Term Project Progress Presentation

MANSUR YEŞİLBURSA
ABDULLAH YILDIZ
CAN DEVECİ

Overview

- Part 1: Pre-processing and Baseline, Mansur Yeşilbursa
- Part 2: Performance Evaluations, Abdullah Yıldız
- Part 3: Future Work, Can Deveci

Preprocessing

Dataset class

- ✓Processes .a1 and .a2 files in the given set (train, dev or test)
- ✓Stores Term number-Term name tuples and Term number-OBT tuples
- ✓Stores position of terms in the txt files
- ✓Creates a vocabulary if training set is given
- ✓Has access to Ontology
- ✓No case-folding

```
['T3', 'selective broths based on hypertonic strontium chloride']
```

```
['T3', '000360']
```

Baseline System

Exact Match

- ✓ **On Training Set**
  - Searches training set to find exact entity match
  - If finds, retrieves OBT of the matched entity
  - **Performance**: 22% correct normalization on dev set
- ✓ **On Ontology**
  - Searches Ontology file to find exact entity match
  - If finds, retrieves OBT of the matched entity
  - Improved performance by 11%
  - **Performance**: 33% correct normalization on dev set

Baseline System

➢ Exact Match

- ✓ **Lemmatization**
  - For given test entities, lemmatization is applied
  - If lemmatized entity is different than original entity
    - Searches for both terms in Training set and Ontology
  - Improved performance by 4%
  - **Performance:** 37% correct normalization in dev set
- ✓ **Abbreviation Resolution**
  - Requires a comprehensive Biomedical Abbreviation Database
  - One that is used was messy and not comprehensive enough [1]
  - Didn’t improved performance for now

Baseline System

- Cosine Similarity
  - Used if exact match fails
  - Measures cosine similarity of given entity with all entities in the Training set
  - Binary vectors of Training set vocabulary size
  - Improves performance by 10%
  - Performance: 47% correct normalization on dev set

Performance Problems

Input Term	Normalized Term	True Normalized Term
CD20-positive cells	phagocyte	lymphocyte
lymphocytic	nutrient broth	lymphocyte
B cells	phagocyte	lymphocyte
T cells	phagocyte	lymphocyte
lymphocytic	nutrient broth	lymphocyte

- Predicts phagocyte instead of lymphocyte although input containts lymphocytic.
- e.g. lymphocytic -> lymphocytic
- Lemmatization could not lemmatize scientific terms
- Cells -> phagocyte
- ..cytic -> nutrient broth

Performance Problems

Input Term	Normalized Term	True Normalized Term
patients with atypical lymphoid infiltrates	elderly person	patient
patients with low-grade MALT lymphoma	elderly person	patient
patients with Helicobacter pylori-chronic active gastritis	elderly person	patient with infectious disease
patients with high-grade primary gastric lymphoma	elderly person	patient
patients with chronic active gastritis	elderly person	patient

- Predicts elderly person instead of patient
- Associates certain words to certain entities

Performance Problems

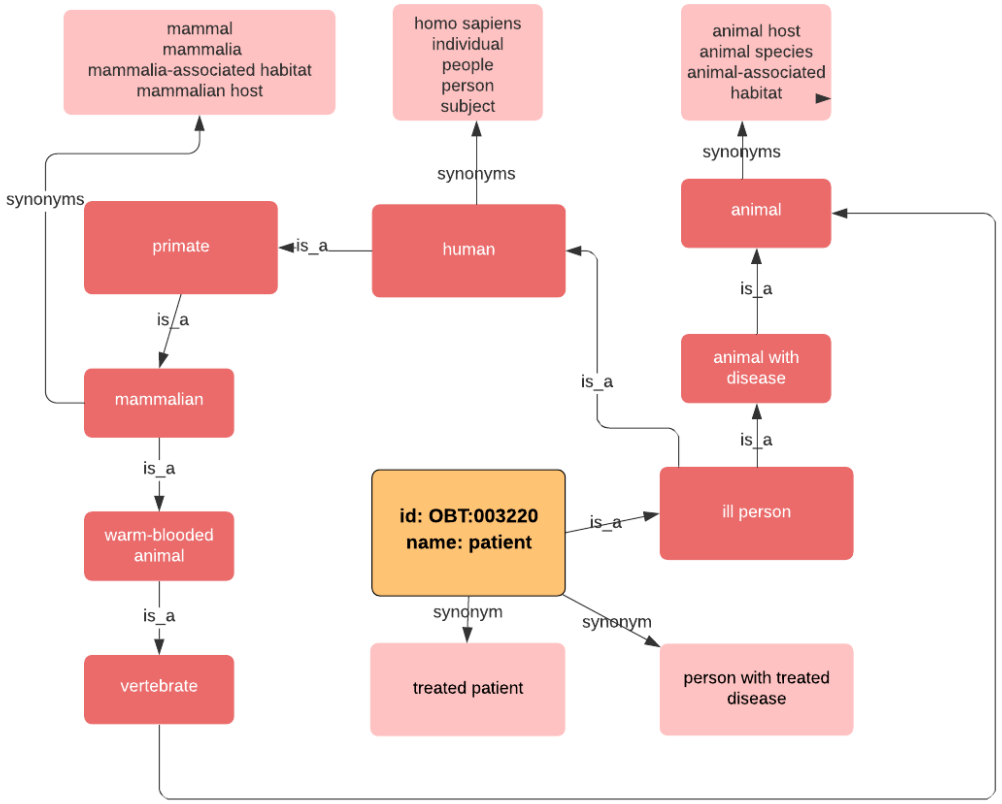
Input Term	Normalized Term	True Normalized Term
placenta of a 38-year-old secondary recurrent aborter	experimental medium	placenta
38-year-old secondary recurrent aborter	experimental medium	pregnant woman

- Lack of Part of Speech tagging
- Lack of phrases

Ontology

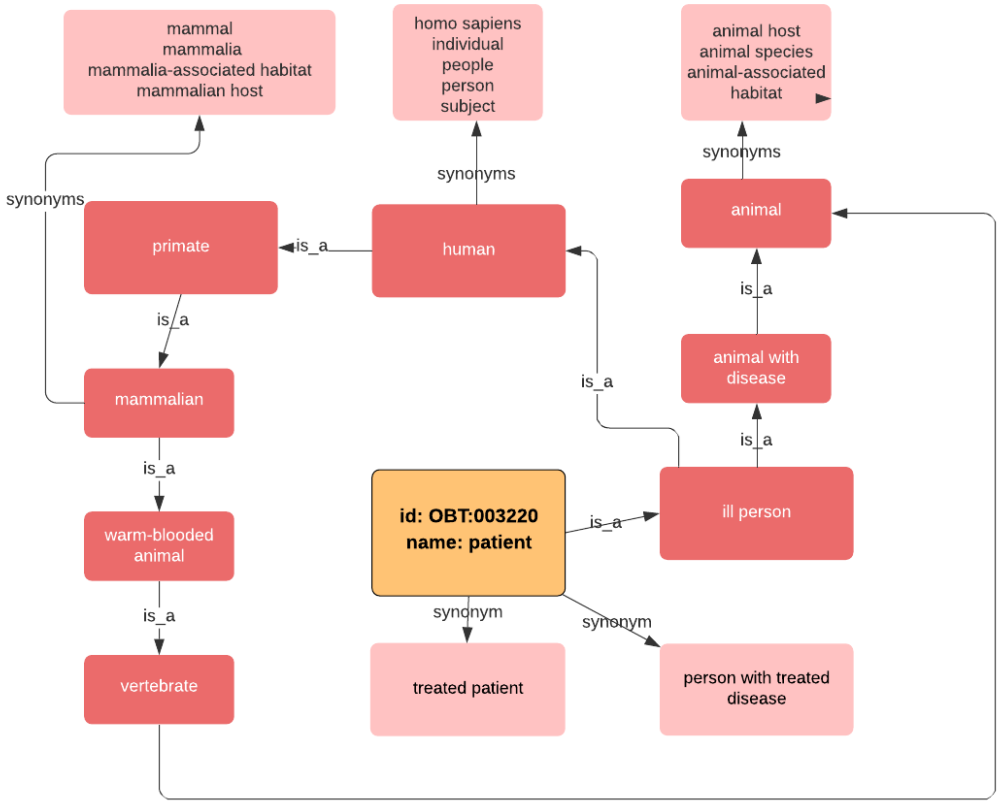
- There is an Ontology implementation in Python written by Martin Larralde
- github.com/althonos
- Given an .obo document
- Creates the graph representation of Ontology Terms by implementing the specifications of the Open Biomedical Ontologies 1.4

Ontology Library - Pronto



- Starts from the Patient (OBT:003220) term
- Whole graph can be induced by following is_a and synonym relationships
- Pronto library has superclasses and subclasses functions to return relevant Ontology terms

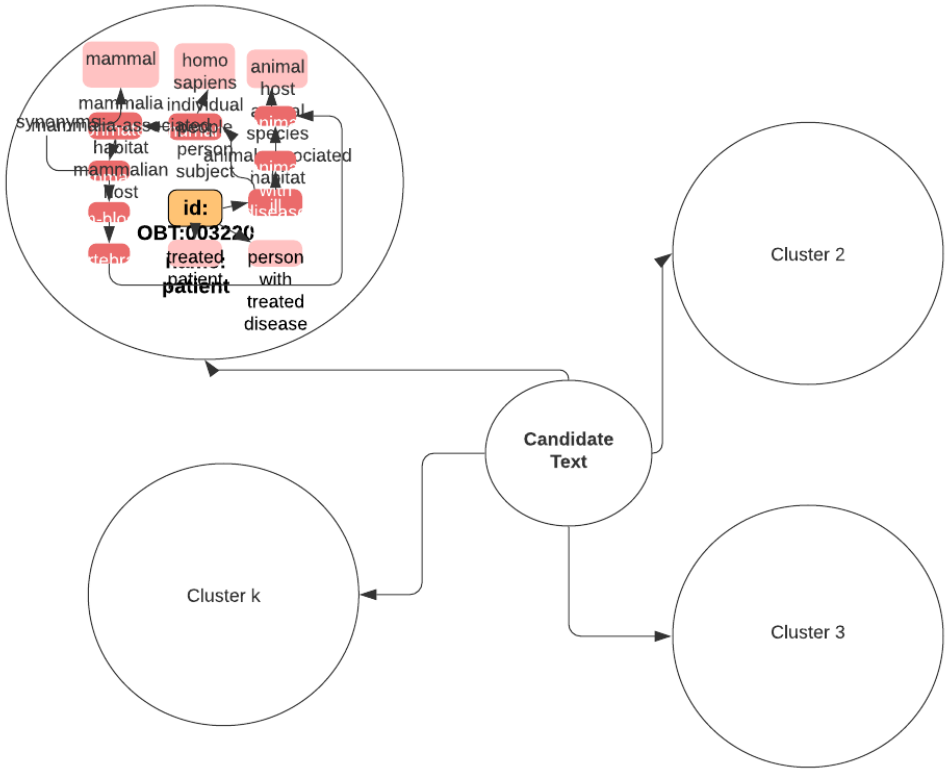
Ontology Library - Pronto



Superclasses of 'Patient'

'root for extraction'
'microbial habitat'
'living organism'
'eukaryote host'
'animal'
'animal with disease'
'vertebrate'
'warm-blooded animal'
'mammalian'
'primate'
'human'
'ill person'

Ontology Library - Pronto



- K-Means Clustering
- Nearest cluster can be selected for candidate text.

Future Work

- Bacteria Habitats Normalization Papers
- Biomedical Named Entity Normalization Papers
- 4 implementable ideas

Expanding Dictionaries

- Brenda Tissue Ontology (BTO)
  - 121,321 habitat synonyms [1]
- Prone to poor precision
- Stop-word lists
  - 2381 stop-words for bacteria
    - E.g. unclassified, scales, root

TF-IDF

- Baseline TF-IDF
  - Habitat names -> a document
    - Represented habitat names a TF-IDF weighted vector
    - Select habitat name -> highest cosine similarity
- Improving TF-IDF
  - BOW model -> character-level, n-grams [2]

CRF

- Conditional Random Field
  - used for tagging sequential data especially in Named Entity Recognition in NLP [3]

- 3 types of features
  - Lexical features
    - current word, its root, its POS tag etc.
  - Orthographic features
    - Substring features
      - first n-characters & the last-n characters
        - already mentioned in classes
    - Word form features
      - case-folding
      - normalizing numbers to ‘0’
  - Dictionary features
    - presence & position of the word in the dictionary

Hybrid Architecture

- Hybrid (rule-based & ML) structure

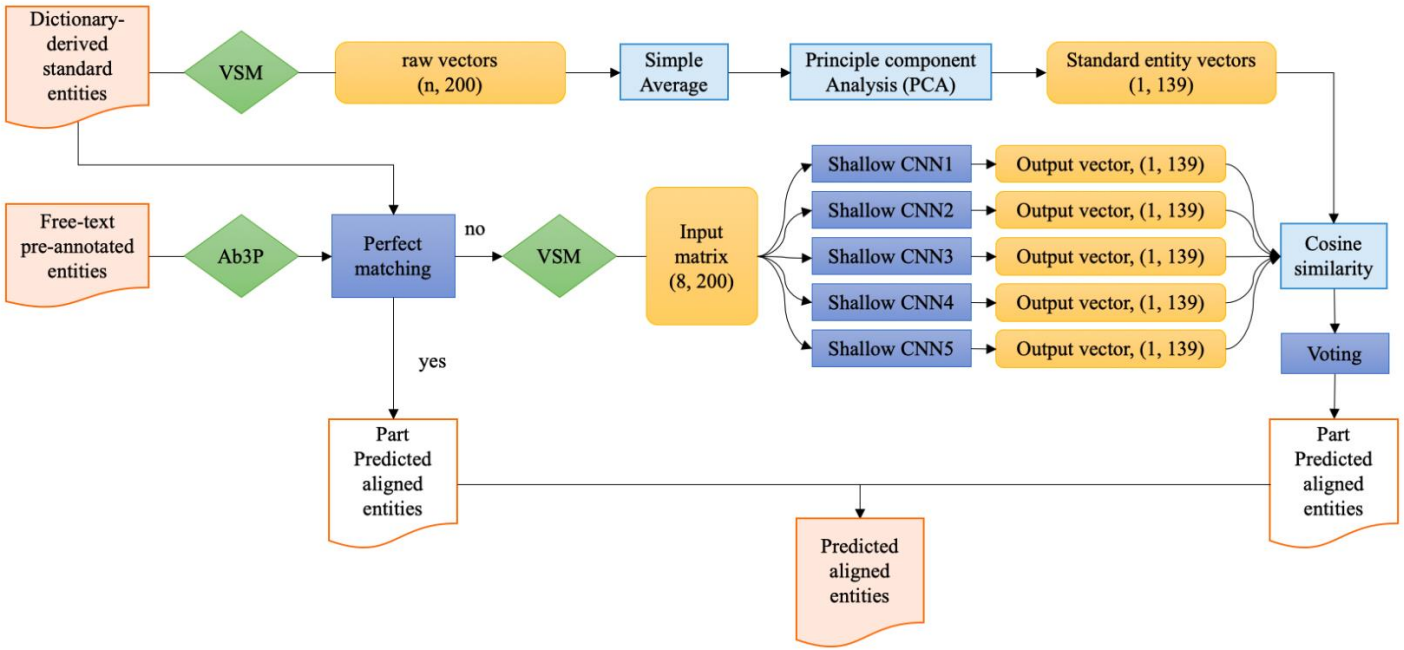


Figure 1: Model Architecture Overview [4]

Our References

- [1] A dictionary- and rule-based system for identification of bacteria and habitats in text
- [2] End-to-End System for Bacteria Habitat Extraction
- [3] Automatic extraction of microorganisms and their habitats from free text using text mining workflows
- [4] An ensemble CNN method for biomedical entity normalization

Thanks