BACTERIA BIOTOPE 2016

Information Standardization via Ontologies

on Biomedical Domain

Rıza ÖZÇELİK Selen PARLAR

Outline

- Task Definition
- Motivation
- Related Work
- Methods
  - Data set
  - Proposed Solution
  - Evaluation
- References

Task Definition

- **Shared task** = Multi-disciplinary approaches + Structured data + Standard evaluation
- BioNLP Shared Task 2016 has 3 sub-tasks:

1. Bacteria and habitat detection and categorization:
    - Mapping bacteria and habitat entities from text to the concepts in the NCBI Taxonomy and OntoBiotope ontology.
2. Entity and event extraction:
    - Extracting events from text to learn interactions between bacteria, habitat and geographical entities.
3. Knowledge Base extraction:
    - Constructing knowledge bases from text automatically.

Task Definition

- **Shared task** = Multi-disciplinary approaches + Structured data + Standard evaluation
- BioNLP Shared Task 2016 has 3 sub-tasks:

1. **Bacteria and habitat detection and categorization:**
   - **Mapping bacteria and habitat entities from text to the concepts in the NCBI Taxonomy and OntoBiotope ontology.**
2. Entity and event extraction:
   - Extracting events from text to learn interactions between bacteria, habitat and geographical entities.
3. Knowledge Base extraction:
   - Constructing knowledge bases from text automatically.

Motivation

- Representing interaction between bacteria and environment in a standard way is valuable.

Motivation

- Representing interaction between bacteria and environment in a standard way is valuable.
- Habitat information is especially critical for applied microbiology.

Motivation

- Representing interaction between bacteria and environment in a standard way is valuable.
- Habitat information is especially critical for applied microbiology.
- Habitat - bacteria information is prevalent in unstructured free text. Ontologies provide normalized and comprehensive information.

Motivation

- Representing interaction between bacteria and environment in a standard way is valuable.
- Habitat information is especially critical for applied microbiology.
- Habitat - bacteria information is prevalent in unstructured free text. Ontologies provide normalized and comprehensive information.
- Shared task **2016** provides structured data, evaluation method and has been studied.

GOAL

Mapping bacteria and habitat mentions in text to large ontologies

Related Work

Representation of complex terms in a vector space structured by an ontology for a normalization task (Ferré, Zweigenbaum, and Nédellec)

- Learn embeddings for both text and ontology concepts.
- Biomedical word embeddings are learned by word2vec. (Available on demand)
- Concept embeddings are learned by PCA.
- Learn word-to-concept mapping by a linear model.

Linking entities through an ontology using word embeddings and syntactic re-ranking (Karadeniz & Ozgur)

- Unsupervised.
- Use word embeddings to represent semantic spaces, and a syntactic parser to give higher weight to the most informative word in the named entity mentions.
- Concept embedding by names and synonyms.

Related Work

Representation of complex terms in a vector space structured by an ontology for a normalization task (Ferré, Zweigenbaum, and Nédellec)

- Learn embeddings for both text and ontology concepts.
- Biomedical word embeddings are learned by word2vec. (Available on demand)
- Concept embeddings are learned by PCA.
- Learn word-to-concept mapping by a linear model.

Linking entities through an ontology using word embeddings and syntactic re-ranking (Karadeniz & Ozgur)

- Unsupervised.
- Use word embeddings to represent semantic spaces, and a syntactic parser to give higher weight to the most informative word in the named entity mentions.
- Concept embedding by names and synonyms.

Data Set

Contains two components

Annotated PubMed Abstracts

Two Ontologies

Data Set

→ Annotated PubMed abstracts

	Train	Dev	Test	Total
Documents	71	36	54	161
Words	16,295	8,890	13,797	38,982
Entities	747	454	720	1,921

(Ferré, Zweigenbaum, and Nédellec)

Data Set

→ Two Ontologies:

Ontology	Number of Classes
OntoBiotope habitat ontology (OBO)	2,320
NCBI taxonomy	1,412,456

PubMed Abstract

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.

Annotated PubMed Abstract

T1 Title 0 80 The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice.

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

Annotation Format (.a1 file)

T1 Title 0 80 The etiologic and epidemiologic spectrum of bronchiolitis in pediatric practice.

T2 Paragraph 81 1213 To develop a broad understanding of the causes and patterns of occurrence of wheezing associated respiratory infections, we analyzed data from an 11-year study of acute lower respiratory illness in a pediatric practice … 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...

T3	Habitat 61 70	pediatric
T4	Habitat 178 189	respiratory
T5	Habitat 256 267	respiratory
T6	Habitat 281 290	pediatric
...		
T17	Habitat 904 933	children under 5 years of age
T18	Bacteria 950 971	Mycoplasma pneumoniae
T19	Habitat 1007 1048	school age children with wheezing illness
T20	Habitat 1124 1138	diagnosticians
T21	Habitat 1146 1157	researchers
T22	Habitat 1193 1204	respiratory

Ontology Linking (.a2 file)

```
T3	Habitat 61 70	pediatric
T4	Habitat 178 189	respiratory
T5	Habitat 256 267	respiratory
T6	Habitat 281 290	pediatric
T7	Habitat 339 372	children less than 2 years of age
...
T19	Habitat 1007 1048	school age children with wheezing illness
T20	Habitat 1124 1138	diagnosticians
T21	Habitat 1146 1157	researchers
T22	Habitat 1193 1204	respiratory
```

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

```
N1	OntoBiotope Annotation:T3 Referent:OBT:002307
N2	OntoBiotope Annotation:T4 Referent:OBT:000164
N3	OntoBiotope Annotation:T5 Referent:OBT:000164
N4	OntoBiotope Annotation:T6 Referent:OBT:002307
N5	OntoBiotope Annotation:T7 Referent:OBT:002307
...
N24	OntoBiotope Annotation:T19 Referent:OBT:002307
N25	OntoBiotope Annotation:T19 Referent:OBT:002187
N26	OntoBiotope Annotation:T20 Referent:OBT:002252
N27	OntoBiotope Annotation:T21 Referent:OBT:002265
N28	OntoBiotope Annotation:T22 Referent:OBT:000164
```

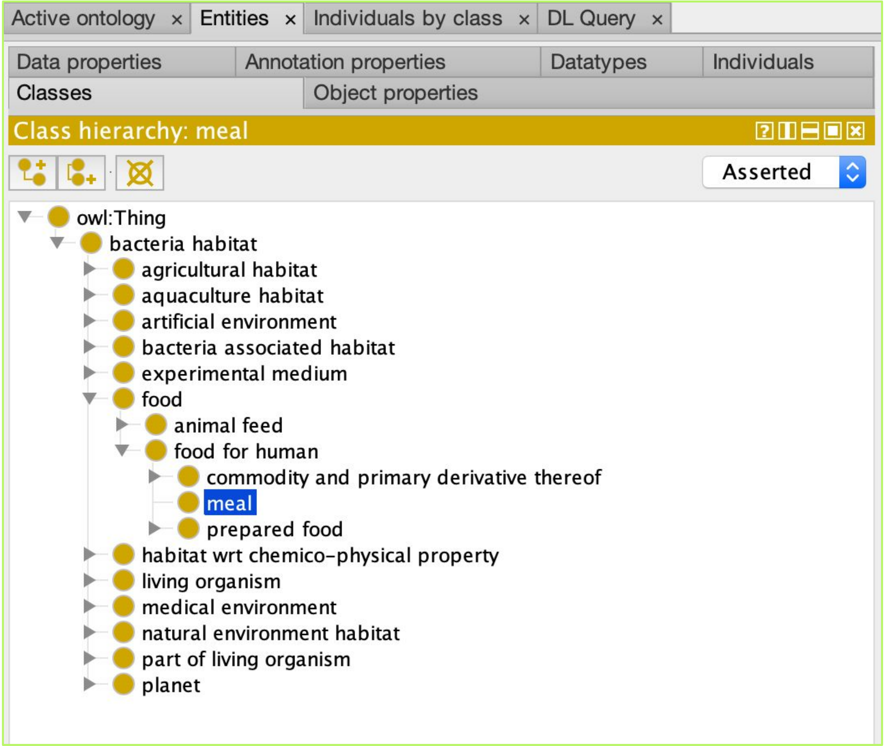
```
T3	Habitat 61 70	pediatric
T4	Habitat 178 189	respiratory
T5	Habitat 256 267	respiratory
T6	Habitat 281 290	pediatric
T7	Habitat 339 372	children less than 2 years of age
...
T21	Habitat 1146 1157	researchers
T22	Habitat 1193 1204	respiratory
```

+

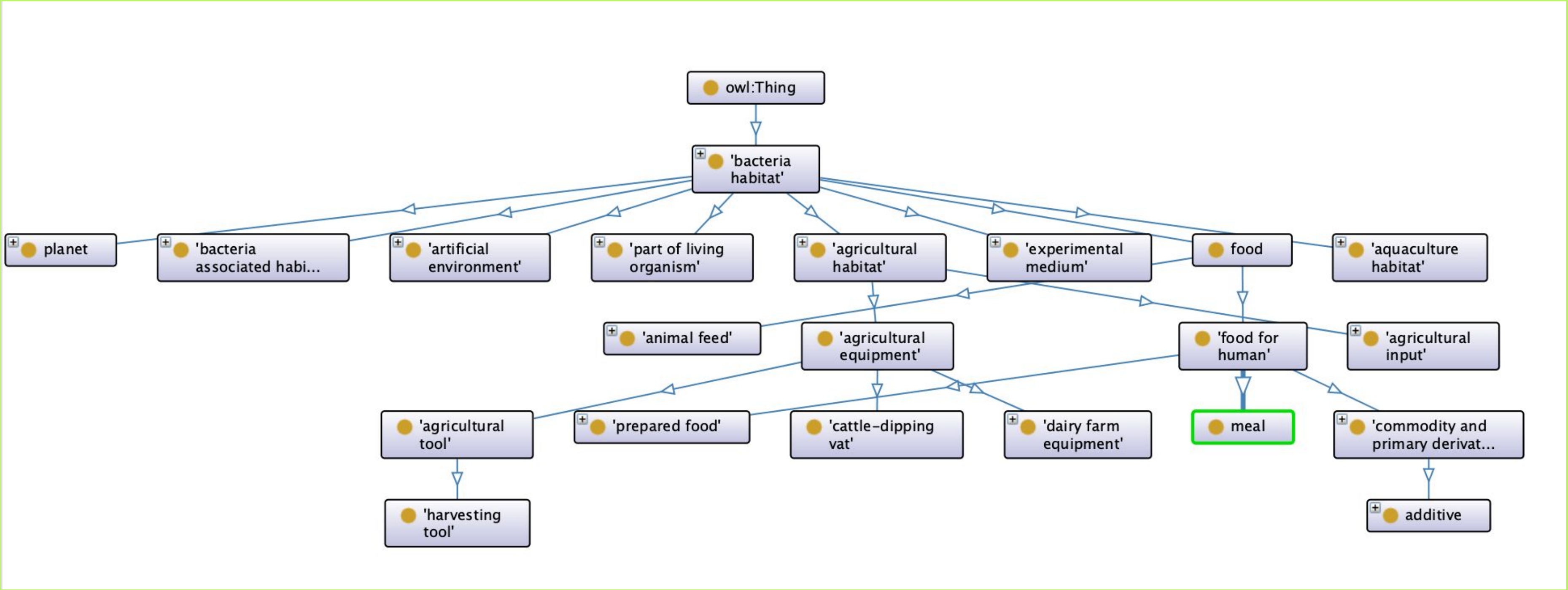
```
N1	OntoBiotope Annotation:T3 Referent:OBT:002307
N2	OntoBiotope Annotation:T4 Referent:OBT:000164
N3	OntoBiotope Annotation:T5 Referent:OBT:000164
N4	OntoBiotope Annotation:T6 Referent:OBT:002307
N5	OntoBiotope Annotation:T7 Referent:OBT:002307
...
N27	OntoBiotope Annotation:T21 Referent:OBT:002265
N28	OntoBiotope Annotation:T22 Referent:OBT:000164
```

Entity Normalization →

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

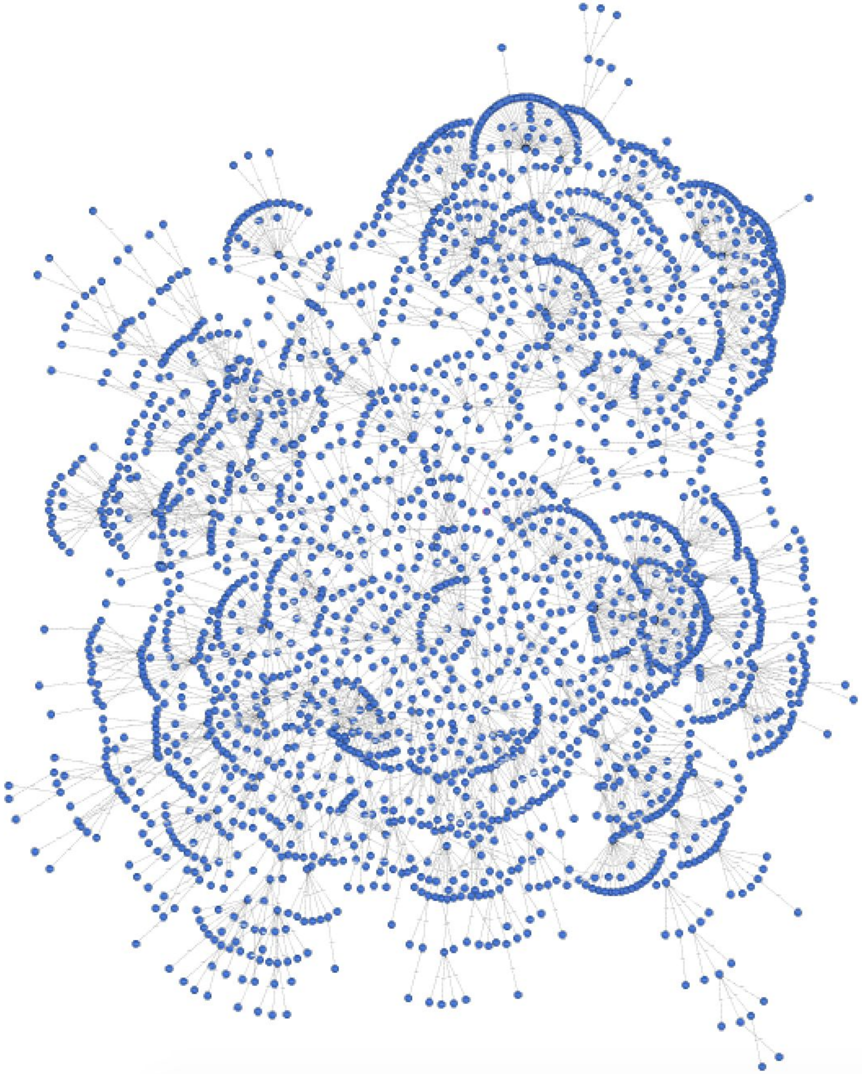


Hierarchy in Protégé



Hierarchy in Protégé

OntoBiotope in Protégé



Our Approach

We will use two different network embedding techniques:

- Learning concept vectors of ontologies by using interlinks in .a2 files (supervised)
- Learning unsupervised concept vectors by treating ontology as graph
- (If time permits) Combining these in an end-to-end manner similar to GraphSAGE

Note: Every method will conduct exact search first!

Evaluation

- The evaluation service is available online and uses a modified precision for tree based evaluation.

Model	Precision
BOUNEL	0.659
TURKU	0.630
BOUN	0.620
CONTES	0.597
LIMSI	0.438

Thanks!

ANY IDEAS?

You can find us at

- riza.ozcelik@boun.edu.tr
- selen.parlar@boun.edu.tr

References

- Presentation template by SlidesCarnival
- BioNLP ST - 2016
- BioNLP ST - 2019