Food-diseases relations extraction using spaCy

Wojciech Konieczkowicz, Marek Jachym, Kacper Duda, Norbert Pilarek

Agenda

1. **Rule-based** food and diseases **entities extraction**
2. **Rule-based** food-diseases **relations extraction**
3. **Snowball algorithm** for food-diseases **relations extraction**
4. **Tools, performance and architecture**
5. **Hands-on presentation**

Rule-based entities extraction - diseases

1. Distinctive “base” words

… *cancer* -> (lung | kidney | …) cancer

2. Distinctive latin suffixes (with exceptions - stop words)

…is -> (Acute Flaccid) Myelitis

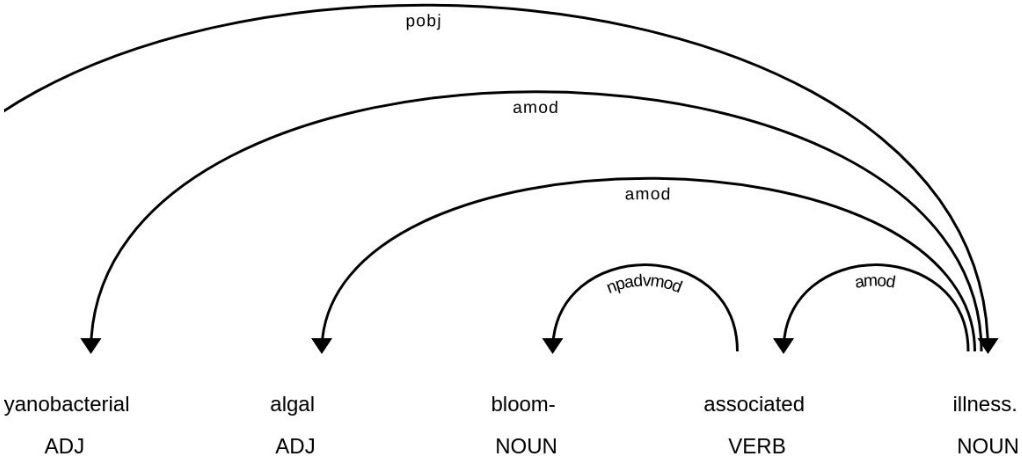
…is -> this

3. Initialisms (using a scraped list)

ADHD, COVID-19, …

Rule-based entities extraction - diseases

4. Grammar dependencies and their multiple combinations for bigger recall



Rule-based entities extraction - food

1. Distinctive words

   diet -> Western diet
   meal -> home cooked meal

2. Distinctive phrases

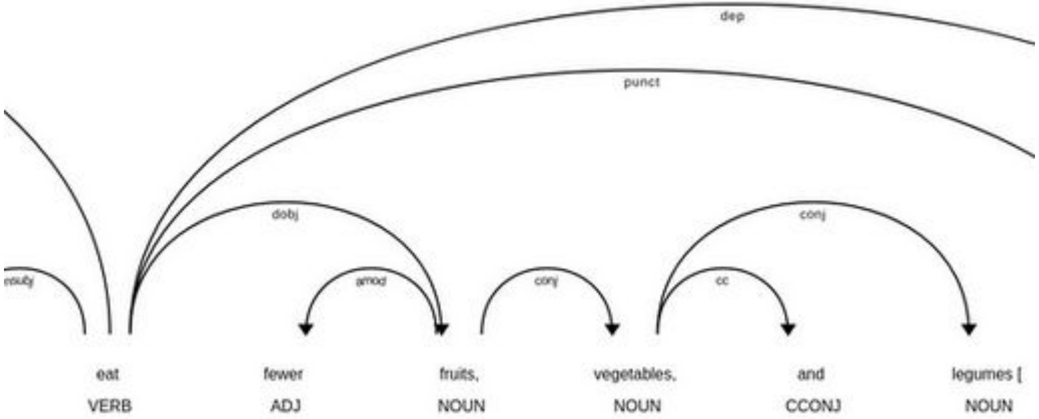
   consumption of **… -> consumption of fruit and vegetables**
   … intake -> fish intake
   eat … -> consume processed meals

3. Initialisms (using a scraped list)

   MDS, DASH, NFI

Rule-based **entities extraction** - food

4. Grammar dependencies and their multiple combinations for bigger recall



Rule-based relations extraction

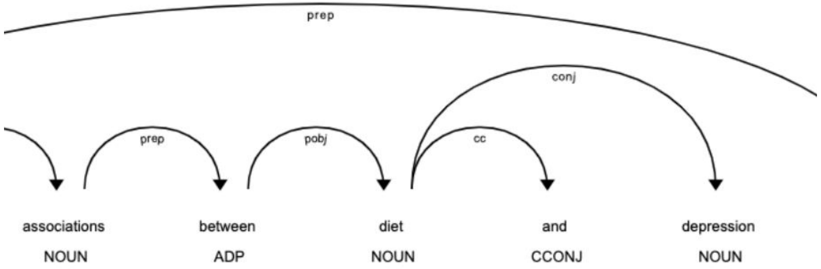
1. Uses food and diseases extractors

2. Uses distinctive words indicating relations

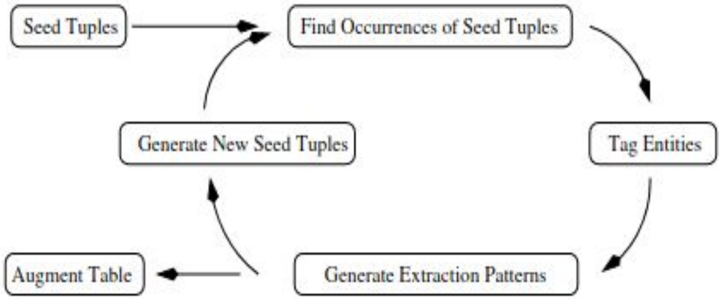
association -> associations between <FOOD> and <DIS>

cause -> <FOOD> consumption causes <DIS>

3. Looks for food and diseases entities being children of the same distinctive node within the same sentence



Snowball algorithm



main loop

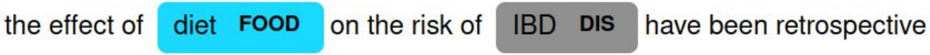
```
articles = load_articles(datapath)

sents = []
for art in articles:
  doc = extract_entities(art)
  sents += leave_food_dis_sents_only(doc)

seed_tuples = get_seed_tuples()
for i in range(n_iter):
  occur = find_ocurrences(seed_tuples)
  patterns = pattern_from_occurences(occur, w_size)
  patterns = single_pass_clustering(patterns, tau_cl)
  patterns = drop_insufficient(patterns, tau_supp)
  seed_tuples += get_new_tuples(patterns, tau_sim)
```

pseudocode
(CLI parameters in bold)

example of seed tuple occurrence with context (w_size = 3)



the effect of diet **FOOD** on the risk of IBD **DIS** have been retrospective

Tools & performance

Python 3.8.10
spaCy 3.0.6
aws
NumPy 1.20.2

tau_cl	tau_supp	tau_sim	iteration 1	iteration 2	iteration 3
3	3	3	42.4	-	-
3	3	3	45s	492.5s	2028.6s
3.5	4	3.5	46.2s	36.2s	-
3.25	3	3.25	50.7s	64.2s	83.7s

**testing setup: Intel Core i5-5200U CPU @ 2.20GHz, 8GB RAM* slidesgo