our language model substantially outperforms other language models and baselines in the task of matching food description and nutrition.

## 1 Introduction

United States Department of Agriculture (USDA) maintains a database called Food and Nutrient Database for Dietary Studies (FNDDS) which provides the nutrient values for foods and beverages reported in what is eaten in the US[1]. Additionally, household and retail scanner data on grocery purchases, such as the Nielsen data available through the Kilts Center for Marketing, have been extensively used in food policy research[2]. Mapping these two databases, i.e., food description found in retail scanner data, to nutritional information database is of utmost importance. This linkage can capture the relationship between the retail food purchase and community health and also the difference between poor and non-poor diets across the whole diet spectrum, and thus it can impact future funding policies that can provide healthy food for low-income households.

In this work, we aim to develop and employ Natural Language Processing (NLP) techniques to find the best linkage between the two databases. A common approach to tackle this kind of problems is semantic matching, which is the task of determining whether two or more elements have similar meaning. Bi-encoders are the most common techniques for semantic matching. A bi-encoder inputs two strings and encodes them in the embedding space and in the final layer calculates the similarity in a supervised fashion. That is why word embedding techniques are a good candidate for this purpose. Word embeddings have been utilized extensively for the task of semantic matching [Kenter and De Rijke, 2015], but recent advancements in contextual word embeddings, in which each word is assigned a vector representation based on the context, have resulted in significant improvements in many NLP tasks, including semantic matching.

Transformer-based language models, e.g., BERT [Devlin *et al.*, 2019], have been widely used in research and practice to study computational linguistics and they have shown superior performance in variety of applications including text classification [Jin *et al.*, 2020], question answering [Yang *et al.*, 2019], and many more. However, these models are not generalizable to every domain when used with their default objectives, i.e., pretrained on generic corpora such as Wikipedia. To address this issue, previous work has attempted to incorporate domain-specific knowledge into the language model by different strategies. One of the prominent approaches is in biomedical domain where a BERT-based language model is pretrained on a large corpus of biomedical literature called BioBERT [Lee *et al.*, 2020]. Motivated by the impressive performance of BioBERT, we use a large corpus of agricultural literature to train a language model for agricultural applications from scratch. The trained model will be further fine-tuned by the downstream tasks.

Another method to incorporate domain knowledge into the language model is to use an external source of knowledge such as a knowledge graph (KG). Knowledge graphs are rich sources of information that are carefully curated around ob-

*Corresponding authors

[1]https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds/

[2]Researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

jects called entities and their relations. A basic building block of a KG is called a triplet which consists of two entities and a relation between the two. Previous work has attempted to inject triples into the sentences [Liu *et al.*, 2020], however, injecting triples can introduce noise to the sentences which will mislead the underlying text encoder. To address this issue, we propose to add $n$ entities from an external knowledge source (i.e., a knowledge graph) based on similarity that can be obtained by various methods such as entity linking. This augments the semantic space but keeps the vocabularies within the domain. We show how changing $n$ can affect the performance of the downstream task, quantitatively and qualitatively.

Moreover, we propose to formulate the task of mapping retail scanner data (also known as Nielsen) to USDA description as an answer selection problem. Given a question and a set of candidate answers, answer selection is the task of identifying which of the candidates answers the question correctly. An answer selection model inputs a pair of question and answer and outputs a binary label (true or false), so it is a binary classification problem. Similarly, we can consider Nielsen product descriptions as the set of all questions and USDA descriptions as the set of all answers, and the goal is to find the best answer for each question. The difference is that in the original answer selection task, the number of answers is limited and usually unique to each question, however, in this setting there is a shared set of answers and its size is much larger. We use our pre-trained language model as the backbone for the answer selection component and we augment both questions and answers during fine-tuning using external knowledge to boost the performance. In summary, we make the following contributions in this paper.

- We collect a large-scale corpus of agricultural literature with more than 300 million tokens. This domain corpus has been instrumental to fine-tune generic BERT into AgriBERT.
- We propose a knowledge graph guided approach to augment the dataset for the answer selection component. We inject related entities to the sentences before the fine-tuning step.
- AgriBERT substantially outperforms existing language models on USDA datasets in the task answer selection. We plan to release our datasets and language models to the community upon publication.

The rest of the paper is organized as follows: in the next section we discuss related works in language modeling in specific domains, next in the Section 3.3 we describe our proposed approach to train a language model in agricultural domain and discuss how we inject external knowledge in the fine-tuning step. In Section 4.5 we introduce different datasets including our corpus for training a language model, the external sources of knowledge, and finally the food dataset to evaluate our language model. We conclude our paper in Section 5.

# 2 Related Works

## 2.1 Pre-trained Language Models

In NLP, Pre-trained language models learn from large text corpora and build representations beneficial for downstream tasks. In recent years, there are two successive generations of languages models. Earlier models, such as Skip-Gram [Mikolov *et al.*, 2013] and GloVe [Pennington *et al.*, 2014], primarily focus on learning word embeddings from statistical patterns, semantic similarities and syntactic relationships at the word level. With this first group of language embedding methods, polysemous words are mapped to the same representation, irregardless of word contexts. For example, the word "bear" in "I see a bear" and "Rising car sales bear witness to population increase in this area" will not be distinguishable in the vector space.

A later group of models, however, recognizes the importance of textual contexts and aims to learn context-dependent representations at the sentence level or higher. For example, CoVe [McCann *et al.*, 2017] utilizes a LSTM model trained for machine translation to encode contextualized word vectors. Another popular model, Bidirectional Encoder Representations from Transformers (BERT) [Devlin *et al.*, 2019] is based on bidirectional transformers and pre-trained with Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks, both ideal training tasks for learning effective contextual representations from unlabelled data. It delivers exceptional performance and can be easily fine-tuned for downstream tasks.

BERT has enjoyed wide acceptance from the NLP community and practitioners from other domains. In particular, domain experts can build domain-specific BERT models that cater to specific environments and task scenarios.

## 2.2 Domain Specific Language Models

BERT has become a fundamental building block for training task specific models. It can be further extended with domain specific pre-training to achieve additional gains over general domain models.

Prior work has shown that language models perform better when the source and target domains are highly relevant [Lee *et al.*, 2020; Gu *et al.*, 2021]. In other words, pre-training BERT models with in-domain corpora can significantly improve overall performance on a wide variety of downstream tasks [Gu *et al.*, 2021].

There is also a correlation between a model's performance and the extent of domain specific training [Gu *et al.*, 2021]. In particular, Gu et al. [Gu *et al.*, 2021] note that training models from scratch (i.e., not importing pre-trained weights from the original BERT model [Devlin *et al.*, 2019] or any other existing BERT-based models) is more effective than simply fine-tuning an existing BERT model with domain specific data.

In this paper, agricultural text such as food-related research papers are considered in-domain while other sources such as Wikipedia and news corpus are regarded as out-domain or general domain. Our primary approach is in line with training-from-scratch with in-domain data.

| Product Description | USDA Description | Label |
|---|---|---|
| domino white sugar granulated 1lb | salsa, red, commercially-prepared | False |
| domino white sugar granulated 1lb | cookie-crisp | False |
| domino white sugar granulated 1lb | sugar, white, granulated or lump | True |

Table 1: An example of how we propose to extend the dataset.

### 2.3 Augmenting Pre-trained Language Models

Data augmentation refers to the practice of increasing training data size and diversity without collecting new data [Feng *et al.*, 2021]. Data augmentation aims to address practical data challenges related to model training. It is applicable to scenarios such training with low-resource languages [Xia *et al.*, 2019], rectifying class imbalance [Chawla *et al.*, 2002], mitigating gender bias [Zhao *et al.*, 2018], and few-shot learning [Wei *et al.*, 2021].

Some data augmentation methods incorporate knowledge infusion. For example, Feng et al. [Feng *et al.*, 2020] used WordNet [Miller, 1995] as the knowledge base to replace words with synonyms, hyponyms and hypernyms. Another study [Grundkiewicz *et al.*, 2019] extracts confusion sets from the Aspell spellchecker to perform synthetic data generation in an effort to enhance the training data, which consists of erroneous sentences used for training a neural grammar correction model.

However, there is limited research on the efficacy of applying data augmentation to large pre-trained language models [Feng *et al.*, 2021]. In fact, some data augmentation methods have been found to have limited benefit for large language models [Feng *et al.*, 2021; Longpre *et al.*, 2020]. For example, EDA [Wei and Zou, 2019], which consists of 4 operations (synonym replacement, random insertion, random swap and random deletion), provides minimal performance enhancement for BERT [Devlin *et al.*, 2019] and RoBERTa.

Nonetheless, researchers [Feng *et al.*, 2021] advocate for more work to explore scenarios in which data augmentation is effective for large pre-trained language models, because some studies [Shi *et al.*, 2021] demonstrate results contrary to the claims of [Longpre *et al.*, 2020].

In this study, we investigate the effectiveness of data augmentation with knowledge infusion and apply our method to the Answer Selection task scenario. We find that our method significantly improves semantic matching performance.

### 2.4 Answer Selection

Answer Selection refers to the task of finding the correct answer among a set of candidate answers for a specific question. For example, given the question "What is the capital of France?", a solution to this task is required to select the correct answer among the following choices:

- A) Paris is the capital of France.
- B) Paris is the most populous city in France.
- C) London and Paris are financial hubs in Europe.

In this case, the first answer should be selected. It is clear that matching words or phrases is not sufficient for this task.

A common approach is to formulate this problem as a ranking problem such that candidate answers are assigned ranking scores based on their relevance to the question. Earlier work primarily relies on feature engineering and linguistic information [Yih *et al.*, 2013]. However, the advance of deep learning introduces powerful models [Rücklé *et al.*, 2019; Laskar *et al.*, 2020] that outperform traditional methods without the need of manual efforts or feature engineering.

In this study, our goal is to establish valid mapping between food descriptions and nutrition data. We formulate this task as an Answer Selection problem and demonstrates the superiority of our method over baselines.

## 3 Methodology

### 3.1 Domain Specific Language Model

Training language models is a powerful tool for a variety of NLP applications, and when it comes to a particular task in a specific domain, it becomes more effective if the language model is trained on a corpora that contain large amount of text in that specific domain. Such practices exist in the literature in various domains, for instance BioBERT and FinBERT are successful examples of training a domain-specific language models in biomedical and financial domains, respectively. Building upon previous research and motivated by the lack of existing corpora or a pre-trained model in agricultural domain and because we are interested in a model that produces vocabulary and word embeddings better suited for this domain than the original BERT, we collect 46,446 articles related to food and agricultural that contain more than 300 million tokens and use it to train a BERT model from scratch (more details about the dataset are provided in Section 4.1). This trained model can be used for various NLP applications in agricultural field. We adopt the standard procedure in training a language model which is masked language modeling. In Masked Language Modelling, a certain fraction of words in a given sentence are masked, and the model is expected to predict those masked words based on other words in that sentence. In this process it learns meaningful representation for each sentence.

### 3.2 Answer Selection Problem Definition

To evaluate the trained language model we require a downstream task in this specific domain. For instance most biomedical language models are evaluated on named entity recognition tasks on medical datasets. Due to the lack of benchmark NLP dataset in this domain, and since the semantic matching problem has practical values, we evaluate our model on this task. Semantic matching is a technique to determine whether two sentences have similar meaning. We

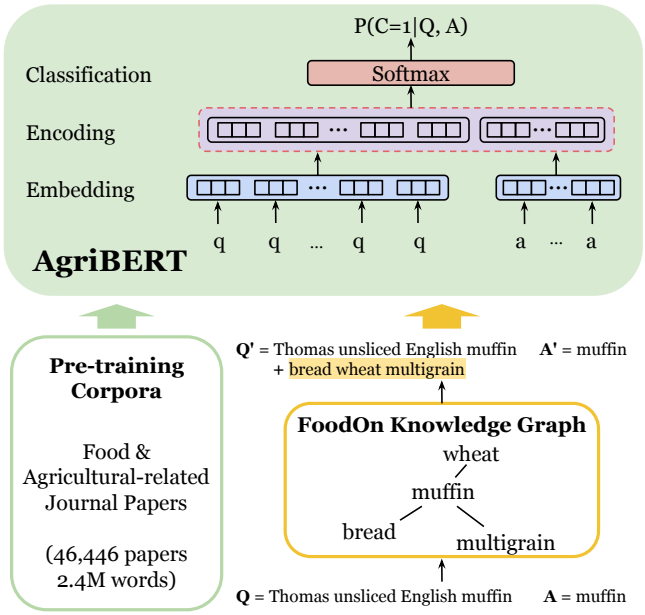


Figure 1: The overall framework of AgriBERT which is trained on agriculture literature from scratch. AgriBERT is evaluated on answer selection task. The answer selection component has two inputs: a question and an answer, and before we input them to the framework, we add new entities to them from an external source of knowledge such as Wikidata or FoodOn. The output of the framework is a score (a probability) which is used for ranking the answers.

express the task at hand as an answer selection problem, we could assign all USDA descriptions (answers) to each Nielsen product description (question) and select $S$ random incorrect USDA description per product description as negative samples where $S << D$. In this case the size of extended dataset is $S \times D$. Table 1 provides an example of how we extend the dataset when $S = 2$.

### 3.3 Knowledge Infused Finetuning

As discussed in 2.4, there have been studies that successfully inject related information from an external source of knowledge to enhance the performance of the downstream task. For instance incorporating facts (i.e., a curated triple extracted from a knowledge graph in the form of `(entity,relation,entity)`) from knowledge graphs [Liu *et al.*, 2020], or injecting refined entities extracted from text to a knowledge graph [Rezayi *et al.*, 2021]. In our setting, since we are dealing with answer selection and the size of training set is small, we propose to append external knowledge to both questions and answers to enhance the performance of the answer selection module.

Finding relevant external knowledge can be fulfilled via different mechanisms such as entity linking, querying, calculating similarity, etc. In this paper we suggest to use entity linking and querying. In entity linking, all the named entities in a text are recognized and then linked to the entities of a knowledge graph which is an ideal solution for our case. However the downside of this approach is that both entity recognition and entity linking algorithms are commonly trained on general text corpora such as Wikipedia and in our case it is not of practical value if we use these tools without reconfiguring them for our purposes. Hence, we consider a domain-specific knowledge graph, e..g, FoodOn [Dooley and Griffiths, 2018] and we obtain new knowledge by querying it using the keywords in the text of question answer pairs. We simply append new entities to the end of questions or answers. More details about this knowledge graph will be provided in Section 4.1. Figure 1 illustrates our proposed framework.

| Dataset | Articles | Tokens | Words | Size |
|---|---|---|---|---|
| Penn Treebank | - | 887,521 | 10,000 | 10MB |
| WikiText-103 | 28,475 | 103,227,021 | 267,735 | 0.5GB |
| Agriculture corpus | 46,446 | 311,101,592 | 2,394,343 | 4.0GB |

Table 2: Basic statistics of our dataset compared with two benchmark datasets in the standard language modeling field.

## 4 Experiment

### 4.1 Datasets

We employ several datasets for training the language models, evaluating the trained language model, and augmenting the downstream dataset. In this section we briefly explain these datasets.

**Language Training Datasets**

Our main dataset is a collection of 46,446 food- and agricultural-related journal papers. We downloaded all published articles from 26 journals and converted the pdf files to text format for use in the masked language modeling task. We also cleaned the dataset by removing URLs, emails, references, and non-ASCII characters. In order to compare the contributions of different components of our model, we consider a secondary dataset for training (WikiText-103) that contains all articles extracted from Wikipedia. This datasets also retains numbers, case, and punctuation, which is similar to our dataset described above. The statistics of the two datasets is provided in Table 2. We also include Penn Treebank dataset [Marcus *et al.*, 1994], which is another common dataset for the task of language modeling, as a reference.

**Answer Selection Dataset**

We use two different data sources for this part. First, we use the consumer panel product from the Nielsen Homescan data. Nielsen provides very granular data on the food purchases from the stores at the product barcode or Universal Product Code (UPC) with detailed attributes for each UPC, including UPC description. While scanner data come with some nutrition-related product attribute variables, this information is not sufficient to examine the nutritional quality. To address this issue, we link product level data from the Nielsen with the USDA Food Acquisition and Purchase Survey (FoodAPS) that supplements scanner data with the detailed nutritional information. The survey contains detailed information about the food purchased or otherwise acquired for consumption during a seven-day period by a nationally representative sample of 4826 US households. The FoodAPS matched 32000+

barcodes with the Food and Nutrient Database for Dietary Studies (FNDDS) food codes of high quality. The linked data set has UPC description for each product, and the corresponding FNDDS food code. In addition, the final data set has full information needed to construct diet quality indexes to evaluate the healthfulness of overall purchases.

**External Source of Knowledge**
We use FoodOn knowledge graph for the augmentation purposes. FoodOn is formatted in the form of OWL ontology. The OWL ontology provides a globally unique identifier (URI) for each term which is used for lookup services and facilitates the query processing system. Much of FoodOn's core vocabulary comes from transforming LanguaL, a mature and popular food indexing thesaurus [Dooley and Griffiths, 2018]. That is why FoodOn is a unique and valuable resource for enhancing our language model.

## 4.2 Metrics

Since the output of the evaluation task on the answer selection dataset is a ranked list of answers per question, we require metrics that take into account the order of results. That is why we propose to use precision at 1 (P@1) and Mean Average Precision (MAP).

**Precision@1 or P@1.** We sort the selected answers based on final similarity score and we count how many times the top answer is correctly selected.

$$P@1 = \sum_{i=1}^{|N|} 1 \text{ if } rank_{a_i} == 1$$

where N is the set of all questions.

**Mean Average Precision (MAP).** MAP measures the percentage of relevant selected answers. Given a ranked list of selected answers per question we mark them as relevant if they are correctly selected and calculate AP as follows:

$$\text{AP} = \frac{1}{n}\sum_{i=1}^{n}(P(i) \times \text{rel}(k)),$$

where n is the set of all selected answers, $\text{rel}(k) \in \{0, 1\}$ indicates if the answer is relevant or not, and $P(i)$ is the precision at $i$ in the ranked list. Once we obtain AP for each question we can average across all questions to find MAP:

$$\text{MAP} = \frac{1}{|N|}\sum_{q=1}^{|N|}\text{AP(q)},$$

where $N$ is the set of all questions.

## 4.3 Baselines

To perform ablation study and make sure that our dataset improves the performance of the downstream task and not using a pretrained model nor training from scratch, we consider following scenarios:

- *k*NN: we compute the embeddings[3] of the Nielsen product descriptions and USDA descriptions and for each vector belonging to the product description embedding space we find the most similar vector from the USDA description embedding space. This naive approach is effective if the number of unique USDA descriptions is small. However, this does not hold in our case.
- We use BERT without any modification as the underlying language model and use an existing answer selection technique to further fine-tune the model on the answer selection dataset.
- We consider BERT as the base language model and further train it with our own corpus. In this case the vocabularies of the final language model is the union of Wikipedia and our corpus. We employ the fine-tuned language model as the backbone of the answer selection tool and apply it on the unmodified answer selection dataset.
- We train a BERT model from scratch using masked language modeling technique on WikiText-103. We employ the fine-tuned language model as the backbone of the answer selection tool and apply it on the unmodified answer selection dataset.
- We train a BERT model from scratch to train a new language model using masked language modeling technique on our own dataset. We employ the fine-tuned language model as the backbone of the answer selection tool and apply it on the unmodified answer selection dataset.

For the trained language model we also consider a scenario where we use entity linking algorithm to find related entities from Wikidata and append them to question and answers. As discussed this approach introduces noise to the text and may harm the performance but we include it as a baseline for the sake of comparison with the case where we query the FoodOn knowledge graph to augment the text of questions and answers.

## 4.4 Experimental Settings

Once the extended dataset is generated we can apply any answer selection method on the dataset. There are a number of studies in the literature on this topic, including COALA [Rücklé *et al.*, 2019], CETE [Laskar *et al.*, 2020], MTQA [Deng *et al.*, 2019], and many more, among which CETE is considered state-of-the-art in the answer selection task by the ACL community[4]. CETE implements a transformer-based encoder (e.g., BERT) to encode the question and answer pair into a single vector and calculates the probability that a pair of question/answer should match or not[5].

For the entity linking process we use the implementation proposed by [Wu *et al.*, 2020], called BLINK. BLINK is an entity linking python library that uses Wikipedia as the target knowledge base. Moreover to send in efficient SPARQL queries to FoodOn knowledge graph which is in the format of

[3]We use sentence-transformer library for this task: https://github.com/UKPLab/sentence-transformers

[4]Reported here: https://aclweb.org/aclwiki/Question_Answering_(State_of_the_art)

[5]The code for this study is open source and available for public use: https://github.com/tahmedge/CETE-LREC

| Sentence | Wikidata entity | FoodOn entity |
|---|---|---|
| nestle nido powder infant formula | nestle | rice powder |
| aunt jemima frozen french toast breakfast entree | aunt jemima | frozen dairy dessert |
| woodys hickory barbecue cooking sauce | woody's chicago style | hickory nut |
| sour punch sour watermelon fruit chew straw | sour punch | sour milk beverage |
| philly steak frozen beef sandwich steak | philly steaks | wagyu steak |
| yoplait original rfg harvest peach yogurt low fat | yoplait | creamy salad dressing |

Table 3: Examples to demonstrate the quality of added entities to the text of product descriptions. For Wikidata we use entity linking and we present here the top linked entity (highest confidence score). For FoodOn we use SPARQL to query the ontology and the first outcome is listed here.

| # | Training Dataset | Model | MAP | P@1 |
|---|---|---|---|---|
| 1 | - | kNN | 26.70 | 14.49 |
| 2 | - | $\text{BERT}^{\text{p}}$ | 27,77 | 10.88 |
| 3 | WikiText-103 | $\text{BERT}^{\text{p}}$ | 28.03 | 11.12 |
| 4 | WikiText-103 | $\text{BERT}^{\text{s}}$+EL (Wikidata) | 27.36 | 10.09 |
| 5 | WikiText-103 | $\text{BERT}^{\text{s}}$+FoodOn (n=1) | 28.78 | 24.83 |
| 6 | Agricultural Corpus | $\text{BERT}^{\text{p}}$ | 29.72 | 12.71 |
| 7 | Agricultural Corpus | $\text{BERT}^{\text{s}}$ | **44.21** | 22.72 |
| 8 | Agricultural Corpus | $\text{BERT}^{\text{s}}$+EL (Wikidata) | 42.33 | 21.52 |
| 9 | Agricultural Corpus | $\text{BERT}^{\text{s}}$+FoodOn (n=1) | 31.54 | 47.89 |
| 10 | Agricultural Corpus | $\text{BERT}^{\text{s}}$+FoodOn (n=3) | 30.65 | 49.80 |
| 11 | Agricultural Corpus | $\text{BERT}^{\text{s}}$+FoodOn (n=5) | 29.91 | **49.98** |

Table 4: Test performances of all models trained on all datasets for the task of answer selection. for kNN model we use sentence-transformers to compute embeddings. EL stands for Entity Linking and bold numbers indicate the best performance. $\text{BERT}^{\text{p}}$ is a pre-trained BERT and $\text{BERT}^{\text{s}}$ means training a BERT model from scratch.

OWL ontology we use ROBOT which is a tool for working with Open Biomedical Ontologies[6]. Additionally, we try to simulate a setting where the number of labeled training samples is small, thus we use 20% of the dataset for training and the remaining 80% for the test. We believe this is a more realistic scenarios in real world applications.

### 4.5 Results

As Table 4 presents, not surprisingly, the best performance is obtained when the language model is trained on the agricultural corpus. We summarize the main observations as follows: First the *k*NN performs surprisingly well compared to other complicated methods, in fact in terms of P@1 it surpasses methods that are trained with BERT and the answers are selected using state-of-the-art answer selection framework. Next, the best MAP score (MAP$= 44.21\%$) is obtained when the language model is trained with agricultural corpus from scratch, and any additional augmentation hurts the performance in terms of MAP. On the other hand, the model performance improves in terms of P@1 when incorporating external knowledge. More specifically, P@1$= 49.98\%$ when 5 new entities are added. This implies that by adding related external knowledge we can find the correct match roughly 50% of the times but if the correct match is not at the top position they ranked very low, and hence the low MAP score.

Moreover, by comparing lines 5 and 9 we can see that the inclusion of external knowledge alone cannot improve the quality of language model in a domain specific task. We also investigate the number of external entities that we include in the text of questions and answers and as lines 9-10 of Table 4 demonstrates, increasing $n$ decreases the MAP score and increases the P@1. This suggests that incorporating related entities from a relevant knowledge source helps to find the correct match in 50% of the times but it mislead the answer selection module and rank the correct match lower that it drops the MAP score. Table 3 provides some examples of augmented sentences by Wikidata and FoodOn knowledge sources. As this table presents, linking the food description to Wikidata entities can easily go wrong, first three rows for instance, where the food descriptions are linked to brand names[7]. However this does not happens in querying method, as the entities are purely food related. Additionally, FoodOn entities contain food-related adjectives such as frozen, creamy, etc. that help in matching the food descriptions to nutrition data.

## 5 Conclusion

In this paper we trained a language model called AgriBERT that will facilitate the NLP tasks in the food and agricultural domain. AgriBERT is a BERT model trained from scratch with a large corpus of academic journal in the field. To evaluate our language model we propose to solve the problem of semantic matching which aims at matching two databases of food description (Nielsen database and USDA database). We reformulate the problem as an answer selection task and used our language model as a backbone of a generic answer selection module to find the best match. Before feeding the pairs of questions and answers to the model we augmented them with external entities obtained from FoodOn knowledge graph, a domain-specific ontology in the field of food. We showed that inclusion of external knowledge can help boost the performance in terms of the more strict P@1 measure but it lowers the performance in terms of mean average precision. As a future direction, we plan to investigate more sophisticated approaches for incorporating external knowledge such as refining the knowledge before including it in the text.

## References

[Chawla *et al.*, 2002] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 2002.

[6]Find it here: https://github.com/ontodev/robot

[7]https://en.wikipedia.org/wiki/Aunt_Jemima

[Deng *et al.*, 2019] Yang Deng, Yuexiang Xie, Yaliang Li, Min Yang, Nan Du, Wei Fan, Kai Lei, and Ying Shen. Multi-task learning with multi-view attention for answer selection and knowledge base question answering. In *AAAI*, volume 33, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[Dooley and Griffiths, 2018] Damion M Dooley and Emma J et al. Griffiths. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *Science of Food*, 2(1), 2018.

[Feng *et al.*, 2020] Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. Genaug: Data augmentation for finetuning text generators. In *DeepIO*, 2020.

[Feng *et al.*, 2021] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. In *ACL-IJCNLP*, 2021.

[Grundkiewicz *et al.*, 2019] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *ACL BEA*, 2019.

[Gu *et al.*, 2021] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), 2021.

[Jin *et al.*, 2020] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *AAAI*, volume 34, 2020.

[Kenter and De Rijke, 2015] Tom Kenter and Maarten De Rijke. Short text similarity with word embeddings. In *CIKM*, 2015.

[Laskar *et al.*, 2020] Md Tahmid Rahman Laskar, Xiangji Huang, and Enamul Hoque. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *LREC*, pages 5505–5514, 2020.

[Lee *et al.*, 2020] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 2020.

[Liu *et al.*, 2020] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *AAAI*, volume 34, 2020.

[Longpre *et al.*, 2020] Shayne Longpre, Yu Wang, and Chris DuBois. How effective is task-agnostic data augmentation for pretrained transformers? In *EMNLP*, 2020.

[Marcus *et al.*, 1994] Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: Annotating predicate argument structure. In *Human Language Technology*, 1994.

[McCann *et al.*, 2017] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *NIPS*, 2017.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *NIPS*, 26, 2013.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 1995.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[Rezayi *et al.*, 2021] Saed Rezayi, Handong Zhao, Sungchul Kim, Ryan Rossi, Nedim Lipka, and Sheng Li. Edge: Enriching knowledge graph embeddings with external text. In *NAACL*, 2021.

[Rücklé *et al.*, 2019] Andreas Rücklé, Nafise Sadat Moosavi, and Iryna Gurevych. Coala: A neural coverage-based approach for long answer selection with small data. In *AAAI*, volume 33, 2019.

[Shi *et al.*, 2021] Haoyue Shi, Karen Livescu, and Kevin Gimpel. Substructure substitution: Structured data augmentation for nlp. In *ACL-IJCNLP*, 2021.

[Wei and Zou, 2019] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, 2019.

[Wei *et al.*, 2021] Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In *NAACL*, 2021.

[Wu *et al.*, 2020] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *EMNLP*, pages 6397–6407, 2020.

[Xia *et al.*, 2019] Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. Generalized data augmentation for low-resource translation. In *ACL*, 2019.

[Yang *et al.*, 2019] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with bertserini. In *NAACL*, 2019.

[Yih *et al.*, 2013] Scott Wen-tau Yih, Ming-Wei Chang, Chris Meek, and Andrzej Pastusiak. Question answering using enhanced lexical semantic models. In *ACL*, 2013.

[Zhao *et al.*, 2018] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*, 2018.