Abbreviation detection for biomedical articles

by Sonja Kenari

Agenda

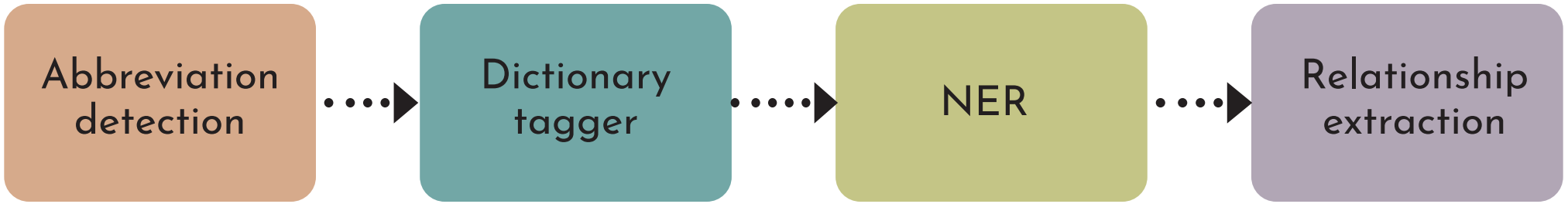
- Introduction
- Background
- Implementation
- Results
- Further Improvents

Introduction

Full project description

COVID-19 Open Research Dataset Challenge (CORD-19):

What do we know about vaccines and therapeutics?



Introduction

Abbreviation Detection

spaCy ⋯⋯→ Python library for NLP

Abbreviation detection

Makes it easier to:

- ⏳ Find articles of interest faster
- ? Keep up with the amount of new abbreviations

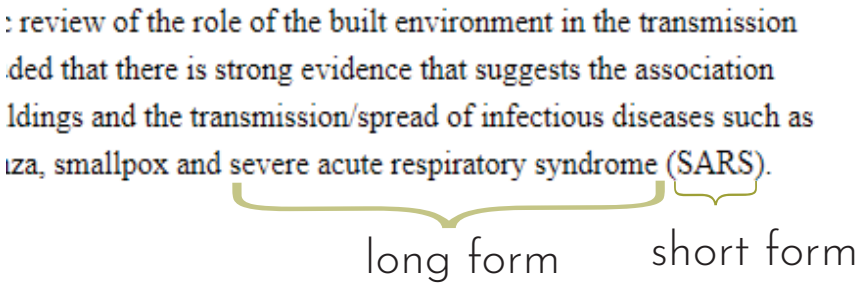
Background

Abbreviation Detection

scispaCy:

AbbreviationDetector

- Pre trained models by spaCy
- Detect: abbreviations & definitions



- Accuracy?

Implementation

Generate Pubannotations

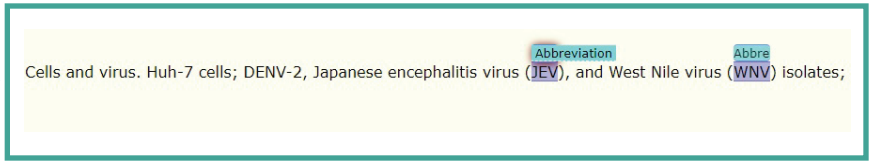
100 out of 60,000 articles

data subset [json]

metadata file [csv]

pubannotation [json]


```
{
    "cord_uid": "0nm7wgf5",
    "source_x": "PMC",
    "pmcid": "PMC5075077",
    "divid": "3",
    "text": "Cells and virus. Huh-7 cells; DENV-2, Japanese encephalitis virus (JEV), and West Nile virus (WNV) isolates;
    "project": "cdlai_CORD-19",
    "denotations": [
        {
            "id": "S-scispacy-abbr_T1",
            "span": {
                "begin": 67,
                "end": 70
            },
            "obj": "Abbreviation"
        },
        {
            "id": "S-scispacy-abbr_T2",
            "span": {
                "begin": 94,
                "end": 97
            },
            "obj": "Abbreviation"
        },
```

Implementation

Generating files of abbreviations

web scraping

metadata file [csv]

↓

url

↓

HTML parser

BeautifulSoup

↓

abbreviation, abbreviations, Abbreviation, Abbreviations

↓

csv files

scispaCy

data subset [json]

↓

full texts

↓

AbbreviationDetector

↓

csv files

output file format

Short form	Long form
CRP	C-reactive protein
FTD	Fast-track Diagnostics
IQR	interquartile range
OR	odds ratio
PCV	pneumococcal conjugate vaccination
PCT	procalcitonin
RSV	respiratory syncytial virus
RT-PCR	real-time multiplex polymerase chain reaction
WHO	World Health Organization

Implementation

Evaluation

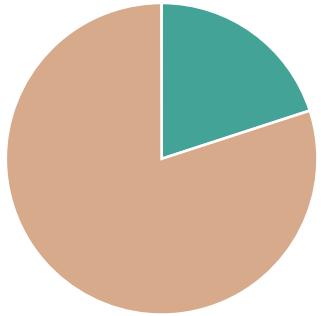
Compare the 2
- detected abbreviations with spaCy [csv]
- detected abbreviations with web scraping [csv]

$$\frac{\text{Number unique short forms detected by spaCy}}{\text{Number short forms detected by web scraping}} = (\%)$$

$$\frac{\text{Number unique long forms detected by spaCy}}{\text{Number long forms detected by web scraping}} = (\%)$$

Result

Abbreviation lists in



20 out of 100

short forms hit rate

Highest: 87.5%

Lowest: 25%

long forms hit rate

Highest: 52.6%

Lowest: 0%

notable faults

- spaCy weak on long form
- text from json files not updated after url articles
- faults in denotation extraction

Further Improvements

spaCy
Improve the results

Optimize programs
Make more time efficient

web scraper
Update data

Extract from Pubannotations
Instead of full text extraction

Thank you for listening!

Questions...?

Sonja Kenari
nat14sta@student.lu.se
2020-05-29