

Received January 31, 2020, accepted February 8, 2020, date of publication February 12, 2020, date of current version February 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973502

A Survey of Named-Entity Recognition Methods for Food Information Extraction

GORJAN POPOVSKI^{1,2}, BARBARA KOROUŠIĆ SELJAK², AND TOME EFTIMOV^{1,2}

¹Jožef Stefan International Postgraduate School, Jožef Stefan Institute, 1000 Ljubljana, Slovenia

²Computer Systems Department, Jožef Stefan Institute, 1000 Ljubljana, Slovenia

Corresponding author: Gorjan Popovski (gorjan.popovski@ijs.si)

This work was supported in part by the Ad Futura Grant for postgraduate study, in part by the Slovenian Research Agency Program under Grant P2-0098, in part by the European Union's Horizon 2020 Research and Innovation Programme SAAM under Grant 769661, and in part by the European Union's Horizon 2020 Research and Innovation Programme FNS-Cloud under Grant 863059.

ABSTRACT As great amounts of food-related information is presented in the form of heterogeneous textual data, computer-based methods are useful to automatically extract such information. One way to do this is to utilize Named-Entity Recognition (NER) methods that are broadly used in computer science for information extraction. Despite the existence of numerous and well-versed NER methods in the biomedical domain, the domain of food science still remains scarcely resourced. In this paper, we provide an overview and a comparison of named-entity recognition methods in the food domain, which can be used for automated extraction of food information from text. Four methods are discussed: FoodIE, NCBO (SNOMED CT), NCBO (OntoFood), and NCBO (FoodON). We compare them using a benchmark data set that consists of 1000 manually annotated recipes initially obtained from Allrecipes, which is the largest social network focused on food. After analysing the results from the evaluation, it is evident that FoodIE obtains very promising results compared to the other food named-entity recognition methods taken into consideration.

INDEX TERMS Benchmarking, food information extraction, food ontology, named-entity recognition.

I. INTRODUCTION

In different food information systems, great amounts of heterogeneous data related to food and nutrition is collected. For example, the European Food Safety Agency (EFSA) maintains the Comprehensive Food Consumption Database as a source of information on food consumption across the European Union (available at <https://www.efsa.europa.eu/en/food-consumption/comprehensive-database>). To assess nutrient intakes of the EU population, its data is matched with the food composition data that is another type and source of food-related data [1]. Beside food consumption and food composition data that is in the form of structured data, we need to mention unstructured data (e.g. text, images, sound), which is also an important source of information. For instance, the mentioned assessment of nutrient intakes requires knowledge about dietary recommendations that is based on evidence collected in peer-reviewed papers and scientific reports. The extraction of such evidence requires extensive work from human experts,

The associate editor coordinating the review of this manuscript and approving it for publication was Xiping Hu.

which can be eased by using state-of-the-art computer techniques [2]. Great efforts have already been done regarding automated extraction of information not only from text, but also from food images [3], food odours, etc.

Extracting food entities from textual data is a challenging task that can be used for filling in the gaps in many practical applications. Automatically extracting food entities as well as all other biomedical entities (i.e. drugs, diseases, treatments, etc.) from scientific papers can help us to follow the knowledge that comes with each new publication. Additionally, this is useful for analyzing relations that exist between these entities. Moreover, machine learning (ML) algorithms can be used to find some hidden (i.e. unknown) relations that exist between food entities and disease entities. This is especially important for food allergy studies. Additionally, automated extraction of food information can be used to fill in missing values that appear in food-related databases (e.g., food composition databases). Another interesting application is where information extraction is used to extract food entities from dietary records for individuals (i.e. written as free-form text), and then map them on a nutrient level. This information can be combined and used by recommender systems.

In this paper, we focus on the task known as information extraction (IE) from food-related text (i.e. publicly published recipes).

For this reason, we provide a survey of the available food-named entity recognition methods as well as compare their performances in order to see which one provides the most promising results. The comparison is made using a data set of 1,000 recipes that were extracted from Allrecipes [4], which is the largest food-focused social network, where everyone plays a part in helping chefs and cooking enthusiasts discover and share home cooking.

II. RELATED WORK

In this section, different methods and resources that can be used for food information extraction are presented. We start by providing an overview of available food ontologies, followed by an overview of different named-entity recognition methods (NERs) that can be used for information extraction from food-related data.

A. INFORMATION EXTRACTION

Nowadays, computer-based methods are extremely important for automated extraction of food information from unstructured textual data. The goal of this is to follow the rapidly increasing knowledge in the food domain which is presented in published scientific papers. These methods are based on natural language processing (NLP) [5] and machine learning (ML) [6]. NLP is a sub-field of computer science and artificial intelligence that is concerned with the interactions between computers and natural human languages, while ML is focused on learning statistical methods based on a sample data set known as “training data” in order to make predictions, classifications, or decisions.

Natural language texts carry textual information in the form of unstructured data, meaning that it has no predefined data model. This means that the information and relations that the data represents are not explicitly stated or formatted. Working with textual data is a difficult task because of its variability - the same entities can be mentioned in various ways regarding the differences in how people express themselves and use diverse writing styles.

Information Extraction (IE) is the task of automatically extracting information from unstructured data and, in most cases, is concerned with the processing of natural language texts [7], [8]. The goal of IE is to provide a structured representation of the extracted information captured from the analyzed text.

The information to be extracted is contained within the texts themselves. It consists of predefined entities of interest, as well as relationships between the entities that are typical for some domain. For example, users may be interested in extracting information about food entities, nutrient entities, quantity/unit entities, population groups, together with the relations between them. Let us assume that we have the dietary recommendation “*Babies need about 10g protein a day*”. Using an IE method, the extracted information should

be “*Babies*” as a population group, “*protein*” as a nutrient entity, “*10g a day*” as a quantity/unit entity, and “*need*” as the relation between the population group and the nutrient entity.

One well known IE task is named-entity recognition (NER), which addresses the problem of identification and classification of predefined concepts in a given domain [9]. It aims to identify words or phrases from the text and then label them into predefined classes (labels) that describe concepts of interest in a given domain.

To illustrate NER, more specifically NER in the food domain with a concrete example, let us consider the following example sentence:

“*Pour the egg mixture over the Cheddar cheese-covered hash browns in the skillet. Sprinkle bacon and sausage pieces on top.*”

For the example sentence, a NER method designed to extract concepts in the food domain would ideally produce the output set: [*egg mixture, Cheddar cheese-covered hash browns, bacon, sausage pieces*].

Various NER methods exist: *terminology-driven, rule-based, corpus-based, methods based on active learning (AL), and methods based on deep neural networks (DNNs)*. An overview of these types of NER methods is presented in Table 1. Additionally, hybrid approaches for named-entity recognition from unstructured textual data exist. One such example can be found in [10], where the authors present a NER method composed of rule-based deep learning as well as clustering-based approaches.

In recent years, numerous NER methods have been developed for the biomedical domain [25]–[27], which are available together with comparison studies on different benchmark data sets [28]–[30]. For example, QuickUMLS is a fast unsupervised technique for medical concept extraction [31]. Clinical Named Entity Recognition system (ClinER) is a named entity recognition method that can be used for extracting clinical entities from electronic health records [25]. However, to the best of our knowledge, there are very few food-named entity recognition methods and no comprehensive comparative studies between them. It is important to mention that it is very challenging to transfer any existing NER method to a different domain. This is the case because many of them are trained on data from a specific domain (i.e. in the case of corpus-based NERs), or they are based on rules that are crucial for one domain, but not important to other domains.

B. FOOD ONTOLOGIES

An ontology is a formal, explicit specification of a shared conceptualization [32]. They represent data models which define formal data representations, concepts of interest, categories, properties and relations. Domain specific ontologies are especially useful as they provide a formal, standardized data model of the domain. Moreover, these standardized data models enable easy extensions and updates of the data represented in the ontologies.

TABLE 1. Overview of different types of NER methods.

NER method type	Description
Terminology-driven [11]	<ul style="list-style-type: none"> • Performs matching on text phrases with concept synonyms from terminological resources. • Combines heuristics, such as generating permutations of concept synonyms, solving disambiguation, etc. • Advantage: frequent updates of the terminological resource with new concepts and synonyms. • Disadvantage: only entities that exist in the resource can be recognized. • Disadvantage: creation of a well-constructed and inclusive terminological resource is a time consuming task.
Rule-based [12] [13]	<ul style="list-style-type: none"> • Usually use regular expressions combining information from terminological resources and characteristics of the entities. • Advantage: frequent updates of the terminological resource with new concepts and synonyms. • Disadvantage: manual construction of the domain-specific rules, which is a time consuming task.
Corpus-based [14] [15]	<ul style="list-style-type: none"> • Utilizes information that is found in annotated corpora provided by human experts in the domain of interest, as well as ML algorithms in order to predict the entities' labels. • Advantage: less affected by terminological resources and manually created rules. • Disadvantage: require the existence of an annotated corpus for the task at hand.
Based on Active Learning (AL) [16]–[18]	<ul style="list-style-type: none"> • Uses a form of semi-supervised learning [19] which includes unlabeled data in the training process. • The AL algorithms can interactively query the user to obtain desired outputs for new data instances. • Usually consists of three components: the annotation interface, the corpus-based NER, and a component for querying samples • Advantage: the number of examples to learn a new concept can often be much lower, compared to supervised learning. • Disadvantage: requires continuous user input.
Based on Deep Neural Networks (DNNs)	<ul style="list-style-type: none"> • Does not rely on hand-crafted, domain-specific features. [20]–[24]. • Advantage: minimizes the effort put into the costly process of feature engineering. • Disadvantage: typically requires a much greater amount of annotated data.

In the food domain, few ontologies have been constructed, with the difference between them being that they are constructed for different application scenarios. Also, it is possible for food instances which exist in one ontology to not exist in others. For example, “apple” can be a part of one ontology, but not part of another ontology. Moreover, the relations between the concepts that exist in different ontologies can be different. Some of them may consist of relations between food and disease concepts, others may consist of relations between recipes and ingredients, and so on. For this paper, we selected the food ontologies that are available in the BioPortal. They are presented in Table 2 in addition to some other food ontologies.

A detailed review of the food ontologies and resources has been provided by [37].

C. FOOD NAMED-ENTITY RECOGNITION

The UCREL Semantic Analysis System (USAS) [36] is a system for automatic semantic analysis of text. Out of the 21 high level categories it consists of, of particular interest in our case is the category “food and farming”. Furthermore, one can obtain additional semantic information about the concept of interest. However, one significant drawback is that USAS only works on a word level. As an example consider a food concept represented, in some free-form text, by two words - “cooked beef”. The goal here is to extract and annotate this food concept. In this case, the USAS system would extract and annotated both words from the food concept representation as separate concepts, i.e. one concept would be “cooked”, and the other “beef”. This does not adequately represent the food concept at hand, as both tokens are semantically represent one food concept.

The NCBO [38] annotator is a web framework that extracts and annotates food concepts from free-form text provided by the user. The domain of the concepts and the performance of the annotator depend on the ontology that chosen to be used in the background. Accordingly, it is able to extract and annotate only the concepts that are present in the ontology. For this reason, any combination of NCBO and an ontology can be assumed as a different NER method. Its annotation workflow is centered around a highly efficient syntactic concept recognition engine and a set of semantic expansion algorithms. The NCBO annotator is available within the BioPortal software services and is able to use ontologies that are available there as well [39].

DrNER [2] is a rule-based named-entity recognition system aimed at extracting information from evidence-based dietary recommendations. Apart from nutritional information, food concepts are also in the domain of this NER system. However, this work was then extended into a rule-based named-entity recognition system specifically tasked with food information extraction. The core of this new food NER system - FoodIE [40], consists of a rule engine. These rules are based on computational linguistics and semantic information which describe each food concept. Unlike USAS, FoodIE takes into account word chunking when extracting and annotating the food concepts, i.e. multiple words (tokens) can be grouped into a single food concept. The evaluation of this method has been performed using two independent benchmark data sets. The first one consists of 200 recipes extracted from Allrecipes [4], including recipes from five categories: Appetizers and snacks, Breakfast and Lunch, Dessert, Dinner, and Drinks. From each recipe category 40 recipes were included in the first benchmark data set. The second benchmark data set consists of 1000 new recipes also extracted from

TABLE 2. Main food ontologies and semantic resources.

(a) Short description

Name	Description
FoodOn [33]	Focused on human-centric categorization and handling of food.
OntoFood	With Semantic Web Rule Language (SWRL) rules of nutrition for diabetic patients.
SNOMED CT [34]	Standardized, multilingual vocabulary of clinical terms which additionally includes food concepts.
Hansard corpus [35] [36]	Part of the SAMUELS project with 37 higher level semantic group, one of which is <i>Food and Drinks</i> .

(b) Resource availability

Name	Availability
FoodOn	https://foodontology.github.io/foodon/
OntoFood	https://bioportal.bioontology.org/ontologies/OF/?p=summary
SNOMED CT	https://confluence.ihtsdotools.org/display/DOC/Technical+Resources
Hansard corpus	https://www.hansard-corpus.org/

the Allrecipes website and consists of 200 recipes from each recipe category. After extracting the food entities, one human expert manually selected what should be extracted while another human expert compared the results with what was extracted using FoodIE. The evaluation that was done using the two different data sets showed that FoodIE's behaviour is consistent, as well as that it achieves very promising evaluation results.

III. METHODOLOGY FOR COMPARING FOOD NAMED-ENTITY RECOGNITION METHODS

In this section we provide an explanation of the methodology used for comparing food named-entity recognition methods. First, the benchmarking data set that is used to compare the selected methods is explained, followed by the selected methods used for comparison. Finally, definitions of the evaluation metrics that are used to compare the selected methods on the benchmarking data set are given. The data analysis script is publicly available at https://github.com/GorjanP/food_NER_comparison_script.

A. BENCHMARKING DATA SET

To evaluate the results of the NER methods we used the recently published FoodBase corpus [41], which consists of 1000 recipes annotated with food concepts. The recipes were taken from the most popular recipe sharing social network - Allrecipes, selecting them from five categories, 200 from each. The categories it provides are: Appetizers and snacks, Breakfast and Lunch, Dessert, Dinner, and Drinks. To the best of our knowledge, this corpus is one of the first annotated corpora with food entities. It was created by extracting the food entities by using a rule-based approach, and then manually evaluated by two subject matter experts. To reduce potential human bias, one subject matter expert extracted all food entities in each recipe manually, while the other checked if each food entity extracted manually is in the list of the food entities extracted by the rule based approach. After this, the False Negatives (FNs) were manually added in the corpus, while the False Positives (FPs), which were usually related to some cooking procedures

and tools, were removed, in order to create a ground truth data set. This data set is presented in the BioC format, which is a simple format to share text data and annotations, with the goals of simplicity, interoperability, and broad use and reuse [42]. The ground truth data set can be found at <http://cs.ijs.si/repository/FoodBase/foodbase.zip>.

B. METHODS

With the goal of extracting food entities from each recipe, two different approaches were used:

- 1) FoodIE - our recently proposed rule-based food named-entity recognition system.
- 2) NCBO annotator - performed three times, each time running on a different ontology (FoodOn, OntoFood, SNOMED CT). As the extractions depends on the underlying ontology that is used in conjunction with the NCBO annotator, every run is considered as a different NER method.

This resulted in a total of four sets of NER methods which are the compared: FoodIE, NCBO (FoodOn), NCBO (OntoFood), and NCBO (SNOMED CT). Before we compare the evaluation results of the NER methods, we are going to explain the methodology behind FoodIE and the NCBO annotator in more detail.

FoodIE is a rule-based food NER, which is based on computational linguistics and rules that incorporate semantic information. It primarily consists of four steps. The first step performs pre-processing of the unstructured text, by removing non-standard characters, removing excess white spaces, as well as performing ASCII transliteration. The second step includes morphological analysis by combining the results of two Part-of-Speech (POS) taggers in order to obtain more robust tags. The next step defines the rules that use computational linguistics information as well as semantic information from the Hansard corpus related to the food entities. This step defines the possible phrases which can be candidates for food entities. The final step classifies the candidate phrases as food entities or non-food entities.

The results from FoodIE are organized in the BioC format, the same format as the ground truth data set. The BioC format

```

<document>
  <id>4recipe761</id>
  <infony key="category">Dinners</infony>
  <infony key="full_text">
    Preheat oven to 350 degrees F (175 degrees C). Spray a baking dish with cooking spray.
    Whisk egg in a shallow bowl. Mix Parmesan cheese and Cajun seasoning together on a plate.
    Dip each pork chop into egg. Press into Parmesan mixture until coated on both sides.
    Place in the prepared baking dish. Bake in the preheated oven until golden and
    an instant-read thermometer inserted into the center reads at least
    145 degrees F (63 degrees C), 35 to 40 minutes.
  </infony>
  <annotation id="1">
    <location offset="18" length="13"/>
    <text>cooking spray</text>
    <infony key="semantic_tags">
      AG.01.t.07 [Cooking];AG.01.f [Fat/oil];
    </infony>
  </annotation>
  <annotation id="2">
    <location offset="22" length="3"/>
    <text>egg</text>
    <infony key="semantic_tags">
      AG.01.g [Eggs];
    </infony>
  </annotation>
  <annotation id="3">
    <location offset="29" length="15"/>
    <text>Parmesan cheese</text>
    <infony key="semantic_tags">
      AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];
    </infony>
  </annotation>
  <annotation id="4">
    <location offset="32" length="15"/>
    <text>Cajun seasoning</text>
    <infony key="semantic_tags">
      AG.01.l [Additive];AG.01.t.05 [Preparation for table/cooking];
    </infony>
  </annotation>
  <annotation id="5">
    <location offset="41" length="9"/>
    <text>pork chop</text>
    <infony key="semantic_tags">
      AG.01.d [Animals for food];AG.01.d.05 [Pork];AG.01.o [Animal food];
    </infony>
  </annotation>
  <annotation id="6">
    <location offset="44" length="3"/>
    <text>egg</text>
    <infony key="semantic_tags">
      AG.01.g [Eggs];
    </infony>
  </annotation>
  <annotation id="7">
    <location offset="48" length="16"/>
    <text>Parmesan mixture</text>
    <infony key="semantic_tags">
      AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];
    </infony>
  </annotation>
</document>

```

Listing 1. Example recipe in the BioC format, as annotated by FoodIE.

for one recipe and its annotations as processed by FoodIE are presented in Listing 1. From it, we can see that each recipe is presented as a document for which the category, description (full text), and food annotations are included. Each annotation consists of the food entity that is extracted, the semantic tags from the Hansard corpus that are assigned to it, and the offset that points the position from the beginning of text where the food entity starts, as well as its length. The offset is expressed on a token level, while the length is expressed as the number of characters in the annotated food entity.

The National Center for Biomedical Ontology (NCBO) Annotator is an ontology-based web service that is used for annotating unstructured textual data with biomedical ontology entities. It consists of two main steps. The first step consists of a selection of an appropriate domain dictionary. The dictionary is constructed by aggregating all entity names that belong to the domain of interest. The second step is the annotator which uses Mgrep [43] to recognize the entities by using string matching on the dictionary.

The same recipe used as an example above, represented by a NCBO annotation, is presented in Table 3. In the table there

TABLE 3. The same example recipe as annotated by NCBO using the SNOMED CT ontology.

	SNOMED CT ID	text	from	to	matchType
1	http://purl.bioontology.org/ontology/SNOMEDCT/226838004	PARMESAN CHEESE	121	135	PREF
2	http://purl.bioontology.org/ontology/SNOMEDCT/102264005	CHEESE	130	135	PREF
3	http://purl.bioontology.org/ontology/SNOMEDCT/227553009	SEASONING	147	155	PREF
4	http://purl.bioontology.org/ontology/SNOMEDCT/226934003	PORK	187	190	PREF

are four columns. Each one provides different information about the annotations: a semantic tag (i.e. id url), the text that represents the food concept, and two numbers pointing to where the annotation begins and ends. These two numbers, referred to as *from* and *to*, are expressed in the offset in characters. It is important to mention that in the process of comparing the NCBO NER methods these character offsets were converted to token offsets, as is the case in the BioC recipe format.

To perform the evaluation, the outputs from each NER method were compared with the annotations found in the ground truth dataset.

C. EVALUATION METRICS

For evaluation, we used a confusion matrix, which is also called an error matrix, that is used for visualization of the performance of NER methods (more generally classification methods). Each row of the matrix represents the entities in a predicted class while each column represents the entities in an actual class. Using it, four metrics can be defined: true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). For evaluation we selected three metrics out of four: TPs, FPs, and FNs. We did not select the true negatives (TNs). A true negative (TN) is a match where the NER method correctly extracts a negative entity. In our case we do not have a negative entity, we are interested only in one class, which is the food class, so every word or phrase that is not extracted as a food entity, and it indeed is not a food entity, can be assumed as a TN.

Additionally, a type of match called “partial” is introduced, as some of the food concepts that were extracted were incomplete, but still contained some semantic information of relevance. This category encompasses all the extracted food entities which were caught, but missed at least one token that belongs to that food entity.

The meanings of these evaluation metrics are:

- True Positives (TPs) - This type of match occurs when all of the tokens from the NER method are an exact match with the same food entity in the ground truth data set (as distinguished by the respective offsets).
- False Negatives (FNs) - This type of match occurs when a certain annotation is not present when it indeed should be classified as a food entity. This happens when a food entity is not correctly extracted by the NER method.
- False Positives (FPs) - This is the inverse type of match from FNs, i.e. this occurs when the extracted food entity is falsely done so. This happens when something that is not a food entity is classified as one.

- Partial - This type of match is very specific, it occurs when only part of the whole food entity has been extracted and annotated. If at least one token (word) is missing or falsely superfluous, this type of match is present.

The bigger the number of TPs, the better, while the number of FPs and FNs should be minimized. Regarding the partial match type, it is better to have a TP match than a partial match, but partial matches are of importance if the alternative is not to match anything at all.

To concretely illustrate these four types of matches, let us consider the following few examples.

The True Positive (TP) type of match is quite straightforward; consider the following sentence: “*Let the water boil until the carrots are tender.*” In it, we have two food entities that should be extracted: *water* and *carrots*. If a NER method extracts both, we would have two TP matches.

Regarding False Negatives (FNs), some food entities indeed occur rarely, but nevertheless carry significant information. Consider the sentence “*Substitute sugar for stevia if desired.*” In it, we have two food entities of interest: *sugar* and *stevia*. However, due to the rarity of the concept *stevia*, many NER methods fail to identify this as a food entity and do not extract it from the raw text. It is apparent that such failures to classify certain food entities can carry significant implications, as *stevia* is supposed to be a safe alternative to *sugar*.

An instance of the False Positive (FP) match type would be if the NER method extracts food entities such as *milk frother* or *coffee mug*. In both cases, the concept is related to the food domain, but it does not represent a food entity in and of itself. Usually, concepts that are FPs represent general objects.

The final type of match is the partial match. These matches can either occur when a word (token) is missing from the extracted food entity, or when an unrelated word (token) is included in the extracted food entity. For example, consider the sentence: “*Empty tropical fruit juice in the glass.*” The only food entity in this sentence is *tropical fruit juice*. However, if a NER method extracts *fruit juice* or *empty tropical fruit juice* a partial match occurs. In the first case, a token which carries important semantic information (*tropical*) is missing. In the second case, a token (*empty*) which is not part of the food entity is present.

Using these match types, three unique statistical evaluation metrics were calculated. Each one is focused on a specific aspect of performance evaluation:

TABLE 4. Evaluation metrics (1,000 recipes).

	FoodIE	NCBO (SNOMED CT)	NCBO (OntoFood)	NCBO (FoodON)
F ₁ Score	96.05%	63.75%	32.62%	63.90%
Precision	97.80%	91.53%	85.48%	79.22%
Recall	94.37%	48.91%	20.16%	53.54%

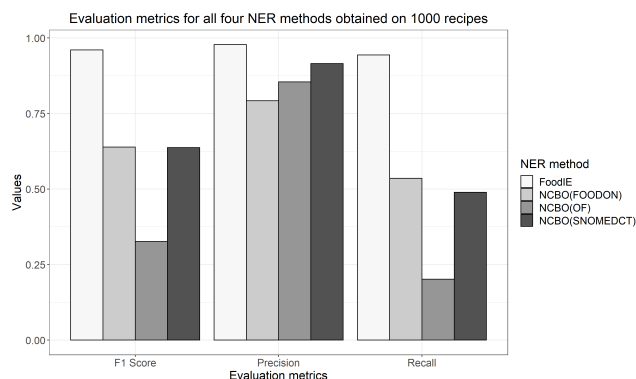
- **Precision** - This evaluation metric is defined as $Precision = \frac{TP}{TP+FP}$. Precision evaluates the fraction of correctly extracted entities over the all the extracted entities (including entities which should not have been extracted as food entities)
- **Recall** - This evaluation metric is defined as $Recall = \frac{TP}{TP+FN}$. Recall evaluates the fraction of the correctly extracted entities over the total amount of ground truth entities (as classified by a human expert).
- **F₁ Score** - This evaluation metric is defined as $F_1\ Score = \frac{2TP}{2TP+FP+FN}$. The F₁ Score is often used as a more robust evaluation metric, as it combines both aspects (Precision and Recall) into a more robust evaluation metric.

All of these evaluation metrics are on a real scale from 0 to 1 (i.e. in the real range [0, 1]), and they should be maximized. The bigger the value for each evaluation metric, the better. Precision gives us the ratio of correctly extracted positive entities to the total extracted positive entities. It highlights the correct positive extracted entities out of all the positive extracted entities. High precision indicates low false positive rate. The recall gives us the ratio of correctly extracted positive entities to the actual positive entities. It highlights the sensitivity of the algorithm i.e. out of all the actual entities how many were caught by the NER. When we have applications where the false negatives are important, recall is a better measure than the precision, while when the false negatives is less of concern, precision is the more appropriate metric. However, if we would like to make a general conclusion that takes both into an account, we should use the F1 score. It is a weighted average of the Precision and Recall, which takes both false positives and false negatives into account. It also has some issues as it is biased to the majority class and it does not take into account the true negatives (TNs). However, reporting it in the cases of one-class classification is not unreasonable, which is the case in our NER (i.e. a single class - food). Precision, Recall, and F1 Score are usually reported instead of accuracy, as they offer more detailed insights about the NER that is analyzed. Accuracy is avoided as it relies on the concept of True Negatives (TNs), which is not of interest to named-entity recognition tasks.

IV. RESULTS

The results from the calculated evaluation metrics are presented in Table 4 and on Figure 1.

It is interesting to note that not all ontologies provided annotations for each recipes. The number of missed recipes for each NER method is given in Table 5.

**FIGURE 1.** Evaluation metrics (1,000 recipes).**TABLE 5.** Missed recipes for each NER method (out of 1000).

NER method	Total missed recipes
FoodIE	0
NCBO (SNOMED CT)	6
NCBO (OntoFood)	71
NCBO (FoodON)	5

These recipes were missed due to the fact that the respective ontologies do not cover the food domain well. This means that if a certain recipe contains information that is not present in a certain ontology, it will produce an empty annotation.

V. DISCUSSION

A. MATCH TYPES

The largest number of TPs is obtained by FoodIE (11,461). The three other methods, by using the NCBO annotator on SNOMED CT, OntoFood, and FoodOn, obtain 5,100, 2,279, and 5,725 TPs respectively. As previously discussed, the number of TPs should be maximized.

Regarding the number of FPs, FoodIE is again the most promising, obtaining only 258 FP instances. In contrast, the three other methods respectively obtain 472, 378, and 1502 FPs. As mentioned in Section III-C, the number of FPs should be minimized.

A similar situation is observed regarding FNs, where FoodIE once again has the lowest (i.e. most preferable) number of instances. It obtains 684, while the others respectively have 5327, 9026, and 4968. As mentioned in the same section as the in the previous paragraph, the number of FNs should be minimized. The reason for such a disparity is due to the domain coverage of the ontologies used by the NCBO annotator, i.e. SNOMED CT, OntoFood and FoodOn. Specifically, this means that many of the entities that are found in the recipes do not exist in the respective ontologies.

In contrast to this, the semantic information that FoodIE utilizes is more representative of the domain, and hence it has a great advantage.

The final and most specific type of match we mention is the partial match type. In contrast to the previous three types of matches, which all have a clearly defined optimization goal, it is not clear whether the partial type of match should be maximized or minimized. Additionally, this type of match strongly depends on the outcome of the other three types of matches, especially TPs and FNs. For instance, an ideal evaluation would be if all food concepts are counted as TPs and none as FNs. Nonetheless, if a TP type match is not made for a food concept instance, it would be better to have it as a partial type match rather than a FN.

B. EVALUATION METRICS

As all of the used evaluation metrics are to be ideally maximized, the comparison of these metrics is quite straightforward. For each evaluation metric, FoodIE outperforms the remaining three NER methods. Specifically, F_1 Scores are: FoodIE (96.05%), SNOMED CT (63.75%), OntoFood (32.62%), and FoodON (63.90%). The metric value is on a scale from 0 to 1, we express it as percentages, as is standard. It is apparent that FoodIE has quite a notable advantage over the other three NER methods, as the absolute differences for the F_1 Score between FoodIE and the remaining three NER methods are: 32.30%, 63.43%, and 32.15%, respectively. The NER method with the worst evaluation metrics is NCBO (OntoFood), which also gives us an indication that the OntoFood ontology does not cover the food domain adequately, i.e., many food entities are not present in it.

VI. CONCLUSION

Evaluating four different NER methods in the food domain: FoodIE, NCBO (SNOMED CT), NCBO (OntoFood), and NCBO (FoodON) on a data set of 1,000 manually annotated recipes, it is evident that FoodIE provides more promising results for each individual evaluation metric, as well as the best overall result.

Additionally, extracting food entities can further be linked with entities from other domains, such as health, bioinformatics, consumer and social sciences etc. This can help in reducing knowledge gaps that inhibit public health goals as well as the optimal development of scientific, agricultural and industrial policies. To work towards this goal, in our future work, we aim to upgrade the FoodIE NER method to support the extraction of information from data relevant for all fields of the food science (e.g. food safety, food authenticity and traceability, food sustainability).

REFERENCES

- [1] B. K. Seljak, P. Korošec, T. Eftimov, M. Ocke, J. van der Laan, M. Roe, R. Berry, S. Crispim, A. Turrini, and C. Krems, "Identification of requirements for computer-supported matching of food consumption data with food composition data," *Nutrients*, vol. 10, no. 4, p. 433, 2018.
- [2] T. Eftimov, B. K. 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, Jun. 2017, Art. no. e0179488.
- [3] S. Mezgec, T. Eftimov, T. Bucher, and B. K. Seljak, "Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment," *Public Health Nutrition*, vol. 22, no. 7, pp. 1193–1202, 2019.
- [4] S. Groves, "How allrecipes.com became the worlds largest food/recipe site. ROI of social media (BLOG)," Tech. Rep., 2013.
- [5] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, vol. 999. Cambridge, MA, USA: MIT Press, 1999.
- [6] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. (Springer Series in Statistics), vol. 1, no. 10. New York, NY, USA: Springer, 2001.
- [7] J. Piskorski and R. Yangarber, "Information extraction: Past, present and future," in *Multi-source, Multilingual Information Extraction and Summarization*. Berlin, Germany: Springer, 2013, pp. 23–49.
- [8] C. C. Aggarwal and C. Zhai, *Mining Text Data*. New York, NY, USA: Springer, 2012.
- [9] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.
- [10] A. Thomas and S. Sangeetha, "An innovative hybrid approach for extracting named entities from unstructured text data," *Comput. Intell.*, vol. 35, no. 4, pp. 799–826, Nov. 2019.
- [11] X. Zhou, X. Zhang, and X. Hu, "MaxMatcher: Biological concept extraction using approximate dictionary lookup," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Berlin, Germany: Springer, 2006, pp. 1145–1149.
- [12] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems," in *Proc. 39th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 426–433.
- [13] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "ProMiner: Rule-based protein and gene entity recognition," *BMC Bioinf.*, vol. 6, no. 1, p. S14, 2005.
- [14] N. Alnazzawi, P. Thompson, R. Batista-Navarro, and S. Ananiadou, "Using text mining techniques to extract phenotypic information from the phenochf corpus," *BMC Med. Informat. Decis. Making*, vol. 15, no. 2, p. 1, 2015.
- [15] R. Leaman, C.-H. Wei, C. Zou, and Z. Lu, "Mining patents with tmChem, GNormPlus and an ensemble of open systems," in *Proc. 5th BioCreative Challenge Eval. Workshop*, 2015, pp. 140–146.
- [16] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, and H. Xu, "A study of active learning methods for named entity recognition in clinical text," *J. Biomed. Informat.*, vol. 58, pp. 11–18, Dec. 2015.
- [17] Y. Chen, T. A. Lask, Q. Mei, Q. Chen, S. Moon, J. Wang, K. Nguyen, T. Dawodu, T. Cohen, and J. C. Denny, "An active learning-enabled annotation system for clinical named entity recognition," *BMC Med. Informat. Decis. Making*, vol. 17, no. 2, p. 82, 2017.
- [18] V. C. Tran, N. T. Nguyen, H. Fujita, D. T. Hoang, and D. Hwang, "A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields," *Knowl.-Based Syst.*, vol. 132, pp. 179–187, Sep. 2017.
- [19] B. Settles, *Active Learning Literature Survey*, vol. 52, nos. 55–66. Madison, WI, USA: Univ. Wisconsin-Madison, 2010, p. 11.
- [20] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*. [Online]. Available: <http://arxiv.org/abs/1508.01991>
- [21] C. N. dos Santos and V. Guimarães, "Boosting named entity recognition with neural character embeddings," 2015, *arXiv:1505.05008*. [Online]. Available: <http://arxiv.org/abs/1505.05008>
- [22] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, *arXiv:1603.01360*. [Online]. Available: <http://arxiv.org/abs/1603.01360>
- [23] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, Jul. 2017.
- [24] M. M. Lopez and J. Kalita, "Deep learning applied to NLP," 2017, *arXiv:1703.03091*. [Online]. Available: <http://arxiv.org/abs/1703.03091>
- [25] W. Boag, E. Sergeeva, S. Kulshreshtha, P. Szolovits, A. Rumshisky, and T. Naumann, "Cliner 2.0: Accessible and accurate clinical concept extraction," 2018, *arXiv:1803.02245*. [Online]. Available: <https://arxiv.org/abs/1803.02245>
- [26] P. J. Gorinski, H. Wu, C. Grover, R. Tobin, C. Talbot, H. Whalley, C. Sudlow, W. Whiteley, and B. Alex, "Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches," 2019, *arXiv:1903.03985*. [Online]. Available: <http://arxiv.org/abs/1903.03985>

- [27] K. Xu, Z. Yang, P. Kang, Q. Wang, and W. Liu, "Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition," *Comput. Biol. Med.*, vol. 108, pp. 122–132, May 2019.
- [28] A. Jagannatha, F. Liu, W. Liu, and H. Yu, "Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0)," *Drug Saf.*, vol. 42, no. 1, pp. 99–111, Jan. 2019.
- [29] A. Stubbs, C. Kotfila, H. Xu, and Ö. Uzuner, "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2," *J. Biomed. Informat.*, vol. 58, pp. S67–S77, Dec. 2015.
- [30] C. Nédellec, R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, "Overview of BioNLP shared task 2013," in *Proc. BioNLP Shared Task Workshop*, 2013, pp. 1–7.
- [31] L. Soldaini and N. Goharian, "QuickUMLS: A fast, unsupervised approach for medical concept extraction," in *Proc. MedIR Workshop, Sigir*, 2016, pp. 1–4.
- [32] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquisition*, vol. 5, no. 2, pp. 199–220, 1993. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1042814383710083>
- [33] E. J. Griffiths, D. M. Dooley, P. L. Buttigieg, R. Hoehndorf, F. S. Brinkman, and W. W. Hsiao, "FoodOn: A global farm-to-fork food ontology," in *ICBO/BioCreative*, 2016, pp. 1–2.
- [34] K. Donnelly, "SNOMED-CT: The advanced terminology and coding system for eHealth," *Stud. Health Technol. Informat.*, vol. 121, p. 279, Jan. 2006.
- [35] F. Nanni, S. Menini, S. Tonelli, and S. P. Ponzetto, "Semantifying the UK hancard (1918–2018)," in *Proc. ACM/IEEE Joint Conf. Digit. Libraries (JCDL)*, Jun. 2019, pp. 412–413.
- [36] P. Rayson, D. Archer, S. Piao, and A. M. McEnery, "The UCREL semantic analysis system," in *Proc. Workshop Beyond Named Entity Recognit. Semantic Labelling NLP Tasks, Assoc. (LREC)*. European Language Resources Association, 2004, pp. 7–12.
- [37] M. N. K. Boulous, A. Yassine, S. Shirmohammadi, C. S. Namahoot, and M. Brückner, "Towards an 'Internet of food': Food ontologies for the Internet of Things," *Future Internet*, vol. 7, no. 4, pp. 372–392, 2015.
- [38] C. Jonquet, N. Shah, C. Youn, C. Callendar, M.-A. Storey, and M. Musen, "NCBO annotator: Semantic annotation of biomedical data," in *Proc. Int. Semantic Web Conf., Poster Demo Session*, vol. 110, 2009, pp. 1–3.
- [39] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, and C. G. Chute, "BioPortal: Ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Res.*, vol. 37, no. 2, pp. W170–W173, 2009.
- [40] G. Popovski, S. Kochev, B. Koroušič Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction," in *Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods, (ICPRAM)*, 2019, pp. 915–922.
- [41] G. Popovski, B. K. Seljak, and T. Eftimov, "FoodBase corpus: A new resource of annotated food entities," *Database*, vol. 2019, p. baz121, Nov. 2019, doi: [10.1093/database/baz121](https://doi.org/10.1093/database/baz121).
- [42] D. C. Comeau, R. I. Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, and M. Torii, "BioC: A minimalist approach to interoperability for biomedical text processing," *Database*, vol. 2013, p. bat064, Sep. 2013.
- [43] M. Dai, N. H. Shah, W. Xuan, M. A. Musen, S. J. Watson, B. D. Athey, and F. Meng, "An efficient solution for mapping free text to ontology terms," *AMIA Summit Transl. Bioinf.*, vol. 21, 2008.



GORJAN POPOVSKI received the bachelor's degree from the Faculty of Computer Science and Engineering, Skopje, North Macedonia. He is currently pursuing the master's degree with the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia.

His main topics of interest include machine learning, deep learning, natural language processing, algorithms and complexity, knowledge engineering, and knowledge representation.



BARBARA KOROUŠIČ SELJAK works with the Computer Systems Department, Jožef Stefan Institute. She has a leading role in several EU-funded projects including REFRESH, SYNERGY, and RICHFIELDS. She acts as an Executive Director of EuroFIR AISBL, which is a non-profit international member-based organization based in Belgium, and she is a Council Member of the Slovenian Society for Clinical Nutrition. She has contributed to the preparation and validation of

the Slovenian dietary guidelines for children in schools and kindergartens, students, workers, and patients and elderly, and she developed a tool for food dietary assessment, which has been used by Slovenian hospitals. She has supervised a few tens of bachelor's, master's, Ph.D. degrees students for completion. Her current main research areas are food and nutrition data management, and knowledge discovery.



TOME EFTIMOV received the Ph.D. degree from the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, in 2018. He was a Postdoctoral Research Fellow of the Department of Biomedical Data Science and the Centre for Population Health Sciences with Stanford University, USA, and a Research Associate with the University of California, San Francisco, USA (UCSF). He is a Research Fellow of the Jožef Stefan Institute, Ljubljana, Slovenia. His main areas of

research include statistics, natural language processing, heuristic optimization, machine learning, and representational learning. His specific topics of research include information extraction of food- and nutrition-related concepts from textual data, food data normalization, and knowledge management for food- and nutrition-related data.

...