Introduction to natural language processing for health and biological questions

Nona Naderi
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique

université PARIS-SACLAY

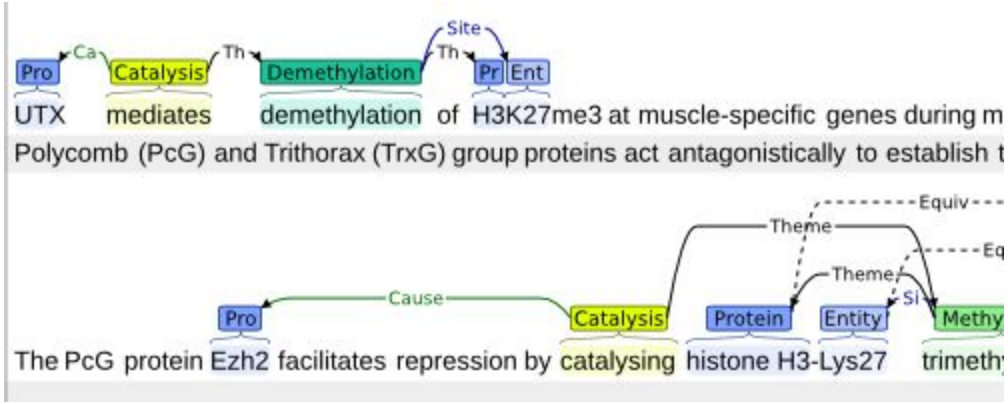
Getting the data

Preparing annotations is expensive and time consuming.

Domain-specific tasks require knowledge of domain.

The annotations need quality control: computing inter-annotator agreement.

Brat rapid annotation tool



Named Entity Recognition

Assigning concept types to the text and classifying the text into a set of predefined categories of interest.

Categories can be domain-specific or not.

For example, person, location, date/time, protein, organism, drugs,...

Challenging, why?

Example : metonomy

England won the world cup.

Other problems in NE

- Variations, abbreviations
- Ambiguity
- Punctuations, spellings,...

Why NER?

- Relation extraction
- Question Answering
- Information search

Simple lookup approach

Only recognizing the entities stored in lists (gazetteers)

Advantage: It is simple and fast

Disadvantage: collection, maintenance, ambiguity

Feature engineering

Designing a set of informative feature, for example POS tags, capitalization, numerals, Greek letters, lexical features, previous words...

Evaluation metrics

- Token accuracy: what percentage of tokens got correct labels. Problematic, why?

- Standard evaluation: per entity not token
- Precision, Recall, F-measure (strict, lenient)

```
File    O
Photo   O
)       O
```

IOB encoding

```
The     O
Defence B-group
Research        I-group
Development     I-group
Organisation    I-group
(       O
DRDO    B-group
)       O
is      O
working O
on      O
four    O
projects        O
to      O
develop O
new     O
technologies    O
```

Normalization

Assigning the extracted information a unique identifier from vocabularies, ontologies, or metathesaurus e.g., UMLS, Gene Ontology, …

Example datasets: MedMentions

27838742	726	738	RT-PCR assay	T062	UMLS:C1709846
27838742	758	771	delicaflavone	T103	UMLS:C1254351
27838742	780	801	autophagic cell death	T038	UMLS:C1326207
27838742	829	835	LC3-II	T103	UMLS:C3711208
27838742	839	844	LC3-I	T103	UMLS:C3714087
27838742	856	882	autophagy-related proteins	T103	UMLS:C4277731
27838742	916	943	acidic vesicular organelles	T017	UMLS:C0029219
27838742	948	961	autolysosomes	T017	UMLS:C0230822
27838742	969	978	cytoplasm	T017	UMLS:C0010834
27838742	988	999	lung cancer	T038	UMLS:C0684249
27838742	1000	1004	A549	T017	UMLS:C4277577
27838742	1009	1019	PC-9 cells	T017	UMLS:C0007634
27838742	1059	1072	Delicaflavone	T103	UMLS:C1254351
27838742	1073	1086	downregulated	T038	UMLS:C0013081
27838742	1091	1116	expression of phospho-Akt	T033	UMLS:C2697945
27838742	1118	1130	phospho-mTOR	T103	UMLS:C0033684
27838742	1136	1150	phospho-p70S6K	T103	UMLS:C0033684
27838742	1217	1226	autophagy	T038	UMLS:C0004391
27838742	1245	1248	Akt	T038	UMLS:C1515844
27838742	1251	1255	mTOR	T038	UMLS:C1515673
27838742	1258	1264	p70S6K	T103	UMLS:C0073337
27838742	1265	1272	pathway	T038	UMLS:C0037080
27838742	1276	1280	A549	T017	UMLS:C4277577
27838742	1285	1295	PC-9 cells	T017	UMLS:C0007634

Exercise

A simple bag-of-words model for NER:

https://github.com/juand-r/entity-recognition-datasets/blob/master/data/WNUT17/

Ethical considerations

Language involves humans:

Corpus collection: Privacy, potential harms, imbalanced

Corpus processing: Avoiding bias processing, reproducibility, generalizability

Corpus distribution: availability, confidentiality, copyright, detailed description, quality assessment

Project

Sentence classification of the PubMed 200k RCT dataset:

https://github.com/Franck-Dernoncourt/pubmed-rct

A baseline model with bag of words.

A model with pre-trained biomedical (word) embeddings.

A deep learning based model. (not needed)

Provide a report with the description of your method, evaluation, comparison of the models performance, results, discussion of your results, and error analysis.