Ontology-based Categorization of Bacteria and Habitat Entities using Information Retrieval Techniques

Mert Tiftikci*, **Hakan Şahin***, **Berfu Büyüköz***, **Alper Yayıkçı***, **Arzucan Özgür**
Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey
{mert.tiftikci,hakan.sahin1,berfu.buyukoz,alper.yayikci,arzucan.ozgur}@boun.edu.tr

Abstract

A database which provides information about bacteria and their habitats in a comprehensive and normalized way is crucial for applied microbiology studies. Having this information spread through textual resources such as scientific articles and web pages leads to a need for automatically detecting bacteria and habitat entities in text, semantically tagging them using ontologies, and finally extracting the events among them. These are the challenges set forth by the Bacteria Biotopes Task of the BioNLP Shared Task 2016. This paper describes a system for habitat and bacteria entity normalization through the OntoBiotope ontology and the NCBI taxonomy, respectively. The system, which obtained promising results on the shared task data set, utilizes basic information retrieval techniques.

1 Introduction

Retrieving useful information from text became increasingly important as numerous data are collected on the Internet (Singhal, 2001). It became even more crucial to be able to reach the desired information from among lots of articles and resources when it comes to studies of science, especially biomedicine (Cohen and Hersh, 2005). The problem tackled in this paper is the semantic categorization of bacteria and habitat entities extracted from scientific paper abstracts. This problem has been addressed as a sub-task of the BioNLP Bacteria Biotope Shared Task 2016 (Deléger et al., 2016).

The Bacteria Biotope Task of the BioNLP Shared Task was previously conducted in 2011 (Bossy et al., 2011; Bossy et al., 2012) and 2013 (Bossy et al., 2013; Bossy et al., 2015). Both machine learning based (Nguyen and Tsuruoka, 2011; Björne et al., 2012; Grouin, 2013; Claveau, 2013) and rule based approaches (Ratkovic et al., 2012; Karadeniz and Ozgür, 2013; Bannour et al., 2013) have been developed to identify and normalize bacteria and habitat entities. The normalization of habitat entities through the OntoBiotope ontology has been first addressed in the 2013 edition of the Entity Categorization subtask, where four teams participated (Bossy et al., 2013; Bossy et al., 2015). The highest F1-score (61%) and lowest Slot Error Rate (SER) (66%) was achieved by the *LIPN* system (Bannour et al., 2013), which used a combination of an ontology projection method and a rule based machine learning algorithm, namely WHISK (Soderland, 1999). The *BOUN* system (Karadeniz and Ozgür, 2013; Karadeniz and Özgür, 2015), which is based on syntactic rules, and the *LIMSI* system (Grouin, 2013), which is based on Conditional Random Fields (CRF), obtained similar SER scores (68%). The *IRISA* system, which obtained a SER value of 93% (Claveau, 2013), used the k nearest neighbor algorithm with the Okapi-BM25 (Robertson et al., 1999) similarity measure.

In this paper we describe the system that we developed for our participation at the "Entity categorization" sub-task of the Bacteria Biotope (BB3) task of BioNLP Shared Task 2016. Motivated by the promising results of rule-based entity categorization approaches in the previous editions of the

*These authors contributed equally to this work.

Proceedings of the 4th BioNLP Shared Task Workshop, pages 56–63,
Berlin, Germany, August 13, 2016. ©2016 Association for Computational Linguistics

shared task, we designed a rule-based approach that makes use of information retrieval and pattern matching techniques for normalizing bacteria and habitat entities through the provided ontologies.

2 System Description

2.1 Overview of the System

We developed an ontology based categorization system for the Entity Categorization sub-task. The system consists of two modules, one for the habitat categorization task and the other for the bacteria categorization task. The habitat categorization module makes use of basic information retrieval techniques including tf-idf scoring and cosine similarity. The bacteria categorization module utilizes string matching methods such as Levenshtein distance. These modules are described in detail in the following sub-sections.

2.2 Categorization of Habitat Entities

The workflow of the habitat categorization module is presented in Figure 1. Given a habitat entity mention, the goal is to identify the corresponding concepts in the OntoBiotope ontology. First, the OntoBiotope ontology is expanded by using the training and development data sets. Next, both exact matching and partial matching approaches are used to identify the ontology concepts relevant to the habitat entity mention. Partial matching is formulated as an information retrieval task, where tf-idf scoring and cosine similarity are used to rank the ontology concepts with respect to the given habitat entity.

2.2.1 Ontology Expansion

The OntoBiotope ontology is an ontology of biotopes organized as a hierarchical structure of concepts. A sample concept in the ontology is shown in Figure 2. An OntoBiotope concept consists of an ID, name, as well as exact and related synonyms. The parent-child relations between concepts are represented with the is_a field.

```
[Term]
id: OBT:000218
name: animal
synonym: “animal host” RELATED []
synonym: “animal-associated habitat” EXACT []
synonym: “animal species” RELATED []
is_a: OBT:000036 ! eukaryote host
```

Figure 2: A sample OntoBiotope ontology concept

The documents in the training and development data sets have habitat mentions labeled with their corresponding OntoBiotope concepts. We expanded the OntoBiotope ontology by including these habitat mentions as *related synonyms* to the associated concepts. Figure 3 shows the expanded version of the “animal” concept in Figure 2, where the concept has been expanded by adding the “animals” and “animal models” as related synonyms.

```
[Term]
id: OBT:000218
name: animal
synonym:“animal host” RELATED []
synonym:“animal-associated habitat” EXACT []
synonym: “animal species” RELATED []
synonym: “animals” RELATED []
synonym: “animal models” RELATED []
is_a: OBT:000036 ! eukaryote host
```

Figure 3: A sample expanded OntoBiotope ontology concept

2.2.2 Normalization

A habitat entity mention is normalized by matching it with one or more concepts in the OntoBiotope ontology. We used exact and partial matching approaches for this task.

Given a habitat entity mention, first the system searches for exact matches with the *names* or *exact synonyms* of the ontology concepts. If an exact match is found, the habitat entity is labeled with the corresponding ontology concepts.

If an exact match is not found, partial matching is performed using information retrieval techniques. Each concept in the ontology is treated as a document and an inverted index of concepts is created. The unigrams and bigrams in the names, exact synonyms, and related synonyms of concepts are represented as tf-idf weighted terms in the inverted index. The habitat entity mention is treated

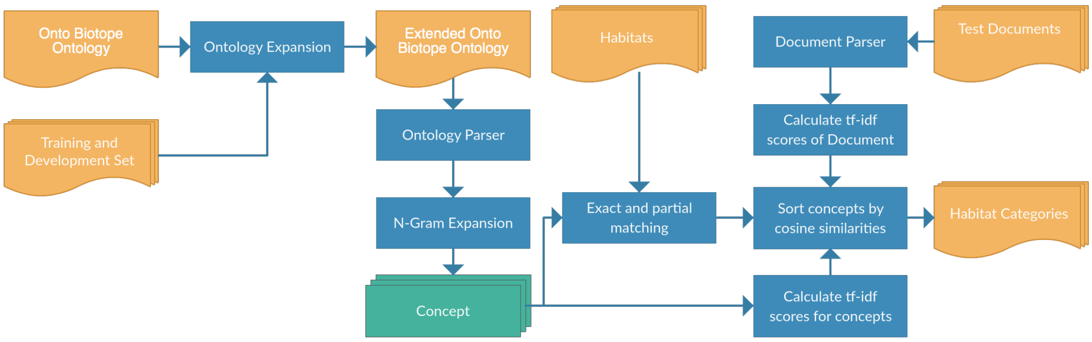
Onto Biotope Ontology
Ontology Expansion
Extended Onto Biotope Ontology
Habitats
Document Parser
Test Documents
Training and Development Set
Ontology Parser
Calculate tf-idf scores of Document
N-Gram Expansion
Exact and partial matching
Sort concepts by cosine similarities
Habitat Categories
Concept
Calculate tf-idf scores for concepts

Figure 1: Categorization of Habitat Entities

as a query. In order to capture more contextual information, the query is expanded by including the unigrams and bigrams of the document where the habitat mention occurs. The cosine similarities between the query and the concepts in the inverted index are computed. The concepts are ranked based on their cosine similarity scores to the habitat query and the habitat mention is annotated with the concept that obtains the highest cosine similarity score.

If the system does not find any relevant concepts based on exact and partial matching, the habitat mention is normalized with the root of the ontology, i.e., with the concept OBT:000001 shown in Figure 4.

[Term]
id: OBT:000001
name: experimental medium
is_a: OBT:000000 ! bacteria habitat

Figure 4: Default normalization

2.3 Categorization of Bacteria Entities

The workflow of the bacteria categorization module is presented in Figure 5. In this module, bacteria entity mentions are normalized with their corresponding taxonomy IDs in the NCBI taxonomy. First, a preprocessing step is applied where punctuation marks are removed and abbreviation and acronyms are expanded. Next, a normalization and matching step is applied where preprocessed bacteria mentions are matched against the NCBI taxonomy using exact and approximate string matching methods.

2.3.1 Preprocessing

In order to increase the possibility of matching bacteria mentions with their correct categories in the NCBI taxonomy, a set of preprocessing techniques described below are developed by examining the documents and the NCBI taxonomy.

2.3.1.1 Punctuation Mark Removal

Some punctuation marks provide no useful information for our task and may hinder the performance of the system for matching bacteria names in the NCBI taxonomy. Therefore, we replaced parenthesis, quotation marks, and multiple white spaces with a single white space character. In addition, lower casing all characters is performed in this step.

The preprocessing steps described below make use of the context information, i.e., the document where the bacteria entity occurs to transform the bacteria entity mention to a more convenient format for matching with the NCBI taxonomy.

2.3.1.2 Abbreviation Expansion

One of the most common challenges for bacteria categorization is that bacteria names frequently occur in abbreviated forms. In general, the first occurrence of a bacteria name in a document is written as a full name (e.g., *"Escherichia coli"*) and the successive mentions are written in abbreviated forms (e.g., *"E. Coli"*). A bacteria mention in abbreviated form is compared with the previous closest[1] bacteria mentions in the document. If a previously occurring bacteria mention starts with the same capital letter as the abbreviated form and

[1] Distance between two bacteria mentions is computed based on the positions of the mentions in the document.

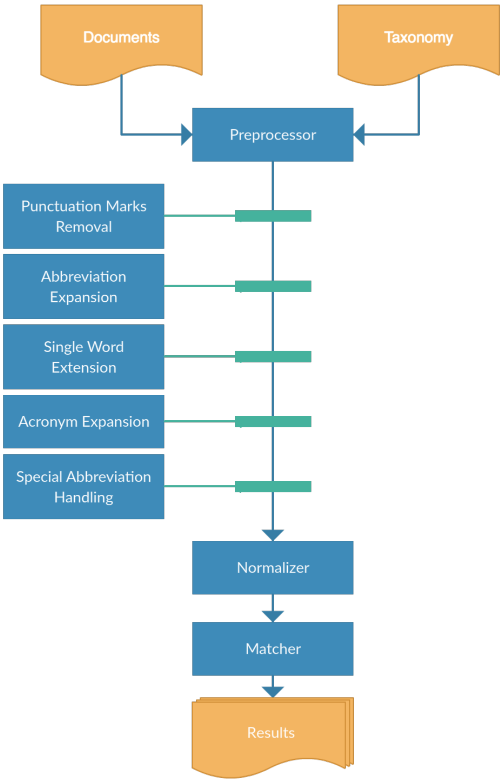

Figure 5: Categorization of Bacteria Entities

contains the remaining sub-string of the abbreviated form, the abbreviated form is converted to the corresponding bacteria mention in full name form before searching in the NCBI Taxonomy. For example, *"E. Coli"* is expanded to *"Escherichia coli"* if there is an occurrence of *"Escherichia coli"* before the abbreviated form in the same document. If there is not a match with previously occurring bacteria mentions in the document, then a search is performed starting from the abbreviated form until the end of the document, to look for the expanded version.

Another commonly occurring abbreviation pattern in documents is that the first terms of two bacteria mentions actually refer to the same word, but one of them occurs as an abbreviation. For example, in the *"Chlamydia trachomatis and C. psittaci"* phrase, *"C."* corresponds to *"Chlamydia"*. The bacteria mentions occurring before and after the abbreviated name in the document are examined. If there is not a bacteria mention in the document matching the sub-string *"psittaci"*, then bacteria mentions starting with the same letter are considered as matches. Search is performed from the abbreviated mention first to the beginning of the document and next to the end of the document. Preference is given to matches that are closer to the abbreviated form in the document. In the provided example, *"C. psittaci"* is expanded to *"Chlamydia psittaci"* before searching the NCBI Taxonomy.

2.3.1.3 Single Word Abbreviation Expansion

Another common abbreviation pattern is when after the full name mention of a bacterium, it is referred to with the first word in its name (i.e., its genus name) in the rest of the document. For example, *"Escherichia coli"* is referred to as *"Escherichia"*. To expand such abbreviations, the single word bacteria mention is compared with the preceding and following bacteria mentions in the document. If the single word bacteria mention is a sub-string of a multi-word bacteria mention, the single word mention is expanded to that multi-word mention. Preceding and closer matches are given higher precedence.

2.3.1.4 Acronym Expansion

Although bacteria entities can be referred to with acronyms like "MRSA" in documents, such acronyms are not directly represented in the NCBI taxonomy, but may appear within the names of multiple bacteria categories with different IDs such as "Staphylococcus aureus MRSA-Lux-1" and "Staphylococcus aureus MRSA-Lux-2". Based on our observation in the training set, in order to resolve ambiguity, we expanded such acronyms consisting of less than five capital letters to the closest bacteria mention in the same document.

2.3.1.5 Handling Other Special Abbreviations

Several special abbreviations including "sp.", "spp.", "strain", "str.", "aff.", "cf.", "subgen.", "gen.", and "nov." are used within species names in biomedical documents. These special abbreviations should be ignored while matching species names against the NCBI Taxonomy, since they are not in general included in the "scientific name" or the name tagged as "authority" in the NCBI Taxonomy. For instance, *"Escherichia (sp.) coli"* in a biomedical document should match with *"Escherichia coli"* in the NCBI Taxonomy. Therefore, we removed such abbreviations from bacteria name mentions in text to improve matching performance.

Another challenge in bacteria categorization is that a bacteria species can have numerous sub-types, each corresponding to a different category in the taxonomy. This makes it hard to match a bacteria mention in text with its corresponding category in the taxonomy. Special rules are designed by analyzing the provided training and development data to enhance matching in such cases. For example, the word "type" is removed from a bacteria mention in text before matching against the NCBI Taxonomy. This enables matching *"Escherichia coli type a"* in text with *"Escherichia coli a"* in the taxonomy. In cases where sub-types are denoted with semi-colon, the sub-string following the semi-colon in the bacteria mention is removed before matching with the terms in the taxonomy. This allows *"Escherichia Coli O8:K88"* in text to match with the category *"Escherichia Coli O8"* in the taxonomy. Other transformations performed to enhance sub-type matching are converting the "ssp" abbreviation to "subsp." and the "ara+" sub-string to "ara+ biotype" in the bacteria mentions in text. We did not remove these sub-species denoting abbreviations, since keeping them resulted in better performance. We converted these abbreviations to their versions occurring in the names tagged as "scientific name" or "authority" in the taxonomy.

2.3.2 Normalization and Matching with Taxonomy

After the preprocessing steps, a bacteria mention in text is converted to a candidate phrase to be matched against the categories in the NCBI Taxonomy. First, an exact match is performed and the phrase is assigned to the matching category in the taxonomy.

If there is no an exact match, then partial phrase matching is performed. In a candidate phrase, it is possible that an irrelevant word, for instance an adjective, appears. That irrelevant word will cause an unsuccessful search in the taxonomy. Therefore, partial matching with the first two words, last two words, and first and last words of the candidate phrase are performed and the partially matching category is assigned to the candidate phrase.

If exact and partial phrase matching do not match with any categories in the taxonomy, then partial string matching using *Levenshtein edit distance* is performed to detect the most similar category to the candidate. We set the edit distance threshold to 2. For taxonomy categories with edit distance less than or equal to the threshold, "error ratio" is computed as follows.

$$\textbf{Error ratio} = \frac{\textit{edit distance}}{\textit{length of the candidate}} \tag{1}$$

The error ratio threshold is set to 0.2. So, the candidate phrase is assigned to a taxonomy category, if edit distance and error ratio are less than or equal to 2 and 0.2, respectively. In this case, if a candidate phrase is of length 4, and if a bacteria name is found in the taxonomy with edit distance 1, this is not accepted as a successful match, since error ratio is 0.25.

Finally, if an exactly or partially matching category is not found, the context of the bacteria mention in the document is used for category assignment. In this case, the bacteria mention is mapped to the same category of the closest bacteria mention for which a category was assigned in the document.

3 Evaluation and Results

Different evaluation metrics are used for habitat and bacteria entities. Wang similarity (Wang et al., 2007) with a weight of 0.65 is used for evaluation of habitat entities by computing the similarity between the reference and the predicted normalization. This metric determines a semantic similarity score between two nodes of a directed acyclic graph (DAG) whose nodes have is_a relations with their parents. This similarity metric takes into account the locations of the terms within the DAG, their distances to the root and to common ancestors. Evaluation of bacteria entities is stricter. If two terms (reference and predicted) are the same, the similarity score is equal to 1, otherwise it is 0. Two systems, namely LIMSI and our system BOUN participated in the BB3-CAT shared task. The official evaluation results on the shared task test data set are presented in Table 1[2]. Among the two participating systems, our system ranked first in the overall task of habitat and bacteria name categorization, as well as in the individual sub-tasks of habitat name categorization and bacteria name categorization.

[2] http://2016.bionlp-st.org/tasks/bb2/bb3-evaluation

Precision	BOUN	LIMSI
Main Scoring	0.679	0.503
Habitats Only	0.620	0.438
Bacteria Only	0.801	0.637

Table 1: Official evaluation results

3.1 Results for Habitat Categorization

This sub-section provides the evaluation results of the system at its major development phases over the training and development data sets. Precision and recall values calculated from true positives, false positives, and false negatives are reported. Habitats that are normalized correctly are considered to be true positive, the ones normalized with wrong categories are considered to be false positives, and if there are no exact or partial matches found for a habitat, it is considered to be a false negative. BB3-CAT Precision corresponds to the entity categorization precision computed using the online evaluation tool provided at the BB3 shared task[3]. BB3-CAT precision is based on the Wang similarity (Wang et al., 2007).

	Development	Training
True Positive	214	309
False Positive	186	349
False Negative	321	516
Precision	0.53	0.47
Recall	0.40	0.37
BB3-CAT Precision	0.58	0.55

Table 2: Results without bigram expansion or normalization to OBT:000001

Table 2 presents the baseline results when only unigrams are used for cosine similarity computation and no normalization to the root concept is performed. Table 3 presents the results after introducing the bigrams to the system and simultaneously increasing the term frequency weights of the unigrams by a factor of two.

	Development	Training
True Positive	224	332
False Positive	176	326
False Negative	311	493
Precision	0.56	0.50
Recall	0.42	0.40
BB3-CAT Precision	0.61	0.59

Table 3: Results with bigram expansion

Although bigram expansion increases the scores, there are still some habitats with no matched categories. In case of computing Wang scores, leaving a habitat without a category is a drawback, since any normalization gains a better score than no normalization. Our results in Table 4 show that normalizing unmatched habitats to the concept OBT:000001 increases the Wang scores.

	Development	Training
True Positive	226	343
False Positive	228	404
False Negative	309	482
Precision	0.50	0.46
Recall	0.42	0.42
BB3-CAT Precision	0.63	0.62

Table 4: Results with bigram expansion and normalization to OBT:000001

3.2 Results for Bacteria Categorization

Our baseline system that only performs punctuation removal and exact matching between candidate name and a bacteria name in the taxonomy obtained precision-F-measure values of 0.39-0.40 over the development set and 0.57-0.58 over the training set. This benchmark was a plain starting point for this study.

We improved the baseline system by applying the preprocessing steps. The most common errors seen in the results were the unmatched abbreviations. Then, we applied the abbreviation expansion step described in Sub-section 2.3.1.2. and our precision-F-measure values increased to 0.59-0.71 over the development set and to 0.74-0.83 over the training set. Thus, this enhancement was the most effective one overall.

After that, we applied the single word abbreviation expansion step described in Subsection 2.3.1.3. This improvement increased the precision-F-measure values to 0.64-0.75 on the de-

[3] http://bibliome.jouy.inra.fr/demo/BioNLP-ST-2016-Evaluation/index.html

velopment set and to 0.78-0.86 on the training set. Finally, we applied the acronym expansion step and it raised the precision-F-measure values to 0.67-0.77 on the development set and to 0.81-0.87 on the training set. This final result is the last benchmark that we got after preprocessing the bacteria names. The results of the preprocessing steps are presented in Table 5.

		Development	Training
Punctuation rem.	P	0.39	0.57
	R	0.41	0.59
	F	0.40	0.58
Abbreviation exp.	P	0.59	0.74
	R	0.89	0.94
	F	0.71	0.83
Single word exp.	P	0.64	0.78
	R	0.90	0.95
	F	0.75	0.86
Acronym exp.	P	0.67	0.81
	R	0.90	0.93
	F	0.77	0.87

Table 5: Results after preprocessing (P: Precision, R: Recall, F: F-measure)

Table 6 summarizes the results of the normalization and matching steps that are performed after the preprocessing steps. Matching with the original bacteria mention first and matching with the preprocessed version if there is no a match with the original version resulted in 0.77 precision and 0.78 F-measure over the development set and 0.87 precision and 0.88 F-measure over the training set. In addition, both partial phrase matching using two-word combinations and partial string matching using Levenshtein distance resulted in improved performance. Finally, assigning unmatched bacteria mentions to the taxonomy of the closest categorized bacteria mention in the same document resulted in considerable improvement in precision and F-measure.

4 Conclusion

This study introduced a system that is developed in the scope of the Entity Categorization sub-task of the BioNLP Bacteria Biotope Shared Task 2016. The system consists of two modules both of which target normalizing entities that have been detected in scientific paper abstracts. While the habitat categorization module operates on habitat mentions expressed in natural language and uses the OntoBiotope ontology for normalization, the bacteria categorization module deals with bacteria mentions expressed as more structured scientific expressions and uses the NCBI Taxonomy for normalization.

		Development	Training
Exact matching	P	0.77	0.87
	R	0.79	0.89
	F	0.78	0.88
Sub-phrases	P	0.81	0.90
	R	0.85	0.92
	F	0.83	0.91
Edit distance	P	0.83	0.91
	R	0.89	0.95
	F	0.86	0.93
Unmatched handling	P	0.89	0.95
	R	0.99	0.99
	F	0.94	0.97

Table 6: Results after the matching and normalization steps (P: Precision, R: Recall, F: F-measure)

Promising results are obtained by both modules, which utilize pattern matching and information retrieval techniques. According to the official evaluations, the habitat categorization module obtained 0.620 precision and the bacteria categorization module obtained 0.801 precision, which led to achieving the highest overall precision of 0.679 in the BB3-CAT sub-task. As future work, integrating WordNet based similarity measures to improve ontology-based matching will be investigated.

Acknowledgments

This work has been supported by Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme. We would like to thank the Bacteria Biotope Task organizers for organizing the shared task and for their help with the questions.

References

Sondes Bannour, Laurent Audibert, and Henry Soldano. 2013. Ontology-based Semantic Annotation: An Automatic Hybrid Rule-based Method. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 139–143. Sofia August.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13(11):1.

Robert Bossy, Julien Jourde, Philippe Bessieres, Maarten Van De Guchte, and Claire Nédellec. 2011. BioNLP Shared Task 2011: Bacteria Biotope. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 56–64. Association for Computational Linguistics.

Robert Bossy, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Maarten Van

De Guchte, Philippe Bessières, and Claire Nédellec. 2012. BioNLP Shared Task-The Bacteria Track. *BMC Bioinformatics*, 13(11):1.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. 2013. BioNLP Shared Task 2013–An Overview of the Bacteria Biotope Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 161–169.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec. 2015. Overview of the Gene Regulation Network and the Bacteria Biotope Tasks in BioNLP'13 Shared Task. *BMC Bioinformatics*, 16(Suppl 10):S1.

Vincent Claveau. 2013. IRISA Participation to Bionlp-ST 2013: Lazy-Learning and Information Retrieval for Information Extraction tasks. In *BioNLP Workshop, Colocated with ACL 2013*, pages 188–196.

Aaron M Cohen and William R Hersh. 2005. A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics*, 6(1):57–71.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessières, and Claire Nédellec. 2016. Overview of the Bacteria Biotope Task at Bionlp Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, Berlin, Germany, August. Association for Computational Linguistics.

Cyril Grouin. 2013. Building a Contrasting Taxa Extractor for Relation Identification from Assertions: Biological Taxonomy & Ontology Phrase Extraction System. *ACL 2013*, 144.

Ilknur Karadeniz and Arzucan Ozgür. 2013. Bacteria Biotope Detection, Ontology-based Normalization, and Relation Extraction using Syntactic Rules. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 170–177.

İlknur Karadeniz and Arzucan Özgür. 2015. Detection and Categorization of Bacteria Habitats using Shallow Linguistic Analysis. *BMC Bioinformatics*, 16(Suppl 10):S5.

Nhung TH Nguyen and Yoshimasa Tsuruoka. 2011. Extracting Bacteria Biotopes with Semi-Supervised Named Entity Recognition and Coreference Resolution. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 94–101. Association for Computational Linguistics.

Zorana Ratkovic, Wiktoria Golik, and Pierre Warnier. 2012. Event Extraction of Bacteria Biotopes: A Knowledge-Intensive NLP-based Approach. *BMC Bioinformatics*, 13(11):1.

Stephen E Robertson, Steve Walker, Micheline Beaulieu, and Peter Willett. 1999. Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive Track. *Nist Special Publication SP*, pages 253–264.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

Stephen Soderland. 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1-3):233–272.

James Z Wang, Zhidian Du, Rapeeporn Payattakool, S Yu Philip, and Chin-Fu Chen. 2007. A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics*, 23(10):1274–1281.