SciFoodNER: Food Named Entity Recognition for Scientific Text

Gjorgjina Cenikj[1], Gašper Petelin[1], Barbara Koroušić Seljak[1], and Tome Eftimov[1]

[1]Computer Systems Department, Jožef Stefan Institute, Ljubljana, Slovenia

February 20, 2023

Abstract

Named Entity Recognition (NER) and Named Entity Linking (NEL) are key Information Extraction tasks, tackling the identification and normalization of entity mentions from raw text. In the domain of food and nutrition, there have been several NER methods already developed, however, when applied to scientific text, they fail to generalize and produce large performance degradation. This introduces the need for new food NER and NEL models, developed specifically for extracting food entities from scientific text. In this paper, we present a scientific food NER and NEL model, SciFoodNER, obtained by fine-tuning transformer models on a corpus of scientific abstracts annotated with food entities. The models can identify mentions of food entites from raw text, and link the food entities to the Hansard Taxonomy, the FoodOn ontology and the Systematised Nomenclature of Medicine Clinical Terms (SNOMEDCT). Out of the evaluated models, the BioBERT model achieves the best results, reaching a median macro-averaged F1 score of 0.90 for the NER task, 0.66 for the NEL task linking to the Hansard Taxonomy, 0.43 for the NEL task linking to the FoodOn ontology and 0.58 for the NEL task linking to the SNOMEDCT ontology.

food, named entity recognition, named entity linking, information extraction

1 Introduction

The interactions of food with other health-related entities such as chemicals, diseases, treatments, and drugs are a commonly studied topic in food science. However, the sheer volume and the unstructured form of the produced scientific publications introduces challenges with following the latest discoveries. In order to alleviate the process of following newly published knowledge, Information Extraction (IE) methods are needed to extract meaningful findings from unstructured scientific text. Named Entity Recognition (NER) is one of the key subtasks of IE, which involves the identification of entities of a specific type in raw textual data. The related task of Named Entity Linking (NEL) addresses the linking to the extracted entities to concepts in existing semantic resources.

Existing NER methods can be broadly distinguished into four groups: *dictionary-based, rule-based, corpus-based methods* and *methods based on active learning* [1].

Dictionary-based methods [2] look for entities that have been defined in domain dictionaries, while rule-based systems [3–5] use dictionaries in combination with rules that describe the characteristics of the domain entities. The performance of these methods is largely dependant on the quality of the used dictionaries, since the methods can only extract entities which are present in the resources they rely on.

On the other hand, corpus-based methods [6, 7], train supervised machine learning models on data annotated for the presence of the entities in question. Methods based on active learning can be used in the event when large annotated corpora are not available. In this case, a domain expert is needed to interactively provide new annotations to the models to iteratively improve the performance.

In the domain of food and nutrition, until very recently, there existed only one corpus which could be used to train such models, namely, the FoodBase corpus [8], which contains recipe texts annotated with mentions of food entities. Although two corpus-based methods have already been trained on this corpus, as we show further on in this paper, these methods fail to generalize to scientific text.

For this reason, in this paper, we target the development of a food NER and NEL model for scientific text, accomplished by fine-tuning transformer-based representation models on a recently created corpus of scientific texts annotated with food entities. The BERT (Bidirectional Encoder Representations from Transformers) [9], RoBERTa (Robustly Optimized BERT Approach) [10], BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [9], and SciBERT [11] models are fine-tuned for the NER task, i.e. the recognition of food entities from raw text, as well as the NEL task, where the extracted entities are linked to the Hansard taxonomy [12], the FoodOn ontology and the Systematised Nomenclature of Medicine Clinical Terms (SNOMEDCT). Each task is treated separately, i.e. different models are trained for NER and each of the NEL tasks. The evaluation results show that BioBERT achieves the best results, with a median macro-averaged F1 score of 0.90 for the NER task, 0.66 for the NEL task linking the entities to the Hansard Taxonomy, 0.43 for the linking to the FoodOn ontology and 0.58 for the linking to the SNOMEDCT ontology.

The remainder of the paper is organized as follows. In Section 2, we introduce the knowledge bases to which the proposed NEL models can link the entities, as well as the existing NER and NEL models in the domain of food and nutrition. In Section 3, we evaluate the existing models on a scientific food NER corpus, which serve as baselines for the proposed NER models. In Section 4, we present our methodology, while the results are presented in Section 5. Section

6 concludes the paper.

Reproducibility: The code is available on Github [1]. The trained models are available at [2].

2 Related Work

2.1 Knowledge Bases in the Food Domain

In this subsection, we introduce the Knowledge Bases in the food domain, to which the proposed method can link the extracted food entities.

FoodOn [13] is a food ontology developed with the aim of solving the issues of incompatibility and ambiguity of food references by acting as a hub that links specialized ontologies. It provides schemes for food and food product categorization and covers animal and plant food sources, in addition to terms related to packaging, cooking and preservation.

The Hansard taxonomy [12] is a hierarchical organization of more than 8,000 different semantic categories, where *Food and Drink* is one of the top-level categories.

SNOMEDCT [14] is a healthcare terminology used for the representation of electronic healthcare records. Apart from clinical concepts, such as body structures, organisms, substances, pharmaceutical products, specimens, symptoms, drugs, and clinical procedures, it also contains food entities.

2.2 Named Entity Recognition in the Food Domain

In the domain of food and nutrition, several NER methods have already been proposed.

2.2.1 Rule-Based Named Entity Recognition

Due to the lack of annotated data required to train corpus-based methods, the first food NER methods were rule-based [4,5] or were using semantic information from various resources [15, 16]. DrNER [4] is a rule-based NER, which is capable of extracting food entities from evidence-based dietary recommendations, among other entities of interest. However, drNER extracts several food entities as one. This was improved by developing the rule-based NER FoodIE [5], where the rules incorporate computational linguistics information in combination with food semantic annotations from the Hansard taxonomy [12]. FoodIE achieved promising results on different independent evaluation datasets and has been used to create the FoodBase corpus [8], which is the first NER corpus in the food domain. The main weakness of the FoodIE method is its dependency on external resources, which have become inaccessible in the time after its publication, rendering the method currently unusable.

[1] https://github.com/gjorgjinac/SciFoodNER

[2] https://portal.ijs.si/nextcloud/s/C3jCDq84TBoE8gY

2.2.2 Corpus-Based Named Entity Recognition

The FoodBase corpus is annotated for the existence of food entities and their corresponding Hansard [12] semantic tags. The gold standard of the FoodBase corpus consists of 1,000 recipes from 5 recipe categories.

The annotation of the FoodBase corpus enabled the development of two corpus-based NER methods for the food domain. BuTTER [1] is the first corpus-based food NER method, which leverages a neural network architecture based on Bidirectional Long Short-Term Memory and Conditional Random Fields. The word embeddings GloVe [17], Word2Vec [18] and FastText [19] are used to represent the recipe data from the FoodBase corpus, on which the neural network architecture is trained.

The second corpus-based NER method for the food domain, FoodNER [20], performs finetuning of BERT [21] and BioBERT [9] for the NER task using the annotations from the FoodBase corpus. Apart from identifying the mentions of food entities from the text, the FoodNER model is also capable of performing NEL, i.e. linking the entity mentions to concepts in the Hansard taxonomy, and the FoodOn and SNOMEDCT ontologies.

When evaluated using cross-fold validation on the FoodBase corpus, BuTTER and FoodNER achieve comparable results, with a macro averaged F1 score of 0.94. However, since these models are trained on recipe texts, their application on scientific text introduces challenges due to the domain-specific vocabulary and the difference in writing style.

3 The Need For A Scientific Food Named Entity Recognition Model

For a long time, the generalization of the BuTTER and FoodNER models could not be evaluated due to a lack of annotated scientific corpus with food entities. In order to demonstrate the need for a food NER model developed specifically for scientific text, we use a recently annotated scientific corpus with food entities in order to evaluate the generalization of BuTTER, FoodNER and two dictionary-based NER models using dictionaries from the FooDB (`http://foodb.ca/`) database.

3.1 Dictionary-based Baselines

We propose two dictionary-based models, which we refer to the FooDB non-scientific and FooDB scientific models. The FooDB non-scientific model performs simple string matching using the common names of 992 foods from the FooDB database. Similarly, the FooDB scientific performs string matching using the 675 scientific names of foods available in the FooDB database, due to the fact that biomedical texts commonly refer to food entities using their scientific names which are not extracted by the other models.

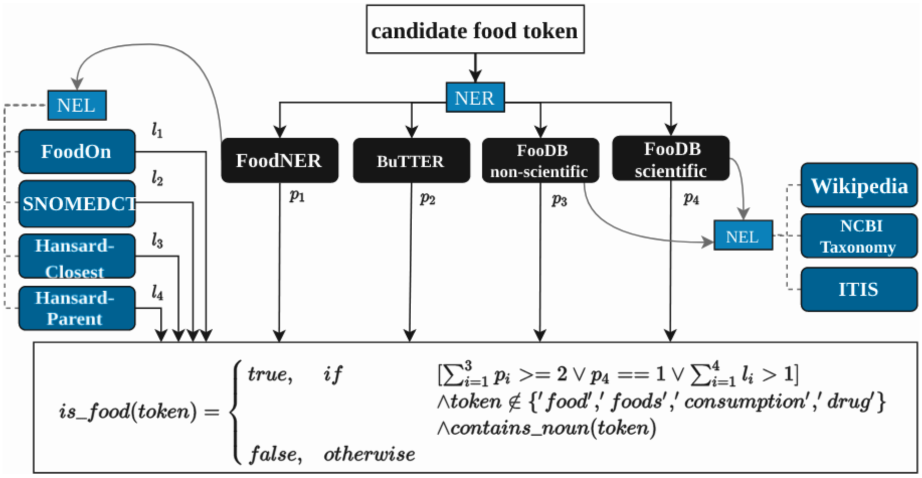

candidate food token
NER
NEL
FoodOn
SNOMEDCT
Hansard-Closest
Hansard-Parent
l_1
l_2
l_3
l_4
FoodNER
BuTTER
FooDB non-scientific
FooDB scientific
p_1
p_2
p_3
p_4
NEL
Wikipedia
NCBI Taxonomy
ITIS
$is\_food(token) = \begin{cases} true, & if \quad [\sum_{i=1}^{3} p_i >= 2 \vee p_4 == 1 \vee \sum_{i=1}^{4} l_i > 1] \\ & \wedge token \notin \{'food','foods','consumption','drug'\} \\ & \wedge contains\_noun(token) \\ false, & otherwise \end{cases}$

Figure 1: Voting scheme for the food NER task.

3.2 Corpus-based Baselines

We use the FoodViz platform [22] to apply the FoodNER models based on the original BERT model, to both extract the food entities and link them to Hansard, FoodOn and SNOMEDCT. We use the lexical lemmatized BuTTER model without character embeddings, which achieved the best results in terms of the averaged macro F1 score in [1].

3.3 Ensemble-based Baseline

In order to benefit from both dictionary-based and corpus-based models, we propose an ensemble-based baseline which combines the annotations of the previously proposed baselines. We refer to this baseline as the Food Voting Scheme. A voting strategy is used to combine the entities extracted using BuTTER, FoodNER and FooDB non-scientific. A token extracted by any of these models is considered to be a valid food entity if at least two of the models nominate the exact same token without any missing or additional words, or if the FoodNER method is able to link it to an external resource. The tokens extracted by the FooDB scientific method are always considered to be valid entities, since no other models are capable of identifying foods using their scientific names. This voting scheme is visually depicted in Figure 1, where p_1, p_2, p_3 and p_4 are binary indicators of whether each of the food NER annotators (FoodNER, BuTTER, FooDB non-scientific and FooDB scientific) extracted the particular token as a food entity, while l_1, l_2, l_3 and l_4 are binary indicators of whether FoodNER managed to link the token to a concept in FoodOn, SNOMEDCT ontologies, or Hansard (using the *Hansard-Closest* or *Hansard-Parent* strategies).

Finally, we perform post-processing, which involves removing a few food-related words (*food, foods, consumption*, and *drug*) that are too general to be

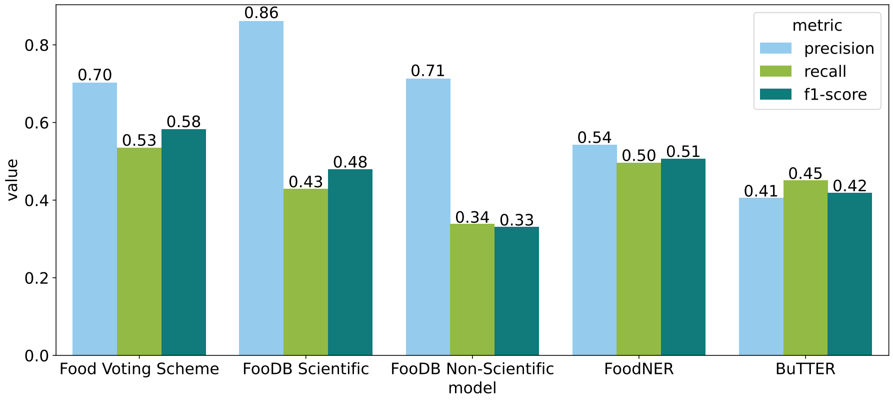


Figure 2: Macro-averaged precision, recall and F1-score for each of the food NER models, evaluated on scientific text

useful, and entities that do not contain any nouns, since these are likely to be false positives.

3.4 Baseline Evaluation on Scientific Text

Each of the baseline models is applied on the 500 annotated scientific abstracts. To be consistent with the evaluation of the BuTTER and FoodNER models, the annotations are then represented into the Inside-Outside-Beginning (IOB) format, and the macro-averaged precision, recall and F1-score are reported in Figure 2. As can be seen from the figure, the BuTTER and FoodNER models experience a large performance degradation when evaluated on scientific text. In particular, the performance of the BuTTER model, in terms of the macro-averaged F1 score drops from 0.94 to 0.42, while the performance of the FoodNER model drops from 0.94 to 0.51. In terms of precision, the FooDB models give the best results, however, they also have the lowest recall scores, since they only perform simple string matching and can only identify foods if they are present in their dictionaries. In terms of recall, the Food Voting Scheme achieves the best results, followed by the FoodNER model. The best performance in terms of the F1-score is achieved by the Food Voting Scheme, which achieves a score of 0.58. This shows that even an ensemble model which combines the previously developed NER models achieves a relatively unsatisfactory performance when applied on scientific text, introducing the need for a novel corpus-based food NER method, intended specifically for the extraction of entities from scientific text.

4 Methodology

4.1 Data

The Food Voting Scheme introduced in 3.3 is applied on abstracts of scientific papers extracted from PubMed. Out of all of the automatically annotated abstracts, 500 were selected, such that each one contains at least eight entities which were automatically extracted. These abstracts were then manually corrected by several human annotators and linked to their appropriate semantic classes from the Hansard taxonomy. The annotated entities have then been automatically linked to the FoodOn and SNOMEDCT ontologies, using The National Center for Biomedical Ontology Annotator [15]. The manually annotated food entity phrases were inputted to the annotator and the tags obtained by the annotator were joined together to form a tag for the entire food entity. Since the tags for the FoodOn and SNOMEDCT ontologies were obtained in an automatic manner, human inspection would be required to check the annotations. The scientific corpus has been created as part of an European Food Safety Authority-funded project called CAFETERIA.

4.2 Text Representation Models

In order to generate vectorized representation of the textual data, we use four pre-trained transformed-based models: BERT [9], RoBERTa [10], BioBERT [9], and SciBERT [11] which achieve state-of-the-art results in Natural Language Processing. These are bidirectional, contextual representation models, which are pre-trained on large amounts of textual data, in an unsupervised manner, on the Mask Language Modeling or Next Sentence Prediction tasks. This kind of pre-training allows the models to be fine-tuned for other downstream tasks, such as NER, Natural Language Inference or Question Answering, with smaller amounts of data, and without substantial modifications in the original architecture. In the simplest case, only the output layer needs to be replaced, depending on the task that the model is intended to perform.

In our experiments, we use the original BERT model, which is pre-trained on the BooksCorpus [23] and English Wikipedia.

The RoBERTa model introduces some improvements to the BERT model, such as longer pre-training, training on a larger amount of data, and the use of dynamic masking. Apart from the data used for the pre-training of BERT model, RoBERTa is trained on data from 3 additional sources: the CommonCrawl News dataset [24], the OpenWebText corpus [25] and the Common Crawl Stories dataset [26].

BioBERT is a version of the BERT model intended to be used specifically for the domain of biomedicine. Since biomedical text contains domain-specific terms that do not appear in more general texts, texts from PubMed abstracts and full-text articles from PubMed Central are used in the pre-training of the BioBERT model. As a result, BioBERT has been shown to outperform BERT in biomedical NER, Relation Extraction and Question-Answering [9].

SciBERT makes use of the same architecture as BERT while being trained on scientific text corpus from Semantic Scholar [27]. Corpus consists of 1.14M full papers and not just abstracts. The SciBERT vocabulary is more specialized for scientific text and differs from BERT vocabulary with 42% token overlap. SciBERT thus outperforms BERT on multi-domain tasks, and in some cases, outperforms BioBERT in biomedical tasks [11].

4.3 Fine-tuning the Models for Named Entity Recognition

The NER task is addressed as a token classification problem, where the classes are the tags from the Inside-Outside-Beginning (IOB) tagging scheme. This means that for each word in the text, the model is supposed to determine whether the word is at the beginning (class B), inside (class I) or outside (class O) of a food entity. Therefore, the fine-tuning of the transformer models is performed by adding a linear layer on top of the hidden-states outputs of the original architecture which performs token classification. During the fine-tuning process, the model parameters are initialized with the values from the pre-training step, and are fine-tuned for the NER task.

4.4 Training Tasks

- NER - In the standard NER task, the models are trained to only identify the mentions of food entities, regardless of their semantic tag, i.e. the multi-class classification is using three classes, ("I", "O" and "B").

- NEL - Hansard - In this task, the Hansard semantic tags of the food entities is supposed to be determined. There are 70 Hansard semantic categories to which the entities can be mapped, and one additional class "X", which represents a term related to the food entity, which is not a food entity in itself. Some of the words that are annotated with this class include: *intake*, *consumption*, *use*, and *dietary.* Combining the semantic categories with the prefixes "B-" and "I-", indicating if the token is at the beginning or inside the phrase matching the category, and including the additional "O" and "X" classes, the multi-class classification is done with a total of 140 classes. Note that since the some of the semantic tags are described only with one token (there are no "I-" prefixes, only "B-") the number of unique classes is not necessarily equal to 2∗`number of semantic tags`+2.

- NEL - FoodOn - In this task, the food entities can be linked to 618 concepts from the FoodOn ontology, and the multi-class classification is done with a total of 1,078 classes.

- NEL - SNOMEDCT - In this task, the entities can be linked to 349 concepts from the SNOMEDCT ontology, and there are a total of 607 classes.

5 Results

5.1 Experimental Setup

The models are trained using the AdamW [28] optimizer with a learning rate of $4 * 10^{-5}$. The models are allowed to train for a maximum of 100 epochs, however, to prevent overfitting, an early stopping strategy is used, meaning that the training is interrupted once the decrease in the validation loss does not exceed the threshold of $5 * 10^{-3}$ for five consecutive epochs. Most of the models reach the early stopping criteria in five to ten epochs. The models are evaluated using 10-fold cross validation, where the splits are done on a sentence level, i.e. 10% of the total sentences are in each test fold. The remaining 90% of the sentences are further split into training (81%) and validation sets (9%).

5.2 Experimental Evaluation

For each model, we report the macro-averaged precision, recall and F1-score, averaged across all of the classes.

Figure 3 features the results for the NER task. As can be seen, the BioBERT model achieves the best performance, with a median macro-F1 score of 0.90 (median across 10 folds). The rest of the models have a similar performance, with median macro-F1 scores of 0.89. The performance is generally stable across the 10 folds. Comparing the proposed models to the baselines introduced in section 3, we can see that all proposed models outperform the best-performing baseline (the Food Voting Scheme) by 0.31 in terms of the macro F1-score. It should be noted, however, that the macro F1-score of the baseline models is evaluated based on a single run on the entire scientific corpus, while the F1-score of the proposed models is calculated by taking the median of the macro F1-scores obtained on each test fold of the cross-fold validation.

The results for the NEL task, linking the food entities to the Hansard taxonomy, are presented in Figure 4. In this case, the best models, in terms of the median macro-F1 score, are the RoBERTa and BioBERT models, with a score of 0.66. Next is the BERT model with a score of 0.65, and finally, the SciBERT with a score of 0.64.

Figure 5 depicts the evaluation results for the FoodOn NEL task. BioBERT again achieves the best median macro-F1 score of 0.43. The BERT and RoBERTa models have scores of 0.42, while the SciBERT model achieves a score of 0.41. In this case, however, the models' performance is somewhat less stable across the ten folds, with one run obtaining macro-F1 scores in the range 0.55-0.60.

Figure 6 contains the results for the NEL task of linking to the SNOMEDCT ontology. The BioBERT model has the highest median macro-F1 score of 0.58, followed by RoBERTa with a score of 0.56, and BERT and SciBERT, both with a score of 0.54. It can also be noted that the results of the FoodOn and SNOMEDCT NEL models are somewhat worse than the Hansard NEL models, which is due to the larger number of classes used in the FoodOn and SNOMEDCT linking and the automatically performed annotation.

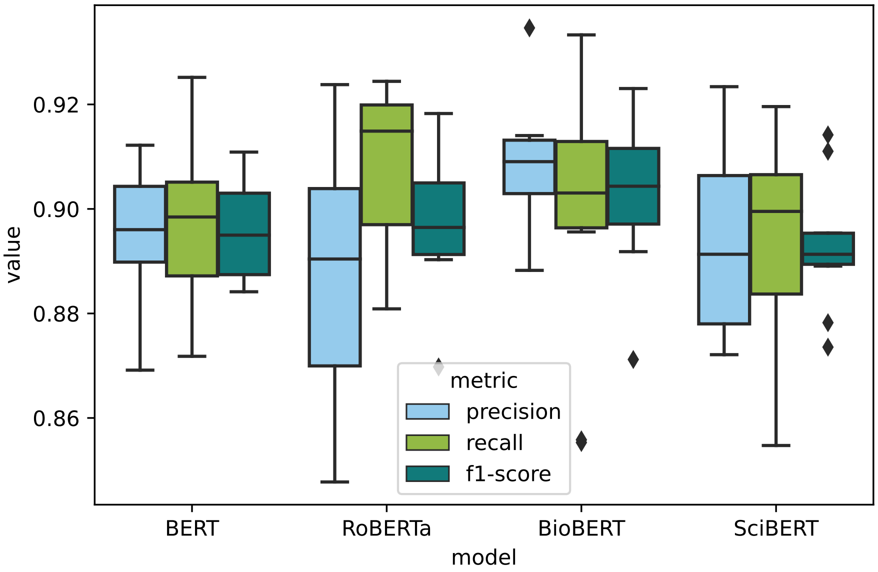


Figure 3: Macro-averaged precision, recall and F1-score for each of the food NER models

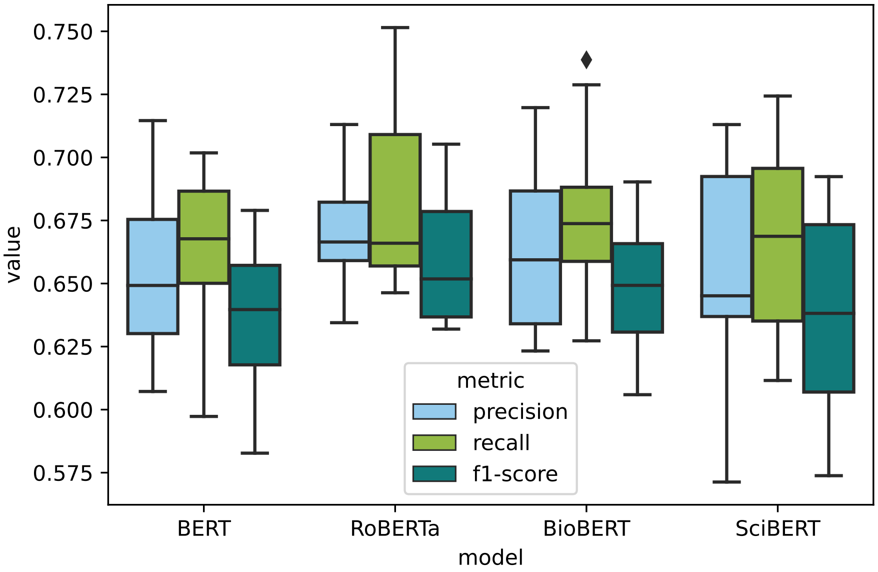


Figure 4: Macro-averaged precision, recall and F1-score for each of the NEL models linking to the Hansard taxonomy

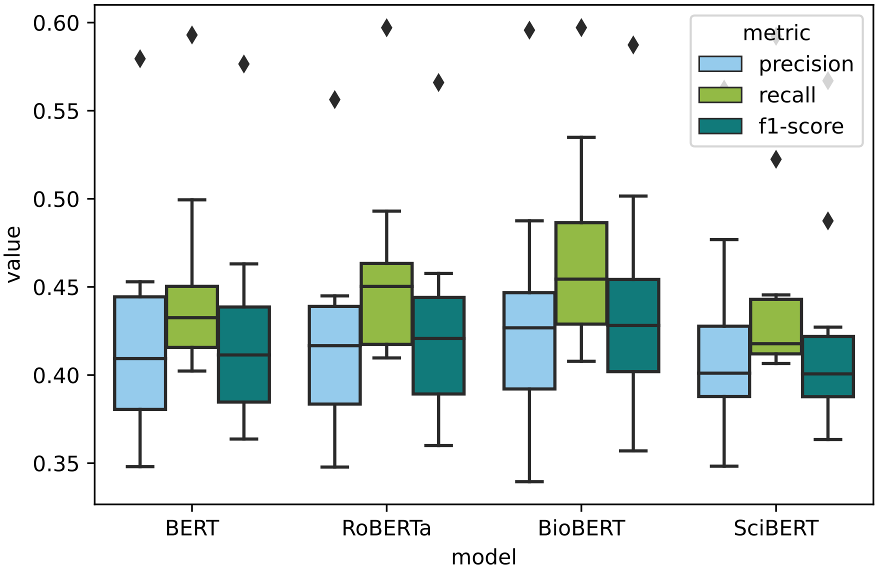


Figure 5: Macro-averaged precision, recall and F1-score for each of the NEL models linking to the FoodOn ontology

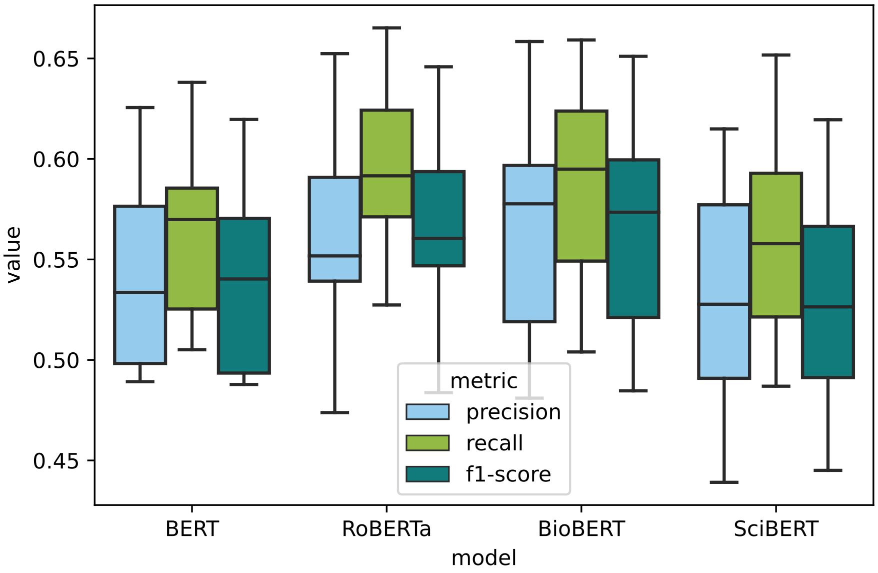


Figure 6: Macro-averaged precision, recall and F1-score for each of the NEL models linking to the SNOMEDCT ontology

Going across the results of the models for all four training tasks, we can observe that all of the models achieve similar results, however, there is a consistent ordering of the models in terms of the median macro F1-score. BioBERT achieves the best results, then RoBERTa, and finally, BERT and SciBERT. The performance of the models for each task is proportional to the number of classes, meaning that the models for the NER task using three classes reach a performance of 0.90, the best model for the Hansard NEL task has a score of 0.66 with 140 classes, the best model for the SNOMEDCT NEL task has a score of 0.58 with 607 classes, while the best model for the FoodOn NEL task has a score of 0.43 with 1,078 classes.

5.3 Per-Class Analysis

In order to identify the semantic tags which the models have difficulty identifying, we analyze the results obtained for each semantic tag. Due to the large amount of tags in the FoodOn and SNOMEDCT ontologies and the difficulties of visualizing all of them, we only present the results for the NEL to the Hansard Taxonomy. For each model, we take into account the medians of the F1-scores of the semantic tags obtained in each fold, regardless of the prefix ”B-” or ”I-”, indicating that the token is at the beginning or inside the semantic tag. This means that for each semantic tag consisting of one token, the median is calculated over 10 values (scores for the class with ”B-” prefix obtained in each of the 10 runs), while for each semantic tag consisting of two or more tokens, the median is calculated over 20 values (scores for the class with ”B-” and ”I-” prefixes obtained in each of the 10 runs).

Figure 7 features the results for each semantic tag from the Hansard taxonomy. We can see that there are several semantic tags for which all of the models have a median F1-score of zero. This is the case for the tags *Manufacture of alcoholic drink*, *Part/joint of animal*, *Mutton*, *Stalk vegetables*, *Puddings*, *Meat dishes*, *Fish dishes*, *Prepared fruit and dishes*, *Pancake/tortilla/oatcake*, and *Egg dishes*.

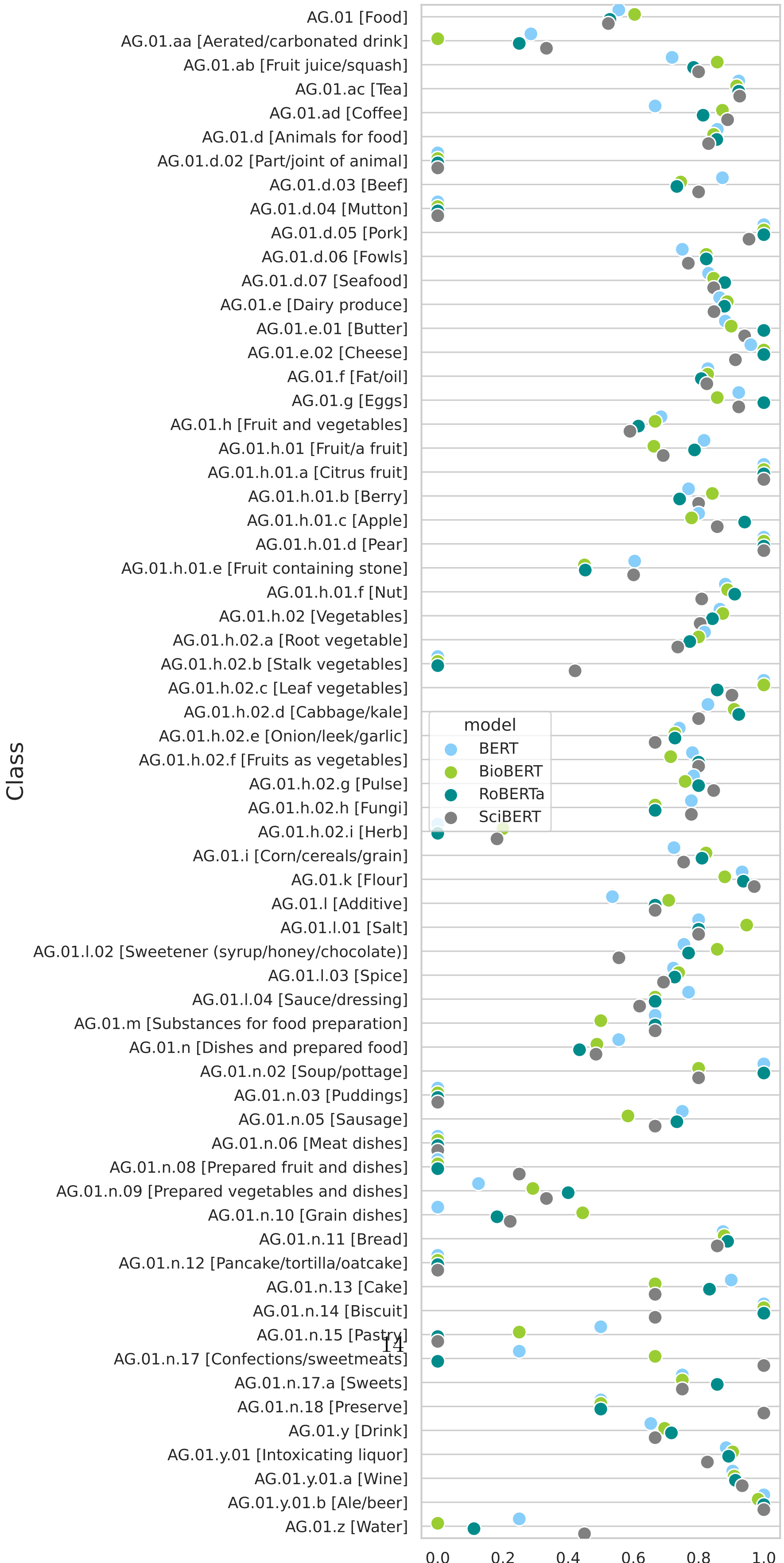
On the other hand, some semantic tags have median F1-score of one. Such is the case with the tags *Ale/beer*, *Ices*, *Biscuit*, *Soup/pottage*, *Pear*, *Citrus fruit*, and *Pork*. Looking at the terms that are linked to these categories, we can observe that the tags which the models have most difficulties identifying are groups of prepared dishes, since these can contain a variety of food items that are not necessarily similar lexically. For instance, the semantic tag *Meat dishes* is used to link the following food items: *bolognese sauce*, *bake-only chicken nuggets*, *grilled beef*, *hamburgers*, *paté or high - fat meat*. The food items *tortillas* and *mung bean pancake* are both linked to the tag *Pancake/tortilla/oatcake*, while the food items *dried plums*, *raisins*, *applesauce*, *apple chips*, *dried fruit*, *dried vine fruits*, *fruit purée*, *dried plum powder* are all linked to the tag *Prepared fruit and dishes*.

On the other hand, the food items linked to semantic tags that are easier to identify almost always contain at least one common word. For instance, all of the food items linked to the tag *Ale/beer* contain the word *beer*, the items linked to the tag *Ices* contain the word *ice*, while the items linked to the tag *Biscuit* contain either the word *biscuit* or *cookie*. Another characteristic of the tags that are easy to identify is that they describe a more specific type of food. Take for instance, the tags *Pear* and *Citrus fruit* which refer to a specific fruit type, as opposed to the class *Prepared fruit and dishes* which can refer to any type of fruit product.

A more detailed breakdown of classification errors can be seen in figures 8, 9, and 10 where we show 10 tags that achieved the worst F1-score and how they were classified. Note that some predicted classes where no misclassification occurred are removed form the plot. As can be seen from Figure 8, for the NEL to the Hansard Taxonomy, most of the mistakes occur due to the entities being linked to a semantic tag that is a parent of the true tag in the taxonomy. Such is the case with the entities belonging to the tags *Part/joint of animal* and *Mutton* being linked to the parent tags *Animals for food*, and the tags *Puddings*, *Pancake/tortilla/oatcake*, *Meat dishes*, *Grain dishes*, *Pastry*, *Prepared vegetables and dishes* being linked to their parent tag *Dishes and prepared food*. A similar mistake is the incorrect linking to a sibling tag in the taxonomy, such as in the case when *Drink* is predicted instead of *Aerated/carbonated drink*. Finally, some tags are mistaken due to some food entities being an ingredient of other food entities, like in the case when *Corn/cereals/grain* is predicted instead of *Grain dishes*.

In Figure 9, we can observe the classification errors for the linking to the FoodOn ontology. An important note to be made here is that in the scientific corpus, in the case when a suitable match for a food entity is not found in the FoodOn ontology, several of the original semantic tags are merged into a new tag, which we refer to as a complex tag. As we can see in Figure 9, most of the incorrectly classified tags are in fact complex tags. Consequently, most of the complex tags are misclassified into their component tags. This means that the models identify partial matches of the food entities and link them to the original component tags, instead of detecting the entire food entity and linking it to the complex tag. Such is the case with the tags *water* and *spinach* are predicted separately instead of the complex tag *water spinach*, the tags *fat*, *sugar* and *salt* being predicted instead of the tag *products rich in salt, fat and sugar*. The same holds for the tags *sugar cane*, *cartenoid-rich fruits and vegetables*, *smoked, pickled or salted food*, *salt water*, *foods preserved in salt*, *products rich in salt and fat*, and *products rich in salt, fat and sugar*.

For the linking to the SNOMEDCT ontology, the mistakes are similar to those of the FoodOn ontology, i.e. the complex tags are mapped to their component tags. Some examples include the entity *chocolate ice cream* classified as *chocolate* and *ice cream*, *chocolate drink* being classified as *chocolate* and *drink*, and *strawberry jam* being classified as *strawberries* and *jam*. This is due to the fact that the ontology does not contain such fine-grained distinctions of food entities. In such cases, the model does capture the ”primary” class, i.e. in the case of *strawberry jam* the model can correctly identify that the entity related to some type of jam. However, since the entire concept match is not present in the ontology, the automatic annotator would create a new artificial class by


Class
AG.01 [Food]
AG.01.aa [Aerated/carbonated drink]
AG.01.ab [Fruit juice/squash]
AG.01.ac [Tea]
AG.01.ad [Coffee]
AG.01.d [Animals for food]
AG.01.d.02 [Part/joint of animal]
AG.01.d.03 [Beef]
AG.01.d.04 [Mutton]
AG.01.d.05 [Pork]
AG.01.d.06 [Fowls]
AG.01.d.07 [Seafood]
AG.01.e [Dairy produce]
AG.01.e.01 [Butter]
AG.01.e.02 [Cheese]
AG.01.f [Fat/oil]
AG.01.g [Eggs]
AG.01.h [Fruit and vegetables]
AG.01.h.01 [Fruit/a fruit]
AG.01.h.01.a [Citrus fruit]
AG.01.h.01.b [Berry]
AG.01.h.01.c [Apple]
AG.01.h.01.d [Pear]
AG.01.h.01.e [Fruit containing stone]
AG.01.h.01.f [Nut]
AG.01.h.02 [Vegetables]
AG.01.h.02.a [Root vegetable]
AG.01.h.02.b [Stalk vegetables]
AG.01.h.02.c [Leaf vegetables]
AG.01.h.02.d [Cabbage/kale]
AG.01.h.02.e [Onion/leek/garlic]
AG.01.h.02.f [Fruits as vegetables]
AG.01.h.02.g [Pulse]
AG.01.h.02.h [Fungi]
AG.01.h.02.i [Herb]
AG.01.i [Corn/cereals/grain]
AG.01.k [Flour]
AG.01.l [Additive]
AG.01.l.01 [Salt]
AG.01.l.02 [Sweetener (syrup/honey/chocolate)]
AG.01.l.03 [Spice]
AG.01.l.04 [Sauce/dressing]
AG.01.m [Substances for food preparation]
AG.01.n [Dishes and prepared food]
AG.01.n.02 [Soup/pottage]
AG.01.n.03 [Puddings]
AG.01.n.05 [Sausage]
AG.01.n.06 [Meat dishes]
AG.01.n.08 [Prepared fruit and dishes]
AG.01.n.09 [Prepared vegetables and dishes]
AG.01.n.10 [Grain dishes]
AG.01.n.11 [Bread]
AG.01.n.12 [Pancake/tortilla/oatcake]
AG.01.n.13 [Cake]
AG.01.n.14 [Biscuit]
AG.01.n.15 [Pastry]
AG.01.n.17 [Confections/sweetmeats]
AG.01.n.17.a [Sweets]
AG.01.n.18 [Preserve]
AG.01.y [Drink]
AG.01.y.01 [Intoxicating liquor]
AG.01.y.01.a [Wine]
AG.01.y.01.b [Ale/beer]
AG.01.z [Water]
model
● BERT
● BioBERT
● RoBERTa
● SciBERT
0.0
0.2
0.4
0.6
0.8
1.0

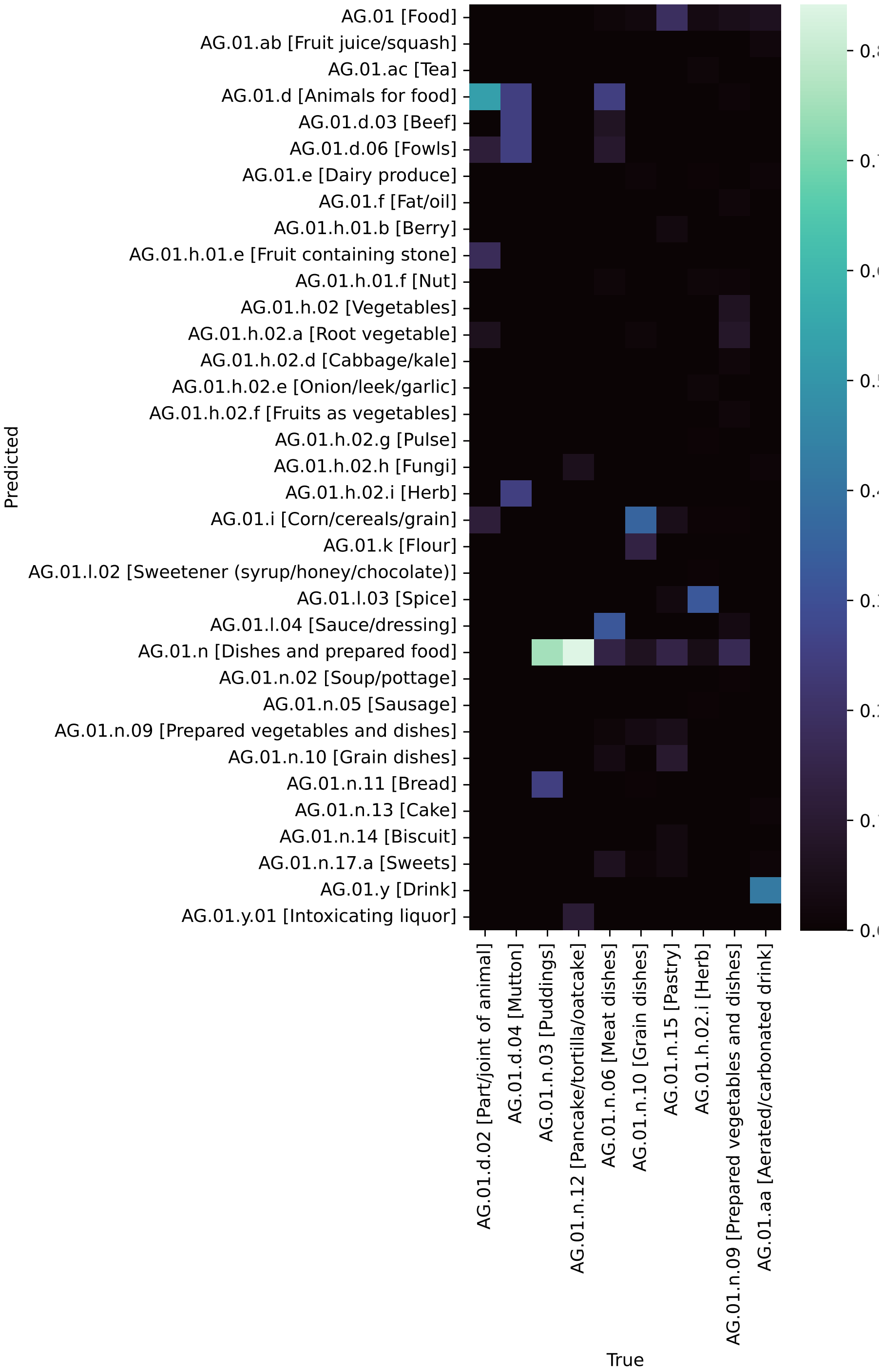


Figure 8: Confusion matrix for different 10 Hansard semantic tags with the lowest F1 score averaged across all the models.

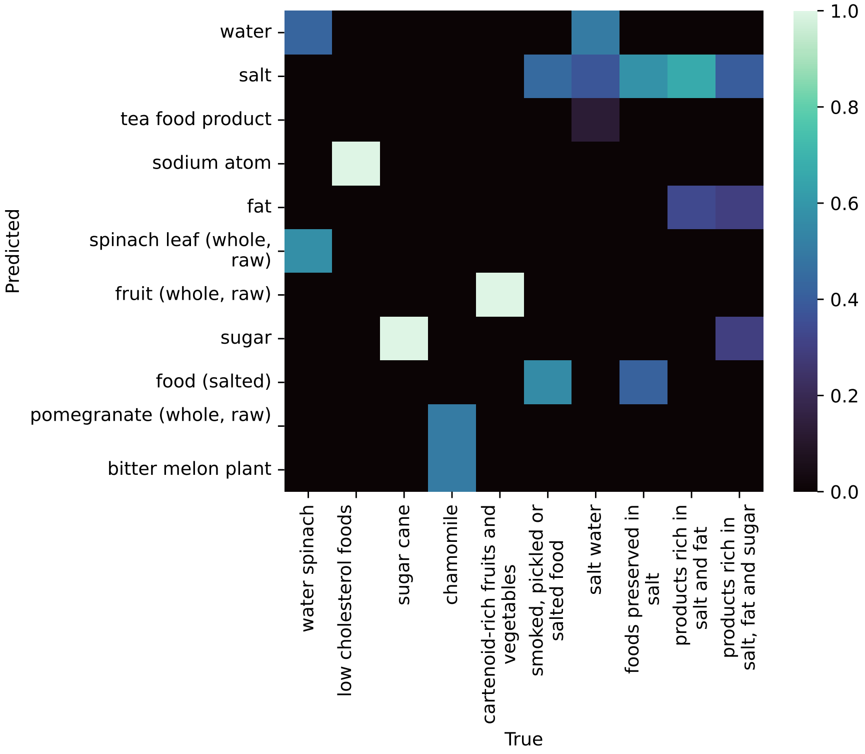


Figure 9: Confusion matrix for different 10 FoodOn semantic tags with the lowest F1 score averaged across all the models.

merging the concepts *strawberry* and *jam* into *strawberry jam*, and when the model predicts them separately, this would be considered incorrect in the evaluation. Other mistakes include the mapping to concepts which are similar but not exactly correct, such as the mapping of the entity *coffee with milk* to *coffee*, *coffee sample without cream* and *tea with milk*.

6 Conclusion

In this paper, we present a corpus-based Named Entity Recognition (NER) model for extracting food entities from scientific text, as well as Named Entity Linking (NEL) models for linking the food entities to concepts in the Hansard Taxonomy, the FoodOn ontology and the Systematised Nomenclature of Medicine Clinical Terms (SNOMEDCT). To accomplish this, we fine-tune the transformer-based models BERT, BioBERT, SciBERT and RoBERTa on a corpus of scientific abstracts annotated with food entities. The models have a similar performance for each of the tasks, with the BioBERT model consistently achieving the best results, which are usually only 0.01 higher than the rest of the models. For the NER task, the best BioBERT model reaches a median macro-averaged F1 score of 0.89. For the NEL task, linking the entities to the Hansard Taxonomy, and the FoodOn and SNOMEDCT ontologies, the best model achieves scores of 0.66, 0.43, and 0.58, respectively.

Acknowledgments

This work has been supported by the Ad Futura grant for postgraduate study; the Slovenian Research Agency [research core funding programmes P2-0098]; the European Union's Horizon 2020 research and innovation programme [grant agreement 863059] (FNS-Cloud, Food Nutrition Security) and [grant agreement 101005259] (COMFOCUS); and the EFSA-funded project under [grant agreement GP/EFSA/AMU/2020/03/LOT2] (CAFETERIA).

References

[1] G. Cenikj, G. Popovski, R. Stojanov, B. Koroušić Seljak, and T. Eftimov, "BuTTER: Bidirectional LSTM for Food Named-Entity Recognition," pp. 3550–3556, 2020.

[2] X. Zhou, X. Zhang, and X. Hu, "Maxmatcher: Biological concept extraction using approximate dictionary lookup," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2006, pp. 1145–1149.

[3] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: rule-based protein and gene entity recognition," *BMC bioinformatics*, vol. 6, no. 1, p. S14, 2005.

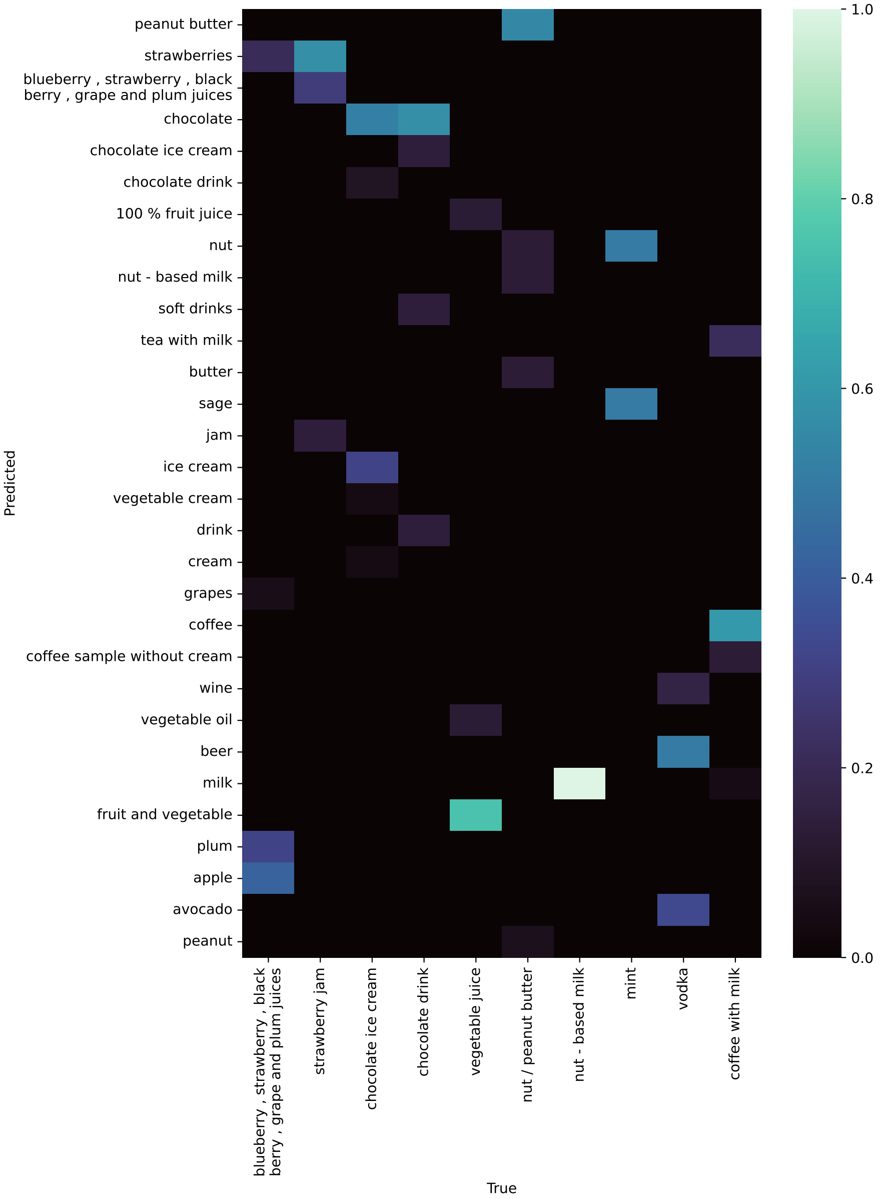


Figure 10: Confusion matrix for different 10 SNOMEDCT semantic tags with the lowest F1 score averaged across all the models.

[4] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS One*, vol. 12, no. 6, p. e0179488, 2017.

[5] G. Popovski, S. Kochev, B. K. Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction," in *ICPRAM*, 2019.

[6] N. Alnazzawi, P. Thompson, R. Batista-Navarro, and S. Ananiadou, "Using text mining techniques to extract phenotypic information from the phenochf corpus," *BMC medical informatics and decision making*, vol. 15, no. 2, p. 1, 2015.

[7] R. Leaman, C.-H. Wei, C. Zou, and Z. Lu, "Mining patents with tmchem, gnormplus and an ensemble of open systems," in *Proce. The fifth BioCreative challenge evaluation workshop*, 2015, pp. 140–146.

[8] G. Popovski, B. K. Seljak, and T. Eftimov, "Foodbase corpus: a new resource of annotated food entities," *Database*, vol. 2019, 2019.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 09 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[11] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://www.aclweb.org/anthology/D19-1371

[12] M. Alexander and J. Anderson, "The hansard corpus, 1803-2003," Glasgow, UK, Other, 2012. [Online]. Available: http://eprints.gla.ac.uk/81804/

[13] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. L. Brinkman, and W. W. L. Hsiao, "Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration," *npj Science of Food*, vol. 2, no. 1, p. 23, 12 2018. [Online]. Available: https://doi.org/10.1038/s41538-018-0032-6

[14] K. Donnelly, “Snomed-ct: The advanced terminology and coding system for ehealth.” *Studies in health technology and informatics*, vol. 121, pp. 279–90, 2006. [Online]. Available: https://app.dimensions.ai/details/publication/pub.1077321040

[15] C. Jonquet, N. Shah, C. Youn, C. Callendar, M.-A. Storey, and M. Musen, “Ncbo annotator: semantic annotation of biomedical data,” in *International Semantic Web Conference, Poster and Demo session*, vol. 110, 2009.

[16] G. Popovski, B. K. Seljak, and T. Eftimov, “A survey of named-entity recognition methods for food information extraction,” *IEEE Access*, vol. 8, pp. 31 586–31 594, 2020.

[17] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[19] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *CoRR*, vol. abs/1607.04606, 2016.

[20] R. Stojanov, G. Popovski, G. Cenikj, B. K. Seljak, and T. Eftimov, “A Fine-Tuned Bidirectional Encoder Representations From Transformers Model for Food Named-Entity Recognition: Algorithm Development and Validation, journal = Journal of Medical Internet Research,” vol. 23, no. 8, p. e28229, Aug. 2021. [Online]. Available: https://doi.org/10.2196/28229

[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[22] R. Stojanov, G. Popovski, N. Jofce, D. Trajanov, B. K. Seljak, and T. Eftimov, “Foodviz: Visualization of food entities linked across different standards,” in *Machine Learning, Optimization, and Data Science*, G. Nicosia, V. Ojha, E. La Malfa, G. Jansen, V. Sciacca, P. Pardalos, G. Giuffrida, and R. Umeton, Eds. Cham: Springer International Publishing, 2020, pp. 28–38.

[23] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 19–27.

[24] S. Nagel, "Cc news," http://commoncrawl.org/2016/10/news-dataset-available/, 2016, accessed: 2021-03-10.

[25] A. Gokaslan and V. Cohen, "Openwebtext corpus," http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[26] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," *CoRR*, vol. abs/1806.02847, 2018. [Online]. Available: http://arxiv.org/abs/1806.02847

[27] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha *et al.*, "Construction of the literature graph in semantic scholar," *arXiv preprint arXiv:1805.02262*, 2018.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.