FOOD INFORMATION EXTRACTION AND DATA NORMALIZATION

Gorjan Popovski

Master Thesis
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Assoc. Prof. Dr. Barbara Koroušić Seljak, Computer Systems Department, Jožef Stefan Institute, 1000 Ljubljana, Slovenia
Co-Supervisor: Dr. Tome Eftimov, Computer Systems Department, Jožef Stefan Institute, 1000 Ljubljana, Slovenia

Evaluation Board:
Asst. Prof. Dr. Panče Panov, Chair, Department of Knowledge Technologies, Jožev Stefan Institute, 1000 Ljubljana, Slovenia
Asst. Prof. Dr. Slavko Žitnik, Member, Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia
Assoc. Prof. Dr. Barbara Koroušić Seljak, Member, Computer Systems Department, Jožef Stefan Institute, 1000 Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Gorjan Popovski

FOOD INFORMATION EXTRACTION AND DATA NORMALIZATION

Master Thesis

LUŠČENJE INFORMACIJ IN NORMALIZACIJA PODATKOV O ŽIVILIH

Magistrsko delo

Supervisor: Assoc. Prof. Dr. Barbara Koroušić Seljak

Co-Supervisor: Dr. Tome Eftimov

Ljubljana, Slovenia, 2020

To my Father Jovo, my Mother Beti and my big Sister Bojana.

Acknowledgments

First and foremost, I would like to express my gratitude to my co-supervisor Tome Eftimov, a mentor, and now a true friend, who introduced me to the world of science, made each discussion motivating and exciting, and supported me in every goal I aimed for. It was this newly found motivation that made me certain of what I want to pursue in my future, and for that I will be eternally grateful.

Next, I would like to express my gratitude to the kindest person, my supervisor Barbara Koroušić Seljak, who unconditionally supported me throughout my whole experience of working, studying and integrating into a new community.

Finally, I would like to acknowledge the funding awarded by the Ad Futura scholarship and thank the Computer Systems Department at the Jožef Stefan Institute for their continuous monetary support.

Abstract

In the last decade, a great amount of work has been done in predictive modelling in healthcare. All this work is made possible by the existence of several available biomedical vocabularies and standards, which play a crucial role in understanding health information. Moreover, there are available systems, such as the Unified Medical Language System, that bring and link together all these biomedical vocabularies to enable interoperability between computer systems. However, there is still no annotated corpus with food concepts, and there are only a few rule-based food named-entity recognition systems for food concepts extraction. There are also several food ontologies that exist, each developed for a specific application scenario. However, there are no links between these ontologies.

Additionally, the application of Natural Language Processing (NLP) methods and resources to biomedical textual data has received growing attention over the past years. Previously organized biomedical NLP-shared tasks (such as, for example, BioNLP Shared Tasks) are related to extracting different biomedical entities (like genes, phenotypes, drugs, diseases, chemical entities) and finding relations between them. However, to the best of our knowledge, there are limited NLP methods that can be used for information extraction of entities related to food concepts.

To address the issues of Information Extraction (IE) and data normalization in the domain of food, we propose several methods for solving these tasks in this thesis.

First, we propose a novel rule-based Named-Entity Recognition (NER) method named FoodIE, whose manual evaluation showed it to be a very promising approach. The benefit of using a rule-based approach is that it does not rely on manually annotated ground truth data corpora. Additionally, as mentioned previously, no such corpora exist. Moreover, we present an experimental evaluation and comparison with other existing food NER methods.

Moving on to the second task, we present the first annotated corpus with recipes annotated with food entities. This is a crucial first step into food data normalization, as it enables corpus-based methods to be developed and according models to be trained. In addition to this, we propose a benchmarking data set as part of the corpus, consisting of 1,000 ground true annotated recipes. This corpus is named FoodBase and can be considered a "silver standard" in the domain.

Furthermore, we explore an existing language for describing foods in order to showcase the need for food data and food concept normalization. With this, we pave the way for FoodOntoMap, a NER-based method for linking food concepts across different food ontologies. It works by mapping the entities recognized by different food NER methods. We further explore data normalization by leveraging lexical and semantic similarity measures in low-researched languages, i.e. Slovenian.

Finally, the thesis presents a food data visualization tool, named FoodViz, where all the methods and data are made available in a web-based framework.

Povzetek

Na področju modeliranja napovedovanja v zdravstvu je bilo v zadnji dekadi narejenega veliko dela. To ne bi bilo možno brez obstoja biomedicinskih slovarjev in standardov, ki so ključni za razumevanje informacij o zdravju. Poleg tega obstajajo sistemi, kot je na primer sistem UML (angl. Unified Medical Language System), ki omogočajo povezovanje biomedicinskih slovarjev in tako zagotavljajo interoperabilnost informacijskih sistemov. Na področju živilstva in prehrane pa žal še nimamo korpusa besedil z označenimi koncepti, ki opisujejo živila in tudi število sistemov za luščenje takšnih konceptov iz imen živil s pomočjo pravil (angl. rule-based food named-entity recognition) je skopo. Poznamo zgolj nekaj ontologij, ki opisujejo živila. Njihova pomanjkljivost je v nezdružljivosti, saj so bile razvite za specifične potrebe izbranih aplikacij.

Na področju biomedicine najdemo številne primere uporabe metod obdelave naravnih besedil (angl. Natural Language Processing) in semantičnih virov. Že pred leti je raziskovalna skupnost zastavila BioNLP naloge, v okviru katerih izvajajo luščenje različnih biomedicinskih entitet (kot so geni, fenotipi, zdravila, bolezni, kemijski elementi) in iskanje relacij med njimi. Obratno, na področju živilstva, kolikor nam je znano, obstaja le omejeno število metod obdelave naravnih besedil, ki omogočajo luščenje konceptov živil.

Magistrska teza obravnava problem luščenja informacij iz podatkov o živilih in njihovo normalizacijo, s predstavitvijo več različnih metod.

Najprej predstavimo novo metodo, imenovano FoodIE, ki omogoča luščenje informacij iz besedilnih podatkov na osnovi imen živil. Ročna evalvacija rezultatov metode je pokazala, da je pristop zelo obetaven. Njegova prednost je, da temelji na pravilih, kar pomeni, da se ne zanaša na ročno označene 'ground-truth' podatke v korpusu besedil. Kot smo že omenili, takšen označen korpus niti še ne obstaja. Poleg ročne evalvacije smo izvedli tudi primerjavo pristopa z drugimi metodami za luščenje informacij iz besedilnih podatkov na osnovi imen živil.

Nadaljujemo z opisom prvega korpusa besedil z označenimi entitetami za opis živil. Besedila smo zajeli iz 1000 receptov, opisanih v angleščini, in jih označili z 'ground-truth' označbami. Korpus je pomembna osnova za normalizacijo podatkov in omogoča nadaljnji razvoj metod, temelječih na korpusu, in učenje pripadajočih modelov. Poimenovali smo ga FoodBase in predstavlja šrebrni standard"na področju živilstva. Zasnovali smo ga tako, da omogoča primerjalno analizo (angl. benchmarking).

Nato raziščemo obstoječe jezike za opis živil, ki so potrebni za normalizacijo podatkov in konceptov, ki opisujejo živila. Predstavimo metodo, imenovano FoodOntoMap, ki omogoča povezovanje takšnih konceptov, opisanih z različnimi ontologijami. Metoda temelji na luščenju informacij iz imen živil. Opis metode FoodOntoMap zaključimo s predstavitvijo pristopa, ki temelji na leksikalni in semantični meri podobnosti, razviti za potrebe slovenskega jezika, kot primera slabo raziskanega jezika.

Nalogo zaokrožimo s predstavitvijo spletnega orodja za vizualizacijo, imenovanega FoodViz, ki omogoča predstavitev rezultatov metod FoodIE in FoodOntoMap ter korpusa FoodBase živilskim strokovnjakom.

Contents

List of Figures

List of Tables

Abbreviations

NLP	...	Natural Language Processing
IE	...	Information Extraction
NER	...	Named-Entity Recognition
ML	...	Machine Learning
DL	...	Deep Learning
RL	...	Representation Learning
TP	...	True Positive
FP	...	False Positive
TN	...	True Negative
FN	...	False Negative

Chapter 1

Introduction

This thesis presents completed work in the Computer Science. More specifically, the thesis presents contributions in the field of Information Extraction and data normalization applied to the domain of Food. It proposes several novel contributions, such as (1) a novel rule-based food Named-Entity Recognition method, (2) a data set with annotated recipes, (3) a methodology for food data normalization based on Named-Entity Recognition models, and (4) a methodology for visualisation of food entities. In this chapter, we provide the problem definition, the list of considered hypotheses, discussion of scientific relevance, scientific contributions, and an overview of the structure of the thesis.

1.1 Problem Definition

Nowadays, a large amount of textual information is available in digital form and published in public web repositories (e.g. online media news, scientific publications, social media posts). The textual information is contained within unstructured data, meaning that the data has no predefined data model. In Computer Science, working with textual data is still a challenge because of its variability - the same concepts can be mentioned in different ways regarding the fact that people express themselves using different writing styles and even dialects.

Information Extraction (IE) is a task of automatically extracting information from unstructured data and, in most cases, is concerned with the processing of human language text by means of Natural Language Processing (NLP) [1] methods. The idea behind IE is to extract information from analyzed text and provide its structured representation, both in an automated way. The information to be extracted is defined by users who write the free-from text, and consists of predefined concepts of interest and related entities (e.g. chemicals, drugs, diseases, foods, etc.), as well as relationships between entities and events.

One of the classic IE tasks is Named-Entity Recognition (NER), which addresses the problem of identification and classification of concepts predefined by subject-matter experts. [2]. It aims to determine and identify words or phrases in text into predefined labels (classes) that describe concepts of interest in a given domain. Various types of NER methods exist: *terminological-driven*, *rule-based*, *corpus-based*, *methods based on active learning (AL)*, and *methods based on deep neural networks (DNNs).*

Terminology-driven NER methods, also called dictionary-based NER methods [3], perform matching text phrases with concept synonyms that exist in the terminological resources (dictionaries). In order to improve the performance of these methods, instead of strict matching they are combined with some heuristics, such as the generation of words that occur in entity mentions, generating permutations of words in concept synonyms, solving disambiguation problem, etc. The main disadvantage of these methods is that only the entity mentions that exist in the resources will be recognized, but the benefit of using them is related to the frequent updates of the terminological resources with new concepts and synonyms.

Rule-based NER methods [4], [5], which use regular expressions that combine information from terminological resources and characteristics of the entities of interest. The main disadvantage of these methods is the manual construction of the rules, which is a time-consuming task and is very domain dependent.

Corpus-based NER methods [6], [7] are based on the evidence that exists in an annotated corpus provided by subject-matter experts from the domain and use of Machine Learning (ML) algorithms to predict the entities' labels. These methods are less affected by terminological resources and manually created rules, but the limitation is the existence of an annotated corpus for the domain of interest. The construction of the annotated corpus for a new domain is a time consuming task and requires effort by the human experts to produce it.

To minimize the annotation cost and to exploit unlabelled data in a research related to NER, *AL* can be used [8]. This is a semi-supervised setting, or iterative supervised learning, in which an algorithm is able to interactively query the user to obtain the desired outputs at new data points. The examples needed to learn a new concept are chosen by the algorithm and their number can often be much lower than the number of examples required for supervised learning. An active learner uses an unlabelled corpus as the input and generates NER models, while iteratively interacting with the users who annotate sentences queried from the corpus. It usually consists of three components: the annotation interface, the corpus-based NER, and the component for querying samples. Recently, several studies have also presented the effectiveness of AL to NER tasks [9]–[11].

Corpus-based NER methods rely on the use of the costly handcrafted features and on the output of other NLP tasks. Over the past few years, some recent work on NER applied *DNNs*, which *minimize* the need of these costly features and have successively advanced the state-of-the-art approaches [12]. However, this typically requires large amounts of annotated data. To reduce the amount of annotated data, deep learning or DNN can be combined with AL [13].

In recent years, the use of predictive modeling in healthcare increases with the large amount of data that is becoming available. One example of such data are Electronic Health Records (EHRs) [14], [15], which represent the largest source of medical data. Analyzing them directly is very difficult as the medical information is presented as natural language text (i.e. unstructured data) and the key challenge is to extract terms that denote different medical concepts (e.g. drugs, diseases, procedures, treatments, etc.). For this reason, a lot of NER methods have been developed [16], which are further used to extract this information for each patient and then trying to find some representation for the patient's information for some predictive study. Besides the unstructured data, there are also resources that consist of structured patient medical information. One such example is the MIMIC-III data [17], which consists of data relating to patients who stayed within the intensive care units at Beth Israel Deaconess Medical Center.

The common thing about the medical data, no matter where it comes from (unstructured or structured data), is that it is further used to find a patient's representation by projecting the data into a continuous vector space [18]–[20]. With this, medical embeddings (medical vector representations) are learned in order to capture the non-linear relationships that exist between the medical concepts. These representations are further used with some advanced ML methods, such as deep learning, to perform predictive studies in healthcare. However, all this happens as a result of the availability of biomedical vocabularies and standards that can be used to normalize the medical concepts before learning the embedding space. One such example is the Unified Medical Language System (UMLS) that brings and links together several biomedical vocabularies to enable interoperability between computer systems [21].

In 2019, the Lancet Planetary Health published that 2019 would be the year of nutrition, where the focus would be on the links between food systems, human health, and the environment. Namely at that time, contrary to the large number of available resources for the biomedical domain, there is still a limited number of resources that can be used in the food domain. In 2020, there is still no annotated corpus with food concepts, and there are few rule-based food named-entity recognition systems that can be used for food concepts extraction [22], [23]. Additionally, a number of food ontologies exist, each developed for a specific application scenario, but there are no links between them [24].

In order to move beyond the state-of-the-art for finding relationships between the human health, food systems and environment, we should have resources that will be used for food concepts normalization. For this reason, we have created the FoodOntoMap resource that consists of food concepts automatically extracted from recipes and for each one semantic tags from four food vocabularies (i.e. Hansard taxonomy and three food ontologies (FoodOn, OntoFood, and SNOMED CT)) are assigned. With this, we have created a resource that provides a link between different food ontologies that can be further reused to develop embedding space for food concepts and applications for understanding the relations between food systems, human health, and the environment.

1.2 Task Definition

The purpose of the thesis is to develop novel methodologies for Information Extraction (IE) and data normalization in the domain of Food. The goal is to work towards creating a unified language for describing foods that can provide a standardized semantic resource. The future goal is to interlink and enrich this data with different aspects of the food domain, such as food availability and nutritional information. Hence the thesis consists of three main tasks:

- The first task is Food Information Extraction from unstructured text. Specifically, it is concerned with Food Named-Entity Recognition from recipe texts. The aim of this task is to automatically extract and annotate food concepts from the texts.

- The second task is creating an annotated data set with extracted food concepts. The utility of such data set is its ability to be used in supervised ML techniques and to enable the use of corpus-based methodologies. The annotated data set to be developed in this work is the first of its kind, containing recipes with annotated food concepts.

- The third and final task is food data normalization. The aim of this task is to interlink the data from different food semantic resources to provide a unified view of them. This would enable interoperability between each semantic resource and serve as a good basis for creating a unified language for describing foods.

1.3 Hypotheses

The hypotheses of this thesis regard the tasks of: i.) Information Extraction in the domain of food, and ii.) data normalization in the domain of Food. The goal of the thesis is to advance the necessary methodologies and data to work towards constructing a unified language for describing foods. To achieve the goals, we consider the following hypotheses:

- **Hypothesis 1:** The development of a rule-based food Named-Entity Recognition method without the need of using pre-annotated corpora from subject-matter experts is possible.

- **Hypothesis 2:** It is possible to construct an annotated semantic resource in the food domain which provides recipes annotated with the food concepts that appear in them.

- **Hypothesis 3:** It is possible to develop a method for data normalization based on NER methods and semantic resources in the domain of Food that provides an explicit mapping between each of the used semantic resources.

1.4 Scientific Relevance

In the last decades, copious computer and data science research has been published regarding the biomedical domain, focusing on drugs, diseases, genes, etc. Different computing-based methods have been proposed to address data and extract information automatically with the goal to provide a unified way of modeling and presenting varying biomedical data. As the nature of the data is inherently unstructured (e.g. biomedical records), these methods are particularly useful in improving the efficiency, readability, and interoperability of the data. Moreover, the existence of such unified data sources enables more advanced ML techniques to be applied. Some examples include automatic annotation of unstructured data, discovery of different types of relationships (e.g. disease-drug, disease-phenotype, gene-protein interactions, etc.), personalized health-records, etc. These methodologies provide yet another tool to enable novel and more efficient research and discoveries in the biomedical domain.

However, despite the high impact food and nutrition has on personal and public health, it remains a less-researched field from this aspect, as compared to the biomedical domain. The effect of food intake and dieting is a crucial aspect of personal and public health. In modern times, debates over what people should eat to avoid certain health risks or to improve some aspects of their health, or even treat a certain ailment, are becoming more prevalent. Moreover, this can be further linked with environmental studies to explore relationships between food systems and the environment. To render studies regarding the effect of a certain dietary habit feasible, data needs to be collected and unified in a way utilizable by state-of-the-art modeling techniques. As this type of data (e.g. dietary records, recipes, etc.) is also unstructured, the first crucial step is to automatically recognize and extract the relevant information from the raw data. Such a methodology would be far more efficient than manual iteration and extraction by subject-matter experts. As most such data is presented in natural languages, this task would be directly concerned with NLP. With this, a foundation for NLP resources in the domain of food and nutrition can be laid.

Additionally, annotated data sets are crucial for supervised ML techniques. As such data sets are not prevalent in the domain of food, constructing it would be greatly beneficial and enable further analyses.

Finally, as different data sets in the same domain are usually constructed with a specific goal in mind or for a specific purpose, they do not follow a common standard of representing the data contained within them. Enabling interoperability between data from different sources would provide an easy way to enrich the cumulative information contained in it, as well as propose a standardized representation for data resulting from future research. For this, data normalization and data harmonization in the domain of food would directly ease the endeavours of future research, as well as increase their potential for new discoveries.

In summary, NLP methods, annotated data sets (especially useful for ML techniques), and methods for data normalization and data harmonization in the domain of food would be greatly beneficial to help answer open scientific questions regarding dietary habits and public health.

1.5 Scientific Contributions

The completed work presented in this thesis follows the aforementioned tasks and proves the defined hypotheses, leading to the following scientific contributions:

1. FoodIE - a novel rule-based Named-Entity Recognition method, whose novelty lies in its rule-based nature as well as the promising results it has obtained. The work completed as part of this contribution is published as one conference [23] paper and one article [25] in the peer-reviewed journal *IEEE Access* with Impact Factor (IF) of 4.098. Personal contributions to the completed work: wholly developing FoodIE, data collection, data pre-processing, topic discussions and manuscript preparation.

2. FoodBase corpus - a new resource of annotated food entities, which additionally can be used as a benchmarking data set. The work completed as part of this contribution is published as an article [26] in the journal *Database* with an Impact Factor (IF) of 3.683. Personal contributions to the completed work: data collection, data pre-processing, manual annotation, manual evaluation, topic discussions and manuscript preparation.

3. FoodOntoMap - linking food concepts across different food ontologies, which provides an explicit mapping between food concepts and is constructed by using various NER methods and semantic resources from the domain of food. The work completed as part of this contribution is published as three conference articles [27]–[29]. Personal contributions to the completed work: data collection, data pre-processing, food concept mapping, training predictive models, learning embeddings, topic discussions and manuscript preparation.

4. FoodViz - visualization of food entities linked across different standards, which represents a tool aimed at facilitating interactions with users and subject-matter experts. The work completed as part of this contribution is published as one conference article [30]. Personal contributions to the completed work: supplying the data used in the visualization tool, topic discussions and manuscript preparation.

1.6 Thesis Structure

In Figure 1.1, the flowchart of the methodologies of the thesis is presented, where for each chapter the corresponding publications are listed.

In the Introduction of the thesis, we provide a basic overview of the concepts that are relevant in the domain of Food. It familiarizes the reader with the problem definition as well as the motivation of the work being done. The goal of this chapter is to enable the reader to internalize and distinguish the basic concepts that are critical for understanding the topic of the thesis.

Chapter 2 provides the necessary background to enable the reader to understand what research has been done in the field relevant to this thesis. Its main focus is providing an insight into the current state of the field of food Information Extraction as well as providing an overview with existing food semantic resources.

The text, figures and tables in the following chapters are adapted from our already published journal articles and conference papers to create a cohesive structure for the purpose of this thesis.

Chapter 3 focuses on a novel food Named-Entity Recognition method, named FoodIE [23], as well as its comprehensive evaluation and comparison with other existing food NERs from the NCBO Portal [25].

Chapter 4 delves into the topic of food data normalization. It begins with presenting an overview of the first data set with annotated food recipes - FoodBase [26], followed by an exploration of an existing food semantic resource (LanguaL) [29]. To conclude this Chapter, we present a novel methodology for food data normalization based on food NER methods - FoodOntoMap [27].

In Chapter 5, a use-case presenting food data normalization in a low-resourced language (Slovenian) is provided to demonstrate its applicability for these non-English languages. [28].

In Chapter 6, we present FoodViz [30], a visualization tool which incorporates all of the novel methods presented in the paper.

Finally, in the last chapter the conclusion and directions for future work are presented.

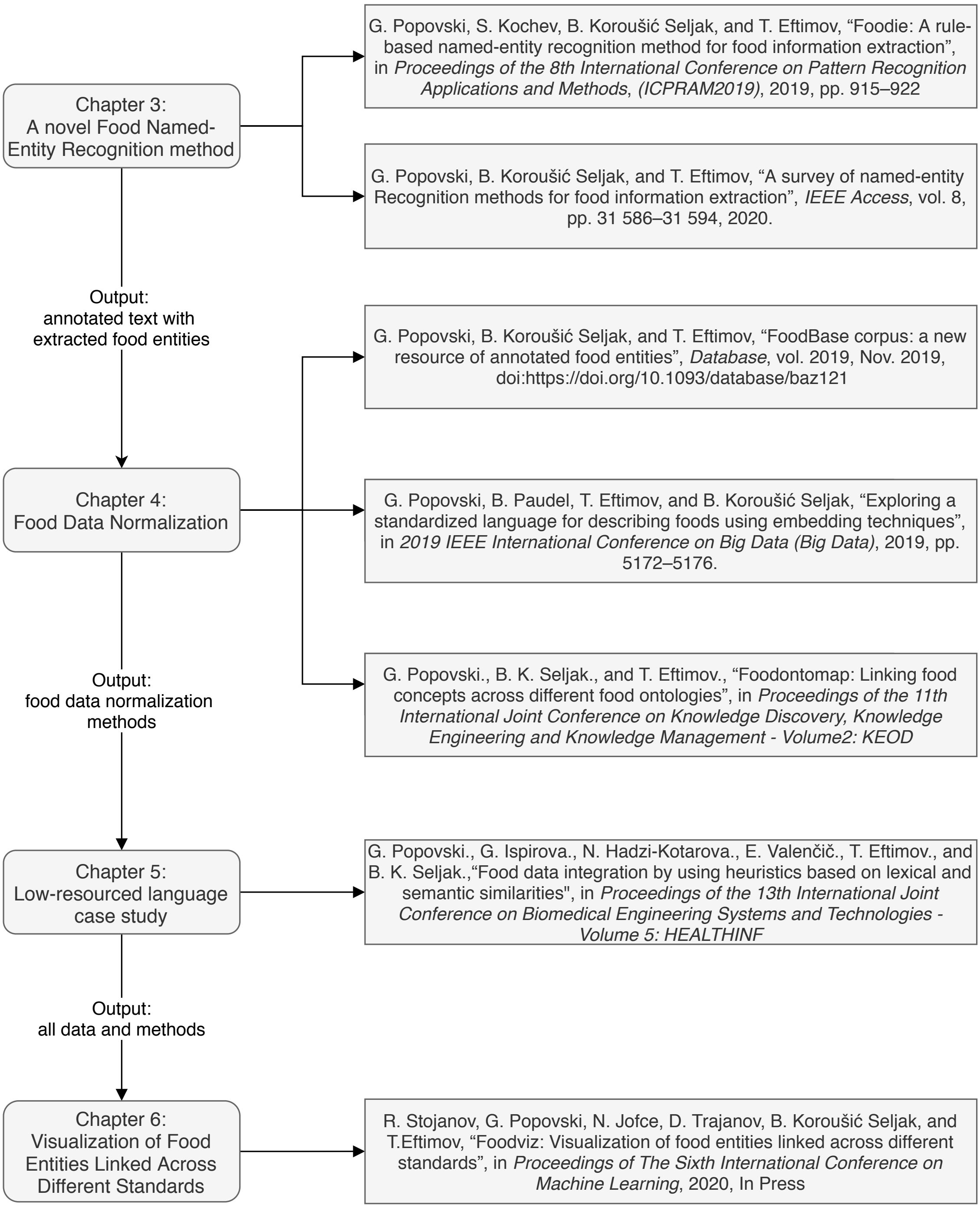

Chapter 3: A novel Food Named-Entity Recognition method
G. Popovski, S. Kochev, B. Koroušić Seljak, and T. Eftimov, “Foodie: A rule-based named-entity recognition method for food information extraction”, in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, *(ICPRAM2019)*, 2019, pp. 915–922
G. Popovski, B. Koroušić Seljak, and T. Eftimov, “A survey of named-entity Recognition methods for food information extraction”, *IEEE Access*, vol. 8, pp. 31 586–31 594, 2020.
Output: annotated text with extracted food entities
Chapter 4: Food Data Normalization
G. Popovski, B. Koroušić Seljak, and T. Eftimov, “FoodBase corpus: a new resource of annotated food entities”, *Database*, vol. 2019, Nov. 2019, doi:https://doi.org/10.1093/database/baz121
G. Popovski, B. Paudel, T. Eftimov, and B. Koroušić Seljak, “Exploring a standardized language for describing foods using embedding techniques”, in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 5172–5176.
G. Popovski., B. K. Seljak., and T. Eftimov., “Foodontomap: Linking food concepts across different food ontologies”, in *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume2: KEOD*
Output: food data normalization methods
Chapter 5: Low-resourced language case study
G. Popovski., G. Ispirova., N. Hadzi-Kotarova., E. Valenčič., T. Eftimov., and B. K. Seljak.,“Food data integration by using heuristics based on lexical and semantic similarities", in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF*
Output: all data and methods
Chapter 6: Visualization of Food Entities Linked Across Different Standards
R. Stojanov, G. Popovski, N. Jofce, D. Trajanov, B. Koroušić Seljak, and T.Eftimov, “Foodviz: Visualization of food entities linked across different standards”, in *Proceedings of The Sixth International Conference on Machine Learning*, 2020, In Press

Figure 1.1: Flowchart of the methodology of the thesis.

Chapter 2

Background

In this chapter, we provide insight into the existing related research in the field of Information Extraction and data normalization. It focuses on (1) Information Extraction methods in the domain of biomedicine and food science, where an overview of models and resources is provided, (2) available food semantic resources, (3) food data normalization, (4) Representation Learning, and (5) text similarity measures.

2.1 Information Extraction in Biomedicine and Food Science

IE from biomedical literature is a very important task with the goal of improving public health. Because NER methods which have the best performances are usually corpus-based NER methods, there is a need for an annotated corpus from biomedical literature that includes the entities of interest. For this purpose, different annotated corpora are produced by shared tasks, where the main aim is to challenge and encourage research teams to work on NLP problems.

BioNLP Shared Task 2013 [31] aimed to provide a common framework for IE in the biomedical domain. The biological questions addressed by this task were related to the domain of molecular biology and its related fields. The BioNLP Shared task 2013 consisted of six tasks: gene event extraction, cancer genetics, pathway curation, corpus annotation with gene regulation ontology, gene regulation networks in bacteria, and bacteria biotopes.

BioNLP Shared Task 2016 consisted of three tasks that address different aspects of knowledge acquisition from text and also encompasses a wide range of biological diversity [32]–[40]. The SeeDev task [32] aimed at extracting the regulation of the seed development in plants using a rich model. The Bacteria Biotopes 3 (BB3) task [36], [38] was used for the construction of a bacteria habitat database using external ontologies. The Genia 4 (GE4) task [34] aimed at delivering a new shared task framework to construct a knowledge base of NFkB synthesis and regulation through information extraction (IE).

BioCreative II gene mention recognition [41] was a task where different systems were designed to identify substrings in sentences corresponding to gene name mentions. The annotated corpus was provided to the participants, on which different methods were used and results in the performances varied. The best system was a semi-supervised learning method known as alternating structure optimization (ASO) [42]. Other systems were developed by using supervised machine learning, ML, algorithms. The second best performing system used conditional random fields, CRFs, [43], the third best performing system used a combination of two support vector machines, SVMs, and one CRF [44], and the fourth best performing system used a multimodal approach with two CRFs [45].

The work on gene mention recognition continued in BioCreative III [46], where the focus was on three tasks: cross-species gene normalization using full text; extraction of protein-protein interactions from full text, including document selection, identification of interacting proteins and identification of interacting protein pairs; and an interactive demonstration task for gene indexing and retrieval task using full text.

In BioCreative IV [47], the gene ontology annotation task was reintroduced along with the following new tasks: interoperability of text mining systems, web service-based NER, and chemical/drug entity name recognition. In the chemical/drug NER two main aspects were covered, the chemical document indexing and the chemical entity mention recognition. The extraction of chemical entities from unstructured text is a very important task for different research areas, since they are related to metabolism, enzymatic reactions, potential adverse effects, etc. The systems presented are based on three general strategies: supervised ML approaches, rule/knowledge-based approaches, and chemical dictionary look-up approaches [48]. The evaluation of the systems performances was made using the CHEMDNER annotated corpus, which was provided as a part of the workshop [49]. Most systems that use supervised ML methods are based on the CRFs, some of them used SVMs, and some a combination of SVMs and CRFs. Systems were presented that use mainly rule-based methods, but these require a deep understanding of both the existing chemical nomenclature standards as well as of the CHEMDNER annotation guidelines. The use of dictionary-lookup based systems required efficient dictionary pruning and post-processing of the results.

The work related to chemical NER continued also in BioCreative V [50]–[52], where the focus was on disease- and symptom-related entities and relations that exist between chemical/drug entities and disease entities.

BioCreative 2016 was focused on four main tasks: applications of text mining methods in areas such as crowdsourcing, database curation, the publication process, and metagenomics; methods for annotations such as disease, phenotype, and adverse reactions in different text sources literature, clinical records and social media; methods to achieve interoperability, generalisability, and scalability in text mining: BioC [53], RDF and semantic web, among others; and the application of ontologies in text mining and text mining as an ontology builder.

BioCreative/OHNLP 2018 was focused on two tasks: family history information extraction and clinical semantic textual similarity [54].

In comparison with the extensive work done for biomedical tasks, in the food science domain the situation is different. Several studies have been conducted, but with different goals. For example, in [55], the authors presented an approach to identify rice protein resistant to *Xanthomonas oryzae pv. oryzae*, which is an approach to enhance gene prioritization by combining text mining technologies with a sequence-based approach. Co-occurrence methods were also used to identify ingredients mentioned in food labels and extracting food-chemical and food-disease relationship [56], [57].

A ML approach to Japanese recipe text processing was proposed in [58], where one task, which was evaluated, was food-named entity recognition. This approach used the r-FG corpus, which is composed solely from Japanese food recipes. Another similar approach for generating graph structures from food recipes was proposed in [59], where the authors manually annotated a recipe corpus that is then used for training a ML model.

DrNER [22] 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 methodology extracts the whole dietary recommendation as opposed to separate food entities.

The UCREL Semantic Analysis System (USAS) is a framework for automatic semantic analysis of text, which distinguishes between 21 major categories, one of which is “food and farming” [60], being heavily utilized in our recently published rule-based system -

FoodIE [23]. The USAS can provide additional information about the food entity, but the limitation is that it works on a token level. For example, if in the text two words (i.e. tokens), like “grilled chicken”, denote one food entity that needs to be extracted and analyzed, the semantic tagger would actually parse the words “grilled” and “chicken” as separate entities and obtain separate semantic tags.

The NCBO Annotator is a Web service that annotates text provided by the user by using relevant ontology concepts [61]. It is available as a part of the BioPortal software services [62]. The annotation workflow is based on a highly efficient syntactic concept recognition (using concept names and synonyms) engine and on a set of semantic expansion algorithms that leverage the semantics in ontologies [61]. The methodology leverages ontologies to create annotations of raw text and returns the annotations by using semantic web standards.

In conclusion, Information Extraction is a crucial task in the domains of biomedicine and food science, which often depends on the time-consuming tasks, such as the creation of annotated resources or manual construction of rules.

2.2 Food Semantic Resources

In the domain of food, several food semantic resources already exist, such as:

FoodWiki provides a model of different types of foods, together with their nutritional information, and the re-commended daily intake [63].

AGROVOC is a large multilingual thesaurus, whose terminology is widely used in practice for subject fields in agriculture, fisheries, forestry, food and related domains [64].

Open Food Facts is an open source global food database that allows users to learn about a food’s nutritional information and compare products from around the world [24]. It is also beneficial for the food industry, where it can be used to track, monitor, and strategically plan food production.

Food Product Ontology describes food products using common representation, vocabulary and language for the food product domain. It is an extended version of a widely used standardized ontology for product, price, store, and company data [65].

FOODS (Diabetics Edition) is an ontology-driven system that delivers a web-based food-menu recommendation system for patients with diabetes in Thailand [66].

FoodOn focuses on the human-centric categorization and handling of food [67]. Its main goal is to develop semantics for food safety, food security, agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes. It uses parts from several ontologies covering domains such as anatomy, taxonomy, geography and cultural heritage. Its usage is related to research and clinical data sets in academia and government.

OntoFood is an ontology with Semantic Web Rule Language (SWRL) rules of nutrition for diabetic patients and is available in the BioPortal.

A detailed review of food ontologies was provided by Boulos et. al. [24].

SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) is a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic health records [68]. Beside the medical concepts that are the main focus of this ontology, there is also a *Food* concept that can be further used for food concept normalization.

The Hansard corpus is a collection of text and concepts created as a part of the SAMUELS project (2014-2016). It consists of nearly every speech given in the British Parliament from 1803-2005. The main benefit is that it allows semantically-based searches of these speeches. More details about semantic tags can be found in [60], [69]. The words are organized in 37 higher level semantic groups, in which one of them is also *Food and Drink* (i.e. AG). The AG category is further split into three subcategories: “Food” (AG:01), “Production of food, farming” (AG:02) and “Acquisition of animals for food, hunting” (AG:03). The “Food” subcategory consists of 125 top level semantic tags, the “Production of food, farming” consists of 36 top level semantic tags, and the “Acquisition of animals for food, hunting” consists of top level 13 semantic tags. In addition to the AG category there are the categories “Animals” (AE) and “Plants” (AF), so that any missing information (semantic tag) for a food entity that is a recipe ingredient could be searched for in AE and AF, as part of nature animal or plant, respectively. The AE category consists of 15 semantic tags, while the AF category consist of 30 semantic tags. There are additional and more specific tags on a deeper hierarchical level within some of these tags, which are also utilised.

However, all of these food ontologies and food semantic resources were developed for a specific purpose. That implies that they cover varying aspects in the domain of food, having no way to be interlinked, which is an important task in order to enable interoperability , as well as extend their overall coverage of the domain.

2.3 Food Concepts Normalization

Concept normalization is the task of mapping free-form expressions to predefined concepts in a certain domain. In the last few years, food concepts normalization is an open research question that is highly researched by the food and nutrition science community, calling it food matching. For this reason, StandFood [70] was recently introduced, which is a semi-automatic system for classifying and describing foods according to a description and classification system, such as FoodEx2 proposed by the European Food Safety Agency (EFSA) [71]. It consists of three parts. The first involves a machine learning approach and classifies foods into four categories, with two for single foods: raw (r) and derivatives (d), and two for composite foods: simple (s) and aggregated (c). The second uses a natural language processing approach and probability theory to perform food concepts normalization. The third combines the result from the first and the second part by defining post-processing rules in order to improve the result for the classification part.

However, the food normalization process was based only on lexical similarity between the food concepts names, avoiding the semantic similarity between them. This means that when matching corresponding food concepts, only their textual representation (i.e. surface form) was taken into account, without including the semantic and contextual meaning that they convey.

2.4 Representation Learning

Representation Learning (RL) is a set of techniques that learn representations of input data by transforming it or extracting features from it. The input data may have some structure (e.g. free-form text or graph data) that either completely does not allow or significantly complicates the utilization of existing ML methods. With RL, the input is usually transformed into structures more suitable (i.e. vector spaces) for complex analyses, i.e. representations are learned.

2.4.1 Text Embedding Methods

In order to include the semantic information from the textual data in the representations, Mikolov et al. [72] proposed the *Word2Vec* model, which learns high-quality distributed vector representations for words. These vector representations are also known as embeddings and capture a large number of precise syntactic and semantic word relationships. Using this model, each token (i.e. word) is represented as a vector of continuous numbers. After introducing this model, multiple methods have followed suit with a focus on learning distributed vector representations.

For example, *GloVe* [73] is an unsupervised learning algorithm for obtaining vector representations for words. It is based on aggregated global word-word co-occurrence statistics from a textual corpus, and the resulting representations are linear substructures of the word vector space.

2.4.2 Graph Embedding Methods

In a similar fashion there are also methods, such as Node2vec [74] and Metapath2vec [75], which can be used for embedding problems that can be represented as graphs, especially social networks. Additionally, vector embeddings of multi-relational data such as TransE [76] are used for Information Extraction tasks.

Most of the embedding methods learn vector representations that are in the Euclidean vector spaces, which do not take into account the hierarchical structure that can be present in the data. For this reason, Nickel and Kela [77] introduced a new approach for learning hierarchical vector representations of data by embedding it into hyperbolic space, specifically a Poincaré ball, appropriately named *Poincaré embeddings*. The main idea behind the Poincaré embeddings is that it produces vector representations of symbolic data in a way such that the distance between two concepts in the embedding space reflects their semantic similarity, while the hierarchy of the concepts is captured by the norms of the vectors. Hence, the data is embedded in a hyperbolic space, which in this case represents a Poincaré ball.

What is common for all these approaches is that they provide more promising results than the traditional representations when they are used in some predictive or descriptive analyses by using ML algorithms.

2.5 Text Similarity Measures

One of the challenges while working on text similarity is that the same concept can be mentioned using phrases with a variety of structures, which is a consequence of how people express themselves. In order to combine the information for the same concept that is represented in different ways, we should apply text normalization methods. Text normalization methods are based on text similarity measures.

Text similarity measures operate on string sequences and give us a metric of similarity (or dissimilarity) between two text strings. Text similarity determines how distant two texts are both in surface (i.e. lexical similarity) and meaning (i.e. semantic similarity).

Normalization methods based on text similarity measures are well presented in [78], [79]. Several normalization methods that are based on ranking technique are available, with the goal to rank the candidate matches and then to find the most relevant match [80]. Normalization methods can also utilize machine learning (ML) algorithms to improve results, which was shown in the gene normalization task as part of BioCreative II [81] and BioCreative III [82]. Regarding the food and nutrition domain, methods for normalization of short text segments (e.g. names or descriptions of nutrients, food composition data, food consumption data) have recently been proposed [70], [83]–[85] by using two approaches: two approaches: (i) standard text similarity measures; and (ii) a modified version of Part of Speech (POS) tagging probability-weighted method, first proposed in [83].

2.5.1 Lexical similarity

Lexical similarity can be calculated either on the character or the word level. Most of the lexical similarity measures do not take into account the actual meaning behind words or the entire phrases in context, but focus on how many characters or words overlap.

Let D_1 and D_2 be two pieces of text. Some of the standard lexical similarity measures are [86]:

- The *Levenshtein distance* counts the number of deletions, insertions and substitutions necessary to turn D_1 into D_2.

- The *Optimal String Alignment distance* is like the Levenshtein distance but also allows transposition of adjacent characters. Each substring may be edited only once.

- The full *Damerau-Levenshtein distance* is like the optimal string alignment distance except that it allows for multiple edits on substrings.

- The *longest common substring* is defined as the longest string that can be obtained by pairing characters from D_1 and D_2 while keeping the order of characters intact.

- A *q-gram* is a subsequence of q consecutive characters of a string. If x (y) is the vector of counts of q-gram occurrences in D_1 (D_2), the q-gram distance is given by the sum over the absolute differences $|x_i - y_i|$.

- The *cosine distance* is computed as $1 - \frac{x \cdot y}{||x|| ||y||}$, where x and y were defined above.

- Let X be the set of unique q-grams in D_1 and Y the set of unique q-grams in D_2. The *Jaccard distance* is defined as $1 - \frac{|X \cap Y|}{|X \cup Y|}$.

- The *Jaro distance* is defined as $1 - \frac{1}{3}(w_1 \frac{m}{|D_1|} + w_2 \frac{m}{|D_2|} + w_3 \frac{(m-t)}{m})$, where $|D_i|$ indicates the number of characters in D_i, m is the number of character matches and t the number of transpositions of matching characters. The w_i are weights associated with the characters in D_1, characters in D_2 and with transpositions.

- The *Jaro-Winkler distance* is a correction of the Jaro distance. It uses a prefix scale p which gives more favourable ratings to strings that match from the beginning for a set prefix length l.

- The *skip-grams* are generalization of *n-grams* in which the components (typically words) need not be consecutive in the text, but may leave gaps that are skipped over.

2.5.2 Semantic similarity

Semantic similarity is a metric that defines the distance between two pieces of text based on their meaning or semantic content. Calculating semantic similarity is related to representation learning (i.e. learning embeddings), which has become an important research task for learning representation of symbolic data. The idea of representation learning is to represent each piece of text (e.g. word, sentence, paragraph, depending on the problem) as a vector of continuous numbers. In the case of learning word embeddings, the learned vector captures the context of a word in a piece of text, as well as semantic and syntactic similarity, relation with other words, etc. To find the similarity between two words, we should calculate the similarity between their vectors. To do this, we can find the angle between their vectors. The cosine distance between two words represented by their vectors $\mathbf{x}$ and $\mathbf{y}$ can be calculated using the following equation:

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}\mathbf{y}}{||\mathbf{x}||^2||\mathbf{y}||^2}. \tag{2.1}$$

Chapter 3

A Novel Food Named-Entity Recognition Method

This chapter presents a novel food Named-Entity Recognition Method called FoodIE [23]. It is a rule-based method, which at its core contains a rule engine based on computational linguistics rules and semantic information regarding the food entities of interest. The construction of the whole methodology is presented, followed by its evaluation and comparison with other existing food NER methods [25].

This chapter is adapted from [23], [25].

3.1 FoodIE: a Rule-Based Named-Entity Recognition Method for Food Information Extraction

To enable food-named entity recognition, in this chapter, we propose a rule-based approach, called FoodIE. It works with unstructured data (more specifically, with a recipe that includes textual data in the form of instructions on how to prepare the dish) and consists of four steps:

- Food-related text pre-processing
- Text POS-tagging and post-processing of the tag data set
- Semantic tagging of food tokens in the text
- Food-named entity recognition

The flowchart of the methodologay is presented in Figure 3.1. Further, we are going to explain each step in more detail.

3.1.1 Food-related text pre-processing

The pre-processing step takes into account the discrepancies that exist between the outputs of the taggers we are utilizing, *coreNLP tagger* from the R programming language [87] and the *UCREL Semantic Analysis System (USAS)* [60]. It is also used to remove any characters that are unknown to the taggers.

Firstly, quotation marks should be removed from the raw text, for the simple reason that they are treated differently by both used NLP libraries, causing a discrepancy.

Secondly, every white space sequence (including tabulation, newlines, etc.) is converted into a single white space to provide a consistent structure to the text.

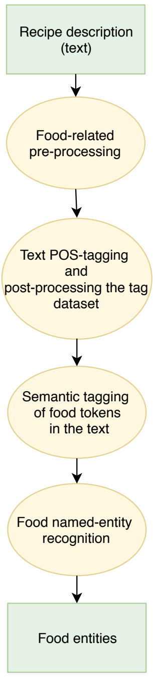

Recipe description (text)
Food-related pre-processing
Text POS-tagging and post-processing the tag dataset
Semantic tagging of food tokens in the text
Food named-entity recognition
Food entities

Figure 3.1: The flowchart of the FoodIE methodology.

Additionally, ASCII transliteration is performed, which means characters that are equivalent to ASCII characters are transliterated. An example of such characters is [è, ö, à], which are transliterated to [e, o, a], respectively.

Finally, fractions should be converted into real numbers. Usually, when a food-related text is written (e.g., recipe), fractions are used when discussing quantities. However, they are usually written in plain ASCII format and in a manner which is confusing to NLP taggers. For example, “2.5” is usually written as “2 1/2” in such texts. This does not bode well with *coreNLP* and the *USAS semantic tagger*. Thus, in the pre-processing step, all fractions are converted into the standard mathematical decimal notation for real numbers.

3.1.2 Text Part-Of-Speech (POS) tagging and post-processing of the tag set

To obtain the morphological information from a textual data, we use UCREL Semantic Analysis System (USAS) and coreNLP. One of the outputs of these tools are Part-Of-Speech tag, which describe the morphological nature of the words (whether it is a noun, verb, adjective, etc.).

The USAS semantic tagger provides word tokens associated with their POS tags, lemmas, and semantic tags. The semantic tags show semantic fields that group together word senses that are related at some level of generality with the same contextual concept. The groups include not only synonyms and antonyms but also hypernyms and hyponyms. More details about semantic tags can be found in [60], [69].

Furthermore, the same is done using the coreNLP library, which includes all of the above except semantic tags.

For example, the sentence “Heat the beef soup until it boils” is processed by both libraries. The results from the coreNLP library for the aforementioned example sentence are presented in Table 3.1, while the results from USAS are presented in Table 3.2. Observing the results presented in the tables, it is obvious that there is a discrepancy between the

Table 3.1: Tags obtained from coreNLP for one recipe sentence

Token ID	Token	Lemma	POS tag
1	Heat	heat	NN
2	the	the	DT
3	beef	beef	NN
4	soup	soup	NN
5	until	until	IN
6	it	it	PRP
7	boils	boil	VBZ
8	.	.	.

Table 3.2: Tags obtained from USAS for one recipe sentence.

Token ID	Token	Lemma	POS tag	Semantic tag 1	Semantic tag2
1	Heat	heat	VV0	O4.6+	AJ.03.c.02 [Heat]; AJ.03.c.02 [Heat]; AJ.03.c.02.a [Heating/making hot/warm];
2	the	the	AT	Z5	ZC [Grammatical Item];
3	beef	beef	NN1	F1	AG.01.d.03 [Beef]; AE.14.m.03 [Subfamily Bovinae (bovines)]; AE.14.m.03 [Subfamily Bovinae (bovines)];
4	soup	soup	NN1	F1	AG.01.n.02 [Soup/pottage]; AA.04.g.04 [Wave]; AA.11.h [Cloud];
5	until	until	CS	Z5	ZC [Grammatical Item];
6	it	it	PPH1	Z8	ZF [Pronoun];
7	boils	boil	VVZ	O4.6+ E3-	AJ.03.c.02.b [Action of boiling]; AJ.03.c.02.b [Action of boiling]; AJ.03.c.02.b [Action of boiling];
8	.	PUNC	YSTP	PUNC	NULL

POS tags for the token "Heat".

As is evident, both the USAS semantic tagger and the coreNLP library, do not provide perfect tags (e.g., sometimes verbs are misclassified as nouns, as is the case with the first token in the example given in Table 3.1). For this reason, the tags returned by both taggers are post-processed and modified using the following linguistic rules:

- If at least one of the taggers classify a token as a verb, mark it as a verb.

- If a discrepancy exists between the tags for a specific token, prioritize the tag given by the USAS semantic tagger.

- If a past participle form or a past simple form of a verb precedes and is adjacent to a noun, and it is classified as a verb, change the tag from verb to adjective.

Finally, we keep two versions of the modified tag set, one in each format. These modified tags in the coreNLP format and USAS format are presented in Table 3.3 and Table 3.4, respectively.

Table 3.3: Modified tags from coreNLP for one recipe sentence.

Token ID	Token	Lemma	POS tag
1	Heat	heat	VB
2	the	the	DT
3	beef	beef	NN
4	soup	soup	NN
5	until	until	IN
6	it	it	PRP
7	boils	boil	VBZ
8	.	.	.

Table 3.4: Modified tags from USAS for one recipe sentence.

Token ID	Token	Lemma	POS tag	Semantic Tag 1	Semantic tag 2
1	Heat	heat	VV0	O4.6+	AJ.03.c.02 [Heat]; AJ.03.c.02 [Heat]; AJ.03.c.02.a [Heating/making hot/warm];
2	the	the	AT	Z5	ZC [Grammatical Item];
3	beef	beef	NN1	F1	AG.01.d.03 [Beef]; AE.14.m.03 [Subfamily Bovinae (bovines)]; AE.14.m.03 [Subfamily Bovinae (bovines)];
4	soup	soup	NN1	F1	AG.01.n.02 [Soup/pottage]; AA.04.g.04 [Wave]; AA.11.h [Cloud];
5	until	until	CS	Z5	ZC [Grammatical Item];
6	it	it	PPH1	Z8	ZF [Pronoun];
7	boils	boil	VVZ	O4.6+ E3-	AJ.03.c.02.b [Action of boiling]; AJ.03.c.02.b [Action of boiling]; AJ.03.c.02.b [Action of boiling];
8	.	PUNC	YSTP	PUNC	NULL

3.1.3 Semantic tagging of food tokens in text

To define phrases in the text related to food entities, we first need to find tokens that are related to food entities. For this purpose, the USAS semantic tagger is utilized. Using it, a specific rule is defined to determine the food tokens in the text. Food tokens are predominantly nouns or adjectives, so we account for this as to improve the false positive rate, i.e. allowing a token to be categorized as a food token if and only if it is either a noun or an adjective. The decision rule combines three conditions using the following Boolean expression $((Condition_1\ OR\ Condition_2)\ AND\ Condition_3)$. If the expression is true, then the token is classified as a food token. For clarity, let us assume that t is a token and s_t is the semantic tag that is assigned to it using the USAS semantic tagger. Each condition is constructed using the following rules:

- $Condition_1$:
  - Food tag F(1|2|3|4), or
  - Living tag L(2|3), or
  - Substance tag (liquid and solid) O1.(1|2).
- $Condition_2$:
  - Body part tag B1, and
  - Not Linear order tag N4, and
  - Not Location and direction tag M6, and
  - Not Texture tag O4.5.
- $Condition_3$:
  - Not General Object tag O2, and
  - Not Quantities tag N5, and

  - Not Clothing tag B5, and
  - Not Equipment for food preparation tag AG.01.t.08, and
  - Not Container for food, place for storing food tag AG.01.u, and
  - Not Clothing tag AH.02.

More formally, using Boolean algebra, we can represent these rules as:

$Condition_1$:

$s_t \in \{\mathrm{F1}, \mathrm{F2}, \mathrm{F3}, \mathrm{F4}\} \vee s_t \in \{\mathrm{L2}, \mathrm{L3}\} \vee s_t \in \{\mathrm{O1.1}, \mathrm{O1.2}\}$

$Condition_2$:

$s_t = \mathrm{B1} \wedge s_t \neq \mathrm{N4} \wedge s_t \neq \mathrm{M6} \wedge s_t \neq \mathrm{O4.5}$

$Condition_3$:

$s_t \neq \mathrm{O2} \wedge s_t \neq \mathrm{N5} \wedge s_t \neq \mathrm{B5} \wedge s_t \neq \mathrm{AG.01.t.08} \wedge s_t \neq \mathrm{AG.01.u} \wedge s_t \neq \mathrm{AH.02}.$

$Rule_1$:

$(Condition_1 \vee Condition_2) \wedge Condition_3$

Additionally, we define one rule to determine object tokens. Determining the object tokens will further help us in the definition of food entities, mainly to avoid false positives. The rule consists of

- General Object tag O2, or
- Clothing tag B5, and
- Not Body Part tag B1, and
- Not Living tag L(2|3), and
- Not a food token as defined by the aforementioned first rule.

Using Boolean algebra, this rule is represented as

$Rule_2$:
$(s_t = \mathrm{O2} \vee s_t = \mathrm{B5}) \wedge s_t \neq \mathrm{B1} \wedge s_t \neq \mathrm{L2} \wedge s_t \neq \mathrm{L3} \wedge \neg \mathrm{Rule}_1.$

If this condition is met, the token is tagged as general object.
The single rule for defining a color noun consists of

- Color tag O4.3.

The rule for defining a color noun is then formally defined as

$Rule_3$:

$s_t = \mathrm{O4.3}.$

These tags are useful when food entities ending on a color, such as “egg whites” or “hash browns”, appear in the text, which indeed are to be treated as food entities.

At the end, one additional rule is constructed for defining what is explicitly disallowed to be the main token in a food entity, and is defined as

- Equipment for food preparation AG.01.t.08, and
- Container for food, place for storing food AG.01.u, and
- Clothing tag AH.02, and
- Temperature tag O4.6, and
- Measurement tag N3.

This rule can be represented as

$Rule_4$:

$$s_t = \text{AG.01.t.08} \wedge s_t = \text{AG.01.u} \wedge s_t = \text{AH.02} \wedge s_t = \text{O4.6} \wedge s_t = \text{N3}.$$

This rule is utilized when isolating entities that could be potential false positives. An example of this would be “oil temperature” or “cake pan”. Additionally, there are some manually added resources in this disallowed category, which frequently occur in the texts.

3.1.4 Food-named entity recognition

To obtain food chunks, we used the modified tag set from the USAS semantic tagger obtained in Subsection 3.1.2 in combination with the food tokens obtained in Subsection 3.1.3. The process of food-named entity recognition consists of three steps.

Firstly, we iterate through every food token which we extracted previously from the text, and for each token we define a set of rules that constitute a food entity.

Adjacent to the left of the food token we allow chaining of adjectives (JJ), nouns (NN), proper nouns (NP), genitive tag (GE), unknown tags (Z99) and general tokens tagged as food, but explicitly omit general objects. The purpose of including the unknown POS tag (Z99) is to catch tokens that do not concisely fall into one of the tags in the standard POS tag set, yet are still of importance to the semantics of the food entity. Such an example would be “Colby-Jack cheese”, whose POS tags are Z99 and NN, respectively.

Adjacent to the right, the logic is the same, differing only by allowing a general object to be a part of the food entity and tokens that have been tagged as a color noun by the rule engine. We also keep track not to use a token twice.

Then, to determine if it truly is a food entity chunk or just a chunk related to food but not a food entity in and of itself, we check the last token of the chunk. The whole chunk is discarded if the last token is:

- A noun (starts with NN) and a general non-food object, or
- in the disallowed category as defined by the rule engine, or
- in the disallowed category as defined by the resources.

Some examples where this would be a false positive are “muffin liner”, “casserole dish” or “egg timer”. If this check passes and the last token is not a general object, we mark each token in the new food chunk with an index unique to the whole chunk and continue iterating through the remaining food tokens.

After the first step, we must now concatenate all relevant information for each food entity. For each indexed food entity, we join all the instances into one entry, thus creating a vector where each token is its own entry, except for the food entities which are represented as one entry. If initially we had a vector of tokens such as [Chop, the, hot, Italian, sausage, into, pieces, .] the output would be [Chop, the, hot Italian sausage, into, pieces, .]. This also applies to other relevant information we might want to track, such as lemmas, POS tags, sentence indexes or even individual token indexes.

For additional robustness, we perform a check to assure that each food chunk we have isolated indeed contains a food token, and that the token is marked under some food chunk. For this we only mark a chunk as a food entity if it contains at least one word that has previously been tagged as a food token and has been indexed as part of the respective chunk as well.

3.1.5 Evaluation

The evaluation was performed manually, since there is no pre-existing method to evaluate such a text corpus. To avoid any kind of bias when evaluating food-related text, one person was tasked with manually performing food chunk extraction from each individual text, while another person cross-referenced those manually obtained chunks with the ones obtained from FoodIE. Using this method, a figure for true positives (TPs), false negatives (FNs) and false positives (FPs) was procured, while it was decided that the category true negative was not applicable to the nature of the problem and its evaluation. Additionally, it was decided that a “partial (inconclusive)” category was necessary, as some of the food chunks were incomplete, but nevertheless caught, thus including significant information. This category encompasses all the extracted food chunks which were caught, but missed at least one token. An example would be “bell pepper”, where FoodIE would only catch “pepper”.

We would like to compare the results using the model presented in [59], but we were unable to obtain the requested model and corpus. We provide a small example of comparing FoodIE with drNER [22], in order to show that they provide food entities on a different level, so a fair comparison cannot be made.

While the evaluation was being done, we kept track of all the False Negative instances and have constructed a resource set that will improve the performance of FoodIE in future implementations.

Data. Firstly, a total of 200 recipes were processed and evaluated. The original 100 recipes, which were analyzed and upon which the rule engine was built, were taken into consideration, as well as 100 new recipes which had not been analyzed beforehand. The recipes were taken from two separate user-based sites, Allrecipes [1] and MyRecipes [2], where there is no standardized format for the recipe description. This was chosen as such to ensure that the linguistic constructs utilized in each written piece varied and had no pattern behind them. The texts were chosen from a variety of topics, as to provide further diversity.

Secondly, we selected 1,000 independently obtained recipes from Allrecipes [88], which is the largest food-focused social network, where everyone plays a part in helping cooks discover and share home cooking. We selected Allrecipes because there is no limitation as to who can post recipes, so we have variability in how users express themselves. The recipes were selected from five recipe categories: Appetizers and snacks, Breakfast and Lunch, Dessert, Dinner, and Drinks. From each recipe category 200 recipes were included in the evaluation set.

[1] `https://www.allrecipes.com/`

[2] `https://www.myrecipes.com/`

The evaluation data sets, including the obtained results, are publicly available at a repository [3] listed in Appendix A.

Results and discussion. The results for TPs, FPs, and FNs of evaluating the FoodIE using the data set of 200 recipes are presented in Table 3.5. The group "Partial (Inconclusive)" was left out of these evaluations, as some would argue they should be counted as TPs, while others that they should be included in the FNs. Some examples included here are: "empty passion fruit juice", "cinnamon" and "soda", where the actual food entity chunks would be "passion fruit juice", "cinnamon sticks" and "club soda", respectively. These are mostly due to the dual nature of words, meaning that a word that is a synonym of both a noun and a verb or an adjective and a verb occurs. For such words, the tagger sometimes incorrectly classifies the tokens. In these examples, "empty" is tagged as an adjective, whereas in context it is, in fact, a verb. The same explanation holds for the other two examples. For these reasons, when the evaluation metrics were calculated, this category was simply omitted. Moreover, even if they are grouped with either TPs or FNs, this does not significantly affect the results.

Regarding the FN category (type II error), there were some specific patterns that produced the most instances. One very simple type of a FN instance is where the author of the text refers to a specific food using the brand name, such as "allspice" or "Jägermeister". These are difficult to catch if there is no additional information following the brand name. However, if the user includes the general classification of the branded food, FoodIE will catch it. An example of this would be by simply writing "Jägermeister liqueur". Another instance of a type II error is when the POS taggers give incorrect tags, as was the case with some "Partial (Inconclusive)" instances. An example of this is when the tagger misses chunks such as "mint leaves" and "sweet glazes", where both "leaves" and "glazes" are incorrectly classified as verbs when, in this context, they should be tagged as nouns. Another example would be when the semantic tagger incorrectly classifies some token within the given context, such as "date" being classified as a noun meaning day of year, as opposed to it being a certain fruit. Furthermore, FNs exist which are simply due to the rarity of the food, such as "kefir", "couscous" or "stevia", the last one being of immense importance to people suffering from diabetes, as it is a safe sugar substitute. Another category of type II errors is due to the fact that some foods are often referred to by their colloquial name, such as "half-and-half" and "spring greens". The final category of this type of error is where spelling variations exist for a single food, such as "eggnog", "egg nog", "egg-nog". These are very difficult, if not impossible, to correctly predict since grammatical and morphological styles vary with each user, which extend as far as including simply improper use of the English language. This is a separate problem in and of itself, i.e. spellchecking and spelling correction.

The second type of error to discuss is the FP category (type I error), which is often due to the existence of objects that are not foods, but are closely related to food entities. These include instances such as "dollop" or "milk frother", where the first example has a meaning very closely related to food, thus making it difficult to distinguish using the semantic tags. The second chunk is simply an instrument related to food and cooking, while being rare enough such that the semantic tagger does not classify it properly as an object.

Using the results reported in Table 3.5, the evaluation metrics for F_1 score, precision, and recall, are presented in Table 3.6.

The results from evaluating FoodIE on the data set with 1000 recipes are reported in Tables 3.7 and 3.8.

Comparing the results obtained from the evaluations (Tables 3.6 and 3.8), we can conclude that FoodIE behaves consistently. Evaluating the data set with 200 recipes,

[3] `http://cs.ijs.si/repository/FoodIE/FoodIE_datasets.zip`

Table 3.5: Predictions from FoodIE on 200 recipes.

True Positive (TP)	3063
False Positive (FP)	75
False Negative (FN)	185
Partial (Inconclusive)	97

Table 3.6: Evaluation metrics for FoodIE on 200 recipes.

F_1 Score	Precision	Recall
0.9593	0.9761	0.9430

which consists of 100 recipes that were analyzed to build the rule engine and 100 new recipes that were not analyzed beforehand, we obtained 0.9761 precision, 0.9430 recall, and 0.9593 F_1 score. Furthermore, by evaluating it on a data set that consists of 1000 new recipes, it obtained 0.9780 for precision, 0.9437 for recall, and 0.9605 for F_1 score. Comparing these results provides that FoodIE gives very promising and consistent results.

We also provided the TPs, FPs, FNs, and Partial predictions, together with the evaluation metrics for each recipe category separately (Table 3.9). Using them, we can see that Dinner category provides most FNs (223), while the Breakfast/lunch category provides the least FNs (82). Regarding the FNs, the Breakfast/lunch category provides the most FPs (108), while the Drinks category provides the least FPs (31). Looking at the results, it is evident that FoodIE retains the aforementioned consistency, even when comparing the evaluation metrics from each category between themselves.

In Table 3.10, we present the results obtained for 10 sentences (i.e evidence-based dietary recommendations) previously used in [22], [89], in order to present the difference between FoodIE and drNER. Semicolon was used to split separate food entities. Using the table, we can see that drNER and FoodIE provide results on a different level. For example, let us consider the sixth recommendation. drNER extracted only one food entity, which is “Milk, cheese, yogurt and other dairy products”, while FoodIE extracted four separate food entities, i.e. “Milk”, “cheese”, “yogurt”, and “other dairy products”. From this, it follows that FoodIE provides more precise results, which means it can also be used as a post-processing tool for drNER in order to extract the food entities on an individual level.

The performance of the rule-based system FoodIE heavily depends on the taggers used, so the improvement of the qualities of the POS-tagging and semantic tagging methods will also improve the evaluation metrics for FoodIE.

Table 3.7: Predictions from FoodIE on 1000 recipes.

True Positive (TP)	11461
False Positive (FP)	258
False Negative (FN)	684
Partial (Inconclusive)	359

Table 3.8: Evaluation metrics for FoodIE on 1000 recipes.

F_1 Score	Precision	Recall
0.9605	0.9780	0.9437

Table 3.9: Predictions from FoodIE and evaluation metrics for each recipe category.

Recipe category	TP	FP	FN	Partial	F_1 Score	Precision	Recall
Appetizers/snacks	2147	27	162	45	0.9578	0.9876	0.9298
Breakfast/lunch	2443	33	82	108	0.9770	0.9876	0.9675
Desserts	2612	87	127	124	0.9607	0.9678	0.9536
Dinner	3176	47	223	51	0.9592	0.9854	0.9344
Drinks	1083	64	90	31	0.9336	0.9442	0.9233

3.1.6 Conclusions

To extract food entities from unstructured textual data, we propose a rule-based named-entity recognition method for food information extraction, called FoodIE. It is a rule engine, where the rules are based on computational linguistics and semantic information that describe the food entities. Evaluation showed that FoodIE behaves consistently using different independent evaluation data sets and very promising results have been achieved.

To the best of our knowledge, there is a limited number of NLP tools that can be used for IE of food entities. Moreover, there is a lack of annotated corpora that can be used to train corpus-based NER methods. Motivated by the evaluation results obtained, we are planning to use it in order to build an annotated corpus that can be further used for extracting food entities together with their relations to other biomedical entities. By performing this, we can easily follow the new knowledge that comes rapidly with each day with new scientifically published papers aimed at improving public health.

Table 3.10: Food entities extracted by drNER and FoodIE.

	Recommendation	**drNER**	**FoodIE**
1.	Good sources of magnesium are: fruits or vegetables, nuts, peas and beans, soy products, whole grains and milk.	fruits or vegetables, nuts, peas and beans; soy products; whole grains and milk	fruits; vegetables; nuts; peas; beans; whole grains; milk
2.	The RDAs for Mg are 300 mg for young women and 350 mg for young men.	-	-
3.	Increase potassium by ordering a salad, extra steamed or roasted vegetables, bean-based dishes fruit salads, and low-fat milk instead of soda.	salad; extra steamed or roasted vegetables; fruit salads; low-fat milk	salad; roasted vegetables; bean-based dishes; fruit salads; low-fat milk; soda
4.	Babies need protein about 10 g a day.	-	-
5.	1 teaspoon of table salt contains 2300 mg of sodium.	table salt	table salt
6.	Milk, cheese, yogurt and other dairy products are good sources of calcium and protein, plus many other vitamins and minerals.	Milk, cheese, yogurt and other dairy products	Milk; cheese; yogurt; other dairy products
7.	Breast milk provides sufficient zinc, 2 mg/day for the first 4-6 months of life.	Breast milk	milk
8.	If you're trying to get more omega-3, you might choose salmon, tuna or eggs enriched with omega-3.	salmon, tuna; eggs	salmon; tuna; eggs
9.	If you need to get more fiber, look to beans, vegetables, nuts and legumes.	beans, vegetables, nuts, and legumes	beans; vegetables; nuts; legumes
10.	Excellent sources of alpha-linolenic acid, ALA, include flaxseeds and walnuts.	flaxseeds and walnuts	alpha-linolenic acid; flaxseeds; walnuts

3.2 Comparison of 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 repository[4] where the data analysis script is publicly available is given in Appendix A.

3.2.1 Benchmarking data set

The aforementioned data set consisting of 1000 recipes is used as the benchmarking data set.

Every mention of a food entity in the recipe was manually extracted and annotated by two human experts. To reduce potential human bias, one person was tasked with extracting each food entity in each recipe, while the other person was given these annotations to independently double check the extracted entities. The result was a ground truth data set consisting of 1,000 recipes, where each recipe consists of all food entities in it [26]. 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 [53]. The repository [5] where the ground truth data set is publicly available is provided in Appendix A.

3.2.2 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), considering every iteration as a different NER method.

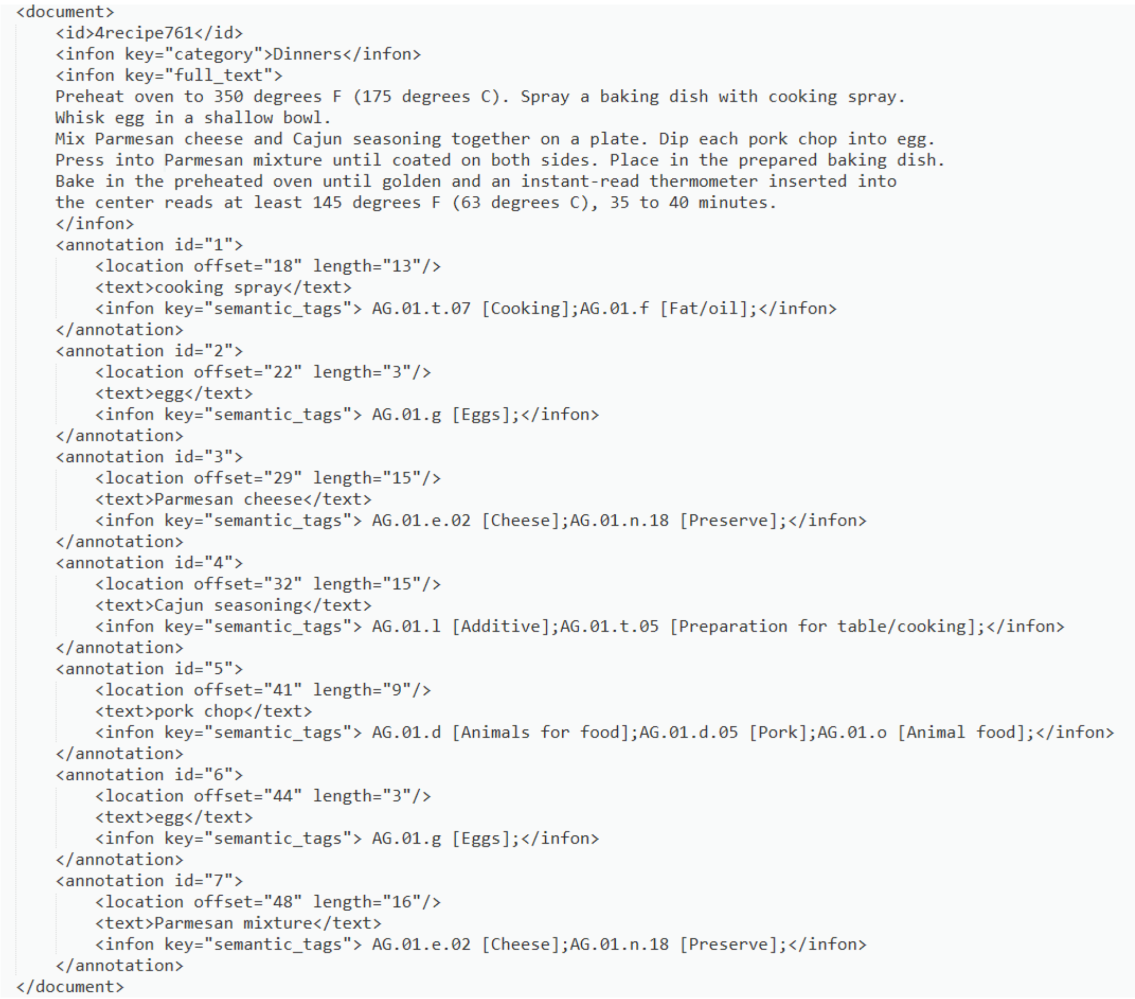
This resulted in a total of four sets of extractions which are the compared: FoodIE, NCBO (FoodOn), NCBO (OntoFood), and NCBO (SNOMED CT).

The results from FoodIE were organized in the BioC format, the same format as the ground truth data set. The BioC format for one recipe and its annotations as processed by FoodIE are presented in Figure 3.2. 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 same recipe used as an example above, represented by a NCBO annotation, is presented in Figure 3.3. In the figure, there 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

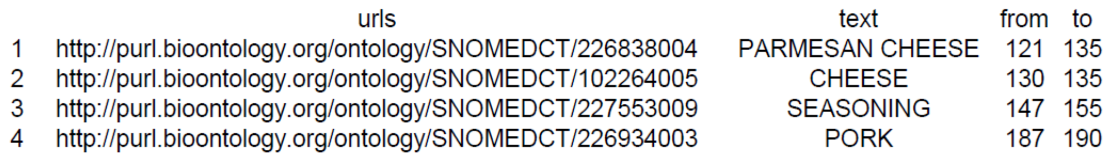
[4] https://github.com/GorjanP/food_NER_comparison_script

[5] http://cs.ijs.si/repository/FoodBase/foodbase.zip



```
<document>
    <id>4recipe761</id>
    <infon key="category">Dinners</infon>
    <infon 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.
    </infon>
    <annotation id="1">
        <location offset="18" length="13"/>
        <text>cooking spray</text>
        <infon key="semantic_tags"> AG.01.t.07 [Cooking];AG.01.f [Fat/oil];</infon>
    </annotation>
    <annotation id="2">
        <location offset="22" length="3"/>
        <text>egg</text>
        <infon key="semantic_tags"> AG.01.g [Eggs];</infon>
    </annotation>
    <annotation id="3">
        <location offset="29" length="15"/>
        <text>Parmesan cheese</text>
        <infon key="semantic_tags"> AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];</infon>
    </annotation>
    <annotation id="4">
        <location offset="32" length="15"/>
        <text>Cajun seasoning</text>
        <infon key="semantic_tags"> AG.01.l [Additive];AG.01.t.05 [Preparation for table/cooking];</infon>
    </annotation>
    <annotation id="5">
        <location offset="41" length="9"/>
        <text>pork chop</text>
        <infon key="semantic_tags"> AG.01.d [Animals for food];AG.01.d.05 [Pork];AG.01.o [Animal food];</infon>
    </annotation>
    <annotation id="6">
        <location offset="44" length="3"/>
        <text>egg</text>
        <infon key="semantic_tags"> AG.01.g [Eggs];</infon>
    </annotation>
    <annotation id="7">
        <location offset="48" length="16"/>
        <text>Parmesan mixture</text>
        <infon key="semantic_tags"> AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];</infon>
    </annotation>
</document>
```

Figure 3.2: Example recipe in the BioC format, as annotated by FoodIE.



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

Figure 3.3: Example recipe as annotated by NCBO using the SNOMED CT ontology.

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 data set.

3.2.3 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 is 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:

- Precision - This evaluation metric is defined as $Precision = \frac{TP}{TP+FP}$. Precision evaluates the fraction of correctly extracted entities over 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).

- F_1 Score - This evaluation metric is defined as $F_1 Score = \frac{2TP}{2TP+FP+FN}$. The F_1 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 a 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 are less of a concern, precision is the more appropriate metric. However, if we wanted to make a general conclusion that takes both into account, we should use the F1 score. It is a weighted average of the Precision and Recall, which takes both false positives and false

Table 3.11: Evaluation metrics when comparing FoodIE, NCBO (SNOMED CT), NCBO (OntoFood) and NCBO (FoodOn) on 1,000 recipes.

	FoodIE	NCBO (SNOMED CT)	NCBO (OntoFood)	NCBO (FoodON)
F_1 Score	0.9605	0.6375	0.3262	0.6390
Precision	0.9780	0.9153	0.8548	0.7922
Recall	0.9437	0.4891	0.2016	0.5354

Table 3.12: Number of missed recipes for FoodIE, NCBO (SNOMED CT), NCBO (OntoFood) and NCBO (FoodOn) out of a total of 1000).

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

negatives into account. It also has some issues since 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, since they offer more detailed insights about the NER that is analyzed.

3.2.4 Results

The results from counting the entities of each match type are presented in Figure 3.4, while the results from the calculated evaluation metrics are presented in Table 3.11 and in Figure 3.5.

It is interesting to note that not all ontologies provided annotations for each recipe. The number of missed recipes for each NER method is given in Table 3.12.

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.

3.2.5 Discussion

Match types. Observing the comparison results in 3.4, 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 3.2.3, 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 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

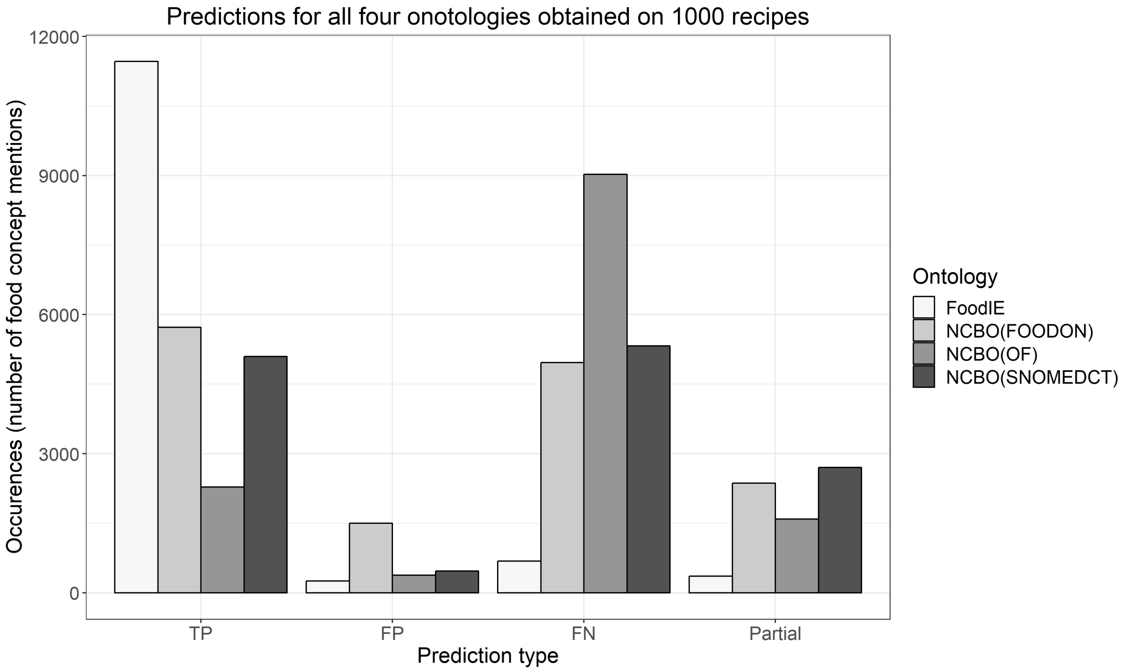


Figure 3.4: Evaluation results graph when comparing FoodIE, NCBO (SNOMED CT), NCBO (OntoFood) and NCBO (FoodOn) on 1,000 recipes.

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.

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 (0.9605), SNOMED CT (0.6375), OntoFood (0.3262), and FoodON (0.6390). Since the metric is on a scale from 0 to 1, 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: 0.323, 0.6343, and 0.3215, 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.

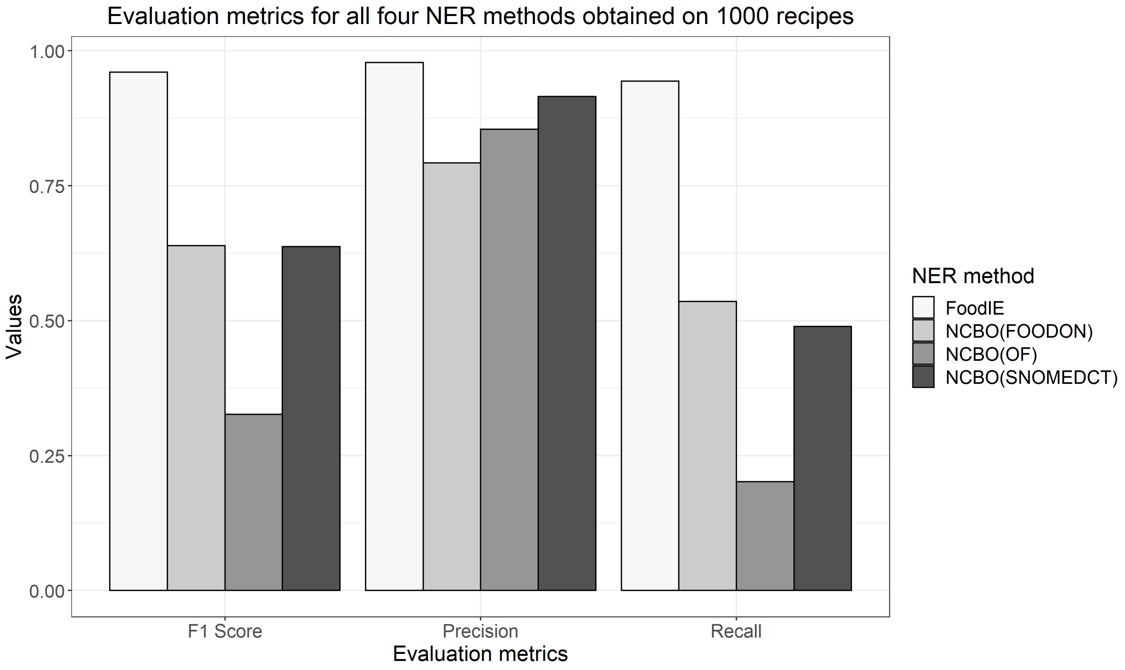


Figure 3.5: Evaluation metrics graph when comparing FoodIE, NCBO (SNOMED CT), NCBO (OntoFood) and NCBO (FoodOn) on 1,000 recipes.

3.2.6 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).

Chapter 4

Food Data Normalization

The first topic of discussion in this chapter is the first data set consisting of annotated food recipes, named FoodBase [26]. It consists of two parts, curated and un-curated. Both are based on the FoodIE NER method, with the difference being that the curated version is manually edited to ensure it contains ground true annotations, while the un-curated version consists of annotations directly given by FoodIE. The former consists of 1,000 annotated recipes, while the latter consists of around 22,000 recipes.

Second, an exploration of the LanguaL language for describing foods is presented. The study utilizes embedding techniques to explore the coverage and to see if the different standards within the LanguaL language are linked together [29].

Finally, a food data normalization method that performs food concept mapping across different food semantic resources is presented. The methodology, named FoodOntoMap [27], is based on the use of food NER methods to perform the mapping.

This chapter is adapted from [26], [27], [29].

4.1 FoodBase Corpus: a New Resource of Annotated Food Entities

4.1.1 Methods and Materials

In this section, we present an extension to the rule-based NER FoodIE [23], which is described in Chapter 3. First, we briefly describe the additional step aimed at semantic annotation of the extracted food entities and then we focus on its evaluation.

The recipe selection process is already described in Chapter 3, while the Hansard corpus [69] and its semantic tags are already described in Chapter 2.

FoodIE extension. For the aims of creating the FoodBase corpus, we added an additional step to the end of the FoodIE pipeline:

- **Semantic annotation of the extracted food entities:** Here, the Hansard semantic tags are grouped within each token for each food chunk, with the goal of representing the food concept in its entirety.

The flowchart of the extended methodology is presented in Figure 4.1. More details about the first four steps have already been presented in our previous work [23] or in Chapter 3. However, in this chapter, we will focus on the evaluation of the extended FoodIE methodology as this is the crucial step in building the annotated corpus. An example of running FoodIE on one recipe is explained in [23], step by step. Then we will describe the new step of semantic annotation of the extracted food entities.

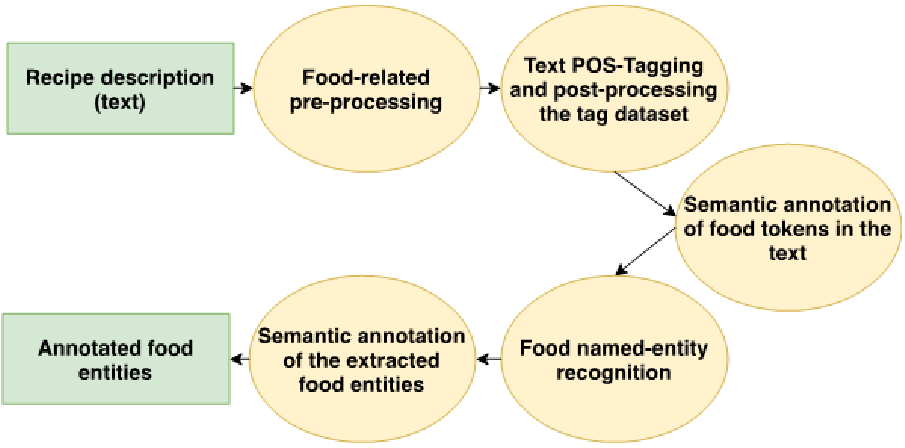

Recipe description (text)
Food-related pre-processing
Text POS-Tagging and post-processing the tag dataset
Semantic annotation of food tokens in the text
Food named-entity recognition
Semantic annotation of the extracted food entities
Annotated food entities

Figure 4.1: Flowchart of the extended FoodIE methodology

4.1.2 Evaluation of the extended FoodIE methodology

Once the information about the recipes was selected from Allrecipes, we asked a person to manually extract food chunks from the description of each recipe. A food chunk is a contiguous collection of tokens which describe a single food concept. Then we run the first two steps of FoodIE to obtain automatically extracted food chunks from the description of the same recipes. To avoid any kind of bias when comparing the food chunks extracted manually and automatically by FoodIE, another person was asked to cross-reference the manually obtained chunks with the ones obtained by FoodIE. Using this method, True Positives (TPs), False Positives (FPs) and False Negatives (FNs) were counted, while it was decided that the category True Negative (TN) is not applicable to the nature of the problem and its evaluation. In our case, TP and FP mean outcomes where FoodIE correctly or incorrectly predicted the positive class, respectively. Similarly, a FN means an outcome where FoodIE incorrectly predicted the negative class. In addition to the results for TP, FP and FN, the results for "Partial (Inconclusive)" are presented. This group of outcomes includes evaluations that could be either TP or FP/FN. For example, in the text segments "empty passion fruit juice", "cinnamon" and "soda" the actual food entity chunks are "passion fruit juice", "cinnamon sticks" and "club soda", respectively. These occurrences are mostly due to the dual nature of words, meaning that a word is a synonym for both a noun and a verb or a synonym for an adjective and a verb. For such words, the FoodIE tagger sometimes incorrectly classifies the tokens. In these examples, "empty" is tagged as an adjective, whereas in this context it is a verb. The explanation is almost identical for the other two examples. For these reasons, when the evaluation metrics were being calculated, the 'Partial (Inconclusive)' category was omitted. Moreover, even if they are classified either as TP or as FP/FN, they would not significantly affect the results. We performed this kind of the evaluation in our previous study [23] since there is no pre-existing method to evaluate such a text corpus.

Checking the concept. First, a subset of 200 recipes out of 1,000 were processed and evaluated. From each category, we selected 40 recipes. More details about the predictions are presented in [23].

Regarding the FN category (type II error), there were some specific patterns that produced the most instances. One very simple type of an FN instance is where the author of the text refers to a specific food using the brand name, e.g. "Jägermeister". These are difficult to catch if there is no additional information following the brand name. However, if the user includes the general classification of the branded food, FoodIE will correctly classify it. An example of this would be by simply writing “Jägermeister liquor”. Another instance of a type II error is when the POS taggers give incorrect tags as was the case with some “Partial (Inconclusive)” instances. An example of this is when the tagger misses chunks such as “mint leaves” and “sweet glazes”, where both “leaves” and “glazes” are incorrectly classified as verbs when in this context they should be tagged as nouns. Another example would be when the semantic tagger incorrectly classifies some token within the given context, such as “date” being classified as a noun meaning day of year, as opposed to it being a certain fruit. Furthermore, FNs exist which are simply due to the rarity of the food, such as “kefir”, “couscous” or “stevia”, the last one being of great importance to people suffering from diabetes, as it is a safe sugar substitute. Another category of type II errors occurs due to the fact that some foods are often referred to by their colloquial name, such as “half-and-half” and “spring greens”. The final category of this type of error is where there exist spelling variations for a single food, such as “eggnog”, “egg nog”, “egg-nog”. These are very difficult, if not impossible, to correctly predict since grammatical and morphological styles vary with each user, which extend as far as including simply improper use of the English language. This is a separate problem in and of itself, i.e. spellchecking and spelling correction.

The second type of error to discuss is the FP category (type I error), which is often due to the existence of objects that are not foods but are closely related to food entities. These include instances, such as “dollop” or “milk frother”, where the first example has a meaning very closely related to food, thus making it difficult to distinguish using the semantic tags. The second chunk is simply an instrument related to food and cooking, while being rare enough such that the semantic tagger does not classify it properly as an object.

Second trial. Once the effectiveness of the concept was evaluated on 200 recipes, the complete set of 1,000 recipes was processed and evaluated, and predictions for are presented in [23].

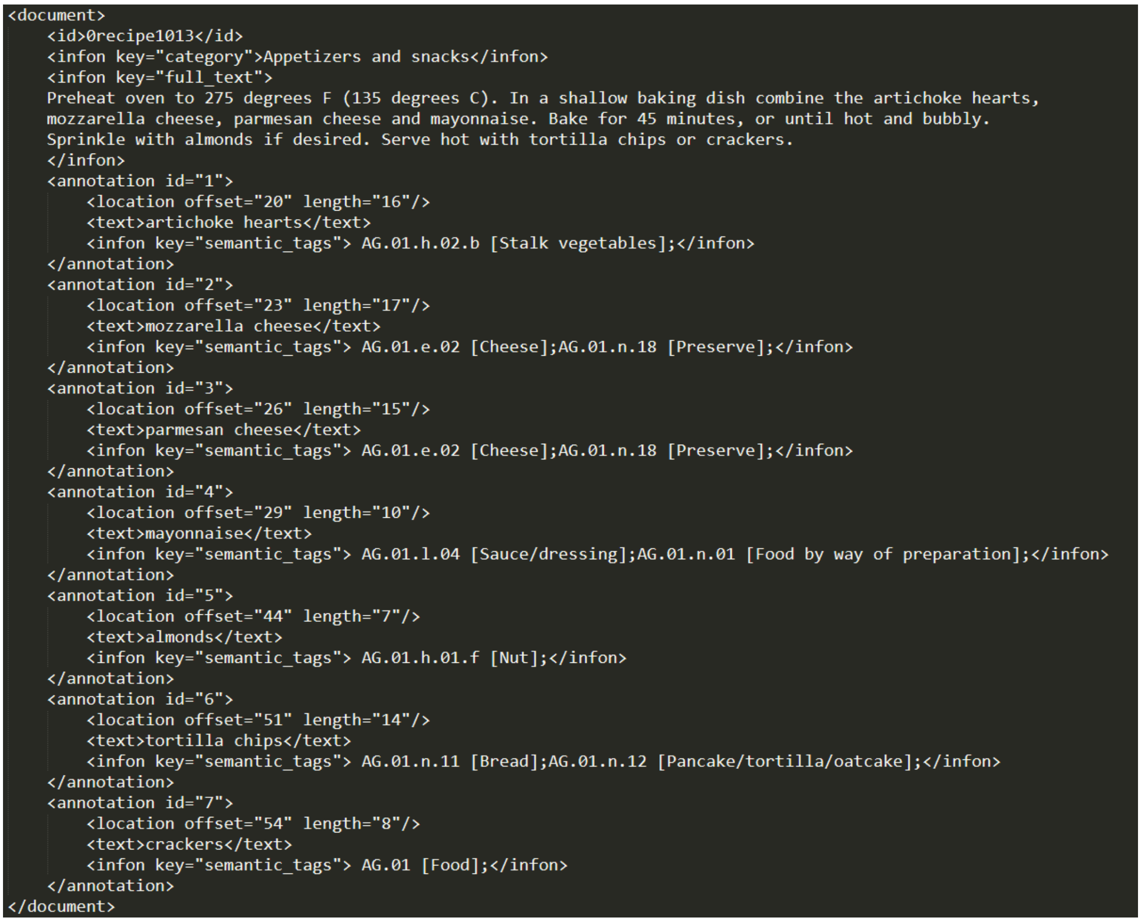
Comparing the evaluation metrics for 200 and 1,000 recipes presented in [23], we can conclude that FoodIE behaves consistently. Evaluating the data set with 200 recipes, which consists of 100 recipes that were analysed to build the rule engine and 100 new recipes that were not analysed beforehand, we obtained a precision of 0.9761, a recall of 0.9430, and a F1 score of 0.9593. Furthermore, by evaluating it on the data set of 1000 new recipes, we obtained 0.9780 for precision, 0.9437 for recall, and 0.9605 for F1 score. From these results we can conclude that FoodIE gives very promising and consistent results.

4.1.2.1 Semantic annotation of the extracted food entities

Once food entities were extracted using FoodIE, we annotated each of them using the semantic tags provided by the Hansard corpus. For this reason, annotations that are assigned to each food chunk are the sematic tags that belong to the tokens from which the chunk is constructed. As we explained before, these tags come only from three general Hansard corpus categories, i.e. “Food and drink” (AG), “Animals” (AE), and “Plants” (AF). When a selected entity recognized as a food entity cannot be annotated with any semantic tag from the “Food and drink category”, a tag from either “Animals” or “Plants” is used. Moreover, when no semantic tag can be associated to the food entity, it is assigned to the top food level hierarchy, i.e. “AG.01[Food]”.

Examples include:

- “grilled chicken” obtains the semantic tags AG.01.t.07[Cooking] / AG.01.d.06[Fowls],
- “tortilla chips” obtains AG.01.n.11[Bread] / AG.01.n.12[Pancake/tortilla/oatcake],
- “dry ranch salad dressing mix” obtains AG.01.h.02 [Vegetables] / AG.01.m [Substances for food preparation] / AG.01.n.09 [Prepared vegetables and dishes], and
- “cauliflower” obtains AG.01.h.02.d [Cabbage/kale].



```
<document>
    <id>0recipe1013</id>
    <infon key="category">Appetizers and snacks</infon>
    <infon key="full_text">
    Preheat oven to 275 degrees F (135 degrees C). In a shallow baking dish combine the artichoke hearts,
    mozzarella cheese, parmesan cheese and mayonnaise. Bake for 45 minutes, or until hot and bubbly.
    Sprinkle with almonds if desired. Serve hot with tortilla chips or crackers.
    </infon>
    <annotation id="1">
        <location offset="20" length="16"/>
        <text>artichoke hearts</text>
        <infon key="semantic_tags"> AG.01.h.02.b [Stalk vegetables];</infon>
    </annotation>
    <annotation id="2">
        <location offset="23" length="17"/>
        <text>mozzarella cheese</text>
        <infon key="semantic_tags"> AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];</infon>
    </annotation>
    <annotation id="3">
        <location offset="26" length="15"/>
        <text>parmesan cheese</text>
        <infon key="semantic_tags"> AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];</infon>
    </annotation>
    <annotation id="4">
        <location offset="29" length="10"/>
        <text>mayonnaise</text>
        <infon key="semantic_tags"> AG.01.l.04 [Sauce/dressing];AG.01.n.01 [Food by way of preparation];</infon>
    </annotation>
    <annotation id="5">
        <location offset="44" length="7"/>
        <text>almonds</text>
        <infon key="semantic_tags"> AG.01.h.01.f [Nut];</infon>
    </annotation>
    <annotation id="6">
        <location offset="51" length="14"/>
        <text>tortilla chips</text>
        <infon key="semantic_tags"> AG.01.n.11 [Bread];AG.01.n.12 [Pancake/tortilla/oatcake];</infon>
    </annotation>
    <annotation id="7">
        <location offset="54" length="8"/>
        <text>crackers</text>
        <infon key="semantic_tags"> AG.01 [Food];</infon>
    </annotation>
</document>
```

Figure 4.2: Annotated recipe from "Appetizers and snacks" category presented in the BioC format. For the recipe presented in this figure, all the extracted food concepts are presented, along with their respective semantic tags and their location in the raw recipe text.

Manual evaluation. Semantic annotations obtained by FoodIE were manually evaluated. Food entities reported as FPs were manually excluded from the corpus, while the food entities reported as FNs were included in the corpus. This was done in order to obtain a good benchmarking data set, which contains all food entities that are present in the data set of 1000 randomly selected recipes from five main dish categories. Furthermore, apart from excluding FPs and including FNs, the annotated semantic tags were double-checked. During this process all the incorrect semantic tags were removed, while all the missing semantic tags were added to specific food entities.

Annotation format. We decided to annotate the extracted information using the BioC format [53], which has been originally proposed by biomedical NLP and text mining tools. It is a simple XML-based format aimed for sharing text data and annotations, with the goals of simplicity, interoperability, and broad use and reuse. In Figure 4.2, a selected recipe is presented in the BioC format.

Post-processing of the annotated semantic tags. While manually evaluating and correcting the semantic annotations generated by FoodIE, the food entities reported as FNs were incorporated into the FoodIE rule engine as a resource, with the goal to improve its performance. This means that the new version of the FoodIE rule engine is more robust, since it does not incorrectly produce the FNs that were manually added as a resource. In addition to this, there were some specific instances where the semantic tags themselves needed a modification in some way. For example, the semantic tag "AG.01.af [Tea manufacture]" incorrectly appears every time the token "mixture" is present in the food entity chunk, so it is removed from the list of semantic tags for that food chunk. Another example of this is the semantic tag "AG.01.ae.03 [Brewing]" that incorrectly appears whenever the token "mashed" is present. If the food entity does not contain any semantic meaning relevant to these semantic tags, the tag is removed. In addition to this, some semantic tags were very vague. The semantic tagger occasionally did not tag the food entity "water" as such, but just provided the tag "AG.01 [Food]". In such cases, the tag "AG.01.z [Water]" was manually added for that food entity, while the original one was removed. Such omissions and inclusions, although rare (<5%), were performed as needed.

Justification of using semantic tags from Hansard corpus. Before we selected the semantic tags from the Hansard corpus that are used to describe the extracted food entities, we also tried several other knowledge resources for identifying food entities using the set of 1000 recipes. Firstly, our recently published rule-based food-named entity recognition method, named FoodIE, was used to extract food entities for each recipe. Using it, semantic tags from the Hansard corpus were assigned to each food concept. Then, additionally the NCBO annotator [61] working with the food ontologies that are available in the BioPortal (i.e. FoodOn [90], OntoFood, and SNOMED CT [68]) was used to extract food entities from each recipe. It is interesting to note here that the SNOMED CT is also a part of the Unified Medical Language Systems (UMLS) [21]. The NCBO Annotator is a web service that annotates text provided by the user by using relevant ontology concepts. It is available as part of the BioPortal software services [62]. The annotation workflow is based on an efficient syntactic concept recognition engine (which utilizes concept names and synonyms), as well as on a set of semantic expansion algorithms that leverage the semantic information found in ontologies. The methodology relies on ontologies to create annotations for textual data and presents them by using semantic web standards. It can also be used for named-entity extraction from food ontologies that are part of the BioPortal software services. Each version of the NCBO annotator working with a different ontology was taken to be as a different NER method (i.e. NCBO (SNOMED CT), NCBO (OF), NCBO (FoodOn)) as to provide a fair comparison. At the end, a total of four different NER methods (FoodIE, NCBO (SNOMED CT), NCBO (OF) and NCBO (FoodOn)) that can be used for food information extraction were compared.

For evaluation we selected three standard types of matches: true positives (TPs), false negatives (FNs) and false positives (FPs), as well as the aforementioned "Partial (Inconclusive)" match type. The results from counting the instances of each match type are presented in Table 4.1. It is important to note that not all ontologies provided annotations for each recipe. More specifically, out of 1,000 recipes: SNOMED CT missed six, OntoFood missed 71, and FoodON missed five.

Looking at the comparison results in Table 4.1, we can see that the number of TPs is substantially larger when using FoodIE (11,461) when compared to the three other ontologies with the NCBO annotator, i.e.: SNOMED CT (5,100), OF (2,279), and FoodON (5,725). The number of TPs should be maximized.

Moving on to the FPs, FoodIE again provides the best results of the four, while FoodON provides significantly more FPs than the other three NER methods. Respectively they provide: FoodIE (258), SNOMED CT (472), OF (378), and FoodON (1,502). The number of FPs should be minimized.

The last of the standard types of matches is FN, where FoodIE once again behaves superiorly to the other three NER methods. The numbers here are: FoodIE (684), SNOMED CT (5,327), OF (9,026), and FoodON (4,968). The number of FNs should be minimized.

Table 4.1: Results of comparing the entities extracted by FoodIE and the NCBO Annotator using SNOMED CT, OF and FoodOn.

	FoodIE	SNOMED CT	OF	FoodOn
TPs	11,461	5,100	2,279	5,725
FPs	258	472	378	1,502
FNs	684	5,327	9,026	4,968
Partials	359	2,705	1,591	2,365

The last type of match we take into account is the partial match type. It is not clear whether this type of match should be maximized or minimized in and of itself, as it heavily depends on the number of other types of matches (especially TPs and FNs). For example, ideally all the food concepts would be matched as TPs and none as FNs. However, if a TP match is not encountered for a specific food concept instance, the second-best occurrence would be to "partially" match it. The worst-case scenario is when the food concept is matched as a FN.

From analysing the results, we can conclude that FoodIE, using the Hansard corpus, provides the most promising results since it can extract a larger number of food concepts as opposed to the NCBO annotator in combination with the selected ontologies. Moreover, the results imply that the three food ontologies (i.e. SNOMED CT, FoodOn, OntoFood) do not represent the food domain exhaustively, as many food concepts are not extracted using the NCBO annotator running on these ontologies, which indicates that they do not exist as entities in the food ontologies themselves.

Food ontologies alignment. To align food concepts in different food ontologies, we have created a resource, named FoodOntoMap, that consists of food concepts extracted from recipes. For each food concept, semantic tags from four food ontologies are assigned. With this, we create a resource that provides a link between different food ontologies, which can further be reused to develop applications for understanding the relation between food systems, human health, as well as the environment.

The results from the FoodOntoMap are four different datasets and one data set mapping. Each data set consists of an artificial id for each unique food concept that is extracted by using each approach, the name of the extracted food concept, and the semantic tags assigned to it. Each data set corresponds to one of the four semantic resources: Hansard corpus, FoodOn, OntoFood, and SNOMED CT. At the end, there is one data set mapping, called FoodOntoMap, where for each concept that appears at least in two datasets, the mapping between them is provided by listing the artificial id of the concepts from each of the datasets in which it is encountered.

An example for one instance from is A000016, B000011, C000012, D000002. The provided codes are unique artificial identifiers that are assigned to the food concept using each of the aforementioned food semantic resources, respectively. If we look at the three separate datasets, we can see that A000016 corresponds to "garlic" with the semantic tag AG.01.h.02.e [Onion/leek/garlic] from the Hansard corpus, B000011 corresponds to "garlic" with the semantic tag `http://purl.obolibrary.org/obo/NCBITaxon_4682` from the FoodOn, C000012 refers to "garlic" with the semantic tag `http://purl.bioontology.org/ontology/SNOMEDCT/735030001` from the SNOMED CT, and D000002 corresponds to "garlic" with the semantic tag `http://www.owl-ontologies.com/Ontology1435740495.owl#Garlic` from the OntoFood.

The datasets consist of 13,205; 1,069; 111; and 582 unique food concepts, obtained using Hansard corpus, FoodOn, OntoFood, and SNOMED CT, respectively. The FoodOntoMap mapping consists of 1,459 food concepts that are found in at least two of the food semantic resources.

The motivation for building such a resource in the food domain comes from the existence of the UMLS, which is extensively used in the biomedical domain. For example, the MRCONSO.RRF table that is a part of the UMLS is used in a lot of semantic web applications since it can map the medical concepts to a variety of different biomedical standards and vocabularies.

4.1.3 FoodBase corpus overview

After applying FoodIE for semantic annotation of 1,000 randomly selected recipes with semantic tags from the Hansard corpus and performing post-processing of the annotated semantic tags, the initial FoodBase corpus was generated. It consists of 12,844 food entities extracted from the selected recipes for dishes from five main groups, with 2,105 unique food entities in total. Because the evaluation of extended FoodIE gave very promising and consistent results, we followed the idea of transfer learning for extracting and annotating food entities for a new, more extensive subset of 21,790 recipes from the same five groups of dishes, i.e. “Appetizers/Snacks”, “Breakfast/Lunch”, “Dessert”, “Dinner”, and “Drinks”. The outcome was the next version of the FoodBase corpus that includes much more recipes and corresponding food entities (274,053 total food entities, with 13,079 unique food entities). However, this version has not been processed manually to exclude the FPs and to include the FNs. To distinguish between the two versions of the FoodBase corpus, we call the manually post-processed version containing 1000 recipes “curated” and the one consisting of 21,790 recipes as annotated by FoodIE, “un-curated”.

The descriptive statistics (i.e. mean, median, mode, and standard deviation) for the number of words per recipe, the number of entities extracted per recipe, and the number of semantic tags assigned per food entity for both versions are presented in Table 4.2. The distribution of the number of words per recipe for the curated and un-curated version of FoodBase is presented in Figure 4.3, while the distribution of the number of food entities extracted per recipe is presented in Figure 4.4. Additionally, the distribution of the number of semantic tags assigned per food entity for both versions is presented in Figure 4.5.

Looking at the descriptive statistics provided for the number of words per recipe, it is apparent that there are no big differences between the descriptive statistics of both versions. The biggest difference appears for the mode. The mode for the curated version is 121.00, while for the un-curated version it is 91.00. If we look closer at the distribution provided for the curated version in Figure 4.3, we can see that these two values for frequency are close and that they differ no more than 15. Moreover, it follows that the distributions have a similar trend. However, the distribution of the un-curated version is much smoother because it includes more recipes. The same conclusion is also true for the descriptive statistics provided for the number of extracted food entities per recipe. The only difference that is apparent is for the modes. It is 5.00 for the curated version and 10.00 for the un-curated version. However, if we look at the distribution of the curated version (Figure 4.4), we can see that the difference between their frequency is less than 10. From Figure 4.4, it is obvious that the distributions have a similar trend, the only difference is that the distribution of the un-curated version is much smoother, which is reasonable, since it includes more recipes. In Figure 4.5, the distribution of the number of assigned semantic tags per food entity is presented for both FoodBase versions, separately. Analyzing it, it follows that both distributions have a similar trend, which is a power law distribution.

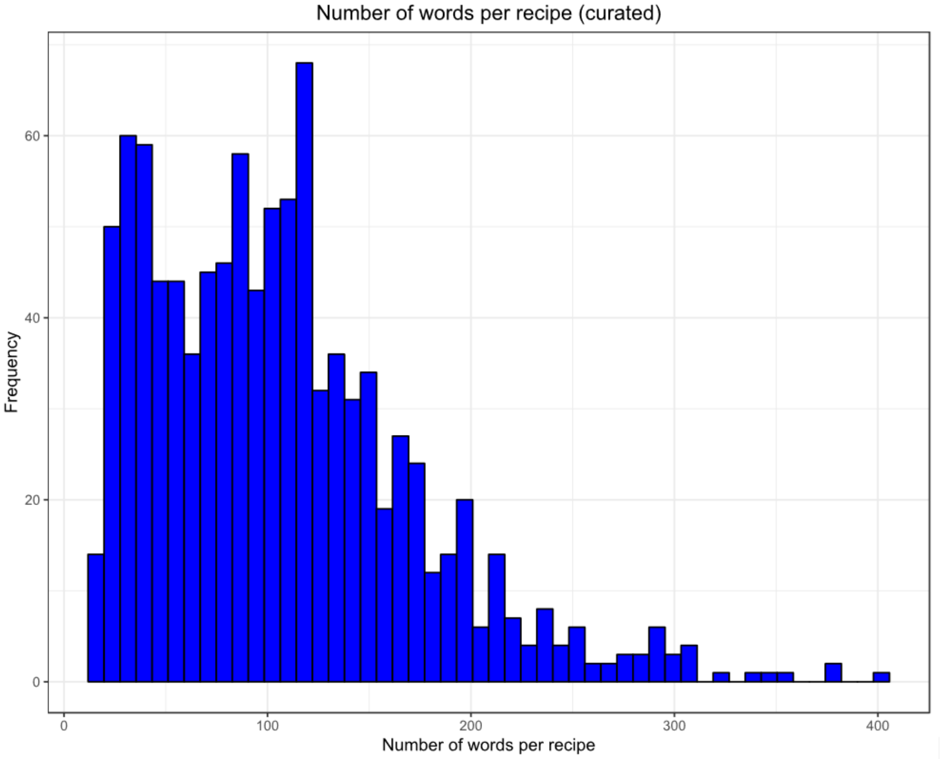
Additionally, the statistics are presented for each category “Appetizers/Snacks”, “Breakfast/Lunch”, “Dessert”, “Dinner”, and “Drinks” separately in Table 4.3. It is evident that in both versions of FoodBase the average number of extracted entities, as well as the standard deviation, are the largest in the "Dinner" category. There is no big difference between the average number of extracted entities from the "Appetizers/Snacks", "Breakfast/Lunch", and "Dessert" and additionally they have similar standard deviations. The "Drinks" category has the smallest average number and standard deviation of extracted entities. If we compare the descriptive statistics for each category between both versions, we can see that there are not big deviations between their values.

The number of food entities per the ten most frequent semantic tags for both FoodBase versions is presented in Figure 4.6. The most frequent tag for both versions is the "AG.01 [Food]", which is the top level in the food category hierarchy in the Hansard corpus. This result comes from the fact that if there is no semantic tag assigned to the entity, but it is recognized as food, then it is automatically assigned to this semantic tag. If we compare the other nine most frequent semantic tags, we can see that the semantic tags "AG.01.l.02 [Sweetener (syrup/honey/chocolate)]", "AG.01.m [Substance for food preparation]", "AG.01.n [Dishes and prepared food]", "AG.01.e [Dairy products]", "AG.01.n.11 [Bread]", and "AG.01.g [Eggs]" appear in both versions of FoodBase. The initial FoodBase based on 1,000 recipes also includes "AG.01.l.03 [Spice]", "AG.01.w [Setting table]", and "AG.01.h.02.e [Onion/leek.garlic]", while the next FoodBase version based on 21,790 recipes includes "AG.01.k [Flour]", "AG.01.j [Meal]", and "AG.01.e.01 [Butter]".

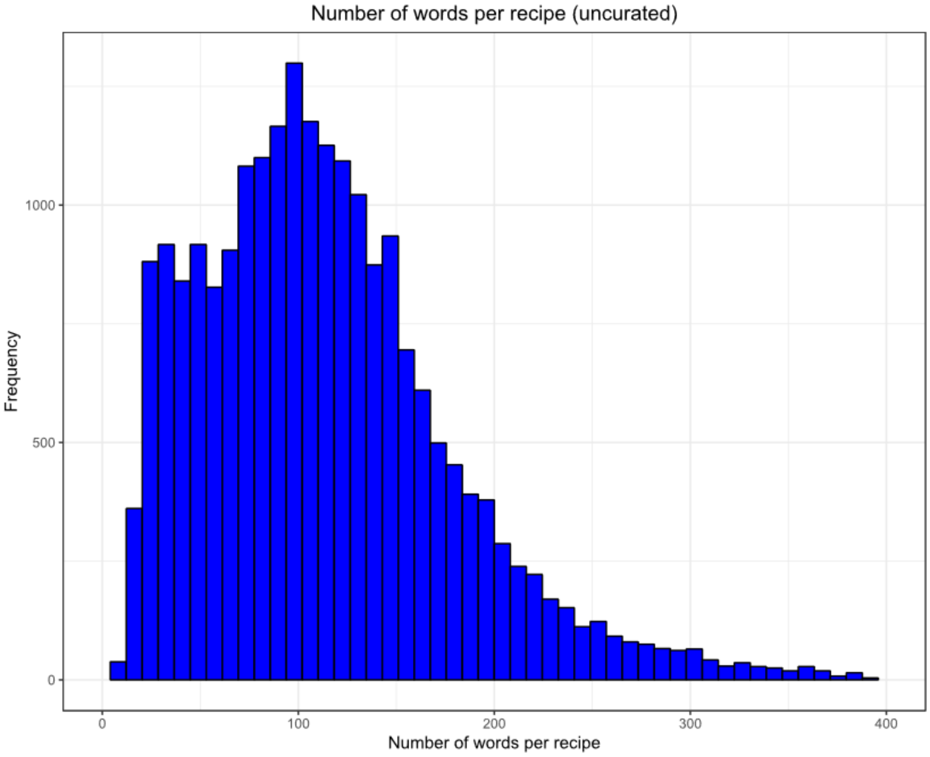
Looking at the 10 most frequent tags for both versions, we can see that seven out of 10 semantic tags are the same. There is a difference between the frequency of the semantic tags in both versions, since the un-curated version consists of more recipes. However, the first idea of building such a corpus is that it can be generally used for bi-classification problem (food vs. not food concept), where the semantic tags are not crucial. Further, using some sampling techniques, different subsets can be generated form the un-curated version with regard to the semantic tags that will be of interest (e.g. top 5 or 10 most frequent) in order to produce more representative data sets for training corpus-based NERs.

In general, the difference between the curated and un-curated version is that the un-curated version consists of FPs, which are in most cases related to objects that are not foods but are concepts closely related to food entities or an instrument for food or cooking. Also, the FNs are not included, and they are related to food entities that are branded food products, or some rare foods that are typical for some cultures.

The availability of data and tools used to create FoodBase corpus is provided in Table 4.4.

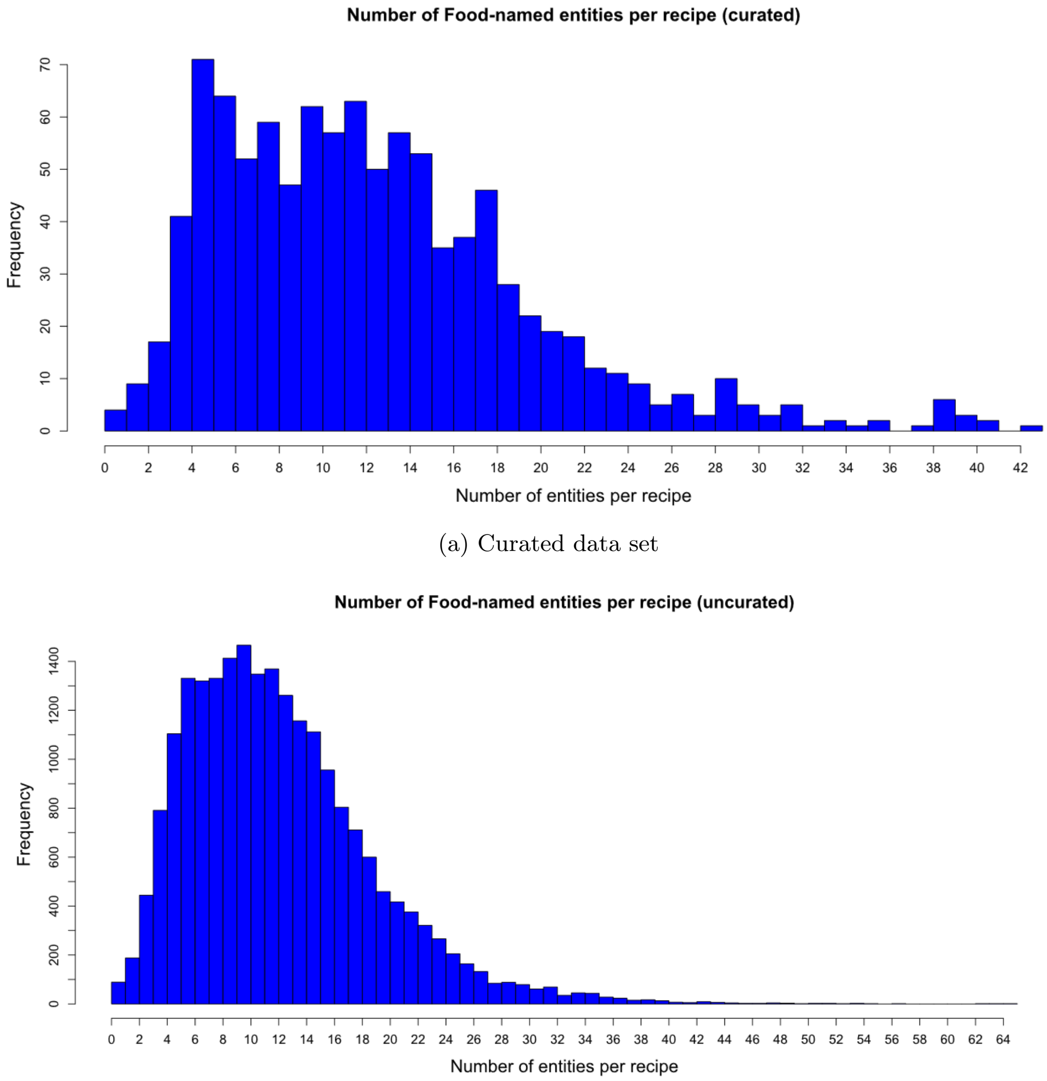


(a) Curated data set



(b) Un-curated data set

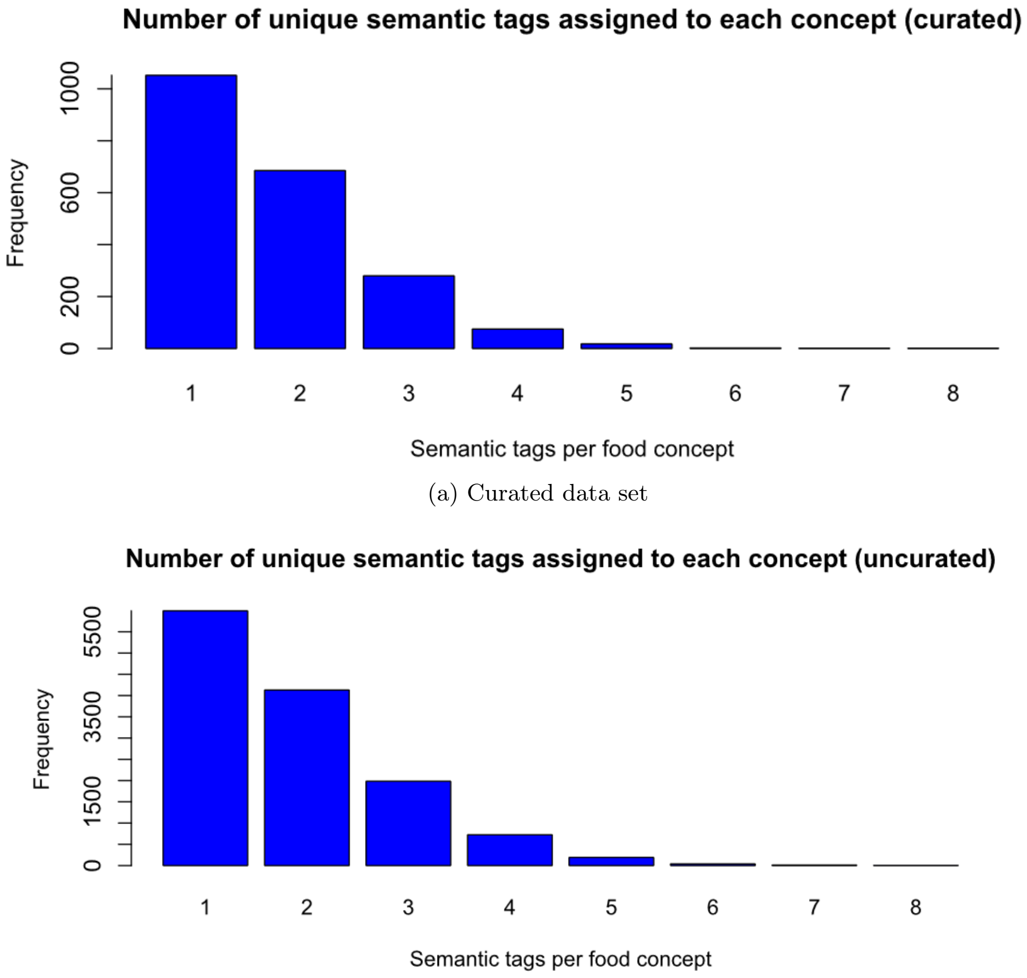
Figure 4.3: Distribution of the number of words per recipe. It is apparent that both distributions have a similar trend. However, the distribution of the un-curated version is much smoother because it includes more recipes.



(a) Curated data set

(b) Un-curated data set

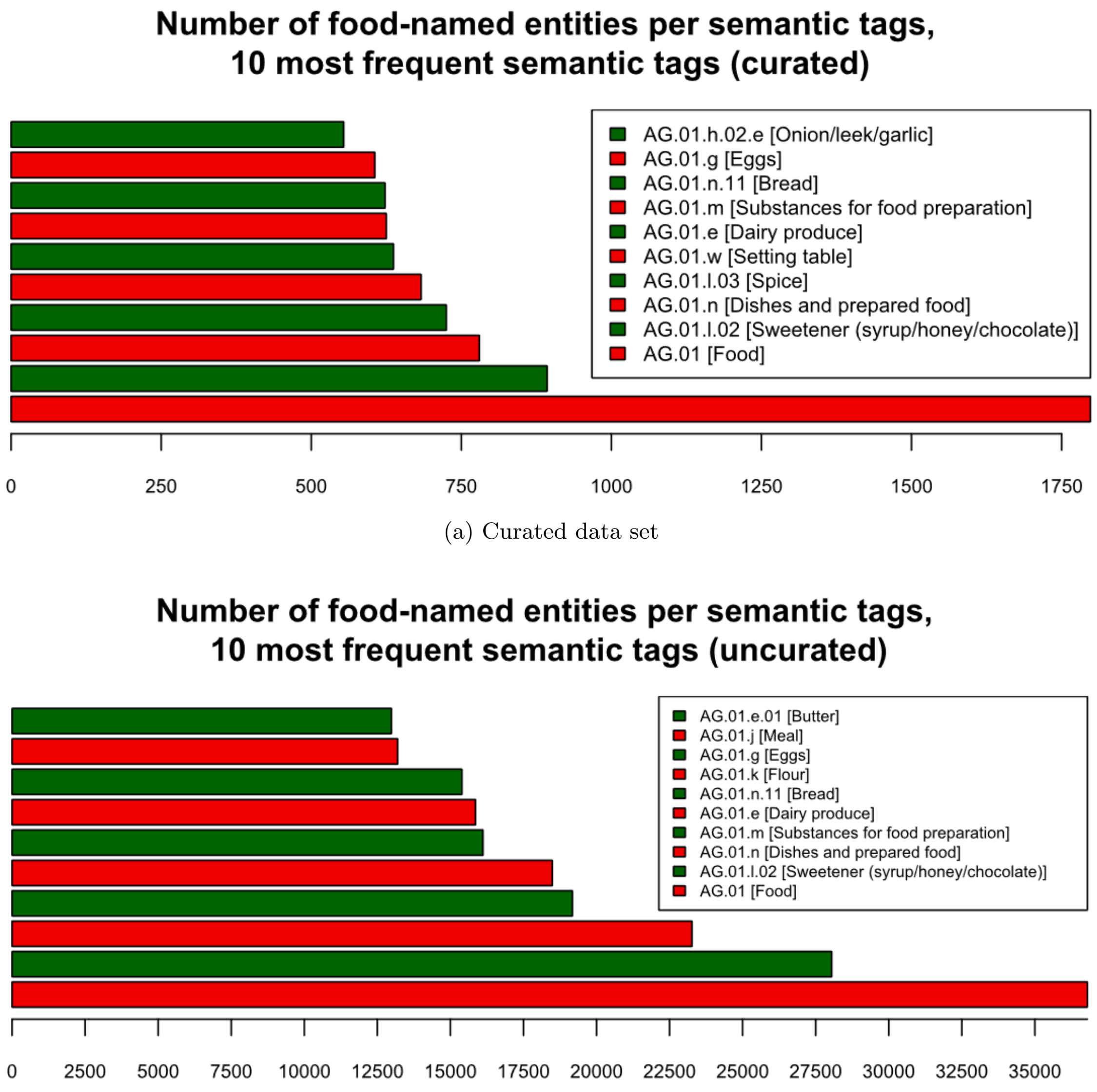
Figure 4.4: Distribution of the number of extracted food entities per recipe. It is apparent that both distributions have a similar trend. However, the distribution of the un-curated version is much smoother because it consists of more recipes.



(a) Curated data set

(b) Un-curated data set

Figure 4.5: Distribution of the number of assigned semantic tags per food entity. Analyzing it, it follows that both distributions have a similar trend, which is a power law distribution.



(a) Curated data set

(b) Un-curated data set

Figure 4.6: Number of food-named entities per ten most frequent semantic tags. From the 10 most frequent semantic tags in both the curated and un-curated version, seven are identical across both versions. The three that differ are due to the difference in the number of recipes in both versions.

Table 4.2: Descriptive statistics for the number of words per recipe, the number of food entities per recipe and the number of semantic tags per food entity.

Number of words per recipe		
	Curated	Un-curated
Mean	106.40	114.78
Median	99.00	106.00
Mode	121.00	91.00
Standard Deviation	64.44	67.61

Number of food entities per recipe		
	Curated	Un-curated
Mean	12.85	12.58
Median	12.00	12.00
Mode	5.00	10.00
Standard deviation	7.22	6.71

Number of semantic tags per food entity		
	Curated	Un-curated
Mean	1.74	1.87
Median	2.00	2.00
Mode	1.00	1.00
Standard deviation	0.91	1.01

Table 4.3: Descriptive statistics for the number of food entities per recipe and the number of semantic tags per food entity, for each category separately.

Number of food entities per recipe								
	Curated				Un-curated			
	Mean	Median	Mode	Sd	Mean	Median	Mode	Sd
Appetizers/Snacks	11.57	10.00	8.00	6.21	11.25	10.00	7.00	6.20
Breakfast/Lunch	13.46	13.00	13.00	6.32	13.04	12.00	11.00	6.33
Dessert	15.15	14.00	15.00	7.03	13.79	13.00	12.00	6.54
Dinner	17.38	16.00	11.00	7.43	17.50	16.00	15.00	8.03
Drinks	6.67	6.00	5.00	3.53	6.52	6.00	6.00	3.31
Number of semantic tags per food entity								
	Curated				Un-curated			
	Mean	Median	Mode	Sd	Mean	Median	Mode	Sd
Appetizers/Snacks	1.59	1.00	1.00	0.77	1.69	1.00	1.00	0.88
Breakfast/Lunch	1.58	1.00	1.00	0.75	1.75	2.00	1.00	0.89
Dessert	1.71	1.00	1.00	0.90	1.88	2.00	1.00	1.01
Dinner	1.65	1.00	1.00	0.84	1.72	2.00	1.00	0.86
Drinks	1.82	1.00	1.00	1.02	1.96	2.00	1.00	1.14

Table 4.4: Availability of data and tools used to create FoodBase.

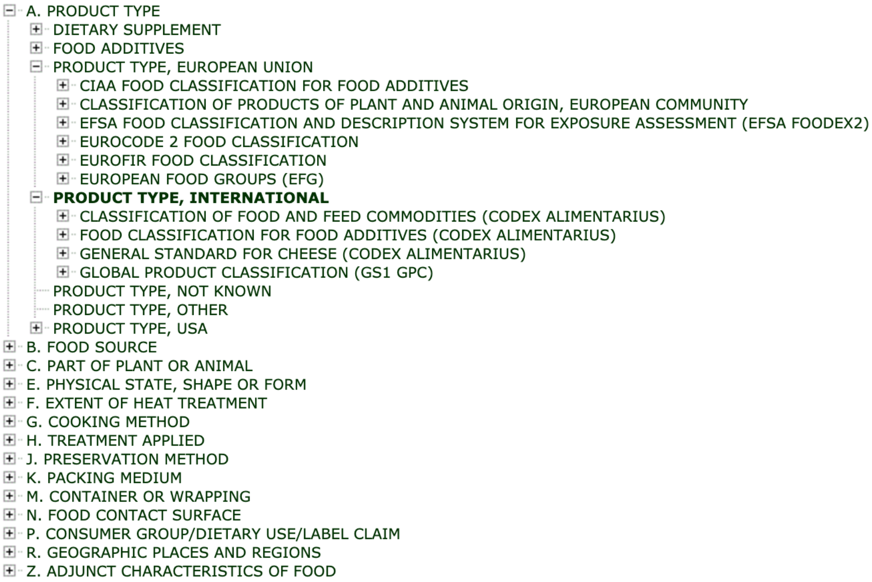
Resource/Tool name	Availability
FoodBase	http://cs.ijs.si/repository/FoodBase/foodbase.zip
FoodIE	https://github.com/GorjanP/foodie
FoodOn	https://foodontology.github.io/foodon/
OntoFood (OF)	https://bioportal.bioontology.org/ontologies/OF/?p=summary
SNOMED CT	https://confluence.ihtsdotools.org/display/DOC/Technical+Resources
Hansard corpus	https://www.hansard-corpus.org/
NCBO annotator	http://bioportal.bioontology.org/annotator
NCBO annotator REST API	http://data.bioontology.org/documentation
FoodOntoMap	https://doi.org/10.5281/zenodo.2635437
FOM mapper	https://github.com/GorjanP/FOM_mapper_client

4.1.4 Conclusions

Our motivation to start building the FoodBase corpus has been to provide the scientific community with a fundamental resource required for learning corpus-based methods that can be used for food-named entity recognition. FoodBase is presented in two versions: curated and un-curated. The curated version is manually evaluated, consisting of 1000 recipes, while the un-curated version consists of 21,790 recipes. For this reason, we can consider FoodBase as a whole as a “silver standard”. It can also be used as a benchmark data set for several ML tasks, such as multi-class classification [91], multi-label classification [92] and hierarchical multi-label classification [93]. Multi-class classification is applied when a food entity may be annotated with several semantic tags (e.g. “Food” (AG:01); “Production of food, farming” (AG:02); and “Acquisition of animals for food, hunting” (AG:03)). Multi-label classification is performed when an output is a more complex structure such as a vector of tags with some dependencies among them (i.e. the food entity can belong to multiple semantic tags simultaneously). Hierarchical multi-label classification is needed when the classes are hierarchically structured and food entities can be assigned to multiple paths of the class hierarchy at the same time (e.g. the food hierarchy from the Hansard corpus). As part of future work, we are working on presenting benchmarking results obtained from corpus-based food-named entity recognition using three datasets of a different scale and quality (e.g. 200 recipes manually annotated, 1000 recipes – manually annotated, and 21,790 recipes – automatically annotated), in order to explore the utility of having such a corpus by applying sensitivity analysis. The FoodBase corpus will enable a further development of more accurate food NERs to be used for the extraction of food entities not only from recipes as presented in this chapter but also from scientific literature. Consequently, the exploration and the extraction of relations between food entities and other biomedical entities such as drug, disease, gene entities, etc., will be supported. Moreover, the FoodBase corpus is a step towards food normalization where semantic, instead of lexical, similarity can also be included. Furthermore, the semantic tags will be able to be used for building food embedding space needed for predictive studies.

4.2 Exploring a Standardized Language for Describing Foods Using Embedding Techniques

LanguaL, also called the “language of food”, is an automated method for describing, capturing and retrieving data about food [94]. It is developed by the US National Cancer Institute (NCI), and its European partners from France, Denmark, Switzerland and Hungary. The thesaurus provides a standardised language for describing and classifying food products. The LanguaL hierarchy of terms is presented in Figure 4.7. From it, we can see that LanguaL supports different terms to describe a food, such as product type, food source, cooking method, etc. Moreover, if we focus on the product type, it supports different food standards, such as EFSA Food Classification and Description System for Exposure Assessment (EFSA FoodEx2) [71], EuroFir Food Calcification [95], European Food Groups (EFG) [96], Global Product Classification (GS1 GPC) [97], which are mapped to the LanguaL. The mapping between these different standards is really important in cases where we want to link different food data sets that are described using different food standards. For example, the LanguaL code 14730 corresponds to “APPLES” from the EFSA FoodEx2 standard with code A01DJ. Additionally, the LanguaL code 10005900 corresponds to “APPLES” from the GS1 GPC food standard with code A1763.



- ⊟ A. PRODUCT TYPE
  - ⊞ DIETARY SUPPLEMENT
  - ⊞ FOOD ADDITIVES
  - ⊟ PRODUCT TYPE, EUROPEAN UNION
    - ⊞ CIAA FOOD CLASSIFICATION FOR FOOD ADDITIVES
    - ⊞ CLASSIFICATION OF PRODUCTS OF PLANT AND ANIMAL ORIGIN, EUROPEAN COMMUNITY
    - ⊞ EFSA FOOD CLASSIFICATION AND DESCRIPTION SYSTEM FOR EXPOSURE ASSESSMENT (EFSA FOODEX2)
    - ⊞ EUROCODE 2 FOOD CLASSIFICATION
    - ⊞ EUROFIR FOOD CLASSIFICATION
    - ⊞ EUROPEAN FOOD GROUPS (EFG)
  - ⊟ **PRODUCT TYPE, INTERNATIONAL**
    - ⊞ CLASSIFICATION OF FOOD AND FEED COMMODITIES (CODEX ALIMENTARIUS)
    - ⊞ FOOD CLASSIFICATION FOR FOOD ADDITIVES (CODEX ALIMENTARIUS)
    - ⊞ GENERAL STANDARD FOR CHEESE (CODEX ALIMENTARIUS)
    - ⊞ GLOBAL PRODUCT CLASSIFICATION (GS1 GPC)
  - PRODUCT TYPE, NOT KNOWN
  - PRODUCT TYPE, OTHER
  - ⊞ PRODUCT TYPE, USA
- ⊞ B. FOOD SOURCE
- ⊞ C. PART OF PLANT OR ANIMAL
- ⊞ E. PHYSICAL STATE, SHAPE OR FORM
- ⊞ F. EXTENT OF HEAT TREATMENT
- ⊞ G. COOKING METHOD
- ⊞ H. TREATMENT APPLIED
- ⊞ J. PRESERVATION METHOD
- ⊞ K. PACKING MEDIUM
- ⊞ M. CONTAINER OR WRAPPING
- ⊞ N. FOOD CONTACT SURFACE
- ⊞ P. CONSUMER GROUP/DIETARY USE/LABEL CLAIM
- ⊞ R. GEOGRAPHIC PLACES AND REGIONS
- ⊞ Z. ADJUNCT CHARACTERISTICS OF FOOD

Figure 4.7: The LanguaL hierarchy.

Taking into account that various food standards have been mapped to the LanguaL hierarchy, in this chapter we focus on providing additional knowledge representation forms of the LanguaL language with vector embeddings. This means that we are trying to learn a vector representation for every food product that is present in the LanguaL hierarchy. The main goal of this study is to see if the same food items represented by different food standards are semantically linked together.

In this chapter, the methodology for exploring the LanguaL hierarchy is explained, which is based on Representation Learning (RL) (explained in Chapter 2), followed by results and discussion. Finally, the conclusions of the chapter are presented.

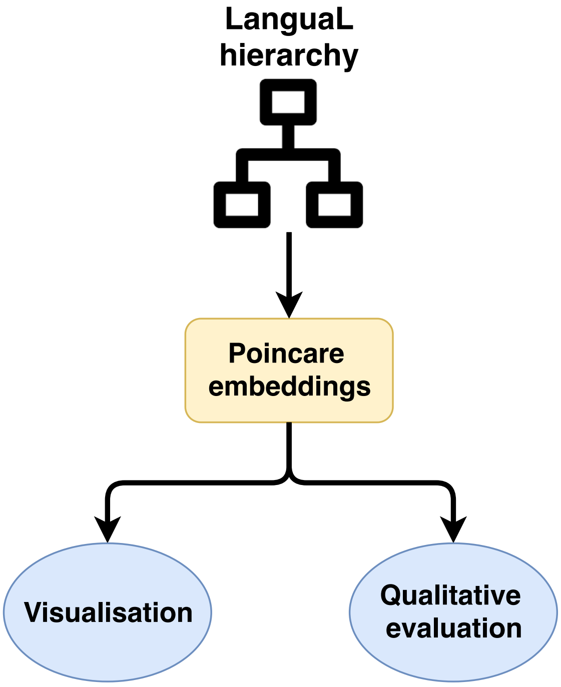


Figure 4.8: Evaluation methodology for exploring LanguaL.

4.2.1 Methodology

To explore the LanguaL hierarchy and to evaluate if the various food standards it is comprised of are linked together, we propose a methodology that is presented in Figure 4.8. It consists of three steps:

- Learning a vector representation (i.e. embedding) for every concept from the LanguaL hierarchy;
- Visualization of the learned representations;
- Qualitative evaluation of the learned representations.

4.2.2 Poincaré embedding

The concepts in LanguaL follow a hierarchical structure. For this reason, the most appropriate embeddings to use are the aforementioned Poincaré embeddings [77], which are described in Chapter 2.

If two concepts $\mathbf{x}$ and $\mathbf{y}$ are represented in Poincaré space with dimensionality D, the distance between these two concept vector representations is calculated with the following function:

$$d(\mathbf{x},\mathbf{y}) = arcosh\left(1 + 2\frac{||\mathbf{x}-\mathbf{y}||^2}{(1-||\mathbf{x}||^2)(1-||\mathbf{y}||^2)}\right) \tag{4.1}$$

In our case, with the food concepts available from LanguaL, the Poincaré embeddings are computed such that semantically similar food concepts are close to one another in the embedding space with regard to their distance in the Poincaré ball. We have decided to call the learned embeddings *LanguaL2vec* because they represent vector representations for the food concepts that are available in LanguaL. More details about the learning process of the Poincaré vector representations can be found in [77].

4.2.3 Visualization methodology

After learning the embeddings for every concept that is present in the LanguaL hierarchy, we used the t-distributed stochastic neighbour embedding (t-SNE) method [98] to visualize them in order to see how they are distributed in the embedded space. t-SNE is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization into a low-dimensional space of two or three dimensions.

4.2.4 Qualitative evaluation

To evaluate if different standards (concepts) which are part of the LanguaL hierarchy are linked together, we performed two experiments.

In the first experiment, we randomly selected equivalent (as defined by their title) food product concepts (e.g. 'Apple') from different standards (e.g. GS1 and FoodEx2), and then calculated the similarity between their vector space embeddings. In our case, each food product is represented as D dimensional vector in the Poincaré ball. To find similarity between two concepts, we calculated the similarity between their vectors. This is done by finding the angle between the vectors. Even though we previously described that the distance between two concepts in the Poincaré ball can be calculated using the $arcosh(\cdot)$ function, in this experiment we used the cosine distance instead. This was motivated by the fact that the Poincaré ball model is conformal, meaning that the angles between vectors are identical to their Euclidean counterparts [99]. Additionally, this formula is simpler and faster to compute. The cosine distance between two concepts represented by their vectors $\mathbf{x}$ and $\mathbf{y}$ can be calculated using the following equation:

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{xy}}{||\mathbf{x}||^2||\mathbf{y}||^2}. \tag{4.2}$$

If the concepts are semantically related, the cosine similarity should be significantly greater than zero, i.e. closer to one than to zero. We computed the cosine similarity for ten randomly selected pairs of food products that represent the same concept from the LanguaL hierarchy.

In the second experiment, we randomly selected 49 different food products from the LanguaL hierarchy and for each we returned the top three most similar concepts from the embedded space based on the cosine similarity. If the resulting concepts for one food product are from the same standard, this means that the standards in the LanguaL hierarchy are not linked together. This occurs since the main contributing factor to the similarity in such cases is their graph structure, rather than what is considered to be more important - the semantic information the food concepts convey.

4.2.5 Results and Discussion

Visualization. Using the aforementioned t-SNE visualisation methodology on the 100 dimensional Poincaré embeddings, presented in Figure 4.9, it is clear that the hierarchical information from LanguaL is captured in the form of a ball in the embedding space.

Qualitative evaluation. In the first experiment, we randomly selected ten different food products and for each of them we calculated the cosine similarity between their embeddings that are learned for different standards. For example, for the food product "APPLES", we used the embeddings learned for the concept found in EFSA FOODEX2 and for the one found in GS1 GPCC. The results for ten different food products are presented in Table 4.5. Using it, we can conclude that the same concepts represented by different sources are not linked together well, i.e. not close in the Poincaré ball.

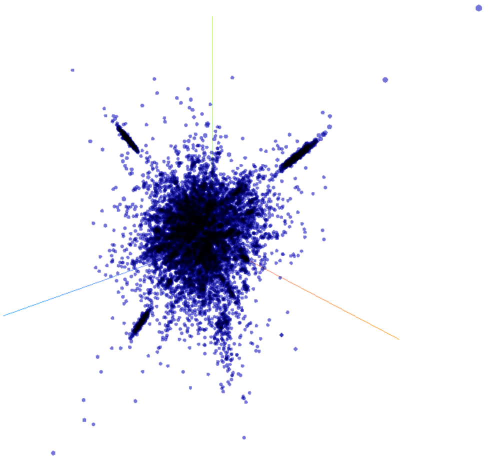


Figure 4.9: Visualization of the LanguaL embeddings (d=100) learned by the Poincaré method.

Table 4.5: Cosine similarity between the same concept from different sources.

	Concept	Database 1	Database 2	Cosine Similarity
1	APPLES	EFSA FOODEX2	GS1 GPCC	0.18924
2	PLUMS	EFSA FOODEX2	GS1 GPCC	-0.06574
3	SALT/SALT SUBSTITUTE	EFSA FOODEX2	US CFR	-0.16501
4	GLUTEN	EFSA FOODEX2	Category C	0.05803
5	MILK	EFSA FOODEX2	EFG	0.01479
6	STEVIA	EFSA FOODEX2	EUROFIR	-0.08529
7	PEANUTS	EFSA FOODEX2	EFSA FOODEX2	0.06698
8	TREE NUTS	CCPR	EC	-0.05245
9	EGGS/EGG PRODUCTS	EFSA FOODEX2	EFG	0.09800
10	CHICKEN FRESH/UNPROCESSED	EFSA FOODEX2	GS1 GPC	0.02552

In the second experiment, we found the top three most similar concepts in the embedded space for 49 randomly selected food products from the LanguaL hierarchy. The results are presented in Table 4.6. Looking at the table, we can conclude that the three most similar concepts for every food product are only concepts that are in the same sub-hierarchy (e.g. FoodEx2, GS1), which shows that the different standards are not semantically linked together.

By performing this kind of analysis, an open gap is revealed that needs to be addressed in order to link different food standards between each other.

4.2.6 Conclusion

In this chapter, we explored the LanguaL hierarchy, which is a resource used to describe foods. It consists of different standards where the same food product can be presented by different codes. To see if the standards are linked together, we learned hierarchical Poincaré embeddings for each food product and calculated the cosine similarity to see if the same concepts are close to each other in the embedded space. Our analyses showed that the LanguaL hierarchy consists of different food standards that are connected in one hierarchy. However, the standards are not linked together, which means that the vector representations of a food product, taken from different standards, are not close together in the embedded space. This suggests that they have a low level of connection, even though their semantic information is identical. Additionally, the most similar concepts for any concept are other concepts from the same standard. This shows us that further effort should be done to link all these standards together adequately in order to provide a unified system for describing and standardizing food products and their corresponding semantic knowledge.

Table 4.6: Top three most similar concepts for 49 randomly selected food products from LanguaL hierarchy.

	Concept	First most similar
1	ACID (CODEX) [A0362]	"COLOUR RETENTION AGENT (CODEX) [A0392]"
2	ALKALI (CODEX) [A0364]	"CODEX ALIMENTARIUS, FUNCTIONAL CLASSES [A0351]"
3	ANTICAKING AGENT (CODEX) [A0367]	"ANTIOXIDANT (CODEX) [A0371]"
4	YEAST FOOD (CODEX) [A0451]	"CODEX ALIMENTARIUS, FUNCTIONAL CLASSES [A0351]"
5	FLAVOURING AGENT (CODEX) [A0413]	"CODEX ALIMENTARIUS, FUNCTIONAL CLASSES [A0351]"
6	ACID (EC) [A0325]	"EMULSIFYING SALT (EC) [A0334]"
7	EMULSIFIER (EC) [A0333]	"FOOD ADDITIVE CLASSIFICATION, EUROPEAN COMMUNITY [A0324]"
8	SWEETENER (EC) [A0349]	"FOOD ADDITIVE CLASSIFICATION, EUROPEAN COMMUNITY [A0324]"
9	0640000 - COCOA (EC) [A1245]	"0620000 - COFFEE BEANS (EC) [A1243]"
10	COLOUR (EC) [A0332]	"GELLING AGENT (EC) [A0339]"
11	FATS AND OILS (CIAA) [A0453]	"OTHER FOODS (CIAA) [A0467]"
12	EGG AND EGG PRODUCTS (CIAA) [A0461]	"OTHER FOODS (CIAA) [A0467]"
13	BEVERAGES (CIAA) [A0465]	"CEREALS AND CEREAL PRODUCTS (CIAA) [A0457]"
14	MEAT AND MEAT PRODUCTS (CIAA) [A0459]	"CEREALS AND CEREAL PRODUCTS (CIAA) [A0457]"
15	SUGAR AND HONEY (CIAA) [A0462]	"EDIBLE ICES (CIAA) [A0454]"
16	01740 - PASTA, DOUGHS AND SIMILAR PRODUCTS (EFSA FOODEX2) [A04QT]	"01980 - RAW DOUGHS AND PRE-MIXES (EFSA FOODEX2) [A008G]"
17	02990 - PANCAKES (EFSA FOODEX2) [A00CL]	"03000 - SCONES AND SIMILAR (EFSA FOODEX2) [A00CM]"
18	03580 - VEGETABLES AND VEGETABLE PRODUCTS (EFSA FOODEX2) [A00FJ]	"05340 - FLOWERS USED AS VEGETABLES (EFSA FOODEX2) [A0ESZ]"
19	37220 - COCOA BEVERAGES (EFSA FOODEX2) [A03KY]	"37220 - COCOA BEVERAGES (EFSA FOODEX2) [A03KY]"
20	11810 - PEANUTS (EFSA FOODEX2) [A015H]	"11800 - PEANUTS AND SIMILAR- (EFSA FOODEX2) [A0DBS]"
21	BREAKFAST CEREAL (EUROFIR) [A0816]	"CEREAL BAR (EUROFIR) [A1330]"
22	RED MEAT (EUROFIR) [A0794]	"SAUSAGE OR SIMILAR MEAT PRODUCT (EUROFIR) [A0798]"
23	NUT OR SEED PRODUCT (EUROFIR) [A0824]	"NUT, SEED OR KERNEL (EUROFIR) [A0823]"
24	FOOD FOR INFANTS (EUROFIR) [A0873]	"FOOD FOR WEIGHT REDUCTION (EUROFIR) [A1205]"
25	04 PASTA (EFG) [A0694]	"EUROPEAN FOOD GROUPS (EFG) [A0690]"
26	13 NUTS (EFG) [A0703]	"EUROPEAN FOOD GROUPS (EFG) [A0690]"
27	20 COFFEE, TEA, COCOA POWDER (EFG) [A0710]	"EUROPEAN FOOD GROUPS (EFG) [A0690]"
28	22 WINE (EFG) [A0712]	"19 NON-ALCOHOLIC BEVERAGES (EFG) [A0709]"
29	28 EGGS (EFG) [A0718]	"EUROPEAN FOOD GROUPS (EFG) [A0690]"
30	01 FRUITS (CCPR) [A0649]	"004 BERRIES AND OTHER SMALL FRUITS (FB) (CCPR) [A0671]"
31	022 TREE NUTS (TN) (CCPR) [A0685]	"024 SEED FOR BEVERAGES AND SWEETS (SB) (CCPR) [A0687]"
32	030 MEAT (FROM MAMMALS OTHER THAN MARINE MAMMALS) (MM) (CCPR) [A0737]	"033 MILKS (ML) (CCPR) [A0740]"
33	05 HERBS AND SPICES (CCPR) [A0653]	"028 SPICES (HS) (CODEX) [A0689]"
34	086 MILK FATS (FM) (CCPR) [A0773]	"084 CRUSTACEANS, PROCESSED (SC) (CCPR) [A0771]"
35	50101600 - FRUIT - UNPREPARED/UNPROCESSED (GS1 GPC) [A0992]	"10000002 - FRUIT - UNPREPARED/UNPROCESSED (FROZEN) (GS1 GPC) [A0993]"
36	50101700 - NUTS/SEEDS - UNPREPARED/UNPROCESSED (GS1 GPC) [A1003]	"10000008 - NUTS/SEEDS - UNPREPARED/UNPROCESSED (SHELF STABLE) (GS1 GPC) [A1005]"
37	10005767 - BEEF - PREPARED/PROCESSED (GS1 GPC) [A1619]	"50240100 - MEAT/POULTRY/OTHER ANIMALS - PREPARED/PROCESSED (GS1 GPC) [A1016]"
38	50131700 - MILK/MILK SUBSTITUTES (GS1 GPC) [A1042]	"10000027 - MILK/MILK SUBSTITUTES (FROZEN) (GS1 GPC) [A1043]"
39	50181900 - BREAD (GS1 GPC) [A0943]	"10000163 - BREAD (FROZEN) (GS1 GPC) [A0944]"
40	COLOR ADDITIVE (US CFR) [A0166]	"PRESERVATIVE [A0317]"
41	EGG OR EGG PRODUCT (US CFR) [A0261]	"EGG PRODUCT ANALOG (US CFR) [A0254]"
42	PRESERVATIVE [A0317]	"ACIDIFIER [A0322]"
43	BREAD (US CFR) [A0178]	"PIZZA CRUST (US CFR) [A0167]"
44	CURED MEAT (US CFR) [A0279]	"MEAT PRODUCT ANALOG (US CFR) [A0132]"
45	0300 BABY FOODS (USDA SR) [A1273]	"PRODUCT TYPE, USDA STANDARD REFERENCE [A1269]"
46	0400 FATS AND OILS (USDA SR) [A1274]	"0900 FRUITS AND FRUIT JUICES (USDA SR) [A1279]"
47	2000 CEREAL GRAINS AND PASTA (USDA SR) [A1290]	"0700 SAUSAGES AND LUNCHEON MEATS (USDA SR) [A1277]"
48	2100 FAST FOODS (USDA SR) [A1291]	"PRODUCT TYPE, USDA STANDARD REFERENCE [A1269]"
49	1200 NUT AND SEED PRODUCTS (USDA SR) [A1282]	"2200 MEALS, ENTREES, AND SIDEDISHES (USDA SR) [A1292]"
	Second most similar	Third most similar
1	"CODEX ALIMENTARIUS, FUNCTIONAL CLASSES [A0351]"	"LIQUID FREEZANT (CODEX) [A0429]"
2	"CLOUD PRODUCING AGENT IN NON-ALCOHOLIC BEVERAGES (CODEX) [A0388]"	"ANTIOXIDANT (CODEX) [A0371]"
3	"SEALING AGENT (CODEX) [A0440]"	"FLAVOURS SOLUBILIZER (CODEX) [A0414]"
4	"PACKING GAS (CODEX) [A0435]"	"FOAMING AGENT (CODEX) [A0416]"
5	"PACKING GAS (CODEX) [A0435]"	"ANTIOXIDANTS SOLUBILIZER (CODEX) [A0373]"
6	"FOOD ADDITIVE CLASSIFICATION, EUROPEAN COMMUNITY [A0324]"	"SEQUESTRANT (EC) [A0347]"
7	"FLOUR TREATMENT AGENT (EC) [A0337]"	"RAISING AGENT (EC) [A0346]"
8	"RAISING AGENT (EC) [A0346]"	"MODIFIED STARCH (EC) [A0342]"
9	"0650000 - CAROB (EC) [A1246]"	"0600000 - 6. TEA, COFFEE, HERBAL INFUSIONS AND COCOA (EC) [A1241]"
10	"PACKAGING GAS (EC) [A0343]"	"FOOD ADDITIVE CLASSIFICATION, EUROPEAN COMMUNITY [A0324]"
12	"DAIRY PRODUCTS (CIAA) [A0452]"	"FOODSTUFFS INTENDED FOR PARTICULAR NUTRITIONAL USES (CIAA) [A0464]"
11	"FOODSTUFFS INTENDED FOR PARTICULAR NUTRITIONAL USES (CIAA) [A0464]"	"FISH AND FISH PRODUCTS (CIAA) [A0460]"
13	"DAIRY PRODUCTS (CIAA) [A0452]"	"MEAT AND MEAT PRODUCTS (CIAA) [A0459]"
14	"BEVERAGES (CIAA) [A0465]"	"FOODSTUFFS INTENDED FOR PARTICULAR NUTRITIONAL USES (CIAA) [A0464]"
15	"CEREALS AND CEREAL PRODUCTS (CIAA) [A0457]"	"FOODSTUFFS INTENDED FOR PARTICULAR NUTRITIONAL USES (CIAA) [A0464]"
16	"02190 - CHOUX PASTRY DOUGH AND SIMILAR (EFSA FOODEX2) [A009G]"	"02020 - YEAST LEAVENED SWEET DOUGHS (EFSA FOODEX2) [A008L]"
17	"03010 - MACAROONS (EFSA FOODEX2) [A00CN]"	"02980 - DUMPLING, SWEET (EFSA FOODEX2) [A00CK]"
18	"05360 - OTHER EDIBLE FLOWERS (EFSA FOODEX2) [A0DGF]"	"05350 - COURGETTE (EDIBLE FLOWERS) (EFSA FOODEX2) [A0DGR]"
19	"37240 - HOT CHOCOLATE WITH ADDED INGREDIENTS (EFSA FOODEX2) [A03LA]"	"37240 - HOT CHOCOLATE WITH ADDED INGREDIENTS (EFSA FOODEX2) [A03LA]"
20	"11860 - SUNFLOWER SEEDS AND SIMILAR- (EFSA FOODEX2) [A0DBP]"	"11770 - OILSEEDS (EFSA FOODEX2) [A015F]"
21	"PASTA AND SIMILAR PRODUCTS (EUROFIR) [A0815]"	"CEREAL OR CEREAL-LIKE MILLING PRODUCTS AND DERIVATIVES (EUROFIR) [A0813]"
22	"PRESERVED MEAT (EUROFIR) [A0797]"	"POULTRY MEAT (EUROFIR) [A0795]"
23	"EUROFIR FOOD CLASSIFICATION [A0777]"	"FRUIT OR FRUIT PRODUCT (EUROFIR) [A0833]"
24	"MEDICAL FOOD (EUROFIR) [A0872]"	"FOOD FOR SPECIAL NUTRITIONAL USE (EUROFIR) [A0871]"
25	"21 BEER (EFG) [A0711]"	"15 VEGETABLES, EXCLUDING POTATOES (EFG) [A0705]"
26	"28 EGGS (EFG) [A0718]"	"08 SUGAR PRODUCTS, EXCLUDING CHOCOLATE (EFG) [A0698]"
27	"21 BEER (EFG) [A0711]"	"15 VEGETABLES, EXCLUDING POTATOES (EFG) [A0705]"
28	"EUROPEAN FOOD GROUPS (EFG) [A0690]"	"15 VEGETABLES, EXCLUDING POTATOES (EFG) [A0705]"
29	"33 PRODUCTS FOR SPECIAL NUTRITIONAL USE (EFG) [A0723]"	"15 VEGETABLES, EXCLUDING POTATOES (EFG) [A0705]"
30	"005 ASSORTED TROPICAL AND SUB-TROPICAL FRUIT - EDIBLE PEEL (FT) (CCPR) [A0672]"	"003 STONE FRUITS (FS) (CCPR) [A0670]"
31	"04 NUTS AND SEEDS (CCPR) [A0652]"	"023 OILSEED (SO) (CCPR) [A0686]"
32	"032 EDIBLE OFFAL (MAMMALIAN) (MO) (CCPR) [A0739]"	"031 MAMMALIAN FATS (MF) (CCPR) [A0738]"
33	"027 HERBS (HH) (CCPR) [A0688]"	"03 GRASSES (CCPR) [A0651]"
34	"087 DERIVED MILK PRODUCTS (LD) (CCPR) [A0774]"	"17 DERIVED EDIBLE PRODUCTS OF ANIMAL ORIGIN (CCPR) [A0665]"
35	"10000003 - FRUIT - UNPREPARED/UNPROCESSED (SHELF STABLE) (GS1 GPC) [A0995]"	"50250000 - FRUITS - UNPREPARED/UNPROCESSED (FRESH) (GS1 GPC) [A0994]"
36	"10000007 - NUTS/SEEDS - UNPREPARED/UNPROCESSED (PERISHABLE) (GS1 GPC) [A1004]"	"10000605 - FRUITS/VEGETABLES/NUTS/SEEDS VARIETY PACKS (GS1 GPC) [A1547]"
37	"10005771 - DUCK - PREPARED/PROCESSED (GS1 GPC) [A1624]"	"10000013 - MEAT/POULTRY/GAME/BATRACHIAN - PREPARED/PROCESSED (FROZEN) (GS1 GPC)"
38	"10000025 - MILK/MILK SUBSTITUTES (PERISHABLE) (GS1 GPC) [A1044]"	"10000026 - MILK/MILK SUBSTITUTES (SHELF STABLE) (GS1 GPC) [A1045]"
39	"10000165 - BREAD (SHELF STABLE) (GS1 GPC) [A0946]"	"10000164 - BREAD (PERISHABLE) (GS1 GPC) [A0945]"
40	"ANTIOXIDANT (US CFR) [A0318]"	"ACIDIFIER [A0322]"
41	"PRODUCT TYPE, U.S. CODE OF FEDERAL REGULATIONS, TITLE 21 [A1270]"	"GLAZE (US CFR) [A0214]"
42	"COLOR ADDITIVE (US CFR) [A0166]"	"ANTIOXIDANT (US CFR) [A0318]"
43	"PASTRY SHELL, UNSWEETENED (US CFR) [A0108]"	"QUICK BREAD, UNSWEETENED (US CFR) [A0283]"
44	"MEAT OR MEAT PRODUCT (FROM MAMMAL) (US CFR) [A0150]"	"SAUSAGE OR LUNCHEON MEAT (US CFR) [A0221]"
45	"0900 FRUITS AND FRUIT JUICES (USDA SR) [A1279]"	"1700 LAMB, VEAL, AND GAME PRODUCTS (USDA SR) [A1287]"
46	"PRODUCT TYPE, USDA STANDARD REFERENCE [A1269]"	"1700 LAMB, VEAL, AND GAME PRODUCTS (USDA SR) [A1287]"
47	"0900 FRUITS AND FRUIT JUICES (USDA SR) [A1279]"	"PRODUCT TYPE, USDA STANDARD REFERENCE [A1269]"
48	"1700 LAMB, VEAL, AND GAME PRODUCTS (USDA SR) [A1287]"	"0200 SPICES AND HERBS (USDA SR) [A1272]"
49	"0400 FATS AND OILS (USDA SR) [A1274]"	"0900 FRUITS AND FRUIT JUICES (USDA SR) [A1279]"

4.3 FoodOntoMap: Linking Food Concepts Across Different Food Ontologies

4.3.1 FoodOntoMap Design

To provide a data set in which food concepts are normalized by semantic tags from different food ontologies, we used the same recipe selection method described in Chapter 3. The data set consists of 22,000 recipes.

To extract and annotate food concepts using different food ontologies we used two approaches. First, we used FoodIE and its extension, which is described in Chapter 3 and in previous paragraphs of this Chapter. With this, each food concept is assigned semantic tags from the Hansard corpus. Then, we also used the NCBO annotator together with the food ontologies that are available in the BioPortal (i.e. FoodOn, OntoFood, and SNOMED CT), as is described in Chapter 3. The semantic tags that are assigned to each food concept are the semantic tags that belong to the tokens from which the concept is constructed, so it is a multi-label annotation process. Further in the chapter, we are going to explain each step in more detail.

After assigning the semantic tags to each food concept, the results were organized in the BioC format which is described in more detail in Chapter 3. The BioC format for one recipe and its annotations are presented in Figure 4.10.

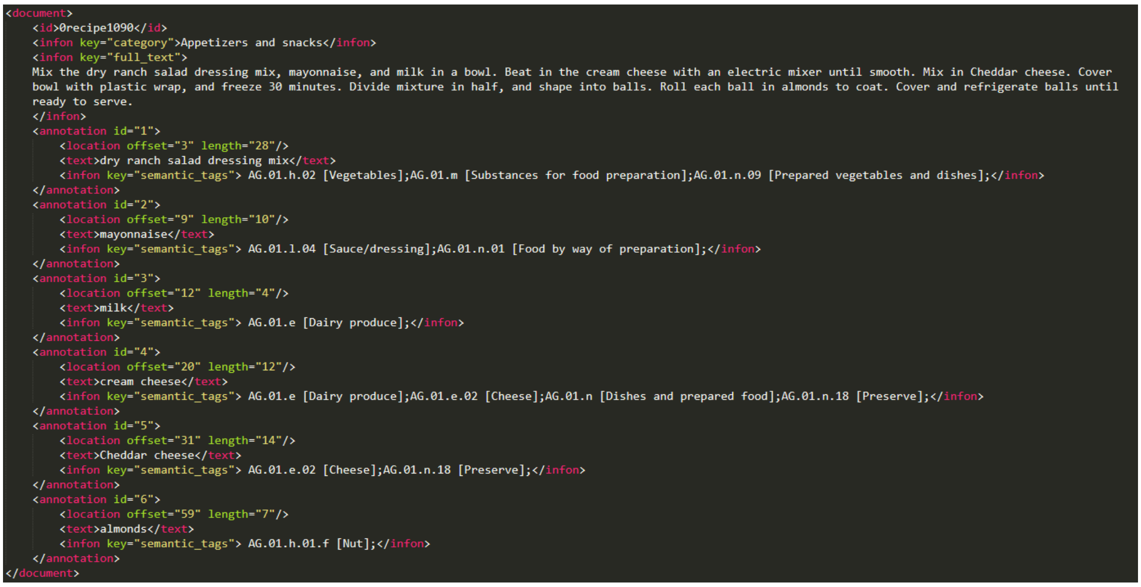
Furthermore, the same recipe data set was processed using the NCBO annotator with the FoodOn, OntoFood, and SNOMED CT ontology. To do this, we used an R client that uses the Annotator API.[1] The annotator API was used with the recipe text as input and filters provided by the ontology id (i.e *FoodOn, OntoFood and SNOMED CT*). When SNOMED CT was used, additionally another filter was applied based on a semantic type for which food was specified (i.e. Food (T168)). An example of an annotation is presented in Figure 4.11. The format and meaning of an NCBO annotation is described in more detail in Chapter 3.

Using these annotations it is determined which food concept mentions extracted by the NCBO annotator refer to the same food concept. More specifically, if some food concept mention is a subset of another mention, their information is aggregated and represented as the superset food concept mention. With these methods, duplicates, synonyms and multiple labels are resolved into a more complete food entity.

The FoodOntoMap data set consists of food concepts extracted from the recipes and normalized to different food ontologies. It also provides a link between the food ontologies. For this reason, the food concepts are matched and for each food concept the semantic information from each data set is assigned.

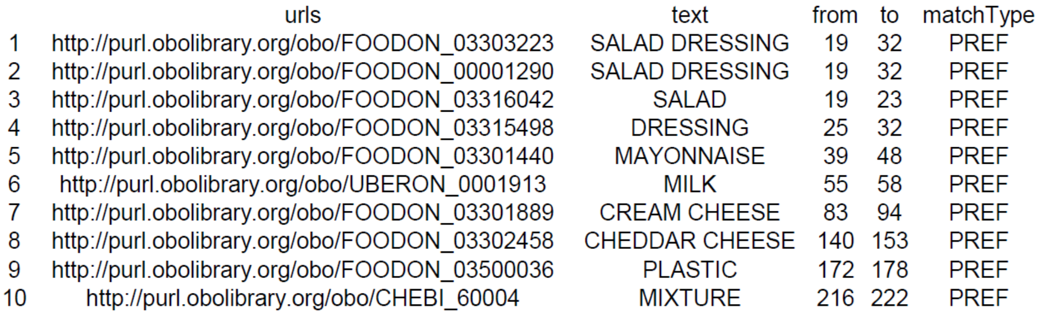
The concept matching is done by iterating through each food concept that is extracted by the NER method FoodIE. If the concept is also recognized wholly or partially by the NCBO annotator in combination with any of the selected ontologies, the semantic tags from that ontology are also assigned to the food concept. However, it is not uncommon while using the NCBO annotator for it to provide semantic tags on a token level instead of on a concept level. Such an example would be when an ontology returns two outputs for the food concept “salad dressing”, instead of classifying it as a single food concept consisting of two tokens. In these cases, each incomplete food concept extracted by the NCBO needs to be matched to its corresponding superset food concept extracted by FoodIE. This was done by checking if the location metrics from the NCBO annotator are in accordance (more specifically, whether they are a subrange) with the location metrics of a food concept provided by FoodIE. If such a match exists, the NCBO food concept is added to the

[1] `http://data.bioontology.org/documentation`



```
<document>
    <id>0recipe1090</id>
    <infon key="category">Appetizers and snacks</infon>
    <infon key="full_text">
    Mix the dry ranch salad dressing mix, mayonnaise, and milk in a bowl. Beat in the cream cheese with an electric mixer until smooth. Mix in Cheddar cheese. Cover
    bowl with plastic wrap, and freeze 30 minutes. Divide mixture in half, and shape into balls. Roll each ball in almonds to coat. Cover and refrigerate balls until
    ready to serve.
    </infon>
    <annotation id="1">
        <location offset="3" length="28"/>
        <text>dry ranch salad dressing mix</text>
        <infon key="semantic_tags"> AG.01.h.02 [Vegetables];AG.01.m [Substances for food preparation];AG.01.n.09 [Prepared vegetables and dishes];</infon>
    </annotation>
    <annotation id="2">
        <location offset="9" length="10"/>
        <text>mayonnaise</text>
        <infon key="semantic_tags"> AG.01.l.04 [Sauce/dressing];AG.01.n.01 [Food by way of preparation];</infon>
    </annotation>
    <annotation id="3">
        <location offset="12" length="4"/>
        <text>milk</text>
        <infon key="semantic_tags"> AG.01.e [Dairy produce];</infon>
    </annotation>
    <annotation id="4">
        <location offset="20" length="12"/>
        <text>cream cheese</text>
        <infon key="semantic_tags"> AG.01.e [Dairy produce];AG.01.e.02 [Cheese];AG.01.n [Dishes and prepared food];AG.01.n.18 [Preserve];</infon>
    </annotation>
    <annotation id="5">
        <location offset="31" length="14"/>
        <text>Cheddar cheese</text>
        <infon key="semantic_tags"> AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];</infon>
    </annotation>
    <annotation id="6">
        <location offset="59" length="7"/>
        <text>almonds</text>
        <infon key="semantic_tags"> AG.01.h.01.f [Nut];</infon>
    </annotation>
</document>
```

Figure 4.10: BioC format for a single recipe.



	urls	text	from	to	matchType
1	http://purl.obolibrary.org/obo/FOODON_03303223	SALAD DRESSING	19	32	PREF
2	http://purl.obolibrary.org/obo/FOODON_00001290	SALAD DRESSING	19	32	PREF
3	http://purl.obolibrary.org/obo/FOODON_03316042	SALAD	19	23	PREF
4	http://purl.obolibrary.org/obo/FOODON_03315498	DRESSING	25	32	PREF
5	http://purl.obolibrary.org/obo/FOODON_03301440	MAYONNAISE	39	48	PREF
6	http://purl.obolibrary.org/obo/UBERON_0001913	MILK	55	58	PREF
7	http://purl.obolibrary.org/obo/FOODON_03301889	CREAM CHEESE	83	94	PREF
8	http://purl.obolibrary.org/obo/FOODON_03302458	CHEDDAR CHEESE	140	153	PREF
9	http://purl.obolibrary.org/obo/FOODON_03500036	PLASTIC	172	178	PREF
10	http://purl.obolibrary.org/obo/CHEBI_60004	MIXTURE	216	222	PREF

Figure 4.11: NCBO tagger output for a single recipe, using the ontology FoodOn.

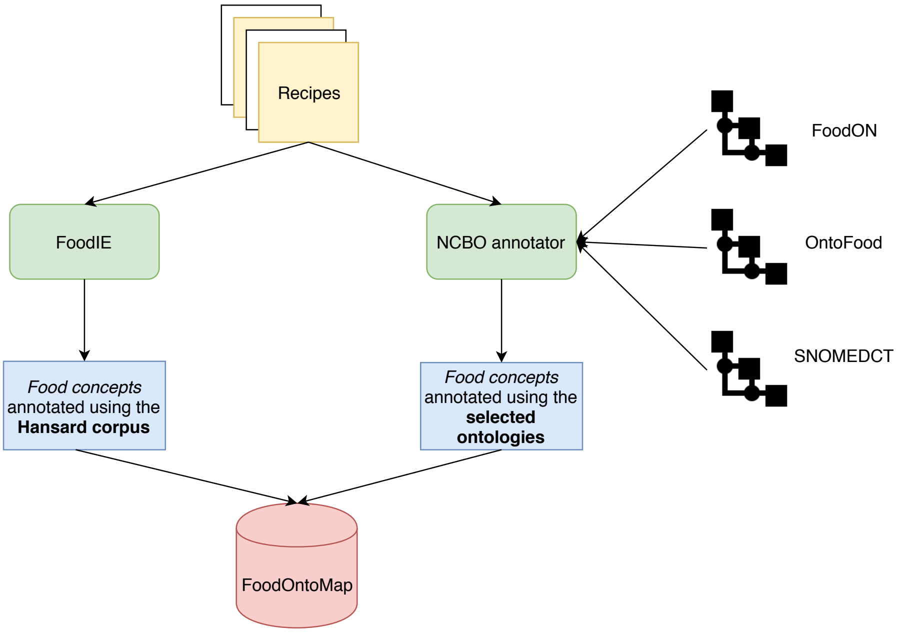

Recipes
FoodIE
NCBO annotator
FoodON
OntoFood
SNOMEDCT
Food concepts annotated using the **Hansard corpus**
Food concepts annotated using the **selected ontologies**
FoodOntoMap

Figure 4.12: FoodOntoMap design.

mapping set of the FoodIE concept. By doing this, the mapping sets aggregate all the corresponding food concept mentions that map to a food concept, along with their semantic information, even if the NCBO annotator does not classify some food concepts wholly.

To illustrate this with an example, in Figure 4.10, it is evident that we have multiple food concept mentions that overlap. Such mentions are “SALAD DRESSING”, starting at 19 and ending at 32 and appearing twice with different semantic information; “SALAD”, starting at 19 and ending at 23; and “DRESSING”, starting at 25 and ending at 32. During the matching step that is mentioned above, all these are aggregated into a single food concept mention that holds all the semantic information of its component food concepts. The sole food concept mention becomes “SALAD DRESSING”, starting at 19 and ending at 32 along with four different semantic tags, one from each component mention. Then, after converting the location metrics to be compatible with the BioC recipe format, the food concept is matched to its corresponding food concept from the BioC recipe data set. This means that the resulting mapping is “dry ranch salad dressing mix” to “SALAD DRESSING”. Notice that not all tokens are caught by the NCBO annotator using FoodOn. With this, the normalization is complete, the code mapping being A000046 to B000027.

In instances where one code from a data set maps to two separate codes from another data set, the NCBO tagger failed to produce one food concept mention which contains the full semantic information. Instead, it produced only token-level mentions, and as such they were not aggregated into one larger food concept mention. Such an example is “SALTED WATER”, which maps to “WATER” and “SALTED” separately, the codes being A000123, B000012 and B000065, respectively.

The flowchart of the FoodOntoMap design is presented in Figure 4.12.

The availability of the resources that are used to create the FoodOntoMap is presented in Table 4.7.

The results from the FoodOntoMap are four different data sets and one mapping. Each data set consists of an artificial id for each unique food concept that is extracted using each approach, the name of the extracted food concept, and the semantic information assigned to it. Each data set corresponds to one of the four semantic resources: Hansard corpus,

Table 4.7: Availability of resources used to create FoodOntoMap.

Resource name	Availability
FoodIE	https://github.com/GorjanP/foodie
BioC format recipe set	http://cs.ijs.si/repository/FoodBase/foodbase.zip
Mapper client and NCBO annotator client	https://github.com/GorjanP/FOM_mapper_client
Hansard corpus	https://www.hansard-corpus.org/
NCBO annotator	http://bioportal.bioontology.org/annotator
NCBO annotator REST API	http://data.bioontology.org/documentation
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

FoodOn, OntoFood, and SNOMED CT. At the end there is one data set mapping, called FoodOntoMap, which, for each concept that appears at least in two data sets, provides the links between them by listing the artificial id of the concepts from each of the data sets in which it is mentioned.

An example mapping for one instance from FoodOntoMap is A000016, B000011, C000012, D000002. This means that this food concept is mapped from the Hansard corpus to the respective ontologies, i.e. FoodOn, SNOMED CT and OntoFood. The provided codes are artificial unique identifiers that are assigned to the food concept using each of the aforementioned food semantic resources, respectively. If we look at the separate data sets, we can see that A000016 corresponds to "garlic" with semantic tag AG.01.h.02.e [Onion/leek/garlic] from the Hansard corpus, B000011 corresponds to "garlic" with semantic tag http://purl.obolibrary.org/obo/NCBITaxon_4682 from the FoodOn, C000012 is for "garlic" with the semantic tag http://purl.bioontology.org/ontology/SNOMEDCT/735030001 from the SNOMED CT, and D000002 corresponds to enquotegarlic with the semantic tag http://www.owl-ontologies.com/Ontology1435740495.owl#Garlic from the OntoFood.

The data sets consist of 13,205; 1,069; 111; and 582 unique food concepts, obtained using Hansard corpus, FoodOn, OntoFood, and SNOMED CT, respectively. The FoodOntoMap data set consists of 1,459 food concepts that are found in at least two food semantic resources.

From the results, we can conclude that FoodIE with the Hansard corpus gives the most promising results because it can extract a larger number of food concepts compared with the NCBO annotator in combination with some of the selected ontologies. Moreover, the results prove that the three food ontologies (i.e. SNOMED CT, FoodOn, OntoFood) do not represent the food domain well, as many food concepts cannot be extracted using the NCBO annotator, which indicates that they do not exist in the food ontologies themselves.

4.3.2 Discussion

Impact. To the best of our knowledge, FoodOntoMap is the first resource that provides normalization of food concepts to different food ontologies, additionally providing a link between them. The motivation for building such a resource in the food domain comes from the existence of the UMLS, which is extensively used in the biomedical domain. For example, the MRCONSO.RRF table that is a part of the UMLS is used in a lot of semantic web applications because it can map the medical concepts to different biomedical standards and vocabularies. To make progress in analysing the large amount of data that is available in order to find these relations, resources for food concepts normalization are extremely valuable and welcome.

The main benefit of using FoodOntoMap is that the food concepts can be normalized by mapping them to a unified system. Furthermore, the semantic tags can be reused to find the non-linear relations that exist between the concepts in the vector space, by learning the embedding space. This can also be done together with some medical and environment concepts. Once the embedding space is learned, the embeddings can be used for predictive studies in order to explain the relations between human health, food systems, and the environment.

Reusability. FoodOntoMap can be used as a resource that represents a normalized data set of food concepts. Additionally, users can also follow the described pipeline of steps used to create the FoodOntoMap mapping in order to create their own new resource where the food concepts will be normalized. Both approaches, FoodIE and NCBO, used for food concepts extraction have already been well-documented and evaluated. The ontologies that are used by the NCBO annotator have also been well documented and are easy to utilize. Furthermore, FoodOntoMap can be easily extended on additional recipes, as well as a wide variety of different ontologies. With this it can provide an ever wider coverage of food concepts. Additionally, the FoodOntoMap pipeline can be used to normalize food concepts that exist in food consumption and food composition databases.

Availability. The resource FoodOntoMap is published and publicly available for download at `https://doi.org/10.5281/zenodo.2635437`. under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) licence. With this, we encourage users to further contribute to this resource and modify it as need be. Zenodo was the platform of choice, as it provided all the framework tools needed for such a data set. Hosted along with the resource's data sets is a DCAT specification file, which briefly describes the structure of the resource. Furthermore, the resource will be actively maintained and extended as new ontologies, annotators, NER methods and NLP methods become available. The goal is to keep the resource relevant and contemporary while also improving its domain coverage with the ever improving NLP tools.

4.3.3 Conclusion

The resource that is presented in this chapter represents a data normalization pipeline. FoodOntoMap targets the domain of food and nutrition science, normalizing food concepts across three different ontologies and one semantic data set. Additionally, it provides a set with semantic information for each food concept present in the data set. This was made possible with the use of NER methods for information extraction from unstructured textual data. FoodOntoMap represents a first of its kind in the domain of Food and Nutrition, with no other such resources being readily available to the best of our knowledge. As such, it is crucial in building tools that improve knowledge in these domains and the public health effects it implies, as well as building data models that can be utilized to further discover relations and links in the food and nutrition domain.

Chapter 5

Case Study: Food Data Normalization and Integration for a Low-Researched Language

In this chapter, we provide a case study on food data normalization and integration by using lexical and semantic similarity heuristics for a low-researched language, i.e. Slovenian. Slovenian is a South Slavic language spoken by around 2.5 million speakers worldwide. The methodology utilizes the Jaccard index for lexical similarity, while using the classic Word2Vec and GloVe text embedding methods for computing semantic similarity [28].

This chapter is adapted from [28].

5.1 Methodology

To match the information about the same food products from different data sources, we first pre-process the data. Next, we match the food products by applying lexical similarity measures, followed by matching them with regard to semantic similarity. Finally, we compare the mapping results by evaluating them on a set of pairs that represent the ground truth, which are pairs matched by their EAN bar-codes.

5.1.1 Lexical similarity

Let D_1 and D_2 be two pieces of text. First POS tagging, also called grammatical tagging, is applied to each of them to identify the part-of-speech tags such as nouns (NN, NNS, NNP, NNPS), verbs (VB, VBD, VBG, VBN, VBP, VBZ), adjectives (JJ, JJR, JJS), cardinal numbers (CD), etc. [100] Let us define

$$Y_i = \{tokens\ from\ D_i\ that\ belong\ to\ one\ word\ class\}, \tag{5.1}$$

where $i = 1, 2$. The word classes are: nouns, adjectives, verbs, adverbs, prepositions, determiners, pronouns, conjunctions, modal verbs, particles, and numerals. For example, Y_i can be a set of all tokens from D_i that are tagged as nouns. In such a case, the set consists of all tokens that are tagged as NN, NNS, NNP, and NNPS.

The next step is to define which of the extracted word classes (morphological POS tags) are significant to describe the domain to which the text belongs. The set of nouns is crucial because nouns carry most of the information in the text, while all other word classes (adjectives, verbs, numbers, etc.) only give an additional explanation. After extracting the set of nouns and the sets of other word classes that are significant for the domain, lemmatization [101] is applied to each of them. To find string similarity between both pieces of text, a probability event is defined as a product of independent events

$$X = N \prod_{j=1}^{k} Z_j, \tag{5.2}$$

where N is the similarity between the sets of nouns found in both pieces of text, k is the number of additional word classes that are selected and are significant for the domain, and Z_j is the similarity between the sets of word class, j, found in both texts. The additional word classes can be adjectives, verbs, etc.

Because these events are independent, the probability of the event X can be calculated as

$$P(X) = P(N) \prod_{j=1}^{k} P(Z_j). \tag{5.3}$$

To calculate it, the probabilities of the independent events need to be defined. Because the problem looks for the similarity between two sets, it is logical to use the Jaccard index, J, which is used in statistics for comparing similarity and diversity of sample sets [102]. For the similarity between the nouns, the Jaccard index is used, while for the similarity between the additional word classes, the Jaccard index in combination with Laplace probability estimate [103] is used. This is because, in some short segments of text, the additional information provided by other word classes can be missed, so there will be no zero probabilities. The probabilities are calculated as

$$\begin{aligned} P(N) &= \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}, \\ P(Z_j) &= \frac{|Z_{j_1} \cap Z_{j_2}| + 1}{|Z_{j_1} \cup Z_{j_2}| + 2}. \end{aligned} \tag{5.4}$$

By substituting Equation 5.4 into Equation 5.3, we obtain a weight for the matching pair.

If we focus on the food domain, or specifically on the food matching problem, let D_1 and D_2 be the (Slovenian) names of two selected food products. As we said before, the nouns carry most of the information, while the additional word classes that describe the food domain are adjectives, which explain the food item in more detail (e.g. frozen, fresh), and the verbs, which are generally related with the method of preparation (e.g. cooked, drained). Let us define

$$\begin{aligned} N_i &= \{nouns\ extracted\ from\ D_i\}, \\ A_i &= \{adjectives\ extracted\ from\ D_i\}, \\ V_i &= \{verbs\ extracted\ from\ D_i\} \qquad (5.5) \\ & \qquad (5.6) \end{aligned}$$

where $i = 1, 2$.

To find the similarity between the names of food products, an event is defined as a product of two other events

$$X = N \cdot (A + V), \tag{5.7}$$

where N is the similarity between the nouns found in N_1 and N_2, and $A+V$ is the similarity between the two sets of adjectives and verbs handled together as $A_1 + V_1$ and $A_2 + V_2$. The adjectives and verbs are handled together to avoid different forms with the same meaning. Additionally, lemmatization is applied for each extracted noun, verb and adjective, and the similarity event uses their lemmas.

Because these two events are independent, the probability of the event X can be calculated as

$$P(X) = P(N) \cdot P(A + V). \tag{5.8}$$

The probabilities are calculated as

$$\begin{aligned} P(N) &= \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|}, \\ P(A + V) &= \frac{|(A_1 \cup V_1) \cap (A_2 \cup V_2)| + 1}{|(A_1 \cup V_1) \cup (A_2 \cup V_2)| + 2} \end{aligned} \tag{5.9}$$

By substituting Equation 5.9 into Equation 5.8, we obtain a weight for each matching pair.

5.1.2 Semantic similarity

For mapping the food products from both data sets considering semantic similarity, we decided to apply two different word embedding techniques – word2vec [104] and GloVe [73]. For the model training we used the lemmas of the words contained in the names of the food products. The reason for learning vector representations for the lemmas and not the whole words is the fact that one word, grammatically, can have different cases in Slovene. Let's assume that fp is the name of the food product, which consists of n words:

$$fp = \Big\{ word_1, word_2, ..., word_n \Big\} \tag{5.10}$$

After obtaining the lemmas of each word:

$$fp = \Big\{ lemma_1, lemma_2, ..., lemma_n \Big\} \tag{5.11}$$

We then apply the two algorithms and obtain vector representations for each lemma (i.e. word) in the food product name:

$$E[lemma_a] = \Big[x_{a1}, x_{a2}, ..., x_{ad} \Big] \tag{5.12}$$

Where $a \in \{1, ..., n\}$, and d is the dimension of the generated word vectors, manually defined for the both of the algorithms. After obtaining the vector representations, the next step is to apply a heuristic for merging the vectors for all the lemmas of a name, in order to obtain the vector representation for the whole food product name. We chose to work with two heuristics:

1. Average – Calculating the vector representation for the food product name as an average from the vector representations of the lemmas of the words from which it consists of:

$$E_{average}[fp] = \left[\frac{x_{a1} + ... + x_{n1}}{n}, ..., \frac{x_{ad} + ... + x_{nd}}{n} \right] \tag{5.13}$$

2. Sum – Calculating the vector representation of the food product name as a sum from the vector representations of the lemmas of the words from which it consists of:

$$E_{sum}[fp] = \Big[x_{a1} + ... + x_{n1}, ..., x_{ad} + ... + x_{nd} \Big] \tag{5.14}$$

Finally, to perform the matching, we calculate the cosine similarity between the vector representations of the food product.

Word2vec embeddings. The only numeric parameters that varied between the different word2vec models were the dimension size and the sliding window size. Values for the sliding window were chosen to be *[2, 3, 5]*, while the dimensions were *[100, 200]*. Additionally, the feature extraction algorithms included Bag of Words and Skip-gram. By combining these parameter values, a total of 12 word2vec models were trained.

GloVe embeddings. Analogous to the word2vec parameter choice, the same values were used for the numeric parameters of GloVe, i.e. *[2, 3, 5]* for the sliding window and *[100, 200]* for the number of dimensions. Thus, a total of six models were trained.

In both cases, the sliding windows were chosen according to the average number of words per food product, which rounded equals to nine.

5.2 Data collection and pre-processing

In this section, we explain the data collection process, after which we elaborate on the data pre-processing step.

The data about food products used in this study were scrapped from the web sites of two food retailers (for convenience purposes let us name them: $Retailer_1$ and $Retailer_2$). Each website contains some, but not complete, information about each food product, such as the food product name in Slovenian, the EAN bar-code, the food label, the lists of ingredients and allergens, and the name of the producer. For the food products for which we have their food product names and EAN codes, we constructed data sets containing these two pieces of information about each product (the format of the data sets is shown in Table 5.1). It needs to be pointed out that the food names were similar, but not the same (e.g. bread is named by one retailer as "bel kruh", i.e "white bread" in English, and by another retailer as "pšenični kruh, bel", i.e. "wheat bread, white"). Where fp is the food product name, bc is the corresponding EAN code, and n is the number of food products in each data set – for $Retailer_1$, $n = 1,836$ and for $Retailer_2$, $n = 6,587$.

Table 5.1: Format of the data set collected from both retailers.

Food product name	EAN code
fp_1	bc_1
$\vdots$	$\vdots$
fp_n	bc_n

Having the data sets in the format presented in Table 5.1, before applying the algorithms for obtaining semantic similarity or calculating lexical similarity with Equation 5.9, the data needed to be pre-processed. The first step was to perform POS tagging on the food product names. Since we are working with words in Slovenian, the POS tagger that is used is for Slovenian [105]. The Slovenian tagger outputs the tokens in three types of data: word form, lemma, and morph-syntactic description or tag. We use the lower case lemmas for each word. The data consists of words spanning across multiple morphological types. However, only the lemmas nouns, adjectives, and verbs convey semantic information. Therefore, these are the only three types that are considered while calculating lexical similarity and training the word embedding models.

5.3 Evaluation

In order to produce a data set consisting of ground truth values, we matched the food products by using their corresponding EAN codes. With this, we obtained 438 food products that are available in both retailers' catalogues.

Since *Retailer*1 has significantly fewer food products to offer, we find the five most similar food products from *Retailer*2 and check whether one of them corresponds to the food product matched by the EAN code. If so, we count this as a positive example, otherwise as a negative one.

For computing the similarity between the food products, we fixed the dimensionality to 200, used a sliding window of 5 for both the word2vec and GloVe models. Additionally, word2vec was trained using CBOW. Lastly, the lexical similarity measure was computed according to formulas 5.8 and 5.9.

The hyper-parameter choice was made after evaluating the models described in Section 5.1.2. The model with the best empirical results proved to be the ones with a dimensionality of 200 and a sliding window of 5. Therefore, we use this model in our final evaluation.

To gain some insight regarding the embedding training process, it is useful to look at the values of the loss function for each training iteration (epoch). In Figure 5.1, these values are plotted. It is evident that the loss improvement plateaus after a certain point, so it is computationally beneficial to stop the training process after this plateau is reached. This also prevents over-fitting the training data, which is important if new data is added for future evaluation. In this case, the plateau is somewhere around iteration 800, which is where it is favorable to stop the training process.

5.3.1 Results and discussion

In Table 5.2, we present the results of the evaluation on the data set of 438 food products having similar, but not the same, food names and the same EAN bar-codes. It is interesting to note that both summing and averaging the vector embeddings provided identical predictive results. Additionally, in Table 5.3, the accuracy of each model is presented.

Table 5.2: Evaluation results for each text embedding model.

	Word2vec	GloVe	Lexical sim.
Positives	271	238	329
Negatives	167	200	109

Table 5.3: Accuracy for each text embedding model.

Model	Accuracy
Word2Vec	0.61872
GloVe	0.54338
Lexical	**0.75114**

Looking at Table 5.2 it follows that out of a total of 438 food products, 271 were in the top five predictions when using the word2vec model; 238 were in the top five predictions when using the GloVe model and 329 were in the top five predictions when using the lexical model.

Further insight into the matching evaluation can be obtained by counting how many food products were not found (Negatives) for all models. Specifically, we count how many

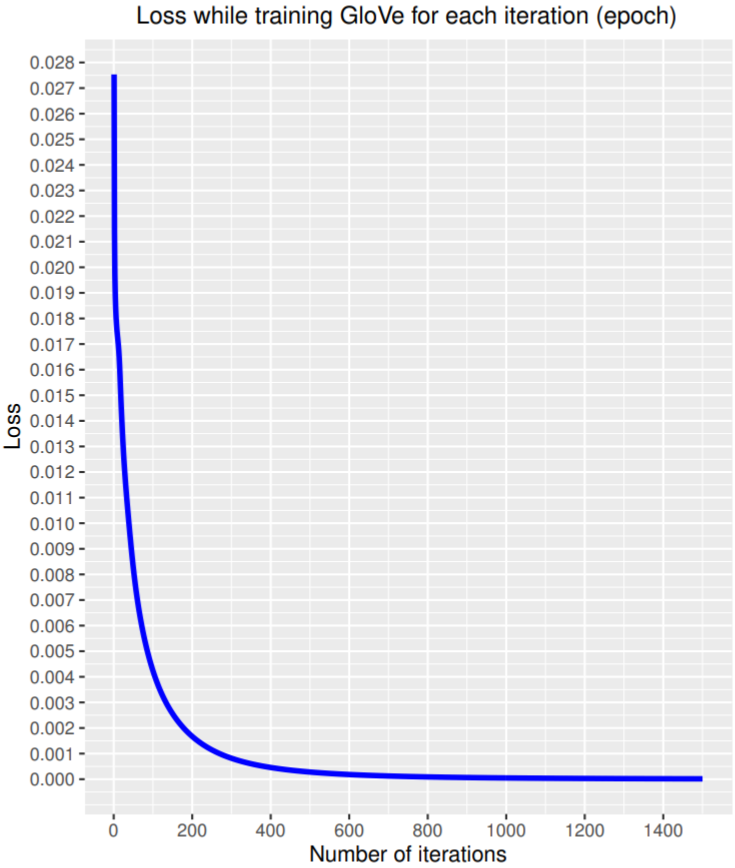


Figure 5.1: The loss function value plotted over the number of iterations (epochs) while training GloVe embeddings for the data set.

food products were not positively matched by any model at all. Taking this into consideration, 355 out of a total of 438 products in the evaluation set were positively matched by at least one of the models. These results are additionally presented in Table 5.4.

Table 5.4: Accuracy for each model

Total positively matched	Accuracy
355	0.81050

For example, in Table 5.5, the top five matches for a food product (in this case "jogurt mu borovnica 1,3mm 1l", i.e blueberry fruit yogurt) from each model is presented. In this example, one of each five matches is a positive match in the ground truth evaluation data set.

Additionally, even if the matches from the semantic models do not include the ground truth product, they still convey significant semantic information about the food products. In Table 5.6, we provide one such example, where it is evident that all five matches are related to the food product of interest. In this example, all food products are related to "sir", i.e. cheese. Therefore, the semantic models are not limited by the lexical information of the food product name and can be used to match food concepts in cases where there is low lexical similarity, but the semantic similarity is high.

Table 5.5: Positive match food product examples when ground-true EAN code match is in top five predictions.

(a) Word2Vec model

Food product:	jogurt mu borovnica 1,3mm 1l
Match 1:	sadni jogurt borovnica 1,3 m. m. mu 500g
Match 2:	sadni jogurt borovnica super 150g
Match 3:	sadni jogurt s chia semeni crni ribez borovnica 1,5 m. m. meggle 330g
Match 4:	tekoci jogurt borovnica mu 1l
Match 5:	sadni bio jogurt s senenim mlekom borovnica 150g

(b) GloVe model

Food product:	jogurt mu borovnica 1,3mm 1l
Match 1:	tekoci jogurt borovnica mu 1l
Match 2:	tekoci jogurt kramar 500g
Match 3:	grski jogurt z borovnico 0 m. m. total 170g
Match 4:	sadni jogurt borovnica 1,3 m. m. mu 500g
Match 5:	lca jogurt nula 150 g borovnica 3,3 m. m.

(c) Lexical model

Food product:	jogurt mu borovnica 1,3mm 1l
Match 1:	tekoci jogurt borovnica mu 1l
Match 2:	sadni jogurt borovnica 1,3 m. m. mu 500g
Match 3:	sadni jogurt borovnica 1,2 m. m. lca 180g
Match 4:	grski jogurt z borovnico 0 m. m. total 170g
Match 5:	sadni jogurt borovnica super 150g

Table 5.6: Top five predictions when no EAN code match is available in the data set.

(a) Word2Vec model

Food product:	topljeni sir kiri navadni 100g
Match 1:	topljeni sir slovenka 200g
Match 2:	topljeni sir 140g
Match 3:	naravni topljeni sir president 140g
Match 4:	topljeni sir camembert president 125g
Match 5:	topljeni sir gauda v listicih kaeserei champignon 150g

(b) GloVe model

Food product:	topljeni sir kiri navadni 100g
Match 1:	topljeni sir slovenka 200g
Match 2:	topljeni sir klasik zdenka 140g
Match 3:	topljeni sir klasik zdenka 280g
Match 4:	topljeni sir cardas zdenka 140g
Match 5:	topljeni sir v listicih klasik 150g

One thing to notice is that using lexical similarity as a heuristic will always yield better results when considering the task of matching branded food products, while semantic similarity as a heuristic can provide more insight when considering other tasks, such as matching food data for imputing missing nutrient values from food composition databases.

One weakness of the semantic models is that they are using embeddings on a word level. For our future work, we are planning to explore more advanced textual representational models such as BERT [106], RoBERTa [107], XLNet [108], and ALBERT [109]. There are pre-trained models for English text for all of these embedding methods. However, in order for these methods to be used with Slovenian text, we should acquire more data and train the corresponding models. Additionally, the same methodology described in this chapter can be generalized and applied to any language, provided sufficient pre-trained models, or data to train the required models on, exist.

5.4 Conclusions

The problem of food data integration becomes especially important with one of the 2030 development goals of the United Nations, which states “End hunger, achieve food security and improved nutrition and promote sustainable agriculture" [110]. With the huge amount of food and nutrition-related data that is collected in the last 10 years, there is a need for data normalization techniques that will link these data sets.

In this chapter, we propose two heuristics that can be used for matching food products represented by their non-English descriptions (i.e. Slovenian). To give a matching score of a pair of food products, the first one is based on lexical similarity, and the matching score is a probability event defined as a product of similarity between the set of nouns that appear in their names and the joint set of adjectives and verbs. The second one is based on semantic similarity and uses word embeddings. For it, first vector representations (i.e. embeddings) for the lemmas of nouns, adjectives, and verbs, which appear in food products names, are learned. After that, the vector representation of a food product name can be calculated as an average or sum from the vector representations of the lemmas of the words from which it consists of. The matching score of a pair of food products is the cosine similarity between the vector representations of their names.

We evaluated the proposed heuristics by mapping food products from two online grocery stores. We compared the results for the proposed heuristics using a data set of 438 food product pairs, which present the ground truth. They were obtained by matching the food products from every pair based on their EAN codes. By applying the proposed heuristics, for the first food product from every pair, we returned the 5 most similar food products, and we checked whether one of them corresponds to the second food product from the pair. Experimental results showed that the best semantic models achieve an accuracy of 62%, while the lexical model outperforms this with an accuracy of 75%. Additionally, if all the models are considered together, an accuracy of 81% is obtained.

For our future work, we are planning to explore more advanced textual representational methods (i.e. embeddings methods), and also use the information from graph-based embeddings to improve the matching process.

Chapter 6

FoodViz: Visualization of Food Entities Linked Across Different Standards

In the penultimate chapter, we present a visualization tool aimed at making the links between different food standards understandable by food subject-matter experts, named FoodViz. It is a web-based framework used to present food annotation results from existing Natural Language Processing and Machine Learning pipelines in combination with different food semantic data models. Using this framework, users would become more familiar with the links between different food semantic data models.

This chapter is adapted from [30].

6.1 FoodViz Overview

To make the links between different food standards understandable by food subject matter experts and to make them familiar with the interoperability process using different standards, we develop FoodViz, which is a web-based framework used to present food annotation results from existing Natural Language Processing and machine learning pipelines in conjunction with different food semantic data resources. Currently, a lot of work can already be done in an automatic way, but it is very important that the results are presented to experts in a concise way so that they can check and approve (or disapprove) the results. To show the utility of FoodViz, we visualize the results that are already published in the FoodOntoMap resource. The results consist of recipes that are coming from the curated and uncurated version of FoodBase, which was constructed by using the food NER method FoodIE.

The FoodViz[1] is a single page application developed with React[2], served by a back-end application programming interface (API) developed in Flask[3]. The back-end API serves pre-processed recipes annotated in our previous work [26] (Chapter 4) and the annotation mappings from [27] (Chapter 4).

The home page of FoodViz is presented in Figure 6.1. Three different parts exist that can be explored: “NER tagger”, “Documents”, and “Test custom document”.

The “Documents” part displays the curated and uncurated recipes of the FoodBase corpus. There are 1000 curated recipes, 200 per each recipe category, and more than

[1] `http://foodviz.ds4food.ijs.si/fbw/#/recipes`

[2] `https://reactjs.org/`

[3] `https://flask.palletsprojects.com/en/1.1.x/`

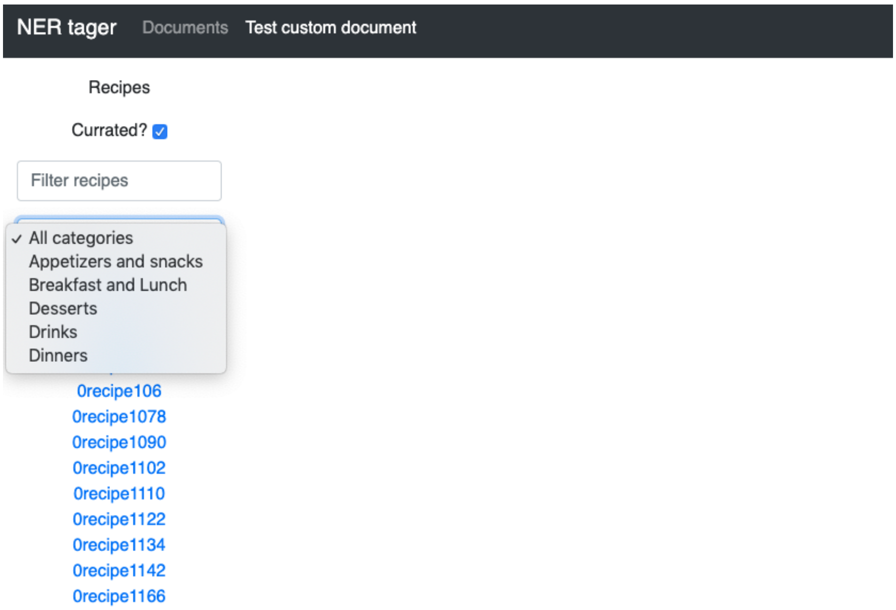


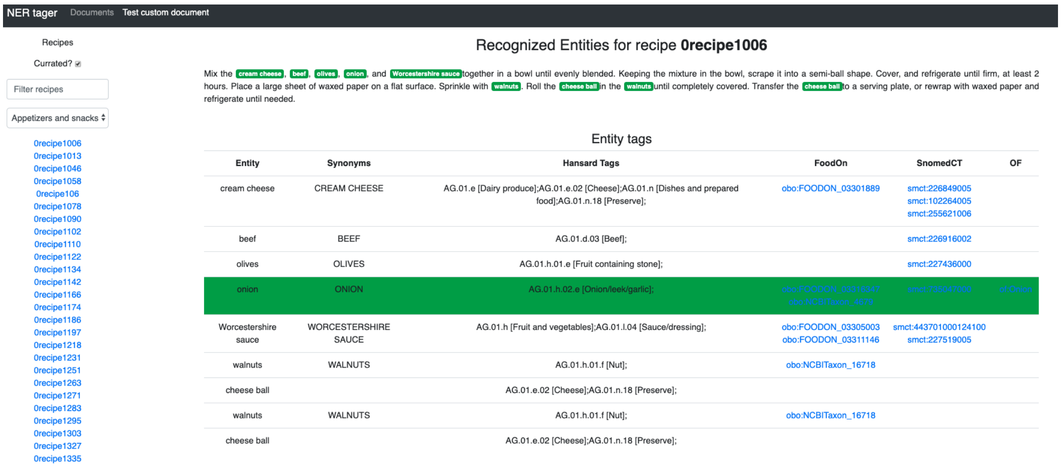
Figure 6.1: FoodViz. A new visualization tool for presenting the results published in the FoodOntoMap resource to food subject matter experts.

22.000 uncurated recipes available in FoodViz. The curated version is a ground truth data set, because in the process of developing it, the missing food entities were manually included, while the false positive entities were manually excluded from the corpus [26].

FoodViz allows users to filter the recipes by name, by the recipe category and between the curated and uncurated recipes. Next, the user can select a recipe, for which the semantic annotations are shown. Figure 6.2 presents an example for a selected curated recipe. The recipe belongs to Appetizers and snacks. Up in the top, the recipe description is presented, where all food entities (nine entities) that are mentioned in it are highlighted. FoodViz allows a selection of an entity, which is displayed in the table below. In our case, we selected “onion”. Further, for each extracted food entity the synonyms are presented, which are the food names available in different food semantic resources, followed by the semantic tags from Hansard corpus, FoodOn, SNOMED CT, and OntoFood. Additionally, users can further explore the semantic tags from the FoodOn, SNOMED CT and OntoFood, which are linked to their original semantic definitions.

Figure 6.3 presents an example for a selected uncurated recipe. The uncurated version of FoodBase does not include the false negatives entities and does not exclude the false positive food entities, since it is created from a collection of around 22,000 recipes. With this, subject matter experts can help the process of annotations, by removing the false positives, and including the false negatives, or the FoodViz tool can also be used as an annotation tool. By applying this, we will be able to create a much bigger annotated corpus that will allow training on more robust NER based on deep neural networks. Therefore, FoodViz allows manual removal of the false positives, and adding of the false negatives. In this process, as shown in Figure 6.3, the removed entities are highlighted in red in the text, and are removed from the table. The only difference in the interface for the curated recipes is that this interaction is not available, since they are already validated.

Using FoodViz, subject matter experts can understand the links between different food standards. It can easily be seen which semantic tag from one food semantic resource is equivalent to a semantic tag from another resource. With this, we also perform food ontology alignment. Additionally, let us assume that information about dietary intake



0recipe1006
0recipe1013
0recipe1046
0recipe1058
0recipe106
0recipe1078
0recipe1090
0recipe1102
0recipe1110
0recipe1122
0recipe1134
0recipe1142
0recipe1166
0recipe1174
0recipe1186
0recipe1197
0recipe1218
0recipe1231
0recipe1251
0recipe1263
0recipe1271
0recipe1283
0recipe1295
0recipe1303
0recipe1327
0recipe1335

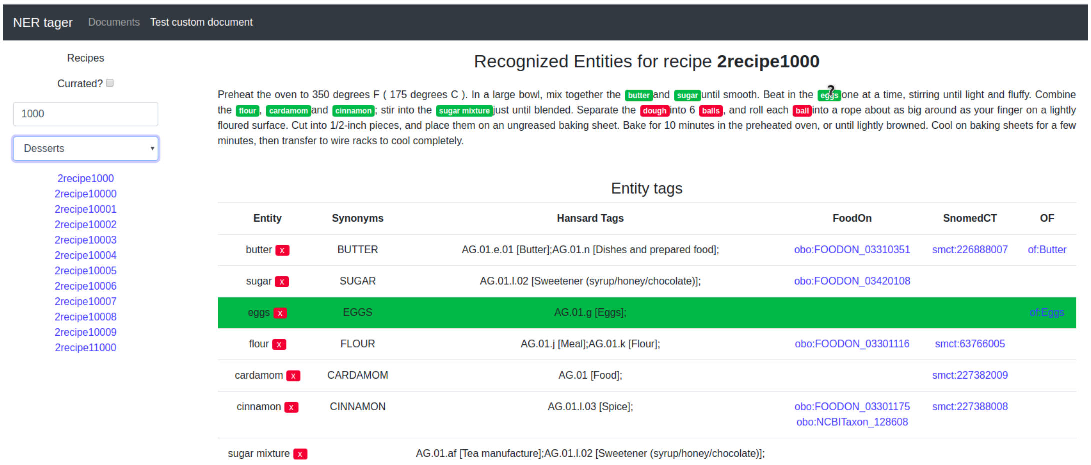
Recognized Entities for recipe **0recipe1006**

Mix the cream cheese, beef, olives, onion, and Worcestershire saucetogether in a bowl until evenly blended. Keeping the mixture in the bowl, scrape it into a semi-ball shape. Cover, and refrigerate until firm, at least 2 hours. Place a large sheet of waxed paper on a flat surface. Sprinkle with walnuts. Roll the cheese ballin the walnutsuntil completely covered. Transfer the cheese ballto a serving plate, or rewrap with waxed paper and refrigerate until needed.

Entity tags

Entity	Synonyms	Hansard Tags	FoodOn	SnomedCT	OF
cream cheese	CREAM CHEESE	AG.01.e [Dairy produce];AG.01.e.02 [Cheese];AG.01.n [Dishes and prepared food];AG.01.n.18 [Preserve];	obo:FOODON_03301889	smct:226849005 smct:102264005 smct:255621006	
beef	BEEF	AG.01.d.03 [Beef];		smct:226916002	
olives	OLIVES	AG.01.h.01.e [Fruit containing stone];		smct:227436000	
onion	ONION	AG.01.h.02.e [Onion/leek/garlic];	obo:FOODON_03316347 obo:NCBITaxon_4679	smct:735047000	of:Onion
Worcestershire sauce	WORCESTERSHIRE SAUCE	AG.01.h [Fruit and vegetables];AG.01.l.04 [Sauce/dressing];	obo:FOODON_03305003 obo:FOODON_03311146	smct:443701000124100 smct:227519005	
walnuts	WALNUTS	AG.01.h.01.f [Nut];	obo:NCBITaxon_16718		
cheese ball		AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];			
walnuts	WALNUTS	AG.01.h.01.f [Nut];	obo:NCBITaxon_16718		
cheese ball		AG.01.e.02 [Cheese];AG.01.n.18 [Preserve];			

Figure 6.2: An example FoodViz annotation for a recipe from the curated corpus. The annotation is ground truth and there is no need for editing.



2recipe1000
2recipe10000
2recipe10001
2recipe10002
2recipe10003
2recipe10004
2recipe10005
2recipe10006
2recipe10007
2recipe10008
2recipe10009
2recipe11000

Recognized Entities for recipe **2recipe1000**

Preheat the oven to 350 degrees F (175 degrees C). In a large bowl, mix together the butterand sugaruntil smooth. Beat in the eggsone at a time, stirring until light and fluffy. Combine the flour, cardamomand cinnamon; stir into the sugar mixturejust until blended. Separate the doughinto 6 balls, and roll each ballinto a rope about as big around as your finger on a lightly floured surface. Cut into 1/2-inch pieces, and place them on an ungreased baking sheet. Bake for 10 minutes in the preheated oven, or until lightly browned. Cool on baking sheets for a few minutes, then transfer to wire racks to cool completely.

Entity tags

Entity	Synonyms	Hansard Tags	FoodOn	SnomedCT	OF
butter x	BUTTER	AG.01.e.01 [Butter];AG.01.n [Dishes and prepared food];	obo:FOODON_03310351	smct:226888007	of:Butter
sugar x	SUGAR	AG.01.l.02 [Sweetener (syrup/honey/chocolate)];	obo:FOODON_03420108		
eggs x	EGGS	AG.01.g [Eggs];			of:Eggs
flour x	FLOUR	AG.01.j [Meal];AG.01.k [Flour];	obo:FOODON_03301116	smct:63766005	
cardamom x	CARDAMOM	AG.01 [Food];		smct:227382009	
cinnamon x	CINNAMON	AG.01.l.03 [Spice];	obo:FOODON_03301175 obo:NCBITaxon_128608	smct:227388008	
sugar mixture x		AG.01.af [Tea manufacture];AG.01.l.02 [Sweetener (syrup/honey/chocolate)];			

Figure 6.3: An example FoodViz annotation for a recipe from the un-curated corpus. The annotation is not manually checked and thus there is an option to manually edit the annotations.

is collected by some dietary assessment tool. This information is normalized using some food semantic data model. Further, if we want this data to be explored and exploited in combination with some health data, its transformation to SNOMED CT semantic tags will be required, since SNOMED CT is one of the most commonly used semantic resources in the biomedical domain.

For future work, the existence of FoodViz opens different directions for future work in order to make the ML results more closer to subject matter experts. In this direction, we are planning to allow users to select which food NER method they want to use for their data in order to extract the food information. Additionally, the part “Test custom document” will allow users to provide their own text and based on the selection of the NER, the extracted results will be shown.

All in all, we are able to aid and understand the process of food data interoperability, which is a crucial task that should be done as a pre-processing step before involving the data in more advanced data analyses.

6.2 Conclusion

Information coming from raw textual data is very important albeit difficult to be understand by both humans and machines. There have been debates who beats whom and it seems that machines are becoming better and better in understanding the written word [111]. Several steps need to be performed to support such a complex task, and one of them is a presentation of entities automatically identified in selected texts by ML and NLP. In this chapter, we presented the user-friendly tool, named FoodViz, whose goal is visualization of automatically annotated text in the domain of food. Additionally, it can be used as an annotation tool where subject matter experts can check the results from automatic entity extraction and data normalization methods.

To illustrate this with an example, let us present the health-related problem that is creating dietary menus in hospitals. For instance, in a menu suitable for patients with an egg allergy, eggs and egg products need to be excluded, which is not so difficult to follow because of the regulation relating to the provision of information on substances or products causing allergies or intolerances (e.g. the Regulation (EU) No 1169/2011 on the provision of food information to consumers). However, people suffering from egg allergy, especially children, frequently need to avoid other foods as well (e.g. honey, stock cube etc.), which are not specified in the list of allergens but can be found on the list of ingredients which are usually written in an unstructured way. Composing a dietary menu requires knowledge of all ingredients of all food items that are to be included in a menu, which can be a challenge for a dietitian and could be facilitated by the FoodViz tool.

Chapter 7

Conclusions and Future Work

In the final chapter, the conclusions of the thesis and directions for future work are provided.

7.1 Conclusions

The completed work presented in this thesis proves the three defined hypotheses:

"*Hypothesis 1: The development of a rule-based food Named-Entity Recognition method without the need of using pre-annotated corpora from subject-matter experts is possible.*" is proved with the creation of the FoodIE [23] NER method and its subsequent comparison with other food NER methods [25].

"*Hypothesis 2: It is possible to construct an annotated semantic resource in the food domain which provides recipes annotated with the food concepts that appear in them.*" is proved with the creation of the FoodBase [26] corpus, the first corpus containing annotated food entities.

"*Hypothesis 3: It is possible to develop a method for data normalization based on NER methods and semantic resources in the domain of Food that provides an explicit mapping between each of the used semantic resources.*" is proved with the creation of FoodOntoMap [27] where entities from different food semantic resources are interlinked. Moreover, food data normalization was further explored by analyses of LanguaL [29] and a low-resourced language case study [28].

Additionally, a food entity visualization tool named FoodViz [30] has been developed, which presents all the aforementioned developed methodologies in a user-friendly way, facilitating subject-matter integration in Machine Learning approaches.

7.2 Future Work

The developed resources, methods and tools presented in this thesis represent a basis for future endeavours in the domain of food, even potentially extending to the domain of bioinformatics and nutrition.

The extraction of food entities from unstructured text can be used to link these entities to entities from other domains such as health, bioinformatics, consumer and social sciences, etc. This would help by reducing the gaps that inhibit public health goals and the optimal development of scientific, agricultural and industrial policies. One goal for the future is to extend this NER method to support extraction of information from all fields of food science, e.g. food safety, food authenticity and traceability, and food sustainability.

Additionally, the creation of an annotated corpus with food entities would benefit further development of more accurate NER methods targeted at extracting entities from recipes, as well as other forms of text, such as scientific literature. Consequently, the exploration and extraction of relations between food entities and other biomedical entities (e.g. drug, disease, nutrients, genes, etc.) would be supported.

Moreover, the work presented in this thesis in the domain of data normalization between different food semantic resources is not restricted only to the domain of food. It can be utilized in any domain provided there exist sufficient semantic resources and corresponding NER methods. With this, it would be easy to follow the new knowledge that comes rapidly with each passing day with new scientifically published papers aimed at improving public health.

Finally, the existence of a food entity visualization tool not only brings subject-matter experts closer to the domain, but can serve as a foundation for a future framework for data annotation, facilitating the integration of advanced Machine Learning methods and with any domain of interest.

Appendix A

Availability of Resources and Methods

- FoodIE code -
  `https://github.com/GorjanP/foodie`
- FoodIE evaluation data sets and results -
  `http://cs.ijs.si/repository/FoodIE/FoodIE_datasets.zip`
- FoodIE data analysis script -
  `https://github.com/GorjanP/food_NER_comparison_script`
- FoodBase resource -
  `http://cs.ijs.si/repository/FoodBase/foodbase.zip`
- FoodOntoMap -
  `https://doi.org/10.5281/zenodo.2635437`
- FoodOntoMap mapper script -
  `https://github.com/GorjanP/FOM_mapper_client`
- FoodViz -
  `http://foodviz.ds4food.ijs.si/fbw/#/recipes`

References

[1] C. C. Aggarwal and C. Zhai, *Mining text data.* Springer Science & Business Media, 2012.

[2] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

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

[4] 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 *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2001, pp. 426–433.

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

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

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

[8] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, no. 55-66, p. 11, 2010.

[9] 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," *Journal of biomedical informatics*, vol. 58, pp. 11–18, 2015.

[10] Y. Chen, T. A. Lask, Q. Mei, Q. Chen, S. Moon, J. Wang, K. Nguyen, T. Dawodu, T. Cohen, J. C. Denny, *et al.*, "An active learning-enabled annotation system for clinical named entity recognition," *BMC medical informatics and decision making*, vol. 17, no. 2, p. 82, 2017.

[11] 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," *Knowledge-Based Systems*, vol. 132, pp. 179–187, 2017.

[12] M. M. Lopez and J. Kalita, "Deep learning applied to nlp," *arXiv preprint arXiv:1703.03091*, 2017.

[13] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," *arXiv preprint arXiv:1707.05928*, 2017.

[14] L. Gligic, A. Kormilitzin, P. Goldberg, and A. Nevado-Holgado, "Named entity recognition in electronic health records using transfer learning bootstrapped neural networks," *arXiv preprint arXiv:1901.01592*, 2019.

[15] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, and P. He, "Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition," *Journal of biomedical informatics*, p. 103 133, 2019.

[16] W. Boag, K. Wacome, T. Naumann, and A. Rumshisky, "Cliner: A lightweight tool for clinical named entity recognition," *AMIA Joint Summits on Clinical Research Informatics (poster)*, 2015.

[17] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160 035, 2016.

[18] A. L. Beam, B. Kompa, I. Fried, N. P. Palmer, X. Shi, T. Cai, and I. S. Kohane, "Clinical concept embeddings learned from massive sources of medical data," *arXiv preprint arXiv:1804.01486*, 2018.

[19] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun, "Multi-layer representation learning for medical concepts," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1495–1504.

[20] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, p. 26 094, 2016.

[21] O. Bodenreider, "The unified medical language system (umls): Integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

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

[23] G. Popovski, S. Kochev, B. Koroušić Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction," in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, (ICPRAM 2019)*, 2019, pp. 915–922.

[24] M. N. K. Boulos, 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.

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

[26] G. Popovski, B. Koroušić Seljak, and T. Eftimov, "FoodBase corpus: a new resource of annotated food entities," *Database*, vol. 2019, Nov. 2019, baz121, ISSN: 1758-0463. DOI: `https://doi.org/10.1093/database/baz121`. eprint: `https://academic.oup.com/database/article-pdf/doi/10.1093/database/baz121/30350820/baz121.pdf`.

[27] G. Popovski., B. K. Seljak., and T. Eftimov., "Foodontomap: Linking food concepts across different food ontologies," in *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 2: KEOD,*, INSTICC, SciTePress, 2019, pp. 195–202, ISBN: 978-989-758-382-7. DOI: `10.5220/0008353201950202`.

[28] G. Popovski., G. Ispirova., N. Hadzi-Kotarova., E. Valenčič., T. Eftimov., and B. K. Seljak., "Food data integration by using heuristics based on lexical and semantic similarities," in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF,*, INSTICC, SciTePress, 2020, pp. 208–216, ISBN: 978-989-758-398-8. DOI: `10.5220/0008990602080216`.

[29] G. Popovski, B. Paudel, T. Eftimov, and B. Koroušić Seljak, "Exploring a standardized language for describing foods using embedding techniques," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 5172–5176.

[30] R. Stojanov, G. Popovski, N. Jofce, D. Trajanov, B. Koroušić Seljak, and T. Eftimov, "Foodviz: Visualization of food entities linked across different standards," in *Proceedings of The Sixth International Conference on Machine Learning*, 2020, In Press.

[31] C. Nédellec, R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, "Overview of bionlp shared task 2013," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, 2013, pp. 1–7.

[32] E. Chaix, B. Dubreucq, A. Fatihi, D. Valsamou, R. Bossy, M. Ba, L. Deléger, P. Zweigenbaum&, P. Bessieres, L. Lepiniec, *et al.*, "Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016," *ACL 2016*, p. 1, 2016.

[33] Y. Luo, Ö. Uzuner, and P. Szolovits, "Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations," *Briefings in bioinformatics*, vol. 18, no. 1, pp. 160–178, 2017.

[34] J.-D. Kim, Y. Wang, N. Colic, S. H. Baek, Y. H. Kim, and M. Song, "Refactoring the genia event extraction shared task toward a general framework for ie-driven kb development," *ACL 2016*, p. 23, 2016.

[35] C. Li, Z. Rao, and X. Zhang, "Litway, discriminative extraction for different bio-events," *ACL 2016*, p. 32, 2016.

[36] H. V. Cook, E. Pafilis, and L. J. Jensen, "A dictionary-and rule-based system for identification of bacteria and habitats in text," *ACL 2016*, p. 50, 2016.

[37] J. Lever and S. J. Jones, "Verse: Event and relation extraction in the bionlp 2016 shared task," *ACL 2016*, p. 42, 2016.

[38] M. Tiftikci, H. Sahin, B. Büyüköz, A. Yayıkçı, and A. Ozgür, "Ontology-based categorization of bacteria and habitat entities using information retrieval techniques," *ACL 2016*, p. 56, 2016.

[39] F. Mehryary, J. Björne, S. Pyysalo, T. Salakoski, and F. Ginter, "Deep learning with minimal training data: Turkunlp entry in the bionlp shared task 2016," *ACL 2016*, p. 73, 2016.

[40] N. C. Panyam, G. Khirbat, K. Verspoor, T. Cohn, and K. Ramamohanarao, "Seedev binary event extraction using svms and a rich feature set," *ACL 2016*, p. 82, 2016.

[41] L. Smith, L. K. Tanabe, R. J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, *et al.*, "Overview of biocreative ii gene mention recognition," *Genome biology*, vol. 9, no. 2, p. 1, 2008.

[42] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.

[43] C.-J. Kuo, Y.-M. Chang, H.-S. Huang, K.-T. Lin, B.-H. Yang, Y.-S. Lin, C.-N. Hsu, and I.-F. Chung, "Rich feature set, unification of bidirectional parsing and dictionary filtering for high f-score gene mention tagging," in *Proceedings of the second BioCreative challenge evaluation workshop*, Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, vol. 23, 2007, pp. 105–107.

[44] H.-S. Huang, Y.-S. Lin, K.-T. Lin, C.-J. Kuo, Y.-M. Chang, B.-H. Yang, I.-F. Chung, and C.-N. Hsu, "High-recall gene mention recognition by unification of multiple backward parsing models," in *Proceedings of the second BioCreative challenge evaluation workshop*, Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, vol. 23, 2007, pp. 109–111.

[45] R. Klinger, C. M. Friedrich, J. Fluck, and M. Hofmann-Apitius, "Named entity recognition with combinations of conditional random fields," in *Proc. of the Second BioCreative Challenge Evaluation Workshop*, 2007, pp. 89–91.

[46] C. N. Arighi, Z. Lu, M. Krallinger, K. B. Cohen, W. J. Wilbur, A. Valencia, L. Hirschman, and C. H. Wu, "Overview of the biocreative iii workshop," *BMC bioinformatics*, vol. 12, no. 8, p. 1, 2011.

[47] C. N. Arighi, C. H. Wu, K. B. Cohen, L. Hirschman, M. Krallinger, A. Valencia, Z. Lu, J. W. Wilbur, and T. C. Wiegers, "Biocreative-iv virtual issue," *Database*, vol. 2014, bau039, 2014.

[48] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, and A. Valencia, "Chemdner: The drugs and chemical names extraction challenge," *Journal of cheminformatics*, vol. 7, no. 1, p. 1, 2015.

[49] M. Krallinger, O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, *et al.*, "The chemdner corpus of chemicals and drugs and its annotation principles," *Journal of cheminformatics*, vol. 7, no. 1, p. 1, 2015.

[50] S. Kim, R. I. Doğan, A. Chatr-Aryamontri, C. S. Chang, R. Oughtred, J. Rust, R. Batista-Navarro, J. Carter, S. Ananiadou, S. Matos, *et al.*, "Biocreative v bioc track overview: Collaborative biocurator assistant task for biogrid," *Database*, vol. 2016, baw121, 2016.

[51] Q. Wang, S. S. Abdul, L. Almeida, S. Ananiadou, Y. I. Balderas-Martınez, R. Batista-Navarro, D. Campos, L. Chilton, H.-J. Chou, G. Contreras, *et al.*, "Overview of the interactive task in biocreative v," *Database*, vol. 2016, baw119, 2016.

[52] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu, "Assessing the state of the art in biomedical relation extraction: Overview of the biocreative v chemical-disease relation (cdr) task," *Database*, vol. 2016, baw032, 2016.

[53] D. C. Comeau, R. I. Doğan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, *et al.*, "Bioc: A minimalist approach to interoperability for biomedical text processing," *Database*, vol. 2013, bat064, 2013.

[54] M. Rastegar-Mojarad, S. Liu, Y. Wang, N. Afzal, L. Wang, F. Shen, S. Fu, and H. Liu, "Biocreative/ohnlp challenge 2018," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, 2018, pp. 575–575.

[55] J. Xia, X. Zhang, D. Yuan, L. Chen, J. Webster, and A. C. Fang, "Gene prioritization of resistant rice gene against xanthomas oryzae pv. oryzae by using text mining technologies," *BioMed research international*, vol. 2013, 2013.

[56] A. B. do Nascimento, G. M. R. Fiates, A. dos Anjos, and E. Teixeira, "Analysis of ingredient lists of commercially available gluten-free and gluten-containing food products using the text mining technique," *International journal of food sciences and nutrition*, vol. 64, no. 2, pp. 217–222, 2013.

[57] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level," *PLoS computational biology*, vol. 10, no. 1, e1003432, 2014.

[58] S. Mori, T. Sasada, Y. Yamakata, and K. Yoshino, "A machine learning approach to recipe text processing," in *Proceedings of the 1st Cooking with Computer Workshop*, 2012, pp. 29–34.

[59] Y. Chen, "A statistical machine learning approach to generating graph structures from food recipes," PhD thesis, 2017.

[60] P. Rayson, D. Archer, S. Piao, and A. M. McEnery, "The ucrel semantic analysis system.," 2004.

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

[62] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, *et al.*, "Bioportal: Ontologies and integrated data resources at the click of a mouse," *Nucleic acids research*, vol. 37, no. suppl_2, W170–W173, 2009.

[63] D. Çelik, "Foodwiki: Ontology-driven mobile safe food consumption system," *The Scientific World Journal*, vol. 2015, 2015.

[64] C. Caracciolo, A. Stellato, S. Rajbahndari, A. Morshed, G. Johannsen, Y. Jaques, and J. Keizer, "Thesaurus maintenance, alignment and publication as linked data: The agrovoc use case," *International Journal of Metadata, Semantics and Ontologies*, vol. 7, no. 1, pp. 65–75, 2012.

[65] M. Kolchin and D. Zamula, "Food product ontology: Initial implementation of a vocabulary for describing food products," in *Proceeding of the 14th Conference of Open Innovations Association FRUCT, Helsinki, Finland*, 2013, pp. 11–15.

[66] C. Snae and M. Brückner, "Foods: A food-oriented ontology-driven system," in *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, IEEE, 2008, pp. 168–176.

[67] 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.

[68] K. Donnelly, "Snomed-ct: The advanced terminology and coding system for ehealth," *Studies in health technology and informatics*, vol. 121, p. 279, 2006.

[69] M. Alexander and J. Anderson, "The hansard corpus, 1803-2003," 2012.

[70] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Standfood: Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2," *Nutrients*, vol. 9, no. 6, p. 542, 2017.

[71] E. F. S. A. (EFSA), "The food classification and description system foodex 2 (revision 2)," *EFSA Supporting Publications*, vol. 12, no. 5, 804E, 2015.

[72] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[73] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[74] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2016, pp. 855–864.

[75] Y. Dong, N. V. Chawla, and A. Swami, "Metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, 2017, pp. 135–144.

[76] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, 2013, pp. 2787–2795.

[77] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in neural information processing systems*, 2017, pp. 6338–6347.

[78] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: The metamap program," 2001.

[79] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[80] N. Collier, A. Oellrich, and T. Groza, "Concept selection for phenotypes and diseases using learn to rank," *Journal of biomedical semantics*, vol. 6, no. 1, p. 24, 2015.

[81] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, *et al.*, "Overview of biocreative ii gene normalization," *Genome biology*, vol. 9, no. 2, S3, 2008.

[82] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki, *et al.*, "The gene normalization task in biocreative iii," *BMC bioinformatics*, vol. 12, no. 8, S2, 2011.

[83] T. Eftimov and B. K. Seljak, "Pos tagging-probability weighted method for matching the internet recipe ingredients with food composition data," in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, IEEE, vol. 1, 2015, pp. 330–336.

[84] G. Ispirova, T. Eftimov, B. Korousic-Seljak, and P. Korosec, "Mapping food composition data from various data sources to a domain-specific ontology.," in *KEOD*, 2017, pp. 203–210.

[85] T. Eftimov, G. Ispirova, P. Finglas, P. Korosec, and B. Korousic-Seljak, "Quisper ontology learning from personalized dietary web services.," in *KEOD*, 2018, pp. 277–284.

[86] D. Metzler, S. Dumais, and C. Meek, "Similarity measures for short segments of text," in *European conference on information retrieval*, Springer, 2007, pp. 16–27.

[87] T. Arnold and L. Tilton, *Corenlp: Wrappers around stanford corenlp tools*, R package version 0.4-2, 2016. [Online]. Available: `https://CRAN.R-project.org/package=coreNLP`.

[88] S. Groves, *How allrecipes. com became the worlds largest food/recipe site. roi of social media (blog)*, 2013.

[89] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "Grammar and dictionary based named-entity linking for knowledge extraction of evidence-based dietary recommendations.," in *KDIR*, 2016, pp. 150–157.

[90] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. Brinkman, and W. W. Hsiao, "Foodon: A harmonized food ontology to increase global food traceability, quality control and data integration," *npj Science of Food*, vol. 2, no. 1, pp. 1–10, 2018.

[91] S. Cheong, S. H. Oh, and S.-Y. Lee, "Support vector machines with binary tree architecture for multi-class classification," *Neural Information Processing-Letters and Reviews*, vol. 2, no. 3, pp. 47–51, 2004.

[92] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[93] P. Pant, A. S. Sabitha, T. Choudhury, and P. Dhingra, "Multi-label classification trending challenges and approaches," in *Emerging Trends in Expert Applications and Security*, Springer, 2019, pp. 433–444.

[94] J. D. Ireland and A. Møller, "Langual food description: A learning process," *European journal of clinical nutrition*, vol. 64, no. S3, S44, 2010.

[95] W. Becker, A. Møller, J. Ireland, M. Roe, I. Unwin, and H. Pakkala, "Proposal for structure and detail of a eurofir standard on food composition data ii: Technical annex," *Danish Food Information*, 2008.

[96] J. Brussaard, M. Löwik, L. Steingrimsdottir, A. Møller, J. Kearney, S. De Henauw, and W. Becker, "A european food consumption survey method–conclusions and recommendations," *European Journal of Clinical Nutrition*, vol. 56, no. S2, S89, 2002.

[97] S. Lockhead, "The global data synchronisation network (gdsn): Technology and standards improving supply chain efficiency," in *First International Technology Management Conference*, IEEE, 2011, pp. 630–637.

[98] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[99] B. Dhingra, C. J. Shallue, M. Norouzi, A. M. Dai, and G. E. Dahl, "Embedding text in hyperbolic spaces," *arXiv preprint arXiv:1806.04313*, 2018.

[100] L. Màrquez and H. Rodrıguez, "Part-of-speech tagging using decision trees," in *European Conference on Machine Learning*, Springer, 1998, pp. 25–36.

[101] T. Korenius, J. Laurikkala, K. Järvelin, and M. Juhola, "Stemming and lemmatization in the clustering of finnish text documents," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM, 2004, pp. 625–633.

[102] S. Kosub, "A note on the triangle inequality for the jaccard distance," *Pattern Recognition Letters*, vol. 120, pp. 36–38, 2019.

[103] B. Cestnik *et al.*, “Estimating probabilities: A crucial task in machine learning.,” in *ECAI*, vol. 90, 1990, pp. 147–149.

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

[105] M. Grcar, S. Krek, and K. Dobrovoljc, “Obeliks: Statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik,” in *Zbornik Osme konference Jezikovne tehnologije, Ljubljana, Slovenia*, 2012.

[106] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[107] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[108] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *arXiv preprint arXiv:1906.08237*, 2019.

[109] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019.

[110] A. Lartey, “End hunger, achieve food security and improved nutrition and promote sustainable agriculture,” *UN Chronicle*, vol. 51, no. 4, pp. 6–8, 2015.

[111] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. (2020). Glue: A multi-task benchmark and analysis platform for natural language understanding, [Online]. Available: `https://gluebenchmark.com` (visited on 04/30/2020).

Bibliography

Publications Related to the Thesis

Journal Articles

G. Popovski, B. Koroušić Seljak, and T. Eftimov, "FoodBase corpus: a new resource of annotated food entities," *Database*, vol. 2019, Nov. 2019, baz121, ISSN: 1758-0463. DOI: https://doi.org/10.1093/database/baz121. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baz121/30350820/baz121.pdf.

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

T. Eftimov, G. Popovski, E. Valenčič, and B. K. Seljak, "Foodex2vec: New foods' representation for advanced food data analysis," *Food and Chemical Toxicology*, vol. 138, p. 111 169, 2020, ISSN: 0278-6915. DOI: https://doi.org/10.1016/j.fct.2020.111169. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278691520300570.

Conference Paper

G. Popovski, S. Kochev, B. Koroušić Seljak, and T. Eftimov, "Foodie: A rule-based named-entity recognition method for food information extraction," in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, (ICPRAM 2019)*, 2019, pp. 915–922.

G. Popovski., G. Ispirova., N. Hadzi-Kotarova., E. Valenčič., T. Eftimov., and B. K. Seljak., "Food data integration by using heuristics based on lexical and semantic similarities," in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF,*, INSTICC, SciTePress, 2020, pp. 208–216, ISBN: 978-989-758-398-8. DOI: 10.5220/0008990602080216.

G. Popovski., B. K. Seljak., and T. Eftimov., "Foodontomap: Linking food concepts across different food ontologies," in *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 2: KEOD,*, INSTICC, SciTePress, 2019, pp. 195–202, ISBN: 978-989-758-382-7. DOI: 10.5220/0008353201950202.

G. Popovski, B. Paudel, T. Eftimov, and B. Koroušić Seljak, "Exploring a standardized language for describing foods using embedding techniques," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 5172–5176.

R. Stojanov, G. Popovski, N. Jofce, D. Trajanov, B. Koroušić Seljak, and T. Eftimov, "Foodviz: Visualization of food entities linked across different standards," in *Proceedings of The Sixth International Conference on Machine Learning*, 2020, In Press.

Biography

The author of this thesis, Gorjan Popovski, was born in 1997 in Bitola, Republic of Macedonia. He completed his undergraduate degree at the Faculty of Computer Science and Engineering in Skopje, Macedonia, in 2019. Later in the same year he started his Master's studies at the Jožef Stefan International Postgraduate School in Ljubljana, Slovenia, with Barbara Koroušić Seljak as his supervisor and Tome Eftimov as his co-supervisor. He additionally took the position of a Research Assistant at the Computer Systems Department at the Jožef Stefan Institute.

His passion for computer science sparked with his introduction to competitive programming while still in primary school, and he continued attending national and international competitions in which he received multiple awards. After starting his undergraduate studies, he continued his passion for competitive programming by preparing high school students for national and international competitions.

While pursuing his undergraduate degree, he was awarded twice for his excellent academic performance throughout the academic years between 2016 and 2018. Additionally, he was awarded a scholarship from the Ministry of Education of the Republic of Macedonia for excellent undergraduate academic success for two years. Finally, he was awarded the Ad Futura Scholarship for studying in Slovenia when he enrolled in his Master's programme.

His main topics of interest include machine learning, deep learning, natural language processing, algorithms and complexity, knowledge engineering and knowledge representation. His works have been published in several international journals and as part of international conferences and workshops.