2018 IEEE International Conference on Information Reuse and Integration for Data Science

Ontology-Based Supervised Concept Learning for the Biogeochemical Literature

Deya M. Banisakher[1], Maria E. Presa Reyes [1], Joshua D. Eisenberg[1]
Joshua Allen[2], Mark A. Finlayson[1], Rene Price[2], and Shu-Ching Chen[1]
[1] *School of Computing and Information Sciences*
[2] *Department of Earth and Environment & Southeast Environmental Research Program*
CREST Center for Aquatic Chemistry and the Environment
Florida International University
Miami, FL 33199
{dbani001,mpres029,jeise003,jalle091,markaf,pricer,chens}@fiu.edu

***Abstract*—Academic literature search is a vital step of every research project, especially in the face of the increasingly rapid growth of scientific knowledge. *Semantic* academic literature search is an approach to scientific article retrieval and ranking using concepts in an attempt to address well-known deficiencies of keyword-based search. The difficulty of semantic search, however, is that it requires significant knowledge engineering, often in the form of conceptual ontologies tailored to a particular scientific domain. It also requires non-trivial tuning, in the form of domain-specific term and concepts weights. As part of an ongoing project seeking to build a domain-specific semantic search system, we present an ontology-based supervised concept learning approach for the biogeochemical scientific literature. We first discuss the creation of a dataset of scientific articles in the biogeochemical domain annotated using the Environment Ontology (ENVO). Next we present a supervised machine learning classifier—a random decision forest—that uses a distinctive set of features to learn ENVO concepts and then label and index scientific articles at the sentence level. Finally, we evaluate our approach against two baseline methods, keyword-based and bag-of-words, achieving an overall performance of 0.76 F_1 measure, an improvement of approximately 50%.**

***Keywords*-Natural Language Processing; Semantic Search; Academic Search; Ontologies; Machine Learning**

I. Introduction

The first step of most scientific research projects is a review of the existing literature. This *academic literature search* allows a researcher to understand what hypotheses have been proposed, what methods or procedures have been tried or tested, and what results have been achieved. In most cases, indexing and retrieval of relevant articles is done using keywords [1]. Although simple and computationally inexpensive, keyword-based search has serious limitations considering the complexity of human language [1], [2]. Furthermore, as scientific knowledge grows exponentially larger, these limitations become more serious and serve to inhibit the ability of researchers to use existing tools to find relevant scientific literature [3].

A solution to this problem that has often been proposed is *semantic search*, that is, systems that can infer the meaning of a user's query and therefore retrieve articles of greater relevance [4]. Ontologies are a key component of this approach, as they provide a specific lists of terms and concepts as well as relationships between those items [5]. The challenge, however, lies in mapping articles and their constituent parts to the relevant parts of the ontologies [6].

Early work on ontology-based concept extraction used regular expressions or exact keywords matching [7], [8]. However, this requires encoding knowledge of all possible tokens that can map to specific ontology entities [9], a problematic task because of the ambiguity of language. Because of this, keyword approaches often miss essential concepts during the recognition and extraction steps. More recent work tackles the problem using matches driven by supervised machine learning (ML), which can automatically learn and judge which ontology concept is indicated by observed text.

The work presented here demonstrates the latter approach specifically for the biogeochemical domain. It is part of a larger domain-specific semantic search engine for the biogeochemical academic literature. In a prior report, we demonstrated the efficacy and feasibility of using ontological concepts to rank articles based on a search query [10]. In this paper, we demonstrate the development of a supervised machine learning (ML) approach that automatically learns ontological concepts, and labels sentences from biogeochemical articles with those concepts using features extracted from the unstructured text. We discuss the features necessary to build such systems and the process by which those features are extracted.

The remainder of this paper is organized as follows: We first review related work on ontology-based concept extraction (§II). Next, we describe our approach including the task definition, the ontology used, as well as the dataset created (§III). We then present and discuss the experiments performed as well as the results obrtained from those experiments (§IV). Finally, we conclude and specify our contributions (§V).

978-1-5386-2659-7/18/$31.00 ©2018 IEEE
DOI 10.1109/IRI.2018.00066

II. Related Work

An ontology provides formal and explicit specifications of conceptualizations, usually with a focus on a particular domain. Ontologies are one of the most recognized methodology of knowledge representation, providing definitions for a particular entities, relationships between entities, and classification of an entity on a class hierarchy. Ontology-based information extraction (OBIE) has been recently coined as a subfield of information extraction. In OBIE, ontologies play a crucial role in providing knowledge representation. The process is a core building block for the implementation of semantic search for large document repositories as well as the development of the Semantic Web [11], [12].

Ontologies have been useful for semantic data mining and search tasks. Ontology-based semantic data mining and search approaches and task include: association rule mining, classification, clustering, information extraction, recommendation systems, and link prediction for social networks [11]. Classification is a common task in data mining as well as other fields which aims at finding a model (or function) to describe and distinguish data classes or concepts [13]. Typical use of classification in ontology-based semantic search is the annotation of classification labels using entities and relations defined within the ontology. Setchi *et al.* [14] proposed a concept indexing algorithm that makes use of general-purpose ontologies. Although the paper uses a supervised approach, the ontology tagging process was done automatically instead of manually. Therefore, the accuracy of the tagged terms is only an approximate.

Some approaches to ontology-based classification of documents or topic modeling use the similarity of semantic graphs. The HITS algorithm [15] works over semantic graphs to identify core entities. Using DBpedia-based ontologies, Allahyari *et al.* [8] identified entities and their relations from test documents. By contrast, for this work, we focus on indexing ontology concepts at the sentence level, other approaches have indexed concepts at the word or the document level [12].

Most related to this work is Textpresso [7], a search engine which promises to enhance the retrieval of biological literature (as opposed to the biogeochemical here) by using an ontology-based approach. In Textpresso, multiple ontologies play essential roles in the retrieval of pertinent information from documents, resulting in significant acceleration of extraction of biological facts. The user can retrieve a set of documents by searching one or a combination of keywords. Ontologies make it possible to create semantic queries, facilitating the search the corpus of text by meaning instead of keyword-match. Textpresso achieves this by first identifying and matching the terms against pre-defined regular expressions.

Additionally, the creation and use of ontologies have been especially relevant in the biomedical domain where they were used for the identification of biological terms within raw text—such as scholarly publications and medical records [16]–[18]. The first step in the extraction of such terms is named entity recognition (NER), where the system can recognize and extract names of genes, drugs, chemical compounds, diseases, and so on. After these terms have been listed and formally defined via ontologies, the next step is defining the relationships between different entities (i.e., identify gene-gene or protein-protein interaction) [18].

III. Approach

The goal of the work presented here is to label the sentences of scientific articles—drawn from the biogeochemical academic literature—with concepts derived from a domain-specific ontology (specifically the *Environment Ontology*, or ENVO). We treated this as a supervised classification problem where we train a classifier using sentences that have been manually labeled (annotated) for their concepts; then, this classifier takes individual sentences found in a new article as input, outputting ontology concepts.

In this section we first describe the task and ENVO in detail, followed by the dataset which we created through manual annotation. Next we discuss the classification training process, starting with data preprocessing, followed by feature extraction, and ending with classifier construction.

A. Task Definition

As noted above, our task was to index academic articles in the biogeochemical domain with concepts derived from ENVO. That is, given a set of academic articles and our domain-specific ontology, the solution is a supervised classification model that can assign ontology concepts to the sentences found in the articles. We created a dataset of articles which was manually labeled and indexed with concepts from ENVO (§III-D discusses this in detail). Each sentence may have any number of concepts and therefore the labels are not mutually exclusive and our solution must admit a multi-label classification, including possibly no label. We identified a set of distinctive features to support this classification, and designed feature extractors to compute these features over article text.

B. The Environment Ontology

In prior work we determined that the most useful ontology for our purposes was the Environment Ontology (ENVO), a community-led, open ontology for various life science disciplines [19]. According to its creators, ENVO is an attempt at establishing a standard annotation scheme for several co-dependent or related disciplines, including, but not limited to, ecology, hydrology, environmental biology, and the geospatial sciences. ENVO contains concepts corresponding to a wide range of natural environments and environmental conditions. It is encoded in the Open Biomedical Ontologies (OBO) syntax, which is a subset of the Web Ontology

Language (OWL). ENVO can be populated, managed, and maintained using the OBO-Edit ontology development tool.

ENVO, like many ontologies, is hierarchical in design. Three of its top-level, most developed branches are *environmental system*, *environmental feature*, and *environmental material*. It's hierarchical structure allows for it to include not only entities, but also higher-level relationships between various concepts, including many standard ontological relationships such as `is-a`, `part-of`, `contained-in`, `connects`, and `has-condition`. ENVO also contains scientific and domain-specific relationships such as `derives-from`, `input-of`, `output-of`, `has-habitat`, and `biomechanically-related-to`. Furthermore, the ontology boasts a well-connected graph of synonymy relationships, encoded using different granularities including `broad`, `exact`, and `narrow`.

ENVO has seen quite a bit of success in adoption and use. It has served as the foundation for the creation and expansion of a number of other ontologies, as well as applied in several annotation projects such as the International Census of Marine Microbes (ICOMM) and the International Nucleotide Sequence Database Collaboration (INSDC) [20]. Additionally, ENVO has been used in data retrieval and query-based systems such as the Genomic Metadata for Infectious Agents Database (GEMINA) [21], while the National Institute for Allergy and Infectious Diseases Bioinformatics Resource Centers (NIAID BRCs) employ ENVO in metadata formulation and manipulation [22].

C. Dataset

To the best of our knowledge there is no corpus of scientific articles annotated with ENVO concepts, so we created our own. We collected a total of 14 articles (62,015 total words) using three search queries that were created by two domain experts (one of which is a co-author on this paper [JA]). Our domain experts ran the queries through Google Scholar and examind from the several hundred results returned, identifying the top four or five most relevant articles for each query. Importantly, several of the articles were not ranked near the top of Google's results, and were rather found many pages deep. We then manually annotated articles at the sentence level using concepts from ENVO (§III-D discusses the annotation study in detail). Table I lists the queries, the corresponding articles returned from the search results, as well as article-specific statistics. The articles have an average of 4,430 tokens, 172 sentences, 31 unique ENVO concepts. Table I presents detailed statistics on the test set.

D. Annotation Study

The purpose of manually annotating concepts from the ontology was twofold: first, to show that the ontological concepts appear in the target texts and, second, to show that it is possible to automatically learn domain-specific concepts from a relevant ontology. Because developing concept detectors is a non-trivial task, in prior work we tested the utility of the ontology, as well as verified that it is feasible to automatically rank articles using detected ontological concepts [10]. The current work expands that effort by creating a larger gold-standard corpus and demonstrating that we can identify the concepts in the articles automatically.

More than 20 years ago, Andren & Harriss (1973) measured relatively high % MeHg (MeHg as a percent of total Hg) in Everglades sediments, noting that samples from the Everglades were comparable to Hg-contaminated Mobile Bay sediments. [30, p. 328]

Text Span	Concept	ID
Everglades sediments	sediment	2007
Everglades	peat swamp	189
Mobile Bay sediments	sediment	2007

Figure 1. Example sentence from article [30, p. 328]. Underlined portions of the text indicate spans that were associated with an ENVO concept; the table shows the associated ENVO concept ID.

As discussed above, we collected a corpus of 14 articles from the biogeochemical domain, aligned with three search queries. Our team of domain trained annotators then annotated the queries and the articles for concepts from ENVO. For each article, annotations were collected at the sentence level.

Annotators used Protégé [37] to search and explore ENVO when deciding what concepts should marked for each sentence of each article. Annotators recorded their annotations in a spreadsheet, where each row represented a sentence, followed by columns representing the span of text containing the concept and the ID of the identified concept.

Figure 1 gives an example sentence from one of the test articles, along with the text spans which were associated with an ENVO concept.

The process of annotation involved several rounds of training, annotating, and revision of the annotation guidelines. Even for a relatively simple sentence as shown in Figure 1, numerous annotation decisions were needed. Below, we walk through this process phrase by phrase:

More than 20 years—This phrase does not need to be annotated, as it is a temporal expression referring to time period of the events mentioned later in the sentence.

...Andren & Harris (1973)—This phrase also does not need to be annotated, because it is a reference to a relevant article, and referring to the scientific literature isn't a concept in ENVO.

...measured relatively high %—This does not need to be annotated, as ENVO does not contain concepts related to specific chemical concentration levels.

...MeHg—This is the chemical formula for *methylmer-*

Query	Title	Citation	Tokens	Sentences	Unique Concepts	κ
Methyl-Mercury concentrations in Everglades water and sediment	Mercury in the Aquatic Environment ...	[23]	5,081	162	26	n/a
	Sulfide Controls on Mercury Speciation ...	[24]	4,133	168	13	n/a
	Sulfate Stimulation of Mercury Methylation ...	[25]	3,642	160	18	n/a
	Effect of Salinity on Mercury Activity ...	[26]	3,421	150	22	n/a
Sulfate reduction occurring in Everglades pore waters and sediments	Anaerobic Microflora of Everglades Sediments ...	[27]	4,651	179	35	0.64
	Constants for mercury binding ...	[28]	4,629	173	17	0.62
	Mercury methylation in periphyton ...	[29]	3,839	159	18	0.75
	Methylmercury Concentrations ...	[30]	4,295	183	26	0.30
	Bacterial Methylmercury Degradation ...	[31]	3,696	199	27	0.44
Sulfur reduction affecting South Florida Everglades soils	Groundwater's significance to changing ...	[32]	9,650	300	73	0.63
	Variation in Soil Phosphorus ...	[33]	3,032	103	39	0.71
	Sulfur in the South Florida ecosystem ...	[34]	3,485	149	37	0.69
	Sulfur in peat-forming systems ...	[35]	3,998	165	35	0.71
	Effects of sulfate amendments ...	[36]	4,463	160	42	0.62
		Max	9,650	300	73	0.75
		Average	4,430	172	31	0.61
		Min	3,032	103	13	0.30
		Standard Deviation	1,604	43	15	0.14

Table I
ARTICLES IN THE TEST SET. LISTED ARE THE NUMBER OF TOKENS IN EACH ARTICLE, THE NUMBER OF SENTENCES OVERALL, THE NUMBER OF UNIQUE CONCEPTS, AND THE ANNOTATOR AGREEMENT EXPRESSED AS COHEN'S κ.

cury, an environmental contaminant. The concepts of *contaminant* and *contamination* are not in ENVO. However, because this concept is relevant to the domain of interest, we did record these text spans and their related ideas so as to begin to build a set of concepts to expand ENVO in future work.

... (MeHg as a percent of total Hg)—Again, we identified the spans *MeHg* and *Hg* as the missing concept *contaminant*.

... in Everglades sediments—This phrase is tricky, because *Everglades* and *sediment* appear as individual concepts in ENVO, but when they appear in succession they form a multiword expression. *Everglades sediment* does not appear directly in ENVO. However, as it is presumably a subclass (or multiple subclasses) of sediments generally, we queried ENVO for the entity *sediment* (ENVO ID 2007), and examined its children for potential matches. *Sediment* has multiple children, namely, specific subtypes such as *lake sediment* or *contaminated sediment*. However, because there is no concept corresponding to the specific collection of different types of sediments that comprise the Everglades, we tagged this with the more general entity *sediment*.

... noting that samples from the Everglades—For this span, we first looked through ENVO to find a concept for *Everglades*. The closest concept is *peat swamp* (ENVO ID 189), which has no children, and so we tag this span using this concept.

... were comparable to Hg-contaminated Mobile Bay sediments.—For this span, we again tagged *Hg* as the missing concept *contaminant*. In the same way as above for *Everglades sediment*, the phrase *Mobile bay sediments* was tagged with the general concept *sediment*.

The first four articles were the result of a previous pilot annotation study [10]. The first three co-authors and a domain expert served as the annotators for those articles, and were annotated as follows: we annotated the first 50 sentences of one of the articles [23] cooperatively to develop the annotation guidelines, while each annotator annotated the remaining 130 sentences individually so as to allow us to calculate inter-rater reliability. This produced a Cohen's κ of 0.57, which is "moderate to substantial" agreement [38]. After these first articles was finished, we then assigned each of the annotators one of the four remaining articles for annotation [24]–[26], [30]. The remaining ten articles were doubly annotated by a new team of trained annotators and domain experts following the developed annotation guidelines. The resulting micro-averaged inter-annotator measure agreement over all annotator groups using Cohen's κ is 0.61 which is "substantial" agreement [38]. We also report per-document κ measures. We report an κ with zeroes columns and rows removed. This refers to the following situation: when analyzing the confusion matrix for a given concept, if there was a row or column that only contained the number 0, we removed it from the calculation of the average κ. We justify this because situations where there is a row or column consisting of only zeroes means that the annotators consistently marked a certain concept as two different things. An example of this is an annotator consistently marking a set of spans as the concept *watercourse*, and the other annotator consistently marking the same span as *watershed*, which are two similar concepts. They were marking the same span as different concepts, and each annotator always made the same decisions, but the problem was with what they called

the concept. They were consistent, which is qualitatively represented by the fact that there is a column or row in the confusion matrix of all 0's. Due to the consistency of the mislabeling, we can justify removing their κ's from the calculation of the average κ.

E. Data Preprocessing

In addition to annotating the data with ENVO concepts as described in the previous section, we performed standard NLP preprocessing tasks to prepare the data for feature extraction and supervised learning. First, we encoded document structure and formatting information such as section and paragraph headers, as well as sentence counts and relative positions of sentences within sections. Next, we cleaned the text by removing in-text citations and standalone mathematical, chemical, and biological formulas. We then tagged each token with its part-of-speech [39], lemmatized tokens using WordNet [40], filtered known stop words using PubMed's list [41], and used the pywsd module to perform word-sense disambiguation [42] to tag words with WordNet senses.

F. Data Balancing

The articles included 192 unique concepts across 3,434 occurrences. More than half of these occurrences (2,049) represented only 10 concepts, while the most frequent 50 concepts (26% of the total) occurred 3,091 times in total. Additionally, 61 concepts (32%) appeared only once. When supervised ML is performed over such distributions, they tend to overfit the classes with higher number of examples. Several solutions have been proposed and used for the problem of imbalanced data such as sampling (undersampling and oversampling) and weight assignment. These techniques are used to help supervised ML classifiers learn more about a class that has a significantly smaller number of examples relative to others. In our case we opted to use the Synthetic Minority Over-sampling Technique (SMOTE) [43]. SMOTE is an hybrid sampling technique that oversamples the minority classes while undersampling the majority classes. We applied resampling to the training set only, leaving the testing set with the original distribution.

G. Feature Extraction

Identifying a useful set of features is integral for an accurate machine learning model. For this task we extracted lexical, syntactic, and semantic features from the articles and their sentences. For lexical features, we used the most frequent distinctive terms for each article using *term frequency-inverse document frequency* (tf-idf) [44]. We used the top 10% of the resulting lists. Additionally, we used global and local sentence positions as features—i.e., the relative position of a sentence in both its section and article, expressed as a real number between 0 and 1, inclusive. Further, we extracted named entities from each sentence by examining parts-of-speech (looking for runs of tokens tagged NNP or NNPS), and used these entities as features. As discussed earlier, recognizing named entities is useful for many IR and NLP tasks. An example of this from our study is the term *Everglades* which is found encoded in ENVO as a synonym and part definition for *peat swamp*.

Finally, for semantic features, we mapped the words in each sentence to a semantic embedding space. As an example of an embedding approach, word2vec [45] is a popular and powerful method to represent high-dimensional word embeddings which reduce the complexity and size of the feature set as opposed to a bag of words (BoW) approach. However, word2vec does not consider words that have multiple senses, mapping them to the same position in the vector space. To address this limitation, we used sense2vec [46], where different senses of the same word are placed differently in the embedding space. We used Sense2vec as implemented in the SpaCy python module [47], and followed the algorithm described in [46] by using the part-of-speech tags and named entity labels assigned to the tokens. Additionally, we merged named entities into single tokens (using hash symbols), so that they were assigned a single vector.

In addition features extracted directly from the raw text, we also used other concepts as features. First, we used concepts identified in the abstract of each article as features for the body of the article. Second, we used the concepts present in a the immediately preceding sentence as features for determining the next sentence's concepts. This feature engineering led to several interesting observations; first that concepts found in the abstract of an article can improve concept labeling performance for the article body; and further, that knowing which concepts came before a sentence (i.e., in sentences preceding the sentence in question) also improves concept labeling performance.

H. Concept Learning

The first stage of classification is model training, followed by a stage of testing on separate (unseen) data. As discussed in §III-C, the original data was randomly split into two portions ten different times (ten folds), 80% in the training set and 20% in testing set (11 and 3 articles, respectively). We built and trained our models using random decision forest models (RDFs). RDFs are ensemble learning methods and are employed in regression and classification applications [48]. They operate through the construction of numerous decision trees during the training stage. The technique outputs the class that contains the mode of the classes of the collection of collection of tress. This technique is influential, especially in data mining applications [49]. A major advantage of RDF over regular decision trees is that the RDF avoids overfitting the training set [50].

We built and trained two separate models using the features discussed in the previous section—a *body-only* model,

which used all features, and an *abstract-only* model, which omitted the abstract concept features as well as the sentence counts and position features. This two-model approach attempts to mimic how human read scientific articles, namely, using the concepts found in the abstract to better guide the understanding concepts found in the rest of the text.

With regard to the parameters of the RDF classifiers, *max_features* was set to the square root of the total number of features in an individual run, *number_of_trees* was arbitrarily set to 50, where this is referring to the number of trees built before taking the average tree votes for predictions. Finally, *min_sample_leaf* was set to 50. To implement these models we used the scikit-learn python ensemble module [51].

IV. Experiments and Results

As discussed above, we randomly split the dataset into training and testing sets across ten folds, resulting in 11 articles for training and 3 for testing in each fold. Our models learned a total of 192 unique concepts. For all experiments, we evaluated the performance of the models on each concept using the F_1 measure averaged across all folds. Here we present our evaluation methods and results, describing our baseline approaches, as well as the performance of both the baselines and our method average averaged across the test sets.

A. Baseline Methods

We compared our approach to two baseline methods. The first baseline was a keyword-based approach, where we matched sentence words directly to the names of ontology concepts. All previously mentioned preprocessing steps were performed on both the text and the ontology, such as lemmatization of both concepts and words in the sentences. This model needed no training. The second baseline was a Bag of Words (BoW) supervised classifier. For this classifier, we trained and tested a support vector machine (SVM) [52] following the same cross-validation splits and multi-label fashion as used for our proposed approach. The SVM classifier was trained using the RBF kernel function and a soft margin C of 10,000—a common setup.

B. Experiments

As noted above we built two models: (1) an *abstract-only* model, and (2) a *body-only* model. Both the models learn concepts using all sentences in the text (including the abstract), but as the names suggest, they only used to label the abstract sentences and the body sentences respectively. Additionally, the the *body-only* model uses the labels produced by the *abstract-only* model as features for labeling the body of an article. In order to compare the efficacy of using the sense2vec approach as a feature, we built trained and tested the same models using a word2vec approach instead.

	Unique Concepts		
Approach	Single	Top 50	All
Keyword Search	0.39	0.35	0.38
SVM_BoW	0.45	0.56	0.50
RDF_word2Vec	0.54	0.69	0.61
RDF_sense2Vec	0.67	0.78	**0.76**

Table II
Average F_1 scores per approach over all concepts, the 50 most frequent concepts, and the 61 least frequent concepts with single occurrences.

Features		Unique Concepts		
Abstract	Position	Single	Top 50	All
Omitted	Omitted	0.52	0.69	0.65
Omitted	Included	0.63	0.71	0.68
Included	Omitted	0.63	0.76	0.70
Included	Included	0.67	0.78	**0.76**

Table III
Average F_1 scores for the feature combination experiments. The first two columns indicate whether the abstract concepts and sentence positions were included or omitted as features in the models.

Table II shows three average F_1 scores over different sets of concepts for all discussed approaches. The first column shows the average F_1 score for the concepts with single occurrence in the original data (61 concepts), while the second column shows the average scores for the top 50 concepts in terms of total occurrences over all the articles. The last column shows the results over all concepts. The proposed approach (RDF_sense2vec) outperforms both baselines as well as the RDF_word2vec models across all concepts. Additionally, Figure 2 shows the frequency of the top 50 ENVO concepts as well as the average F_1 score of each for each of the concepts. As shown, the score drops with the frequency of the concept in the dataset, although not dramatically. This is expected as it is a result of the original class imbalance. Finally, the *abstract-only* model performed similarly well with a 0.69 F_1 over all concepts present in the abstract sections, which were relatively small in number.

As discussed in §III-G, we proposed that the model's performance would improve when (1) abstract concepts are used as features for the body concept extraction, and (2) sentence positions are also included as features (i.e., sentence positions relative to the article as a whole and individual sections). To evaluate this, we performed four experiments, testing the inclusion of abstract concepts and sentence position features. Table III shows three average F_1 scores over different sets of concepts per experiment.

In the last row in Table III, both abstract concepts and sentence positions were included as features in the models. The results confirm our proposal, in that the inclusion of both of those features yields better labeling results across

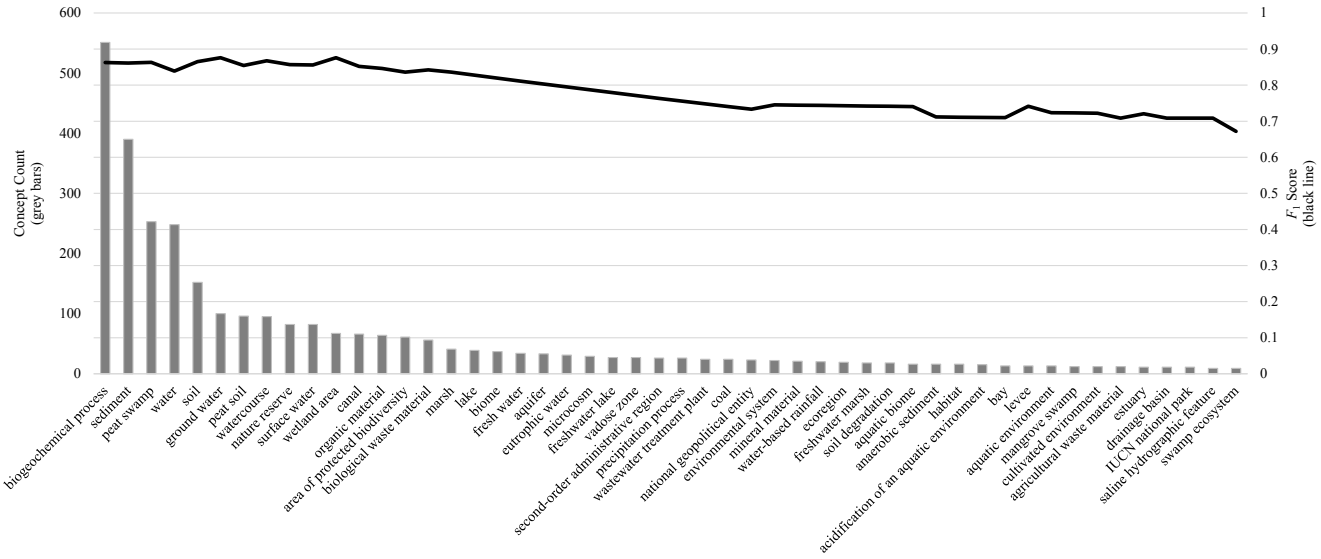

Figure 2. Frequency of the top 50 ENVO concepts (grey bars) and the average 10-fold testing results (F_1 scores) for each of the concepts (black line).

all concepts. The second-to-last row shows the results for including the abstract concepts as features, but omitting the position features. This resulted in lower results overall, but significantly impacted the average score for the single occurrence concepts. Interestingly, we investigated this and found that deeper concepts (i.e., in terms of the ontology hierarchy) are found at higher densities close to the middle of the articles as well as the centers of article sections. In retrospect this makes sense, as the methodology section of a scientific paper (located around the middle) would normally contain detailed concepts rather than abstract ones. To put this together, most of the single occurrence concepts are deeper, low-level concepts, hence the low occurrence frequency in the original data. The second row shows the results for only omitting the abstract concepts as features when labeling the rest of the text in the articles. Again, the models' performance dropped overall, but less so than for single occurrence concepts. This can be attributed again to including the sentence position features, which aid the labeling for less frequent concepts. Finally, the first row shows the results for omitting both features with the models performing the worst across all concepts.

V. Contributions

In this paper we present a system for learning to identify domain-specific ontology concepts in the academic literature, specifically for the biogeochemical domain. We created a dataset of academic articles that we manually annotated. We then used the annotated dataset to build a supervised machine learning model—a random decision forest classifier—which was trained and tested using cross-validation. Further, we identified a set of useful features and evaluated their efficacy in training and testing the models. Our model significantly outperformed the the baseline methods discussed, however we do believe that the model could be further improved, in particular by performing additional preprocessing and including additional features such as multi-word expressions. It is also important to note the small size of the annotated corpus used for training; more data will likely improve the result. In this vein our annotation is ongoing and our team is well on its way to double the size of the dataset. This will provide our models with more training examples for the concepts as well as additional unseen concepts. Additionally, using a larger dataset will allow for further tuning of the model's parameters which may yield better performance.

Acknowledgment

Our team would like to thank Kalli Unthank for her effort in the original pilot annotation study, creating the search queries, and identifying the relevant articles used in the current annotation study. We also thank our domain-expert annotators Francisca Olmos de Aguilera, Rodrigo Tavara, Shannon Joseph, Thomas Zerquera, Jose Llaguno, and Daniel Infante. Finally, our team would like to thank the CREST CAChE leadership at FIU for their continued support of the annotation study and the team's collaborative project. This material is based upon work supported by the National Science Foundation under Grant No. HRD-1547798, which was awarded to Florida International University as part of the Centers of Research Excellence in Science and Technology (CREST) Program. This is contribution number 870 from the Southeast Environmental Research Center in the Institute of Water and Environment at Florida International University.

REFERENCES

[1] D. Lewandowski, "Evaluating the retrieval effectiveness of web search engines using a representative query sample," *Journal of the Association for Information Science and Technology*, vol. 66, no. 9, pp. 1763–1775, 2015.

[2] L. Martínez-Sanahuja and D. Sánchez, "Evaluating the suitability of web search engines as proxies for knowledge discovery from the web," *Procedia Computer Science*, vol. 96, pp. 169–178, 2016.

[3] J. Brophy and D. Bawden, "Is google enough? Comparison of an internet search engine with academic library resources," in *Aslib Proceedings*, vol. 57, no. 6. Emerald Group Publishing Limited, 2005, pp. 498–512.

[4] T. Leyba, "Semantic search by means of word sense disambiguation using a lexicon," Apr 19, 2016, US Patent 9,317,589.

[5] J. Huang, F. Gutierrez, H. J. Strachan, D. Dou, W. Huang, B. Smith, J. A. Blake, K. Eilbeck, D. A. Natale, Y. Lin, B. Wu, N. de Silva, X. Wang, Z. Liu, G. M. Borchert, M. Tan, and A. Ruttenberg, "Omnisearch: A semantic search system based on the ontology for microRNA target (OMIT) for microRNA-target gene interaction data," *Journal of Biomedical Semantics*, vol. 7, no. 1, p. 25, 2016.

[6] J. Dang, M. Kalender, C. Toklu, and K. Hampel, "Semantic search tool for document tagging, indexing and search," Jun 20, 2017, US Patent 9,684,683.

[7] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature," *PLoS Biology*, vol. 2, no. 11, p. e309, 2004.

[8] M. Allahyari, K. J. Kochut, and M. Janik, "Ontology-based text classification into dynamically defined topics," in *Proceedings of the 2014 IEEE International Conference on Semantic Computing (ICSC)*, Newport Beach, CA, 2014, pp. 273–278.

[9] G. Sriharee, "An ontology-based approach to auto-tagging articles," *Vietnam Journal of Computer Science*, vol. 2, no. 2, pp. 85–94, 2015.

[10] J. D. Eisenberg, D. Banisakher, M. Presa, K. Unthank, M. A. Finlayson, R. Price, and S.-C. Chen, "Toward semantic search for the biogeochemical literature," in *Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration (IRI)*, San Diego, CA, 2017, pp. 517–525.

[11] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," in *Proceedings of the 2015 IEEE International Conference on Semantic Computing (ICSC)*, Anaheim, CA, 2015, pp. 244–251.

[12] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *Journal of Information Science*, vol. 36, no. 3, pp. 306–323, 2010.

[13] H. Jaiwei and M. Kamber, *Data mining: concepts and techniques*. San Francisco: Morgan Kaufmann, 2006.

[14] R. Setchi and Q. Tang, "Concept indexing using ontology and supervised machine learning," *Transactions on Engineering, Computing and Technology*, vol. 19, pp. 221–226, 2007.

[15] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The web as a graph: Measurements, models, and methods," in *Proceedings of the International Computing and Combinatorics Conference*, Tokyo, Japan, 1999, pp. 1–17.

[16] S. Žitnik, M. Žitnik, B. Zupan, and M. Bajec, "Sieve-based relation extraction of gene regulatory networks from biological literature," *BMC Bioinformatics*, vol. 16, no. 16, p. S1, 2015.

[17] H. Gurulingappa, A. Mateen-Rajpu, and L. Toldo, "Extraction of potential adverse drug events from medical case reports," *Journal of Biomedical Semantics*, vol. 3, no. 1, p. 15, 2012.

[18] M.-F. Moens, *Information extraction: Algorithms and prospects in a retrieval context*. Dordrecht, The Netherlands: Springer Netherlands, 2006.

[19] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, and S. E. Lewis, "The environment ontology: Contextualizing biological and biomedical entities," *Journal of Biomedical Semantics*, vol. 4, no. 1, p. 43, 2013.

[20] D. Field, L. Amaral-Zettler, G. Cochrane, J. R. Cole, P. Dawyndt, G. M. Garrity, J. Gilbert, F. O. Glöckner, L. Hirschman, and I. Karsch-Mizrachi, "The genomic standards consortium," *PLoS Biology*, vol. 9, no. 6, p. e1001088, 2011.

[21] L. M. Schriml, C. Arze, S. Nadendla, A. Ganapathy, V. Felix, A. Mahurkar, K. Phillippy, A. Gussman, S. Angiuoli, E. Ghedin, O. White, and N. Hall, "GeMInA, genomic metadata for infectious agents, a geospatial surveillance pathogen database," *Nucleic Acids Research*, vol. 38, Suppl. 1, pp. D754–D764, 2010.

[22] "The national institute for allergy and infectious diseases (NIAID), microbiology and infectious diseases resources, DMID metadata standards core sample," https://www.niaid.nih.gov/research/dmid-metadata-standards-core-sample, 2017, retrieved on Jun 19, 2018.

[23] S. M. Ullrich, T. W. Tanton, and S. A. Abdrashitova, "Mercury in the Aquatic Environment: A Review of Factors Affecting Methylation," *Critical Reviews in Environmental Science and Technology*, vol. 31, no. 3, pp. 241–293, 2001.

[24] J. M. Benoit, C. C. Gilmour, R. P. Mason, and A. Heyes, "Sulfide controls on mercury speciation and bioavailability to methylating bacteria in sediment pore waters," *Environmental Science & Technology*, vol. 33, no. 6, pp. 951–957, 1999.

[25] C. C. Gilmour, E. A. Henry, and R. Mitchell, "Sulfate stimulation of mercury methylation in freshwater sediments," *Environmental Science & Technology*, vol. 26, no. 11, pp. 2281–2287, 1992.

[26] G. C. Compeau and R. Bartha, "Effect of salinity on mercury-methylating activity of sulfate-reducing bacteria in estuarine sediments," *Applied and Environmental Microbiology*, vol. 53, no. 2, pp. 261–265, 1987.

[27] H. L. Drake, N. G. Aumen, C. Kuhner, C. Wagner, A. Griesshammer, and M. Schmittroth, "Anaerobic microflora of Everglades sediments: Effects of nutrients on population profiles and activities." *Applied and Environmental Microbiology*, vol. 62, no. 2, pp. 486–493, 1996.

[28] J. Benoit, R. P. Mason, C. C. Gilmour, and G. R. Aiken, "Constants for mercury binding by dissolved organic matter isolates from the Florida Everglades," *Geochimica et cosmochimica acta*, vol. 65, no. 24, pp. 4445–4451, 2001.

[29] L. B. Cleckner, C. C. Gilmour, J. P. Hurley, and D. P. Krabbenhoft, "Mercury methylation in periphyton of the Florida Everglades," *Limnology and Oceanography*, vol. 44, no. 7, pp. 1815–1825, 1999.

[30] C. C. Gilmour, G. Riedel, M. Ederington, J. Bell, G. Gill, and M. Stordal, "Methylmercury concentrations and production rates across a trophic gradient in the northern Everglades," *Biogeochemistry*, vol. 40, no. 2-3, pp. 327–345, 1998.

[31] M. C. Marvin-DiPasquale and R. S. Oremland, "Bacterial methylmercury degradation in Florida Everglades peat sediment," *Environmental Science & Technology*, vol. 32, no. 17, pp. 2556–2563, 1998.

[32] J. W. Harvey and P. V. McCormick, "Groundwater's significance to changing hydrology, water chemistry, and biological communities of a floodplain ecosystem, Everglades, South Florida, USA," *Hydrogeology Journal*, vol. 17, no. 1, pp. 185–201, 2009.

[33] R. M. Chambers and K. A. Pederson, "Variation in soil phosphorus, sulfur, and iron pools among South Florida wetlands," *Hydrobiologia*, vol. 569, no. 1, pp. 63–70, 2006.

[34] W. Orem, C. Gilmour, D. Axelrad, D. Krabbenhoft, D. Scheidt, P. Kalla, P. McCormick, M. Gabriel, and G. Aiken, "Sulfur in the South Florida ecosystem: Distribution, sources, biogeochemistry, impacts, and management for restoration," *Critical Reviews in Environmental Science and Technology*, vol. 41, no. S1, pp. 249–288, 2011.

[35] D. J. Casagrande, K. Siefert, C. Berschinski, and N. Sutton, "Sulfur in peat-forming systems of the Okefenokee swamp and Florida Everglades: Origins of sulfur in coal," *Geochimica et Cosmochimica Acta*, vol. 41, no. 1, pp. 161–167, 1977.

[36] F. E. Dierberg, T. A. DeBusk, N. R. Larson, M. D. Kharbanda, N. Chan, and M. C. Gabriel, "Effects of sulfate amendments on mineralization and phosphorus release from South Florida (USA) wetland soils under anaerobic conditions," *Soil Biology and Biochemistry*, vol. 43, no. 1, pp. 31–45, 2011.

[37] M. A. Musen, "The Protégé project: A look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015.

[38] R. Artstein and M. Poesio, "Inter-Coder Agreement for Computational Linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[39] S. Bird and E. Loper, "NLTK: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, Barcelona, Spain, 2004, p. 31.

[40] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[41] PubMed Help, "Stopwords table," https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/, National Center for Biotechnology Information, Bethesda, MD, 2005, accessed on Jun 19, 2018.

[42] L. Tan, "Pywsd: Python implementations of word sense disambiguation (WSD) technologies [software]," https://github.com/alvations/pywsd, 2014, accessed on Jun 19, 2018.

[43] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[44] K. Church and W. Gale, "Inverse document frequency (idf): A measure of deviations from Poisson," in *Natural Language Processing Using Very Large Corpora*. New York: Springer, 1999, pp. 283–295.

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[46] A. Trask, P. Michalak, and J. Liu, "sense2vec – A fast and accurate method for word sense disambiguation in neural word embeddings," *arXiv Computing Research Repository (CoRR)*, 2015, abs/1511.06388.

[47] E. AI, "SpaCy: A library for advanced natural language processing in python and cython [software]," https://github.com/explosion/spaCy, 2015, accessed on Jun 19, 2018.

[48] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, Montreal, Canada, 1995, pp. 278–282.

[49] J. Franklin, "The elements of statistical learning: Data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[50] A. Criminisi, J. Shotton, E. Konukoglu *et al.*, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2–3, pp. 81–227, 2012.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[52] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.