

ONTOLOGY-BASED ENTITY TAGGING AND NORMALIZATION IN THE  
BIOMEDICAL DOMAIN

by

Zeynep İlknur Karadeniz Erol

B.S., Department of Computer Engineering, Yeditepe University, 2004

M.S., Department of Computer Engineering, Istanbul Technical University, 2007

Submitted to the Institute for Graduate Studies in  
Science and Engineering in partial fulfillment of  
the requirements for the degree of  
Doctor of Philosophy

Graduate Program in Computer Engineering

Boğaziçi University

2019

## ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Assoc. Prof. Dr. Arzucan Özgür not only for her continuous support, patience, motivation, knowledge but also for her kind and positive attitude from the beginning to the end. Her guidance helped me in all the time of research and writing of this thesis. I feel so lucky that I had a chance to work with her.

Besides my advisor, I would like to thank the rest of my thesis progress committee: Prof. Dr. Tunga Güngör, Assoc. Prof. Dr. A. Cüneyd Tantı, Prof. Dr. Lale Akarun, Prof. Dr. Olcay Taner Yıldız for their insightful comments, encouragement, and suggestions.

This research is supported by Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme and by Turkish State Planning Organization (DPT) under the TAM Project, number 2007K120610. I want to specially thank Prof. Dr. Ufuk Çağlayan, Prof. Dr. Lale Akarun, Prof. Dr. Cem Ersoy, and all TAM project members. Without their precious support, it would not have been possible to conduct this research. I would like to acknowledge BÜVAK for providing financial support to attend conferences.

My sincere thanks also go to all contributors especially Prof. Dr. Kutlu Ülgen, Dr. Saliha Durmuş Tekir, Dr. Junghuk Hur, and Assoc. Prof. Dr. Yongqun He for their considerable contributions. I would also like to thank the BioNLP Shared Task Bacteria Biotope organizers especially Robert Bossy and Claire Nedellec for their answers to the questions about the data set and the evaluation tool.

I want to thank all the instructors of the courses that I have taken during my PhD, and all my inspiring teachers throughout my education especially to Prof. Dr. Eşref Adalı. I also want to thank Dr. Suzan Üsküdarlı for the motivative talks we had.

I thank my fellow labmates for the stimulating discussions, and for all the fun we have had in the six years. I especially thank Arda Çelebi for sharing his computer resources with me, Mert Tiftikçi and Berfu Büyüköz for their supports for BOUNEL, and Hakime Öztürk for her feedback about the algorithms.

Last but not the least, I would like to thank my family. I gratefully thank my mother, Zehra for her unbelievable support throughout my education; my father, Ali İhsan who inspires me as an engineer; and to both of them for their condition-less love throughout my life. I thank my two-year-old baby girl, Ada for her miracle presence and patience, and for teaching me to use the time wisely. I thank my true love, Emre for being such a supportive husband and interested father with our daughter. I would like to thank my brother, Emre for all the fun we have had all the times. I would like to thank my mother-in-law Saadet, my father-in-law Mehmet, and my sister-in-law Ece for their love and sensibility. I also would like to express my very special thanks to my cousins Aynur for being a real sister, and Aydın for being a real brother.

## ABSTRACT

# ONTOLOGY-BASED ENTITY TAGGING AND NORMALIZATION IN THE BIOMEDICAL DOMAIN

One of the challenges for scientists in the biomedical domain is the huge amount and the rapid growth of information buried in the text of electronic resources. Developing text mining methods to automatically extract biomedical entities from the text of these electronic resources and identifying the relations between the extracted entities is crucial for facilitating research in many areas in the biomedical domain. Two main problems, which have to be solved to accomplish this goal, are the extraction and normalization of entities, and the identification of the relations between them from a given text.

In this thesis, we proposed two approaches with two different perspectives for the extraction and normalization of biomedical named entities. The first approach makes use of shallow linguistic knowledge to extract entities and normalize them through an ontology. On the other hand, the second approach makes use of word embeddings, which convey semantic information, for the normalization of the entities in a given text. The word-embedding based approach obtained the state-of-the-art results on the BioNLP Shared Task 2016 Bacteria Biotope data set. Both of the proposed methods are unsupervised and can be adapted to different domains. We also developed two applications, one of which is a pipeline, which is composed of modules based on the approaches that we proposed in this thesis, for the extraction of bacteria biotope information from scientific abstracts. The other application is developed for extracting Brucella-host interaction relevant data from the biomedical literature, whose results reveal the importance of using a wider context than a sentence for biomedical relation extraction.

## ÖZET

### BİYOMEDİKAL ALANDA ONTOLOJİ TABANLI VARLIK İSMİ ETİKETLEME VE NORMALİZASYONU

Biyomedikal alandaki zorluklardan biri, elektronik kaynakların ve bu kaynaklardaki gömülü bilgilerin fazla olması ve hızla artmaya devam etmesidir. Biyomedikal varlıkların isimlerini bu elektronik kaynaklardaki metinlerde otomatik olarak belirlemek için metin madenciliği yöntemleri geliştirmek ve bu varlıklar arasındaki ilişkileri belirlemek, birçok alandaki araştırmayı kolaylaştırmak için çok önemlidir. Bu hedefe ulaşmak için çözülmesi gereken iki ana sorun, belirli bir metindeki varlık isimlerinin belirlenmesi ile normalizasyonu ve bu varlıkların arasındaki ilişkilerin tanımlanmasıdır.

Bu tezde, biyomedikal alandaki varlık isimlerinin metinlerden çıkarılması ve normalizasyonu için iki farklı bakış açısına sahip iki yeni yaklaşım önerilmiştir. Birinci yaklaşımda, metinlerdeki varlık isimlerini belirlemek ve onların bir ontoloji yoluyla normalizasyonunu sağlamak için sığ dilbilimsel bilgiden yararlanılmıştır. Öte yandan, ikinci yaklaşımda, metindeki varlık isimlerinin normalizasyonu için anlamsal bilgi içeren sözcük gömme işlemleri kullanılmıştır. Sözcük gömme temelli yaklaşım, BioNLP 2016 Bakteri Biyotop veri kümesi üzerinde mevcut yöntemlerden daha başarılı sonuçlar elde etmiştir. Önerilen yöntemlerin her ikisi de denetimsizdir ve farklı alanlara uyarlanabilir. Ayrıca bu tezde, iki ayrı uygulama sunulmuştur. Birinci uygulama, bakterilerin biyotop bilgilerinin bilimsel özetlerden çıkarılması için önerdiğimiz yaklaşımlara dayanan modüllerden oluşan bir sistemdir. Diğer uygulama ise, biyomedikal literatürden Brusellakonak etkileşimi ile ilgili verileri çıkarmak için geliştirilmiştir; bu uygulamanın sonuçları, biyomedikal ilişki çıkarımı için bir cümleden daha geniş bir bağlam kullanmanın önemini ortaya koymaktadır.

# TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . .	iii
ABSTRACT . . . . .	v
ÖZET . . . . .	vi
LIST OF FIGURES . . . . .	xi
LIST OF TABLES . . . . .	xiv
LIST OF SYMBOLS . . . . .	xvii
LIST OF ACRONYMS/ABBREVIATIONS . . . . .	xviii
1. INTRODUCTION . . . . .	1
1.1. Problem Statement . . . . .	2
1.2. Challenges . . . . .	5
1.3. Motivation . . . . .	6
1.4. Publication Notes . . . . .	7
1.5. Thesis Overview . . . . .	8
2. BACKGROUND . . . . .	10
2.1. Related Tasks . . . . .	10
2.1.1. Named Entity Recognition . . . . .	10
2.1.2. Named Entity Normalization . . . . .	13
2.2. Related Techniques . . . . .	14
2.2.1. Word Embeddings . . . . .	14
2.2.2. Support Vector Machines . . . . .	16
2.2.2.1. Cosine Kernel . . . . .	17
2.2.2.2. Edit Kernel . . . . .	18
3. ENTITY TAGGING AND NORMALIZATION USING SHALLOW LINGUISTIC ANALYSIS . . . . .	19
3.1. Entity Detection using Exact String Matching . . . . .	19
3.1.1. Background . . . . .	19
3.1.2. Methods and Materials . . . . .	20
3.1.2.1. Dictionary of Experimental Methods . . . . .	20
3.1.2.2. Abstract Retrieval Module . . . . .	20

3.1.2.3.	Exact Matching Algorithm . . . . .	21
3.1.3.	Results and Discussion . . . . .	22
3.2.	Entity Detection and Normalization using Rules based on Shallow Lin- guistic Analysis . . . . .	23
3.2.1.	Background . . . . .	24
3.2.2.	Related work . . . . .	26
3.2.3.	Methods . . . . .	29
3.2.3.1.	Preprocessing . . . . .	29
3.2.3.2.	Ontology expansion from the training data . . . . .	30
3.2.3.3.	Noun phrase extraction and simplification . . . . .	30
3.2.3.4.	Discontinuous entity handling . . . . .	31
3.2.3.5.	Entity modifier handling . . . . .	32
3.2.3.6.	Ontology mapping . . . . .	33
3.2.4.	Results for the BioNLP Shared Task 2013 Data Set . . . . .	35
3.2.4.1.	Data set . . . . .	35
3.2.4.2.	Evaluation metrics . . . . .	35
3.2.4.3.	Results . . . . .	37
3.2.5.	Results for the BioNLP Shared Task 2016 Data Set . . . . .	39
3.2.5.1.	Data Set . . . . .	39
3.2.5.2.	Evaluation Metrics . . . . .	40
3.2.5.3.	Results . . . . .	40
4.	ENTITY NORMALIZATION USING WORD EMBEDDINGS AND SYNTAC- TIC ANALYSIS . . . . .	42
4.1.	Background . . . . .	42
4.2.	Related Work . . . . .	43
4.3.	Methods . . . . .	46
4.3.1.	Data Sets . . . . .	47
4.3.1.1.	Bacteria Biotope Entity Normalization . . . . .	47
4.3.1.2.	Adverse Drug Reaction Normalization . . . . .	48
4.3.2.	Preprocessing . . . . .	48
4.3.3.	Word representations . . . . .	48
4.3.4.	Identifying the Semantically Similar Ontology Concepts . . . . .	49

4.3.5.	Syntactic Re-ranking . . . . .	50
4.4.	Results and Discussion . . . . .	54
4.4.1.	Evaluation Metrics . . . . .	54
4.4.1.1.	Evaluation for Bacteria Biotopes . . . . .	54
4.4.1.2.	Evaluation for Adverse Drug Reaction . . . . .	55
4.4.2.	Results . . . . .	56
4.4.2.1.	Bacteria Biotopes . . . . .	56
4.4.2.2.	Adverse Drug Reactions . . . . .	58
4.4.3.	Discussion . . . . .	58
4.4.3.1.	Bacteria Biotopes . . . . .	58
4.4.3.2.	Adverse Drug Reactions . . . . .	61
5.	APPLICATIONS . . . . .	63
5.1.	An application for the Bacteria Biotopes domain . . . . .	63
5.1.1.	Retrieval of the related abstracts . . . . .	63
5.1.2.	Preprocessing . . . . .	63
5.1.3.	Named Entity Tagging . . . . .	63
5.1.4.	Named Entity Normalization . . . . .	64
5.1.5.	Relation Extraction . . . . .	64
5.1.5.1.	Related Work . . . . .	65
5.1.5.2.	Methods . . . . .	67
5.1.5.3.	Results . . . . .	71
5.2.	An application for Brucella-Host Relevant Interaction Extraction . . . . .	74
5.2.1.	Motivation . . . . .	74
5.2.2.	Methods and Materials . . . . .	78
5.2.2.1.	Data set collection . . . . .	79
5.2.2.2.	Identifying gene names . . . . .	79
5.2.2.3.	Mapping genes to pathogen and host species . . . . .	80
5.2.2.4.	Gene-gene interaction extraction . . . . .	80
5.2.2.5.	Ontology modeling . . . . .	82
5.2.3.	Results and discussion . . . . .	83
5.2.3.1.	Results . . . . .	83
5.2.3.2.	Discussion . . . . .	86



6. CONCLUSIONS . . . . .	89
6.1. Discussion . . . . .	89
6.2. Future Work . . . . .	92
REFERENCES . . . . .	94

## LIST OF FIGURES

Figure 1.1.	Exponential growth of the number of published articles in Pubmed	1
Figure 1.2.	Exponential growth of the number of published articles related to bacteria in Pubmed . . . . .	2
Figure 1.3.	Sample text with annotated named entities . . . . .	2
Figure 1.4.	Sample text. Sample abstract of [1] with habitat entity mentions annotated . . . . .	3
Figure 1.5.	Sample ontology. A sample portion from the Onto-Biotope ontology	3
Figure 3.1.	Sample portion of our experimental method dictionary . . . . .	21
Figure 3.2.	Sample abstract . . . . .	21
Figure 3.3.	Pseudo-code of Exact Matching Algorithm . . . . .	22
Figure 3.4.	Sample PHI data without experimental method information . . . .	22
Figure 3.5.	Sample text. A sample input file containing bacteria and habitat entities. . . . .	25
Figure 3.6.	Workflow of the Sub-task 1 System . . . . .	29
Figure 3.7.	Sample output of the preprocessing, and the noun phrase extractor and simplifier . . . . .	30

Figure 3.8.	Discontinuous entity handling for the sample phrase “ <i>pharyngeal and gut mucosa</i> ” . . . . .	32
Figure 3.9.	Continuous entity handling for the sample phrase “ <i>iron-rich and wet environment</i> ” . . . . .	33
Figure 3.10.	Ontology mapping example . . . . .	34
Figure 4.1.	System Work-flow. Work-flow of the Named Entity Normalization System . . . . .	47
Figure 4.2.	Sample multi-word expression. Computation of the corresponding real-value vector for a sample multi-word expression “ <i>a day-care center</i> ”, where $\vec{e}(t)$ is the word embedding vector for token $t$ . . .	50
Figure 4.3.	Sample syntactic parse Syntactic parse of the Stanford Parser for the sample named entity mention “ <i>children attending a day-care center</i> ” . . . . .	52
Figure 4.4.	Tree view of the sample parse Tree view of the syntactic parse of the sample named entity mention “ <i>children attending a day-care center</i> ” . . . . .	52
Figure 4.5.	Pseudo-code Algorithm for finding the most informative word in an entity mention whose syntactic parse is given as input. NP: Noun Phrase; NN: Noun singular; NNS: Noun plural ;NNP: Proper noun singular; NNPS: Proper Noun plural . . . . .	53
Figure 5.1.	Workflow of the Sentence-based Sub-task 2 System . . . . .	69
Figure 5.2.	Sample host-pathogen interaction describing sentence. . . . .	76

Figure 5.3.	Workflow of the host-Brucella interaction extraction approach. . .	78
Figure 5.4.	The dependency parse tree of a sample sentence. . . . .	82
Figure 5.5.	Literature-mined host-Brucella gene-gene interaction results. . . .	85
Figure 5.6.	The ontology hierarchy of literature mined INO interaction types.	86

## LIST OF TABLES

Table 3.1.	Detailed results on the test set for Sub-task 1 ( <i>Entity Boundary Detection &amp; Ontology Categorization</i> ) . . . . .	37
Table 3.2.	Comparison with the other systems that participated in the BB Sub-task 1 ( <i>Entity Boundary Detection &amp; Ontology Categorization</i> ). The results obtained on the test set are reported. . . . .	38
Table 3.3.	Effect of discontinuous entity handling (DEH). The results are reported on the training, development, and test sets. . . . .	38
Table 3.4.	Effect of entity modifier handling. The results are reported on the training, development, and test sets. . . . .	39
Table 3.5.	Results for BB-cat+ner-task ( <i>Entity Recognition and Categorization</i> ) on BioNLP Shared Task 2016 Data Set. The results obtained on the test set are reported. . . . .	40
Table 3.6.	Comparison of BB-cat+ner-task Results ( <i>Entity Recognition and Categorization</i> ) on BioNLP Shared Task 2016 Data Set. The results obtained on the test set are reported. MM:Mismatches M:Matches I:Insertions D:Deletions R:Recall P:Precision Pred:Predictions . . .	41
Table 4.1.	Semantically most similar concepts to the entity mention “ <i>children attending a day-care center</i> ” with/without re-ranking. . . . .	51
Table 4.2.	Results for the system with and without syntactic re-ranking. Precision values for the training and development data sets are reported. $k$ is set to 5 and $w$ is set as 0.25 for the re-ranking module. . . .	57

Table 4.3.	Comparison with previous 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 (BOUNEL) based on the results on the training and development sets. . . . .	57
Table 4.4.	Results of the proposed method with/without re-ranking on the adverse drug reaction normalization task. Precision, recall and f-score values for the training and test sets are reported. . . . .	59
Table 4.5.	Prediction performance of our system without syntactic re-ranking among the semantically most similar top ( $k = 1, 5, 10, 20, 25, 50$ ) concepts. Precision values for the training and development data sets are reported when the reference concept is among the top $k$ . .	59
Table 4.6.	Results for the system with syntactic re-ranking for the different semantically most similar top ( $k = 5, 10, 15, 20, 25, 50$ ) concepts. Precision values for the training and development data sets are reported when the reference concept is at the first rank after re-ranking the semantically most similar top ( $k = 5, 10, 15, 20, 25, 50$ ) concepts. .	60
Table 4.7.	Results for the system with different weights for the most informative words ( $w = 0, 0.25, 0.50, 0.75$ ). Precision values for the training and development data sets are reported. . . . .	60
Table 5.1.	Results of BB Sub-task 2 ( <i>Localization and PartOf Event Extraction</i> ). The results obtained on the test set are reported. . . . .	72
Table 5.2.	Comparison with the other systems that participated in the BB Sub-task 2 ( <i>Localization and PartOf Event Extraction</i> ). The results obtained on the test set are reported. . . . .	72

Table 5.3.	Effects of Anaphora Resolution Module and Syntax Rules ( <i>Localization and PartOf Event Extraction</i> ). The results obtained on the training set are reported. . . . .	73
Table 5.4.	Effects of Anaphora Resolution Module and Syntax Rules ( <i>Localization and PartOf Event Extraction</i> ). The results obtained on the development set are reported. . . . .	73
Table 5.5.	Effects of Anaphora Resolution Module and Syntax Rules ( <i>Localization and PartOf Event Extraction</i> ). The results obtained on the test set are reported. . . . .	73
Table 5.6.	Co-occurrence and machine learning-based host-Brucella gene-gene interaction results. TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative. . . . .	83

## LIST OF SYMBOLS

$c$	candidate concept $c$
$c_{head}$	the head word of the candidate concept $c$
$D$	Number of deletions
$d$	Distance between the corresponding concept and the ancestor
$E$	a set of entities
$e$	corresponding entity in the ontology
$I$	Number of insertions
$J$	Jaccard Coefficient Similarity
$M$	Similarity between two entities
$m$	entity mention
$m_{head}$	the head word of the entity mention $m$
$N$	Total number of habitats habitats in the reference
$N_p$	Total number of predicted entities
$S$	Number of substitutions
$S_p$	total Wang similarity $W$ for all predictions
$S_{RR}$	the re-ranked similarity score
$S_S$	the semantic similarity score
$s$	A parameter which can take values between 0 and 1
$W$	Wang similarity that measures the similarity between the ontology concepts of the reference and the predicted entities
$w$	A weighting parameter which can take values between 0 and 1
$X$	a set of entity mentions
$x$	entity mention



## LIST OF ACRONYMS/ABBREVIATIONS

ADR	Adverse Drug Reaction
BB	Bacteria Biotope
BHI	Brucella-Host Interaction
BioNLP	Biomedical Natural Language Processing
BNER	Biomedical Named Entity Recognition
BNEN	Biomedical Named Entity Normalization
CBOW	Continuous Bag-of-Words
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
DEH	Discontinuous Entity Handling
FN	False Negatives
FP	False Positives
GeniaSS	Genia Sentence Splitter
HMM	Hidden Markov Models
HPI	Host-Pathogen Interaction
IMT	Interaction Method Task
kNN	k Nearest Neighbor
LSTM	Long-Short Term Memory
MedDRA	Medical Dictionary for Regulatory Activities
NE	Named Entity
NEN	Named Entity Normalization
NER	Named Entity Recognition
NLP	Natural Language Processing
NN	Noun Singular
NNS	Noun Plural
NNP	Proper Noun Singular
NNPs	Proper Noun Plural
NP	Noun Phrase
PHI	Pathogen-Host Interaction

PHISTO	Pathogen-Host Interaction Search Tool
PPI	Protein-Protein Interaction
PPP	Possessive Prepositional Phrase
prep	preposition
RNN	Recurrent Neural Networks
SER	Slot Error Rate
SIDER	Side Effect Resource
SVM	Support Vector Machines
TAC	Text Analytics Conference
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negatives
TP	True Positives
UMLS	Unified Medical Language System
WMD	Word Mover's Distance

# 1. INTRODUCTION

The number of published articles in the biomedical domain is increasing every day. PubMed is a free search engine that presents more than 28 million references and abstracts on biomedical topics [2]. In PubMed, each year, approximately 1 million new articles are being included and this number is growing exponentially (see Figure 1.1). Even for a restricted search keyword *bacteria*, the number of published articles is growing exponentially (see Figure 1.2) and approximately 100,000 new articles are being published each year. Human annotation can not keep up with this increasing amount of information, which leads to increased demand for automation.

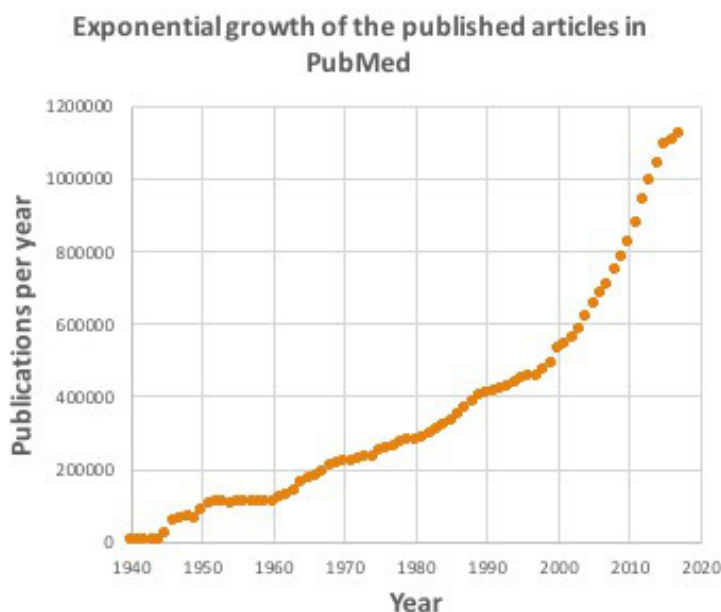


Figure 1.1. Exponential growth of the number of published articles in Pubmed

As the number of published articles is increasing rapidly each day, the need for systems that automatically extract information from biomedical literature becomes more important. Due to the reason that the majority of these information is in natural language, natural language processing (NLP) techniques are required to extract and categorize the relevant information.

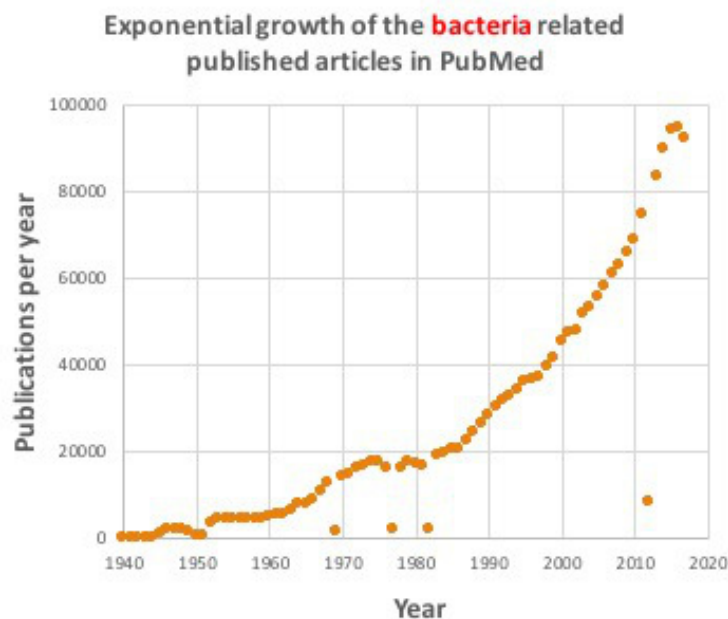


Figure 1.2. Exponential growth of the number of published articles related to bacteria in Pubmed

### 1.1. Problem Statement

The main problems tackled in this thesis are the extraction and categorization of biomedical named entities from the biomedical literature. These problems are called Named Entity Recognition (NER) and Named Entity Normalization (NEN), which are defined in detail in this section.

Boğaziçi University is a major research university located in Istanbul, Turkey. It has four faculties and two schools offering undergraduate degrees, and six institutes offering graduate degrees.

Organization Location Number

Figure 1.3. Sample text with annotated named entities

The term Named Entity (NE) was first used in the Sixth Message Understanding Conference (MUC-6), whose ultimate aim was to extract structured information such as person, organization, or location of company activities and defense activities from unstructured text. Since then, a NE is defined as a portion of any text, which defines a name of an abstract object or that of a physical object. In other words, named entities can be defined as entity instances (e.g., “Istanbul” is an instance of a city).

The task of identifying string portions with their boundaries as named entities from a given text is called Named Entity Recognition (NER) (see Figure 1.1 for a sample text with annotated named entities).

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

Figure 1.4. Sample text. Sample abstract of [1] with habitat entity mentions annotated

```
[Term]
id: OBT:002307
name: pediatric patient
is_a: OBT:002133 ! patient
is_a: OBT:002146 ! child

[Term]
id: OBT:000124
name: respiratory tract part
is_a: OBT:000065 ! animal part

[Term]
id: OBT:002146
name: child
synonym: "children" EXACT []
is_a: OBT:001804 ! human
is_a: OBT:000889 ! animal with life stage property
```

Figure 1.5. Sample ontology. A sample portion from the Onto-Biotopo ontology

In the biomedical domain, there are a variety of named entity types such as drugs, proteins, genes, species, cell types, diseases, adverse drug reactions, and bacteria biotopes. For example, “*Diaformin*” is an instance of a drug and “*blood cancer*” is an instance of a disease.

After the identification of named entities with their boundaries, a further standardization step, which is called Named Entity Normalization (NEN), is in general required to disambiguate the extracted named entities [3]. When an ontology/dictionary containing a set of entities  $E$  and a text containing a set of entity mentions  $X$  are given, NEN is the task of mapping each named entity mention  $x$  in the given text to its corresponding entity  $e$  in the given ontology/dictionary, where  $x \in X$  and  $e \in E$  [4]. This task is also called entity linking, entity grounding, or entity categorization, which are used interchangeably throughout this thesis.

In this thesis, we addressed the NER and NEN tasks, experimenting with the bacteria biotope and adverse drug reaction biomedical entity types. Figure 1.1 demonstrates a sample text with annotated bacteria habitat (biotope) mentions, which are represented in bold and Figure 1.1 demonstrates a sample portion from Onto-Biotope, which is an ontology for bacteria habitats. Given a sample text with annotated habitat mentions, the aim of habitat entity normalization is to link the mentions through the Onto-Biotope Ontology. For instance, “*pediatric*”, “*respiratory*”, and “*children less than 2 years of age*” are habitat entity mentions. The concept that is associated with the “*pediatric*” habitat mention in the Onto-Biotope ontology is “*pediatric patient*”, the one associated with the “*respiratory*” habitat mention is “*respiratory tract part*”, and for “*children less than 2 years of age*” it is “*pediatric patient*”. Entity normalization can also be performed through a dictionary. For instance, the sample sentence “*In Study 3, 67% of patients treated with ADCETRIS experienced any grade of neuropathy.*” states a relation between the drug mention “*ADCETRIS*” and adverse drug reaction mention “*neuropathy*”. The adverse drug reaction mention “*neuropathy*” can be normalized to the “*peripheral neuropathy*” term in the Medical Dictionary for Regulatory Activities (MedDRA) [5].

Named entity recognition and normalization are preliminary tasks that should be handled for many information extraction and retrieval tasks. For instance, in document retrieval, queries and documents often contain named entities, whose detection and categorization are fundamental for the success of the retrieval task. Another major information extraction task where named entity recognition and normalization is crucial is relation extraction, since the relations, which are intended to be extracted from text, occur among the named entities. The named entity recognition and normalization tasks pose many challenges, which are explained in detail in the following section.

## 1.2. Challenges

There are many challenges for named entity recognition and normalization in the biomedical domain, some of which are summarized as followings:

- Ambiguity : Two named entities with the same surface form may have different semantic meanings. These ambiguities are generally caused by abbreviations (e.g., “*ten*” may refer to a number entity “*ten*” or an adverse drug reaction entity “*toxic epidermal necrolysis*”).
- Variety : A named entity may appear in different surface forms in a given text (e.g., “*GIT*”, “*GI tract*”, “*git*”, “*intestinal tract*”, “*gastro-intestinal tract*” are all used to refer to the “*gastro intestinal tract*” named entity).
- Out of dictionary words : Construction of a dictionary is not an adequate technique due to the rapid growth of the vocabulary in the biomedical domain (e.g., “*pediatric patient*” is a habitat entity that exist in the Onto-Biotope ontology, but “*children less than 2 years-of-age with a respiratory illness*”, which does not exist in the ontology in this surface form, is another entity mention that refers to the same habitat entity, namely “*pediatric patient*”).
- Multi-word biomedical named entities : Biomedical named entities are often not single words, instead they generally consist of multiple-words in a text (e.g., “*gastrointestinal tract*” is a habitat entity, and “*Stevens-Johnson syndrome*” is an adverse drug reaction entity).

- Overlapping biomedical named entities : There may be overlapping named entities in an entity mention such as “*human gastrointestinal tract*”, where there are two overlapping habitat entities “*human gastrointestinal tract*” and “*human*”. Therefore, the detection of the boundaries of the named entities in a text is not a trivial task.
- Relatively small data sets : In the biomedical domain, the training data is relatively smaller compared to many other domains in natural language processing (e.g., in the training data set of the BioNLP Shared Task Bacteria Biotope Task, there are only 747 entity mentions).
- Relatively large number of categories : In the biomedical domain, the number of the semantic categories that should be considered is in general larger compared to many other domains in natural language processing (e.g., there are 2,221 semantic categories in the Onto-Biotope ontology and 22,499 dictionary terms in the MedDRA dictionary).

### 1.3. Motivation

Many text mining methods have been implemented to extract and categorize the biomedical named entities that are buried in biomedical texts. Most of the previous methods require manually annotated training data, which makes the adaptation to different kinds of biomedical entities difficult. Furthermore, even if they can be adapted, the performances in general drop in the adapted domain. Even in the same domain, any change such as changing the entity type, may result in decrease in performance for the NER and NEN systems. For example, considering the biomedical domain, many named entity recognition systems have been proposed for proteins and genes [6–10], diseases [11, 12], chemical compounds and drugs [13]. Nevertheless, a system that is developed for the identification of gene names, may not achieve high performances for the identification of drug names.

In this thesis, we investigate the problems of the detection of named entities mentioned in a biomedical text and the normalization of these mentions to a dictionary or an ontology. We proposed two approaches with two different perspectives for the



extraction and normalization of biomedical named entities. The first approach makes use of shallow linguistic knowledge to extract entities and normalize them through an ontology. On the other hand, the second approach makes use of word embeddings, which convey semantic information, for the normalization of entities in a text. We applied the shallow linguistic based approach to extract and normalize bacteria biotope entities, and the word-embedding based approach to normalize bacteria biotope entities and adverse drug reaction entities. Although these types of entities have been used for testing purposes of the developed methods, both of the proposed methods are unsupervised and can be adapted to different biomedical entity types, since they do not require entity-specific manually annotated data.

#### 1.4. Publication Notes

Parts of the work in this thesis have appeared in the following publications:

- (i) PHISTO: pathogen-host interaction search tool, S. Durmuş Tekir, T. Çakır, E. Ardiç, A.S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür, F.E. Sevilgen, K. Ülgen, *Bioinformatics*, 2013 [14]. (Chapter 3.1)
- (ii) Detection and categorization of bacteria habitats using shallow linguistic analysis, İ. Karadeniz and A. Özgür, *BMC Bioinformatics*, 2015 [15]. (Chapter 3.2) and (Chapter 5.1.5)
- (iii) Bacteria biotope detection, ontology-based normalization, and relation extraction using syntactic rules, İ. Karadeniz and A. Özgür, *Proceedings of the BioNLP Shared Task*, 2013 [16]. (Chapter 3.2) and (Chapter 5.1.5)
- (iv) Linking named entities through an ontology using word embeddings and syntactic re-ranking, İ. Karadeniz and A. Özgür, *BMC Bioinformatics*, 2019 (Under Review). (Chapter 4)
- (v) Literature Mining and Ontology based Analysis of Host-Brucella Gene-Gene Interaction Network, İ. Karadeniz, J. Hur, Y. He, A. Özgür, *Frontiers in microbiology*, 2015 [17]. (Chapter 5.2)

## 1.5. Thesis Overview

In this thesis, we focused specifically on the recognition of biomedical named entities and their normalization through an ontology. The application domains that we targeted are extracting information regarding bacteria biotopes (i.e., bacteria habitats), adverse drug reactions, and pathogen-host interactions.

This thesis demonstrates that the linguistically-motivated rule-based and unsupervised data-driven methods developed for named entity recognition and normalization are promising alternatives to supervised machine learning algorithms in the biomedical domain, where manually labeled data are in general scarce. The targeted text is domain-specific and mainly comprise published scientific articles. Therefore, rule-based and unsupervised data-driven approaches are able to capture some of the available regularity and achieve promising results. The main contributions of this thesis are summarized below.

- (i) A text-mining module is implemented and integrated to the Pathogen-Host Interaction Search Tool (PHISTO) to find the missing experimental method information of Pathogen-Host Interaction (PHI) data (Chapter 3.1) [14].
- (ii) A rule-based method, which makes use of shallow linguistic knowledge, is proposed for the ontology-based tagging and normalization of the biomedical named entities. The method is evaluated for the task of bacteria habitat detection and categorization and promising results are obtained compared to supervised machine learning based algorithms in the 2013 edition of the BioNLP Shared Task on Bacteria Biotopes (Chapter 3.2) [16] [15].
- (iii) A data-driven unsupervised approach is proposed for the ontology-based normalization of biomedical named entities (Chapter 4). This approach is novel in the sense that it makes use of syntactic information of the entity mention phrase while representing the mentions using the embeddings of the constituent words. The proposed approach is applied to the normalization problem of the habitat entities through the Onto-Biotope ontology and the adverse drug reaction entities to the MedDRA dictionary and state-of-the-art results are obtained. Although promis-

ing results were obtained in Chapter 3.2 using the shallow linguistic knowledge based approach for the detection and categorization of bacteria habitat entities, the need for the manually crafted syntax-rules makes the method's adaptation harder to other types of biomedical entities. The newly proposed word embedding based normalization method is unsupervised, since it does not require training data manually annotated with entity mentions and their corresponding concepts in the ontology. Therefore, it can be easily adapted for normalizing different types of biomedical entities. (Chapter 4)

- (iv) A rule-based method, which makes use of anaphora resolution, is proposed for extracting the relations between biomedical named entities. The method is applied for the extraction of bacteria-habitat localization and part-of relations. Despite of the simplicity of the approach, promising results have been achieved over the BioNLP Shared Task 2013 Bacteria Biotopes test data set (Chapter 5.1.5) [16] [15].
- (v) A pipeline, which retrieves the related documents with bacteria localization information from PubMed, identifies and normalizes the bacteria and habitat named entities in these documents, and finally extracts the relations between the identified entities, is developed (Chapter 5.1) [16] [15].
- (vi) Co-occurrence and supervised machine-learning based methods have been applied for extracting Brucella-host interactions from PubMed. The results show that incorporating context is essential for the extraction of biomedical relations, which will be addressed as future work (Chapter 5.2) [17].

## 2. BACKGROUND

This chapter consists of two sections. The first section introduces the tasks that are closely related to the problems studied in this thesis, while the second section briefly explains the machine learning techniques that we apply in this thesis.

### 2.1. Related Tasks

The tasks that are closely related to the problems studied in this thesis are Named Entity Recognition and Named Entity Normalization.

#### 2.1.1. Named Entity Recognition

The approaches proposed in the literature for the problem of named entity recognition can be classified as dictionary-based approaches, rule-based approaches, machine-learning based approaches, and deep learning approaches.

Early dictionary-based approaches tried to identify the entity mentions by utilizing dictionary look-up and string matching algorithms by comparing the entity and the dictionary terms [18–20]. These approaches can deal with morphological variations of the named entity mentions at the character-level and word-level by providing a comprehensive dictionary. On the other hand, they are not able to capture the word-order variations (e.g. “*integrin alpha 4*” or “*alpha 4 integrin*”) in the named entity mentions [21]. Furthermore, dictionary-based approaches are restricted with the completeness of the dictionary. In other words, if the dictionary is not complete in the applied domain, it will be unable to recognize some named entity mentions, which are composed of the words that do not exist in the dictionary. For these reasons, dictionary-based approaches in general achieve high precision performances, but low recall performances.

Rule-based methods [22–24], which usually utilize manually defined rules, have also been proposed as a solution for biomedical named entity recognition. Although, with the rule-based methods, word-order variations in the named entity mentions can also be captured, they need hand-crafted rules, which are time-consuming to build. Moreover, it is difficult to apply the rule-based approaches to new named entity types, especially in the biomedical domain [21].

To overcome the problems of rule-based and dictionary-based approaches, a variety of machine-learning based approaches have been proposed. Classical machine learning based approaches can be grouped as feature-based supervised approaches and unsupervised approaches. Feature-based supervised machine-learning approaches can be categorized as classification-based approaches and sequence-labeling approaches. Both of these approaches generally consist of two phases: the training and the test phases. In the training phase, features are extracted representing each training example given annotated data sets and machine-learning algorithms are utilized to learn a model. In the testing phase, the previously learned model is utilized to label the named entities from unseen data.

Classification-based approaches generally handle the NER task as a multi-class classification task, where each word in a sentence is classified as being part of a certain type of a named entity or not. Support Vector Machines (SVM) [25, 26], and Naive Bayes are among the mostly used classifiers.

Sequence-based approaches include Hidden Markov Models (HMM) based [27–29], and Conditional Random Fields (CRF) based [30–32] approaches. In these approaches, whole sequences of words are taken into consideration instead of single words or phrases. HMM-based NER approaches try to find the most likely tag “ $T$ ” that maximizes  $P(T|W)$ , where “ $W$ ” is the given token sequence. Collier *et al.*, (2000) proposed a HMM based model [33] for the identification of protein and DNA names in the text [27] and Zhou *et al.*, (2004) applied the model proposed in a previous study [28] to the biomedical domain [29]. Conditional Random Fields (CRF) [34] have also become popular in the biomedical domain for the named entity recognition task. Settles *et*

*al.*, (2004) presented a biomedical named entity recognition framework, which utilizes CRFs, for the recognition of protein, DNA, RNA, cell-line and cell-type entity classes from biomedical articles [30]. The study showed that CRFs with even simple orthographic features achieves comparable performances to the state-of-the-art systems at that time.

Zhang & Elhadad proposed an unsupervised approach for biomedical named entity recognition and experimented the approach on two different benchmark data sets [35]. In the study, they utilized features such as noun phrase (NP) chunks and inverse document frequencies. They showed that it is theoretically applicable to the biomedical entity types other than the experimented entity types.

With the increased popularity of deep learning [36], a variety of supervised neural architectures such as neural networks with word and character embeddings [37], convolutional neural networks [38], recurrent neural networks [39], and long-short term memory - conditional random fields (LSTM-CRF) based models [40] have been proposed for the NER task in the biomedical domain. Although promising results are obtained by these approaches, deep learning approaches require extensive amounts of labeled data. In the biomedical domain, for many named entity types, either labeled data does not exist or exists in small amounts that makes the application of these methods to different kinds of named entities difficult.

A number of community-wide challenges including the BioCreative Challenges [41–45] and BioNLP Shared Tasks [46–49], which have been conducted to assist the progress of research in biomedical text mining, also addressed the task of biomedical named entity recognition. The systems developed for these challenges are further explained in the related work subsection of Chapter 3.2.

### 2.1.2. Named Entity Normalization

Several approaches have been proposed for biomedical entity normalization for different types of biomedical entities including genes/proteins [7–10, 50, 51], bacteria biotopes [15, 48, 49, 52, 53], and diseases [11, 12].

Early systems tried to link the entity mentions to the knowledge base entities by utilizing dictionary look-up and string matching algorithms [7, 50, 51, 54]. For dictionary based approaches, automatically extracted dictionaries were also utilized for the normalization of named entities such as genes and proteins [50]. Similarly to the NER studies, dictionary-based approaches for named entity normalization are also restricted with the completeness of the dictionary. In other words, unseen entities in the dictionary can not be captured and normalized by dictionary-based approaches [55].

Similar to the NER studies, rule-based approaches that rely on manually defined rules [11] or automatically extracted rules [56] have also been proposed for entity normalization. When the context is not defined by the rules, these kinds of approaches are not be able to normalize the entities in the context. As a result, the rule-based approaches are difficult to adapt to new entity types.

Feature-based supervised machine-learning approaches, which learn the similarities between biomedical entity mentions and ontology concept names from labeled training data have also been proposed and applied as a solution to the normalization task of various biomedical entities. For example, GeNo is a gene name normalization system, which utilizes logistic regression for learning a string similarity measure from a dictionary [57] and DNORM is the first supervised machine-learning based disease name normalization system, which utilizes pairwise learning to rank with the aim of the normalization of the disease mentions [12].

Approaches that rely on convolutional neural networks have also been proposed for the normalization of biomedical entities from biomedical literature [58]. Experiments on two benchmark datasets (the ShARe/CLEF eHealth dataset and the NCBI disease dataset) resulted in promising performances. However, the need for the manually annotated training data makes the adaptation of such methods to new domains difficult. Cho *et al.*, (2017) proposed a semi-supervised approach that facilitates word embeddings to represent semantic spaces for normalizing biomedical entities such as disease names and plant names [59]. Together with unlabeled data, this method also makes use of labeled domain specific data, which makes its adaptation to other domains difficult, if there are no such resources available.

The BioCreative Challenges [41–45] and BioNLP Shared Tasks [46–49] also addressed the task of biomedical entity normalization. The systems developed for these challenges are further explained in the related work section of Chapter 4.

## 2.2. Related Techniques

The machine learning techniques that we apply in this thesis are word embeddings and support vector machines, which are briefly explained in this section.

### 2.2.1. Word Embeddings

Word embeddings are the techniques that are used to represent the words in a vocabulary as vectors of real numbers [60]. The terms word representation and distributional representation are also used to refer to word embeddings.

In text mining, the extraction of semantic knowledge is an important issue to make sense of the documents and the sentences. The motivation of representing the words as real value vectors gives the opportunity to observe the semantic knowledge between the represented words. To extract semantic knowledge, the conversion of words to vectors of real numbers, which are machine processable formats, is in general needed. There are many approaches for the representation of a word in NLP, the easiest



of which is called one-hot encoding, in which 1 stands for the position where the word exists in a vocabulary and 0 for the other positions. For a sample sentence “*Stomach is a part of human gastrointestinal tract*” the words in the vocabulary are “*Stomach, is, a, part, of, human, gastrointestinal, tract*”. According to one-hot encoding, “1, 0, 0, 0, 0, 0, 0, 0” represents the word “*Stomach*” and “0, 0, 0, 0, 0, 0, 1, 0” represents the word “*gastrointestinal*”.

In one-hot encoding, word vectors are calculated without considering the other words in the context. As a result, this kind of symbolic representation can not capture the semantic similarity between the words such as “*Stomach*” and “*gastrointestinal*”, although we know that there is a semantic similarity between them and “*Stomach*” is part of the “*gastrointestinal*” system.

On the other hand, another type of word representations, which is called distributional representations, describe the meanings of words by understanding the context in which they appear. In these representations, the aim is to build dense vectors for each word type such that they are good at predicting the other words in the context. As a result, if we represent the vectors of words in the coordinate space, the vectors of the words with similar context are observed to occupy close positions to each other. Mathematically, the cosines of the angles between such vectors are close to 1.

Distributional representations became popular in NLP with word2vec, which is a popular learning model proposed by Mikolov *et al.*, (2013) [61], to induce word embeddings from large unlabeled corpora. Skip gram and continuous bag-of-words (CBOW) are the two different schemes that are proposed for word2vec model to learn word representations. Both schemes are based on neural network models. For the skip-gram model, the surrounding context is predicted given the current word. On the other hand, for the CBOW model, the current word is predicted given the surrounding context. Both schemes have their own advantages and disadvantages. Skip gram scheme is observed to perform better with small amount of data. Furthermore, it is able to represent rare words better. On the other hand, the CBOW scheme is faster than skip gram scheme and able to represent frequent words better.

Chiu *et al.*, (2016) tried to find answers to the research questions of how to train good word embeddings for biomedical NLP and how the quality of the embeddings change according to the input corpora, model architectures, and hyper-parameter settings [62]. In the study, to assess the effect of the input corpora, three variants of corpora (PubMed, PMC, PubMed & PMC) were used to learn word embeddings. The evaluation results showed that using a larger input corpus do not always guarantee a higher score in the biomedical domain. Furthermore, the skip-gram model achieved higher performance than the CBOW model. It was also shown that optimum window sizes to define the context varies according to the type of the task. For NER tasks, a window size of 2 is adequate, while for similarity and relatedness tasks, a window size of 30 achieves better results.

As a consequence, word embedding models are promising approaches for capturing semantic information and have been successfully used in several recent NLP tasks such as named entity recognition [40, 63], word-sense disambiguation [64, 65], information retrieval [66, 67], and machine translation [68]. Word embedding models have also led to promising results in the biomedical domain [62, 69–71].

### 2.2.2. Support Vector Machines

Support Vector Machines (SVM) are one of the most popular machine learning techniques that are used for classification and regression problems. In this part of the thesis, classification based SVMs will be covered in detail.

The ultimate goal of the classification based SVMs is to separate the data points that belong to different classes with an optimal decision surface (or hyper-plane). In other words, the aim is to find a hyper-plane, that separates the instances into different classes in the best way. In two dimensions, the decision surface used to classify instances is a line. On the other hand, the decision surface used to classify the instances in three-dimensional space is a plane, whereas in higher dimensions, it is a hyper-plane.

Although, in general, there are lots of possible solutions that separate the features, the aim is to find the optimal hyper-plane. The optimal hyper-plane is the hyper-plane for which the distance between the hyper-plane and the closest data points (or support vectors) is maximized. The intuition is that the closest data points to the hyper-plane are the data points that are most difficult to be classified. The SVM classifier takes into consideration these closest data points, which are called support vectors, and neglects the other data points.

However, in reality, not all the patterns are linearly separable. In this case, the original data points are required to be transformed into a new space, where the training set is separable by utilizing a similarity function, which is called a kernel function. Edit kernel and cosine kernel are two of the kernel functions that can be used to implicitly realize this transformation in SVMs in the NLP domain.

2.2.2.1. Cosine Kernel. The cosine kernel defines the similarity between two vectors as the cosine of the angle between them and is formulated as follows:

$$\text{cos\_sim}(\mathbf{a}, \mathbf{b}) = \text{cos}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \bullet \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \quad (2.1)$$

which is the dot product of the two vectors ( $\mathbf{a}$  and  $\mathbf{b}$ ) normalized by the lengths of them.

The cosine similarity takes values in the range of  $[0, 1]$ , where the similarity takes the maximum value of 1 if all the terms are the same. On the other hand, if none of the terms are the same, then the cosine similarity takes the minimum value of 0.

2.2.2.2. Edit Kernel. Edit kernel is defined by the edit distance between two strings, where it is the minimum number of operations that have to be performed to transform a string to another. In edit distance, there are three types of operations, which are insertion, deletion, or substitution of a single character in a string. For example, the edit distance between “dog” and “dogs” is 1. We insert “s” to the first word “dog” to convert it to the second word “dogs”. The edit distance between “bad” and “sad” is also 1. We substitute “b” with “s” to convert the first word “bad” to the second word “sad”.

In the edit kernel, edit distance is normalized by dividing it by the number of characters in the longer string, so that it takes values in the range of [0,1]. The edit distance is converted into a kernel (or similarity) function as follows:

$$edit\_sim(\mathbf{a}, \mathbf{b}) = e^{-\gamma(edit\_distance(\mathbf{a}, \mathbf{b}))} \quad (2.2)$$

where  $\gamma$  is a parameter that is a positive real number, which is used to obtain a positive definite kernel function [72].

### 3. ENTITY TAGGING AND NORMALIZATION USING SHALLOW LINGUISTIC ANALYSIS

In this chapter, first an exact string matching based method for entity recognition is described. The method has been integrated as a text mining module to the Pathogen-Host Interaction Search Tool (PHISTO) for identifying the experimental methods of pathogen-host protein-protein interactions. Recognizing the limitations of an exact string matching based approach, next a linguistically motivated rule-based approach is developed and evaluated for bacteria biotope detection and categorization. Although the rule-based approach obtained promising performance, it is difficult to adapt to other domains due to the hand-crafted rules. Therefore, in the next chapter an unsupervised data-driven method for entity normalization is presented, which makes use of the word embeddings of the entity mentions as well as their syntactic parses.

#### 3.1. Entity Detection using Exact String Matching

In this section the string matching based entity detection module developed to detect protein-protein interaction experimental methods and integrated to PHISTO is described.

##### 3.1.1. Background

The interactions between the proteins of infectious microorganisms, pathogens and their human hosts allow the microorganisms to manipulate human cellular mechanisms to their own advantage, resulting in infection in the host organism. The recent advances in high-throughput protein interaction detection methods have led to the production of large-scale inter-species protein-protein interaction (PPI) data of pathogen-human systems. Currently, there are a number of pathogen-host interaction (PHI) resources that are specific to some pathogens. The only available resource to access all PHI data in a single database [73] does not offer any additional functionality

to analyze PHI networks. Pathogen–host interaction search tool (PHISTO) serves as an up-to-date and functionally enhanced source of PHI data through a user-friendly interface. Text mining is used to label PHIs extracted without any information on interaction detection method.

23,661 PHIs between human and pathogens are stored in PHISTO. Among the PHI data extracted from the PPI databases, there were 12,751 PHI data that were not labeled with the experimental methods used to detect these interactions. In order to make such PHI data available to the users, a fully automatic text mining module for experimental method extraction is implemented using JAVA. The following subsections describe this module in more detail.

### **3.1.2. Methods and Materials**

3.1.2.1. Dictionary of Experimental Methods. We compiled a dictionary which includes interaction detection methods with general method names and the synonyms from the PSI-MI Ontology version 2.5 [74]. The dictionary consists of 115 different experimental method names with 159 synonyms.

A sample portion from our experimental method dictionary is shown in Figure 3.1. MI: ID is the ID of the term in the PSI-MI ontology. The rows in bold colors indicate the general names of the experimental methods, and the other rows correspond to the synonyms. Both the general names and the synonyms of the methods are used to assign methods to PHI data without experimental method information. In order to have consistent data in PHISTO, the methods are represented with their general names. Therefore, when a method’s synonym is matched in text, it is converted to its corresponding general name using the dictionary before being stored in PHISTO.

3.1.2.2. Abstract Retrieval Module. An abstract retrieval module is implemented to automatically download the abstracts of the articles that contain PHIs without experimental method information from PubMed [2]. The PHI data downloaded from the PPI

	MI:ID	Method Name
<b>General Name</b>	MI:0004	affinity chromatography technology
Synonym1	MI:0004	affinity chrom
Synonym2	MI:0004	Affinity purification
<b>General Name</b>	MI:0006	anti bait coimmunoprecipitation
Synonym1	MI:0006	anti bait coip
<b>General Name</b>	MI:0007	anti tag coimmunoprecipitation
Synonym1	MI:0007	anti tag coip

Figure 3.1. Sample portion of our experimental method dictionary

databases contain the PubMed IDs of the articles from where they were curated. Using this information a list of PubMed IDs associated with PHIs without experimental method information is created.

Jsoup [75], which is an open source Java library for HTML source parsing, is used to parse the HTML source code of the related article pages from PubMed. All abstracts, whose PubMed IDs are in PubMed ID list generated above, are downloaded one by one from PubMed. A sample abstract with PubMed ID: “11129635” that is extracted from PubMed is shown in Figure 3.2.

TITLE: The vaccinia virus E3L protein interacts with SU... [Virus Genes. 2000] - PubMed - NCBI  
 ABSTRACT: We report the results of a two-hybrid study which identified clones from a HeLa cDNA library that interact with the vaccinia virus protein E3L. These clones encode the nuclear protein SUMO-1 (also known as PIC-1, sentrin or GMP-1); the cytoplasmic ribosomal protein L23a; and a small peptide sequence of unknown significance.

Figure 3.2. Sample abstract

3.1.2.3. Exact Matching Algorithm. Exact string matching is used to assign experimental methods to PHIs without experimental method information. For each PHI data without experimental method, the related abstract is fetched from the abstracts’ corpus by using the PubMed ID associated with the PHI. All the general method names and their synonyms in the experimental method dictionary (which is compiled from PSI-MI Ontology) are matched against the text of the abstract. If an experimental method name from the dictionary occurs in the abstract, it is assigned to the current PHI. The pseudo-code of this algorithm is shown in Figure 3.3.

A sample PHI, which does not originally contain experimental method information (represented with “not specified”), is given below (Figure 3.4). Its corresponding abstract (shown in Figure 3.2) is obtained from PubMed using the abstract extractor

For each PHI data without experimental method
Find the related <b>Abstract</b> from the corpus
(using PubMed ID of the current PHI data)
For each <b>Method</b> in the experimental methods list
Scan the related abstract <b>with exact matching</b>
(for the current method in the list)

Figure 3.3. Pseudo-code of Exact Matching Algorithm

module described in Section 3.1.2.2. By using the exact string matching technique, the “two-hybrid” method is assigned to this PHI (see Figure 3.2.).

```
Vaccinia virus STRAIN WR, 10254,P21605,VE03 VACCV --- Protein E3(p25)
,P62750,RL23A HUMAN --- 60S ribosomal protein L23a,"not specified",11129635
```

Figure 3.4. Sample PHI data without experimental method information

### 3.1.3. Results and Discussion

We evaluated the text mining module using a test set consisting of 5104 PHIs that contain experimental method information in PHISTO. The test set was created by removing the 8162 PHIs whose experimental method was specified as Reactome-curated, since Reactome-curated is not an experimental method name occurring in the PSI-MI ontology. It rather denotes that these PHIs were obtained from the Reactome database. We also removed 8903 PHIs that were curated from the article with PubMed ID 20711500 [76]. Even though our text mining module was successfully able to determine their experimental method, we decided to remove these PHIs from our test set, since including them would result in unrealistically high performance scores.

The text mining module assigned experimental methods to 2331 of the 5104 PHIs in our test set. 1715 of these were correctly identified experimental methods. Thus, the module achieves a promising precision of 74%. In other words, 74% (1715 / 2331) of the assigned methods are correct. The recall and the F-score of the module are 34% (1715 / 5104) and 47% (harmonic mean of recall and precision), respectively.

The text mining module is applied to PHI data that do not have experimental method information. By utilizing the text mining module, 2952 experimental method



names are extracted for 2109 unique PHIs. Finally, these results are stored in PHISTO, which are presented with an asterisk to indicate that these experimental methods are obtained by the text mining module.

### **3.2. Entity Detection and Normalization using Rules based on Shallow Linguistic Analysis**

In this section, we introduce a linguistically motivated rule-based approach for entity recognition and normalization and apply the approach for tagging names of bacteria habitats (i.e., biotopes) in biomedical text by using an ontology. Our approach is based on the shallow syntactic analysis of the text that include sentence segmentation, part-of-speech (POS) tagging, partial parsing, and lemmatization.

Information regarding bacteria biotopes is important for several research areas including health sciences, microbiology, and food processing and preservation. One of the challenges for scientists in these domains is the huge amount of information buried in the text of electronic resources. Developing methods to automatically extract bacteria habitat relations from the text of these electronic resources is crucial for facilitating research in these areas.

We participated in the Bacteria Biotope (BB) Task of the BioNLP Shared Task 2013. Our system (Boun) achieved the second best performance with 68% Slot Error Rate (SER) in Sub-task 1 (Entity Detection and Categorization) This section of the thesis reports the system that is implemented for the shared task, including the novel methods developed and the improvements obtained after the official evaluation. The extensions include the expansion of the OntoBiotope ontology using the training set for Sub-task 1, which resulted in promising results for Sub-task 1 with a SER of 68%.

Our results show that a linguistically-oriented approach based on the shallow syntactic analysis of the text is as effective as supervised machine learning approaches for the detection and ontology-based normalization of habitat entities.

### 3.2.1. Background

Identifying and characterizing the habitats where bacteria live (i.e. bacteria biotopes) is crucial for gaining a better understanding of bacterial infections, which in turn can lead to the development of novel disease prevention, prediction, and treatment methods. Besides health sciences, information about the relations of bacteria with their environments is also important for research areas such as microbiology, agronomy, and food processing and preservation. One of the challenges that researchers in these areas face is the absence of a comprehensive database that stores the relationships among bacteria and their habitats in a structured format. Most of the bacteria habitat information is only available in unstructured textual format in electronic resources such as scientific publications and web pages of bacteria sequencing projects [77]. For example, even a limited search in PubMed for “*bacteria AND (habitat OR localization OR environment)*”, which probably barely covers all relevant documents, returns 177,000 documents (Search date: January 29, 2014). This illustrates the difficulty of manual curation for creating a comprehensive database of bacteria and habitat relations. An important step towards the creation and population of such a database is developing text mining methods to automatically recognize and normalize mentions of bacteria and habitats in text, as well to identify the relations among them.

The Bacteria Biotope (BB) Task in the BioNLP Shared Task 2013 addressed the problems of identifying locations where bacteria live and semantically annotating them using an ontology [52, 77, 78]. Unlike most previous biomedical information extraction challenges which target extracting information from publications in PubMed (e.g. [43, 79, 80]), the documents targeted in the BB task are scientific web pages intended for a general audience. In addition, these documents are richer in terms of both the number and the variety of habitats, compared to the ones in PubMed [77].

The BB task consisted of three sub-tasks. Sub-task 1 involved the identification of habitat mentions in text and the assignment of them to the concepts in the OntoBiotope (MBTO) Ontology [81]. Figure 3.5 shows a sample text file from the training set provided by the organizers. The bacteria and habitat entities are shown in bold. For

**Bifidobacterium longum NCC2705**  
 Description  
**Bifidobacterium**. Representatives of this genus naturally colonize the **human gastrointestinal tract (GIT)** and are important for establishing and maintaining homeostasis of the **intestinal ecosystem** to allow for normal digestion. Their presence has been associated with beneficial health effects, such as prevention of diarrhea, amelioration of lactose intolerance, or immunomodulation. The stabilizing effect on **GIT microflora** is attributed to the capacity of **bifidobacteria** to produce bacteriocins, which are bacteriostatic agents with a broad spectrum of action, and to their pH-reducing activity. Most of the ~30 known species of **bifidobacteria** have been isolated from the **mammalian GIT**, and some from the **vaginal and oral cavity**. All are obligate anaerobes belonging to the Actinomycetales, branch of Gram-positive bacteria with high GC content that also includes **Corynebacteria**, **Mycobacteria**, and **Streptomyces**.  
 Description  
**Bifidobacterium longum**. This organism is found in **adult humans** and **formula fed infants** as a normal component of **gut flora**.  
 Description  
**Bifidobacterium longum strain NCC2705**. This strain was isolated from **infant feces**. The genome of this strain is being sequenced for comparative genomics.

Figure 3.5. Sample text. A sample input file containing bacteria and habitat entities.

instance, “*Bifidobacterium*” is a bacteria entity, whereas “*human gastrointestinal tract*” and “*human*” are habitat entities. The concept that is associated with the “*human gastrointestinal tract*” habitat in the OntoBiotope ontology is “*digestive tract*”, and the one associated with the “*human*” habitat is “*human*”.

Given the names, types (i.e. *Bacteria*, *Habitat*, *Geographical*), and positions of the entities in text the goal of Sub-task 2 was to extract the localization relations between bacteria and habitat (i.e. *Habitat*, *Geographical*) pairs, as well as PartOf relations between habitat pairs. A PartOf relation between a pair of habitats holds if one of them is a living organism (called host), and the other one is a part of this organism (called host part). The relation between “*Bifidobacterium*” and “*human gastrointestinal tract*”, as well as the one between “*Bifidobacterium*” and “*human*” are among the localization relations described in the text shown in Figure 3.5. The relation between the host “*human*” and the host part “*human gastrointestinal tract*” is one of the PartOf relations described in Figure 3.5. One of the challenges in the relation extraction task is the high frequency of bacteria anaphors and relations that cross sentence boundaries.

Sub-task 3 was the same as Sub-task 2, except that the gold standard entities were not provided to the participants. In other words, the participants were also expected

to detect the bacteria and habitat entities.

In the following sections of this chapter, our proposed linguistically-oriented rule-based approach for entity detection and categorization is explained. We describe our submissions to Sub-task 1 (Entity Detection and Categorization) [16], as well as the new methods that we developed and the improvements that we obtained after the official evaluation. Our approach is based on the shallow syntactic analysis of the text including sentence-splitting, tokenization, lemmatization, POS tagging, and shallow (partial) parsing. Manually designed syntactic rules that utilize the noun phrases in the sentences and the POS tags of the words are used to recognize the habitat entities and map them to the corresponding concepts in the OntoBiotope ontology. Our approach also tackles the problem of handling discontinuous entities such as the two distinct entities “*nasal cavity*” and “*oral cavity*” in the phrase “*nasal and oral cavity*”.

As improvements to the Sub-task 1 system, we investigate expanding the OntoBiotope ontology using the training set and extending the noun phrases with their modifiers including the ones that are attached with the prepositions *in*, *of*, and *with* (e.g. “*infected child in Germany*”).

### 3.2.2. Related work

Due to the continued rapid increase in the number of scientific articles published in the biomedical domain, it has become difficult for scientists to reach and make use of the knowledge contained in the biomedical scientific literature. Therefore, developing text mining systems for automatically extracting the biologically useful information from biomedical text has become crucial [21]. A number of shared tasks including the LLL and BioCreative Challenges, as well as the BioNLP Shared Tasks have been conducted, which have facilitated research in biomedical text mining [43, 79, 82, 83]. Most of these shared tasks addressed the problems of relation or event extraction among bio-molecular entities such as proteins and genes.

The Bacteria Biotope Task is the first shared task targeting the extraction of information about bacteria and their habitats. This task was first conducted in the BioNLP Shared Task 2011 [47, 84, 85]. Among the three teams that participated in the Bacteria Biotope Task 2011 [84, 85], Bibliome INRA [86] obtained the best F-score performance (45%) on the task of identifying habitat entities. They made use of resources including a list of Agrovoc geographical names [87], the NCBI Taxonomy [88], as well as an ontology for location types, and developed a system that is based on ontology-based reasoning and linguistic features. UTurku [89] developed a generic supervised machine learning based system that can be used for all the main tasks in the BioNLP Shared Task 2011 with minor modifications. They incorporated this generic system with additional named entity recognition patterns and external resources for identifying the named entities and their types in the Bacteria Biotope Task. JAIST [90] also used a supervised machine learning approach based on Conditional Random Fields (CRFs) [34] for this task.

The Bacteria Biotope (BB) Task in the BioNLP 2013 Shared Task gave another opportunity to scientists to address the task of extracting information about bacteria and their habitats from text and evaluate their approaches on a common platform [52, 77]. This task maintained the primary objective of the 2011 edition of the BB task of extracting bacteria and localization relations. In addition, it introduced a new task that targeted a more fine-grained categorization (i.e. normalization) of habitat entities through the OntoBiotope ontology. Five teams participated in the 2013 edition of the BB Task [52, 77]. For Sub-task 1 the systems were ranked according to their slot error rates (SER). The first three systems obtained similar SER performances for this Sub-task despite their different approaches to the problem [52, 77]. The LIPN system [91] based on a supervised machine learning approach achieved the best SER score (66%) in Sub-task 1. The IRISA system used a supervised machine learning approach based on the k-Nearest Neighbor (kNN) method and obtained a SER score of 93% in Sub-task 1 [92]. LIMSI [93] was the only team that participated in all three BB sub-tasks. They used a method based on Conditional Random Fields [94] for the official submissions, while they utilized Maximum Entropy models for later improvements. They utilized various additional resources such as NCBI taxonomy for the detection of

bacteria names, the Cocoa [95] annotations for the categorization of bacteria, habitat, and geographical entities, and OntoBiotope Ontology for the identification of habitat names. They obtained a SER value of 68% in the official submissions for Sub-task 1. We participated in Sub-task 1 and Sub-task 2 of the BB Task 2013. Our system *Boun* ranked second in Sub-task 1 with a SER score of 68% and third in Sub-task 2 with an F-score of 27% in the official evaluation [16]. The Sub-task 1 module of the *Boun* system utilized the shallow syntactic analysis of the text and linguistically-motivated rules. The extended system *Boun 2* obtained 68% SER on Sub-task 1. The details of our official submission as well as the improvements developed after the shared task are described in the following sections.

Sub-task 1 of the BB Task is related to the general problem of named entity recognition (NER) and automatic semantic annotation by ontologies. Rule-based approaches (e.g. [96]), as well as machine-learning based methods (e.g. [97, 98]) have been developed for biomedical NER. While state-of-the-art NER systems for proteins and genes achieve performance levels that enable their use in practice, the problem of recognizing bacteria habitat names in text has not been tackled prior to the 2011 and 2013 editions of the BB Task, and there is still a lot of room for improvement. Different approaches for the semantic annotation of entities using ontologies have been proposed in the literature. Our approach is related to rule-based methods that make use of the syntactic and semantic analysis of the terms [99, 100]. A problem related to ontology-based semantic tagging has also recently been addressed in the Biocreative III Interaction Method Task (IMT) [101]. The goal was to identify the interaction methods in the articles and normalize them through the PSI-MI ontology [74]. The best performing systems in the shared task employed supervised machine learning methods [102, 103]. However, they formulated the problem as classifying the entire articles to the ontology concepts, and did not address the problem of identifying the boundaries of the named entities. The relatively smaller training set size in the BB Task and the large number of classes (i.e. 1700 concepts) pose challenges for supervised machine learning based classifiers in this domain.

### 3.2.3. Methods

We developed a linguistically motivated rule-based system for Sub-task 1 (Entity Detection and Categorization), the workflow of which is displayed in Figure 3.6. The input text is first pre-processed by splitting into sentences and performing shallow syntactic analysis including POS tagging, lemmatization, and partial parsing. Based on our observation in the training set, we assume that most habitat entities are noun phrases. Before normalizing through the OntoBiotope ontology, the candidate habitat entities are identified by extracting and simplifying the noun phrases in the sentences. In addition, the OntoBiotope ontology is expanded by using the training set. We also investigate handling discontinuous entities and entity modifiers. The details of our approach are described in the following subsections.

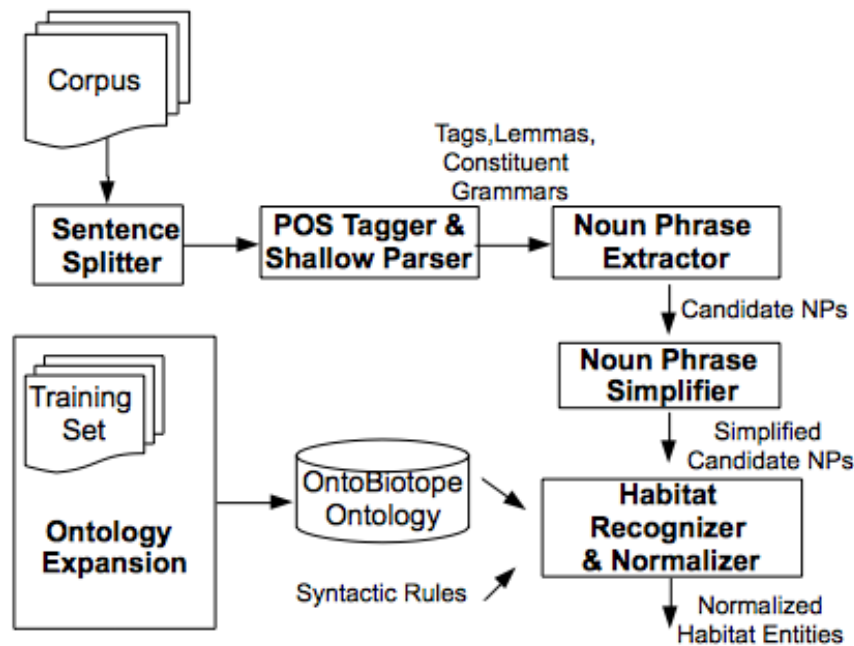


Figure 3.6. Workflow of the Sub-task 1 System

**3.2.3.1. Preprocessing.** In the preprocessing step, we used the Genia Sentence Splitter (GeniaSS) [104] to segment the text into sentences and the Genia Tagger [97, 105] to obtain the shallow linguistic features of these sentences including the POS tags, the lemmas, and the constituent categories of the words. Figure 3.7 shows a sample sentence and the output obtained by the preprocessing module (on the left-hand side of the figure). These shallow syntactic analysis results are then used in the following steps

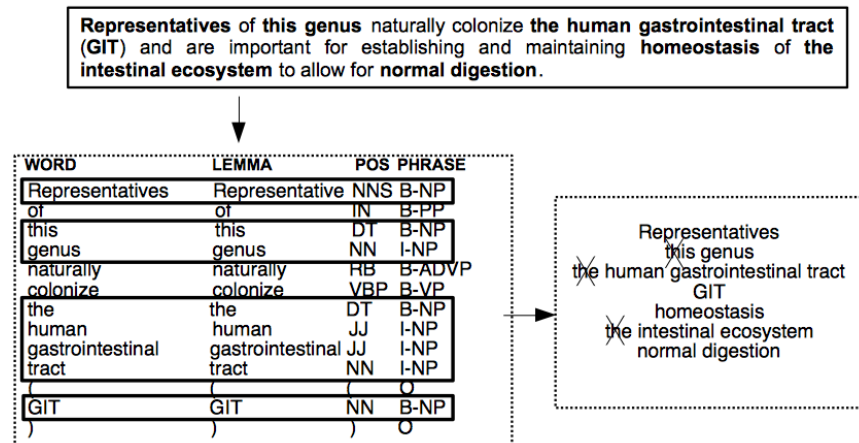


Figure 3.7. Sample output of the preprocessing, and the noun phrase extractor and simplifier

of our system to extract and simplify the noun phrases (as shown on the right-hand side of Figure 3.7), as well as to map them to the OntoBiotope ontology.

**3.2.3.2. Ontology expansion from the training data.** In this step, the annotated training data set is used to expand the OntoBiotope ontology. If a term in the training set is labeled with an OntoBiotope ontology concept, it is included to the ontology as a synonym of that concept, unless it is already defined as a name or as a synonym of that concept. For example, the ontology concept with ID *MBTO:00001875* has the name “*mummy tissue*” in the ontology. This entry does not have any synonyms. However, in the training set the term “*tissues of ancient mummies*” is labeled with this concept. Therefore, “*tissues of ancient mummies*” is added as a synonym of the “*mummy tissue*” concept in the ontology.

**3.2.3.3. Noun phrase extraction and simplification.** In the noun phrase extraction and simplification step, first, the noun phrases are extracted based on the constituent categories of the words identified by the Genia Tagger. Next, the extracted noun phrases are simplified by removing the words that do not contain informative information regarding bacteria habitats. The non-informative words are identified based on their POS tags. For instance, determiners and possessive pronouns are non-informative and thus, are not included to the boundaries of the habitat entities. Consider the noun phrases “*the*



*mummy tissue*” and *“its small intestine”*. The simplified noun phrases are obtained by removing the determiner *“the”* from the first noun phrase and the possessive pronoun *“its”* from the second noun phrase. Thus, the simplified noun phrases are *“mummy tissue”* and *“small intestine”*, respectively. The preprocessing, noun phrase extraction and simplification processes are illustrated in Figure 3.7 for a sample sentence.

**3.2.3.4. Discontinuous entity handling.** Some habitat entity spans in text may be discontinuous. For example, the phrase *“ground and surface water”* contains two overlapping entities, namely *“ground water”* and *“surface water”* [77]. Our system includes a mechanism to handle discontinuous entities, which are represented with noun phrases containing the conjunction *“and”*. Such noun phrases are split into two sub-phrases from the conjunction *“and”*. If the two sub-phrases map to two concepts in the OntoBiotope ontology, which have the same direct ancestor represented with a common is-a relation, then the habitats are identified according to the structure of the noun phrase as follows. Each sub-phrase is considered to be a separate habitat entity, if both of the sub-phrases consist of single words tagged as nouns. Otherwise, the two sub-phrases constituting the noun phrase are identified as a single habitat entity. On the other hand, if the mapped two concepts in the OntoBiotope ontology don’t have a common direct ancestor, then the corresponding two sub-phrases are considered to be two separate habitat entities. Our approach for discontinuous entity handling is described in more detail below through the example phrases *“pharyngeal and gut mucosa”*, *“iron-rich and wet environment”*, *“plants and animals”*, and *“mouse and cheese”*.

- Given the phrase *“pharyngeal and gut mucosa”*, the two generated sub-phrases are *“pharyngeal mucosa”* and *“gut mucosa”*. The direct ancestor of *“pharyngeal mucosa”* in the OntoBiotope ontology is *“respiratory tract part”*, whereas the direct ancestors of *“gut mucosa”* are *“digestive tract part”* and *“mucosal tissue”*. Since the OntoBiotope ontology concepts corresponding to the two sub-phrases don’t have a common direct ancestor, these sub-phrases are identified as two different habitat entities, namely *“pharyngeal mucosa”* and *“gut mucosa”* (See Figure 3.8).

- Given the phrase “*iron-rich and wet environment*”, the two generated sub-phrases are “*iron-rich environment*” and “*wet environment*”. The two concepts corresponding to these sub-phrases in the OntoBiotope ontology have a common direct ancestor, which is “*habitat wrt chemico-physical property*”. Therefore, a single habitat entity (i.e., “*iron-rich and wet environment*”) corresponding to the entire noun phrase is generated (See Figure 3.9).
- Given the phrase “*plants and animals*”, the two generated sub-phrases are “*plants*” and “*animals*”. The two concepts corresponding to these sub-phrases in the OntoBiotope ontology have the “*eukaryote host*” direct ancestor. However, since both sub-phrases consist of single words, which are tagged as noun, two different habitat entities are identified, namely “*plants*” and “*animals*”.
- Given the phrase “*mouse and cheese*”, the two generated sub-phrases are “*mouse*” and “*cheese*”. The concepts corresponding to these sub-phrases in the OntoBiotope ontology don’t have a common direct ancestor. Therefore, two different habitat entities, namely “*mouse*” and “*cheese*”, are identified.

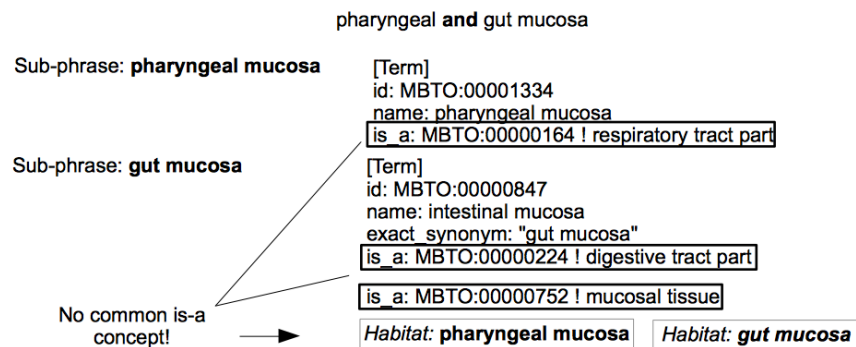


Figure 3.8. Discontinuous entity handling for the sample phrase “*pharyngeal and gut mucosa*”

**3.2.3.5. Entity modifier handling.** The data set for the Bacteria Biotores shared task has been annotated by including the modifiers that describe the habitats in the boundaries of the habitat entities [77]. Consider the phrase “*infected infant in Germany*”. The ontology concept that this phrase is mapped to is “*infant*” (MBTO:00000778). However, the boundary of the habitat entity is the entire phrase, namely “*infected infant in Germany*”. The shallow parser labels “*infected infant*” and “*Germany*” as two separate noun phrases and “*in*” is labeled as a preposition. After the official evaluation,

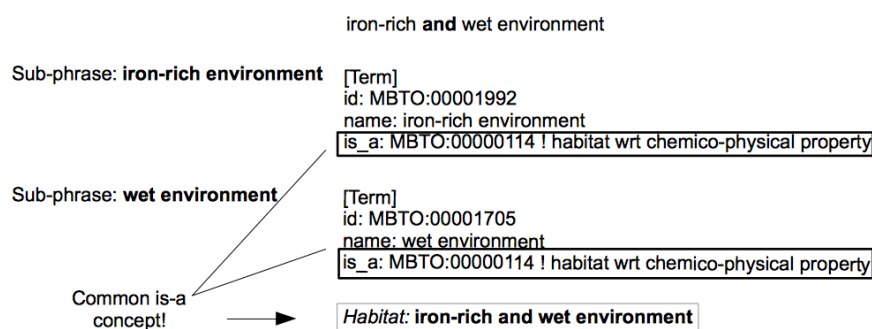


Figure 3.9. Continuous entity handling for the sample phrase “*iron-rich and wet environment*”

our system has been extended to handle the habitat entities that contain modifiers. If a noun phrase (*NP*) is followed by a preposition (*prep*) and then by another noun phrase, the entire *NP prep NP* sequence is identified by the noun phrase extraction and simplification module as a candidate habitat entity. Besides the prepositional phrases that contain “*in*”, the ones that contain “*of*” (e.g. “*respiratory tract of animals*”) and “*with*” (e.g. “*2-year-old girl with tick-bourne relapsing fever*”) are also handled using the same approach. However, as discussed in the Results section this extension degraded the performance of the system.

**3.2.3.6. Ontology mapping.** To identify whether the phrases extracted in the previous steps correspond to habitat entities and to determine the boundaries of the habitat entities, exact or partial matching against the names and synonyms of the concepts in the OntoBiotope ontology is performed.

Consider the extracted noun phrase “*the animal bodily fluid*”. In the noun phrase simplification step, this phrase is simplified as “*animal bodily fluid*”, which is searched against the OntoBiotope ontology for exact or partial matches. As shown in Figure 3.10, this candidate phrase is mapped to two ontology concepts. It is mapped to the concept “*body fluid*” due to the partial match with the *exact\_synonym*: “*bodily fluid*”. Similarly, it is mapped to the concept “*animal*” due to the partial match with the concept *name*: *animal*.

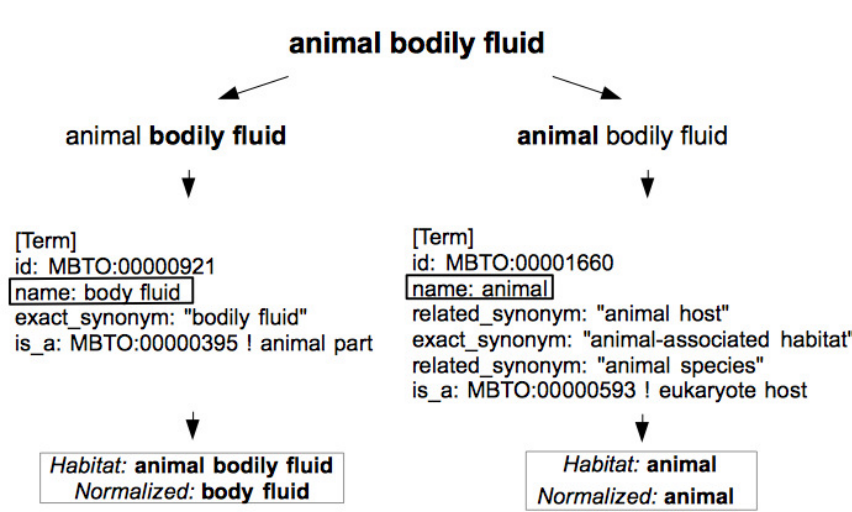


Figure 3.10. Ontology mapping example

The boundaries of the habitat entities are identified by using the following manually designed syntactic rules.

- If there is an exact match between an ontology concept and a candidate phrase, the candidate phrase is identified as a habitat entity and the entity boundary is set as the boundary of the candidate phrase.
- If there is a partial match between a candidate phrase and an ontology concept such that the match begins from the first word of the candidate phrase, but does not cover the entire phrase, the matching sub-phrase of the candidate phrase is identified as a habitat entity and the entity boundary is set as the boundary of the matching sub-phrase. For instance, as shown in Figure 3.10, the first word of the candidate phrase “*animal bodily fluid*” matches with the *name*: “*animal*” of the ontology concept for *animal*. Therefore, the habitat entity “*animal*” is identified and normalized to the ontology concept with *MBTO:00001660*.
- If there is a partial match between a candidate phrase and an ontology concept such that the match does not begin from the first word of the candidate phrase, the candidate phrase is identified as a habitat entity and the boundary of the entity is set as the boundary of the phrase. For instance, in Figure 3.10, the candidate phrase “*animal bodily fluid*” matches with the *exact\_synonym*: “*bodily fluid*” of the ontology concept for *body fluid*, starting with the second word of the candidate phrase. Therefore, the entire candidate phrase “*animal bodily fluid*”

is identified as a habitat entity and normalized to the ontology concept with *MBTO:00000921*.

In order to match the different inflected forms of the habitat names such as matching the habitat name “*animal*” against its plural form “*animals*”, we performed lemmatization on the candidate phrases by using the Genia Tagger, and applied the same methodology that is explained above not only to the surface forms of the candidate phrases, but also to the lemmatized forms of them.

### 3.2.4. Results for the BioNLP Shared Task 2013 Data Set

3.2.4.1. Data set. The training, development, and test sets provided by the BB shared task organizers contain 52, 26, and 26 documents, respectively. The gold standard annotations for the training and development sets were provided to the participants, whereas the evaluations on the test set were performed by using the online evaluation tool released by the shared task organizers. The documents in the corpus consist of web pages obtained from a number of web sites such as from the web sites of bacteria sequencing projects or MicrobeWiki [77].

3.2.4.2. Evaluation metrics. The main evaluation metric used for Sub-task 1 is Slot Error Rate (*SER*) [77]. Lower SER values denote better performance, since SER is an error measure. The computation of SER is shown in Equation 3.1, where  $S$ ,  $D$ , and  $I$  correspond to the number of substitutions, deletions, and insertions, respectively.  $N$  is the total number of habitats in the reference. If a reference entity does not match exactly or partially with any of the predicted entities, then this corresponds to a deletion, i.e., to a false negative. On the other hand, if a predicted entity does not match exactly or partially with any of the reference entities, then this corresponds to an insertion, i.e., to a false positive.  $D$  and  $I$  are the numbers of false negatives and false positives, respectively.

$$SER = \frac{S + D + I}{N} \quad (3.1)$$

The computation of  $S$  is shown in Equation 3.2.

$$S = 1 - M \quad (3.2)$$

Here,  $M$  is the similarity between two entities. It is computed by using Equation 3.3. The more similar two entities are, the lower their substitution score is.

$$M = J \cdot W \quad (3.3)$$

$J$  in Equation 3.3 is the Jaccard coefficient similarity between the predicted and reference entities [85]. If the boundary of the predicted entity is exactly the same as the boundary of the reference entity, then  $J$  equals 1 for the pair. The less the entities overlap, the lower the value of  $J$  is.  $W$  is a parameter that measures the similarity between the ontology concepts of the reference and the predicted entities [106]. It is based on the Jaccard coefficient of the sets of ancestors corresponding to the reference and predicted entities. The value of  $W$  is 1 if the predicted entity and the reference entity are assigned to the same concept in the ontology, and it is less than 1 if they are assigned to different entities. The higher the value of  $W$ , the more similar the two concepts are to each other.

**3.2.4.3. Results.** Table 3.1 shows the detailed results obtained on the test set for Sub-task 1 by the *Boun* and *Boun 2* systems. The workflows of both systems are the same (Figure 3.6), except the ontology expansion module, which is only available in the *Boun 2* system and a new additional rule for discontinuous entity handling. Both systems perform discontinuous entity handling, and neither of them perform entity modifier handling. These results show that expanding the OntoBiotope ontology using the training set, did not lead to improvements in the performance of the system. Since the concepts in the ontology are enriched by including more synonyms, more entities in the test set are matched to their concepts in the ontology. This resulted in a lower number of false negatives (i.e., lower D) and higher number of matches, which leads to higher recall and F-score values. While the SER value does not change, due to the increase in the number of false positives (i.e., insertions), the precision of the system decreases.

Table 3.1. Detailed results on the test set for Sub-task 1 (*Entity Boundary Detection & Ontology Categorization*)

<b>Evaluation Metrics</b>	<b>Boun</b>	<b>Boun 2</b>
<b>S</b>	112.70	115.24
<b>I</b>	141	158
<b>D</b>	89	74
<b>M</b>	305.30	317.75
<b>SER</b>	0.68	0.68
<b>Recall</b>	0.60	0.63
<b>Precision</b>	0.59	0.57
<b>F-score</b>	0.59	0.60

Table 3.2 presents a comparison of the results obtained by the *Boun* and *Boun 2* systems, and the other systems that participated in the Bacteria Biotope 2013 Sub-task 1. The *Boun* system that we submitted to the official evaluation ranked second among four systems in terms of the SER evaluation metric. The *Boun 2* system also achieves a SER value (68%) which is close to the *LIPN* system that ranked first in the shared task. In addition, the precision and recall values of the *Boun* and *Boun 2* systems are

relatively more balanced compared to the other systems except the *LIPN* system.

Table 3.2. Comparison with the other systems that participated in the BB Sub-task 1 (*Entity Boundary Detection & Ontology Categorization*). The results obtained on the test set are reported.

<b>System</b>	<b>SER</b>	<b>Recall</b>	<b>Precision</b>	<b>F-score</b>
<b>LIPN</b>	0.66	0.61	0.61	0.61
<b>Boun</b>	0.68	0.60	0.59	0.59
<b>LIMSI</b>	0.68	0.35	0.62	0.44
<b>Boun 2</b>	0.68	0.63	0.57	0.60
<b>IRISA</b>	0.93	0.72	0.48	0.57

Table 3.3 shows the effect of the discontinuous entity handling (DEH) module. The first column displays the results obtained by the *Boun 2* system, whereas the second column shows the results obtained by removing the discontinuous entity handling module from the system. These results demonstrate that performing discontinuous entity handling leads to a lower SER value, i.e., to a better performance on the development set. On the other hand, the discontinuous entity handling module does not make any particular change in the SER values of the system on the training and test sets.

Table 3.3. Effect of discontinuous entity handling (DEH). The results are reported on the training, development, and test sets.

	<b>Boun 2</b>	<b>Boun 2 - DEH</b>
<b>SER Train</b>	0.66	0.67
<b>SER Dev</b>	0.67	0.68
<b>SER Test</b>	0.68	0.68

Table 3.4 demonstrates the effect of the entity modifier handling module. The first row presents the results obtained by the *Boun 2* system, whereas the subsequent rows show the results obtained by extending the *Boun 2* system by including a mechanism to handle the modifiers attached to the noun phrases with the prepositions *in*, *of*, and *with*. Due to the fact that the SER values obtained by the system with the entity



modifier handling module are not lower than the *Boun 2* system for the training and development sets, this module is not included to the final system. The results reveal that the introduced entity modifier handling approach reduces the performance of the system, due to the prepositional phrase attachment ambiguity problem. For example, consider the sentence “*This species was isolated from a Lyme disease patient in Europe*”. Our entity modifier handling approach correctly identifies the habitat “*Lyme disease patient in Europe*” by extending the “*Lyme disease patient*” noun phrase with its modifier “*in Europe*”. However, given the sentence “*This species was isolated from a Lyme disease patient in 1993*”, the habitat is incorrectly identified as “*Lyme disease patient in 1993*”. The prepositional phrase “*in 1993*” is incorrectly attached to the noun phrase, whereas it should have been attached to the verb “*isolated*”. Handling complex nominals and resolving such prepositional phrase attachment problems can be possible by using a full syntactic parser, rather than a partial parser.

Table 3.4. Effect of entity modifier handling. The results are reported on the training, development, and test sets.

	SER Train	SER Dev	SER Test
<b>Boun 2</b>	0.66	0.67	0.68
<b>Boun 2 + in</b>	0.68	0.67	0.70
<b>Boun 2 + of</b>	0.72	0.72	0.72
<b>Boun 2 + with</b>	0.67	0.67	0.68

### 3.2.5. Results for the BioNLP Shared Task 2016 Data Set

This section provides the evaluation results obtained by the BOUN 2 system (Boun 2) on the BioNLP Shared Task 2016 Data Set.

**3.2.5.1. Data Set.** The training, development, and test sets provided by the BB shared task organizers contain 71, 36, and 54 documents, respectively. The gold standard annotations for the training and development sets were provided to the participants, whereas the evaluations on the test set were performed by using the online evaluation tool released by the shared task organizers. The documents in the corpus consist of

scientific abstracts obtained from the PubMed database.

**3.2.5.2. Evaluation Metrics.** The main evaluation metric slot error rate (SER), which is used for Sub-task 1 of BioNLP Shared Task 2013, is also used for BB-cat+ner Task of BioNLP Shared Task 2016. Lower SER values denote better performance, since SER is an error measure. Precision, recall and f-score measures are used for the evaluation of BB-event task. Higher values denote better performance.

**3.2.5.3. Results.** Table 3.5 shows the detailed results obtained on the test set for BB-cat+ner task by the Boun 2 system. Table 3.6 presents a comparison of the Boun 2 systems with the official results of the other teams that participated in the BioNLP Shared Task 2016 Bacteria Biotope Task BB-cat+ner Sub-task. The results show that the Boun 2 system ranks second in terms of SER value, while achieving better precision and recall scores than the other systems. Although Boun 2 was developed by considering documents that are web pages written for the general public, for scientific abstracts it is able to achieve comparable performance without any adaptation.

Table 3.5. Results for BB-cat+ner-task (*Entity Recognition and Categorization*) on BioNLP Shared Task 2016 Data Set. The results obtained on the test set are reported.

System	Habitats Only	Ignore Boundaries	Multiple Normalizations
<b>SER</b>	0.82	0.71	0.50
<b>Mismatches</b>	122.6	55.75	12.20
<b>Matches</b>	202.4	269.2	25.8
<b>Insertions</b>	92.0	92.0	0
<b>Deletions</b>	697.0	296.0	14
<b>Recall</b>	0.33	0.43	0.50
<b>Precision</b>	0.49	0.65	0.68
<b>Predictions</b>	417.0	417.0	38.0

Table 3.6. Comparison of BB-cat+ner-task Results (*Entity Recognition and Categorization*) on BioNLP Shared Task 2016 Data Set. The results obtained on the test set are reported. MM:Mismatches M:Matches I:Insertions D:Deletions R:Recall

P:Precision Pred:Predictions

<b>System</b>	<b>SER</b>	<b>MM</b>	<b>M</b>	<b>I</b>	<b>D</b>	<b>R</b>	<b>P</b>	<b>Pred</b>
<b>TagIt</b>	0.775	199.960	188.040	49	233	0.30	0.43	437
<b>BOUN 2</b>	0.821	122.600	202.400	92	697	0.33	0.49	417
<b>LIMSI</b>	0.862	192.307	152.693	67	276	0.25	0.37	412
<b>whunlp</b>	0.950	226.358	119.642	89	275	0.19	0.28	435

## 4. ENTITY NORMALIZATION USING WORD EMBEDDINGS AND SYNTACTIC ANALYSIS

### 4.1. Background

Currently, the vast majority of the biomedical resources are in unstructured form which originate from an assortment of contrary sources that incorporate nonstandard naming conventions, which makes the required information difficult to use and understand [107]. Ontologies help researchers to overcome these kinds of difficulties and help researchers facilitate the vast amounts of biomedical knowledge available [108]. An ontology can provide a unique identifier for describing information for each entity, which solves the heterogeneity problem and provides standardized and homogeneous data [109]. Linking named entities in text through an ontology is an essential process to make sense of the identified named entities [3].

In the Figure 1.1, the association between the entity mention “*pediatric*” and the ontology concept term name “*pediatric patient*” can be relatively more easily detected due to the lexical similarity between them. Similarly, the habitat mention “*respiratory*” and the ontology concept “*respiratory tract part*” also share a common word, making them lexically similar. However, lexical similarity may not always exist between entity mentions and concept term names or concept synonyms. For example, there is no lexical similarity between the habitat mention “*children less than 2 years of age*” and ontology concept term name “*pediatric patient*”, which calls for the utilization of semantic similarity.

Even if the named entities are given, linking the identified named entities to a unique concept identifier in an ontology/dictionary is not a trivial task in the biomedical domain. There are many challenges in the task of named entity linking through an ontology or a dictionary, two of which are the variety and ambiguity problems of the named entities [110]. A named entity may appear in different surface forms in a given text, which is called the variety problem. Furthermore, two named entities with the

same surface form may have different semantic meanings, which is called the ambiguity problem. Linking of named entities for the biomedical domain has another big challenge besides these two common problems in the general natural language processing domain. In the biomedical domain, the training data is relatively smaller and the number of the ontology/dictionary categories that should be considered is larger compared to many other domains in natural language processing [52]. This poses a challenge for the standard supervised classification algorithms. For example, there are 2,221 semantic categories in the Onto-Biotope ontology, while the available training set contains only 747 entity mentions, and 16,295 words. For adverse drug reaction normalization, this situation is worse since there are 22,499 MedDRA dictionary terms.

In this part of the thesis, for the ontology based normalization of the named entity mentions in text, we propose an unsupervised approach, which utilizes both semantic and syntactic information. The proposed approach uses word embeddings learned from large unlabeled text to capture semantic information and syntactic parsing information to re-rank the candidate ontology/dictionary concept terms. The proposed approach is tested on two different data sets, which are the BioNLP Shared Task 2016 Bacteria Biotopes (BB3) categorization sub-task data to normalize habitat entities through the Onto-Biotope ontology and the Text Analysis Conference 2017 Adverse Drug Reaction data to normalize adverse drug reaction mentions through the MedDRA dictionary. On both data sets, the proposed normalization method with syntactic re-ranking achieved better performance than the normalization method without syntactic re-ranking. Furthermore, we obtained the new state-of-the-art results with 2.9 percentage points above the previous best result for the Bacteria Biotopes (BB3) categorization sub-task.

## 4.2. Related Work

Several approaches have been proposed for biomedical entity normalization for different types of biomedical entities including genes/proteins [7–10], bacteria biotopes [15, 48, 49, 52, 53], and diseases [11, 12]. Early systems tried to link the entity mentions to the knowledge base entities by utilizing dictionary look-up and string matching algorithms [7, 54]. Some studies [11, 15] used hand-written rules to measure the mor-

phological similarity between entity mentions and ontology/dictionary entities, while others [56] automatically learned patterns of variations of the entities. Supervised machine-learning based approaches, which learn the similarities between biomedical entity mentions and ontology concept names from labeled training data have also been proposed and applied as a solution to the normalization task of various biomedical entities such as diseases [12].

Most previous studies focused on utilizing morphological information for named entity normalization. However, morphological similarity alone is not adequate to normalize biomedical entities, which generally have forms different from the concept terms that they should be tagged with [52]. Word embedding models, which learn distributional representations of words from large unlabeled corpora, are promising approaches for capturing semantic information [61]. They have been successfully used in several recent NLP tasks including the biomedical domain [62, 69–71]. Recently, word embeddings have also been used for the task of biomedical named entity normalization. Li *et al.*, (2017) proposed a convolutional neural network (CNN) architecture leveraging semantic and morphological information, which handles the biomedical entity normalization task as a ranking problem [58]. In the proposed method, firstly candidates are generated using hand-crafted rules, and then they are ranked according to semantic and morphological information, which are represented by a CNN-based model. Experiments on two benchmark datasets (the ShARe/CLEF eHealth dataset and the NCBI disease dataset) showed that semantic information is beneficial for the biomedical entity normalization task as well as morphological information. However, the requirement of hand-crafted rules and labeled data makes the adaptation of this method to different domains harder and time-consuming. Cho *et al.*, (2017) proposed a semi-supervised approach that facilitates word embeddings to represent semantic spaces for normalizing biomedical entities such as disease names and plant names and obtained promising performance [59]. This method requires a domain specific corpus and dictionary. Therefore, the adaptation of it to other domains is not easy, if there are no such resources available.

A number of community-wide challenges including the BioCreative Challenges [41–45] and BioNLP Shared Tasks [46–49], which have been conducted to assist the progress of research in biomedical text mining, also addressed the task of biomedical entity normalization. The Bacteria Biotope task, whose ultimate aim is information extraction regarding bacteria and their habitats, was first addressed in the BioNLP Shared Task 2011 [47, 111], and has been conducted in 2013 [48, 52] and 2016 again since then. We evaluated our proposed approach on the BB-cat subtask of the 2016 edition of the Bacteria Biotope task, which addressed the normalization of habitat entity mentions in PubMed abstracts using the OntoBiotope ontology [49]. In the official task, the teams TagIt [112] and LIMSI [113] proposed rule-based methods, while BOUN [53] proposed a similarity-based method that utilizes both approximate string matching and cosine similarity of word-vectors weighted with Term Frequency-Inverse Document Frequency (TF-IDF). According to the official results, the best precision (62%) for habitat mention normalization was obtained by the BOUN system.

The bacteria habitat mention normalization problem continued to attract the attention of the researchers after the shared task. CONTES is a recently proposed semi-supervised method for linking habitat entity mentions through the Onto-Biotope ontology [114]. The system is based on word embeddings that are induced from PubMed by utilizing the Word2Vec tool. The cosine similarities between term vector representations and concept vector representations are calculated to find the most similar ontology concept to the given entity mention. They applied the proposed normalization method to the test dataset of the Bacteria Biotope 2016 Task 3 (BB-cat), and obtained comparable results to that of the state-of-the-art for the task of Bacteria Biotopes categorization. CONTES contains a transformation step to make comparable the term vectors and the entity vectors which are represented in different dimensions. The need for the transformation step makes the method semi-supervised, since it requires labeled data for training the prediction model. Recently, Mehryary et al. (2017) used TF-IDF weighted vector space representation for the named entity categorization of bacteria biotopes [115]. Each ontology concept name and each entity mention is represented with a TF-IDF weighted vector considering each concept name in the ontology as a separate document and calculating IDF weights based on these names. The ontology

concept with the highest cosine similarity is assigned to a given entity mention. Although they achieved state-of-the-art results in the normalization task, the TF-IDF based scheme has limitations in capturing the semantic relations between the ontology concepts and entity mentions, since it is primarily based on the surface forms of the words.

Besides the Bacteria Biotopes normalization task, we also evaluate our approach on the task of normalizing Adverse Drug Reaction (ADR) mentions in drug labels to the MedDRA terms. We use the recently provided data set from the Text Analysis Conference (TAC) 2017. Different types of data sources such as electronic health records [116], scientific publications, and social media data [117] and different types of lexicons such as the Unified Medical Language System (UMLS) [118] and the side effect resource (SIDER) [119] have been used to extract ADRs from text. Many of these studies proposed a lexicon-based matching approach for ADRs recognition. Although a number of studies have been conducted to automatically identify ADRs in text and map them through a dictionary using NLP techniques, as far as we know the normalization of the ADRs through a dictionary has not been studied as a separate task without named entity recognition.

### 4.3. Methods

We developed a semantic similarity based unsupervised method for entity linking through an ontology/dictionary, the work-flow of which is displayed in Figure 4.1. Given a set of documents with annotated named entities and a corresponding ontology, the normalization task is done in two steps. In the first step, the semantically most similar ontology concepts are generated as candidates, and in the second step, the candidates are re-ranked according to the syntactic-based weighted semantic similarities. The details of our approach are described in the following subsections.



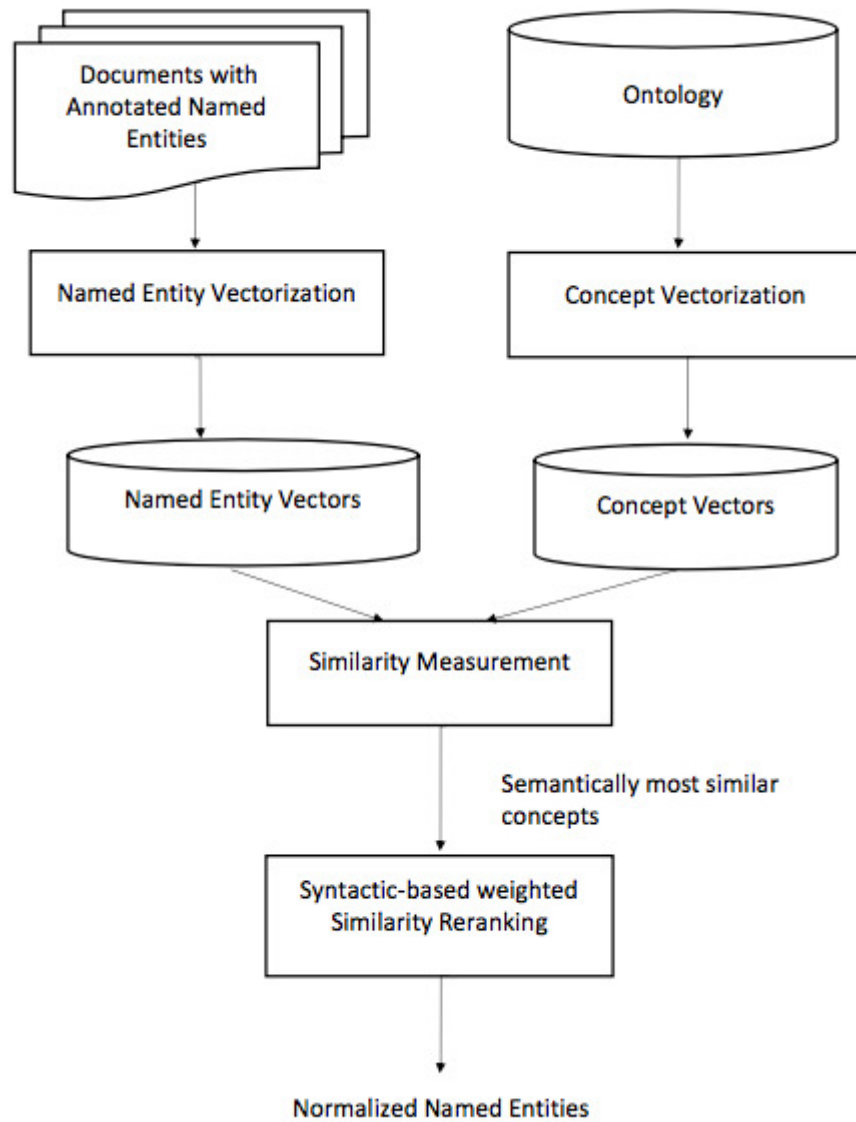


Figure 4.1. System Work-flow. Work-flow of the Named Entity Normalization System

#### 4.3.1. Data Sets

4.3.1.1. Bacteria Biotope Entity Normalization. In this study, we used the official data set that is provided by the BioNLP Shared Task 2016 organizers for the Bacteria Biotope categorization subtask. Since our proposed approach is unsupervised and does not require any training data, the training and development sets are used for error analysis during the development of the system, and the separate test set is used for evaluating the performance of the proposed system. The data set provided by

the shared task organizers was created by collecting titles and abstracts from PubMed, which contain general information about bacteria and habitats. The data set, consisting of 71 training, 36 development, and 54 test documents, was manually annotated by the bioinformaticians of the Bibliome team of MIG Laboratory at the Institut National de Recherche Agronomique (INRA) [49].

**4.3.1.2. Adverse Drug Reaction Normalization.** For Adverse Drug Reaction Normalization, we used the official data set that is provided by the Text Analysis Conference (TAC) 2017 organizers. The test set is used for evaluating the performance of the proposed system. The data set contains general information about drug labels consisting of 101 training and 99 test documents, which were manually annotated by the organizers.

### **4.3.2. Preprocessing**

In the preprocessing step, the annotated named entities and the ontology concept names with their synonyms are tokenized, and the stop words are removed from the named entity mentions and the ontology concept names. Furthermore, all non-ASCII characters are stripped from both the named entities and the ontology concept names.

### **4.3.3. Word representations**

Our proposed approach is mainly based on the assumption that semantically similar words have similar vector spaces. Based on this assumption, if the semantic similarity of named entity mentions and ontology concept terms can be computed, the most similar concept in the ontology can be assigned as the normalized concept to the named entity mention.

To compute the semantic similarity, each word is represented in the vector space as a real-valued vector using a pre-trained word embedding model that is publicly available [62]. The model has been trained leveraging word vectors that were induced

from PubMed by the Word2Vec tool [61]. The trained model is applied to each word to obtain the corresponding word vector. We used the model variant with window size of 30, since it has been shown to obtain higher performance in the biomedical concept similarity and relatedness tasks in [62].

#### 4.3.4. Identifying the Semantically Similar Ontology Concepts

The vectors of the ontology concept terms and the reference named entities (i.e., the named entity mentions in text) are computed in the same way as described below. For each word in the named entities and ontology concept terms, the vector representations are obtained by the pre-trained model as explained in the previous subsection. For the multi-word named entities and ontology concepts, the vector representations are computed by averaging the vectors of their composing words. Figure 4.2 presents the computation of the vector representation for a sample multi-word named entity “*a day-care center*” and shows how the averaging is done. In the preprocessing step, the stop-word “*a*” and the hyphen character are removed. The tokens “*day*”, “*care*”, and “*center*” are considered and used for averaging to compute the vector representation of the multi-word named entity. Each token is represented with a real-valued vector using the pre-trained word embedding model that is explained in the previous subsection. The real-valued vectors of the tokens comprising the multi-word entity mention are summed to create a real-valued vector, which is called  $\vec{sum}$ . At the end,  $\vec{sum}$  is divided by the number of tokens other than the stop-words, which is 3 for the example entity mention, to obtain a normalized real-valued vector for the multi-word named entity.

For each reference entity and for each ontology concept term, a cosine similarity score is calculated to get the semantic similarity between the related entity and the ontology concept term. Since the vectors of ontology concept terms and reference named entities are computed in the same way, unlike the CONTES system, there is no need for a transformation step for the vectors in order to compute the similarity between them. For each reference entity, ontology terms are ranked according to the semantic similarity scores, the top  $k$  of which are the candidates for syntactic weighting

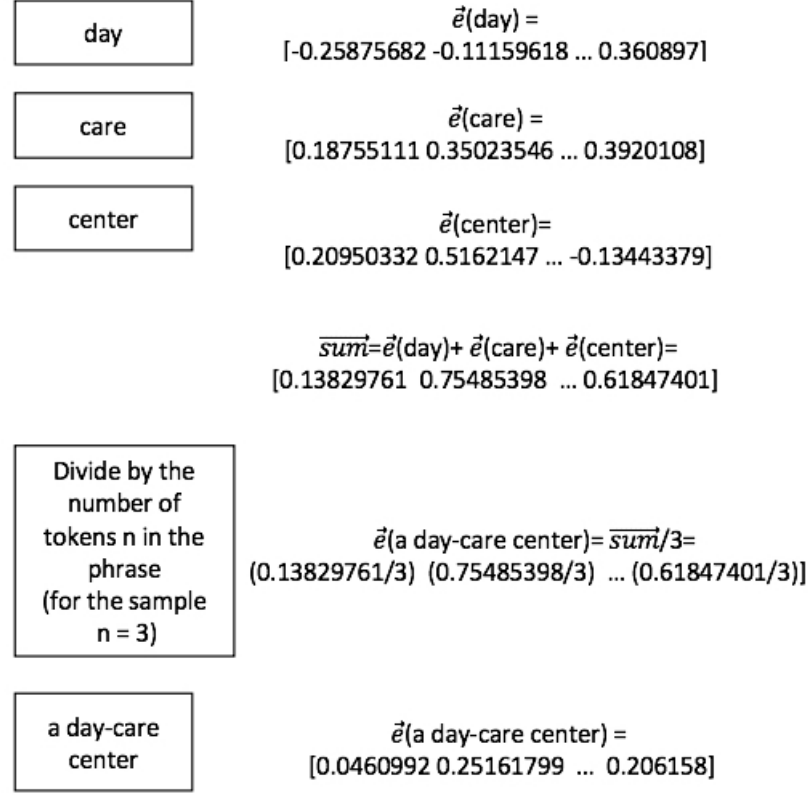


Figure 4.2. Sample multi-word expression. Computation of the corresponding real-value vector for a sample multi-word expression “*a day-care center*”, where  $\vec{e}(t)$  is the word embedding vector for token  $t$

based re-ranking.

We also investigated using word mover’s distance (WMD), instead of cosine similarity. WMD is a distance metric which represents text documents as a weighted point cloud of embedded words and computes the distance between documents as the minimum cumulative distance that words from a document need to travel to another [120]. It is based on the idea that documents without common words may convey similar meanings and bag-of-words (BOW) is not enough to detect this kind of similarity.

#### 4.3.5. Syntactic Re-ranking

Our system without syntactic analysis is not adequate alone to normalize entity mentions like “*children attending a day-care center*”. Table 4.1 (Before re-ranking

part) shows the output of our system without syntactic re-ranking for the sample entity mention “*children attending a day-care center*”. The semantically most similar concepts to the mention are found as “*OBT:001423 medical center*”, “*OBT:001801 clinic*”, and “*OBT:000259 research and study center*”, which are false positives. The correct concept is “*OBT:002146 child*”, which is very similar to the head word “*children*” of the mention “*children attending a day-care center*”. As this example shows, if the system can identify the most informative word in the reference entity mention, the correct concept can be assigned to it (see Table 4.1 (After re-ranking part)).

We proposed a re-ranking module based on syntactic parsing to identify the correct concept from among the top  $k$  candidates returned by the word-embedding based similarity ranking. The re-ranking module makes use of the Stanford Parser (version 3.8.0) [121] to detect the most informative word in the reference entity mention. It computes the semantic similarity between the most informative words of the reference mention and the candidate ontology concept, and re-ranks the top  $k$  semantically most similar concepts.

Table 4.1. Semantically most similar concepts to the entity mention “*children attending a day-care center*” with/without re-ranking.

Before Re-ranking		
Rank	Concept	Similarity score
1	OBT:001423 medical center	0.8297
2	OBT:001801 clinic	0.7917
28	OBT:002146 child	0.6979
After Re-ranking		
Rank	Concept	Similarity score
1	OBT:002146 child	0.7484
3	OBT:001801 clinic	0.6519
24	OBT:001423 medical center	0.5460

The intuition behind our re-ranking approach is that the entity mentions are noun phrases and the heads of the noun phrases are the most informative words in the mentions. To obtain the corresponding head words, the part-of-speech tags and

```
(ROOT
  (FRAG
    (NP (NNS children))
    (S
      (VP (VBG attending)
        (NP (DT a) (NN day) (NN care) (NN center))))))
```

Figure 4.3. Sample syntactic parse Syntactic parse of the Stanford Parser for the sample named entity mention “*children attending a day-care center*”

syntactic parses of the entity mentions are required. We used the Stanford Parser by providing the entity mentions as input and obtaining the syntactic parses composed of their constituent phrases as output. Next, the syntactic parses are processed to find the most informative words in the mentions by utilizing the algorithm whose pseudo-code is given in Figure 4.5. According to this algorithm, the top level rightmost “*noun*” is searched in the tree structured syntactic parse and assigned as the head of the mention phrase. For example, for the sample mention “*children attending a day-care center*”, the Stanford Parser generates the syntactic parse, which is shown in Figures 4.3 and 4.4. Figure 4.3 demonstrates the syntactic parse with its constituent phrases and Figure 4.4 shows the tree view. The head of the sample mention is found as “*children*” and the head of the concept name “*OBT:001423 medical center*” is found as “*center*” by leveraging the algorithm.

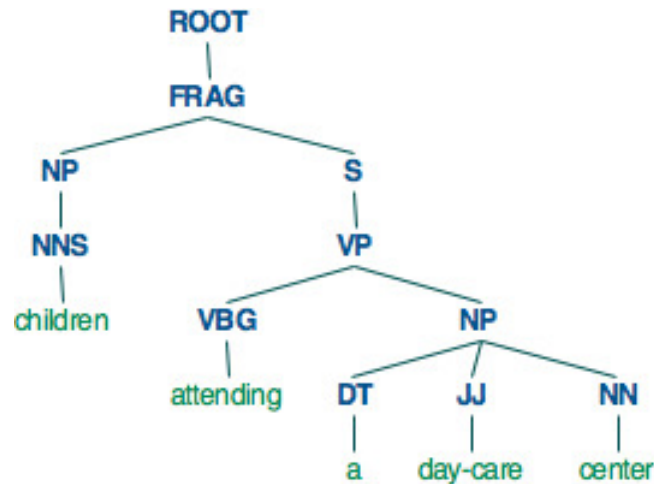


Figure 4.4. Tree view of the sample parse Tree view of the syntactic parse of the sample named entity mention “*children attending a day-care center*”

After the detection of the head words of the phrases as “*children*” for the “*children attending a day-care center*” entity mention and “*center*” for the “*OBT:001423 medical*

---

**Algorithm 1** Algorithm head finder for the entity mentions
 

---

**Input:** *strparse* = syntactic parse of the entity mention

**Output:** *most\_informative\_word* = head word of the entity mention

```

1: procedure FINDHEAD(strparse)
2:   all_np_pares  $\leftarrow$  extract all NPs from strparse
3:   for all np_parse in all_np_pares do
4:     return FINDHEADOFNP(np_parse)

```

---

**Input:** *np\_parse* = syntactic parse of a noun phrase in the entity mention  
**Output:** *head* = head word of the noun phrase whose syntactic parse is *np\_parse*

```

5: procedure FINDHEADOFNP(np_parse)
6:   noun_tags_list  $\leftarrow$  ["NN", "NNS", "NNP", "NNPS"]
7:   np_subtrees  $\leftarrow$  the subtrees of np_parse
8:   top_level_nouns  $\leftarrow$  noun sublist of np_subtrees
   , where for each member subtree, the root label is in noun_tags_list
9:   if len(top_level_nouns) > 0 then
10:    return top_level_nouns[rightmost]
11:  else
12:    top_level_nps  $\leftarrow$  noun phrase sublist of np_subtrees
   , where for each member subtree, the root label is NP
13:    if len(top_level_nps) > 0 then
14:      return FINDHEADOFNP(top_level_nps[rightmost])
15:    else
16:      nouns  $\leftarrow$  all nouns in np_parse
17:      if len(nouns) > 0 then
18:        return nouns[rightmost]
19:      else
20:        np_parse_leaves  $\leftarrow$  all terminal words in the np_parse
21:        return np_parse_leaves[rightmost]

```

---

Figure 4.5. Pseudo-code Algorithm for finding the most informative word in an entity mention whose syntactic parse is given as input. NP: Noun Phrase; NN: Noun singular; NNS: Noun plural ;NNP: Proper noun singular; NNPS: Proper Noun plural

*center*” ontology concept name, the semantic similarities are recomputed based on these new information. The similarity scores of the concepts with unrelated head words (e.g. “OBT:001423 medical center”) will be lower and those of concepts with related head words (e.g. “OBT:002146 child”) will be higher after the re-ranking phase (see Table 4.1).

The mathematical formulation of the syntactic weighting based similarity used for re-ranking is shown in Equation 4.1, where  $S_{RR}(m, c)$  is the final computed similarity between mention  $m$  and 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 4.3.4, and  $w$  is a weighting parameter which can take values between 0 and 1.

$$S_{RR}(m, c) = (w * S_S(m_{head}, c_{head})) + ((1-w) * S_S(m, c)) \quad (4.1)$$

#### 4.4. Results and Discussion

In this section, the results of the proposed systems both with and without re-ranking are presented. In addition, comparison with prior work is performed.

##### 4.4.1. Evaluation Metrics

4.4.1.1. Evaluation for Bacteria Biotopes. For evaluation of the bacteria biotopes entity normalization predictions, we used the official on-line evaluation service to compute the precision score, which is the official measure used to rank the submissions in the BioNLP Shared Task 2016 Bacteria Biotopes categorization sub-task.

In the BioNLP Shared Task 2016 Bacteria Biotopes categorization sub-task, entities have been given and the participants were required to predict the normalization of the entities. In the official on-line evaluation, for each normalized Habitat entity, Wang similarity  $W$  [122] is calculated with  $s = 0.65$  to measure the similarity between the reference and the predicted normalization. Wang similarity is the Jaccard index between the two sets of the predicted and the reference concept ancestors with a weighted factor  $d^s$ , where  $d$  is the distance between the corresponding concept and the ancestor, and  $s$  is a parameter between 0 and 1. The submissions are evaluated with their Precision values:



$$Precision = \sum S_p / N \quad (4.2)$$

where  $S_p$  is the total Wang similarity  $W$  for all predictions [49], and  $N$  is the number of predicted entities.

4.4.1.2. Evaluation for Adverse Drug Reaction. For evaluation of the adverse drug reactions entity normalization predictions, we computed the macro-averaged and micro-averaged scores for precision, recall and f-score measures. True positives (TP), false positives (FP), and false negatives (FN) are calculated by comparing the predicted normalization concept with the reference normalization concept in the gold standard via exact matching.

To compute Micro-average scores, the true positives, false positives, and false negatives of the system are summed up for all drug labels to get the statistics (Equations 4.3 and 4.4).  $N$  is the total number of drug labels in the data set.

$$Micro\text{-}average\ Precision = \frac{\sum_{c=1}^N (TP_c)}{\sum_{c=1}^N (TP_c + FP_c)} \quad (4.3)$$

$$\text{Micro-average Recall} = \frac{\sum_{c=1}^N (TP_c)}{\sum_{c=1}^N (TP_c + FN_c)} \quad (4.4)$$

On the other hand, the macro-averaged scores are computed as the average of the individual precision and recall values obtained on each drug label (Equations 4.5 and 4.6).

$$\text{Macro-average Precision} = \frac{\sum_{c=1}^N (Precision_c)}{(N)} \quad (4.5)$$

$$\text{Macro-average Recall} = \frac{\sum_{c=1}^N (Recall_c)}{(N)} \quad (4.6)$$

#### 4.4.2. Results

4.4.2.1. Bacteria Biotopes. Table 4.2 shows the results of our proposed approach with and without syntactic re-ranking. The results show that the system with the syntactic re-ranking module achieves a higher performance. Recall that the proposed system

without re-ranking computes the vector representations for the multi-word entities by averaging the vectors of their composing words. On the other hand, the proposed system with syntactic re-ranking computes the vector representations by giving higher weights to the head words. This means that instead of averaging the vector representations, giving higher weights to the most informative words is a more suitable way for vector representations of multi-word entities.

Table 4.2. Results for the system with and without syntactic re-ranking. Precision values for the training and development data sets are reported.  $k$  is set to 5 and  $w$  is set as 0.25 for the re-ranking module.

System	Train	Dev
Before Re-ranking	0.601	0.629
After Re-ranking	0.648	0.677

Table 4.3. Comparison with previous 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 (BOUNEL) based on the results on the training and development sets.

System	Precision
BOUNEL(Our system)	0.659
TURKU [115]	0.630
BOUN [53]	0.620
CONTES [114]	0.597
LIMSI [113]	0.438
BASELINE-2	0.322
BASELINE-1	0.225

Table 4.3 presents a comparison of the proposed system, named as BOUNEL (BOUN Named Entity Linker), with the prior work on the task of habitat named entity normalization. We compared our results with the previous systems that are tested on the BioNLP Shared Task 2016 BB cat subtask test set. We computed two different baseline results; the BASELINE-1 assigns the exact match of the term in the ontology. In case of non-existence of an exact match, BASELINE-1 assigns the term to the root concept of the Onto-Biotope ontology hierarchy, which is “bacteria habitat”

concept. On the other hand, BASELINE-2 assigns all terms to the “bacteria habitat” concept without searching for an exact match. The results show that our system obtained a score of 65.9% which is higher than both of the baselines BASELINE-1 and BASELINE-2. Our proposed method also obtained higher scores than all other previously proposed methods on the bacteria biotope normalization task, achieving the new state-of-the-art results.

4.4.2.2. Adverse Drug Reactions. Table 4.4 presents the results of the proposed system before and after syntactic re-ranking for the task of adverse drug reactions entity normalization on the Text Analysis Conference 2017 Adverse Drug Reaction training and test data sets. We used the same values for the parameters of the re-ranking module as the bacteria biotope normalization task ( $k=5$  and  $w=0.25$ ). Since there is no prior work on the task of adverse drug reactions entity normalization task on the same data set, we compared our results with the baseline. We computed baseline results by assigning the mention to the exact match of the term in the MedDRA dictionary. As the results on Table 4.4 demonstrate, the new system with syntactic re-ranking obtained higher precision, recall, and f-measure scores on both the training and test data sets than the system without syntactic re-ranking. Furthermore, the new system with syntactic re-ranking achieved significantly higher recall than the baseline, as a result achieving higher f-measure scores.

### 4.4.3. Discussion

4.4.3.1. Bacteria Biotopes. Table 4.5 shows the performance of the proposed system without syntactic re-ranking for returning the correct concept from the ontology among the top  $k$  ranked candidates. For example, when  $k = 1$ , the concept assignment is considered correct, only if the correct concept is ranked first by the system. On the other hand, when  $k = 10$ , the concept assignment is considered correct, if the correct concepts is ranked in the top ten by the system. These results motivated the development of the re-ranking module, since as  $k$  increases, the precision of the system also increases. The goal of syntactic re-ranking is to re-rank the top  $k$  retrieved

Table 4.4. Results of the proposed method with/without re-ranking on the adverse drug reaction normalization task. Precision, recall and f-score values for the training and test sets are reported.

Training set			
	Baseline	Before Re-ranking	After Re-ranking
Macro-average Precision	0.999	0.737	0.742
Macro-average Recall	0.522	0.732	0.736
Macro-average F-score	0.686	0.735	0.739
Micro-average Precision	0.999	0.728	0.730
Micro-average Recall	0.513	0.723	0.725
Micro-average F-score	0.665	0.726	0.728
Test set			
	Baseline	Before Re-ranking	After Re-ranking
Macro-average Precision	0.999	0.683	0.687
Macro-average Recall	0.494	0.677	0.681
Macro-average F-score	0.661	0.675	0.684
Micro-average Precision	0.999	0.682	0.686
Micro-average Recall	0.489	0.675	0.680
Micro-average F-score	0.657	0.678	0.684

candidate concepts, so that the correct concept moves to the first rank, as in the example shown in Table 4.1.

Table 4.5. Prediction performance of our system without syntactic re-ranking among the semantically most similar top ( $k = 1, 5, 10, 20, 25, 50$ ) concepts. Precision values for the training and development data sets are reported when the reference concept is among the top  $k$ .

k	1	5	10	15	20	25	50
Train	0.614	0.656	0.672	0.711	0.726	0.738	0.831
Dev	0.655	0.683	0.725	0.753	0.789	0.804	0.894

Table 4.6 demonstrates the results of our proposed approach with syntactic re-ranking, when the top  $k$  candidates retrieved by the system without re-ranking are

Table 4.6. Results for the system with syntactic re-ranking for the different semantically most similar top ( $k = 5, 10, 15, 20, 25, 50$ ) concepts. Precision values for the training and development data sets are reported when the reference concept is at the first rank after re-ranking the semantically most similar top ( $k = 5, 10, 15, 20, 25, 50$ ) concepts.

k	5	10	15	20	25	50
Train	0.648	0.634	0.637	0.639	0.640	0.643
Dev	0.677	0.668	0.667	0.667	0.668	0.632

provided as input to the re-ranking module. As the results show, for values of  $k = 10$ ,  $k = 15$ ,  $k = 20$  and  $k = 25$ , the results are nearly the same on the training and development sets, which means that after a threshold of  $k = 5$ , different values of  $k$  make no big difference in the results. Therefore, based on the results on the training and development sets,  $k$  is chosen as 5 empirically.

We also investigated the effects of using different similarity/distance metrics, word mover’s distance (WMD) and cosine similarity. The results show that the system with cosine similarity achieved better precision scores than the system with WMD on both the training (WMD: 58.6%; Cosine: 60.1%) and development (WMD: 49.0%; Cosine: 62.9%) data sets.

Table 4.7. Results for the system with different weights for the most informative words ( $w = 0, 0.25, 0.50, 0.75$ ). Precision values for the training and development data sets are reported.

w	Train	Dev
0	0.614	0.655
0.25	0.648	0.677
0.50	0.648	0.669
0.75	0.632	0.661

Table 4.7 shows the effect of the parameter  $w$ , which is used in Equation 4.1 to give weights to the most informative words (head of the noun phrase) with the ultimate aim to calculate the similarity between the named entity mention phrases

and the reference ontology terms. As the results show, for  $w = 0.25$  our proposed approach obtains higher precision on both the training and the development sets.

During the error analysis of the proposed system with syntactic re-ranking on the training and development sets, we realized the existence of falsely normalized mentions, which are possessive prepositional phrases (PPP). These phrases include compound noun phrases in the “*NP of NP*” form. For example, the entity mention “*throats of two healthy children*” is composed of two noun phrases “*throats*” and “*two healthy children*”, where the first NP “*throats*” is the only informative NP for normalizing the entity mention to the correct concept “*OBT:000374 throat*”. As a result of this fact, a syntax rule is added before re-ranking to strip the non-informative words following “of” from the entity mentions, if they are possessive prepositional phrases.

4.4.3.2. Adverse Drug Reactions. Although experimental results showed that the new system with syntactic re-ranking obtained higher precision scores on both data sets than the system without syntactic re-ranking, the improvement of the new system on the Text Analysis Conference 2017 Adverse Drug Reaction (ADR) data set is lower compared to the improvement that is achieved on the BioNLP Shared Task 2016 Bacteria Biotopes data set. Error analysis revealed two main sources of errors, which are more prevalent in the ADR data set. The first source of errors is the usage of abbreviations and acronyms as entity mentions, which are hard to normalize without incorporating the context of the mentions. For example, in the training set, there are entity mentions such as “*sjs*” and “*ten*”, which are acronyms that should be normalized to the corresponding concepts “*Stevens-Johnson syndrome*” and “*Toxic epidermal necrolysis*” in the MedDRA dictionary. Rare words are the second source of errors. Although the word embedding model, which is used to calculate the semantic similarities, has been learned from PubMed articles, there may still exist out of vocabulary words, which are rare. For example, for the ADR mention “*Neoscytalidium infections*”, the “*Neoscytalidium*” word does not exist in the model that is used to calculate the word embeddings. In that case, the semantically most related concepts are found incorrectly by the proposed system considering only the existing word “*infections*” as “*Nosocomial*

*infection*”, “*Opportunistic infection*” and “*Granulicatella infection*”, while the correct concept is “*Neoscytalidium infection*”.



## 5. APPLICATIONS

### 5.1. An application for the Bacteria Biotopes domain

#### 5.1.1. Retrieval of the related abstracts

An abstract retrieval module is implemented to automatically download the abstracts of the articles related to bacteria habitats from PubMed. We searched in PubMed for “*bacteria*”, which returned 2,141,243 documents (Search date: December 2018). The first 1,000 abstracts from this set of documents are automatically downloaded for further processing.

#### 5.1.2. Preprocessing

Firstly, each input file is split into sentences using the Genia Sentence Splitter (GeniaSS) [123]. The outputs of the splitter are given to the Genia Tagger [97, 105] as input files with the aim of obtaining the lemmas, the part-of-speech (POS) tags, and the constituent categories of the words in the given biomedical text (e.g., surface form: ticks; lemma: tick; POS tag: NNS; phrase structure: I-NP). We utilized these syntactic information at the following steps of our system.

#### 5.1.3. Named Entity Tagging

We assume that bacteria habitats are embedded in text as noun phrases, and all noun phrases are possible candidates for habitat entities and bacteria entities.

The Noun Phrase Extractor and Simplifier module firstly detects the noun phrases in the text by using the Genia Tagger and then post-processes these noun phrases by using the syntactic rules that are explained in detail in the previous chapters. To determine whether a candidate noun phrase is a habitat entity or not, the Habitat Name Recognizer module searches all ontology entries, which compose the OntoBiotope

Ontology, to find an exact match with the candidate noun phrase or with parts of it. In the same way, Bacteria Name Recognizer module searches all the ontology entries, which compose NCBI Taxonomy. In this step, the names, exact synonyms, and related synonyms of ontology entries (ontology entry features) are compared with the candidate noun phrase. After running this module, an output file, which contains the predicted habitat entities and their positions in the input text, and a corresponding output file for the predicted bacteria entities and their positions in the input text, are created.

#### **5.1.4. Named Entity Normalization**

While our system detects entities and their boundaries (as explained in detail in Chapter 3.2), it also assigns ontology concepts to the retrieved entities. Bacteria entities and habitat entities are normalized respectively through NCBI Taxonomy and OntoBiotope ontology.

Although promising results are obtained for the ontology normalization of habitat entities by using the approach, which is explained in detail in Chapter 3.2, the syntax rules makes the adaptation to different biomedical entities harder. For the normalization of habitat entities, our proposed approach (explained in detail in Chapter 4), which is mainly based on the assumption that semantically similar words have similar vector spaces, is utilized. Based on this assumption, the semantic similarity of habitat entity mentions and ontology concept terms are computed, and the most similar concept in the ontology is assigned as the normalized concept to the habitat entity mention.

#### **5.1.5. Relation Extraction**

We propose two methods for identifying bacteria habitat localization relations. The underlying assumption for the first method is that discourse changes with a new paragraph. Therefore, it operates on a paragraph-basis. The second method performs a more fine-grained analysis of the text and operates on a sentence-basis. We also develop a novel anaphora resolution method for bacteria coreferences and incorporate

it with the sentence-based relation extraction approach.

We participated in the Bacteria Biotope (BB) Task of the BioNLP Shared Task 2013. Our system ranked third with an F-score of 27% in Sub-task 2 (Localization Event Extraction). At this part of the thesis, we report the system that is implemented for the shared task, including the novel methods developed and the improvements obtained after the official evaluation. The extensions include the novel sentence-based relation extraction method incorporated with anaphora resolution for Sub-task 2. These extensions resulted in state-of-the-art performance for Sub-task 2 with an F-score of 53%.

Our results show that the newly developed sentence-based relation extraction system with the anaphora resolution module significantly outperforms the paragraph-based one, as well as the other systems that participated in the BB Shared Task 2013.

5.1.5.1. Related Work. In this section, previous work that is related to the extraction of relations between bacteria entities and habitat entities (Localization Relation Extraction) and of relations between two habitat entities (Part Of Relation Extraction) are covered in detail.

The participants of the first shared task (The BioNLP Shared Task 2011), which targets the extraction of information about bacteria and their habitats, UTurku and JAIST adapted machine learning approaches for detecting the Localization and Part-of relations among bacteria and habitats. On the other hand, another team Bibliome developed a rule-based system based on the co-occurrence of entities with a trigger word in the same sentence. Only the Bibliome team performed coreference resolution. UTurku’s system was based on sentence level processing, whereas JAIST’s system was based on paragraph level processing. Therefore, Uturku’s system was most affected from not performing coreference resolution [84, 85].

Sub-task 2 of the Bacteria Biotope (BB) Task in the BioNLP Shared Task 2013, which gave another opportunity to scientists to address the task of extracting information about bacteria and habitats, focused on the aim to extract the Localization and Part Of relations. For Sub-task 2 the LIPN system [91] used a k-NN based approach by building language models for each example relation. The best F-score (42%) for Sub-task 2 was obtained by the TEES 2.1 system [124], which used multi-step Support Vector Machine classification. TEES 2.1 obtained the best F-score of 14% and a relaxed score of 49% in Sub-task 3 as well. TEES 2.1 is a generalized tool for relation extraction that was implemented to apply to many tasks in the BioNLP Shared Task. It did not tackle Sub-task 1 of the BB task that aimed at identifying the habitat entities and assigning them to the corresponding OntoBiotope ontology concepts. The IRISA system used a machine learning approach based on the k-Nearest Neighbor (kNN) method and ranked second with an F-score of 40% in Sub-task 2 [92]. LIMSI [93] was the only team that participated in all three BB sub-tasks, but their results for Sub-task 2 and Sub-task 3 were relatively lower compared to Sub-task 1 results. We also participated in Sub-task 2 of the BB Task 2013. Our system *Boun* ranked third in Sub-task 2 with an F-score of 27% in the official evaluation [16]. The Sub-task 2 system submitted to the official evaluation was based on a paragraph-based relation extraction approach, where the habitat entities were assumed to be related to the bacteria entity that occur first in the paragraph. After the shared task we developed a novel method for Sub-task 2, which operates on a sentence basis. In order to handle relations that span multiple sentences a new anaphora resolution approach for the bacteria biotopes domain has been developed as well. These improvements led to state-of-the-art results in Sub-task 2. The extended system *Boun 2* obtained 53% F-score on Sub-task 2. The details of our official submission as well as the improvements developed after the shared task are described in the following sections.

Sub-task 2 of the BB Task is related to the general problem of relation extraction. A number of different methods including entity co-occurrence based approaches [125, 126] and pattern matching based approaches [127–129] have been developed for extracting relations among biomedical entities including genes, proteins, drugs, and diseases. The state-of-the-art techniques for biomedical relation extrac-

tion are in general based on using the syntactic analyses of the sentences, usually in conjunction with supervised machine learning methods [123, 130–133]. Most relation extraction systems operate on a sentence-level. The underlying assumption is that the majority of the relations are contained within a single sentence. This assumption holds for some domains. For example, it has been shown that only 5% of the relations in the Genia event corpus [83] span multiple sentences [134]. However, a challenge in the Bacteria Biotopes domain is the vast amount of relations that span multiple sentences and the abundance of bacteria anaphora in the text. Despite this fact, only one of the systems that participated in the BB Shared task 2011 tackled the anaphora resolution problem in this domain [86], and none of the systems in the BB Task 2013 included anaphora resolutions modules [77].

We will describe the systems that we developed for extracting bacteria localization and habitat PartOf relations in the following subsections.

#### 5.1.5.2. Methods. Localization relation extraction

One of the two types of relations that have to be extracted for Sub-task 2 is the localization relations between bacteria and habitat entities. For example, the following excerpt from an input text file “*Bordetella. This group of organisms is capable of invading the respiratory tract of animals and causing severe diseases.*”, contains information about “*Bordetella*” bacteria that lives in the “*respiratory tract of animals*”. Therefore, there are localization relations between the “*Bordetella*” bacteria entity and the “*respiratory tract of animals*” and “*animals*” habitat entities, which must be extracted automatically.

In order to extract localization relations between bacteria and habitat entities, we propose two different systems. The paragraph-based system is the official system which was submitted to BioNLP Shared Task 2013 [16]. The sentence-based system with the anaphora resolution module was developed after the official evaluation. In the following subsections, each system is explained in detail.

*Paragraph-based system:* This system is based on the assumption that the bacteria name that occurs first in a paragraph is the topic of that paragraph. Therefore, after identifying the bacteria and habitat entities in a paragraph, the bacterium that appears first in the paragraph is associated with all habitat entities in that paragraph. If this bacterium entity occurs earlier in the document as well, then its first occurrence in the document is associated with the habitat entities in the paragraph. A special rule is applied to bacteria names that contain the term “*strain*”. In this case, the habitat entities are associated with the first occurrence of the corresponding bacterium name that does not contain the “*strain*” term. For example, in a paragraph that starts with the sentence “*Bordetella petrii strain DSM12804 was initially isolated from river sediment*”, a relation is set between the habitat entity “*river sediment*” and the bacterium entity “*Bordetella petrii DSM12804*” that occurs earlier in the document, instead of “*Bordetella petrii strain DSM12804*”, which is the first bacterium name in the given paragraph.

*Sentence-based system:* The workflow of the sentence-based system is shown in Figure 5.1. This system operates on a sentence-basis and performs a more fine-grained analysis of the text compared to the paragraph-based system. First, the text is segmented into sentences. Then, the bacteria and habitat entities that occur in the given sentence are identified. The assumption is that there is a relation between bacteria and habitat entities that occur in the same sentence, if there is a specific bacteria name in the considered sentence. For example, “*Bordetella petrii DSM12804*” is a specific bacteria name, whereas the terms “*bacteria*” and “*bacterium*” are not specific bacteria names, even though they are tagged as bacteria entities in the text documents.

*Anaphora resolution:* One of the challenges for extracting bacteria localization relations is that the corpus contains a large number of anaphora. In general, each document in the corpus is about a specific bacterium species [77]. After an explicit mention of the name of this species in a sentence, it is often referred to by using anaphors in the subsequent sentences. Therefore, several localization relations span multiple sentences. To tackle this problem we developed an anaphora resolution module and integrated it with the sentence-based localization relation extraction system. The anaphora resolu-

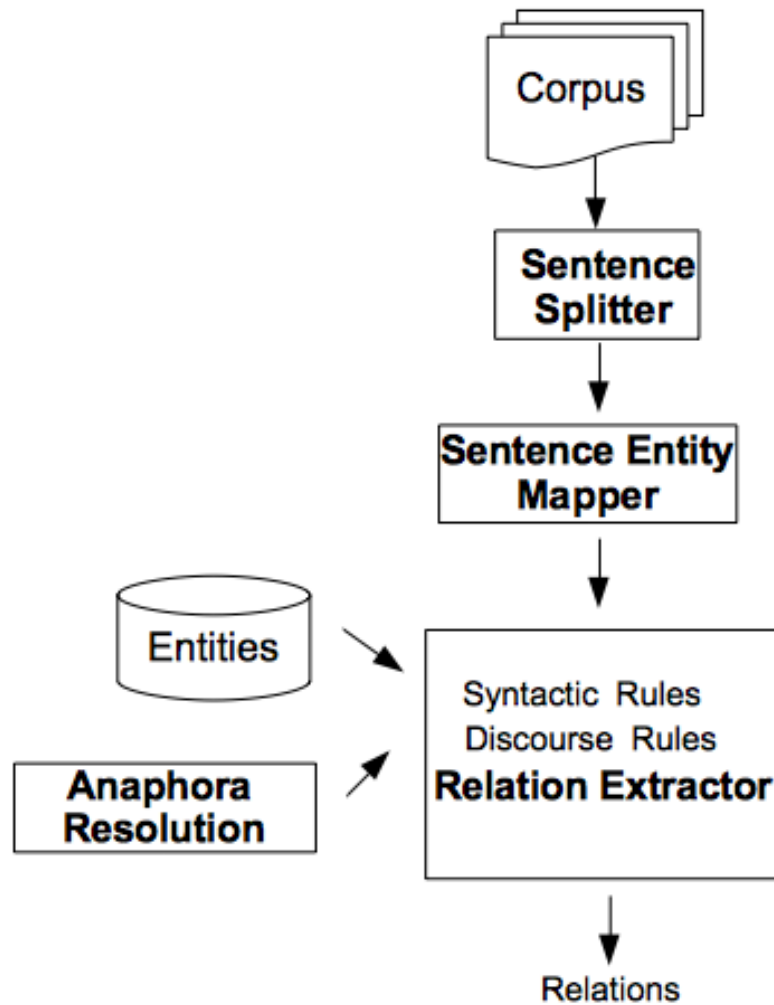


Figure 5.1. Workflow of the Sentence-based Sub-task 2 System

tion module detects sentences that do not include any bacteria entities, but contain coreferences to bacteria entities. There are three types of anaphoric expressions which are handled in different ways by our system:

*Anaphora type 1:* We compiled a keyword list consisting of 23 anaphoric expressions such as “the bacterium”, “this organism”, “this species”, “this genus”, and “this group of organisms” by manually analyzing the training set. If a sentence does not contain a bacteria name, but contains an anaphoric expression included in the keyword list, the antecedent of the anaphor is set as the first bacteria name that occurs in the previous sentence. Then, localization relations are identified between the habitats in the sentence and the detected antecedent of the anaphoric expression. For example, although the sentence “This bacterium is highly infectious, and can be spread

*through the contact with the infected animal products or through the air.*” does not include any explicit bacteria entity names, it describes localization relations between the bacteria anaphor “*This bacterium*” and the habitats “*animal products*” and “*air*”. In this case, the anaphora resolution module looks at the previous sentence, which is “*Brucella canis.*” and assigns the habitat entities to this bacteria entity. If there is no bacteria name in the previous sentence, then the first bacteria entity in the document is assigned to the habitat entities, since in general each document is about a specific bacterium species, and the mention of this species occurs first in the document.

*Anaphora type 2:* If there is no specific bacteria name in the given sentence, but the sentence begins with the anaphoric pronoun “*it*”, then our system looks at the previous sentence and a localization relation is set with the first bacteria in the previous sentence and the habitats in the given sentence. For example, given the sentence “*It was isolated from Ixodes scapularis in 1982.*”, our system looks at the previous sentence “*Borrelia burgdorferi.*” and sets a localization relation between the “*Borrelia burgdorferi*” bacteria entity and the “*Ixodes scapularis*” habitat entity.

*Anaphora type 3:* If a sentence begins with the “*This strain*” anaphoric expression, then similarly to the paragraph-based system, the bacteria entity that occurs first in the document is assigned as the antecedent of the anaphor. Consequently, the habitat entities in the sentence are assigned to this antecedent.

## Part-Of Relation Extraction

PartOf relations between habitat entities is the second relation type targeted in the BB Shared Task. For example, in the sentence “*This strain was isolated from infant feces*”, the habitat entity “*infant feces*” is a part of the habitat “*infant*”. For habitat PartOf relation extraction we introduce a shallow syntactic analysis dependent rule-based approach. The first rule with the preposition “*of*” was developed for the official shared task submission. The remaining rules were developed after the shared task. Our rules are based on the assumption that a habitat is likely to be a part of another habitat, if the mention of the second habitat in text contains the mention of



the first habitat, and in addition the syntactic rules described below are met.

*Syntax rule 1:* If one habitat contains the other one, and the second habitat follows one of the prepositions “of”, “in”, “from”, then the relation that the first habitat is PartOf the second habitat is extracted. For example, the habitat mention “*rhizosphere of plants*” contains the “*plants*” habitat mention. Since the first habitat phrase contains the preposition “of”, and the second habitat phrase “*plants*” occurs right after this preposition, the relation “*rhizosphere of plants*” is PartOf “*plants*” is extracted. As another example, the habitat mention “*oral cavity in humans*” contains the “*humans*” habitat mention. Since the first habitat mention contains the preposition “in”, and the second habitat mention “*humans*” follows this preposition, the relation “*oral cavity in humans*” is PartOf “*humans*” is extracted. Finally, “*skin lesion from a Lyme disease patient in Europe*” and “*Lyme disease patient in Europe*” are overlapping habitat entities, one of which contains “from”, which is succeeded by the second habitat mention. Then, the relation “*skin lesion from a Lyme disease patient in Europe*” is PartOf “*Lyme disease patient in Europe*” is extracted.

*Syntax rule 2:* If two habitat mentions overlap in text like in the example “*Aeschynomene stem nodule*” and “*Aeschynomene*”, by looking at their positions we infer a PartOf relation between them. For example, “*Aeschynomene stem nodule*” is PartOf “*Aeschynomene*”.

5.1.5.3. Results. The evaluation metrics used for Sub-task 2 are precision, recall, and f-score. The details of the evaluation metrics and the official evaluation results are available in [77]. In the following subsections, the results of the system (*Boun*) with which we participated in the BB shared task and the results of the improved system (*Boun 2*) developed after the official evaluation are presented.

This section provides the evaluation results obtained by the paragraph-based system (*Boun*) with which we participated in the BB Shared Task Sub-task 2 (Localization and PartOf Event Extraction), as well as the newly developed sentence-based system

Table 5.1. Results of BB Sub-task 2 (*Localization and PartOf Event Extraction*). The results obtained on the test set are reported.

System	Type	Recall	Precision	F-score
<b>Boun 2</b>	Localization	0.61	0.54	0.57
	PartOf	0.20	0.32	0.25
<b>Boun</b>	Localization	0.23	0.38	0.29
	PartOf	0.15	0.40	0.22

with the anaphora resolution module (*Boun 2*). Table 5.1 presents a comparison of the *Boun* and *Boun 2* systems with each other. The results demonstrate that the *Boun 2* system performs significantly better than the *Boun* system.

Table 5.2. Comparison with the other systems that participated in the BB Sub-task 2 (*Localization and PartOf Event Extraction*). The results obtained on the test set are reported.

System	Recall	Precision	F-score
<b>Boun 2</b>	0.53	0.52	0.53
<b>TEES 2.1</b>	0.28	0.82	0.42
<b>IRISA</b>	0.36	0.46	0.40
<b>Boun</b>	0.21	0.38	0.27
<b>LIMSI</b>	0.04	0.19	0.06

Table 5.2 presents a comparison of the *Boun* and *Boun 2* systems with the other systems that participated in the shared task. According to the official results, the *Boun* system ranked third among the four systems that participated in the event detection task. The new *Boun 2* system achieves 53% F-score on the test set, which is significantly higher than the 27% F-score obtained by the *Boun* system. The F-score of the *Boun 2* system is even higher than the F-score of the system that ranked first in the official evaluation.

Tables 5.3, 5.4, and 5.5 show the effects of the anaphora resolution module for localization extraction and the syntax rules for PartOf relation extraction on the

Table 5.3. Effects of Anaphora Resolution Module and Syntax Rules (*Localization and PartOf Event Extraction*). The results obtained on the training set are reported.

System	Recall	Precision	F-score
<b>Boun 2</b>	0.46	0.42	0.44
- Anaphora	0.36	0.45	0.40
- Syntax rule 1	0.45	0.42	0.43
- Syntax rule 2	0.46	0.42	0.44

Table 5.4. Effects of Anaphora Resolution Module and Syntax Rules (*Localization and PartOf Event Extraction*). The results obtained on the development set are reported.

System	Recall	Precision	F-score
<b>Boun 2</b>	0.55	0.40	0.46
- Anaphora	0.50	0.44	0.47
- Syntax rule 1	0.54	0.40	0.46
- Syntax rule 2	0.53	0.42	0.47

Table 5.5. Effects of Anaphora Resolution Module and Syntax Rules (*Localization and PartOf Event Extraction*). The results obtained on the test set are reported.

System	Recall	Precision	F-score
<b>Boun 2</b>	0.53	0.52	0.53
- Anaphora	0.46	0.56	0.50
- Syntax rule 1	0.52	0.52	0.52
- Syntax rule 2	0.50	0.55	0.52

training, development, and test sets, respectively. The first rows of these tables show the results obtained by the *Boun 2* system. The second row shows the results obtained by removing the anaphora resolution module from the system, and the third and fourth rows show the results obtained by removing the first and second syntax rules from the system, respectively. The anaphora resolution module achieves a considerable increase in recall on all data sets (training, development, and test), which leads to improved F-score performances on the training and test sets. The two syntax rules have similar

effects. In general, they lead to an increase in recall, which can improve F-score if the drop in precision is relatively less (e.g. on the training and test sets).

These results demonstrate that performing a more fine-grained analysis of the text at the sentence level and incorporating an anaphora resolution module to handle relations that span multiple sentence is an effective approach for extracting relations in the bacteria biotopes domain. Our improved system achieves state-of-the-art results. However, there is still a lot of room for improvement. Our current approach assumes that if a specific bacteria (or its coreference) occur in the same sentence with a habitat entity, there is a localization relation between them. Deeper syntactic and semantic analysis of the sentences by using full or dependency parsing strategies can enhance the accuracy of the system. The PartOf relation extraction method that we proposed is only able to identify PartOf relations between habitat entities that overlap (e.g. *“human gastrointestinal tract”* and *“human”*). A deeper syntactic analysis can enable identifying long-distance relations between habitat entities (e.g. the PartOf relation between *“human”* and *“gut”* in the sentence *“This organism is found in humans as a normal component of gut flora.”*). Furthermore, the lower accuracy of the PartOf relations may also be caused by the fact that our system does not take into account whether the candidate habitat entities are hosts or host parts. For example, the habitat entity *“fresh water”* is neither a host nor a host-part. Therefore, it should not be considered for a PartOf relation. Including a module that can pre-identify the habitats which can act as hosts or host-parts in advance, may improve the performance of the system for PartOf relation extraction.

## 5.2. An application for Brucella-Host Relevant Interaction Extraction

### 5.2.1. Motivation

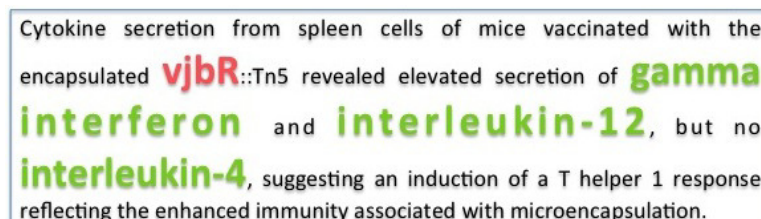
Brucella is a Gram-negative intracellular bacterium that causes zoonotic brucellosis in humans and various animals. Brucellosis is one of the most common zoonotic diseases worldwide, causing approximately half a million new human brucellosis each year. There are 10 species of Brucella based on the preferential host specificity: Bru-

*Brucella melitensis* (goats), *B. abortus* (cattle), *B. suis* (swine), *B. canis* (dogs), *B. ovis* (sheep), *B. neotomae* (desert mice), *B. cetaceae* (cetacean), *B. pinnipediae* (seal), *B. microti* (voles), and *B. inopinata* (unknown) [135]. Among them, *B. melitensis*, *B. abortus*, *B. suis*, and *B. canis* are pathogenic to human. The other *Brucella* species are non-pathogenic to humans.

The genome sequences of all *Brucella* species are strikingly similar with nearly identical genetic content and gene organization [136]. Humans can be infected with *Brucella* by contact with infected animals, by inhalation of an aerosol, or by ingestion of contaminated animal products (e.g., infected milk and meat). Upon entry into animals, the bacteria invade the blood stream and lymphatics where they multiply inside phagocytic cells and eventually cause septicemia. Symptoms include undulant fever, abortion, asthenia, endocarditis and encephalitis. In spite of a long documented history (Corbel, 1997), the treatment of human brucellosis remains difficult and requires antibiotics that penetrate macrophages and can act in an acidic intracellular environment. While currently used live attenuated *Brucella* animal vaccines (e.g., RB51, strain 19, and Rev. 1) have the ability to protect animals, they are still pathogenic to humans. No safe and effective *Brucella* vaccine is available for human use. To develop safe and effective preventive and therapeutic measures against *Brucella* infections, it is critical to understand the host-*Brucella* mechanisms that lead to *Brucella* pathogenesis and host immunity against *Brucella* infection. Although extensive studies have been undertaken, the systematic understanding of the host-*Brucella* interactions is still missing.

Currently, there is very limited information regarding host-*Brucella* interactions in the host-pathogen interaction databases such as PHIDIAS [137], PHISTO [138], and HPIDB [73]. Most of the relevant information is only available in a textual format in the published scientific articles. In this study, our goal is to utilize text mining methods to extract host-*Brucella* gene interactions from the biomedical literature. In order to extract host-pathogen gene interactions, first the pathogen and host gene names should be identified in text, then the interactions among the host and pathogen genes should be detected. For example, the sentence shown in Figure 5.2 [139] contains three

host genes (gamma interferon, interleukin-12, and interleukin-4) and one pathogen gene (vjbR). This sentence states that there are two pathogen–host gene interactions: (gamma interferon, vjbR) and (interleukin-12, vjbR). On the other hand, there is no an interaction between the host gene interleukin-4 and pathogen gene vjbR.



Cytokine secretion from spleen cells of mice vaccinated with the encapsulated **vjbR**::Tn5 revealed elevated secretion of **gamma interferon** and **interleukin-12**, but no **interleukin-4**, suggesting an induction of a T helper 1 response reflecting the enhanced immunity associated with microencapsulation.

Figure 5.2. Sample host-pathogen interaction describing sentence taken from [139].

The pathogen gene is shown in red and the host genes are shown in green.

Different methods have been proposed for literature mining of gene–gene interactions. One of the simplest and widely used methods is based on the co-occurrence statistics of the proteins in text [125]. Another common approach is matching pre-specified patterns and rules over the sequences of words and/or their parts of speech in the sentences [127, 140]. More recently, machine learning methods that integrate the linguistic, syntactic, and/or semantic analysis of the sentences as kernel functions have been proposed and shown to achieve state-of-the-art results for gene/protein interaction extraction from text [141–144]. Similarly to previous literature mining studies, in this study we used the commonly applied GENETAG-style named entity annotation [145]. In other words, a gene interaction can involve genes or gene products such as proteins.

A number of rule-based and machine learning based methods have been proposed for identifying gene/protein mentions in text [22, 146, 147] [148]. In our previous studies, we developed dictionary- and rule-based named entity recognition tools, SciMiner [149] and Vaccine Ontology (VO)-SciMiner [150], which are designed to identify genes/proteins and Vaccine Ontology (VO) terms in the biomedical literature. Conventional Medical Subject Headings (MeSH) terminology has been frequently used for literature mining, such as GenoMesh studies [151]. The usage of ontologies enhances the chances of retrieving gene–gene interactions. For example, in our recent studies we have shown that the VO facilitates the retrieval of vaccine-associated IFN-gamma interaction network [152], fever-related network [153], and Brucella vaccine

interaction network [153]. Recently, we have developed an Interaction Network Ontology (INO) which is used to classify the interaction keywords such as up-regulation, inhibition, association, and binding in an ontology structure [154]. The classified interaction hierarchy makes us not only retrieve gene–gene interactions, but also the types of gene–gene interactions [154]. We hypothesize that such a strategy can also be used in host–pathogen gene–gene interaction literature retrieval.

Currently, the research in host–pathogen interactions literature mining mostly focuses on the retrieval of host gene–gene interaction under a particular pathogen infection (e.g., influenza) or pathogen gene–gene interactions [e.g., our *Brucella* vaccine interaction network analysis [153]]. There are only a few studies on the retrieval of both host and pathogen genes and the inter-species interactions among them [reviewed in [155]]. Machine learning based methods were proposed for classifying abstracts of scientific articles as being relevant to host–pathogen interactions or not [156, 157]. In addition, Thieu et al. (2012) proposed a rule-based approach that is based on the link-grammar representations of the sentences for extracting host–pathogen protein interactions from text.

In this study, we use kernel-based methods for extracting host–pathogen gene interactions, which have been shown to achieve promising results for extracting intra-species protein interactions [132] [158]. One main issue in host–pathogen interaction literature mining is the confusion of a gene being a host gene or pathogen gene, since many gene names are shared in both hosts and pathogens. This is one main research topic in our current study. We extended the SciMiner mammalian gene name identification tool to recognize and distinguish between host and *Brucella* genes. In addition, we used an INO-based method to model various gene–gene interactions under different experimental conditions. Our results show that our combinatory strategy is able to successfully retrieve and analyze host–pathogen gene–gene interaction networks.

### 5.2.2. Methods and Materials

The main focus of this study is to identify the interactions between host and *Brucella* genes. Many eukaryotic organisms act as the host of *Brucella* infections, including human, cattle, goat, sheep, pig, etc. As a laboratory animal model, mice can also be infected with *Brucella*. Our literature mining study covers these different host species. Meanwhile, there are 10 different *Brucella* species.

The overall design and workflow of our approach is shown in Figure 5.3. All PubMed papers are used as our data sources. They are filtered based on their relevance to *Brucella*. The selected abstracts are processed by splitting into sentences and identifying the host and *Brucella* gene name mentions using SciMiner. Next, co-occurrence and machine learning based methods are used to extract the interactions among the host and *Brucella* genes. A literature-mined and manually verified host-*Brucella* gene-gene interaction network is created. Finally, ontology based modeling of host-pathogen gene-gene interactions is performed by utilizing the Interaction Network Ontology (INO). The details of the methods are presented in the following subsections.

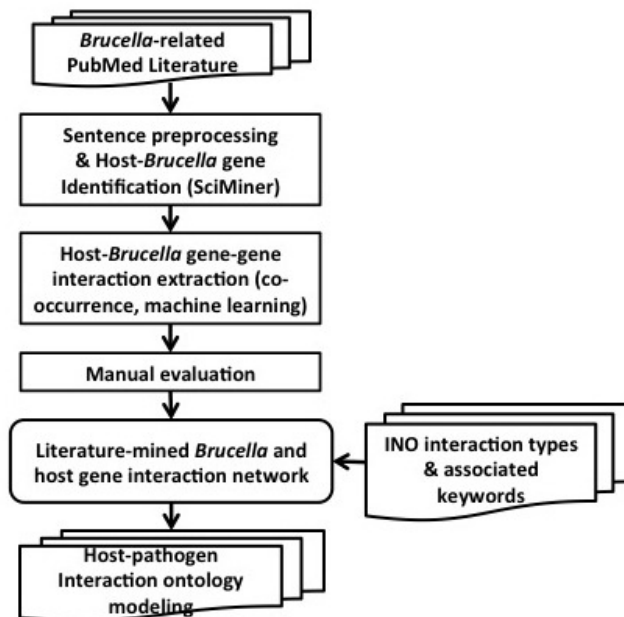


Figure 5.3. Workflow of the host-*Brucella* interaction extraction approach.



5.2.2.1. Data set collection. The 2015 MEDLINE/PubMed Baseline Distribution database consisting of 23,343,329 records was downloaded from the US National Library of Medicine and processed using our established literature mining pipeline. Briefly, the title, abstract, and MeSH term of each record were parsed out from the downloaded XML files. The collected abstracts were split into sentence level using LBJ2.nlp.SentenceSplitter Java module. Then, enhanced version of our named entity recognition tools, SciMiner [149] and VO-SciMiner [150], were used to identify host genes and pathogen genes, and the results were populated into a local MySQL database. To define the Brucella-specific context, we used a PubMed query, Brucella OR Brucellosis, which resulted in a list of 16,699 PubMed IDs as of 2/1/2015.

5.2.2.2. Identifying gene names. To identify the mentioned host genes and Brucella genes in the abstracts of articles, we used SciMiner [149] and VO-SciMiner [150]. SciMiner and VO-SciMiner are both dictionary- and rule-based literature mining tools. SciMiner focuses on identification of mammalian genes, reported in terms of the official human genes based on the HUGO Gene Nomenclature Committee (HGNC) database [159], while VO-SciMiner identifies vaccine ontology (VO) terms and Brucella genes.

To improve the identification accuracy of host and pathogen genes, we enhanced the mining rules in both SciMiner and VO-SciMiner. First, the enhanced version of SciMiner uses a stringent case match of gene symbols. In the original version of SciMiner, which included dictionary of only human genes names and symbols, a relaxed matching of symbol was employed to maximize the gene identification (high recall). This relaxed case matching resulted in misidentifications such as recA, recombinase A gene, being identified as the human RAD51 recombinase (RAD51), whose aliases include RECA. Since the majority of the Brucella gene symbols start with a lower-case character and usually end with an upper-case or numeric character, SciMiner excluded symbols with this pattern. In case of the genes identified by both SciMiner as a host gene and VO-SciMiner as a pathogen gene, the priority is given to the VO-SciMiner identification considering the current context of Brucella-related literature.

5.2.2.3. Mapping genes to pathogen and host species. In order to further improve the overall accuracy of host gene identification, we used potential host species-related Medical Subject Headings (MeSH) terms, including *humans, rats, mice, cattle, guinea pigs, swine, goats, and sheep* to filter the genes identified by SciMiner. Only the host genes identified from PubMed documents whose MeSH terms included at least one of these selected terms were selected.

5.2.2.4. Gene-gene interaction extraction. In this study, co-occurrence based and supervised machine-learning based approaches are used for extracting host-pathogen gene-gene interactions. Both sentence-level and abstract-level co-occurrence approaches, as well as a machine learning-based approach are investigated for this task. These approaches are described in the following subsections.

#### Co-occurrence based host-pathogen interaction extraction

We used two different contexts to extract the interactions based on the co-occurrences of the host and pathogen genes: sentence-based context and abstract-based context. In the sentence-based co-occurrence approach, if one pathogen and one host gene occur in the same sentence, an interaction pair is extracted consisting of the corresponding pathogen and host genes. For example, in the sentence shown in Figure 5.2 [139], the SciMiner tool identifies two host genes (*interleukin-12* and *interleukin-4*) and one pathogen gene (*vjbR*). The sentence-level co-occurrence approach extracts the interactions (*interleukin-12, vjbR*) and (*interleukin-4, vjbR*) from the sample sentence, where (*interleukin-12, vjbR*) is a true interaction and (*interleukin-4, vjbR*) is an incorrectly extracted interaction. In the sample sentence, gamma interferon is also a host gene. However, since this gene is not detected by SciMiner, it is not considered in the interaction extraction step. In the abstract-based co-occurrence approach, an abstract is taken into consideration as the context window instead of a single sentence. In other words, all pairs of host and pathogen genes that occur in the same abstract are extracted as interacting pairs regardless of the sentence boundaries.

## Machine learning based host-pathogen interaction extraction

We utilized a machine learning based approach to classify whether a host and pathogen gene pair occurring in the same sentence is described as interacting in the sentence or not. We used support vector machines (SVM) (in particular the SVMlight package [160]) as our classification algorithm with the cosine and edit kernels introduced in [142]. These kernels make use of the dependency parse trees of the sentences that represent the syntactic and semantic relations among the words. We used the Stanford Parser [161] to obtain the dependency parse trees of the sentences in our *Brucella* specific data set. We only processed sentences for which SciMiner identified at least one host and one pathogen gene. The cosine and edit kernels are defined over the path between the host gene and pathogen gene in the dependency parse tree of the corresponding sentence.

The underlying assumption is that the dependency path between two entities is a good description for the semantic relation between them. For example, the dependency parse tree for the sample sentence “*Furthermore, gap associated with murine IL-12 gene in a DNA vaccine formulation partially protected mice against experimental infection.*”, taken from (Rosinha et al., 2002), is shown in Figure 5.4. The dependency path between the host gene IL-12 and the pathogen gene gap, which are described as interacting in the sample sentence, is “*nn gene prep-with associated vmod*”. On this path we have the word associated as well as the dependency relation type preposition with (prep-with), which provide clues for the interaction between gap and IL-12. The cosine kernel is computed by taking the cosine similarity between two dependency paths, whereas the edit kernel is computed by taking the edit distance between them and then converting the distance measure to a similarity measure [142]. If two paths are similar, they are likely to belong to the same class (the interaction class or the non-interaction class).

To the best of our knowledge, there are no publicly available manually labeled host-pathogen gene-gene interaction corpora. Therefore, we trained the SVM classifier with edit and cosine kernels by using corpora labeled for intra-species protein-protein interactions. Specifically, we used the Christina Brun (CB) corpus provided as resource

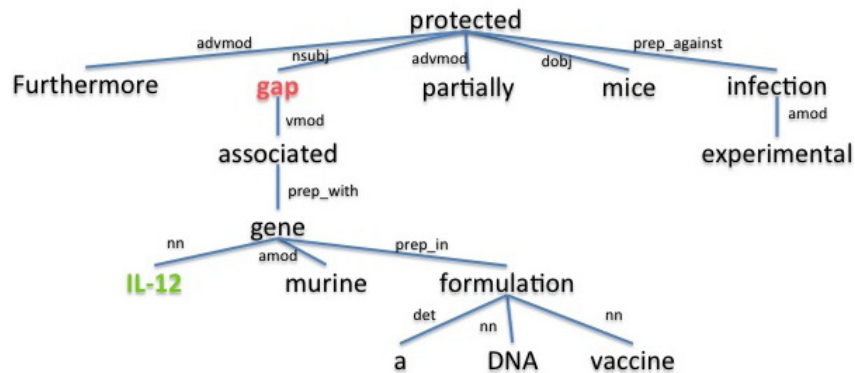


Figure 5.4. The dependency parse tree of a sample sentence. The tree generated for the sentence Furthermore, gap associated with murine IL-12 gene in a DNA vaccine formulation partially protected mice against experimental infection. from the abstract of [162]. Host and pathogen genes identified by SciMiner are shown in green and red, respectively.

at the BioCreative II challenge [163] and the AIMED corpus [164], which is a standard corpus for evaluating intra-species protein-protein interactions. The learned cosine and edit kernel based SVM models are used to classify each sentence as an interaction-describing sentence or not for each host and pathogen gene pair identified by SciMiner in the corresponding sentence.

**5.2.2.5. Ontology modeling.** The Interaction Network Ontology (INO) focuses on the ontological representations of hierarchical biological interaction types and networks [154]. INO has been proven to enhance the literature mining of gene-gene interaction types [154]. In this study, we applied INO to analyze different interaction types between host and *Brucella* at different experimental conditions. Furthermore, different conditions of host-*Brucella* interactions were represented and analyzed through ontology-based modeling.

### 5.2.3. Results and discussion

#### 5.2.3.1. Results. Identification of host and Brucella gene names

SciMiner and VO-SciMiner were enhanced to identify host and pathogen genes, respectively. First, SciMiner has been modified to use stringent case match. In the context of Brucella, consisting of 16,699 PubMed abstracts, the enhanced versions of SciMiner and VO-SciMiner identified 47 unique pairs of potential host gene and Brucella gene interactions using the improved symbol-based identification method and confliction resolution between host and Brucella gene. Out of these 47 pairs, manual examination confirmed that 24 unique pairs were true interactions, indicating an overall accuracy of 51%.

#### Identification of host-Brucella gene-gene interactions

After identifying the host and Brucella gene names in sentences co-occurrence and machine learning based methods are used to classify each pair in a sentence as an interaction (positive class) or not (negative class). We performed manual evaluation for the classification decisions of the methods for each host-Brucella gene pair in each sentence. For the abstract-level co-occurrence approach, manual evaluation is performed for each host-Brucella gene pair in each abstract.

Table 5.6. Co-occurrence and machine learning-based host-Brucella gene-gene interaction results. TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

	TP	TN	FP	FN	Precision	Recall	F score
<b>Co-occurrence (Sentence-based)</b>	29	0	25	0	0.54	1.0	0.70
<b>Co-occurrence (Abstract-based)</b>	55	0	61	0	0.47	1.0	0.64
<b>SVM (edit kernel)</b>	15	12	12	14	0.56	0.52	0.54
<b>SVM (cosine kernel)</b>	12	19	5	17	0.71	0.41	0.52

The results obtained are summarized in Table 5.6. TP (True Positives) is the number of host-pathogen interactions correctly classified as positive; FP (False Positives) is the number of negative host-pathogen interactions that are incorrectly classi-

fied as positive by the classifier; TN (True Negatives) is the number of host-pathogen interactions classified correctly as negative (no interaction); and FN (False Negatives) is the number of positive host-pathogen interactions that are incorrectly classified as negative by the classifier.

Precision, recall, and F-score are used as our metrics to evaluate the performances of the utilized methods. Precision is the ratio of correctly identified positive host-pathogen interactions over all interactions classified as positive by the classifier (i.e.,  $TP/(TP+FP)$ ). Recall is the ratio of correctly classified positive host-pathogen interactions over all positive host-pathogen interactions (i.e.,  $TP/(TP+FN)$ ). F-score is the harmonic mean of these two measures.

Co-occurrence based methods classify all pairs of host-pathogen genes as positive, if they occur in the same sentence or abstract. Therefore, they obtain the maximum level of recall, i.e., 100%. Not all co-occurring gene pairs are true interaction pairs. For example, in the sample sentence shown in Figure 5.2, there is no an interaction between the pathogen gene vjbR and the host gene interleukin-4. However, the co-occurrence methods incorrectly classify this pair as interacting, since they occur in the same sentence. This leads to drop in precision.

SVM with edit and cosine kernel obtain higher precision compared to the co-occurrence based approach. The precision obtained by the cosine kernel (71%) is significantly higher than the precision values of the co-occurrence and edit kernel approaches. Edit kernel, on the other hand, obtains more balanced precision and recall levels compared to the other methods.

Both edit kernel and cosine kernel operate on sentence-level. Therefore, they are not able to identify interactions whose descriptions cross sentence boundaries. The significantly higher number of true positive interactions retrieved by the abstract-level co-occurrence approach indicates the importance of the use of abstracts (or scopes wider than sentences) as context.

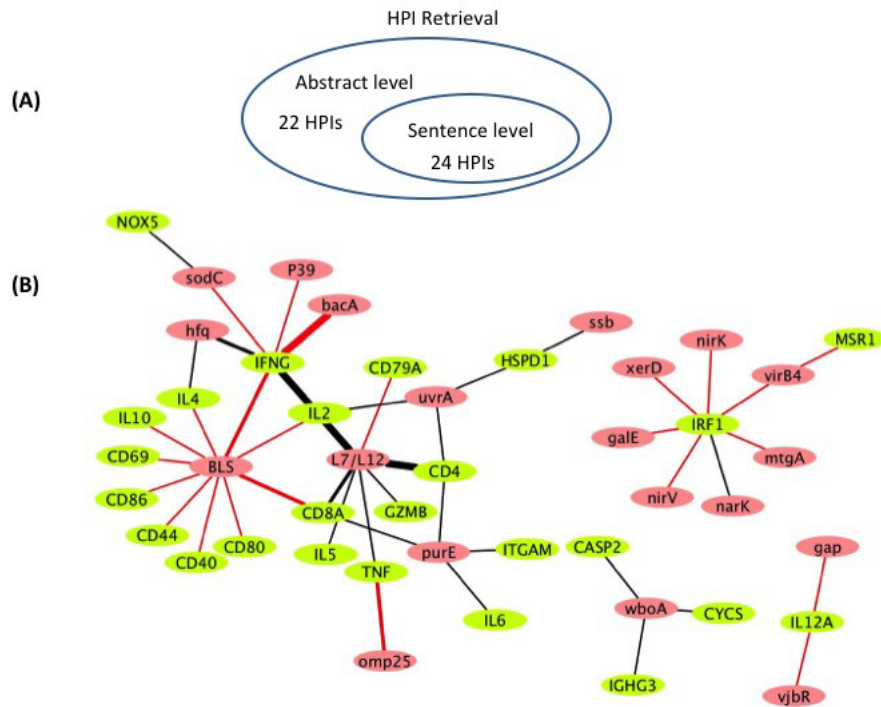


Figure 5.5. Literature-mined host-Brucella gene-gene interaction results. (A) Venn diagram (B) The literature-mined and manually verified host-Brucella gene-gene interaction network.

Figure 5.5 shows the literature mined and manually verified unique host-Brucella gene-gene interactions. In Figure 5.5 (A), Venn diagram shows the number of unique host-Brucella interaction gene pairs retrieved and manually verified from sentence-level and abstract-level processing. In Figure 5.5 (B), The literature-mined and manually verified host-Brucella gene-gene interaction network is shown. Host genes are shown in green and Brucella genes are shown in red. Red edges correspond to interactions retrieved from sentence-level processing. Black edges correspond to interactions retrieved from abstract level processing. The more sentences/abstracts describe an interaction between a gene pairs the thicker the edge connecting them. A total of 46 unique interaction pairs are retrieved. 24 of these were identified using sentence-level processing. Abstract-level analysis enabled the retrieval of 22 additional unique interaction pairs (Figure 5.5 A). The identified host-Brucella gene-gene interactions are represented as a network, which consists of 20 Brucella genes and 25 host genes (Figure 5.5 B). The interactions between host and Brucella gene pairs are represented as edges. The edges are weighed based on the number of sentences/abstracts that state the corresponding

interaction. BLS and L7/L12 are the most connected Brucella genes, whereas IFNG and IRF1 are the most connected host genes.

### Ontology Modeling of Host-Brucella Gene-gene Interactions

We used INO to analyze the types of interactions between the extracted host and Brucella genes. The results of this analysis are shown in Figure 5.6. In total, six different INO interaction types, all of which are sub-types of regulation, are identified from this literature mining study. The ‘induction of production’ type is the most common type identified. For instance, the sentence “The P39 and the bacterioferrin (BFR) antigens of *B. melitensis* 16M were previously identified as T dominant antigens able to induce both delayed-type hypersensitivity in sensitized guinea pigs and in vitro gamma interferon (IFN-gamma) production by peripheral blood mononuclear cells from infected cattle” [165] is an example sentence that describes an interaction of type ‘induction of production’ between pathogen and host genes. The sentence states that Brucella gene P39 is able to induce in vitro host IFN-gamma production.

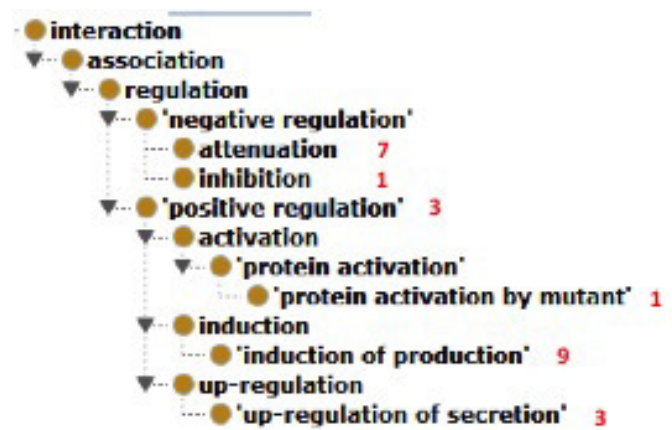


Figure 5.6. In total, six different INO interaction types were identified from this literature mining study. The number of interactions of a specific type is shown in red next to the interaction type. The ‘induction of production’ type is the most common type identified.

**5.2.3.2. Discussion.** Using Brucella as an example pathogen, this study utilized literature mining and ontology analysis approaches to examine the interactions between host genes/proteins and Brucella genes/proteins. Since genes encode for proteins, our



host-Brucella gene-gene interactions also include protein-protein interactions. Our approach identified 46 pairs of host-Brucella gene-gene interactions from the literature, and the ontology modeling analysis identified different types of interactions and provided deeper insights on how the host and Brucella genes/proteins interact at different experimental conditions.

One challenge in host-pathogen interaction literature mining is the difficulty in differentiating host genes and pathogen genes. In the current version of SciMiner and VO-SciMiner we did not use any of the name (longer description)-based identification results in the analysis. This is due to our manual evaluation of the preliminary results suggesting it is far more difficult to distinguish between host and pathogen genes using longer description protein names as they are more redundant than gene symbols. For example, the protein name “Superoxide dismutase [Cu-Zn]” may represent a human/host gene name (*SOD1* or *SODC*) or a Brucella/pathogen protein (*SodC*). In general, the gene names are more unique than the gene symbols; therefore, use of only short gene symbols resulted in decreased numbers of identified genes by the current versions of SciMiner and VO-SciMiner. We will examine these missed genes and further improve the sensitivity and accuracy of the gene name-based identification.

We investigated using co-occurrence and machine learning based methods for extracting host-pathogen gene-gene interactions. The co-occurrence based methods classify each pair of host and pathogen genes as interacting, if they occur in the same sentence/abstract. Therefore, they obtain high recall by retrieving all interacting pairs of genes. However, they also classify many gene pairs incorrectly as interacting, since not all co-occurring gene pairs are true interactions. This leads to drop in performance in terms of precision. The SVM classifiers with the dependency tree based edit and cosine kernels make use of the syntactic analysis of the sentences. These methods achieved higher precision compared to the co-occurrence based methods. To the best of our knowledge, there does not exist a large manually labeled host-pathogen gene-gene interaction data set. Therefore, the edit and cosine kernel based SVM classifiers were trained by using generic (intra-species) protein-protein interaction data sets. Training these classifiers with host-pathogen gene-gene interaction data might improve

their performances. A drawback of most (if not all) currently available machine learning based interaction extraction methods is that they operate on sentence-level and therefore, are not able to identify interactions that cross sentence boundaries. As our sentence-level and abstract-level co-occurrence analysis revealed, many host-*Brucella* interactions span multiple sentences. These results suggest that developing text mining methods that operate on scopes wider than a sentence would be useful for extracting host-pathogen gene-gene interactions.

Our ontology modeling studies demonstrate its value in further identifying the nature and insights of host-pathogen gene-gene interactions. A simple gene-gene interaction may miss many details, especially in the setting of a host-pathogen interaction. A gene-(interaction type)-gene would provide more details since the interaction type could indicate how the two genes interact. The INO provides a way to classify hundreds of interaction keywords into logically defined interaction types under a hierarchical ontology setting [154]. The usage of INO interaction types and its hierarchy allows us to detect the distribution of the interaction types from our literature mining study (see Figure 5.6). INO-based modeling also provides a novel way to identify interaction types that are represented by multiple keywords in sentences [166].

Compared to model pathogens such as *Escherichia coli* and *Salmonella*, *Brucella* is a less studied pathogen. However, the results obtained from this study provide the first example of opportunities and challenges in the literature mining of the host-pathogen gene-gene interactions.

## 6. CONCLUSIONS

### 6.1. Discussion

In this thesis, we proposed two different methods for the two problems: named entity recognition and named entity normalization. For the first one, a rule-based approach is proposed, while for the second method, a data-driven approach, which is based on word-embeddings is utilized. Both of the approaches are unsupervised and do not need labeled data, thus both can be applied to different named entities in the biomedical domain, where labeled data are scarce. In addition, for this thesis, two applications are implemented: the first one is a pipeline for the extraction of bacteria biotope information from Pubmed, and the other one is an application for the extraction of Brucella-host related data from Pubmed abstracts. The details about the contributions and the results are further discussed below.

For this thesis, a fully automatic text mining module for experimental method extraction is implemented and integrated to Pathogen-Host Interaction Search Tool (PHISTO), which serves as an up-to-date and functionally enhanced source of PHI data through a user-friendly interface. To implement this module, a dictionary of interaction detection methods is compiled from the PSI-MI ontology and the abstracts of the articles that contain PHIs without experimental method information are obtained from PubMed. An exact string matching-based approach was used to assign 2952 experimental method names to 2109 unique PHIs. The experimental method detection module was evaluated by using the PHIs with experimental method information in PHISTO. Although the module achieved a promising precision of 74%, the recall of the module is 34%, which is lower compared to precision. This study demonstrates the crucial need for new text-mining methodologies, which are more enhanced than exact matching, for the detection and categorization of biomedical entities.

Following the text-mining module that is implemented for PHISTO, Chapter 3.2 presents a novel unsupervised method, which makes use of shallow linguistic knowl-

edge and syntax rules, for the detection and categorization of biomedical named entities through an ontology. We introduce a linguistically-motivated rule-based approach for named entity recognition and normalization that targets identifying and normalizing habitat entities through an ontology. We participated at the Bacteria Biotope Task in the BioNLP Shared Task 2013 with our system (named as the Boun system) and obtained promising results in the official evaluation. With the developments after the shared task (named as the Boun 2 system), several extensions are proposed for named entity recognition and normalization of the bacteria habitat entities. Extending the candidate noun phrases by their modifiers resulted in lower performance, due to the prepositional phrase attachment ambiguity problem. Incorporating an ontology expansion module to our system (Boun) did not lead to improvement in the performance in terms of SER score. The Boun and Boun 2 systems achieved the same SER value (68%), which is close to the SER value of the system that ranked first in the shared task. Our results show that our approaches based on the shallow syntactic analysis of the text and linguistically-motivated hand-coded rules are as effective as supervised machine learning approaches for named entity detection and ontology-based normalization in the bacteria biotopes domain.

Although promising results are obtained in Chapter 3.2 for the detection and categorization of bacteria biotope entities, the need for the manually designed syntax-rules makes the method’s adaptation harder to other types of biomedical entities. In Chapter 4, we introduce an unsupervised data-driven approach for biomedical entity normalization through an ontology by utilizing word embeddings and syntactic re-ranking. The proposed approach is applied to the normalization problem of habitat entities through the Onto-Biotope ontology and the adverse drug reaction entities through the Med-DRA dictionary, and tested on the BioNLP Shared Task 2016 Bacteria Biotopes data set and the Text Analysis Conference 2017 Adverse Drug Reaction data set, respectively. The new system based on word embeddings and syntactic re-ranking obtained higher precision scores on both data sets than the system without syntactic re-ranking. Furthermore, the system achieved a precision score of 65.9% on the BioNLP Shared Task 2016 Bacteria Biotopes data set, which is 2.9 percentage points above the current state-of-the-art. Our proposed approach with syntactic re-ranking (named as the

BOUNEL system) uses the Stanford Parser, which is a supervised parser. However, BOUNEL is unsupervised in the sense that it does not require training data manually annotated with entity mentions and their corresponding concepts in the ontology. Furthermore, the Stanford Parser has not been re-trained using biomedical data, but the off-the-shelf parser pre-trained with the Penn Treebank has been used. Therefore, the proposed approach can be easily adapted for normalizing different types of biomedical entities. This thesis also shows that our approach based on syntactic based weighted semantic similarity is as effective as supervised and semi-supervised approaches for biomedical named entity normalization.

Chapter 5 presents two different applications, one of which is a pipeline for the ontology-based entity tagging/normalization and relation extraction of bacteria biotopes, and the other one is about extracting Brucella-host interactions and demonstrates the importance of context in the text mining problems, which may also be considered in the task of normalization as a future work to improve the results.

The first application, which is a pipeline for extracting information regarding bacteria biotopes from scientific abstracts, is presented in Chapter 5.1. For this pipeline, an abstract module, which automatically downloads biomedical abstracts related to bacteria from PubMed is implemented. By utilizing the named entity recognition and normalization modules, which are explained in detail in Chapter 3.2, the bacteria entities and habitat entities are extracted from the relevant abstracts. At the end, by utilizing the relation extraction module, which is explained in detail in Chapter 5.1.5, localization relations and Part-Of relations are extracted between the identified entities.

Finally, the study in Chapter 5.2 reveals that many relations among biomedical entities span multiple sentences. Our sentence-level and abstract-level co-occurrence analysis results suggest that developing text mining methods that operate on scopes wider than a sentence would be useful for biomedical relation extraction.

In this thesis, both rule-based and supervised as well as unsupervised machine-learning based approaches are utilized. The drawback of most supervised machine-

learning based approaches in the biomedical domain is that they require labeled data, which are generally not available. On the other hand, the rule-based approaches proposed in this thesis do not require labeled data. The word embeddings based entity normalization approach is also unsupervised, since it does not require labeled data, but utilizes the large PubMed corpus to learn the word embeddings.

According to a recent study by Chiticariu et al. (2013), although the general belief in academia that the rule-based systems, which require *“tedious manual labor”* to build the rules, in information extraction appear to be *“dead”*, there is still room for improvement in the rule-based approaches [167]. The results of the studies in this thesis also support the idea that rule-based approaches can be as useful as supervised and semi-supervised machine-learning based algorithms and there is still room for improvement in these approaches.

## 6.2. Future Work

Our future directions for research include employing a full syntactic parsing approach to better identify the modifiers of the entities for the biomedical named entity recognition task.

Furthermore, rule-based systems will be utilized to annotate new unlabeled data, creating a new labeled data set. Then, this new labeled data set will be used for training the supervised machine-learning based approaches when the labeled data is scarce.

For the normalization of entities, as future work, we will investigate incorporating the context of the reference entity mentions in text into the vector representations. Error analysis over the training sets revealed that the proposed approach (BOUNEL) is more successful for the normalization of entity mentions whose constituent words have semantic meanings, compared to the entity mentions which contain abbreviations, acronyms, or rare words. We believe that incorporating context information may improve the performance of the system for such entity mentions.

As far as we know, a comprehensive database of locations where bacteria live, is currently not available. As future work, we consider building such a database and a web-based tool which will present the pipeline developed in this thesis as a service to the users.

## REFERENCES

1. Henderson, F. W., W. A. Clyde, A. M. Collier, F. W. Denny, R. Senior, C. Sheaffer, W. Conley and R. Christian, “The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice”, *The Journal of pediatrics*, Vol. 95, No. 2, pp. 35–39, 1979.
2. NCBI, *PubMed*, 2018, <http://www.ncbi.nlm.nih.gov/pubmed>, accessed at December 2018.
3. Cohen, K. B., G. K. Acquah-Mensah, A. E. Dolbey and L. Hunter, “Contrast and variability in gene names”, *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pp. 14–20, Association for Computational Linguistics, 2002.
4. Shen, W., J. Wang and J. Han, “Entity linking with a knowledge base: Issues, techniques, and solutions”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 2, pp. 443–460, 2015.
5. Brown, E. G., L. Wood and S. Wood, “The medical dictionary for regulatory activities (MedDRA)”, *Drug safety*, Vol. 20, No. 2, pp. 109–117, 1999.
6. Leser, U. and J. Hakenberg, “What makes a gene name? Named entity recognition in the biomedical literature”, *Briefings in bioinformatics*, Vol. 6, No. 4, pp. 357–369, 2005.
7. Morgan, A. A., Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg *et al.*, “Overview of BioCreative II gene normalization”, *Genome biology*, Vol. 9, No. 2, p. S3, 2008.
8. Hakenberg, J., C. Plake, R. Leaman, M. Schroeder and G. Gonzalez, “Interspecies normalization of gene mentions with GNAT”, *Bioinformatics*, Vol. 24,



- No. 16, pp. i126–i132, 2008.
9. Lu, Z., H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki *et al.*, “The gene normalization task in BioCreative III”, *BMC bioinformatics*, Vol. 12, No. 8, p. S2, 2011.
  10. Wei, C.-H. and H.-Y. Kao, “Cross-species gene normalization by species inference”, *BMC bioinformatics*, Vol. 12, No. 8, p. S5, 2011.
  11. D’Souza, J. and V. Ng, “Sieve-Based Entity Linking for the Biomedical Domain.”, *ACL (2)*, pp. 297–302, 2015.
  12. Leaman, R., R. Islamaj Doğan and Z. Lu, “DNorm: disease name normalization with pairwise learning to rank”, *Bioinformatics*, Vol. 29, No. 22, pp. 2909–2917, 2013.
  13. Krallinger, M., F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal and A. Valencia, “Overview of the chemical compound and drug name recognition (CHEMDNER) task”, *BioCreative challenge evaluation workshop*, Vol. 2, p. 2, 2013.
  14. Tekir, S. D., T. Cakir, E. Ardic, A. S. Sayilirbas, G. Konuk, M. Konuk, H. Sariyer, A. Ugurlu, I. Karadeniz, A. Ozgur *et al.*, “PHISTO: pathogen–host interaction search tool”, *Bioinformatics*, Vol. 29, No. 10, pp. 1357–1358, 2013.
  15. Karadeniz, İ. and A. Özgür, “Detection and categorization of bacteria habitats using shallow linguistic analysis”, *BMC bioinformatics*, Vol. 16, No. 10, p. S5, 2015.
  16. Karadeniz, I. and A. Özgür, “Bacteria biotope detection, ontology-based normalization, and relation extraction using syntactic rules”, *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 170–177, 2013.
  17. Karadeniz, I., J. Hur, Y. He and A. Özgür, “Literature Mining and Ontology based Analysis of Host-Brucella Gene–Gene Interaction Network”, *Frontiers in*

- microbiology*, Vol. 6, p. 1386, 2015.
18. Krauthammer, M., A. Rzhetsky, P. Morozov and C. Friedman, “Using BLAST for identifying gene and protein names in journal articles”, *Gene*, Vol. 259, No. 1, pp. 245–252, 2000.
  19. Hanisch, D., J. Fluck, H.-T. Mevissen and R. Zimmer, “Playing biology’s name game: identifying protein names in scientific text”, *Biocomputing 2003*, pp. 403–414, World Scientific, 2002.
  20. Tsuruoka, Y. and J. Tsujii, “Improving the performance of dictionary-based approaches in protein name recognition”, *Journal of biomedical informatics*, Vol. 37, No. 6, pp. 461–470, 2004.
  21. Ananiadou, S. and J. McNaught, *Text mining for biology and biomedicine*, Cite-seer, 2006.
  22. Fukuda, K.-i., T. Tsunoda, A. Tamura, T. Takagi *et al.*, “Toward information extraction: identifying protein names from biological papers”, *Pac symp biocomput*, Vol. 707, pp. 707–718, 1998.
  23. Proux, D., F. Rechenmann, L. Julliard, V. Pillet and B. Jacq, “Detecting gene symbols and names in biological texts”, *Genome Informatics*, Vol. 9, pp. 72–80, 1998.
  24. Tanabe, L. and W. J. Wilbur, “Tagging gene and protein names in biomedical text”, *Bioinformatics*, Vol. 18, No. 8, pp. 1124–1132, 2002.
  25. Kazama, J., T. Makino, Y. Ohta and J. Tsujii, “Tuning support vector machines for biomedical named entity recognition”, *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pp. 1–8, Association for Computational Linguistics, 2002.
  26. Lee, K.-J., Y.-S. Hwang, S. Kim and H.-C. Rim, “Biomedical named entity recog-

- dition using two-phase model based on SVMs”, *Journal of Biomedical Informatics*, Vol. 37, No. 6, pp. 436–447, 2004.
27. Collier, N., C. Nobata and J.-i. Tsujii, “Extracting the names of genes and gene products with a hidden Markov model”, *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pp. 201–207, Association for Computational Linguistics, 2000.
  28. Zhou, G. and J. Su, “Named entity recognition using an HMM-based chunk tagger”, *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 473–480, Association for Computational Linguistics, 2002.
  29. Zhou, G., J. Zhang, J. Su, D. Shen and C. Tan, “Recognizing names in biomedical texts: a machine learning approach”, *Bioinformatics*, Vol. 20, No. 7, pp. 1178–1190, 2004.
  30. Settles, B., “Biomedical named entity recognition using conditional random fields and rich feature sets”, *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pp. 104–107, Association for Computational Linguistics, 2004.
  31. Kuo, C.-J., Y.-M. Chang, H.-S. Huang, K.-T. Lin, B.-H. Yang, Y.-S. Lin, C.-N. Hsu and I.-F. Chung, “Rich feature set, unification of bidirectional parsing and dictionary filtering for high F-score gene mention tagging”, *Proceedings of the second BioCreative challenge evaluation workshop*, Vol. 23, pp. 105–107, Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, 2007.
  32. Klinger, R., C. M. Friedrich, J. Fluck and M. Hofmann-Apitius, “Named entity recognition with combinations of conditional random fields”, *Proceedings of the second biocreative challenge evaluation workshop*, 2007.
  33. Rabiner, L. R. and B.-H. Juang, “An introduction to hidden Markov models”, *ieee assp magazine*, Vol. 3, No. 1, pp. 4–16, 1986.

34. Lafferty, J., A. McCallum and F. C. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, , 2001.
35. Zhang, S. and N. Elhadad, “Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts”, *Journal of biomedical informatics*, Vol. 46, No. 6, pp. 1088–1098, 2013.
36. LeCun, Y., Y. Bengio and G. Hinton, “Deep learning”, *nature*, Vol. 521, No. 7553, p. 436, 2015.
37. Santos, C. N. d. and V. Guimaraes, “Boosting named entity recognition with neural character embeddings”, *arXiv preprint arXiv:1505.05008*, 2015.
38. Yao, L., H. Liu, Y. Liu, X. Li and M. W. Anwar, “Biomedical named entity recognition based on deep neutral network”, *corpus*, Vol. 8, No. 8, pp. 279–288, 2015.
39. Li, L., L. Jin, Z. Jiang, D. Song and D. Huang, “Biomedical named entity recognition based on extended recurrent neural networks”, *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pp. 649–652, IEEE, 2015.
40. Lample, G., M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer, “Neural Architectures for Named Entity Recognition”, *Proceedings of NAACL-HLT*, pp. 260–270, 2016.
41. Hirschman, L., A. Yeh, C. Blaschke and A. Valencia, “Overview of BioCreAtIvE: critical assessment of information extraction for biology”, *BMC bioinformatics*, Vol. 6, No. 1, p. S1, 2005.
42. Leitner, F., S. A. Mardis, M. Krallinger, G. Cesareni, L. A. Hirschman and A. Valencia, “An overview of BioCreative II. 5”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, No. 3, pp. 385–399, 2010.

43. Arighi, C., Z. Lu, M. Krallinger, K. Cohen, W. Wilbur, A. Valencia, L. Hirschman and C. Wu, “Overview of the BioCreative III workshop”, *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S1, 2011.
44. Wu, C. H., C. N. Arighi, K. B. Cohen, L. Hirschman, M. Krallinger, Z. Lu, C. Mattingly, A. Valencia, T. C. Wiegiers and W. John Wilbur, “BioCreative-2012 virtual issue”, *Database*, Vol. 2012, 2012.
45. Arighi, C. N., C. H. Wu, K. B. Cohen, L. Hirschman, M. Krallinger, A. Valencia, Z. Lu, J. W. Wilbur and T. C. Wiegiers, “BioCreative-IV virtual issue”, *Database*, Vol. 2014, 2014.
46. Kim, J.-D., T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, “Overview of BioNLP’09 shared task on event extraction”, *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 1–9, Association for Computational Linguistics, 2009.
47. Kim, J.-D., S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen and J. Tsujii, “Overview of BioNLP shared task 2011”, *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 1–6, Association for Computational Linguistics, 2011.
48. Nédellec, C., R. Bossy, J.-D. Kim, J.-j. Kim, T. Ohta, S. Pyysalo and P. Zweigenbaum, “Overview of BioNLP Shared Task 2013”, *ACL 2013*, p. 1, 2013.
49. Deleger, L., R. Bossy, E. Chaix, M. Ba, A. Ferre, P. Bessieres and C. Nédellec, “Overview of the bacteria biotope task at bionlp shared task 2016”, *Proceedings of the 4th BioNLP Shared Task Workshop*, pp. 12–22, 2016.
50. Cohen, A. M., “Unsupervised gene/protein named entity normalization using automatically extracted dictionaries”, *Proceedings of the acl-ismb workshop on linking biological literature, ontologies and databases: Mining biological semantics*, pp. 17–24, Association for Computational Linguistics, 2005.

51. Hanisch, D., K. Fundel, H.-T. Mevissen, R. Zimmer and J. Fluck, “ProMiner: rule-based protein and gene entity recognition”, *BMC bioinformatics*, Vol. 6, No. 1, p. S14, 2005.
52. Bossy, R., W. Golik, Z. Ratkovic, D. Valsamou, P. Bessi eres and C. N edellec, “Overview of the gene regulation network and the bacteria biotope tasks in BioNLP’13 shared task”, *BMC bioinformatics*, Vol. 16, No. 10, p. S1, 2015.
53. Tiftikci, M., H.  ahin, B. B uy  k  z, A. Yayık  ı and A.   zg  r, “Ontology-based Categorization of Bacteria and Habitat Entities using Information Retrieval Techniques”, *Proceedings of the 4th BioNLP Shared Task Workshop*, pp. 56–63, 2016.
54. Fluck, J., H. T. Mevissen, H. Dach, M. Oster and M. Hofmann-Apitius, “ProMiner: recognition of human gene and protein names using regularly updated dictionaries”, *Proceedings of the second BioCreAtIvE challenge evaluation workshop*, pp. 149–151, Centro Nacional de Investigaciones Oncologicas, CNIO, 2007.
55. Cohen, K. B. and L. Hunter, “Natural language processing and systems biology”, *Artificial intelligence methods and tools for systems biology*, pp. 147–173, Springer, 2004.
56. Ghiasvand, O. and R. J. Kate, “UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns.”, *SemEval@ COLING*, pp. 828–832, 2014.
57. Wermter, J., K. Tomanek and U. Hahn, “High-performance gene name normalization with GeNo”, *Bioinformatics*, Vol. 25, No. 6, pp. 815–821, 2009.
58. Li, H., Q. Chen, B. Tang, X. Wang, H. Xu, B. Wang and D. Huang, “CNN-based ranking for biomedical entity normalization”, *BMC bioinformatics*, Vol. 18, No. 11, p. 385, 2017.

59. Cho, H., W. Choi and H. Lee, “A method for named entity normalization in biomedical articles: application to diseases and plants”, *BMC bioinformatics*, Vol. 18, No. 1, p. 451, 2017.
60. Bengio, Y., R. Ducharme, P. Vincent and C. Jauvin, “A neural probabilistic language model”, *Journal of machine learning research*, Vol. 3, No. Feb, pp. 1137–1155, 2003.
61. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, “Distributed representations of words and phrases and their compositionality”, *Advances in neural information processing systems*, pp. 3111–3119, 2013.
62. Chiu, B., G. Crichton, A. Korhonen and S. Pyysalo, “How to train good word embeddings for biomedical NLP”, *Proceedings of BioNLP16*, p. 166, 2016.
63. Sienčnik, S. K., “Adapting word2vec to named entity recognition”, *Proceedings of the 20th nordic conference of computational linguistics, nodalida 2015, may 11-13, 2015, vilnius, lithuania*, 109, pp. 239–243, Linköping University Electronic Press, 2015.
64. Chen, X., Z. Liu and M. Sun, “A unified model for word sense representation and disambiguation”, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1025–1035, 2014.
65. Taghipour, K. and H. T. Ng, “Semi-supervised word sense disambiguation using word embeddings in general and specific domains”, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 314–323, 2015.
66. Ganguly, D., D. Roy, M. Mitra and G. J. Jones, “Word embedding based generalized language model for information retrieval”, *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 795–798, ACM, 2015.

67. Diaz, F., B. Mitra and N. Craswell, “Query expansion with locally-trained word embeddings”, *arXiv preprint arXiv:1605.07891*, 2016.
68. Zou, W. Y., R. Socher, D. Cer and C. D. Manning, “Bilingual word embeddings for phrase-based machine translation”, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1393–1398, 2013.
69. Moen, S. and T. S. S. Ananiadou, “Distributional semantics resources for biomedical text processing”, *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pp. 39–43, 2013.
70. TH, M., S. Sahu and A. Anand, “Evaluating distributed word representations for capturing semantics of biomedical concepts”, *Proceedings of BioNLP 15*, pp. 158–163, 2015.
71. Aydın, F., Z. M. Hüsünbeyi and A. Özgür, “Automatic query generation using word embeddings for retrieving passages describing experimental methods”, *Database*, Vol. 2017, No. 1, 2017.
72. Cortes, C., P. Haffner and M. Mohri, “Rational kernels: Theory and algorithms”, *Journal of Machine Learning Research*, Vol. 5, No. Aug, pp. 1035–1062, 2004.
73. Kumar, R. and B. Nanduri, “HPIDB-a unified resource for host-pathogen interactions”, *BMC bioinformatics*, Vol. 11, No. Suppl 6, p. S16, 2010.
74. Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering *et al.*, “The HUPO PSI’s molecular interaction format-a community standard for the representation of protein interaction data”, *Nature biotechnology*, Vol. 22, No. 2, pp. 177–183, 2004.
75. Hedley, J., *JSoup*, 2009, <http://jsoup.org>, accessed at December 2018.
76. Dyer, M. D., C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. Murali and B. W. Sobral, “The human-bacterial pathogen protein interaction



- networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*”, *PloS one*, Vol. 5, No. 8, p. e12089, 2010.
77. Bossy, R., W. Golik, Z. Ratkovic, P. Bessi eres and C. N edellec, “BioNLP shared Task 2013 – An Overview of the Bacteria Biotope Task”, *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 161–169, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.
  78. N edellec, C., R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo and P. Zweigenbaum, “Overview of BioNLP Shared Task 2013”, *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 1–7, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.
  79. N edellec, C., “Learning language in logic-genic interaction extraction challenge”, *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, Vol. 7, Citeseer, 2005.
  80. Kim, J.-D., T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, “Overview of BioNLP’09 Shared Task on Event Extraction”, *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 1–9, Association for Computational Linguistics, Boulder, Colorado, June 2009.
  81. INRA, *Onto-Biotope Ontology*, 2013, <http://bibliome.jouy.inra.fr/MEM-OntoBiotope>, accessed at December 2018.
  82. Krallinger, M., F. Leitner, C. Rodriguez-penagos and A. Valencia, “Overview of the protein-protein interaction annotation extraction task of BioCreative II”, *Genome Biology*, pp. 2–4, 2008.
  83. Kim, J.-D., T. Ohta, S. Pyysalo, Y. Kano and J. Tsujii, “Overview of BioNLP’09 shared task on event extraction”, *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP ’09, pp. 1–9, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

84. Bossy, R., J. Jourde, P. Bessi eres, M. van de Guchte and C. N edellec, “BioNLP Shared Task 2011: bacteria biotope”, *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task ’11, pp. 56–64, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011.
85. Bossy, R., J. Jourde, A. P. Manine, P. Veber, E. Alphonse, M. van de Guchte, P. Bessieres and C. Nedellec, “BioNLP Shared Task - The Bacteria Track”, *BMC Bioinformatics*, Vol. 13, No. Suppl 11, pp. S3+, 2012.
86. Ratkovic, Z., W. Golik and P. Warnier, “Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach”, *BMC Bioinformatics*, Vol. 13, pp. S8+, 2012.
87. FAO, *Agrovoc*, 2018, <http://aims.fao.org/vest-registry/vocabularies/agrovoc>, accessed at December 2018.
88. NCBI, *NCBI Taxonomy*, 2018, <http://www.ncbi.nlm.nih.gov/Taxonomy/>, accessed at December 2018.
89. Bj orne, J., F. Ginter and T. Salakoski, “University of Turku in the BioNLP’11 Shared Task”, *BMC Bioinformatics*, Vol. 13 Suppl 11, p. S4, 2012.
90. Nguyen, N. T. H. and Y. Tsuruoka, “Extracting bacteria biotopes with semi-supervised named entity recognition and coreference resolution”, *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task ’11, pp. 94–101, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011.
91. Bannour, S., L. Audibert and H. Soldano, “Ontology-based semantic annotation: an automatic hybrid rule-based method”, *ACL 2013*, pp. 139–143, 2013.
92. Claveau, V., “IRISA participation to BioNLP-ST 2013: lazy-learning and information retrieval for information extraction tasks”, *Proceedings of the BioNLP Workshop*, pp. 188–196, 2013.

93. Grouin, C., “Building A Contrasting Taxa Extractor for Relation Identification from Assertions: BIOlogical Taxonomy & Ontology Phrase Extraction System”, *ACL 2013*, p. 144, 2013.
94. Sutton, C. and A. McCallum, “An introduction to conditional random fields for relational learning”, *Introduction to statistical relational learning*, Vol. 93, pp. 142–146, 2007.
95. Technologies, R., *Cocoa*, 2012, <http://npjoint.com/annotate.php>, accessed at May 2013.
96. Fukuda, K., A. Tamura, T. Tsunoda and T. Takagi, “Toward information extraction: identifying protein names from biological papers”, *Pac Symp Biocomput*, pp. 707–718, 1998.
97. Tsuruoka, Y., Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou and J. Tsujii, “Developing a Robust Part-of-Speech Tagger for Biomedical Text”, *Advances in Informatics*, Vol. 3746, chap. 36, pp. 382–392, Springer, Berlin Heidelberg, 2005.
98. Leaman, R. and G. Gonzalez, “BANNER: an executable survey of advances in biomedical named entity recognition”, *Pac Symp Biocomput*, pp. 652–663, 2008.
99. Jacquemin, C. and E. Tzoukermann, “NLP for term variant extraction: synergy between morphology, lexicon, and syntax”, pp. 25–74, 1999.
100. Aronson, A. R., “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.”, *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.
101. Krallinger, M., M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli *et al.*, “The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking

- bio-ontology concepts to full text”, *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S3, 2011.
102. Schneider, G., S. Clematide and F. Rinaldi, “Detection of interaction articles and experimental methods in biomedical literature”, *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S13, 2011.
  103. Wang, X., R. Rak, A. Restificar, C. Nobata, C. Rupp, R. T. B. Batista-Navarro, R. Nawaz and S. Ananiadou, “Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature”, *BMC bioinformatics*, Vol. 12, No. Suppl 8, p. S11, 2011.
  104. Saetre, R., K. Yoshida, A. Yakushiji, Y. Miyao, Y. Matsubayashi and T. Ohta, “AKANE System: Protein-Protein Interaction Pairs in the BioCreAtIvE2 Challenge, PPI-IPS subtask”, L. Hirschman, M. Krallinger and A. Valencia (Editors), *Proceedings of the Second BioCreative Challenge Workshop*, 2007.
  105. Tsuruoka, Y. and J. Tsujii, “Bidirectional inference with the easiest-first strategy for tagging sequence data”, *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 467–474, Association for Computational Linguistics, 2005.
  106. Wang, J. Z., Z. Du, R. Payattakool, P. S. Yu and C. F. Chen, “A new method to measure the semantic similarity of GO terms”, *Bioinformatics*, Vol. 23, No. 10, pp. 1274–1281, May 2007.
  107. Cohen, A. M. and W. R. Hersh, “A survey of current work in biomedical text mining”, *Briefings in bioinformatics*, Vol. 6, No. 1, pp. 57–71, 2005.
  108. Spasic, I., S. Ananiadou, J. McNaught and A. Kumar, “Text mining and ontologies in biomedicine: making sense of raw text”, *Briefings in bioinformatics*, Vol. 6, No. 3, pp. 239–251, 2005.

109. Rubin, D. L., N. H. Shah and N. F. Noy, “Biomedical ontologies: a functional perspective”, *Briefings in bioinformatics*, Vol. 9, No. 1, pp. 75–90, 2007.
110. Blaschke, C., L. Hirschman and A. Valencia, “Information extraction in molecular biology”, *Briefings in Bioinformatics*, Vol. 3, No. 2, pp. 154–165, 2002.
111. Bossy, R., J. Jourde, P. Bessieres, M. Van De Guchte and C. Nédellec, “Bionlp shared task 2011: bacteria biotope”, *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 56–64, Association for Computational Linguistics, 2011.
112. Cook, H. V., E. Pafilis and L. J. Jensen, “A dictionary-and rule-based system for identification of bacteria and habitats in text”, *ACL 2016*, Vol. 50, 2016.
113. Grouin, C., “Identification of mentions and relations between bacteria and biotope from pubmed abstracts”, *ACL 2016*, p. 64, 2016.
114. Ferré, A., P. Zweigenbaum and C. Nédellec, “Representation of complex terms in a vector space structured by an ontology for a normalization task”, *BioNLP 2017*, pp. 99–106, 2017.
115. Mehryary, F., K. Hakala, S. Kaewphan, J. Björne, T. Salakoski and F. Ginter, “End-to-End System for Bacteria Habitat Extraction”, *BioNLP 2017*, pp. 80–90, 2017.
116. Gurulingappa, H., 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.
117. Nikfarjam, A., A. Sarker, K. O Connor, R. Ginn and G. Gonzalez, “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features”, *Journal of the American Medical Informatics Association*, Vol. 22, No. 3, pp. 671–681, 2015.
118. Lindberg, D. A., B. L. Humphreys and A. T. McCray, “The unified medical

- language system”, *Methods of information in medicine*, Vol. 32, No. 04, pp. 281–291, 1993.
119. Wadhwa, S., A. Gupta, S. Dokania, R. Kanji and G. Bagler, “A hierarchical anatomical classification schema for prediction of phenotypic side effects”, *PloS one*, Vol. 13, No. 3, p. e0193959, 2018.
  120. Kusner, M., Y. Sun, N. Kolkin and K. Weinberger, “From word embeddings to document distances”, *International Conference on Machine Learning*, pp. 957–966, 2015.
  121. Klein, D. and C. D. Manning, “Accurate unlexicalized parsing”, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 423–430, Association for Computational Linguistics, 2003.
  122. Wang, J. Z., Z. Du, R. Payattakool, P. S. Yu and C.-F. Chen, “A new method to measure the semantic similarity of GO terms”, *Bioinformatics*, Vol. 23, No. 10, pp. 1274–1281, 2007.
  123. Sætre, R., K. Sagae and J. Tsujii, “Syntactic Features for Protein-Protein Interaction Extraction.”, *LBM (Short Papers)*, 2007.
  124. Björne, J. and T. Salakoski, “TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task”, *ACL 2013*, p. 16, 2013.
  125. Jelier, R., G. Jenster, L. C. Dorssers, C. van der Eijk, E. M. van Mulligen, B. Mons and J. A. Kors, “Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes”, *Bioinformatics*, Vol. 21, No. 9, pp. 2049–2058, 2005.
  126. He, M., Y. Wang and W. Li, “PPI finder: a mining tool for human protein-protein interactions”, *PloS one*, Vol. 4, No. 2, p. e4554, 2009.
  127. Blaschke, C. and A. Valencia, “The frame-based module of the Suiseki information

- extraction system”, *IEEE Intelligent Systems*, , No. 17, pp. 14–20, 2002.
128. Tari, L., S. Anwar, S. Liang, J. Cai and C. Baral, “Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism”, *Bioinformatics*, Vol. 26, No. 18, pp. i547–i553, 2010.
  129. Chang, D. T.-H., C.-H. Ke, J.-H. Lin and J.-H. Chiang, “AutoBind: automatic extraction of protein–ligand-binding affinity data from biological literature”, *Bioinformatics*, Vol. 28, No. 16, pp. 2162–2168, 2012.
  130. Rosario, B. and M. A. Hearst, “Classifying semantic relations in bioscience texts”, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 430, Association for Computational Linguistics, 2004.
  131. Fundel, K., R. Kuffner and R. Zimmer, “RelEx–Relation extraction using dependency parse trees”, *Bioinformatics*, Vol. 23, No. 3, pp. 365–371, February 2007.
  132. Erkan, G., A. Özgür and D. R. Radev, “Semi-Supervised Classification for Extracting Protein Interaction Sentences using Dependency Parsing”, *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 228–237, 2007.
  133. Bui, Q.-C., S. Katrenko and P. M. Sloot, “A hybrid approach to extract protein–protein interactions”, *Bioinformatics*, Vol. 27, No. 2, pp. 259–265, 2011.
  134. Björne, J., J. Heimonen, F. Ginter, A. Airola, T. Pahikkala and T. Salakoski, “Extracting complex biological events with rich graph-based feature sets”, *BioNLP ’09: Proceedings of the Workshop on BioNLP*, pp. 10–18, Association for Computational Linguistics, Morristown, NJ, USA, 2009.
  135. O’Callaghan, D. and A. M. Whatmore, “Brucella genomics as we enter the multi-genome era”, *Briefings in functional genomics*, Vol. 10, No. 6, pp. 334–341, 2011.

136. Halling, S. M., B. D. Peterson-Burch, B. J. Bricker, R. L. Zuerner, Z. Qing, L.-L. Li, V. Kapur, D. P. Alt and S. C. Olsen, “Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*”, *Journal of Bacteriology*, Vol. 187, No. 8, pp. 2715–2726, 2005.
137. Xiang, Z., Y. Tian, Y. He *et al.*, “PHIDIAS: a pathogen-host interaction data integration and analysis system”, *Genome Biol*, Vol. 8, No. 7, p. R150, 2007.
138. Durmuş Tekir, S., T. Çakır, E. Ardiç, A. S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür *et al.*, “PHISTO: pathogen–host interaction search tool”, *Bioinformatics*, Vol. 29, No. 10, pp. 1357–1358, 2013.
139. Arenas-Gamboa, A. M., T. A. Ficht, M. M. Kahl-McDonagh and A. C. Rice-Ficht, “Immunization with a single dose of a microencapsulated *Brucella melitensis* mutant enhances protection against wild-type challenge”, *Infection and immunity*, Vol. 76, No. 6, pp. 2448–2455, 2008.
140. Ono, T., H. Hishigaki, A. Tanigami and T. Takagi, “Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature”, *Bioinformatics*, Vol. 17, No. 2, pp. 155–161, 2001.
141. Giuliano, C., A. Lavelli and L. Romano, “Exploiting shallow linguistic information for relation extraction from biomedical literature.”, *EACL*, Vol. 18, pp. 401–408, Citeseer, 2006.
142. Erkan, G., A. Özgür and D. R. Radev, “Semi-supervised classification for extracting protein interaction sentences using dependency parsing.”, *EMNLP-CoNLL*, Vol. 7, pp. 228–237, 2007.
143. Airola, A., S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter and T. Salakoski, “All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning”, *BMC bioinformatics*, Vol. 9, No. Suppl 11, p. S2, 2008.



144. Tikk, D., I. Solt, P. Thomas and U. Leser, “A detailed error analysis of 13 kernel methods for protein–protein interaction extraction”, *BMC bioinformatics*, Vol. 14, No. 1, p. 12, 2013.
145. Tanabe, L., N. Xie, L. H. Thom, W. Matten and W. J. Wilbur, “GENETAG: a tagged corpus for gene/protein named entity recognition”, *BMC bioinformatics*, Vol. 6, No. 1, p. S3, 2005.
146. Tsai, R. T., C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung and W.-L. Hsu, “NER-Bio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition”, *BMC bioinformatics*, Vol. 7, No. Suppl 5, p. S11, 2006.
147. Hsu, C.-N., Y.-M. Chang, C.-J. Kuo, Y.-S. Lin, H.-S. Huang and I.-F. Chung, “Integrating high dimensional bi-directional parsing models for gene mention tagging”, *Bioinformatics*, Vol. 24, No. 13, pp. i286–i294, 2008.
148. Mcdonald, R. and F. Pereira, “Identifying gene and protein mentions in text using conditional random fields”, *BMC Bioinformatics*, Vol. 6, No. Suppl 1, pp. S6+, 2005.
149. Hur, J., A. D. Schuyler, E. L. Feldman *et al.*, “SciMiner: web-based literature mining tool for target identification and functional enrichment analysis”, *Bioinformatics*, Vol. 25, No. 6, pp. 838–840, 2009.
150. Hur, J., Z. Xiang, E. L. Feldman and Y. He, “Ontology-based Brucella vaccine literature indexing and systematic analysis of gene-vaccine association network”, *BMC immunology*, Vol. 12, No. 1, p. 49, 2011.
151. Xiang, Z., T. Qin, Z. S. Qin and Y. He, “A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks”, *BMC systems biology*, Vol. 7, No. Suppl 3, p. S9, 2013.

152. Özgür, A., Z. Xiang, D. R. Radev and Y. He, “Mining of vaccine-associated IFN-g gene interaction networks using the Vaccine Ontology”, , 2011.
153. Hur, J., A. Özgür, Z. Xiang and Y. He, “Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining.”, *J. Biomedical Semantics*, Vol. 3, p. 18, 2012.
154. Hur, J., A. Özgür, Z. Xiang and Y. He, “Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions”, *Journal of Biomedical Semantics*, Vol. 6, No. 1, p. 2, 2015.
155. Durmuş, S., T. Çakır, A. Özgür and R. Guthke, “A review on computational systems biology of pathogen–host interactions”, *Frontiers in microbiology*, Vol. 6, 2015.
156. Yin, L., G. Xu, M. Torii, Z. Niu, J. M. Maisog, C. Wu, Z. Hu and H. Liu, “Document classification for mining host pathogen protein–protein interactions”, *Artificial intelligence in medicine*, Vol. 49, No. 3, pp. 155–160, 2010.
157. Thieu, T., S. Joshi, S. Warren and D. Korkin, “Literature mining of host–pathogen interactions: comparing feature-based supervised learning and language-based approaches”, *Bioinformatics*, Vol. 28, No. 6, pp. 867–875, 2012.
158. Tikk, D., P. Thomas, P. Palaga, J. Hakenberg and U. Leser, “A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature”, *PLoS computational biology*, Vol. 6, No. 7, p. e1000837, 2010.
159. HGNC, *HUGO Gene Nomenclature Committee (HGNC) database*, 2018, <http://www.genenames.org/>, accessed at December 2018.
160. Joachims, T., *Making large scale SVM learning practical*, Tech. rep., Universität Dortmund, 1999.
161. De Marneffe, M.-C., B. MacCartney, C. D. Manning *et al.*, “Generating typed

- dependency parses from phrase structure parses”, *Proceedings of LREC*, Vol. 6, pp. 449–454, 2006.
162. Rosinha, G. M., A. Myioshi, V. Azevedo, G. A. Splitter and S. C. Oliveira, “Molecular and immunological characterisation of recombinant *Brucella abortus* glyceraldehyde-3-phosphate-dehydrogenase, a T-and B-cell reactive protein that induces partial protection when co-administered with an interleukin-12-expressing plasmid in a DNA vaccine formulation”, *Journal of medical microbiology*, Vol. 51, No. 8, pp. 661–671, 2002.
  163. Krallinger, M., *Biocreative II*, 2006, <http://biocreative.sourceforge.net/biocreative2.html>, accessed at December 2018.
  164. Bunescu, R., R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani and Y. W. Wong, “Comparative experiments on learning information extractors for proteins and their interactions”, *Artificial intelligence in medicine*, Vol. 33, No. 2, pp. 139–155, 2005.
  165. Al-Mariri, A., A. Tibor, P. Mertens, X. De Bolle, P. Michel, J. Godefroid, K. Walravens and J.-J. Letesson, “Protection of BALB/c mice against *Brucella abortus* 544 challenge by vaccination with bacterioferritin or P39 recombinant proteins with CpG oligodeoxynucleotides as adjuvant”, *Infection and immunity*, Vol. 69, No. 8, pp. 4816–4822, 2001.
  166. Özgür, A., J. Hur and Y. He, “Extension of the Interaction Network Ontology for Literature Mining of Gene-gene Interaction Networks from Sentences with Multiple Interaction Keywords.”, *BDM2I@ ISWC*, Citeseer, 2015.
  167. Chiticariu, L., Y. Li and F. R. Reiss, “Rule-based information extraction is dead! long live rule-based information extraction systems!”, *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 827–832, 2013.