université PARIS-SACLAY

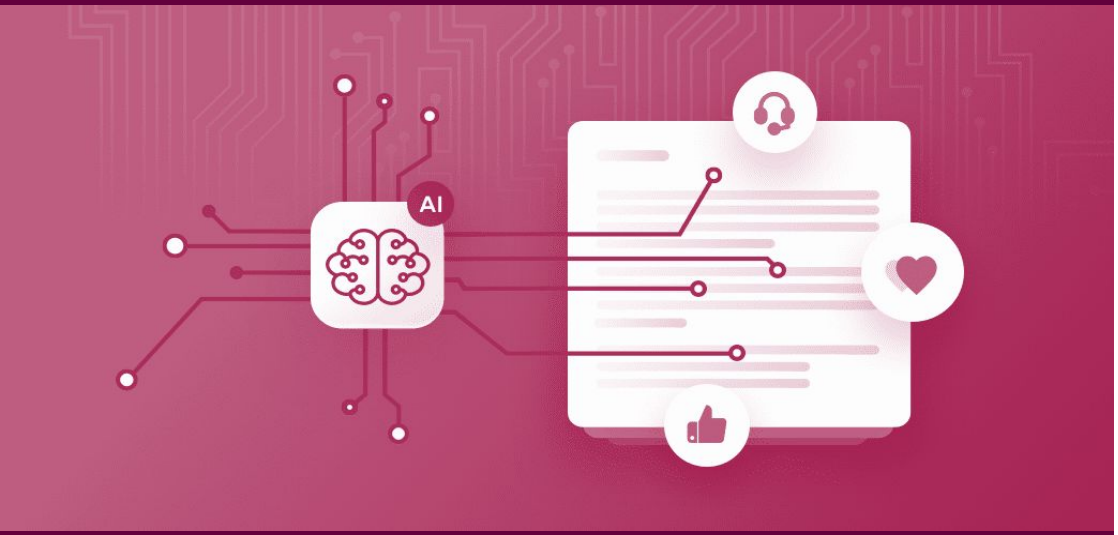
Class on
Information Retrieval
AI Master, first year

Instructor
Kim Gerdes

Information Retrieval Project

C12N - Microorganisms, enzymes and more...

Jacobo Ruiz Ocampo
Diego Andres Torres Guarin



TERM DETECTION

- Domain specific NER TT
- Pipeline
- Annotations
- Example and limits

université PARIS-SACLAY

Domain specific NER TT

- Domain:

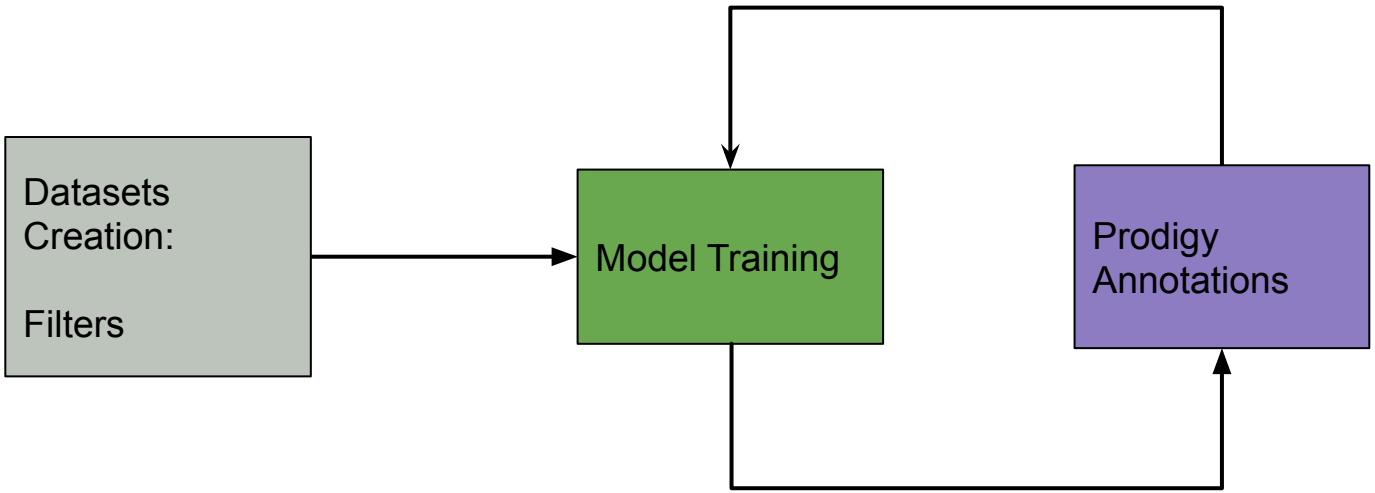
  MICROORGANISMS OR ENZYMES; COMPOSITIONS THEREOF; PROPAGATING, PRESERVING, OR MAINTAINING MICROORGANISMS; MUTATION OR GENETIC ENGINEERING; CULTURE MEDIA

- NER TT:

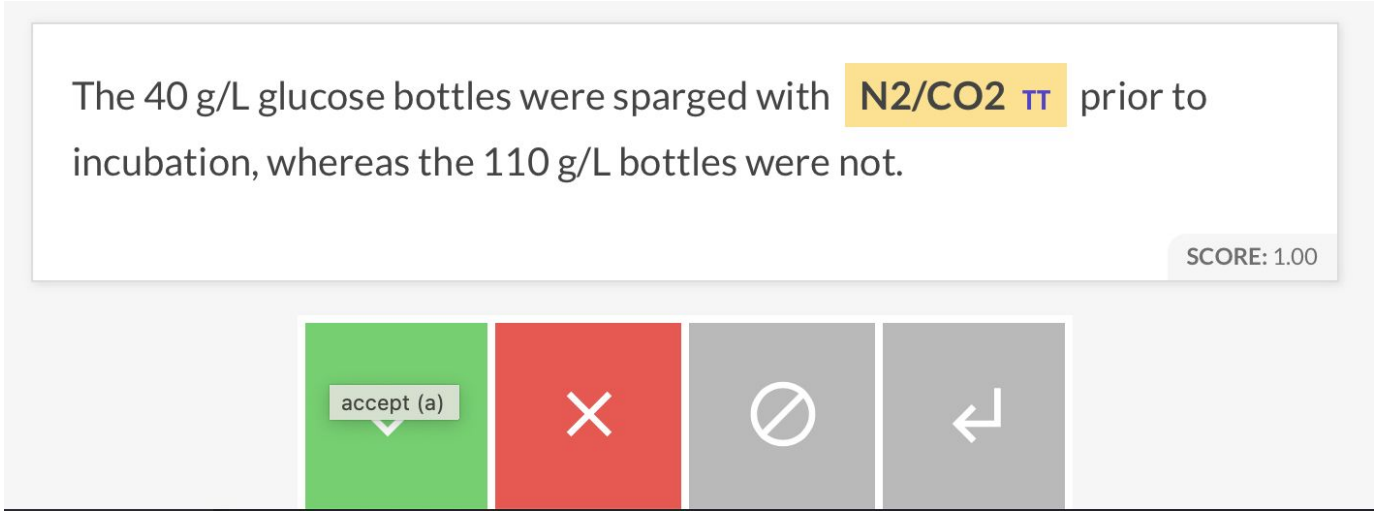
```python
# Load best model
nlp_ner = spacy.load("./results/spacy_output/model-best")

# Just text snippet
doc = nlp_ner(sample)

# Show NER results
spacy.displacy.render(doc, style="ent", jupyter=True)
```

Table TT 5. The retinal cell markers TT and dilutions TT used in the studyCell TypeCell MarkerDilutionsMüller cellAnti-CRALBP1:1000Anti-Vimentin1:100Anti-GFAP1:1000PhotoreceptorAnti-Opsin Red/Green1:250Anti-Opsin Blue1:250Neuron TT in INLAnti-PKCa1:200AstrocytesAnti-GFAP1:1000 TT

Pipeline


Datasets
Creation:
Filters
Model Training
Prodigy
Annotations

Annotating with prodigy

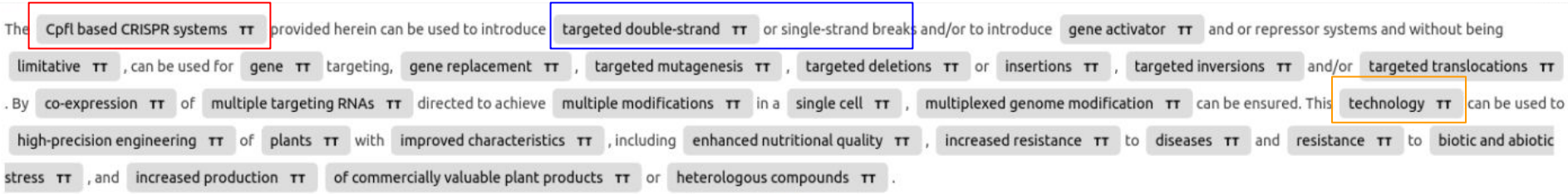



```
Correct annotations: 488
Incorrect annotations: 376
Correct percentage: 56.48%
Number of sentences annotated: 125
```

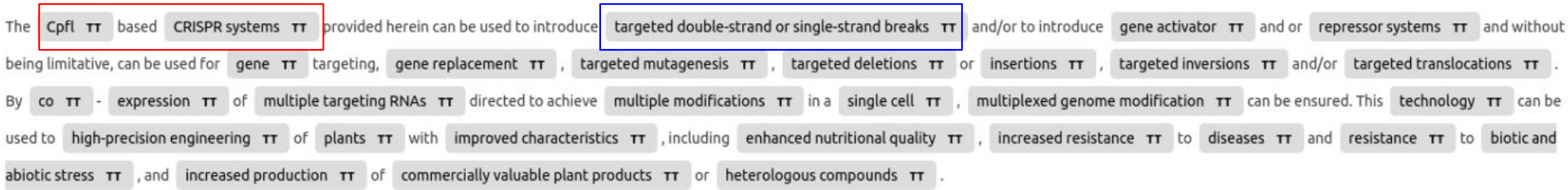

université PARIS-SACLAY

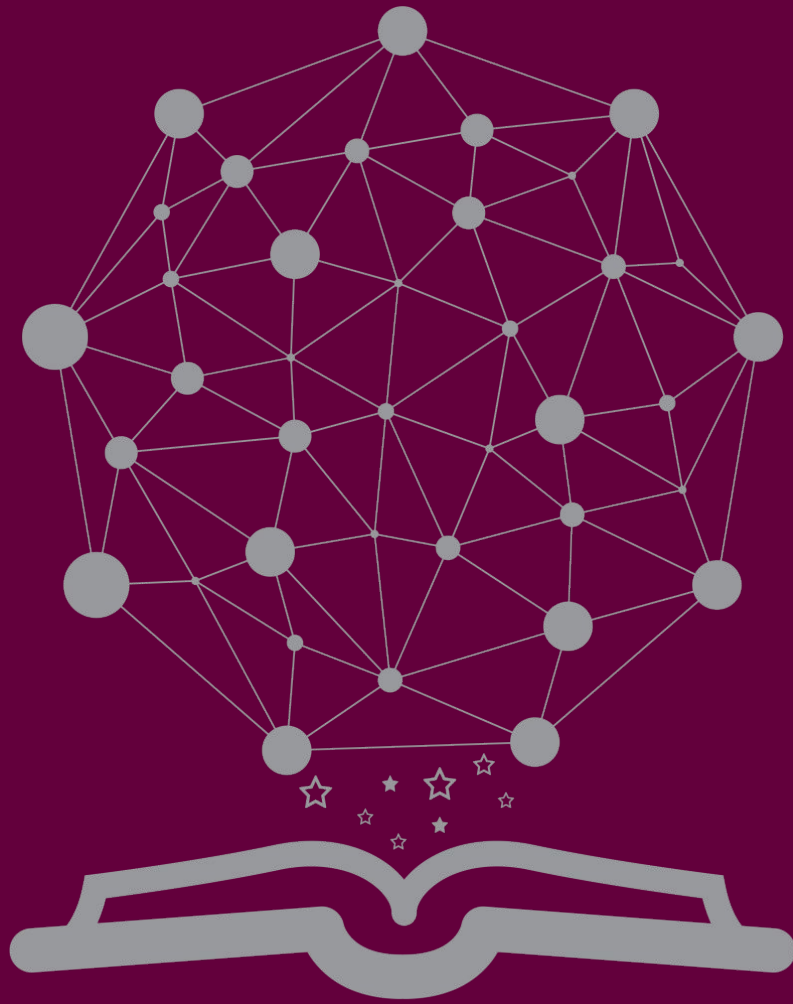
Examples

Fine tuned model on one iteration of annotations:



Model trained once:





RELATION DETECTION

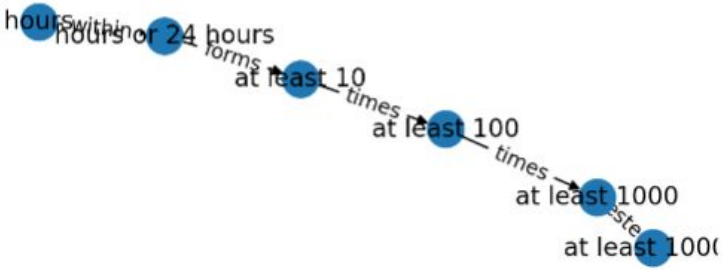
- Failed attempt
- Types of relations
- Implementation
- Example and limits
- Outlook

université PARIS-SACLAY

Failed attempt

Use NB5 with our entities = 

source	target	edge
hours	hours or 24 hours	within
hours or 24 hours	at least 10	forms
at least 10	at least 100	times
at least 100	at least 1000	times
at least 1000	at least 10000	ester
one	HLAME	produce



université PARIS-SACLAY

Types of Relations (0.94 per line)

Verb:

Subj + Verb + Obj

- Composition includes molecule
- Gene encodes protein
- Cells express antigen

To be:

Subj + To be + Attr

- Oxidation is a pretreatment
- Protein may be a polypeptide
- Carrier be a liquid

Preposition:

Subj + Prep + Pobj

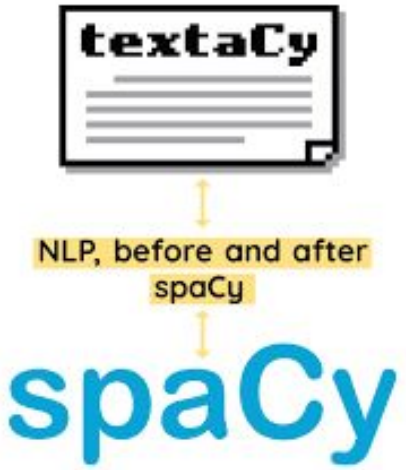
- Fragment consists of acid
- Cells used for screening
- HBeAg function as toleragen

Implementation

Part of Speech tagging and Dependency parsing

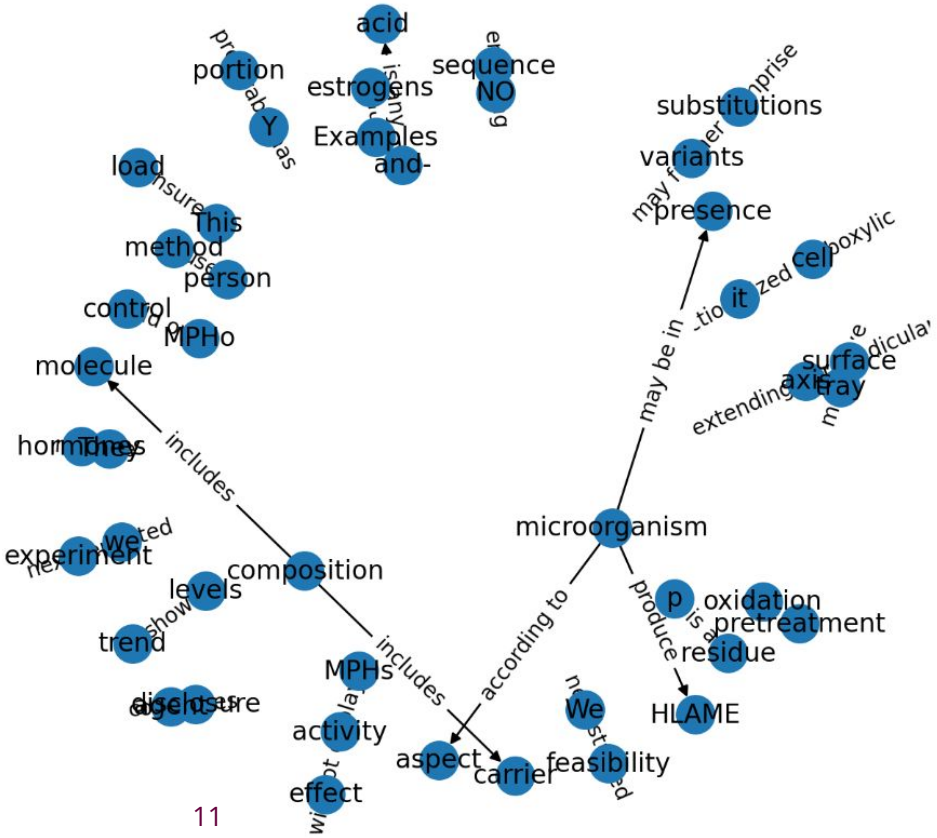
For each consecutive pair of entities:

1. *Check if they correspond to (subj, obj) (subj, attr) (subj, pobj)*
2. *Verify that they are in the same sentence*
3. *Check if the text in between has any (verb, aux, prep)*
4. *If match, clean the relation and add it*



https://realpython.com/natural-language-processing-spacy-python/#verb-phrase-detection

Example and limitations



Limitations:

- No evaluation metric
- Term extraction errors propagate (they, we)
- Some relations don’t make sense
- Only considering short term relations

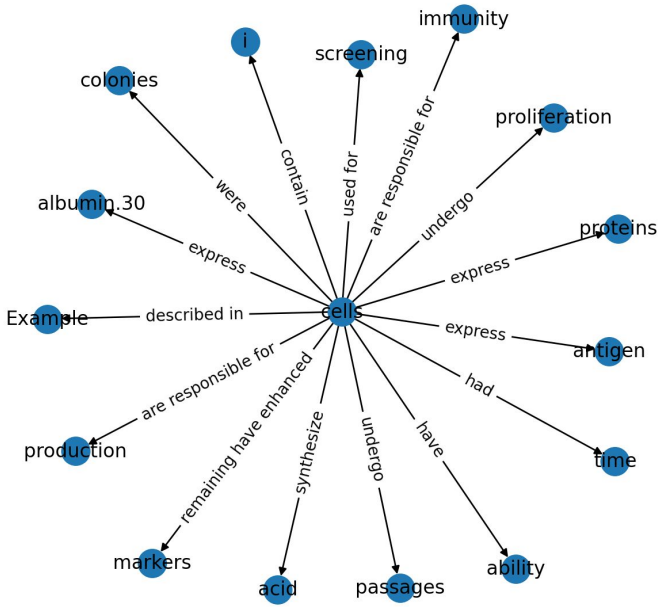
Beyond microorganisms

Term extractor (TE):
- Trained on specific data
- Will only generalize to similar text

Relation extractor (RE):
- No domain-specific rules
- Generalizes to other data if TE does

Search engine?:
1. Create the graph
2. Extract the entities in the query
3. Retrieve the subgraph containing entities
4. Maybe allow for n-n search

QUERY: Source=cells



université PARIS-SACLAY

THANK YOU!