BOUN-ISIK Participation: An Unsupervised Approach for the Named Entity Normalization and Relation Extraction of Bacteria Biotopes

İlknur Karadeniz
Işık University
ilknur.karadeniz
@isikun.edu.tr

Ömer Faruk Tuna
Işık University
218DCS8074
@isik.edu.tr

Arzucan Özgür
Boğaziçi University
arzucan.ozgur
@boun.edu.tr

Abstract

This paper presents our participation at the Bacteria Biotope Task of the BioNLP Shared Task 2019. Our participation includes two systems for the two subtasks of the Bacteria Biotope Task: the normalization of entities (BB-norm) and the identification of the relations between the entities given a biomedical text (BB-rel). For the normalization of entities, we utilized word embeddings and syntactic re-ranking. For the relation extraction task, pre-defined rules are used. Although both approaches are unsupervised, in the sense that they do not need any labeled data, they achieved promising results. Especially, for the BB-norm task, the results have shown that the proposed method performs as good as deep learning based methods, which require labeled data.

1 Introduction

The amount of electronic resources in the biomedical domain and its rapid growth are major challenges for the scientists who make research in this domain. Text mining methods which aim to automatically extract useful information from the text of these electronic resources provide convenience to the researchers.

A number of shared tasks, including the BioNLP Shared Tasks, have been conducted with the goal of developing biomedical text mining methods. In 2011, the Bacteria Biotope Task has been conducted for the first time as a part of the BioNLP Shared Task targeting the extraction of useful information regarding bacteria and their habitats (Bossy et al., 2011). Since then, the participant teams of the following shared task series developed various solutions for the problem of bacteria biotopes (Bossy et al., 2015; Deleger et al., 2016).

The Bacteria Biotope Task of the BioNLP Shared Task 2019 (Bossy et al., 2019) is the final version of the tasks that have been conducted until now readdressing the problem of extraction of the information regarding the bacteria biotopes. This year's task has presented the opportunity to the participants to develop solutions for three subproblems: normalization (BB-norm), relation extraction (BB-rel), and knowledge base extraction (BB-kb). For the BB-norm task of the Bacteria Biotope Task of the BioNLP Shared Task 2019, the participants are expected to develop systems to link the named entities (*Microorganism*, *Habitat*, and *Phenotype*) in a given text through a given ontology, when the entities are given with their boundaries. For instance, the sample sentence *"Atypical mycobacteria causing non-pulmonary disease in Queensland."* consists of the following mentions: *"mycobacteria"* microorganism mention, *"causing non-pulmonary disease"* phenotype mention, and *"pulmonary"* habitat mention, which should be normalized to the *"Mycobacteria"* term in the NCBI taxonomy, and *"human pathogen"* and *"lung"* terms in the Onto-Biotope ontology, respectively. For the BB-rel task of the Bacteria Biotopes Task of the BioNLP Shared Task 2019, the participants are required to extract the relations between the entities when the entities are given. There are two types of relations: *Lives_in* relation, which indicates a localization relation between a *Microorganism* entity and a *Habitat/Geographical* entity, and *Exhibits* relation, which indicates a property relation between a *Phenotype* entity and a *Microorganism* entity. For instance, the sample sentence above indicates two relations: a *Lives_in* relation between the *"Mycobacteria" Microorganism* entity and the *"Queensland" Geographical* entity, and an *Exhibits* relation between the *"Mycobacteria" Microorganism* entity and the *"causing non-pulmonary disease" Phenotype* entity.

We participated at the Bacteria Biotope Task in the BioNLP Shared Task 2019 with our system (named as the BOUN-ISIK system) and ob-

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 150–157

tained promising results in the official evaluation. This paper presents our participating system for two sub-tasks: one for the BB-norm (Entity Normalization) sub-task and one for the BB-rel (Relation Extraction) sub-task. For the entity normalization sub-task, we utilized word embeddings and syntactic re-ranking to normalize the entities. On the other hand, for the relation extraction sub-task, we proposed a rule-based method. Although both systems are unsupervised, they achieved promising results. For the BB-norm sub-task, the official results of our system achieved state-of-the-art results on the BioNLP Shared Task 2019 Bacteria Biotope task test data set. The results have shown that our unsupervised approach, which does not require labeled data, performs as good as the deep learning based methods, which require labeled data.

1.1 Related Work

1.1.1 Named Entity Normalization

Among the previous series (2011, 2013, 2016) of the BioNLP Shared Task, the Bacteria Biotope Task in 2013 is the first shared task that addressed the problem of normalization of the entities in the bacteria biotopes domain. In 2013, the participant teams proposed rule-based methods and similarity-based methods. According to the official results of the Bacteria Biotope Task of 2013, for the habitat mention normalization, the best precision was obtained by the BOUN system, which utilized syntactic rules and shallow linguistic knowledge (Karadeniz and Ozgür, 2013; Karadeniz and Özgür, 2015).

In the following series of the Bacteria Biotopes task, the habitat mention normalization sub-task continued to attract the attention of the researchers. In the Bacteria Biotope task of the BioNLP Shared Task 2016, the best precision for the habitat normalization task was obtained by the BOUN system, which utilized both approximate string matching and cosine similarity of word-vectors weighted with Term Frequency-Inverse Document Frequency (TF-IDF) (Tiftikci et al., 2016).

After the Shared Tasks, the researchers continued to search for a solution for the problem of Bacteria Biotopes normalization (Ferré et al., 2017; Mehryary et al., 2017; Karadeniz and Özgür, 2019). Although promising results have been obtained by these approaches, the results showed that there is still room for improvement for the normalization task of bacteria biotopes.

Besides the bacteria biotopes, there exist a significant amount of prior work on biomedical named entity normalization for different types of biomedical entities including genes/proteins (Morgan et al., 2008; Hakenberg et al., 2008; Wermter et al., 2009; Lu et al., 2011; Wei and Kao, 2011) and diseases (Leaman et al., 2013; Li et al., 2017). However, the need for manually annotated training data makes the adaptation of such methods to new entities difficult.

1.1.2 Relation Extraction

Several approaches, which consider the extraction of relations between various biomedical entities such as protein/protein (Giuliano et al., 2006; Airola et al., 2008; Choi, 2018), drug/drug (Segura-Bedmar et al., 2011; Kim et al., 2015), and gene/disease (Bravo et al., 2015) from biomedical text, have been presented in the literature. Relation extraction in the bacteria biotopes domain has also attracted considerable attention owing to the BioNLP Bacteria Biotope Shared Tasks.

Previous work in the bacteria biotopes domain consists of the extraction of relations between bacteria entities and habitat entities (Localization Relation Extraction) and of relations between two habitat entities (Part Of Relation Extraction). The participants of the BioNLP Shared Task 2011, which is the first shared task that addressed the relation extraction task of bacteria biotopes, utilized both machine learning and rule-based approaches for detecting the Localization and Part-of relations among bacteria and habitats (Bossy et al., 2011).

Sub-task 2 of the Bacteria Biotope (BB) Task in the BioNLP Shared Task 2013 also gave another opportunity to scientists to address the task of extracting the Localization and Part Of relations in the bacteria biotopes domain. For this subtask, the best F-score (42%) was obtained by the TEES 2.1 system (Björne and Salakoski, 2013), which used support vector machine classification. After the shared task, a new sentence-level co-occurrence approach with an anaphora resolution component in order to handle relations that span multiple sentences has been developed in (Karadeniz and Özgür, 2015), which resulted in an improved F-score performance of 53% on Sub-task 2.

In the BioNLP Shared Task 2016, the VERSE

team (Lever and Jones, 2016) achieved the best F-score, which is 56%, on the relation extraction sub-task of Bacteria Biotopes by utilizing support vector machines.

2 Data Set

The data set, which was created by collecting titles and abstracts related to microorganisms from PubMed and extracts from full-text articles related to microorganisms living in food products, is provided by the BioNLP Shared Task 2019 BB Task organizers to the participants. The data set, consisting of 132 training, 67 development, and 97 test documents, was annotated by the bioinformaticians of the Bibliome team of MIG Laboratory at the Institut National de Recherche Agronomique (INRA).

For the training and development phases of BB-norm, document texts with manually annotated named entities and the concepts assigned to them through the OntoBiotope ontology (INRA, 2013) and NCBI taxonomy (NCBI, 2018) were provided, while in the test phase, only the entity boundaries and the entity types were given by the task organizers.

For the training and development phases of BB-rel, document texts with manually annotated *Microorganism*, *Habitat*, *Phenotype* and *Geographical* entities, as well as the *Lives_in* and *Exhibits* relations were provided, while in the test phase, document texts annotated only for *Microorganism*, *Habitat*, *Phenotype* and *Geographical* entities were given.

Since our system for the named entity normalization and relation extraction of bacteria biotopes is based on unsupervised approaches and does not require any labeled training data, the errors of the developed system are analyzed on the provided training and the development sets. The test set is used for the evaluation of the performance of the system.

3 Named Entity Normalization

In this section of the paper, the utilized methods for the BB-norm task are explained in detail. The BB-norm task includes the normalization of *Habitat* entities and *Phenotype* entities in a given set of documents through the Onto-Biotope ontology and the normalization of *Microorganism* entities through the NCBI Taxonomy.

The methods developed for the normalization of the named entities can be categorized into two according to the type of the entities: Habitat and Phenotype Normalization and Microorganism Normalization.

3.1 Habitat and Phenotype Entities

For the normalization of semantically meaningful entities such as Habitat and Phenotype entities, a two-step approach that we have previously proposed in (Karadeniz and Özgür, 2019) is adapted to this new data set. According to this approach, for the normalization of an entity mention, the top k semantically most similar ontology concepts are found at the first step using the word embedding representations of the entity mention and the ontology concepts. At the second step, these top k semantically most similar concepts are re-ranked according to a similarity metric that utilizes the constituency parses of the entity mention and ontology concept phrases. The resulting most similar ontology concept is assigned as the normalized concept for the corresponding mention. The details of this approach are explained in the following subsections.

3.1.1 Named Entity and Ontology Concept Representations

In the pre-processing step, the named entity mentions and the ontology concept names are tokenized, and the stop-words are removed from the mentions and the ontology concept names.

The intuition behind the adapted method is that semantically similar words have similar word vectors. Following this intuition, the semantic similarity between named entity mentions and ontology concept terms would be higher for the similar pairs, and lower for the dissimilar pairs, if the words can be converted into a machine processable format such as real-valued vectors.

After pre-processing, to convert each word into a real-valued vector, we utilized a pre-trained word embedding model (Chiu et al., 2016), which has been trained on PubMed by using the Word2Vec tool (Mikolov et al., 2013). The corresponding word vectors are obtained for each word by using this previously trained model. For the multi-word named entity mentions and ontology concept terms, the vector representations are obtained by averaging the real-valued vectors of their composing words.

3.1.2 Semantic Filtering

After the vector representations are obtained for each entity mention and for each ontology concept term, the semantic similarity between each pair is computed by using the cosine similarity. For each entity mention, the top k most similar ontology concepts are retained as candidates for further processing, i.e., for syntactic weighting based re-ranking. k is chosen as 5 based on the results obtained in our previous study (Karadeniz and Özgür, 2019).

3.1.3 Syntactic Re-ranking

For our re-ranking approach, the assumption is that the entity mentions are noun phrases and the most informative words in the mentions are the heads of the noun phrases. We used the Stanford Parser (version 3.8.0) (Klein and Manning, 2003) to obtain the corresponding head words of the entity mentions by providing the entity mentions as input and extracting the syntactic parses of the mentions as output. Next, the top level rightmost *"noun"* is searched in the tree structured syntactic parse and assigned as the head of the mention phrase.

The semantic similarities are recomputed using the mathematical formulation shown in Equation (1), which considers also the similarity between the head words of the entity mention and ontology concept pair. In Equation (1), $S_{RR}\ (m, c)$ is the final computed similarity between mention m and the candidate concept c, and S_S is the semantic similarity, in which m_{head} is the head word of the mention m and c_{head} is the head word of the concept c, $S_S\ (m, c)$ is the similarity between mention m and concept c computed as described in Section 3.1.1 , and w is a weighting parameter which can take values between 0 and 1. w is chosen as 0.25 based on the results reported in our previous study (Karadeniz and Özgür, 2019).

$$S_{RR}\ (m, c) = \\ (w * S_S(m_{head}, c_{head})) + ((1\text{-}w) * S_S(m, c)) \quad (1)$$

3.2 Microorganism Entities

The normalization of *Microorganism* entities component of our system is based on exact matching against the names and synonyms of the concepts in the NCBI taxonomy. Error analysis on the training and developments data sets revealed that applying some rules may improve the results. For instance, *"Escherichia coli"* has an exact match that can be successfully normalized to the referent concept with an ID *"562"* in the NCBI taxonomy. In the following parts of the document, although the *"E. coli"* mention indicates a clear reference to the same concept, it can not be normalized to the *"Escherichia coli"* concept with an exact matching approach. In this kind of cases, if an exact match does not exist, the previously mentioned similar entities in the text are searched. If a match is found, the same concept is assigned as the normalized concept for the corresponding mention *"E. coli"*. If there does not exist a match with the previously normalized concepts, the root concept with an ID *"2"* is assigned.

4 Relation Extraction

4.1 Localization Relation Extraction

Our system for the relation extraction sub-task is based on the naive assumption that the related entities for most of the relations appear within the same sentence. Therefore, firstly, the input texts are split into sentences using the NLTK library. For the extraction of *Lives_in* relations, all the sentences in the related document are searched to determine whether there exists a *Microorganism* entity and a *Habitat* entity or a *Microorganism* entity and a *Geographical* entity in the corresponding sentence. If there exists such a pair, this will be a sign of a *Lives_in* relation.

For any given sentence, there can be more than one *Habitat* entity and *Microorganism* entity. For this kind of sentences, two different approaches, which are called **smart matching** and **distributed matching**, are applied. In smart matching, each *Habitat* entity is paired with the closest *Microorganism* entity. In other words, the locations of each type of entities in the sentences are checked, and then the pairing process of the *Microorganism* and the *Habitat* entities are done based on the proximity criteria. In distributed matching, on the other hand, each *Habitat* entity is paired with every *Microorganism* entity in the sentence. Distributed matching can be seen as a type of N x N matching, while smart matching 1 x 1 matching. The performance of each approach is tested on the development data set. While there is slight increase in the precision, the recall is observed to decrease considerably for the smart matching method (see Table 1). As a result, the distributed matching approach is used in the final submission.

Table 1: Distributed vs Smart Matching for relation extraction. Precision, Recall, F-measure values for the development data set are reported.

	Distributed Matching	Smart Matching
Precision	0.491	0.576
Recall	0.785	0.515
F-measure	0.604	0.544

For the overlapping entities in which one entity contains another, some relations can be ignored. For instance, for the sample sentence *"An example of this fact is the presence of Psychrobacter DNA on the surface of Formaggio di Fossa cheeses"*, the *Habitat* entity *"surface of Formaggio di Fossa cheeses"*, *Habitat* entity *"Formaggio di Fossa cheeses"*, and *Habitat* entity *"cheeses"* are overlapping entities. In this case, it would not be appropriate to build three relations such as *"Psychrobacter"* - *"surface of Formaggio di Fossa cheeses"*, *"Psychrobacter"* - *"Formaggio di Fossa cheeses"*, and *"Psychrobacter"* - *"cheeses"*. Instead of extracting multiple relations, *"cheeses"* can be ignored and two relations between *"Psychrobacter"* - *"surface of Formaggio di Fossa cheeses"* and *"Psychrobacter"* - *"Formaggio di Fossa cheeses"* are extracted. This strategy, where the shortest overlapping entity is ignored, is called as the **soft filter** operation. On the other hand, the strategy when only the longest overlapping entity is retained and the remaining ones are ignored, is named as the **hard filter** operation. In hard filtering, *"Psychrobacter"* - *"Formaggio di Fossa cheeses"* and *"Psychrobacter"* - *"cheeses"* are ignored and only one relation between *"Psychrobacter"* - *"surface of Formaggio di Fossa cheeses"* is extracted. The performance of each approach is tested on the development data set (see Table 2).

Table 2: Soft Filter vs Hard Filter for relation extraction. Precision, Recall, F-measure values for the development data set are reported.

	Soft Filter	Hard Filter
Precision	0.584	0.575
Recall	0.768	0.639
F-measure	0.616	0.561

Since our rule-based system for relation extraction is based on the assumption that most of the relations appear within the same sentences, our system is not able to catch the relations that cross sentence boundaries. To overcome this problem, a new rule, which is called **remote matching**, is integrated into the system. According to this rule, if there exists only one entity type (Microorganism) in a sentence, and within a context window of three sentences there exists only one entity (Habitat or Geographical), then there is a relation between these two entities. The performance of the remote matching rule is tested on the development data set. The results show that the number of the predicted relations increased, which also led to an increase in recall. The obtained precision and recall values are 51.4% and 78.5%, respectively.

4.2 Exhibits Relation Extraction

Similar to the extraction of localization relations, for the extraction of *Exhibits* relations, all the sentences are searched for whether there exist a *Microorganism* entity and a *Phenotype* entity. The same rules that are explained in the previous subsection are applied for the extraction of the *Exhibits* relations.

5 Evaluation

In the BioNLP Shared Task 2019 Bacteria Biotopes normalization sub-task, entities are given with their boundaries in the text and the participants are required to predict the normalization of the entities. In the official evaluation, for each normalized Habitat/Phenotype entity, Wang similarity W (Wang et al., 2007) is calculated to measure the similarity between the reference concept and the predicted concept for the normalization. The performances of the submitted systems are evaluated with their Precision values, which are calculated as:

$$Precision = \sum S_p / N \quad (2)$$

where S_p indicates the total Wang similarity W for all predictions (Deleger et al., 2016), and N is the number of predicted entities.

In the BioNLP Shared Task 2019 Bacteria Biotopes relation extraction sub-task, entities are given with their boundaries in the text and the participants are asked to predict the relations between the entities. The performances of the submitted systems are evaluated with their F1 (F-measure), recall and precision values.

5.1 Results of BB-norm

The official results obtained by our system and the other participants for the BB-norm sub-task are shown in Table 3. Our system (BOUN-ISIK-2) achieved the best performance with 67.9% Precision in the BB-norm sub-task (Entity Normalization).

Table 3: Comparison with the participant systems for the normalization task of bacteria biotopes. Precision values for the test data set are reported. k is set to 5 and w to 0.25 for the proposed system (BOUN-ISIK).

System	Precision
BOUN-ISIK-2 (Our system)	0.679
BLAIR_GMU-2	0.678
BOUN-ISIK-1 (Our system)	0.675
BLAIR_GMU-1	0.661
PADIA_BacReader-1	0.633
BASELINE-1	0.531
AmritaCen_healthcare-1	0.514

As the results in Table 4 demonstrate, our system performs significantly better than the other systems for the normalization of new Phenotype entities in the test set (Precision: 70.8%).

Table 4: Comparison with the participant systems for the normalization task considering only Phenotype entities. Precision values for the test data set are reported.

System	Phenotypes	Phenotypes (new in test)
BOUN-ISIK (Our system)	0.566	**0.708**
PADIA_BacReader-1	**0.758**	0.156
BASELINE-1	0.582	0.116
BLAIR_GMU-2	0.646	0.03
BLAIR_GMU-1	0.628	0.03
AmritaCen_healthcare-1	0.646	0.0

5.2 Results of BB-rel

The official results obtained by our system and the other participants for the BB-rel task are demonstrated in Table 5.

Table 5: Comparison with the participant systems for the relation extraction task of bacteria biotopes. F1, Recall and Precision values for the test data set are reported.

System	F1	Recall	Precision
whunlp-1	0.664	0.702	0.629
AliAI-1	0.650	0.620	0.682
BASELINE-1	0.635	0.801	0.525
Yuhang_Wu-1	0.605	0.670	0.551
BOUN-ISIK-1 (soft filter)	0.604	0.731	0.514
BLAIR_GMU-2	0.594	0.650	0.548
BOUN-ISIK-2 (hard filter)	0.575	0.601	0.552
BLAIR_GMU-1	0.549	0.496	0.617
UTU-2	0.550	0.474	0.655
UTU-1	0.529	0.428	0.694
Amrita_Cen-1	0.499	0.617	0.419
Amrita_Cen-2	0.493	0.610	0.414

6 Conclusion

In this study, we presented two systems that are implemented in the scope of the BioNLP Shared Task 2019 - Bacteria Biotope Task. The aim of the first system is the normalization of the entity mentions in a biomedical text through the corresponding ontology, whereas the goal of the second system is the extraction of localization and property relations between the related entities when the entities are given. Both systems are unsupervised in the sense that they do not require domain-specific labeled data, while the normalization system makes use of word embeddings and syntactic re-ranking. According to the official evaluation, both of our systems achieved promising results, which have shown that the proposed methods are comparable to or better than the labeled data driven deep learning based approaches used in the shared task.

Acknowledgments

We would like to thank the BioNLP shared task organizers, especially, Robert Bossy for their help with the questions.

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 1–9. Association for Computational Linguistics.

Jari Björne and Tapio Salakoski. 2013. Tees 2.1: Automated annotation scheme learning in the bionlp 2013 shared task. In *Proceedings of the BioNLP shared task 2013 workshop*, pages 16–25.

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nedellec. 2019. Bacteria Biotope at BioNLP Open Shared Tasks 2019. In *Proceedings of the BioNLP Open Shared Tasks 2019 Workshop*.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessieres, and Claire

Nédellec. 2015. Overview of the gene regulation network and the bacteria biotope tasks in bionlp'13 shared task. *BMC bioinformatics*, 16(10):S1.

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.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):55.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. *Proceedings of BioNLP16*, page 166.

Sung-Pil Choi. 2018. Extraction of protein–protein interactions (ppis) from the literature by deep convolutional neural networks with various feature embeddings. *Journal of Information Science*, 44(1):60–73.

Louise Deleger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferre, Philippe Bessieres, and Claire Nedellec. 2016. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22.

Arnaud Ferré, Pierre Zweigenbaum, and Claire Nédellec. 2017. Representation of complex terms in a vector space structured by an ontology for a normalization task. *BioNLP 2017*, pages 99–106.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder, and Graciela Gonzalez. 2008. Inter-species normalization of gene mentions with gnat. *Bioinformatics*, 24(16):i126–i132.

INRA. 2013. *Onto-Biotope Ontology*. Accessed at December 2018.

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(10):S5.

Ilknur Karadeniz and Arzucan Özgür. 2019. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC bioinformatics*, 20(1):156.

Sun Kim, Haibin Liu, Lana Yeganova, and W John Wilbur. 2015. Extracting drug–drug interactions from literature using a rich feature-based linear kernel approach. *Journal of biomedical informatics*, 55:23–30.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Jake Lever and Steven JM Jones. 2016. Verse: Event and relation extraction in the bionlp 2016 shared task. In *Proceedings of the 4th BioNLP shared task workshop*, pages 42–49.

Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18(11):385.

Zhiyong Lu, Hung-Yu Kao, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hong-Jie Dai, Naoaki Okazaki, et al. 2011. The gene normalization task in biocreative iii. *BMC bioinformatics*, 12(8):S2.

Farrokh Mehryary, Kai Hakala, Suwisa Kaewphan, Jari Björne, Tapio Salakoski, and Filip Ginter. 2017. End-to-end system for bacteria habitat extraction. *BioNLP 2017*, pages 80–90.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. 2008. Overview of biocreative ii gene normalization. *Genome biology*, 9(2):S3.

NCBI. 2018. *NCBI Taxonomy*. Accessed at December 2018.

Isabel Segura-Bedmar, Paloma Martinez, and Cesar de Pablo-Sánchez. 2011. Using a shallow linguistic kernel for drug–drug interaction extraction. *Journal of biomedical informatics*, 44(5):789–804.

Mert Tiftikci, Hakan Şahin, Berfu Büyüköz, Alper Yayıkçı, and Arzucan Özgür. 2016. Ontology-based categorization of bacteria and habitat entities using information retrieval techniques. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 56–63.

James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. 2007. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281.

Chih-Hsuan Wei and Hung-Yu Kao. 2011. Cross-species gene normalization by species inference. *BMC bioinformatics*, 12(8):S5.

Joachim Wermter, Katrin Tomanek, and Udo Hahn. 2009. High-performance gene name normalization with geno. *Bioinformatics*, 25(6):815–821.