Toward Robust Food Ontology Mapping

Riste Stojanov
Faculty of Computer Science
and Engineering
Ss. Cyril and Methodius, University
Skopje, North Macedonia
riste.stojanov@finki.ukim.mk

Ilija Kocev
Faculty of Computer Science
and Engineering
Ss. Cyril and Methodius, University
Skopje, North Macedonia
ilija.kocev@students.finki.ukim.mk

Sasho Gramatikov
Faculty of Computer Science
and Engineering
Ss. Cyril and Methodius, University
Skopje, North Macedonia
sasho.gramatikov@finki.ukim.mk

Gorjan Popovski
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
gorjan.popovski@ijs.si

Barbara Koroušić Seljak
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
barbara.korousic@ijs.si

Tome Eftimov
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

***Abstract*—Data normalization methodologies are extremely welcome to link extracted information from textual data to different semantic resources. These methodologies have been previously well researched especially in the biomedical domain, where health concepts were normalized and described using semantic tags. Recently, a methodology for normalizing food concepts has been proposed, based on Named-Entity Recognition methods resulting in the FoodOntoMap semantic resource. In this paper, we propose and evaluate a new architecture for linking phrases (i.e. textual name for foods) to concepts from semantic resources in the Food and Nutrition domain. We represent the food phrases (i.e. their textual name) in continuous vector space using state-of-the-art Natural Language Processing (NLP) embedding algorithms, and evaluate their proximity with respect to the annotated semantic food concepts. Additionally, indexing was incorporated to improve efficiency.**

The GloVe embedding with mean pooling provided best evaluation results, with maximum recall of 74% for the Snomed CT semantic dataset, which is promising result, but also opens a space for future improvement of the phrase representations, and their incorporation in this system.

***Index Terms*—Natural Language Processing, Text representation, Embeddings, Data normalization and linking**

I. Introduction

In the Food and Nutrition domain, textual data contains a huge amount of information to be considered in order to foster the development of research in the domain. As textual data is unstructured, advanced methodologies are required for the automated extraction of information from this data.

To work with textual data, Natural Language Processing (NLP) techniques are extremely welcome. NLP is a subfield of artificial intelligence, where main tasks address analyzing and processing large amount of natural language data in order to make it understandable by computers. One crucial task here is how the textual data can be represented to be interpretable by the machines. Even though the ground work for *word* representation was made with Word2Vec [24] and Glove [27] embeddings, the transformer architecture [35] mainly represented with Elmo [29] and USE [12] in the early days provided more intelligent ways to represent *sentences*. Shortly after that, Google's BERT [13] found a way to contextually represent both words and sentences. From this point on, huge advancement has been made in multiple NLP tasks due to the use of these transformer based language models.

One task in NLP is Information Extraction (IE), where so-called Named Entity Recognition (NER) are developed, which automatically detect and identify text phrases that represent domain entities. Language models have enabled the development of state-of-the-art NER systems that are able to extract entities such as Organizations, Locations, Deceases, Drugs, Foods etc [5], [9], [16]. However, once this kind of information is extracted, there is another problem of normalizing it [25], even though its general category is known. Today, for most of these categories, there already exist detailed semantic resources, i.e. taxonomies and ontologies [1], [3], [10], [14], but still, the systematic knowledge is mainly used for rule-based matching based on synonyms, hypernyms and edit distance matching [18], [31], [32].

In this paper, we are investigating a new approach of linking phrases from the Food and Nutrition domain with their conceptual representations using the similarity of different phrase representations. We should note here that based on the semantic resource that is selected as a linking source, one food concept can simultaneously belong to more semantic tags (i.e. food concept mapping).

The remainder of the paper is organized as follows: in Section II, related work on semantic resources from the Food and Nutrition domain are described. In Section III, the methodology for food concept mapping is introduced, while in sections IV, results of the evaluation on the FoodOntoMap dataset are presented and discussed. The paper presents conclusions and future work in Section VI.

II. Related Work

In this section, first an overview of food and nutrition semantic resources is presented, followed by a summary of text representations using state-of-the-art embedding methods.

A. Semantic resources

SNOMED CT [3] is a growing and evolving collection of semantically organized medical terms structured as uniquely defined concepts with multiple descriptions and synonyms, associated between each other with relationships. It is a resource of scientifically validated clinical content providing a standardized way to represent clinical phrases captured by the clinicians worldwide. Therefore, it is broadly used in clinical documentation and reporting, enabling clinicians to record data with enhanced accuracy and consistency. Although the focus of SNOMED CT are clinical terms, it also contains a considerate amount of food-related data (e.g., data on food allergens) [4].

FoodOn [1] is a new ontology built to represent food-related entities and to provide vocabulary for nutrition, diet, and plant and animal agricultural rearing research. An ontology is a formal description of knowledge as a set of concepts within a domain and relationships that hold between them [8]. It interoperates with the Open Biological and Biomedical Ontology (OBO) Library [2], but also imports material from several ontologies covering anatomy, taxonomy, geography and cultural heritage. The ontology is aiming to cover gaps in the representation of food-related products and processes and is being applied to research and clinical datasets in academia and government.

BioPortal is a web tool that provides access to an open repository of biomedical ontologies [26]. Apart from browsing, searching and visualizing the ontologies, it also offers integrated search of biomedical data from different resources that can be further indexed and annotated using its variety of ontologies. Therefore, it is widely used by investigators, clinicians, and developers to access biomedical ontologies and to integrate data from a variety of biomedical resources, which, among other, contain considerate amount of food-related data.

The Hansard corpus [7], [33] is a collection of text annotated with concepts, created as part of the SAMUELS project (2014-2016). A defining feature of the Hansard corpus is the possibility to perform semantic searches on its data. It consists of 37 higher level semantic groups.

FoodOntoMap [31] is a resource that contains normalized food concepts extracted from recipes. The semantic information for each concept is linked between different food semantic resources, using a total of four food semantic resources: Hansard corpus, FoodOn, parts of SNOMED CT and OntoFood. OntoFood[1] is an ontology with SWRL rules of nutrition for diabetic patient.

B. Text representation with embedding vectors

Word embeddings are numerical representations of words or phrases used in NLP. Unlike humans, computers do not have the ability to capture the context of a word based on its text representation. Word embeddings emerge as a necessity to convert numerous words in a numerical form that will depend on their semantics and will be adequate for processing using neural networks. A word embedding is a mapping of a word into a real number vector in a vector space of reduced dimensions with the order of tens to few hundreds. In general, the values of the embedding of each word are assigned based on their usage and are obtained by using neural networks trained with a vast amount of text, capturing not only the contextual relationship of the words, but also their syntactic or grammar-based relationship. Similar words are represented with vectors that are close to each other in the vector space, i.e. they point to a similar direction in the predefined vector space.

There are many approaches for word embeddings generation using neural networks that aim to capture all linguistic and semantic aspects of the words to certain degree. The embeddings may be classified in the following manner[2]:

- **Word embeddings:** This class of embeddings represents a closed set of words in a continuous n-dimensional vector space. Word2vec [24] is one of the pioneer algorithms widely used for text representation. It is obtained from the single hidden layer of a shallow neural network, traind to predict a certain word in a given context window (referred to as Continuous Bag of Word - BOW) or trained to predict the context based on the word (referred to as Skip-gram). The GloVe algorithm [27] goes one step further by using the statistical occurrences of word pairs to capture their global context. Both approaches can provide pre-trained embeddings from a large corpus, however, they fail to capture different meanings of a single word in different positions of a sentence. They include all different meanings in the same embedding vector.
- **Character level embeddings:** Another way to create a language model that depends on the character co-occurrence only is to train a neural network and use its hidden layer as a representation for the characters [19].
- **Sub-word level embeddings:** The word embeddings have the open vocabulary shortcoming, i.e., if there is a word in the text that is not present in the embedding vocabulary, that word can not be represented in the continuous embedding space. This problem is first solved by fastText [11], where the text is split into smaller chunks, and a shallow neural network is trained to represent this sub-word elements. The same idea is also used in [13].
- **Pooled embeddings:** Once the words or sub-words are represented as vectors, finding a meaningful way to represent the sentences these sub-words belong to is a challenging task. One of the early ways of sentence representation is to use pooling, i.e., to select the minimum, maximum or the mean of the sentence word vectors.
- **Sentence embeddings:** are used when the meaning of the entire sentence needs to be encoded in order to

[1] https://bioportal.bioontology.org/ontologies/OF/?p=summary

[2] Note that the classes are not disjoint and there may be approaches that belong to multiple classes.

understand the context of the words. The representation of the meaning of a sentence is important because it enables understanding of its intention without individually calculating each word embedding. It also enables comparison of sentences, their clustering based on mutual similarity and predicting certain properties of the sentences, such as sentiment. Most common way of obtaining a sentence encoder is using the special $[CLS]$ token from the BERT model [23], [38]. Another example of the state-of-the-art algorithm for sentence embeddings is Sentence-BERT [34].

- **Contextualized embeddings:** The contextual embeddings, such as ELmo [28] and BERT [13], represent each word with different embeddings, depending on the context of use. This approach solves the problem of using same representation for the fruit ”Apple” and the company ”Apple”. However, the contextual embeddings can not be cached and used without computational cost, since they must be generated for each new context.
- **Transformer embeddings:** The transformer networks [35] opened a new era in the way we can obtain text representation. These networks [13], [20], [21], [37] are trained on large text corpora on various tasks related to text representation, and their last layers are usually used to represent the words in the sentence. These embeddings may represent words, sub-words (in the case of [13]) and usually capture the context of the word. They are also able to represent sentence as the value of the [CLS] token pre-pended in front of each sentence.

C. Food Concepts Normalization

In past few years, the question for food normalization became popular among the food and nutrition research communities, which refer it as food matching. StandFood [15] is one of the systems that tries to solve the food mapping problem using semi-automatic classification in order to describe the foods according their description. FoodOntoMap [31] is another system that provides concept alignment against multiple different databases.It is data normalization pipeline that is build upon FoodIE [30] and NCBO annotator [18]. Since FoodOntoMap provides food concept mapping to multiple semantic ontologies and taxonomies, we are using its dataset as a baseline for comparison of our work.

III. Methodology

Aiming to provide more robust solution for food concept mapping, we start from the hypothesis that the language representation of the phrases and its concepts (i.e. semantic tag) should be close in the space of their embeddings.

To validate this hypothesis, we need a method to project the phrases (i.e. textual name of the food) and concepts (i.e. semantic tags) in the same embedding space. In this paper, we use ”phrases” to represent a set of words that describe some thing (i.e. in our case food), while the word ”concept” is used to denote a member of some semantic resource (e.g., ontology or taxonomy). In Table I, we show a sample of the FoodOntoMap dataset, which we use as a baseline for the phrase mapping, consisting of phrases and their corresponding concepts, which may belong to one of the aforementioned semantic resources (Hansard, FoodOn, SNOMED CT, and OntoFood)[3]. The table also shows a sample of the Hansard Corpus as a representative of the semantic resources we use in our work. Each concept is defined with a unique concept id and a label that describes its meaning.

TABLE I
Datasets samples

FoodOntoMap phrases for Hansard Corpus

Phrase	Concept ID
	...
CREAM CHEESE	**AG.01.e.02**, AG.01.n, AG.01.e, AG.01.n.18
OLIVES	AG.01, AG.01.d.03
BEEF	AG.01.h.01.e
	...

Hansard Corpus

Concept ID	Label
...	
AG.01.e.01	Butter
AG.01.e.02	Cheese
AG.01.e.03	Fat/Oil
...	

By using the above-mentioned embedding algorithms to create embedding from multiple words, for both phrases and concept labels, we actually make projections of the phrases and concepts in the same vector space. Hence, our goal is to investigate if a phrase and its related concepts from any semantic resource are close in that vector space. We achieve this goal by choosing a phrase from the FoodOntoMap and searching the top most similar concept embeddings in the different semantic datasets. Then, we check the number of these resulting concepts that match the concepts used to describe that phrase and evaluate the precision, recall and F1 score.

IV. System architecture

The proposed methodology is used to create a system that enables searching of the most relevant food concepts for an arbitrary phrase. The concepts supported by our systems are the ones defined in Hansard, Corpus [7], FoodOn ontology [1] and Snomed CT taxonomy [3].

The datasources used in our work, the embedding algorithms as well as the flow of data in the mapping process are shown in Figure 1.

In order to provide a real-time search of arbitrary phrases, we have pre-computed embeddings for each concept from the above mentioned datasets using multiple different embedding algorithms, including GloVe word embeddings with mean pulling, BERT last layer embeddings with mean pooling, Bert sentence embeddings of last four layers, openAI GPT2 embeddings with mean pooling, XLNet embeddings with mean pooling, Distil Roberta and Albert sentence embeddings[4]. For implementation of the embedding algorithms we use the Flair library [6], which provides a convenient wrappers for Word, Sentence and Document embeddings around the Huggingface Transformers library [36].

[3] There is separate file in FoodOntoMap for each semantic dataset

[4] We tested a wider variety of embeddings, but for clarity, we report only a subset of these results. The full results can be found at: https://gitlab.com/ristes/relation-extraction/-/tree/master/data/evaluation

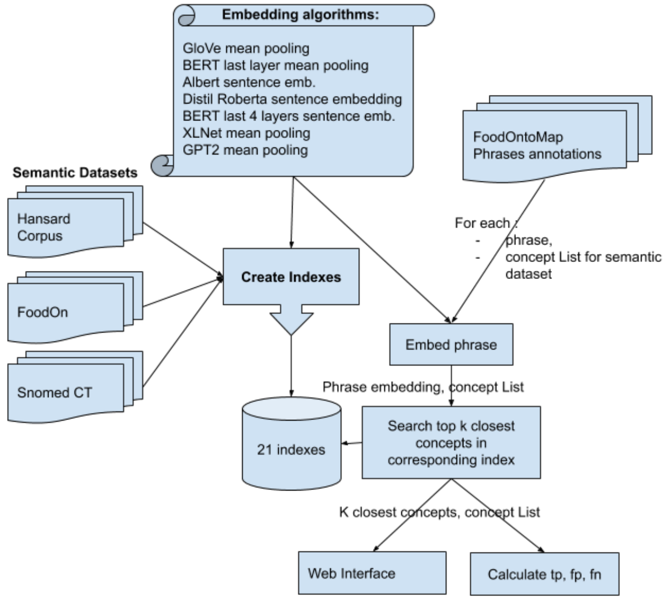

Embedding algorithms:
GloVe mean pooling
BERT last layer mean pooling
Albert sentence emb.
Distil Roberta sentence embedding
BERT last 4 layers sentence emb.
XLNet mean pooling
GPT2 mean pooling
Semantic Datasets
Hansard Corpus
FoodOn
Snomed CT
Create Indexes
21 indexes
FoodOntoMap Phrases annotations
For each :
- phrase,
- concept List for semantic dataset
Embed phrase
Phrase embedding, concept List
Search top k closest concepts in corresponding index
K closest concepts, concept List
Web Interface
Calculate tp, fp, fn

Fig. 1. System Architecture

Finding a similar embedding among a large set of embeddings is computationally expensive task considering the fact that each embedding has thousands of dimensions in the vector space. To accelerate this process, we use the Facebook's Faiss project [17] for creating one inverted index per dataset/embedding algorithm pair. We particularly use the *IndexIVFFlat* index, which stores each vector in an inverted file (IVF) as is, without any compression or quantisation (the *Flat* part) . The basic idea behind the concept is that the vectors are divided into clusters using k-means [22]. Each centroid, which is a central gravity point of all points of the cluster, is chosen as a representative of the vectors in that cluster. At search time, instead of calculating the similarity of a given vector to the entire dataset of vectors, its similarity to the k centroids is checked. Once the closest centroid is found, the similarity check is focused only on the reduced set of vectors within the cluster it represents.

We use *Inner Product* as metric for similarity during index creation and searching, but, before each embedding is stored in the index, it is normalized using L2 norm, which in combination with the Inner Product metric, results in Cosine Similarity [17].

In our work, we pre-compute 21 different indexes using different embedding algorithm for the same dataset. This number can be easily extended or reduced based on the embedding algorithms of interest. Using the indexes, we created a REST API which takes a search phrase and number k as an input and returns the k closest concepts to that phrase in Java Script Object Notation (JSON) format. Additionally, we provide a simple interface for visual representation and interpretation of the search results.

TABLE II
EVALUATION RESULTS

	FoodOn			**Hansard Corpus**			**Snomed CT**		
#res	**prec**	**rec**	**F1**	**prec**	**rec**	**F1**	**prec**	**rec**	**F1**
GloVe with mean pooling									
1	.75	.47	.58	.22	**.12**	.16	.80	.56	.66
5	.18	**.56**	.27	.07	**.21**	.11	.21	**.72**	.32
10	.10	**.60**	.17	.05	**.25**	.08	.11	**.74**	.18
20	.05	**.62**	.09	.03	**.33**	.05	.05	**.74**	.10
30	.03	**.63**	.06	.02	**.39**	.04	.04	**.74**	.07
Bert last layer with mean pooling									
1	.79	**.50**	.61	.01	.04	.02	.82	**.58**	.68
5	.16	.51	.25	.03	.02	.02	.17	.59	.26
10	.08	.52	.14	.01	.05	.01	.08	.59	.15
20	.04	.53	.08	.01	.06	.01	.04	.59	.08
30	.03	.53	.05	.00	.07	.01	.03	.59	.05
Alberta sentence embedding									
1	.79	.49	.61	.02	.01	.01	.81	.57	.67
5	.16	.50	.24	.01	.02	.01	.16	.58	.26
10	.08	.50	.14	.01	.04	.01	.08	.58	.14
20	.04	.51	.08	.00	.05	.01	.04	.58	.08
30	.03	.51	.05	.00	.06	.01	.03	.58	.05
Distil Roberta sentence embedding									
1	.07	.05	.06	.03	.02	.02	.18	.12	.15
5	.02	.06	.03	.01	.03	.02	.05	.17	.08
10	.01	.07	.02	.01	.04	.01	.03	.19	.05
20	.01	.08	.01	.00	.05	.01	.01	.20	.03
30	.00	.09	.01	.00	.07	.01	.01	.21	.02
GPT2 sentence embedding									
1	.00	.00	.00	.00	.00	.00	.03	.02	.02
5	.00	.00	.00	.00	.00	.00	.02	.05	.02
10	.00	.01	.00	.00	.01	.00	.01	.07	.02
20	.00	.00	.00	.00	.01	.00	.01	.10	.01
30	.00	.01	.00	.00	.00	.00	.01	.12	.01

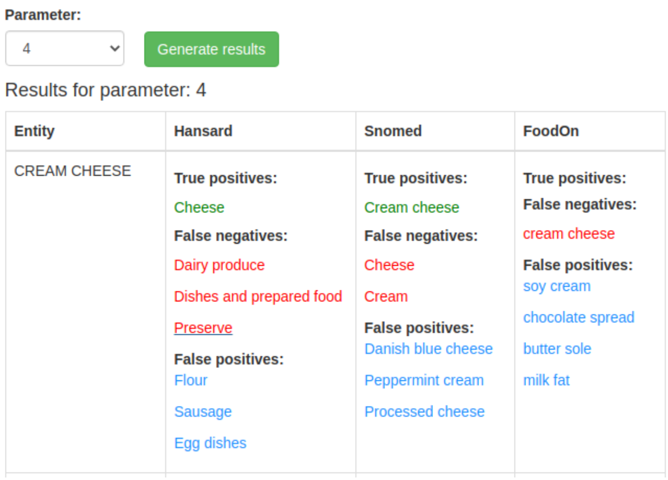
V. EVALUATION

In order to evaluate our methodology, we used FoodOntoMap dataset as ground truth for mapping phrases to their corresponding semantic concepts in each of the semantic datasets Hansard Corpus, FoodOn ontology and Snomed CT. Table I shows that in FoodOntoMap, for each phrase, there can be multiple mapped concepts ids, separated with comma. In our evaluation, we iterate over the FoodOntoMap's phrases and use its embedding to find the *n* most similar concepts in the previously generated semantic datasets' indexes. Once we obtain the results, we compare them with the FoodOntoMap's concept ids of the corresponding phrase, and we calculate the true positives (concept ids present in the results and in the gold standard), false positives (present in the results, but not in the gold standard) and the false negatives (present in the gold standard, but not in the results).

To make the results easier to observe, we created a web application [5] that displays the true positives, false positives and true negatives for each phrase in FoodOntoMap and a number of closest matching concepts k. An example of the results obtained for the phrase *CREAM CHEESE* with number of matching embeddings $k = 4$ is shown in Figure 2.

Once we obtain the number of true positives, false negatives, and false positives, we evaluate the precision, recall

[5] The web applications is available at: wp.finki.ukim.mk/food-concept-index/



Parameter:

4

Results for parameter: 4

Entity	Hansard	Snomed	FoodOn
CREAM CHEESE	**True positives:** Cheese **False negatives:** Dairy produce Dishes and prepared food Preserve **False positives:** Flour Sausage Egg dishes	**True positives:** Cream cheese **False negatives:** Cheese Cream **False positives:** Danish blue cheese Peppermint cream Processed cheese	**True positives:** **False negatives:** cream cheese **False positives:** soy cream chocolate spread butter sole milk fat

Fig. 2. Evaluation result interface

and F1 score in search scenarios with $n = 1..10, 15, 20$ and 30 closest concepts from each index[6] The evaluated results using the GloVe algorithm are shown in Table II. The complete list of the results for each algorithm and dataset can be found at: https://gitlab.com/ristes/relation-extraction/-/tree/master/data/evaluation.

During the evaluation, we came to the following observations:

- We obtain best results in terms of **recall** for the GloVe word embeddings with mean pooling for all 3 datasets, no matter how many results are obtained by the algorithm. We believe that this is due to the small number of words in each of the phrases, where there is very little contextual information. For instance, when 5 results are returned by the search, the recall of the GloVe embedding with mean pooling is 72%, 56% and 21% for the Snomed CT, FoodOn and Hansard concepts correspondingly.
- In terms of F1 score, the best results are by the GloVe embeddings with mean pooling. The BERT contextual embeddings of its last layer with mean pooling give better results only when one result is obtained by the search for the Snomed CT and FoodOn concepts.
- The sentence embeddings of BERT and Alberta return almost identical results, which is not surprising knowing that Alberta transformer is an extension of BERT.
- The sentence embeddings of XLNet, GPT2 and Distil Roberta give significantly worse results than the other embedding algorithms. These results may be expected for Distil Roberta, since it is designed to compress the representation of the Roberta transformer, but XLNet and GPT2 were used without any restrictions.
- One general observation is that all evaluated algorithms are failing to represent the multiword phrases close to the words of which they are composed. One obvious example is given in Figure 2, where *Cream* and *Cheese* concepts are not in the top 4 results for *CREAM CHEESE*.

VI. Conclusion

Although text is relatively easy to be understood by a humans, it requires advanced methodologies for its structuring as it is insufficient to represent its basic elements such as words or sentences only. It is far more important to be able to relate these basic elements with underlying phrases, thus preserving the semantic and syntactic information of the text. In this paper, we presented a new approach of linking phrases to their conceptual representations using the similarity of different phrase representations.

According to the evaluation, the best results were obtained using the GloVe embedding algorithm with mean pooling. Even though these embeddings have shortcomings, they can be really helpful in the process of classification of new text phrases with respect to concepts from ontologies and taxonomies (i.e. semantic tags). This approach is robust and does not depend on any hand-crafted rules that are domain dependant. Furthermore, this approach can leverage the improvements in the area of phrases representation, where the only thing that is needed to use new representation is to build an index of the concepts of interest using that embedding algorithm.

In our future work, we plan to improve the search algorithm to include the separate words in each concept, in order to better represent the multi word phrases. Also, we plan to include the concepts relationships in their representation.

Acknowledgment

This work is supported by the project FNS-Cloud, which is funded by the Horizon 2020 Framework Programme of the European Union under Grant Agreement no. 863059 (2019-2023). Another financial support that this project received was from the Slovenian Research Agency (research core funding No. P2-0098). The authors acknowledge the financial support from both funding parties.

This work was also supported by the NLP4Food (Use of NLP in the food domain) project funded by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje, N. Macedonia

References

[1] Foodon: A field to fork ontology. http://foodon.org. Accessed: 2019-10-07.

[2] Obo: The open biological and biomedical ontology. http://obofoundry.org. Accessed: 2019-10-07.

[3] Snomed clinical trials. https://www.snomed.org/snomed-ct/sct-worldwide. Accessed: 2019-10-07.

[4] Snomed ct - allergy to food. http://bioportal.bioontology.org/ontologies/SNOMEDCT ?p=classesconceptid=414285001. Accessed: 2019-10-07.

[5] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[6] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

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

[6] In the remaining of the text, for brevity we refer the number of the selected closest candidates as *number of search results*.

[8] Grigoris Antoniou and Frank Van Harmelen. *A semantic web primer*. MIT press, 2004.

[9] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*, 2019.

[10] Christian Bizer, Lehmann Jens, Kobilarov Georgi, Soren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7 (3):154–165, 2009.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[12] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[14] Tome Eftimov, Gordana Ispirova, Doris Potočnik, Nives Ogrinc, and Barbara Koroušić Seljak. Iso-food ontology: A formal representation of the knowledge within the domain of isotopes for food science. *Food chemistry*, 277:382–390, 2019.

[15] Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. Standfood: standardization of foods using a semi-automatic system for classifying and describing foods according to foodex2. *Nutrients*, 9(6):542, 2017.

[16] Yufan Jiang, Chi Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. Improved differentiable architecture search for language modeling and named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3585–3590, Hong Kong, China, November 2019. Association for Computational Linguistics.

[17] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[18] Clement Jonquet, Nigam Shah, Cherie Youn, Chris Callendar, Margaret-Anne Storey, and M Musen. Ncbo annotator: semantic annotation of biomedical data. In *International Semantic Web Conference, Poster and Demo session*, volume 110, 2009.

[19] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[20] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[22] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[23] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. *CoRR*, abs/1903.10561, 2019.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[25] A Miranda-Escalada, E Farré, and M Krallinger. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*, 2020.

[26] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173, 2009.

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[28] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[29] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[30] Gorjan Popovski, Stefan Kochev, Barbara Korousic-Seljak, and Tome Eftimov. Foodie: A rule-based named-entity recognition method for food information extraction. In *ICPRAM*, pages 915–922, 2019.

[31] Gorjan Popovski, Barbara Koroušić Seljak, and Tome 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,*, pages 195–202. INSTICC, SciTePress, 2019.

[32] Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. Foodbase corpus: a new resource of annotated food entities. *Database*, 2019, 2019.

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

[34] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 20193.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[37] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019.

[38] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.