FoodViz: Visualization of Food Entities Linked Across Different Standards

Riste Stojanov[1], Gorjan Popovski[2,3], Nasi Jofce[1], Dimitar Trajanov[1], Barbara Koroušić Seljak[2], and Tome Eftimov[2]

[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius, University, Skopje, North Macedonia
{riste.stojanov,nasi.jofce,dimitar.trajanov}@finki.ukim.mk
[2] Computer Systems Department, Jožef Stefan Institute, Ljubljana, Slovenia
{gorjan.popovski,barbara.korousic,tome.eftimov}@ijs.si
[3] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Abstract. Many research questions from different domains involve combining different data sets in order to explore a research hypothesis. One of the main problems that arises here is that different data sets are structured with respect to different domain standards and ensuring their interoperability is a time-consuming task. In the biomedical domain, the Unified Medical Language System supports interoperability between biomedical data sets by providing semantic resources and Natural Language Processing tools for automatic annotation. This allows users also to understand the links between different biomedical standards. While there are extensive resources available for the biomedical domain, the food and nutrition domain is relatively low-resourced. To make the links between different food standards understandable by food subject matter experts we propose the 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.

Keywords: visualization · food named-entity recognition · big data on food and nutrition

1 Introduction

Recently published studies have shown that food is one of the most important environmental factors that is interrelated with human health [1], [2], [19]. Due to this, the United Nations have determined “End hunger, achieve food security and improved nutrition and promote sustainable agriculture” as one of its sustainable development goals to be achieved by the target date of 2030 [22]. In support of this goal, many research studies are conducted across the globe, producing big amounts of heterogeneous data whose analysis may find answers to complex questions related to agri-food, nutrition/health, and the environment. However, a synergy between already conducted studies first needs to be achieved. This requires a fusion of their data sets that is a challenge due of their diversity and the diversity of their meta data. Although food is a common every-day concept, it is described by different concepts and entities in the domains of agri-food, nutrition/health, and environment.

One of the most challenging issues that should be solved before starting the modelling process is to make food- and nutrition-related data sets from the domains of agri-food, nutrition/health, and environment interoperable. This implicates that each of them uses different, already established, food standards to describe the food data. However the links between these different standards are not explored in detail [29].

Conversely, in the past two decades, a huge amount of work and effort have been done in the biomedical domain. Many standards have already been interlinked by developing the Unified Medical Language System (UMLS) [9]. The UMLS is a collection of several health and biomedical vocabularies and standards, supported with lexical and semantic similarity tools for data normalization of biomedical entities [7], as well as with different visualization tools to make subject matter experts familiar with the information that is presented in it. All of this contributes to the rapid developments in predictive healthcare, which utilize these resources in combination with state-of-the-art artificial intelligence (AI)-based methods.

Compared to the biomedical domain, the domains of agri-food, nutrition/health, and environment have several food standards that can be used to describe the food data depending on the research question that is being addressed. Some of them are either still being developing or under exploration in order to discover their utility. However, it is a clear goal that links between them should be established in order to combine different data sets (e.g., on food intake and lifestyle, food composition and bioactivity, food safety, food authenticity and traceability, dietary guidelines available as text, electronic health records, agricultural data, environmental data etc.) and to answer more research questions (e.g. from foodomics). Additionally, the links between different food standards should be well-understood by food subject matter experts in order to know how to code their collected research data, to make it further interoperable with other research data sets.

To make subject matter experts familiar with the links that exist between different food standards, we present FoodViz, a user-friendly tool for visualization of automatically annotated text on foods. It helps the process of food data interoperability and additionally offers a possibility of correcting the results given by automatic extraction and data normalization.

The rest of the paper is organized as follows. Section 2 provides an overview of food named-entity recognition methods, followed by Section 3 in which food semantic resources along with food data normalization methods are presented. Section 4 presents the FoodViz tool. Finally, the conclusions of the paper are presented in Section 5.

2 Food Named-entity Recognition

To extract information related to a specific domain from raw textual data, information extraction (IE) should be applied. IE is a task from natural language processing (NLP) [3], where the main goal is to automatically extract domain entities from text, which are specified by subject matter experts. This further helps the process of extracting relations between different entities and additionally eases the process of extracting events associated with the extracted entities.

Named-entity recognition (NER) is a task from IE, which deals with automatically detecting and identifying phrases from the text that represent domain entities. With regard to the pipeline according to which they are developed, there exist several types:

- Terminological- or dictionary-based [39] - where the extraction is related with a domain specific dictionary. This means that only the entities that are mentioned in the dictionary can be extracted. Additionally, different search heuristics are also applied in order to solve problems that appear with synonyms and different lexical forms of the same entity.
- Rule-based [18,26] - where apart from the dictionary, additional rules in form of regular expressions are developed. These rules define the characteristics of the entitites, which are also domain specific (e.g., in the chemical domain many of the entities names have known prefixes and suffixes).
- Corpus- or Machine Learning (ML)-based [5,23] - where the extraction is done using a trained machine learning model. It requires an annotated corpus, whose creation can be a time-consuming task where subject matter experts need to manually annotated the entities of interest in the textual data. Once the corpus is prepared, supervised ML methods are used to train a model.
- Active Learning (AL)-based [34] - where semi-supervised learning is applied. It starts with a small annotated corpus to train a NER model. Then, an unannotated corpus is used as input and the NER model is retrained. In these cases, the training process iteratively interacts with subject matter experts, who annotate sentences queried from the corpus.
- Deep Learning (DL)-based [24] - where deep neural networks are used to train a model. They typically require large amounts of annotated data.

Despite the existence of several types of NER methods, their application in different domains depends on the available resources (e.g., dictionaries, semantic data models, annotated corpus) in that specific domain. If we look at the biomedical domain, there is a huge amount of work done in this direction [8,16,36]. This would not be possible without the existence of diverse biomedical vocabularies and standards [9], which play a crucial role in understanding biomedical information, coupled with the collection of a large amount of biomedical data (e.g., drug, diseases and other treatments) from numerous sources and shared NLP workshops [6,21,32,37,38].

However, if we look at food and nutrition from the perspectives of agriculture, health and environment, it seems to be low-resourced. There are only a few food semantic data models (i.e. ontologies) that are developed only as a solution to a specific application problem [10]. Additionally, the links between them are not well explored, which causes interoperability difficulties in combining different food-related data sets that are described using them. As a consequence of the resource limitation, there are only a few NERs that can be used for food information extraction. A rule-based NER, known as drNER [14], is developed to extract information from evidence-based dietary recommendations. Within the entities of interest, drNER additionally extracts phrases that consist of food entities. It was further extended and improved by developing FoodIE [27], where the focus is only on food entities. FoodIE is also a rule-based NER, where the rules are combination of computational linguists properties with food semantic information from Hansard corpus [4]. Another way to extract food entities is to use the NCBO Annotator, which is a web service that annotates text by using relevant ontology concepts [20]. It is a part of the BioPortal software services [25], which means that in the food domain it can be combined with food-related ontologies such as FoodOn [17], OntoFood, and SNOMED-CT [11]. More details about a comprehensive comparison of four food NERs (i.e. FoodIE, NCBO (SNOMED CT), NCBO (OntoFood), and NCBO (FoodON)) are available in [30].

3 Food Data Normalization

Data normalization is a crucial task that allows interoperability between data sets that are described using different standards. By applying data normalization, we are mapping entities between different vocabularies, standards, or semantic resources.

In the food domain, there are several resources that can be used for food data normalization. Some of them are:

- FoodEx2 - a description and classification system, proposed by the European Food Safety Agency (EFSA) [12].
- FoodOn - provides semantics for food safety, food security, agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes [17].
- OntoFood - a nutrition ontology for diabetes.
- SNOMED CT - a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic health records [11]. It also consists of a *Food* concept.
- The Hansard corpus - a collection of text and concepts created as a part of the SAMUELS project [4,33]. It consists of 37 higher level semantic groups; one of them is *Food and Drink*.

There are also different food data normalization methods that are developed in a combination with some of the above-mentioned resources.

StandFood [13] is a semi-automatic system for classifying and describing foods according to FoodEx2. It is combination of machine learning and Natural Language Processing methods in order to address the lexical similarity in the food domain. Further, it was also extend in the context of semantic similarity, by applying graph-based embedding methods in order to find a unique representation for each food entity available in the FoodEx2 hierarchy [15].

The information extraction done with FoodIE is also related to the Hansard corpus, since FoodIE uses rules that include the Hasnard food semantic information. Using this information, the FoodBase corpus [31] was recently created, which is one of the first annotated corpora with food entities. It consists of 1000 recipes (curated version), and for each one all food entities that are mentioned are extracted and annotated using the Hansard food semantic tags. The recipe categories that are included are: Appetizers and snacks, Breakfast and Lunch, Dessert, Dinner, and Drinks. This version was manually checked by subject matter experts, so the false positive food entities were removed, while the false negative entities were manually added in the corpus. It additionally provides an uncurated version that consists of annotations for around 22,000 recipes, which were not manually checked by subject matter experts.

Additionally, the results presented in [30] show that the NCBO annotator in combination with food ontology can also be used for food data normalization process. For example, using it with FoodOn, the extracted food entities can also be annotated with the FoodOn semantic tags.

Beside different method that are developed and can be used for food data normalization to a specific food standard, the interoperability still remains an open question. To support an initiative similar to the UMLS in the food domain, a recently published study proposed FoodOntoMap [28]. The FoodOntoMap data set consists of food entites extracted from recipes and normalized to different food standards (i.e. Hansard corpus, FoodOn, OntoFood, and SNOMED CT). It also provides a link between the food ontologies, where to each food entity the semantic information from each resource is available. The FoodOntoMap is available in machine readable format, which supports and enables interoperability between computer systems.

4 FoodViz

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.

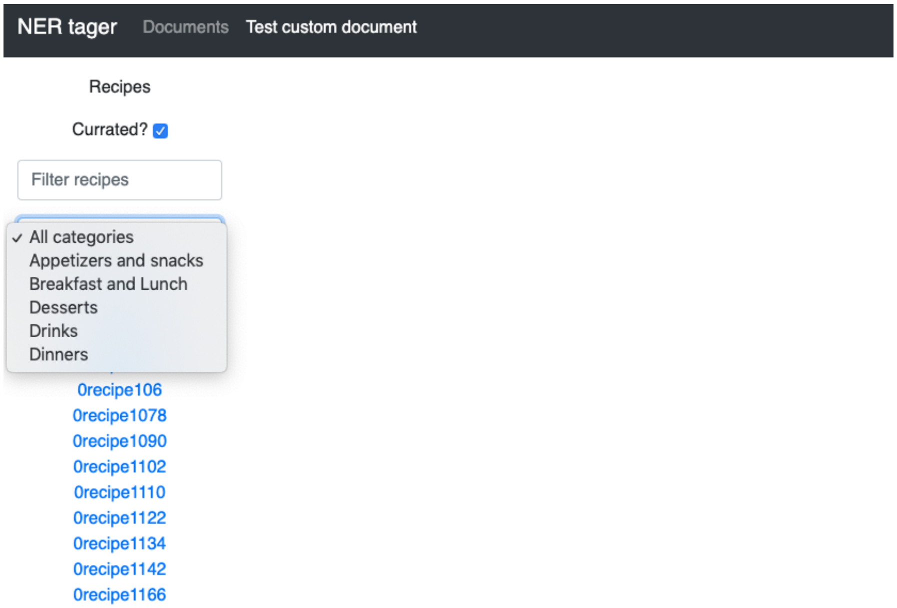

Fig. 1: FoodViz. A new visualization tool for presenting the results published in the FoodOntoMap resource to food subject matter experts.

The FoodViz[4] is a single page application developed with React[5], served by a back-end application programming interface (API) developed in Flask[6]. The back-end API serves pre-processed recipes annotated in our previous work [31] and the annotation mappings from [28].

The home page of FoodViz is presented in Figure 1. There exist three different parts 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 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 [31].

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 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)

4 http://foodviz.ds4food.ijs.si/fbw/#/recipes
5 https://reactjs.org/
6 https://flask.palletsprojects.com/en/1.1.x/

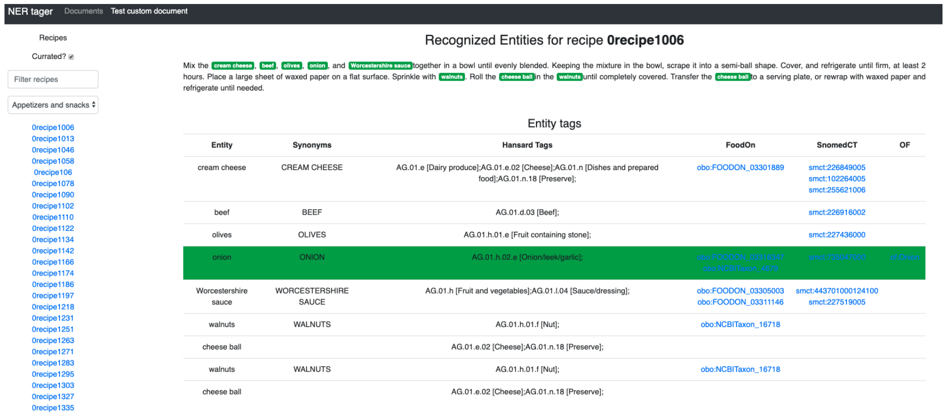

Fig. 2: FoodViz annotation for a recipe from curated corpus.

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 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 be also used as 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 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 with a semantic tag from another resource. With this, we also perform food ontology alignment. Additionally, let us assume that information about dietary intake 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,

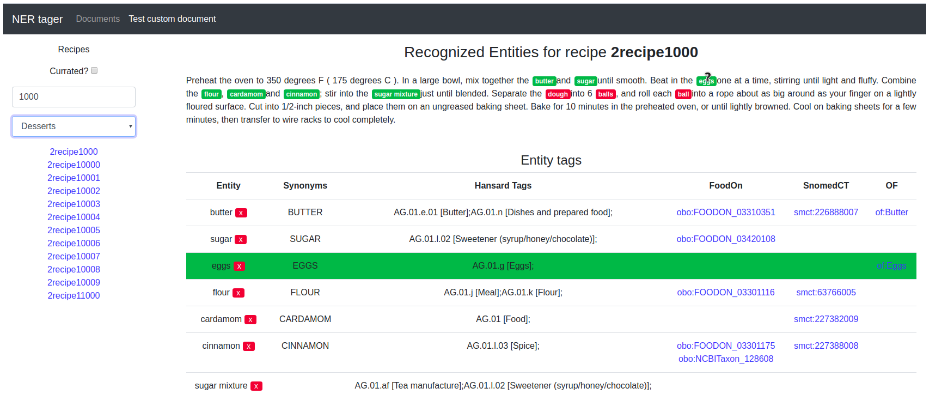

Fig. 3: FoodViz annotation for a recipe from uncurated corpus.

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.

5 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 [35]. 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 paper, 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 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 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 the 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 be 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 challenge for a dietitian and could be facilitated by the FoodViz tool.

Acknowledgements

This work was supported by projects from the Slovenian Research Agency [research core grant number P2-0098], and the European Union's Horizon 2020 research and innovation programme (FNS-Cloud, Food Nutrition Security) [grant agreement 863059].

The information and the views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use that may be made of the information contained herein.

References

1. Food in the anthropocene: the eat–lancet commission on healthy diets from sustainable food systems. The Lancet **393.10170**, 447–492 (2019)
2. Climate change and land: an ipcc special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems (2020), https://www.ipcc.ch/site/assets/uploads/2019/08/4.-SPM_Approved_Microsite_FINAL.pdf
3. Aggarwal, C.C., Zhai, C.: Mining text data. Springer Science & Business Media (2012)
4. Alexander, M., Anderson, J.: The hansard corpus, 1803-2003 (2012)
5. Alnazzawi, N., Thompson, P., Batista-Navarro, R., Ananiadou, S.: Using text mining techniques to extract phenotypic information from the phenochf corpus. BMC medical informatics and decision making **15**(2), 1 (2015)
6. Arighi, C.N., Lu, Z., Krallinger, M., Cohen, K.B., Wilbur, W.J., Valencia, A., Hirschman, L., Wu, C.H.: Overview of the biocreative iii workshop. BMC bioinformatics **12**(8), 1 (2011)
7. Aronson, A.R.: Metamap: Mapping text to the umls metathesaurus. Bethesda, MD: NLM, NIH, DHHS **1**, 26 (2006)
8. Boag, W., Wacome, K., Naumann, T., Rumshisky, A.: Cliner: A lightweight tool for clinical named entity recognition. AMIA Joint Summits on Clinical Research Informatics (poster) (2015)
9. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)

10. Boulos, M.N.K., Yassine, A., Shirmohammadi, S., Namahoot, C.S., Brückner, M.: Towards an "internet of food": Food ontologies for the internet of things. Future Internet **7**(4), 372–392 (2015)
11. Donnelly, K.: Snomed-ct: The advanced terminology and coding system for ehealth. Studies in health technology and informatics **121**, 279 (2006)
12. (EFSA), E.F.S.A.: The food classification and description system foodex 2 (revision 2). EFSA Supporting Publications **12**(5), 804E (2015)
13. Eftimov, T., Korošec, P., Koroušić Seljak, B.: Standfood: Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. Nutrients **9**(6), 542 (2017)
14. Eftimov, T., Koroušić Seljak, B., Korošec, P.: A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PloS One **12**(6), e0179488 (2017)
15. Eftimov, T., Popovski, G., Valenčič, E., Seljak, B.K.: Foodex2vec: New foods' representation for advanced food data analysis. Food and Chemical Toxicology **138**, 111169 (2020)
16. Gligic, L., Kormilitzin, A., Goldberg, P., Nevado-Holgado, A.: Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. arXiv preprint arXiv:1901.01592 (2019)
17. Griffiths, E.J., Dooley, D.M., Buttigieg, P.L., Hoehndorf, R., Brinkman, F.S., Hsiao, W.W.: Foodon: A global farm-to-fork food ontology. In: ICBO/BioCreative (2016)
18. Hanisch, D., Fundel, K., Mevissen, H.T., Zimmer, R., Fluck, J.: Prominer: rule-based protein and gene entity recognition. BMC bioinformatics **6**(1), S14 (2005)
19. Johan, F., Owen, G.: Exponential roadmap (2020), https://exponentialroadmap.org/wp-content/uploads/2019/09/Exponential-Roadmap-1.5-September-19-2019.pdf
20. Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M.A., Musen, M.: Ncbo annotator: semantic annotation of biomedical data. In: International Semantic Web Conference, Poster and Demo session. vol. 110 (2009)
21. Kim, S., Doğan, R.I., Chatr-Aryamontri, A., Chang, C.S., Oughtred, R., Rust, J., Batista-Navarro, R., Carter, J., Ananiadou, S., Matos, S., et al.: Biocreative v bioc track overview: collaborative biocurator assistant task for biogrid. Database **2016**, baw121 (2016)
22. Lartey, A.: End hunger, achieve food security and improved nutrition and promote sustainable agriculture. UN Chronicle **51**(4), 6–8 (2015)
23. Leaman, R., Wei, C.H., Zou, C., Lu, Z.: Mining patents with tmchem, gnormplus and an ensemble of open systems. In: Proce. The fifth BioCreative challenge evaluation workshop. pp. 140–146 (2015)
24. Lopez, M.M., Kalita, J.: Deep learning applied to nlp. arXiv preprint arXiv:1703.03091 (2017)
25. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research **37**(suppl_2), W170–W173 (2009)
26. Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D.: 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. pp. 426–433. Association for Computational Linguistics (2001)

27. Popovski, G., Kochev, S., Koroušić Seljak, B., Eftimov, T.: 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). pp. 915–922 (2019)
28. Popovski, G., Koroušić Seljak, B., Eftimov, T.: 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,. pp. 195–202. INSTICC, SciTePress (2019). https://doi.org/10.5220/0008353201950202
29. Popovski, G., Paudel, B., Eftimov, T., Seljak, B.K.: Exploring a standardized language for describing foods using embedding techniques. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 5172–5176. IEEE (2019)
30. Popovski, G., Seljak, B.K., Eftimov, T.: A survey of named-entity recognition methods for food information extraction. IEEE Access **8**, 31586–31594 (2020)
31. Popovski, G., Seljak, B.K., Eftimov, T.: FoodBase corpus: a new resource of annotated food entities. Database **2019** (11 2019). https://doi.org/10.1093/database/baz121, https://doi.org/10.1093/database/baz121, baz121
32. Rastegar-Mojarad, M., Liu, S., Wang, Y., Afzal, N., Wang, L., Shen, F., Fu, S., Liu, H.: Biocreative/ohnlp challenge 2018. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. pp. 575–575. ACM (2018)
33. Rayson, P., Archer, D., Piao, S., McEnery, A.: The ucrel semantic analysis system. (2004)
34. Settles, B.: Active learning literature survey. University of Wisconsin, Madison **52**(55-66), 11 (2010)
35. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2020), https://gluebenchmark.com
36. Wang, Q., Zhou, Y., Ruan, T., Gao, D., Xia, Y., He, P.: Incorporating dictionaries into deep neural networks for the chinese clinical named entity recognition. Journal of biomedical informatics p. 103133 (2019)
37. Wang, Q., Abdul, S.S., Almeida, L., Ananiadou, S., Balderas-Martínez, Y.I., Batista-Navarro, R., Campos, D., Chilton, L., Chou, H.J., Contreras, G., et al.: Overview of the interactive task in biocreative v. Database **2016**, baw119 (2016)
38. Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C., Lu, Z.: Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. Database **2016**, baw032 (2016)
39. Zhou, X., Zhang, X., Hu, X.: Maxmatcher: Biological concept extraction using approximate dictionary lookup. In: Pacific Rim International Conference on Artificial Intelligence. pp. 1145–1149. Springer (2006)