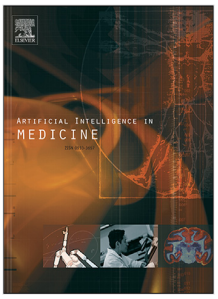
Journal Pre-proof



FooDis: A food-disease relation mining pipeline

Gjorgjina Cenikj, Tome Eftimov, Barbara Koroušić Seljak

PII: S0933-3657(23)00100-8
DOI: https://doi.org/10.1016/j.artmed.2023.102586
Reference: ARTMED 102586

To appear in: *Artificial Intelligence In Medicine*

Received date : 11 April 2022
Revised date : 7 April 2023
Accepted date : 16 May 2023

Please cite this article as: G. Cenikj, T. Eftimov and B. Koroušić Seljak, FooDis: A food-disease relation mining pipeline. *Artificial Intelligence In Medicine* (2023), doi: https://doi.org/10.1016/j.artmed.2023.102586.

FooDis: A food-disease relation mining pipeline

Gjorgjina Cenikj[a,b,*], Tome Eftimov[a], Barbara Koroušić Seljak[a]

[a] *Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[b] *Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000, Ljubljana, Slovenia*

Abstract

Nowadays, it is really important and crucial to follow the new biomedical knowledge that is presented in scientific literature. To this end, Information Extraction pipelines can help to automatically extract meaningful relations from textual data that further require additional checks by domain experts. In the last two decades, a lot of work has been performed for extracting relations between phenotype and health concepts, however, the relations with food entities which are one of the most important environmental concepts have never been explored.

In this study, we propose FooDis, a novel Information Extraction pipeline that employs state-of-the-art approaches in Natural Language Processing to mine abstracts of biomedical scientific papers and automatically suggests potential *cause* or *treat* relations between food and disease entities in different existing semantic resources.

A comparison with already known relations indicates that the relations predicted by our pipeline match for 90% of the food-disease pairs that are common in our results and the NutriChem database, and 93% of the common pairs in the DietRx platform. The comparison also shows that the FooDis pipeline can suggest relations with high precision. The FooDis pipeline can be further used to dynamically discover new relations between food and diseases that should be checked by domain experts and further used to populate some of the existing resources used by NutriChem and DietRx.

*Corresponding author.
Email addresses: gjorgjina.cenikj@ijs.si (Gjorgjina Cenikj), tome.eftimov@ijs.si (Tome Eftimov), barbara.korousic@ijs.si (Barbara Koroušić Seljak)

Keywords: text mining, relation extraction, named entity recognition, named entity linking, food-disease relations

1. Introduction

By a definition of the UN Food Systems' Scientific Group (Joachim von Braun et al., 2021), *food systems embrace the entire range of actors and their interlinked value-adding activities involved in the production, aggregation, processing, distribution, consumption, and disposal (loss or waste) of food products that originate from agriculture (incl. livestock), forestry, fisheries, and food industries, and the broader economic, societal, and natural environments in which they are embedded.* In order to understand such a complex concept, food systems' boundaries, as well as their building blocks and linkages among them, while simultaneously being connected to systems such as health, ecologyclimate, economicalgovernance, and scienceinnovation systems, need to be formalised. Once this is achieved, various approaches, including computer-based techniques, can be applied to start searching for relevant relations between these systems.

Resolution of issues related to healthcare is made possible by the existence of several available biomedical vocabularies and standards, which play a crucial role for understanding health information, together with a large amount of health data. However, in 2019, the Lancet Planetary Health noted that the focus of future improvements in our well-being and societies will depend on investigating the links between food systems and other systems including health systems as presented above. Despite the large number of available resources and work done in the health domain, there is a lack of resources that can be utilized in the food and nutrition domain, as well as their interconnections. In particular, this has become highly relevant since the pandemics of COVID-19 (Eftimov et al., 2020), when food provision and security, as well as healthy nutrition and environment, are tremendously needed for quick recovery and long-term sustainable development of our societies.

In this paper, we focus on the formalisation of food systems in relation to health systems because there are still many open research questions and gaps in evidence on how to transform food systems so that they benefit both human nutrition and health. Functional relationships among food systems and health systems need to be systematically specified, and this can be done either in manual or (semi-)automated ways. Systematic specification of a system requires a definition of all entities that relate to expressions used by domain experts to describe knowledge. Once the entities are specified, they can be used to infer relationships between entities that make sense based on the data context.

For example, in (Schoeneck & Iggman, 2021) there is a statement in the abstract saying '*With high evidence, foods high in unsaturated and low in saturated and trans fatty acids (e.g. rapeseed/canola oil), with added plant sterols/stanols, and high in soluble fiber (e.g. oats, barley, and psyllium) caused at least moderate (i.e. 0.20-0.40 mmol/L) reductions in LDL cholesterol.*'. In this textual data, the expressions *rapeseed/canola oil, oats, barley*, and *psyllium* need to be related to food entities (as specified by a selected food-related semantic resource), and the expression *LDL cholesterol* has to be related to a health entity (i.e., diseases as specified by a selected biomedical semantic resource). Finally, relationships between the identified entities can be inferred (e.g., between unrefined grains and diseases related to LDL cholesterol). However, to make the relationship even more useful, it needs to be defined whether the food entities cause or treat the related diseases. Having a (semi-) automatic system that will be able to extract such knowledge is extremely welcome to dynamically follow the new knowledge that will allow medical professionals to stay up-to-date.

To help dietitians and medical professionals stay up to date and relieve domain experts from the burden of manual work, a pipeline for (semi-)automated information extraction (IE), as an example of a computer-assisted approach, can immensely speed up the process of relation extraction (RE) and enable constant updating. In the food and nutrition domain, scientific literature is an important source that requires the extraction of structured information from unstructured text. Notwithstanding the pipeline's utility to extract and structure the textual

information into meaningful relations, the relations extracted by such a pipeline should be checked by domain experts since the extraction is done by utilizing artificial intelligence (AI) approaches that are stochastic in nature and are not able to extract all information.

Our contribution: In this paper, we introduce a novel IE pipeline for semi-automatic mining of scientific literature for findings that will support the existence of both causal and treatment relations between food and disease entities, known as FooDis. To the best of our knowledge, this is the first attempt in RE, where entities related to food from all food categories have been considered. The novelty of the FooDis pipeline is that it can identify both causal and treatment relations between **food entities** specified in seven knowledge bases containing information from the food and nutrition domain, and **disease entities** specified in seven knowledge bases from the biomedical domain. The full list of knowledge bases to which the pipeline can link the entities can be found in Section 3. The relations are established solely on the basis of raw text from research abstracts, without any additional data that is available in the various resources, meaning that the pipeline is not aiming to derive its own conclusions through the help of Machine Learning (ML) models, but only to structure the information from the text that can further be explored by domain experts. The difference with other state-of-the-art IE pipelines is that they can not be utilized since the food domain is low-resourced with semantic resources compared to the biomedical domain.

The outline of the paper is structured as follows: Section 2 presents the related work; Section 3 describes all steps from the FooDis pipeline; Section 4 is split into three parts with regard to the evaluation i.e., providing descriptive statistics about the extracted relations from a large corpus of abstracts, ground-truth evaluation on a small manually annotated corpus of abstracts, and comparing the extracted relations with two baselines that contain food-disease relations; discussion related to the novelties in the proposed pipeline and directions for future work are presented in Section 5; finally the conclusions are presented in Section 6.

Table 1: Events that support biomedical relations extraction.

Event	Year	Focus
BioCreative VII	2020/2021	Drug-Protein relations
BioCreative VI	2017	Chemical-Protein relations
BioCreative V	2015	Chemical-Disease relations
DDIExtraction	2011/2013	Drug-Drug relations
BioNLP	2011	Gene-Gene relations
i2b2	2010	Clinical relations
BioCreative III	2010	Protein-Protein relations
BioCreative II	2006	Protein-Protein relations

2. Related Work

In the last two decades, a great effort has been done in developing various IE pipelines for the biomedical domain, where the focus was on genotype and phenotype entities, together with health-related entities such as disease, treatments, drugs, etc. For this purpose, a lot of shared workshops have been organized as part of conference events such as BioNLP (Nédellec et al., 2013), BioCreative (Leitner et al., 2010), i2b2 (Sun et al., 2013), and DDIExtraction (Segura-Bedmar et al., 2011) with different focuses.

Table 1 presents a list of events related to biomedical RE topics through the years. Using the table, it is obvious that the focus is on biological and clinical entities, where the food entities are not even included. In the domain of food and nutrition, several efforts have been made toward the structuring of relations between food and disease entities. Few of them Yang et al. (2011); Miao et al. (2012); Ben Abdessalem Karaa et al. (2018) are focused on classical text mining approaches where based on hand-written features sentiment analysis are performed. However, ground truth food-disease relation corpora do not exist.

DietRx(https://cosylab.iiitd.edu.in/dietrx/index) is an example of a platform that provides insight into various relations not only between food and diseases, but also among chemicals and genes. Food entities are organized into

25 food categories and linked to the National Center for Biotechnology Information Taxonomy (NCBIT) (Federhen, 2011), while the disease entities are extracted using the MEDIC disease vocabulary (Davis et al., 2012) and are associated with Medical Subject Headings (MESH) (Rogers, 1963) identifiers. However, despite the great effort of structuring knowledge in such a platform, the weakness is that the collected data are static, requiring regular updates to follow the new knowledge appearing with new scientifically published papers. The update process is a time-consuming task, which involves a lot of human effort to find the relevant papers and then to extract all relevant information related to food-disease relations.

There exists the NutriChem database (Jensen et al., 2014; Ni et al., 2017), which links plant-based foods with their small molecule components, drugs and disease phenotypes using an automated approach of text mining MEDLINE abstracts. However, NutriChem limits its scope to plant-based foods and does not cover the complete range of relevant food categories.

3. Food-Disease Mining Pipeline

The proposed FooDis pipeline can extract causal and treatment relations between **food entities** specified with respect to the state-of-the-art semantic resources, like Food Database (FooDB) (https://foodb.ca/), Systematised Nomenclature of Medicine Clinical Terms (SNOMEDCT) (Donnelly, 2006), the Hansard Corpus (Alexander & Anderson, 2012), the Integrated Taxonomic Information System (ITIS)(https://www.itis.gov/), Wikipedia(https://en.wikipedia.org/) articles and the National Center for Biotechnology Information Taxonomy (NCBIT) (Federhen, 2011), and **disease entities** specified in the Disease Ontology (DO) (Schriml et al., 2018), SNOMEDCT, the Unified Medical Language System (UMLS) (Humphreys et al., 1998), the National Cancer Institute thesaurus (NCIT) (Fragoso et al., 2004), the Online Mendelian Inheritance in Man (OMIM) database (Hamosh et al., 2000), the Experimental Factor Ontology (EFO) (Malone et al., 2010), and the Medical Subject Headings (MESH) vocabulary (Rogers, 1963). More

details about the above-mentioned semantic resources are presented in Appendix A.

Figure 1 depicts an overview of our proposed pipeline for semi-automatic food-disease relation mining. It works with textual data that in our case is a scientific abstract. It consists of four steps, each one described below:

- Food and disease entities extraction, where food and disease entities are automatically extracted from the textual data using named-entity recognition (NER) methods. Next, to obtain unique identifiers, food and disease entities normalization is performed, where the extracted entities are linked to biomedical and food semantic resources by using named-entity linking (NEL) methods.

- Sentence relevance filtering, where the goal is to select relevant sentences that should be included for RE. It works on two levels. First, select only those sentences from the abstracts that contain information about previously established facts or analyses of the conducted research. Second, select only a subset of the fact or analysis sentences that contain at least one pair of food and disease entities.

- Learning relation classification models that are able to classify a relation between food and disease entities as *treat* or *cause*.

- By using the evidence from the relation extraction models which is the number of different sentences from which the same relation is extracted, the final relations are defined. The relations supported by a higher number of sentences are considered to be the more relevant.

  In the remainder of this section, we are going to present each step of the FooDis pipeline in more detail.

3.1. Food and Disease Named Entity Recognition and Linking

Next, we are going to provide information for food and disease NER and NEL methods that are used to extract the information related to them.

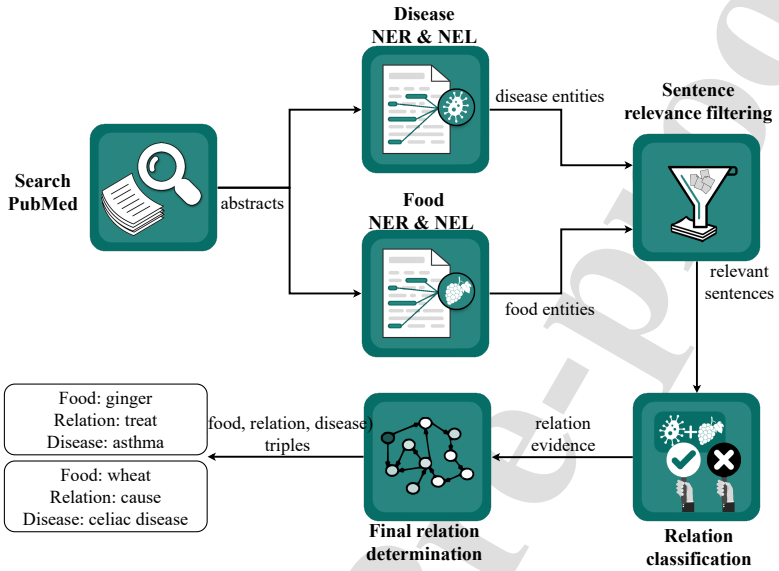


Figure 1: The full FooDis pipeline.

3.1.1. Disease named entity recognition

To extract disease entities, the *DISO* pre-trained model is used, which is provided by the Sequence Annotator for Biomedical Entities and Relations (SABER) (Giorgi & Bader, 2019) tool. This model is capable of identifying the following entity types, which we refer to as diseases as an umbrella term: *Acquired Abnormality, Anatomical Abnormality, Cell or Molecular Dysfunction, Congenital Abnormality, Disease or Syndrome, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Sign or Symptom.*

The SABER is a tool for biomedical NER, which combines several strategies for improving the generalization ability of the BiLSTM-CRF architecture: the inclusion of variational dropout for increased regularization of the recurrent layers, multi-task learning for joint training of the hidden

layers with different corpora, and transfer learning (TL) implemented by first training on a large, semi-automatically generated silver corpus, and continuation of the training with a gold corpus (Giorgi & Bader, 2019).

3.1.2. Disease named entity linking

The SABER tool also has integrated support for the NEL task, performed with a dictionary-based tagger built on top of the DISEASES database (Pletscher-Frankild et al., 2015). This allows us to link the recognized entities to the DO (Schriml et al., 2018), which in turn, provides ids to SNOMEDCT, UMLS, NCIT, OMIM, EFO, and MESH. This means that only a subset of the entities linked to the DO can be linked to the other disease resources, i.e. only the ones that are linked to entries in the DO.

3.2. Food named entity recognition and linking

3.2.1. Food named entity recognition

To extract the food entities from the abstracts, we employ several food NER methods. BuTTER (Cenikj et al., 2020) and FoodNER are corpus-based food NER methods trained on the FoodBase corpus (Popovski et al., 2019). BuTTER is based on the BiLSTM-CRF architecture that has been commonly used for the NER task, while the second one leverages fine-tuning of BERT models (Devlin et al., 2018a). Apart from identifying food entities from raw text, FoodNER is also capable of annotating the entities with semantic concepts from the Hansard corpus (Alexander & Anderson, 2012), the FoodOn (Dooley et al., 2018) and SNOMEDCT (Donnelly, 2006) ontologies.

We choose to use these methods since they are more robust than the rule-based methods that precede them and are not dependent on the availability of external resources, such as ontologies, taxonomies and web taggers.

However, since these models are trained on recipe texts, their application on biomedical text introduces challenges due to the domain-specific vocabulary and the difference in style, which is manifested in the form of increased amount of false positive entities. To overcome these challenges and obtain more reliable results, we combine the annotations of the following models:

- FoodNER - We use the FoodViz platform (Stojanov et al., 2020) 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.
- BuTTER - We use the lexical lemmatized BiLSTM-CRF architecture without character embeddings, which achieved the best results in terms of the averaged macro F1 score in (Cenikj et al., 2020).
- FooDB non-scientific - We perform simple string matching using the common names of 992 foods from the FooDB database.
- FooDB scientific - Due to the fact that biomedical texts commonly refer to food entities using their scientific names, which are not extracted by the other models, we also perform string matching using the 675 scientific names of foods available in the FooDB database.

The combination heuristic or ensemble will be further presented. First, we are going to explain the food NEL methods that are also used in the ensemble.

3.2.2. Food named entity linking

The NEL task is accomplished with the FoodNER and the FooDB methods. FoodNER models can be used to link the extracted food entities to semantic concepts from the Hansard corpus (Alexander & Anderson, 2012), the FoodOn (Dooley et al., 2018) and SNOMEDCT (Donnelly, 2006) ontologies. The entities can be linked to 205 different semantic tags from the FoodOn ontology and 207 tags from the SNOMEDCT ontology.

The linking to the Hansard corpus can be accomplished using two strategies. The first one involves determining the first parent of the food phrase in the Hansard hierarchy, in which case the entities can be categorized to 48 different classes. We refer to the annotations extracted in this manner as the *Hansard-Parent* tags. The second strategy involves linking the entity to the closest tag from the hierarchy, determined using the minimum cosine distance between the BERT embedding of the food phrase and each of the Hansard tags. In this case, the entity can be assigned to 92 different tags.

The NEL done using the FooDB methods (FooDB non-scientific and FooDB scientific) is possible because these methods perform simple string matching, and every extracted food phrase can be completely matched to the entries in the database. Additionally, since the FooDB contains mappings of the food concepts to Wikipedia, ITIS, and NCBIT, once the extracted food phrase is matched to FooDB, it can also be linked to these resources. Out of the 992 total food concepts in FooDB, 838 have links to Wikipedia articles, 601 are linked to ITIS, and 600 are linked to NCBIT.

3.2.3. Food named entity recognition ensemble

For the combination heuristic for food named entity recognition, we have proposed a voting strategy. The voting strategy is used to combine the entities extracted using BuTTER, FoodNER, FooDB non-scientific, and FooDB scientific method.

A token extracted by any of the BuTTER, FoodNER, and FooDB non-scientific 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.

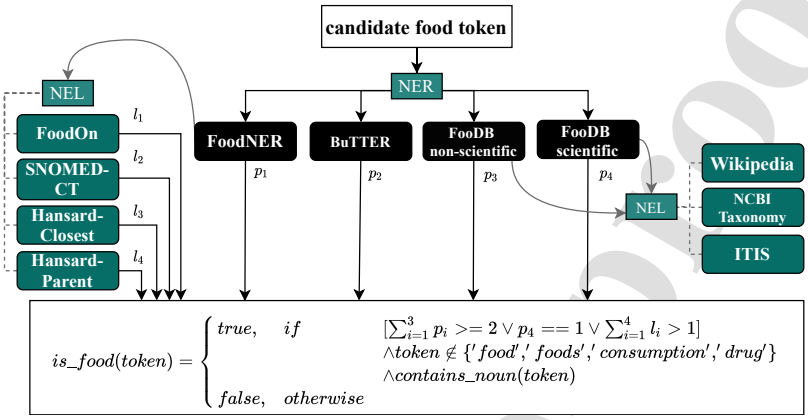


Figure 2: Voting scheme for the food NER task.

The flowchart of the voting scheme is presented in Figure 2. The 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. In addtion, the 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, respectively).

Finally, we perform post-processing rules, which involves removing a few food-related words (*food, foods, consumption*, and *drug*) that are too general to be useful, and entities that do not contain any nouns, since these are likely to be false positives.

3.3. Sentence relevance filtering

The pipeline extracts information from single sentences that contain at least one food and one disease entity, which are non-overlapping. Such sentences are retrieved from the abstracts, and are used for determining the relation between the food and disease entities.

Abstracts typically describe the topic, objective, hypothesis, methodology, and main findings of the research. However, not all of these pieces of information are a reliable source for drawing conclusions, since if the authors' hypothesis was untrue, and we were to extract information from the sentence that describes that hypothesis, our findings would be incorrect. For this purpose, it is necessary to distinguish these sentences and filter out the irrelevant ones.

A relevant sentence is one that contains at least one pair of food and disease entities, and expresses a previously established fact or a finding of a study. To find them, we have trained a binary classifier that can classify a single sentence from each abstract either as relevant or irrelevant.

The classifier is trained using the GENIA Meta-knowledge event corpus (Thompson et al., 2011). The corpus is a collection of MEDLINE articles, annotated with meta-knowledge annotations, such as the general information type of the event (whether it is a fact, experimental result, investigation, etc.), the level of certainty, the polarity of the event (positive or negative), etc. To train the classifier for sentence relevance, we use the *Knowledge Type* annotation, which can take one of the following values:

- *Investigation* (764 samples): Planned or already conducted investigations and inquiries, often accompanied by lexical clues like *examined, investigated* and *studied*
- *Observation*: Direct observations, expressed with clues such as *found*, and *observed*
- *Analysis* (2,876 samples): Inferences or interpretations typically accompanied by *suggest, indicate* or *conclude*.
- *Method* (524 samples): Events that describe experimental methods
- *Fact* (1,444 samples): Events that describe established knowledge
- *Other* (4,814 samples): Events that do not belong to any of the previously described categories

Even though the corpus is annotated with events, we disregard the annotations related to events, because we work at a sentence level. Since a sentence might contain several events, we extract unique, individual sentences that have the *Knowledge Type* annotation, and only use the raw text, without any event annotations.

We train a binary classifier to distinguish between relevant sentences (ones that are annotated with *Analysis* or *Fact*) and irrelevant sentences (annotated with *Investigation* or *Method*). We do not include the *Other* class, since its meaning is rather ambiguous, or the *Observation* class, since no samples were found with this class. Merging the samples from the relevant and irrelevant classes results in 4,320 positive and 1,288 negative samples.

For training the classifier, we perform end-to-end fine-tuning of a BERT model (Devlin et al., 2018b), pre-trained on the BooksCorpus and Wikipedia. The last layer of the BERT model is replaced with a dropout and a linear layer which performs binary classification. During fine-tuning, the model parameters are initialized with the values from the pre-training step, and are fine-tuned using the labeled data from the GENIA Meta-knowledge corpus. When it comes to the choice of this model over other transformer-based models, we were guided by our previous experience that other BERT-variants (for instance, RoBERTa or BioBERT) may give slightly better results than the original BERT model for this task, however, we do not believe that this will have a big impact on the pipeline's overall performance. The summary of the sentence relevance filtering procedure is presented in Figure 3.

The model is trained using stratified 10-fold cross-validation, where 10% of the training portion of each fold is removed and used for validation. The AdamW optimizer is used with a learning rate of $4 * 10^{-5}$. An early stopping strategy is applied to prevent overfitting. The model is trained for a maximum of 10 epochs, or until the decrease in validation loss of two consecutive epochs did not surpass $5 * 10^{-3}$.

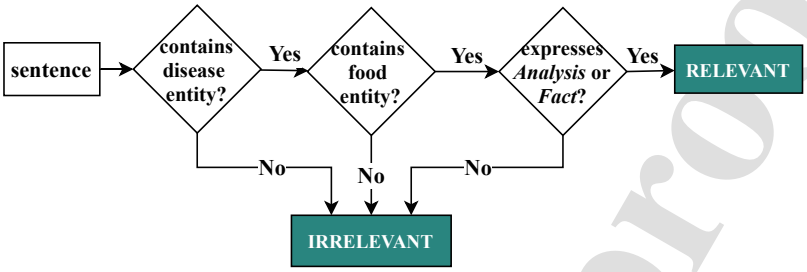


Figure 3: Sentence relevance filtering pipeline.

Due to the imbalanced nature of the dataset, the model is evaluated using the macro-averaged F1 score. Since the model is used to find relevant sentences which will be used for IE by the next components in the pipeline, we are also interested in the precision with which the model identifies the positive class. The averaged macro averaged F1 scores from the 10 folds is 0.81, while the averaged precision for the positive class is 0.90. The model used in practice is trained on 90% of the whole data, with the remaining 10% being used for validation.

3.4. Relation classification

The relevant sentences that contain at least one food and one disease entity are used for determining whether the food entity causes the disease, is used for its treatment, or neither. To accomplish this, we represent the RE task as a binary classification problem, meaning that we use separate classifiers that detect the presence of each relation type (i.e., *cause* and *treat*). This implementation was impeded by the fact that there was no annotated data in the food domain for this specific task.

To train the RE classifiers, we have extended our previous work done in (Cenikj et al., 2021b). There, we have developed the SAFFRON models for RE. Those models are trained using transfer learning, which involves improving a learner from one domain by transferring information

from a related domain (Weiss et al., 2016; Zhuang et al., 2019). The models were trained on data that is annotated for the existence of *cause* and *treat* relations between different types of biomedical entities in the CrowdTruth dataset (Dumitrache et al., 2017, 2015b,a). The occurrences of the biomedical entities in each annotated sentence are masked to prevent the models from learning relations between specific entities, and teach them to recognize relations based on the context words used to express the relation, so they can successfully generalize to the task of recognizing the relations between food and disease entities. Further, the models were evaluated to detect relations between food and disease entities in the FoodDisease dataset (Cenikj et al., 2021a).

The extension in this work is going beyond a limitation of the SAFFRON models. Actually, the SAFFRON models are trained to recognize *cause* and *treat* relations, assuming that the evaluated sentence expresses a fact. The models are not evaluated on sentences which might not express facts, for instance, sentences that express hypotheses, so they cannot distinguish between the sentence: "We hypothesize that excessive salt intake increases the risk of heart disease" and "It has been shown that excessive salt intake increases the risk of heart disease". In fact, because of the specific preprocessing step, which is referred to as *Context extraction* in the original SAFFRON paper, it is possible that the models get the exact same input for the two sentences which were previously given as examples and will produce the same output. The sentence relevance filtering is necessary to make sure that the sentences which express hypothesis are never given as inputs to the SAFFRON model, so that we avoid the previously described issue.

Here, we choose to use the Single Sequence Classifier (SSC) introduced in (Cenikj et al., 2021b), over the Sequence Pair Classifier (SPC), since it provides consistently good results and lower time complexity. We select the SSC models trained by fine-tuning BioBERT (Lee et al., 2019)

Table 2: A macro averaged F1 score of the RE classifiers.

Model	Dataset	Treat	Cause
BioBERT	CrowdTruth	0.87	0.80
RoBERTa	CrowdTruth	0.88	0.80
BioBERT	FoodDisease	0.87	0.84
RoBERTa	FoodDisease	0.88	0.71

and RoBERTa (Liu et al., 2019) models to perform the RE task on the CrowdTruth and FoodDisease datasets. The datasets were selected, since they are annotated for the existence of both the *cause* and the *treat* relation, unlike the models trained on the Adverse Drug Effects dataset, which can only identify the *cause* relation. The macro averaged F1 scores for the RE models are presented in Table 2.

On each relevant sentence, we apply eight classifiers (i.e., four combinations (RoBERTa, CrowdTruth), (BioBERT, CrowdTruth), (RoBERTa, FoodDisease), (BioBERT, FoodDisease) trained for each relation type, *cause* and *treat*, separately). Next, these predictions were combined to find a final one. The predictions are combined with a voting scheme, presented in Figure 4. For each *(food, disease, sentence)* triple, four models are applied for the identification of both the *cause* and the *treat* relation, producing binary predictions, c_1, c_2, c_3, c_4, and t_1, t_2, t_3, t_4, respectively. In order a relation to be accepted as positive, at least M out of the four RE models that identify the relation need to generate a positive prediction, and a maximum of N RE models that identify the opposite relation are allowed to generate a positive prediction. Here, M and N are input parameters to the pipeline.

In our experiments, we set M to three, and N to one. This means that for a given sentence to express a *cause* relation, three out of four RE models trained for the *cause* relation should predict that the *cause* relation is present, and at the same time, only one out of four RE models used for

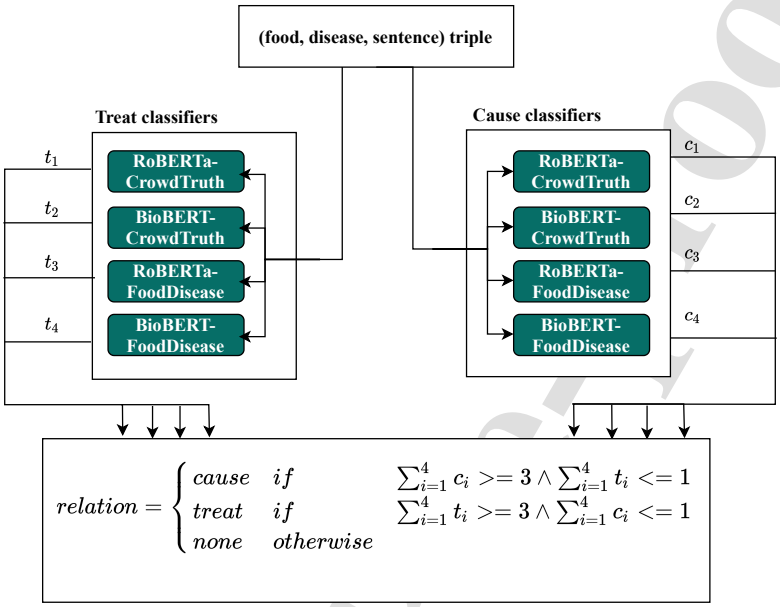


Figure 4: Voting scheme for the RE task.

the *treat* relation can predict the same sentence as a *treat* relation. The vice-versa is also true, where the *cause* and the *treat* relations swap. Each (food, disease, sentence) triple gets assigned one of the labels *cause*, *treat*, or *none*. We refer to each triple with a *cause* or *treat* label as evidence for the existence of that relation between the food and disease entities.

3.5. Combining evidence

To generate the final relation of each food-disease pair, we take into account all of the evidence for that pair. We assign a relation to each pair only if there are at least X sentences that support that relation and a maximum of Y sentences that support the existence of the opposite relation. X and Y are parameters that can be given as input to the pipeline. In our experiments, we set X to one and Y to zero. These conditions

can be strengthened or relaxed, depending on whether the pipeline will be used for completely automatic IE, or the findings will be curated by experts (i.e., semi-automatic), since the process described in this subsection is essentially a tradeoff between the pipeline's precision and recall.

4. Results

In this section, we present the results obtained in our experiments. The source code is publicly available at (Cenikj et al., 2021c) and can be reused for future studies. We need to mention once more that our sentence relevance model was learned on GENIA Meta-knowledge corpus, while the RE models were trained using the CrowdTruth and FoodDisease corpora. To evaluate the approach, we collected a set of scientific abstracts that are used only for testing purposes of the whole FooDis pipeline.

4.1. Data collection used for evaluation

Here, we provide details about the implementation and the heuristic behind the process of collecting the data that is used for testing and evaluating the FooDis pipeline. We need to mention here that this data is used only for evaluation purposes and not learning the models that are part of the FooDis pipeline.

4.1.1. Implementation details and search terms

The scientific paper abstracts involved in the evaluation are collected from PubMed using the Entrez Programming Utilities (Sayers, 2010). The ESearch utility produces a set of unique identifiers of papers corresponding to an input search term. It also allows the specifying the maximal number of identifiers that are going to be returned, which is done by setting the *retmax* parameter in the API call. The EFetch API call is then used to retrieve the paper data for each of the identifiers returned by the ESearch

utility. Apart from the abstract, the EFetch utility can be used to collect other information such as the paper title, the year the paper was published, the journal it was published in (in case of conference publications, we do not have an information about the source), as well as MeSH terms related to it. The ESearch and the EFetch APIs are used only for the data acquisition process in the FooDis pipeline, to collect data further involved to train the food-disease relation models.

After several discussions with a domain expert, the following 17 phrases are used as search terms for obtaining the initial set of abstracts: *asthma food, arthritis food, parkinson disease food, bronchitis food, stroke food, food allergy, heart disease food, diabetes food, kidney stone food, anemia food, osteoporosis food, pneumonia food, alzheimer food, skin disease food, tuberculosis food, hypertension food, influenza food.* We need to mention here, that these search terms were used only as a use case to collect data to show the utility of the pipeline. The definition of the search term is a subject of information retrieval methodologies that are not focus in this subject.

4.1.2. Data collection heuristic

The collection phase of the FooDis pipeline consists of two phases: an initial one and one to acquire more evidence for each extracted relation.

Initial phase: The first phase is the initial phase where the whole FooDis pipeline is executed on the data collected with initial search terms described above (14,712 abstracts). Here, we limit the number of abstracts to a maximum of 1,000 for each search term, which is done by setting the *retmax* parameter in the ESearch API call. This results in the collection of 14,712 abstracts of papers from 3,035 distinct journals. We need to note here that the terms searched in the initial stage consist of "food" as a general concept and do not distinguish between different food items.

Additional phase: Once the whole pipeline is executed on the initial

data and relations are established between certain food and disease entities, the second phase is performed. In this phase, we repeat the entire procedure four times, where the search terms are obtained by concatenating each extracted food-disease pair from the previous iteration. Let us assume that some of the extracted (food, relation, disease) triples when the whole pipeline is executed on the initial data are ("peanut", "cause", "allergy"), and ("pork", "cause", "cardiovascular disease"). Further, for each extracted triple, the ESearch and EFetch APIs are used to search and retrieve abstracts related to the more detailed relations. In this instance, "peanut allergy" and "pork cardiovascular disease" would be used as search keywords, without the word "cause". The pipeline would then retrieve 100 additional paper abstracts for each search keyword, to find additional evidence to support the *cause* relation, or possibly even find opposing papers claiming a *treat* relation. The reasoning behind this is that we aim to have each relation supported by several sources, to account for the possible unreliability of the source and the errors made by the pipeline components.

In each of the four iterations, apart from gathering more evidence for the already extracted relations, the pipeline identifies relations between novel pairs of food and disease entities, thus, in each iteration, new keywords are used to search the PubMed database, and retrieve novel abstracts. This allows the pipeline to both confirm the validity of the relations for which it has insufficient evidence, and extend the scope of the search for new relations. In each subsequent data collection phase, we limit the number of retrieved abstracts per search term to 100.

4.2. Extracted relations

Table 3 presents the relation extraction results from both phases separately. The column *Search terms* contains the number of search terms that were used to search the PubMed database in each phase. The col-

umn *Retrived abstracts* indicates the total number of abstracts of scientific papers that were retrieved from the PubMed database with all of the keywords used in that phase. The *Distict journals* column contains the number of different journals where the retrieved papers were published. The processed abstracts can also come from conference publications, however, for such abstracts we do not have information about the name of the conference in which the paper was published.

Table 3: Extraction results.

	Search terms	Retrieved abstracts	Distinct journals
Initial phase	17	14,712	3,035
Additional phase	1,387	65,229	6,515

In total, 79,941 abstracts are retrieved from all phases. All of the abstracts are taken to the next steps involving the application of NER and NEL models. However, only 8,756 abstracts contain sentences in which both a food and a disease entity are mentioned. There are 82,273 such sentences, out of which 18,438 are removed in the sentence relevance filtering step due to the additional filtering conditions requiring the sentence to express a *Fact* or *Analysis*. This means that 63,835 sentences get to the *Relation Classification* step.

In total, 931 *cause* and 2,059 *treat* relations are extracted between 674 unique food entity mentions and 923 unique disease mentions, some of which are not linked to any semantic resource. The *cause* relations are supported by 6.78 evidence sentences on average, while the *treat* relations are supported by an average of 6.70 evidence sentences.

The number of unique food-disease entity pairs that can be linked to the various resources and between which a *cause* or a *treat* relation was established are listed in tables 4 and 5, respectively. The numbers presented in these two tables are meant to inform the creators of each semantic resource

Table 4: Number of unique pairs linked with a *cause* relationship for each combination of food and disease resource.

	DO	UMLS	NCIT	OMIM	EFO	MESH	SNOMED CT
FoodOn	167	157	120	70	53	142	152
SNOMED CT	159	153	116	74	57	138	149
Hansard Closest	222	207	151	100	67	187	202
Hansard Parent	221	206	150	101	68	186	201
FooDB	148	140	109	66	50	122	134
ITIS	80	75	61	41	31	65	70
Wikipedia	139	131	103	63	47	113	125
NCBIT	82	77	63	43	33	66	72

utilized in this study about the number of links that can be curated by experts and added to the resource. There are an average of 112.94 *cause* relations where the entities are linked to existing semantic resources. In case of the *treat* relation, there are 296.82 such relations extracted on average.

A few relations examples extracted between the food entities in FooDB and disease entities in the EFO are presented in Table 6. Apart from the information given in this table, the pipeline can also provide the evidence for each relation, i.e. all of the sentences where the relation was identified and the corresponding abstracts to which these sentences belong to.

In order to get additional insight into the words used to describe the two relations, we conduct an additional analysis of the verbs that appear in the sentences from which the pipeline extracted the relations. To this

Table 5: Number of unique pairs linked with a *treat* relationship for each combination of food and disease resource.

	DO	UMLS	NCIT	OMIM	EFO	MESH	SNOMED CT
FoodOn	280	269	251	101	103	257	264
SNOMED CT	352	339	320	133	141	329	334
Hansard Closest	438	424	397	165	165	407	418
Hansard Parent	438	423	394	167	167	404	415
FooDB	503	484	432	173	184	464	472
ITIS	314	300	268	105	107	290	293
Wikipedia	423	405	363	144	147	389	396
NCBIT	313	299	267	104	107	289	292

Table 6: Examples of relations extracted by the pipeline.

Food name	FooDB identifier	Disease name	EFO identifier	Relation
barley	FOOD00088	celiac disease	0001060	cause
wheat	FOOD00561	celiac disease	0001060	cause
peanut	FOOD00016	eosinophilic esophagitis	0004232	cause
barley	FOOD00088	asthma	0000270	cause
garlic	FOOD00008	hypertension	0000537	treat
olive oil	FOOD00909	type 2 diabetes	0001360	treat
sweet potato	FOOD00092	renal cell carcinoma	0000681	treat
grapefruit	FOOD00256	parkinson's disease	0002508	treat
ginseng	FOOD00219	type 2 diabetes	0001360	treat
ginger	FOOD00206	obesity	0001073	treat
ginger	FOOD00206	migraine	0003821	treat
ginger	FOOD00206	asthma	0000270	treat

end, we use the spacy spa python library to extract the verbs that appear between the food and disease entities in each sentence describing one of the *cause* and *treat* relations. We then perform lemmatization of the extracted verbs to bring them to a normal form. The following 50 verbs are the most frequently used to describe the *cause* relation: *cause*, *trigger*, *be*, *gluten*, *increase*, *lead*, *report*, *contain*, *develop*, *result*, *include*, *induce*, *characterize*, *mediate*, *have*, *find*, *occur*, *ingest*, *know*, *associate*, *contaminate*, *affect*, *precipitate*, *drink*, *implicate*, *relate*, *contribute*, *sneeze*, *expose*, *show*, *continue*, *predispose*, *involve*, *transmit*, *consider*, *flush*, *decrease*, *eat*, *believe*, *threaten*, *complain*, *diagnose*, *bring*, *follow*, *suffer*, *thrive*, *produce*, *provoke*, *exacerbate*, *experience*, *guarantee*. For the *treat* relation, the following 50 verbs are most frequently used: *use*, *reduce*, *have*, *treat*, *include*, *show*, *be*, *prevent*, *associate*, *improve*, *report*, *cure*, *decrease*, *induce*, *know*, *lower*, *relate*, *find*, *protect*, *contain*, *inhibit*, *help*, *suggest*, *consume*, *exhibit*, *increase*, *prove*, *leave*, *demonstrate*, *base*, *possess*, *promote*, *recommend*, *attenuate*, *exert*, *derive*, *ameliorate*, *see*, *mitigate*, *diffusa*, *propose*, *reverse*, *develop*, *belong*, *offer*, *consider*, *play*, *claim*, *suppress*, *come*.

4.3. Evaluation on ground truth data

The previous results only provide statistics when the pipeline is used for extracting relations from a large corpus of abstracts that are not annotated and do not represent a ground truth corpus. To have more insights into the performance of the FooDis pipeline, we have manually annotated 125 randomly sampled abstracts for the existence of (food, relation, disease, sentence) tuples. We have done this since there is no annotated data provided by domain experts on which we could evaluate the results of the pipeline. The annotations of the relations have been performed by three domain experts. The majority vote of the labels provided by all three annotators has been taken as a final result. We need to mention here that for all relations the annotators agree on the label of the relation. We need to point out that when we create the ground-truth data, we annotate

each relation that is supported only by one sentence (i.e., a single piece of evidence) as a ground truth. However, the evaluation is done using the *Combining evidence* step, since this step takes into account evidence from multiple sentences.

4.3.1. Evaluation scenarios

To analyze the types of errors produced by the FooDis pipeline, its performance is evaluated using four evaluation scenarios. This will further provide us some insight into the performance that can be achieved by the pipeline and where the errors happen. This kind of insight can guide decision-making in the organization of the curation process, that is, how much effort is expected to be required for correcting the relations produced by the pipeline, and the mistakes made by the NER and NEL models.

The four evaluation scenarios are defined based on whether partial matches of entity mentions (i.e., not full name of the entity is extracted) are considered as correct:

- Scenario 1 - A (food, relation, disease, sentence) extraction is considered to be correct, if both of the entities are fully extracted as in the ground truth annotation, and the prediction of the extracted relation and sentence match the annotated ones.
- Scenario 2 - A (food, relation, disease, sentence) extraction is considered to be correct, if the prediction of the extracted relation and sentence match the annotated ones, the disease entity is fully extracted, and the food entity can be either fully or partially extracted (some words can be added or missing, i.e. the annotated entity can be a sub-string of the extracted entity, or the extracted entity can be a sub-string of the annotated entity).
- Scenario 3 - A (food, relation, disease, sentence) extraction is considered to be correct, if the prediction of the extracted relation and

sentence match the annotated ones, the food entity is fully extracted, and the disease entity can be either fully or partially extracted.

– Scenario 4 - A (food, relation, disease, sentence) extraction is considered to be correct, if the prediction of the extracted relation and sentence match the annotated ones, and both the food and the disease entity can be either fully or partially extracted.

4.3.2. Results and sensitivity analysis

To perform sensitivity analysis on the results based on the parameters that are used in the voting scheme in the RE model, we evaluate the pipeline using different values for the configuration parameters M and N. More specifically, M can take the values of two, three, and four, while N can take the value of zero or one. We believe that these values are the most reasonable for the two parameters. Figure 5 depicts the precision, recall, and F1-scores of each pair of values for the M and N parameters, indicated on the x-axis. The green plots refer to the precision metric, the pink plots refer to the recall metric, while the blue plots refer to the f1 score.

In the most rigorous evaluation scenario, the first one, the maximal precision achieved is 0.69, while the maximal recall is 0.36. According to the most flexible evaluation scenario, the fourth one, the maximal precision is 0.80, while the maximal recall is 0.47. These results indicate the performance in the case where only a single piece of evidence (sentence) is considered for each relation. The consideration of multiple pieces of evidence (sentences) for determining the final relation between a pair of food and disease entities, as explained in the *Combining evidence* step is expected to further improve these results.

Figure 5 shows that the precision metric is proportional to the M parameter, while the recall metric is inversely proportional. Each increase of the M parameter results in an increase of the precision metric by at least

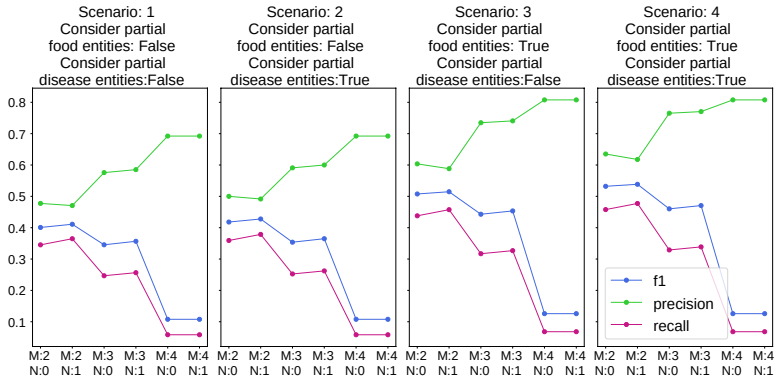


Figure 5: Evaluation of the FooDis pipeline with different configuration settings and evaluation strategies. The M parameter represents the minimum number of RE models which need to generate a positive prediction for the existence of the relation, in order for that relation to be accepted as positive. The N parameter represents the maximum number of RE models which are allowed to generate a positive prediction for the opposite relation, when a relation is accepted as positive.

0.10, and a corresponding decrease to the recall metric. The N parameter is proportional to the recall metric, and inversely proportional to the precision metric, however, this relation is less pronounced. Regardless of the evaluation strategy used, the highest precision is achieved with the configuration settings $M = 4$, $N = 0$ and $M = 4$, $N = 1$, while the highest recall is achieved with the configuration settings $M = 2$, $N = 1$. In terms of the F1 score, the configuration setting $M = 2$, $N = 1$, yields the best results.

The choice of configuration parameters depends on the use case, and the availability of experts available for curation. Using the pipeline with the configuration settings that produce a higher recall would extract a larger number of relations, which might be preferable when a large number of experts are available to curate the relations. On the other hand, setting the configuration for a higher precision would provide a higher reliability of the extracted relations, which would minimize the effort required by experts and might be preferable when a smaller number of experts are available for curation, but would provide a lower number of relations.

4.3.3. Directions for further improvement

Comparing the first and second evaluation scenario, we can note that considering partially matched food entities as correct has a much greater positive impact on the performance than considering partially matched disease entities as correct. This is likely an indicator that the food NER methods produce more partial matches than the disease NER methods, and that they may thus need further improvement.

Table 7 presents some examples of the mistakes that the pipeline can make. The presented examples are categorized according to the type of the error, into three categories: incorrect relation, incorrect food entity, and incorrect disease entity. The incorrect relation error type typically occurs in complex sentences containing multiple relations between different food

and disease entities, as in the first example in Table 7. The incorrect food entity error can occur due to various reasons.

In the second example in Table 7, *smoking* is extracted as a false positive food entity. In the third example, the food NER method confuses the name of the animal *turkey* with the name of the country, while in the fourth example, the method misinterprets the organ *liver* as a food entity, since the word *liver* can also be a food entity in other contexts. These are likely consequences of the use of dictionary-based methods, which do not take into account the context in which the entity occurs. However, the corpus-based food NER methods, BuTTER and FoodNER, that do take into account the context, can also produce such types of false positive entities, since they are trained on recipes, which do not typically contain annotations that would help the model learn the difference in the use of these words according to the context, since they almost always refer to food entities in recipe texts. The disease NER methods typically do not produce many false positive entities, however, one instance is presented in the sixth example, where the word *deaths* is extracted as a disease entity. This happens because the biomedical domain has been already well explored and a lot of resources are available to help these processes, which is not the case of the food domain.

4.4. Comparing the relation extraction results with two baselines

We compare the relations extracted by the FooDis pipeline to relations extracted by two existing resources containing food-disease relations, NutriChem (Jensen et al., 2014; Ni et al., 2017) and DietRx(https://cosylab.iiitd.edu.in/dietrx/index) (already described in Section 2). Table 8 provides the number of food-disease relations that are extracted by FooDis and the two existing resources, NutriChem and DietRx.

Table 9 presents the comparison results between FooDis and each of the other resources. The *Overlapping* column presents the number of rela-

Table 7: Examples of errors made by the pipeline

Food entity	Disease entity	Relation	Sentence	Error type
garlic	nausea	treat	Current evidence suggests that Asian ginseng, garlic, tomatoes and soy intake as part of the diet may be useful in preventing cancer; additional research is needed to determine the efficacy of primrose oil and turmeric as cancer treatments; and ginger may be effective in treating chemotherapy-induced nausea.	incorrect relation
smoking	cardiovascular disease	cause	Cardiovascular disease (CVD) events due to atherosclerosis cause one-third of worldwide deaths and risk factors include physical inactivity, age, dyslipidemia, hypertension, diabetes, obesity, smoking, and red meat consumption.	incorrect food entity
turkey	pain	treat	Papaver rhoeas L. (Papaveraceae) corn poppy, widely distributed in Turkey, is used to make a cough syrup for children, as a tea for disturbed sleep, for pain relief and as a sedative in folk medicine.	incorrect food entity
liver	hyperlipidemia	cause	Lipid accumulation in the liver and pancreas is primarily caused by combined hyperlipidemia.	incorrect food entity
red meat	deaths	cause	Cardiovascular disease (CVD) events due to atherosclerosis cause one-third of worldwide deaths and risk factors include physical inactivity, age, ...	incorrect disease entity

Table 8: Number of food-disease relations for each resource.

Resource	Number of relations
FooDis	931
NutriChem	6,246
DietRx	21,207

tions that can be found as the intersection between both resources. The *Equal evidence* column presents the number of relations that are in the intersection, but the NutriChem or DietRx provides the same number of positive (e.g., treat) and negative (e.g., cause) evidence for those relations, so we do not have enough evidence to classify them. This is due to the fact that NutriChem and DietRx provide the number of sentences that support each relation, which we consider as a level of confidence in the truthfulness of the relation. We do not take into account relations for which there is equal confidence in both relation types for the same food-disease entity pair. For instance, if there are 5 sentences that claim that ginger causes heart disease, and 5 sentences that claim that ginger treats heart disease, then we are not considering any of the two relations as true and simply ignore them in the evaluation. The *Accuracy* column presents the accuracy of the extracted relation by looking at their existence in the other resources.

Table 9: Comparison results between FooDis and the other two resources.

	Overlapping	Equal evidence	Accuracy
FooDis - NutriChem	57	2	90%
FooDis - DietRx	209	47	93%

In case of NutriChem, we search the resource by querying the web service by disease concepts (932 diseases extracted by FooDis). After obtaining

all relations for those diseases, the intersection has been selected by linking the food concepts from those relations returned by NutriChem with the 674 food concepts extracted by the FooDis pipeline using exact string matching.

The comparison with DietRx has been done in a similar manner (i.e., in the opposite way), by querying the web service by the food entity name (674 food entities extracted by FooDis), and finding the common disease entities using exact string matching. It is worth mentioning that the RE classifiers used by DietRx, which are not publicly available, have a reported F1 score of 0.84.

The accuracy results of the comparison show really promising results (i.e., 90% and 93% with NutriChem and DietRx, respectively), pointing out that relations that are extracted by the FooDis have enough evidence and also exist in the other resources. The other fact is that FooDis can extract food-disease relations that do not exist in the other resources, which points out that these resources are not frequently updated (i.e., their sustainability can be an issue). On the other side, FooDis can find relations that are reported as new state-of-the-art research results in scientific publications and can be further used as semi-automatic tool for updating already existing resources. Before updating the resources, the relations should be checked by domain experts. The small amount of overlapping between the FooDis and the other two resources is a result only of the limited coverage of the corpora that is analyzed by the FooDis and results on the search terms used to collect the scientific abstracts. Further, we are planning to use the relations that exist in the DietRx and NutriChem as search terms and then make a more comprehensive evaluation between the resources.

5. Discussion

In this section, we discuss the results of the proposed pipeline and the differences to previous attempts at extracting food-disease relations.

5.1. Principal Results

The comparison of the relations produced by the pipeline to the relations in DietRx and NutriChem indicates that the pipeline can suggest relations with a high precision, using the specified parameterization setting for the relation classification step. It is important to note that these parameters can be tweaked in order to perform a trade-off between the pipeline's precision and recall, based on whether the goal is to extract a larger number of relations (which might be preferable when a large number of experts are available to curate the relations) or to achieve a higher reliability of the extracted relations (which would minimize the effort required by experts, at the cost of providing a lower number of relations). The extracted relations can be used to extend and link the resources, with a reduced effort on the experts' part. With the help of the pipeline, the experts need only to check if the entities and the relation have been correctly identified based on the shown textual evidence in a single sentence, instead of reading the entire abstract, identifying the food and disease mentions in the text, linking them to one of the many potentially corresponding entities in the KB, and finally, determining the relation.

5.2. Comparison with prior work

To the best of our knowledge, NutriChem and DietRx are the only two resources which contain relations between food and disease entities extracted in an automated manner.

A critical difference between NutriChem and our pipeline is the fact that NutriChem limits its scope to plant-based foods, while we aim to extract relations from a broader range of food categories, and link them to various resources.

Even though DietRx provides a large amount of extracted relations, it only provides the extracted results through a web service. The details of the implementation are not disclosed, and the employed source code, methods

and datasets are not publicly available, which impedes reproducibility and reusability. Our pipeline, on the other hand, is open-sourced, completely based on publicly available resources, and can be reused for processing new papers as they are being published, or for conducting studies for specific foods or diseases.

Compared to the SAFFRON method, this work has several contributions. While the BuTTER and SABER models were used to generate the dataset on which the SAFFRON models were trained on, the annotations of these two NER models were manually corrected, primarily because of the large amount of false positive entities produced by the BuTTER model, which result from the fact that this model was trained on recipe data, and does not generalize well on scientific text. The ensemble voting scheme implemented in the food NER and NEL component of the FooDis pipeline is a novel contribution, designed to overcome the generalization issues of the BuTTER model, and the limitations of the simple dictionary-based models. A vital part of the food NER and NEL component is the FoodNER model, which enables the linking of the food entities to the Hansard corpus, and the FoodOn and SNOMEDCT ontologies. This model is not used in the dataset generation procedure of the SAFFRON models, and thus, the extracted food entities in the dataset on which the models are trained, are not normalized, nor linked to any existing resources.

5.3. Directions for future work

In this study, we have limited the types of relations to *cause* and *treat*, and these are meant to refer to a broad positive or negative impact of the consumption of food on the development or progression of the disease. We have not considered more fine-grained relations, since the existing annotated data does not allow for it. Currently, the SAFFRON model is the only publicly available relation extraction model which identifies relations between food and disease entities, and it is only capable of extracting *cause*

and *treat* relations. Extracting more fine-grained relations would require expert engagement for defining the relation types and annotating data for developing the models which can identify such relations. According to the obtained results, the food NER methods used in the pipeline could also use further improvement.

The FooDis pipeline is a data mining pipeline for food-disease RE. It is a pipeline that extracts all relations that are mentioned in the text with the limitation of treating all evidence equally. In the future, we are planning to explore and extract information about evidence-based quality criteria for each extracted relation.

Additional work is required to fully evaluate all aspects of the pipeline. In this paper, we evaluate the pipeline's results on 125 manually annotated abstracts and we perform a comparison with relations found in the NutriChem and DietRx resources. This was a best-effort attempt to provide insight into the pipeline's performance, the types of mistakes it can make, and the precision-recall tradeoff that can be achieved by changing the pipeline parameters. However, this is still a small ground truth corpus, which does not allow for a fine-grained analysis of the pipeline's ability to correctly extract and link diverse entities and to correctly determine the relation in sentences of different complexities. Unfortunately, a more detailed evaluation of the pipeline's results requires a larger annotation effort. Since medical texts contain technical and domain-specific terminology, medical experts with specialized domain knowledge are required to understand and correctly label the data, and the size of the annotated data is limited by expert availability.

Additional effort on the experts' part would be required to completely evaluate the performance of the FooDis pipeline on the NEL task. Since the pipeline links the entities to multiple resources, a ground truth corpus which would be used for evaluation, should contain a link for each food entity to the FooDB, the SNOMEDCT, the Hansard Corpus, and the

FoodOn ontology. The ground truth can optionally link the food entities to the ITIS, Wikipedia articles and the NCBIT, however, this can also be done using the FooDB database, since it contains manually defined mappings between these resources. Similarly, the ground truth corpus should link the disease entities at least to the DO. Since the DO already contains the manually defined mappings to the rest of the resources, the mappings of the disease entities to the rest of the resources mentioned in the paper are optional.

The FooDis pipeline is evaluated on scientific abstracts since this is the part of the paper which is publicly available in most cases. Unfortunately, full texts of the papers are often not free for public use. However, the proposed methodology is general enough to be adapted to extract the relations from the full texts also. In addition, the current extraction has been performed on a sentence level. In the future, we plan to extend it on a document level by combining it with some text summarization and topic detection methodologies.

6. Conclusion

The FooDis pipeline captures the effect of the consumption of food on the development of different diseases, based on findings in biomedical scientific literature. This is accomplished by identifying the food and disease entities in the text, linking them to existing semantic resources and associating them with a *cause* and *treat* relation. The experimental results show that the pipeline can reliably contribute to the process of relation extraction between food and disease entities, decreasing the time and effort required on the domain experts' part. Even though the pipeline works completely autonomously, we deem it to be semi-automatic, since it is highly advisable that the outputs are checked and validated by a domain expert.

Acknowledgements

This work has been supported by the Ad Futura grant for postgraduate study to GC; the Slovenian Research Agency [research core funding programme P2-0098, young researcher grant PR-12393 to GC]; 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

(). spaCy python library. `https://spacy.io/api`. Accessed: 2023-02-14.

Alexander, M., & Anderson, J. (2012). *The Hansard Corpus, 1803-2003*. Technical Report. URL: `http://eprints.gla.ac.uk/81804/`.

Ben Abdessalem Karaa, W., Mannai, M., Dey, N., Ashour, A. S., & Olariu, I. (2018). Gene-disease-food relation extraction from biomedical database. In *Soft Computing Applications: Proceedings of the 7th International Workshop Soft Computing Applications (SOFA 2016), Volume 1 7* (pp. 394–407). Springer.

Joachim von Braun, J., Afsana, K., Fresco, O. L., Hassan, M., & Torero, M. (2021). Food systems – definition, concept and application for the un food systems summit. *The Scientific Group for the UN Food Systems Summit*, . URL: `https://sc-fss2021.org/`.

Cenikj, G., Eftimov, T., & Koroušić Seljak, B. (2021a). Food - Disease Relation Extraction dataset. `https://github.com/gjorgjinac/food-disease-dataset`. Accessed: 2021-05-11.

Cenikj, G., Eftimov, T., & Koroušić Seljak, B. (2021b). SAFFRON: tranSfer leArning For Food-Disease RelatiOn extractioN. In *Proc. Workshop on Biomedical Natural Language Processing at the North American Chapter of the Association for Computational Linguistics*.

Cenikj, G., Koroušić Seljak, B., & Eftimov, T. (2021c). Foodis - A relation mining pipeline, source code. https://github.com/gjorgjinac/foodis_pipeline.

Cenikj, G., Popovski, G., Stojanov, R., Koroušić Seljak, B., & Eftimov, T. (2020). Butter: Bidirectional lstm for food named-entity recognition. In *Proc. Big Food and Nutrition Data Management and Analysis at IEEE BigData 2020* (pp. 3550–3556). doi:10.1109/BigData50022.2020.9378151.

Davis, A. P., Wiegers, T., Rosenstein, M. C., & Mattingly, C. (2012). MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database (Oxford)*, *2012*, bar065.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, .

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018b). Bert: Pre-training of deep bidirectional transformers for language understanding. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

Donnelly, K. (2006). Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, *121*, 279–90. URL: https://app.dimensions.ai/details/publication/pub.1077321040.

Dooley, D., Griffiths, E., Gosal, G., Buttigieg, P., Hoehndorf, R., Lange, M., Schriml, L., Brinkman, F., & Hsiao, W. (2018). Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Science of Food*, *2*.

Dumitrache, A., Aroyo, L., & Welty, C. (2015a). Achieving expert-level annotation quality with crowdtruth: The case of medical relation extraction. In *BDM2I@ISWC*.

Dumitrache, A., Aroyo, L., & Welty, C. (2015b). Crowdtruth measures for language ambiguity: The case of medical relation extraction. *CEUR Workshop Proceedings*, *1467*, 7–19.

Dumitrache, A., Aroyo, L., & Welty, C. (2017). Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, *8*. URL: http://arxiv.org/abs/1701.02185. arXiv:1701.02185.

Eftimov, T., Popovski, G., Petković, M., Seljak, B. K., & Kocev, D. (2020). Covid-19 pandemic changes the food consumption patterns. *Trends in food science & technology*, *104*, 268–272.

Federhen, S. (2011). The NCBI taxonomy database. *Nucleic Acids Research*, *40*, D136–D143. URL: https://doi.org/10.1093/nar/gkr1178. doi:10.1093/nar/gkr1178.

Ferreira, J. D., Teixeira, D. C., & Pesquita, C. (2021). Biomedical ontologies: Coverage, access and use. In O. Wolkenhauer (Ed.), *Systems Medicine* (pp. 382–395). Oxford: Academic Press. URL: https://www.sciencedirect.com/science/article/pii/B9780128012383116642. doi:https://doi.org/10.1016/B978-0-12-801238-3.11664-2.

Fragoso, G., de Coronado, S., Haber, M., Hartel, F., & Wright, L. (2004). Overview and utilization of the NCI thesaurus. *Comparative and Functional Genomics*, *5*, 648–654. URL: https://doi.org/10.1002/cfg.445. doi:10.1002/cfg.445.

Giorgi, J. M., & Bader, G. D. (2019). Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, *36*, 280–286. URL: https://doi.org/10.1093/bioinformatics/btz504. doi:10.1093/bioinformatics/btz504. arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/1/280/31813710/btz504.

Hamosh, A., Scott, A. F., Amberger, J., Valle, D., & McKusick, V. A. (2000). Online mendelian inheritance in man (OMIM).

Human Mutation, *15*, 57–61. URL: https://doi.org/10.1002/(sici)1098-1004(200001)15:1<57::aid-humu12>3.0.co;2-g. doi:10.1002/(sici)1098-1004(200001)15:1<57::aid-humu12>3.0.co;2-g.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, *33*, D514–D517. URL: https://doi.org/10.1093/nar/gki033. doi:10.1093/nar/gki033. arXiv:https://academic.oup.com/nar/article-pdf/33/suppl_1/D514/7621651/gki033.pdf.

Humphreys, B. L., Lindberg, D. A. B., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, *5*, 1–11. URL: https://doi.org/10.1136/jamia.1998.0050001. doi:10.1136/jamia.1998.0050001. arXiv:https://academic.oup.com/jamia/article-pdf/5/1/1/2300717/5-1-1.pdf.

Jensen, K., Panagiotou, G., & Kouskoumvekaki, I. (2014). NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Research*, *43*, D940–D945. URL: https://doi.org/10.1093/nar/gku724. doi:10.1093/nar/gku724. arXiv:https://academic.oup.com/nar/article-pdf/43/D1/D940/7330366/gku724.pdf.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*, 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682. arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682

Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L. A., & Valencia, A. (2010). An overview of biocreative ii. 5. *IEEE/ACM*

Transactions on Computational Biology and Bioinformatics, *7*, 385–399.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

Malone, J., Adamusiak, T., Holloway, E., & Parkinson, H. (2009). Developing an application ontology for annotation of experimental variables – experimental factor ontology. *Nature Precedings*, . URL: https://doi.org/10.1038/npre.2009.3806.1. doi:10.1038/npre.2009.3806.1.

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., & Parkinson, H. (2010). Modeling sample variables with an experimental factor ontology. *Bioinformatics*, *26*, 1112–1118. URL: https://doi.org/10.1093/bioinformatics/btq099. doi:10.1093/bioinformatics/btq099.

Miao, Q., Zhang, S., Meng, Y., & Yu, H. (2012). Polarity analysis for food and disease relationships. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (pp. 188–195). IEEE volume 1.

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-J., Ohta, T., Pyysalo, S., & Zweigenbaum, P. (2013). Overview of bionlp shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop* (pp. 1–7).

Ni, Y., Jensen, K., Kouskoumvekaki, E., & Panagiotou, G. (2017). Nutrichem 2.0: exploring the effect of plant-based foods on human health and drug efficacy. *Database: The Journal of Biological Databases and Curation*, *2017*. doi:10.1093/database/bax044.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., & Jensen, L. J. (2015). Diseases: Text mining and

data integration of disease–gene associations. *Methods*, *74*, 83–89. URL: https://www.sciencedirect.com/science/article/pii/S1046202314003831. doi:https://doi.org/10.1016/j.ymeth.2014.11.020.

Popovski, G., Seljak, B. K., & Eftimov, T. (2019). FoodBase corpus: a new resource of annotated food entities. *Database*, *2019*. URL: https://doi.org/10.1093/database/baz121. doi:10.1093/database/baz121. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baz121/303

Rogers, F. B. (1963). Medical subject headings. *Bull Med Libr Assoc*, *51*, 114–116.

Sayers, E. (2010). A general introduction to the e-utilities. https://www.ncbi.nlm.nih.gov/books/NBK25497/. Accessed: 2021-03-15.

Schoeneck, M., & Iggman, D. (2021). The effects of foods on ldl cholesterol levels: A systematic review of the accumulated evidence from systematic reviews and meta-analyses of randomized controlled trials. *Nutr Metab Cardiovasc Dis*, *31*, 1325–1338. doi:10.1016/j.numecd.2020.12.032. arXiv:https://pubmed.ncbi.nlm.nih.gov/33762150/.

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C. P., Kibbey, S., Sreekumar, P., Le, C., Giglio, M., & Greene, C. (2018). Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research*, *47*, D955–D962. URL: https://doi.org/10.1093/nar/gky1032. doi:10.1093/nar/gky1032.

Segura-Bedmar, I., Martínez Fernández, P., & Sánchez Cisneros, D. (2011). The 1st ddiextraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts, .

Stojanov, R., Popovski, G., Jofce, N., Trajanov, D., Seljak, B. K., & Eftimov, T. (2020). Foodviz: Visualization of food entities linked across

different standards. In G. Nicosia, V. Ojha, E. La Malfa, G. Jansen, V. Sciacca, P. Pardalos, G. Giuffrida, & R. Umeton (Eds.), *Machine Learning, Optimization, and Data Science* (pp. 28–38). Cham: Springer International Publishing.

Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, *20*, 806–813.

Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, *12*, 393 – 393.

Weiss, K., Khoshgoftaar, T., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*, 9. URL: https://doi.org/10.1186/s40537-016-0043-6. doi:10.1186/s40537-016-0043-6.

Yang, H., Swaminathan, R., Sharma, A., Ketkar, V., & D 'Silva, J. (2011). Mining biomedical text towards building a quantitative food-disease-gene network. *Learning structure and schemas from documents*, (pp. 205–225).

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2019). A comprehensive survey on transfer learning. *CoRR*, *abs/1911.02685*. URL: http://arxiv.org/abs/1911.02685. arXiv:1911.02685.

Appendix A. Semantic resources

In this section, we introduce the semantic resources utilized by the FooDis pipeline for extracting relations between the food and disease entities.

Appendix A.1. Biomedical semantic resources

Next, several semantic resources that are used for linking disease entities are explained in more detail.

UMLS (Humphreys et al., 1998) is the largest available compendium of biomedical vocabularies. Its main vocabulary, Metathesaurus, integrates over 200 biomedical vocabularies and thesauri with over 1 million concepts (Ferreira et al., 2021). It links alternative names referring to the same concept and identifies useful relations.

MESH (Rogers, 1963) is hierarchically-organized vocabulary, used for indexing and searching PubMed (https://pubmed.ncbi.nlm.nih.gov/) and MEDLINE (https://www.nlm.nih.gov/medline) records, as well as other biomedical and health-related information.

The DO (Schriml et al., 2018) provides standardized representation of human diseases and semantic connection of related phenotypic, gene and genetic information. It enables cross mapping and integration of disease and medical vocabularies such as MESH, ICD, NCIT, SNOMEDCT and OMIM disease-specific terms and identifiers.

SNOMEDCT (Donnelly, 2006) is a standardized, multilingual clinical terminology for consistent representation of electronic health records. It includes relations between different types of biomedical entities such as body structures, organisms, substances, pharmaceutical products, physical objects, physical forces, specimens, symptoms, drugs, food and surgical, therapeutic and diagnostic procedures.

The NCIT (Fragoso et al., 2004) is a reference terminology aimed at facilitating cancer research. It covers areas of basic and clinical science, including concepts such as diseases, anatomy, genes, drugs, biomedical techniques, and biological processes. Each concept has multiple annotations, such as synonyms, preferred names, textual definitions and references to external sources.

OMIM (Hamosh et al., 2000) is a KB targeting human genes and genetic disorders. It contains summaries of genes or genetic phenotypes and links to genetic resources such as DNA sequences, protein sequences, PubMed references, and mutations (Hamosh et al., 2005).

The EFO (Malone et al., 2010, 2009) is an application-focused ontology modeling the experimental factors in ArrayExpress, a public repository for functional genomics datasets. It covers aspects of diseases, anatomy, cells, and compounds.

Appendix A.2. Food semantic resources

Next, several semantic resources that are used for linking food entities are explained in more detail.

FooDB (http://foodb.ca/) is a database containing food names, descriptions, macronutrient and micronutrient information, including constituents that give different foods their taste, color, texture and aroma. Each chemical is described by more than 100 compositional, biochemical and physiological attributes. The FooDB also contains mappings of food concepts to Wikipedia, ITIS and NCBIT (Federhen, 2011).

FoodOn (Dooley et al., 2018) is a farm-to-fork food ontology developed with the aim of solving the issues of incompatibility and ambiguity of food references. FoodOn acts as a hub that interfaces with more specialized ontologies, providing schemes for food categorization and covering basic raw animal and plant food sources, as well as terms related to packaging, cooking and preservation.

The Hansard corpus (Alexander & Anderson, 2012) is a collection of speeches given in the British Parliament, which can be searched through using semantic tags. One of the features of the corpus search tool [1] is a hierarchical organization of more than 8,000 different semantic categories, where *Food and Drink* is one of the top-level categories.

The ITIS (https://www.itis.gov/) includes documented taxonomic information about the scientific names, synonyms, and common names of

[1] https://www.english-corpora.org/hansard/

aquatic and terrestrial flora and fauna, such as plants, animals, fungi, and microbes.

The NCBIT (Federhen, 2011) is the standard nomenclature and classification repository for the International Nucleotide Sequence Database Collaboration (INSDC). It contains names and phylogenetic lineages of more than 160,000 organisms with molecular data in the NCBI databases. It is aimed at documenting nomenclature and systematics, rather than the description of taxa. It is manually curated by NCBI specialists.

Apart from clinical concepts, SNOMEDCT (Donnelly, 2006) also contains hierarchically organized food entities.

Highlights:

- a novel Information Extraction pipeline for mining scientific literature
- suggestion of cause or treat relations between food and disease entities
- linking food and disease entities from various knowledge bases

We declare no conflicts of interest for the publication of our manuscript “FooDis: A food-disease relation mining pipeline”.

Gjorgjina Cenikj
Tome Eftimov
Barbara Koroušić Seljak