Ontology alignment using Named-Entity Recognition methods in the domain of food

Gorjan Popovski[1,2*] , **Tome Eftimov**[1] , **Dunja Mladenić**[1,2] and **Barbara Koroušić Seljak**[1,2]
[1]Jožef Stefan Institute, 1000 Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
{gorjan.popovski, tome.eftimov, dunja.mladenic, barbara.korousic}@ijs.si

Abstract

In recent years, a great amount of research has been done in predictive modeling in the domain of healthcare. Such research is facilitated by the existence of various biomedical vocabularies and standards which play a crucial role in understanding healthcare information. In addition, the Unified Medical Language System (UMLS) links together biomedical vocabularies to enable interoperability. However, in the food domain such resources are scarce. To address this issue, this paper explores a methodology for ontology alignment in the domain of food by leveraging Named-Entity-Recognition (NER) methods based on different semantic resources. It is based on a recently published rule-based NER method named FoodIE, whose semantic annotations are based on the Hansard corpus, as well as a NER tool called Wikifier, from which DBpedia URIs are extracted. To perform the alignment we use the FoodBase corpus, which consists of recipes annotated with food entities and includes a ground truth version which is additionally used for evaluation.

1 Introduction

Information Extraction (IE) is the 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) [Aggarwal and Zhai, 2012]. The main idea behind IE is to provide a structure to the information extracted from the unstructured data.

One of the core IE tasks is named-entity recognition (NER), which addresses the problem of identification and classification of predefined concepts [Nadeau and Sekine, 2007]. 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 NER methods exist: *terminology-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 [Zhou *et al.*, 2006], match text phrases against concept synonyms that exist in the terminological resources (dictionaries). 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 [Hanisch *et al.*, 2005] 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 depends on the domain.

Corpus-based NER methods [Alnazzawi *et al.*, 2015; Leaman *et al.*, 2015] are based on an annotated corpus provided by subject-matter experts as well as the use of ML techniques to predict the entities' labels. These methods are less affected by terminological resources and manually created rules. However, their limitation is their dependence on an 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 subject-matter experts to produce it.

To exploit unlabelled data in constructing NER methods, *AL* can be used [Settles, 2010; Tran *et al.*, 2017]. This represents semi-supervised learning in which an algorithm is able to interactively query the user to obtain the desired labels/outputs at new data points. Which examples are sent to the user for labelling is chosen by the algorithm and their number is often much lower than the number of examples required for supervised learning. It usually consists of three components: (1) the annotation interface, (2) the corpus-based NER, and (3) component for querying samples.

2 Related work

2.1 Hansard corpus

The Hansard corpus is a collection of text and concepts created as a part of the SAMUELS project [Alexander and Anderson, 2012; Rayson *et al.*, 2004]. It contains 37 higher level semantic groups, one of which is our topic of interest — *Food and Drink*.

[*] Contact Author

2.2 FoodIE

FoodIE is a rule-based food Named-Entity Recognition method [Popovski *et al.*, 2019a]. As it is rule-based, it consists of a rule-engine in which the rules are based on computational linguistics and semantic information that describe the food entities.

2.3 Wikifier

Wikifier is a tool that uses an efficient approach for annotating documents with relevant concepts from Wikipedia [Brank *et al.*, 2017]. It is based on a pagerank method to identify a set of relevant concepts. As it provides the location in the document where the annotation occurs, it is effectively a Named-Entity Recognition method. It provides Wikipedia concepts as annotations, additionally assigning DBpedia concepts if they exist.

3 Data

A recent publication provides one of the first annotated corpora, named FoodBase [Popovski *et al.*, 2019b], containing food entities. It consists of two version, a ground truth set referred to as "curated" (containing 1,000 annotated recipes), as well an "un-curated" version, consisting of around 22,000 recipes. The recipe categories that are included are: *Appetizers and snacks*, *Breakfast and Lunch*, *Dessert*, *Dinner*, and *Drinks*. In this paper, we use the *curated* version to perform the ontology alignment as well as evaluate the methodology. 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. An example of a recipe can be found on Figure 1.

4 Ontology alignment

Using FoodIE and the Wikifier tool, we obtain annotations for all 1,000 recipes from the FoodBase.

FoodIE extracts and annotates each recipe with semantic tags from the Hansard corpus. Each annotation contains the location of the extracted entity, i.e. where in the raw text the surface form representing the concept occurs, and its corresponding semantic tags from the Hansard corpus.

The Wikifier tool is used to annotate the recipes with DBpedia URIs. As these are general DBpedia concepts, additional information to filter out food concepts from non-food concepts is required. Webscraping the pages for the URIs provides useful information that can be used to distinguish food from non-food concepts, such as the broader concept/class to which the concept of interest belongs. The post-processing of the DBpedia URIs checks the entity type of the concept and checks if it is one of: "FOOD", "FOODS", "DISH", "INGREDIENT", "FOOD AND DRINK", "BEVERAGE", "PLANT", "ANIMAL", or "FUNGUS". If it does not belong to one of the above entity types, the page is checked for mentions of other URIs which are semantically related to food: "FOOD", "PLANT", "ANIMAL", or "FUNGUS". These URI mentions can occur anywhere in the page and if one of these matches is satisfied, the entity is assumed to be a food entity.

A post-processed example of such an annotation can be found on Figure 2.

Having annotated the recipes with both methods, we can perform the ontology alignment by using the location information for each annotation in each recipe. Each unique concept from both methods (semantic resources) is assigned its unique ID, and then a table is constructed for each concept mapping containing the IDs.

5 Evaluation and experimental setup

5.1 Match types

- True Positives (TP) — these are matches where the whole food concept is correctly annotated;
- False Positives (FP) — these are matches where a non-food concept is annotated as a food concept;
- False Negatives (FN) — these are matches where a food entity is not properly annotated;
- Partial match — these are matches where only some tokens from a food concepts are properly annotated.

5.2 Evaluation metrics

Using the concept of True Positives, False Positives and False Negatives, we compute the widely used evaluation metrics: Precision (P), Recall (R) and F1 Score (F1). They are defined as:

- $P = \frac{TP}{TP+FP}$
- $R = \frac{TP}{TP+FN}$
- $F1 = 2\frac{P \cdot R}{P+R}$

6 Results and discussion

After running the evaluation, we obtain the following results. The matches for both methods are presented in Table 1, while the evaluation metrics are presented in Table 2.

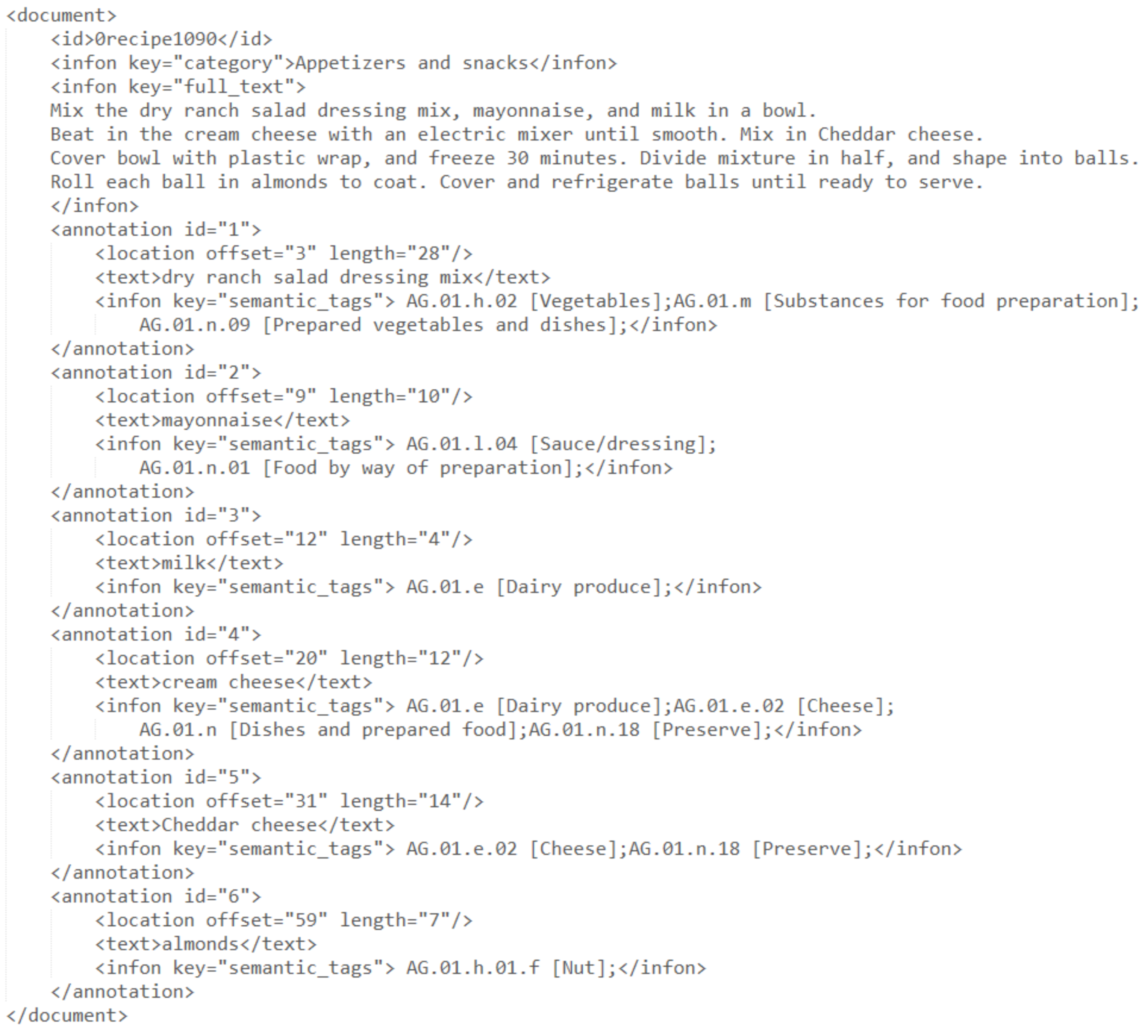
Table 1: Match types.

	FoodIE	Wikifier
TPs	11461	6380
FNs	684	4121
FPs	258	5861
Partial	359	3297

Table 2: Evaluation metrics.

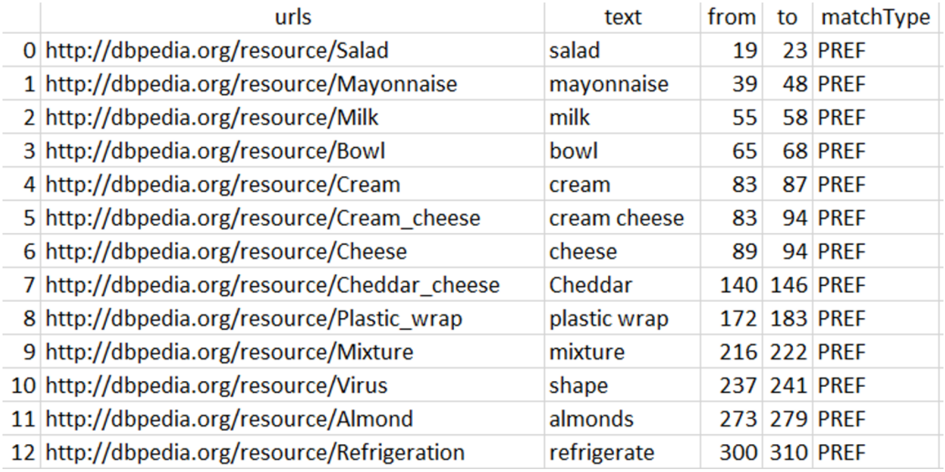
	FoodIE	Wikifier
F_1 Score	0.9605	0.5611
Precision	0.9780	0.5212
Recall	0.9437	0.6076

From the results in the tables it is evident that FoodIE provides more promising results. However, this was expected as this NER method was specifically constructed to only cater to the domain of food. Of especial interest are the matches of type *partial*, since they represent a match where only a subset of the tokens in a food entity are correctly recognized. For example, looking at Figure 1, the first extracted food entity



```
<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 1: Example recipe from the “curated” part of FoodBase.



	urls	text	from	to	matchType
0	http://dbpedia.org/resource/Salad	salad	19	23	PREF
1	http://dbpedia.org/resource/Mayonnaise	mayonnaise	39	48	PREF
2	http://dbpedia.org/resource/Milk	milk	55	58	PREF
3	http://dbpedia.org/resource/Bowl	bowl	65	68	PREF
4	http://dbpedia.org/resource/Cream	cream	83	87	PREF
5	http://dbpedia.org/resource/Cream_cheese	cream cheese	83	94	PREF
6	http://dbpedia.org/resource/Cheese	cheese	89	94	PREF
7	http://dbpedia.org/resource/Cheddar_cheese	Cheddar	140	146	PREF
8	http://dbpedia.org/resource/Plastic_wrap	plastic wrap	172	183	PREF
9	http://dbpedia.org/resource/Mixture	mixture	216	222	PREF
10	http://dbpedia.org/resource/Virus	shape	237	241	PREF
11	http://dbpedia.org/resource/Almond	almonds	273	279	PREF
12	http://dbpedia.org/resource/Refrigeration	refrigerate	300	310	PREF

Figure 2: Wikifier annotation example on a single recipe

should be “dry ranch salad dressing”, which is correctly extracted by FoodIE. Looking at Figure 2, the same food entity is only extracted as “salad”. Such match types do not factor in the calculation of the evaluation metrics, as it is debatable whether to count them as TPs or FNs. Nevertheless, they are interesting to compare, since even partial matches convey at least some semantic meaning regarding the food entity. Moreover, FP annotations on the same figure are “bowl” and “shape” which are not food entities. Additionally, a recent comparison of existing food NER methods can be found in [Popovski *et al.*, 2020], where the authors compare the performance of FoodIE with NER methods using other food ontologies available in the BioPortal.

Regarding the mapping of the concepts, a total of 348 explicit concept mappings were discovered by the methodology. An example mapping for the concept “garlic” would be:

- A000016: ‘garlic’, AG.01.h.02.e [Onion/leek/garlic].
- E000029: ‘garlic’, http://dbpedia.org/resource/Garlic

7 Conclusion and future work

In this work we propose a methodology for ontology alignment by using Named-Entity Recognition methods in the domain of food. It utilizes the newly proposed FoodIE NER method and the Wikifier text annotation tool. Our experimental results show that FoodIE provides more promising results than Wikifier, achieving an $F1$ score of 0.9605, compared to 0.5611. This is expected since FoodIE is specifically designed for the food domain, while Wikifier uses general vocabulary and annotates text with Wikipedia concepts.

For future work, recursive webscraping can be performed to more accurately distinguish between food and non-food annotated concepts from the Wikifier tool. Specifically, this would mean repeating the steps to check if the entity is a food entity or not on the parent nodes in DBpedia. Additionally, more food semantic resources can be included to provide mapping between multiple ontologies. Doing this is dependent on the existence of a NER method that works with concepts from the desired food semantic resource.

Acknowledgements

This research was supported by 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

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

[Alexander and Anderson, 2012] Marc Alexander and J Anderson. The hansard corpus, 1803-2003. 2012.

[Alnazzawi *et al.*, 2015] Noha Alnazzawi, Paul Thompson, Riza Batista-Navarro, and Sophia Ananiadou. Using text mining techniques to extract phenotypic information from the phenochf corpus. *BMC medical informatics and decision making*, 15(2):1, 2015.

[Brank *et al.*, 2017] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*, 2017.

[Hanisch *et al.*, 2005] Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6(1):S14, 2005.

[Leaman *et al.*, 2015] Robert Leaman, Chih-Hsuan Wei, Cherry Zou, and Zhiyong Lu. Mining patents with tmchem, gnormplus and an ensemble of open systems. In *Proce. The fifth BioCreative challenge evaluation workshop*, pages 140–146, 2015.

[Nadeau and Sekine, 2007] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[Popovski *et al.*, 2019a] Gorjan Popovski, Stefan Kochev, Barbara Koroušić Seljak, and Tome 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)*, pages 915–922, 2019.

[Popovski *et al.*, 2019b] Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. FoodBase corpus: a new resource of annotated food entities. *Database*, 2019, 11 2019. baz121.

[Popovski *et al.*, 2020] G. Popovski, B. K. Seljak, and T. Eftimov. A survey of named-entity recognition methods for food information extraction. *IEEE Access*, 8:31586–31594, 2020.

[Rayson *et al.*, 2004] Paul Rayson, Dawn Archer, Scott Piao, and AM McEnery. The ucrel semantic analysis system. 2004.

[Settles, 2010] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[Tran *et al.*, 2017] Van Cuong Tran, Ngoc Thanh Nguyen, Hamido Fujita, Dinh Tuyen Hoang, and Dosam Hwang. A combination of active learning and self-learning for named entity recognition on twitter using conditional random fields. *Knowledge-Based Systems*, 132:179–187, 2017.

[Zhou *et al.*, 2006] Xiaohua Zhou, Xiaodan Zhang, and Xiaohua Hu. Maxmatcher: Biological concept extraction using approximate dictionary lookup. In *Pacific Rim International Conference on Artificial Intelligence*, pages 1145–1149. Springer, 2006.