Semantic Search of Bacteria Habitats on PubMed Abstracts via Ontology[*]

Rıza Özçelik, Selen Parlar

December 23, 2019

1 Introduction

Bacteria are a type of biological living-being that is present in most of the environments since the first life forms. The environments they can live in are called *bacteria habitats*. Bacteria and their habitats are investigated by researchers for ages since bacteria functions can have both positive effects on other species such as easing digestion or negative effects such as causing lethal diseases. These investigations resulted in a high volume of publications in PubMed, a website that store medicine-related publications. Therefore, PubMed stores a vast of valuable information in free-text. The unstructured form of the free-text data creates challenges in querying and processing the existing knowledge base since the same information can be expressed in infinitely many different ways in natural languages. On the other hand, successfully searching and retrieving the information in the literature is of vital importance for the researchers that would base their work on the previous studies. This creates the need for computational methods that can search the literature and retrieve the relevant publications. This can be achieved by semantically structuring the existing information and ontologies are a great structure to support search processes.

Ontologies are frequently used in computer science to semantically structure and normalize information in a computer-understandable format. Ontologies group together different concepts by also specifying their relations. In this sense, they are invaluable data representation schemes and they contain unambiguous, heterogeneous and semantically rich information that can be processed by computers. Thanks to ontologies, a computer program can know that basketball is a sport, documents have titles and abstracts, a human is a bacteria habitat and much more. In this work, we propose a search method that leverages ontologies to structurally represent the information in PubMed abstracts on the bacteria habitat domain and use this structure to retrieve relevant PubMed abstracts to a user query.

[*] https://github.com/rizaozcelik/semantic-search-using-ontobiotope

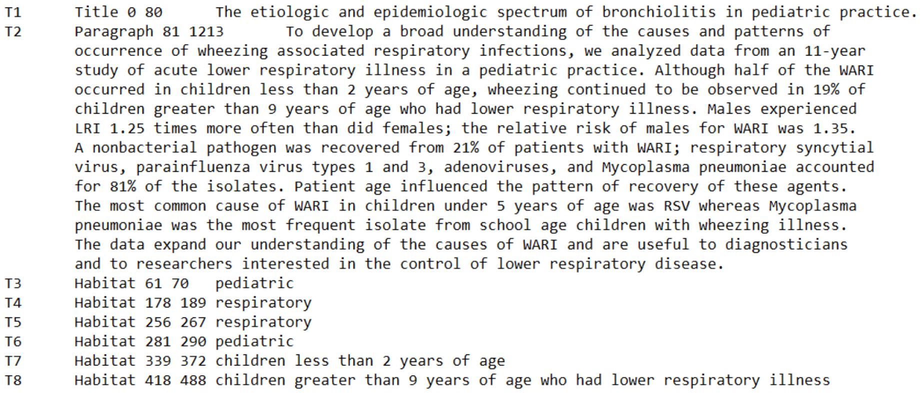
To do so, we learn a mapping from the bacteria habitat mentions in the PubMed abstracts to the classes in a bacteria habitats ontology: OntoBiotope. In parallel, we build an index that maps each class in OntoBiotope to the abstracts they are mentioned. When a query is provided, we first utilize the learned mapping to map the query to an OntoBiotope class and then retrieve the related documents of this class by index. Thanks to the semantic structure and hierarchy in OntoBiotope, we can infer the other related habitats and retrieve documents that mention them as well. For instance, the proposed search technique can retrieve abstracts related to *glial cells* given the query *pathogen in eyes*, thanks to the semantic structure. It maps the query to the habitat *peripheral nervous system* and retrieves documents that mention this habitat. Though the mapping is not exactly correct, the search algorithm can infer from the OntoBiotope that *glial cells* are also related to *peripheral nervous system* and retrieve the documents mention *glial cells*. Note that glial cells exist in the eyes.

This document is structured as follows: In the next section, we will provide details on the PubMed abstracts we used and OntoBiotope. In the upcoming section, we will discuss our two-staged methodology by referring the firs step as *entity normalization* and the second step as *semantic search*. We then discuss the ups and downs of the system and conclude.

2 Data Set

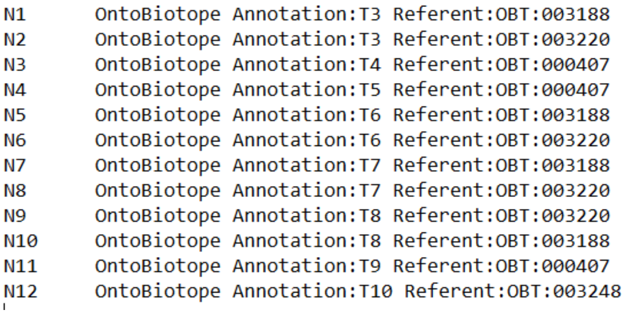
The dataset used in the work is provided by the BioNLP Shared Task 2019's Bacteria Biotope task. Though the task contains additional material, we focused only on the materials related to bacteria habitats. The data of our interest consist of PubMed titles, abstracts, and OntoBiotope.

PubMed abstracts are related to microorganisms whose habitats are mostly food products. The abstracts are manually annotated by the microbiologist experts. Figure 1 shows an example annotation. The direct mapping between entity mentions and bacteria habitats is also provided in annotation files. An example annotation file can be seen in Figure 2. On the other hand, OntoBiotope is a taxonomy of bacteria habitats that provide the semantic hierarchy of habitats. It contains 3602 bacteria habitats and 3984 relations. OntoBiotope is the main focus of this work.



```
T1      Title 0 80      The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice.
T2      Paragraph 81 1213       To develop a broad understanding of the causes and patterns of
        occurrence of wheezing associated respiratory infections, we analyzed data from an 11-year
        study of acute lower respiratory illness in a pediatric practice. Although half of the WARI
        occurred in children less than 2 years of age, wheezing continued to be observed in 19% of
        children greater than 9 years of age who had lower respiratory illness. Males experienced
        LRI 1.25 times more often than did females; the relative risk of males for WARI was 1.35.
        A nonbacterial pathogen was recovered from 21% of patients with WARI; respiratory syncytial
        virus, parainfluenza virus types 1 and 3, adenoviruses, and Mycoplasma pneumoniae accounted
        for 81% of the isolates. Patient age influenced the pattern of recovery of these agents.
        The most common cause of WARI in children under 5 years of age was RSV whereas Mycoplasma
        pneumoniae was the most frequent isolate from school age children with wheezing illness.
        The data expand our understanding of the causes of WARI and are useful to diagnosticians
        and to researchers interested in the control of lower respiratory disease.
T3      Habitat 61 70   pediatric
T4      Habitat 178 189 respiratory
T5      Habitat 256 267 respiratory
T6      Habitat 281 290 pediatric
T7      Habitat 339 372 children less than 2 years of age
T8      Habitat 418 488 children greater than 9 years of age who had lower respiratory illness
```

Figure 1: Annotated PubMed abstract

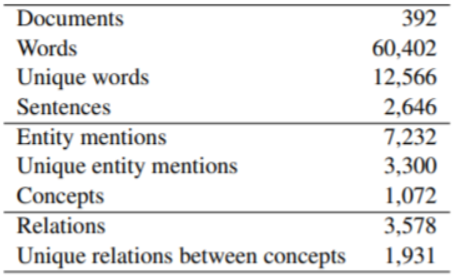


```
N1      OntoBiotope Annotation:T3 Referent:OBT:003188
N2      OntoBiotope Annotation:T3 Referent:OBT:003220
N3      OntoBiotope Annotation:T4 Referent:OBT:000407
N4      OntoBiotope Annotation:T5 Referent:OBT:000407
N5      OntoBiotope Annotation:T6 Referent:OBT:003188
N6      OntoBiotope Annotation:T6 Referent:OBT:003220
N7      OntoBiotope Annotation:T7 Referent:OBT:003188
N8      OntoBiotope Annotation:T7 Referent:OBT:003220
N9      OntoBiotope Annotation:T8 Referent:OBT:003220
N10     OntoBiotope Annotation:T8 Referent:OBT:003188
N11     OntoBiotope Annotation:T9 Referent:OBT:000407
N12     OntoBiotope Annotation:T10 Referent:OBT:003248
```

Figure 2: Mention - Bacteria Habitat mapping

2.1 Data Set Statistics

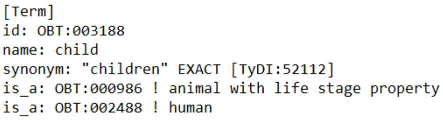
Since OntoBiotope is our main focus, we collected graph statistics from OntoBiotope. For the sake of completeness, we also provide the textual statistics computed in [1] in Figure 3.



Documents	392
Words	60,402
Unique words	12,566
Sentences	2,646
Entity mentions	7,232
Unique entity mentions	3,300
Concepts	1,072
Relations	3,578
Unique relations between concepts	1,931

Figure 3: PubMed abstract statistics [1]

Given the relational nature of the ontologies, we employ graphs in this project and represent OntoBiotope as a graph. The ontology itself has 3602 entities (bacteria habitats) and 3984 distinct *is-a* relation between these entities. For instance, Figure 4 shows an entity instance from OntoBiotope ontology, the name of the term is *child* and that term has two *is-a* relationships, i.e., child is a *human* and child is an *animal with life stage property.* We directly map these concepts to the graphs and represent each term with a node and each relation with an edge between two nodes. Once we represent ontology as a graph, we visualize it as in Figure 5 and then collect some graph statistics.



```
[Term]
id: OBT:003188
name: child
synonym: "children" EXACT [TyDI:52112]
is_a: OBT:000986 ! animal with life stage property
is_a: OBT:002488 ! human
```

Figure 4: Example entity from OntoBiotope

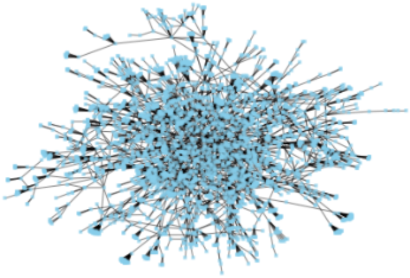


Figure 5: Visual representation of OntoBiotope

A clique is a subset of the vertices, such that every two distinct vertices are adjacent. In a taxonomy, we expect the graph to be a tree and maximum clique size to be 2. However, when we search the maximum clique in OntoBiotope, we find that it is a triangle. Figure 6 shows one of the cliques with the maximum size.

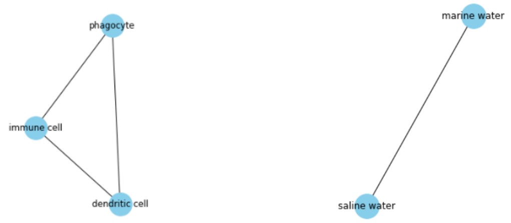


Figure 6: A maximum clique K_3 (on the left) and a K_2 in OntoBiotope.

We also measured the diameter and maximum distance from the root to other nodes. The diameter is the maximum distance between any vertex pair in the graph and is an important indicator of how wide the taxonomy is. On the other hand, the maximum distance from the root is an indicator of the depth of the graph and shows how depth the hierarchy is. Mentioned statistics are summarized in Table 1.

	# Nodes	# Edges	Diameter	Max Dist to Root	Max Clique Size
Original Graph	3602	3984	23	14	3

Table 1: Overall graph statistics for OntoBiotope

Apart from the overall graph statistics, we computed node statistics to gain insights on the importance of different habitats. We used centrality metrics as the importance indicator. We first computed *degree centrality*, which assigns a centrality score to a node by dividing its degree to the total degree of the graph. Thus, the nodes with more neighbors will have higher scores. For instance, *fermented cheese* has lots of subclasses and superclasses in OntoBiotope. This makes *fermented cheese* an important habitat, in terms of degree centrality. The distribution of degree centralities is shown in Figure 7. Note that the histogram is in a log scale for the sake of visual clarity. In the histogram, we can see that there is a high density of nodes with low degree centrality. In other words, OntoBiotope contains only a few high degree nodes.

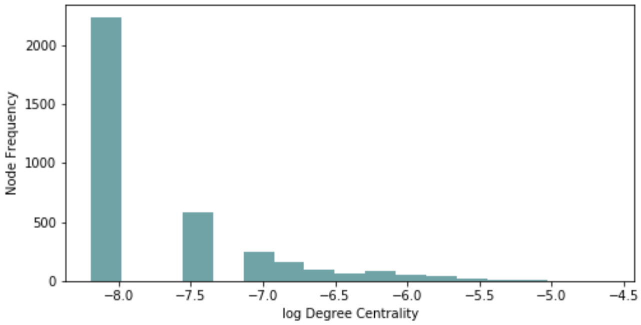


Figure 7: Degree centrality histogram

In addition to degree centrality, we measured *eigenvector centrality* that attributes an importance score to a node based on its neighbors' importance. In this sense, being adjacent to nodes with high eigenvector centrality is more valuable than being adjacent to low scoring ones. Since this is a recursive definition, eigenvector centrality is computed via an alternating method. The distribution of log eigenvector centrality scores is displayed in Figure 8. In this histogram we see that most of the nodes are populated in the middle of the histogram and tails are more sparse. This can be interpreted as OntoBiotope's containing lots of nodes with similar characteristics and certain small habitat groups have high importance.

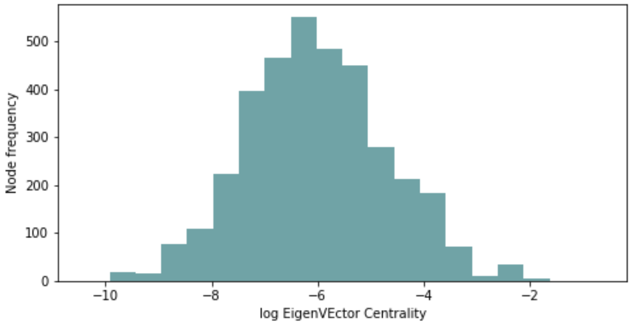


Figure 8: Eigenvector centrality frequency

2.2 Graph Database

We further performed some analysis on OntoBiotope by importing to a graph database, namely GraphDB. GraphDB implements RDF4J (Resource Description Framework for Java) framework interfaces, the W3C (World Wide Web

Consortium) SPARQL (SPARQL Protocol and RDF Query Language) Protocol specification, and supports all RDF (Resource Description Framework) serialization formats. It conducts semantic inference to enable the derivation of semantic results from existing knowledge in the database. Once we imported OntoBiotope to GraphDB, we are able to visualize the class hierarchy as shown in Figure 9. The bigger circles are the parent classes, and the nested ones are their children. Here we can observe that classes form clusters.

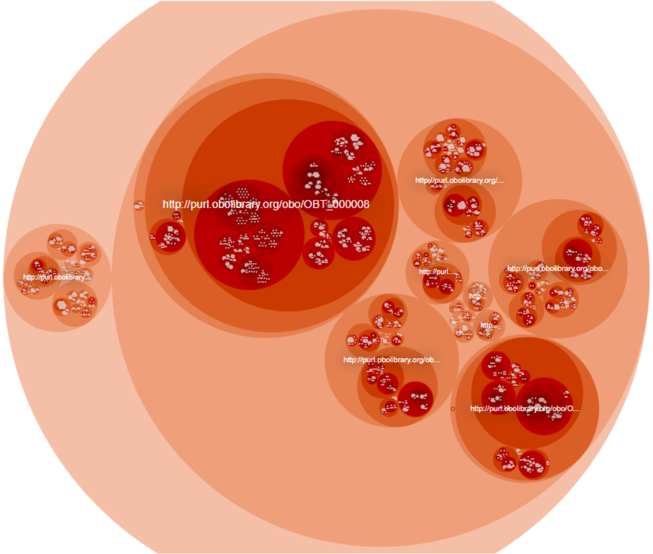

http://purl.obolibrary.org/obo/OBT_000008
http://purl.obolibrary.org/...
http://purl.obolibrary...
http://purl...
http://purl.obolibrary.org/obo...
http...
http...
http://purl.obolibrary.org/ob...
http://purl.obolibrary.org/obo/O...

Figure 9: Class Hierarchy

In order to enrich our database with the class relationships, we include 2 other ontologies that can be related to OntoBitope; The Environment Ontology and Wildlife Ontology . With this operation, we create a database of 15775 classes. The class hierarchy can be seen in Figure 10.

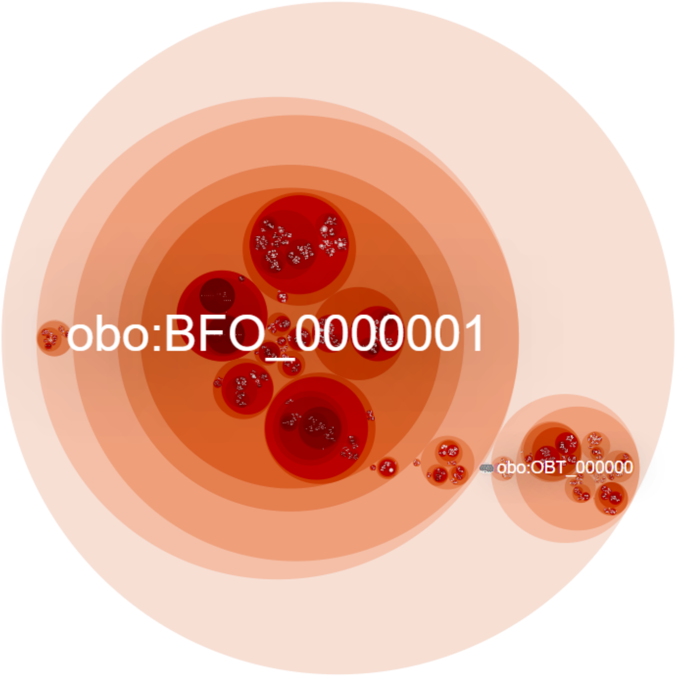


Figure 10: Combined Class Hierarchy

GraphDB did not display any class relationship when OntoBiotope is imported alone. However, when we add new ontologies into our database we get the following class relationship diagram shown in Figure 11. Each bundle shows links between the individual instances of two classes. Each link is an RDF statement where the subject is an instance of one class, the object is an instance of another class, and the link is the predicate.

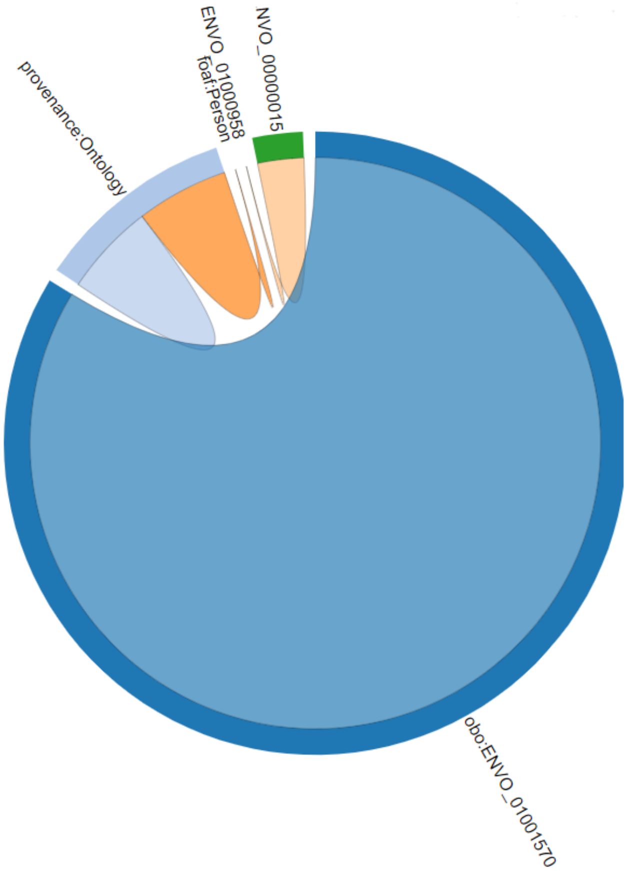


Figure 11: Class Relationship

After the visualization, we run simple queries on the database. For instance, we query the distinct entities and their super and subclasses and get the following result is shown in Table 2.2.

id	name	super class name	sub class name
obo:OBT_000003	animal habitat	root for extraction	apiary
obo:OBT_000003	animal habitat	root for extraction	breeding site
obo:OBT_000003	animal habitat	root for extraction	nest
obo:OBT_000003	animal habitat	root for extraction	rodent nest
obo:OBT_000003	animal habitat	root for extraction	livestock habitat
obo:OBT_000003	animal habitat	root for extraction	chicken coop

Table 2: Super class - Sub class

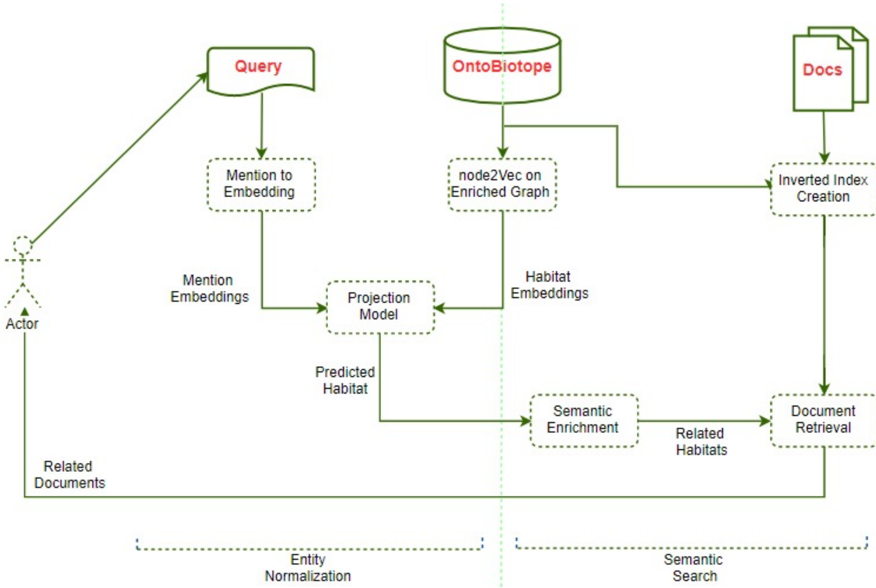


Figure 12: The search flow

3 Methodology

To search unstructured PubMed abstracts, we follow a two-step methodology. In the first step, our goal is to represent the abstracts in a more structured way. We use the annotated abstracts to serve our goal. We both learn a model to map free-text mentions to OntoBiotope classes and construct an index that maps OntoBiotope classes to the PubMed abstracts. The former lets us find the OntoBiotope class (i.e. bacteria habitat) that the query is relevant to, whereas the latter enables us to retrieve the abstracts that mention the relevant habitat. Note that the former one is called *entity normalization* since it maps differently expressed but semantically the same entities to the same ontology class, i.e. normalize them.

In the second step, we aim to leverage OntoBiotope to infer relevant habitats to enrich the search results. Given a user-defined query, we first map it to an OntoBiotope class by the mapping model to find the related bacteria habitat. Then thanks to OntoBiotope, we find related habitats to the found one and enrich the search result. We display the abstracts that mention the mapped and related habitats to the user. We name this step as *Semantic Search*, since it utilizes the semantic hierarchy in OntoBiotope. The flow of the methodology can be seen in 12. Now let us discuss each step in more detail.

3.1 Entity Normalization

As mentioned earlier, in this part we want to structure the information in the free-text PubMed abstracts to query them more easily. Here, there are two problems to solve. First, we need to find the habitat that the query is relevant to and then retrieve the related abstracts. Let us focus on our solution to the first problem now.

To tackle the first problem, we follow a supervised approach and use annotated abstracts to learn a projection from the mentions to the OntoBiotope classes. Thus, we need to represent both mentions and habitats in a vector space. To represent mentions, we use pre-trained word embeddings. We split mentions to their words and average their embeddings to create a mention embedding. To represent habitats, we leverage the graph structure of OntoBiotope and use a graph embedding algorithm named *node2vec* to learn a vector for each habitat.

Thanks to our graph analysis, we know that OntoBiotope is a sparse graph that contains a single information type: hierarchy. Thus, as the result of node2vec, only the nodes that are close to each other in the hierarchy would have similar vector representations. However, for the search purposes, we would be better of if we also retrieve documents that mention a habitat that co-occurs with the one in consideration. To this purpose, we add edges between habitats if they co-occur in an abstract. In other words, we semantically enrich OntoBiotope and run node2vec on this *enriched OntoBiotope* to obtain similar embeddings for the co-occurring habitats.

Thanks to pre-trained word embeddings that we use to vectorize mentions and running node2vec on enriched OntoBiotope to vectorize habitats, we obtain two different vector spaces. Now the problem becomes learning a projection between these two spaces, using mention-habitat annotations in the PubMed abstracts. Since the purpose of this project is not to find the best machine learning model to increase the accuracy but working on the social semantic concepts, we learn a simple projection matrix that projects a mention vector to its corresponding habitat vector. We use gradient descent to optimize the weights of the projection matrix and does not use any activation. Note that this problem can be framed as multi-class classification, but this would create too many classes with few samples. More importantly, the relations between the classes would not be captured due to the independence assumption, In the framework we proposed, we represent the semantic relations between habitats in the vector space as well, thanks to OntoBiotope.

Having learned the projection model, we can find related habitat to a given query by treating the query as a mention and mapping it to the habitat space by the projection model. This process would produce a vector in the habitat space and we can find the related habitat name by simply finding the most similar habitat vector, in terms of cosine similarity. Now the task becomes retrieving the documents that mention this habitat. To do so, we create an index from habitats to the PubMed abstracts that mention them. This is simply parsing the annotated documents and storing them in a dictionary that contains habitat ids as keys and corresponding documents as a list. Now that we know both the related habitat to the query and the documents that mention this habitat, we created structured information from the free-text abstracts. Now we can exploit the semantic structure in OntoBiotope to enrich the search results.

3.2 Semantic Search

OntoBiotope provides the semantic relations between the habitats. Thanks to OntoBiotope, we know that pediatric patients are human, throat is in the respiratory system and so on. Though we utilize this information by learning node embeddings for each habitat in OntoBiotope, we can use it further. To do so, we parameterize the search by adding a *maximum distance* parameter. Instead of retrieving the abstracts mention to the mapped habitat only, we retrieve abstracts that mention any habitat at a distance less than *maximum distance*. The distance computation is conducted both on initial and enriched OntoBiotope, to observe the effect of co-occurrence based enrichment. For the interpretability, the user is also shown why the abstract is retrieved as well.

With this structure, we can map an unstructured query to OntoBiotope and traverse OntoBiotope to infer other related habitats. Thus, the user is displayed with the results related to *throat* as well, when he/she queries *respiratory system*. Note that when the maximum distance is set to 0, only the results related to the mapped habitats are shown.

4 Results & Discussion

In this work, we proposed a semantic search technique that can conduct inference on OntoBiotope ontology to enrich the search results. To do so, we used entity normalization and mapped a given query to a habitat in OntoBiotope. To facilitate the search, we enriched OntoBiotope by adding edges between habitats that cooccur in an abstract. This operation caused significant changes in OntoBiotope. For instance, maximum clique size increased to 16, whereas diameter and maximum distance to root dropped to 18 and 11 respectively. Given that we introduced more edges, it is expected that the distances are shortened in the graph. The maximum clique and the induced subgraph of initial OntoBiotope that contains the same nodes can be seen in Figures 13 and 14. Notice that lots of edges are introduced between similar concepts. The summary statistics are shown in Table 3.

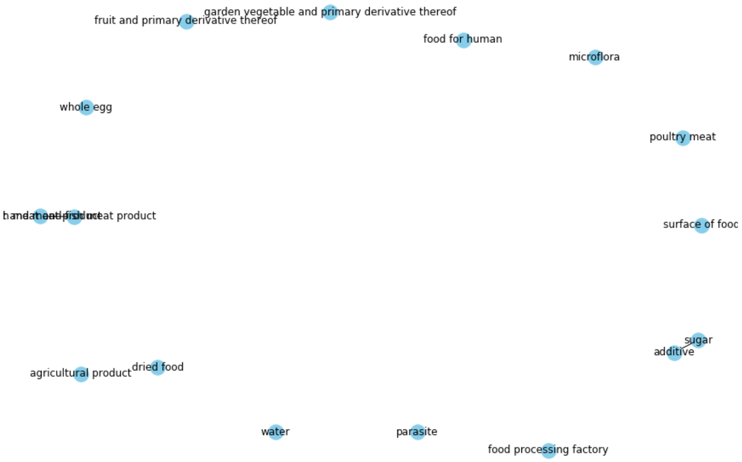


Figure 14: Induced subgraph of maximum clique nodes in the initial graph

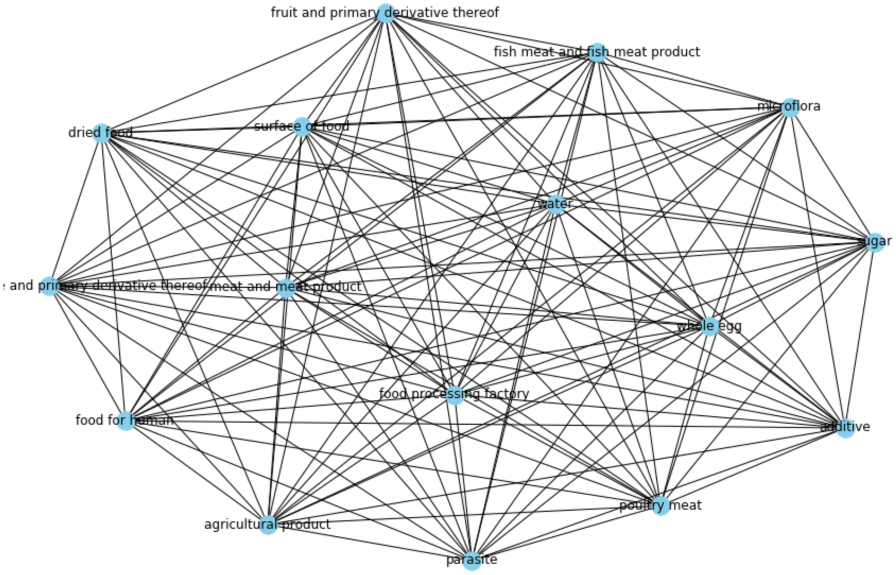


Figure 13: Maximum clique of enriched graph

Observing the significant changes in overall graph statistics, we performed node statistics analysis on enriched OntoBiotope as well. We computed degree and eigenvector centralities on OntoBiotope. From Figures 15 and 16, we can

	# Nodes	# Edges	Diameter	Max Dist to Root	Max Clique Size
Enriched Graph	3602	6115	18	11	16

Table 3: Overall graph statistics for enriched graph.

see that the number of nodes with high centrality increased. Thus, we can infer that certain nodes are mentioned more frequently than the rest.

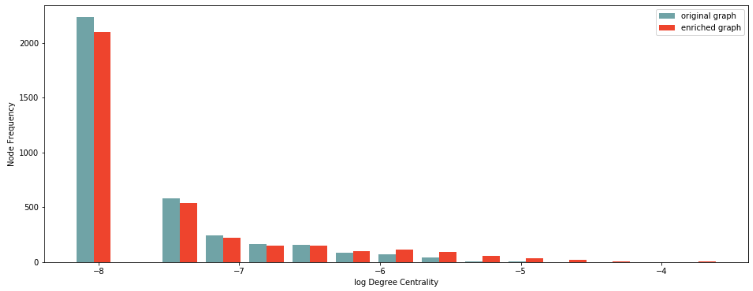


Figure 15: Degree centrality comparison between normal and enriched graphs

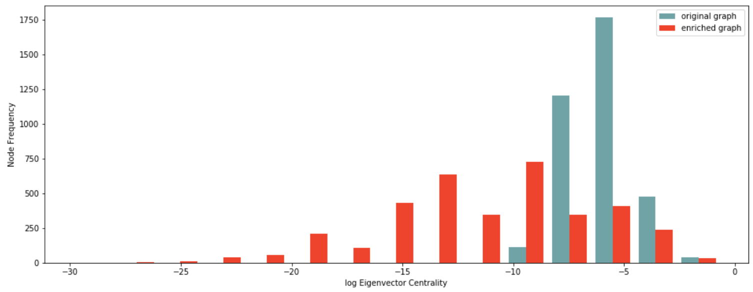


Figure 16: Eigenvector centrality comparison between raw and enriched OntoBiotope

Based on the comparative analysis of enriched and initial OntoBiotope, we run node2vec on the enriched version to learn habitat embeddings. Then we learned the projection from mention embeddings to the corresponding habitat embeddings. We used the train set provided by the organizing committee for the training and development set for testing.

To evaluate model performance, we measured how frequently the model maps a mention to correct habitat, i.e. accuracy. As the result, the training accuracy

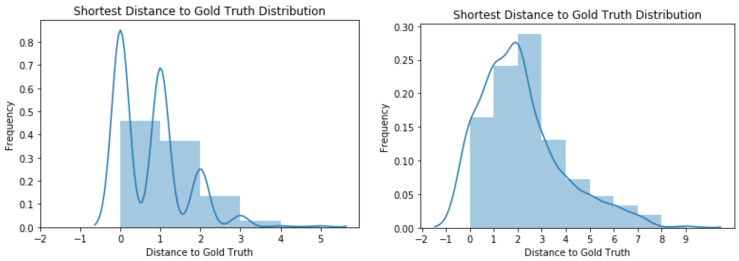


Figure 17: Shortest path distances of predicted habitats to the correct ones on the training (on the left) and test sets (on the right).

Query	Returned by Co-occurrence	Returned by Taxonomy
children with less than 5 years old	welfare center, medical sample, human pathogen, microflora, respiratory tract, nasopharynx, throat, child, healthy person, baby, hospital, infant, clinic, patient with infectious disease, pharynx, patient	welfare center, clinic, hospital
diseased cow	animal pathogen, gram-negative, phenotype wrt metabolic activity, plant, phytopathogen, pathogen, cell, plant part, animal	animal pathogen, pathogen
pathogen in eyes	peripheral nervous system, adult human, human, wound, glial cell, nerve	peripheral nervous system
child with respiratory illness	nasopharynx, medical sample, human pathogen, microflora, respiratory tract, child, healthy person, throat, welfare center, baby, hospital, infant, clinic, patient with infectious disease, pharynx, patient	nasopharynx, throat, pharynx
brain damage	bone fracture, drug resistant, head, pathogen, intensive care unit, blood, wound, central nervous system, patient, brain	bone fracture, wound
baby with cough fever and runny nose	welfare center, medical sample, human pathogen, microflora, respiratory tract, nasopharynx, throat, child, healthy person, baby, hospital, infant, clinic, patient with infectious disease, pharynx, patient	welfare center, clinic, hospital

Table 4: Example search queries and related habitats at maximum distance 2.

is measured as 0.46 and test accuracy is computed as 0.17. Considering that the habitats are not independent of each other, the failures should not be treated equally as well. For instance, predicting a habitat's parent instead of the habitat itself is tolerable more than the most distant habitat. Based on this derivation, we measured predictions' shortest path distances to the gold labels. This metric also shows how much the model can fail. The results can be seen in 17.

Both the accuracy and shortest path distance measurements indicate overfitting. To focus more on semantic search, instead of tuning the hyper-parameters of the node2vec and projection model to alleviate overfitting, we used the existing model for entity normalization before the semantic search.

During the search, we leveraged OntoBiotope to semantically extend the search results. Based on the shortest distance distributions, we can see that approximately 75% percent of the queries are matched with correct habitats during test time. This means that retrieving the documents related to habitats at distance at most 2 would retrieve a document set including the correct ones. Note that shortest path distance calculations can be performed either in initial OntoBiotope or enriched OntoBiotope. We display both results separately to the user for clarity. An example query list and returned habitats at maximum distance 2 are shown in Table 4.

5 Conclusion

In this work, we proposed utilizing an ontology to facilitate document search. More specifically, we used OntoBiotope, a bacteria habitat ontology, to retrieve related PubMed abstracts given a user query. The proposed system is two-staged. It first finds which habitat is relevant to the user query and then retrieves the relevant documents based on OntoBiotope. Thanks to OntoBiotope, we semantically structured PubMed abstracts by matching OntoBiotope habitats with the habitat mentions and enabled easier and more effective querying.

Our work is prominent in the sense that it brings together techniques from the social semantic web, information retrieval, and deep learning. Utilizing ontologies for semantic inference and automated annotation of free-text PubMed abstracts is an idea borrowed from the social semantic web domain, in which ontologies play a central role. We also utilized semantic relations between classes during projection model training. Constructing an inverted index for efficient search is frequently used in information retrieval, whereas we used graph embedding algorithms from the deep learning domain. In this sense, this study leverages approaches from various fields.

In terms of semantics, we utilized the semantic relations in OntoBiotope in three different aspects. Firstly, we used OntoBiotope to learn habitat embeddings with node2vec, which reflects the relations between habitats to the resulting habitat vectors. Secondly, we framed the entity normalization problem as a projection, rather than classification. Thanks to OntoBiotope, we were able to break the independence assumption between habitats and project mentions to habitat space. Lastly, we utilized OntoBiotope to group PubMed abstracts with respect to the habitats they mention. We created an index from the habitats in OntoBiotope to the documents that mention them. With the help of this index, we were able to retrieve related documents of a habitat with a dictionary lookup.

The proposed model can be improved different aspects. Firstly, using ontologies enables the merging of different types of data. During the analysis on GraphDB, we observed that different data sources can be integrated with OntoBiotope, to create a single and larger database. This means that one can use more and more data from different resources and linking them may result in more accurate search results, again from different resources. This would improve the semantic aspect of our work significantly.

Secondly, the performance of the projection model is quite important to map queries to related habitats. In this work, the projection model is deliberately chosen as a simple matrix and it was observed that the model overfits the training data. Thus, smarter model selection processes are needed to improve search performance.

A downside of utilizing ontologies is the approach's generalizability to different domains. Despite we focused only on bacteria habitats for this work, the need for search engines is prevalent in almost every domain. Yet, we lack comprehensive ontologies for most domains and they require expertise and time to build. This renders the generalizability of our approach questionable.

References

[1] R. Bossy, L. Deléger, E. Chaix, M. Ba, and C. Nédellec, “Bacteria biotope at bionlp open shared tasks 2019,” in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 2019, pp. 121–131.