BUNDESINSTITUT FÜR RISIKOBEWERTUNG

Bundesinstitut für Risikobewertung

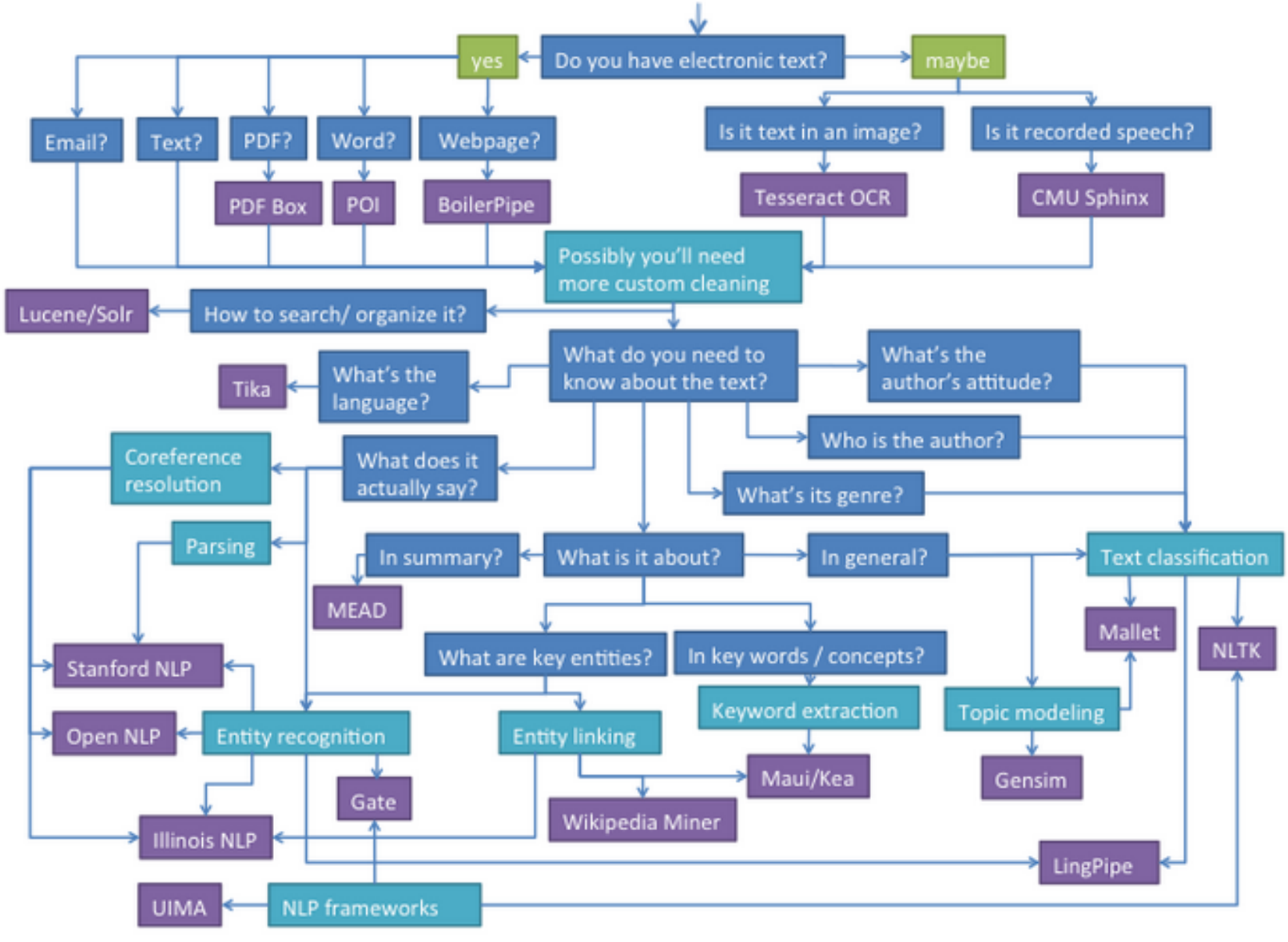
Semesterprojekt WS 17/18

Natural Language Processing

(Mariana Neves)

24. November 2017

Overview of NLP:



(http://entopix.com/so-you-need-to-understand-language-data-open-source-nlp-software-can-help.html)

Motivation:

- Biomedical NLP (BioNLP) and text mining:
  - Highly complex and large terminology (hundreds of thousands of genes, chemicals, cells, diseases, etc.)
  - Large collections of documents: publications, clinical trials, news, social media, clinical reports, etc.
  - Domain knowledge: databases, ontologies, terminologies
  - Needs to obtain fast answers, involving many entities in short time
  - Support for literature curation

Motivation:

- Applications:
  - Document classification
  - Named-entity extraction
  - Information extraction: slot filling, relationships, events
  - Question answering
  - Document summarization

Document classification:

- Automatically classification of a document into one (or more) categories

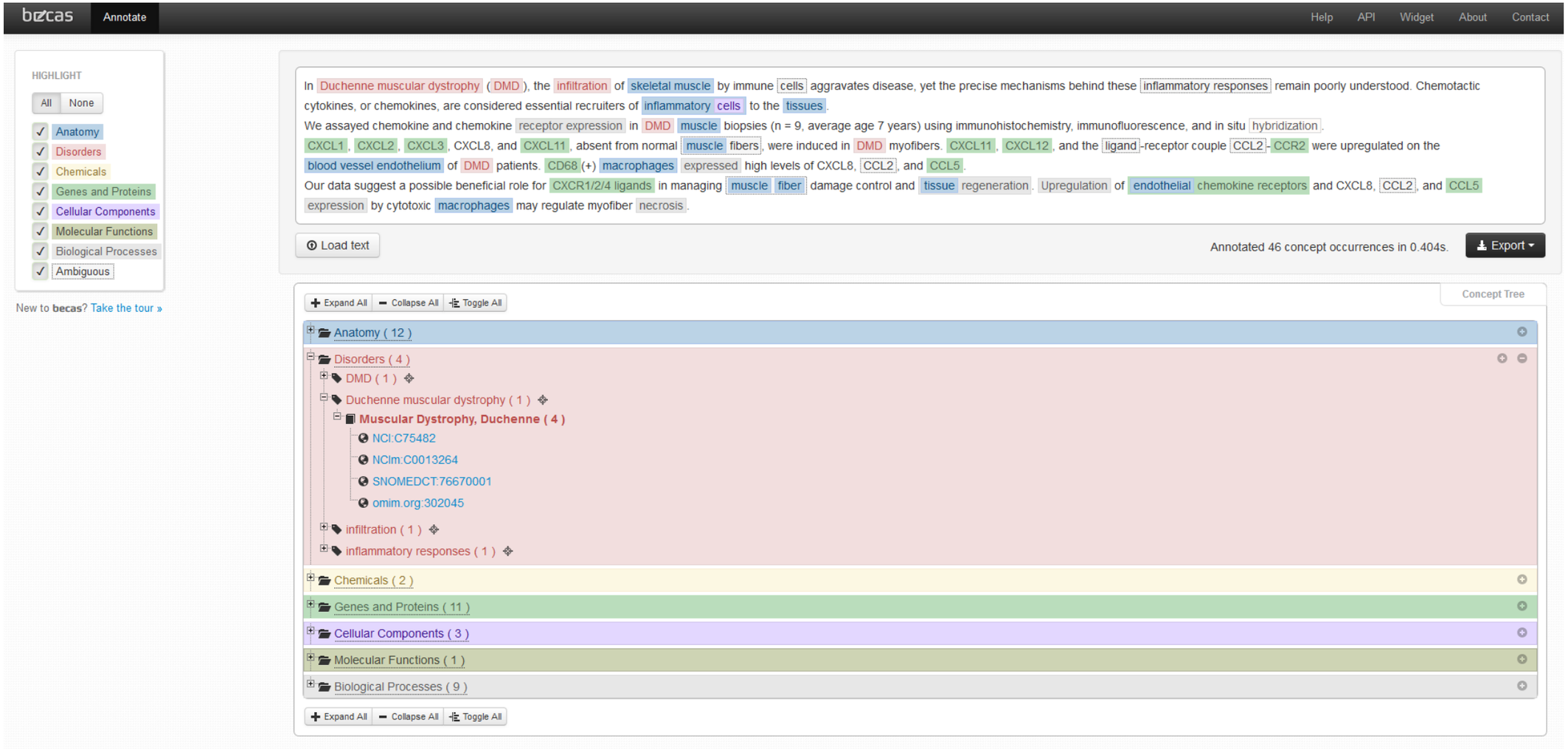
Document classification:

- Automatically classification of a document into one (or more) categories

- Binary, triage
  - Filtering of relevant documents
  - Ex.: animal experiments or not

- Multi-class, multi-label
  - Assign class and label
  - Ex.: i2b2 Obesity challenge: 15 diseases (classes) and 4 labels (present, absent, questionable, unmentioned)
  - Ex.: MeSH Terms
  - Ex.: 3R labels (replacement, reduction, refinement)

Named-entity recognition (NER):



(Source: http://bioinformatics.ua.pt/software/becas/)

Named-entity recognition (NER):

- Localization in text, e.g., “Duchenne muscular dystrophy”
- Document level, e.g., “Muscular Dystrophy, Duchenne”

- Classification of semantic type, e.g., “Disorders” (becas), “T047 - Disease or Syndrome” (UMLS Semantic Types)

- Normalization, e.g., “NCI:C75482”, “NCIm:C0013264”, “SNOMEDCT:76670001”, “omim.org:302045”

UMLS Semantic Types:

ENTITY

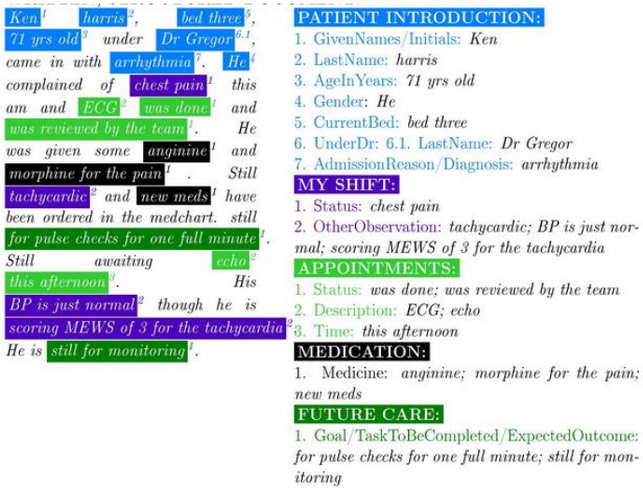
- Physical Object
  - Organism
    - Plant
    - Fungus
    - Virus
    - Bacterium
    - Archaeon
    - Eukaryote
      - Animal
        - Vertebrate
          - Amphibian
          - Bird
          - Fish
          - Reptile
          - Mammal
            - Human
  - Anatomical Structure
    - Embryonic Structure
    - Anatomical Abnormality
      - Congenital Abnormality
      - Acquired Abnormality
    - Fully Formed Anatomical Structure
      - Body Part, Organ, or Organ Component
      - Tissue
      - Cell
      - Cell Component
      - Gene or Genome
  - Manufactured Object
    - Medical Device
      - Drug Delivery Device
    - Research Device
    - Clinical Drug
  - Substance
    - Chemical
      - Chemical Viewed Functionally
        - Pharmacologic Substance
          - Antibiotic
        - Biomedical or Dental Material
        - Biologically Active Substance
          - Hormone
          - Enzyme
          - Vitamin
          - Immunologic Factor
          - Receptor
        - Indicator, Reagent, or Diagnostic Acid
        - Hazardous or Poisonous Substance

- [Physical Object] (continued)
  - [Substance] (continued)
    - [Chemical] (continued)
      - Chemical Viewed Structurally
        - Organic Chemical
          - Nucleic Acid, Nucleoside, or Nucleotide
          - Amino Acid, Peptide, or Protein
        - Inorganic Chemical
        - Element, Ion, or Isotope
    - Body Substance
    - Food
- Conceptual Entity
  - Idea or Concept
    - Temporal Concept
    - Qualitative Concept
    - Quantitative Concept
    - Functional Concept
      - Body System
    - Spatial Concept
      - Body Space or Junction
      - Body Location or Region
      - Molecular Sequence
        - Nucleotide Sequence
        - Amino Acid Sequence
        - Carbohydrate Sequence
      - Geographic Area
  - Finding
    - Laboratory or Test Result
    - Sign or Symptom
  - Organism Attribute
    - Clinical Attribute
  - Intellectual Product
    - Classification
    - Regulation or Law
  - Language
  - Occupation or Discipline
    - Biomedical Occupation or Discipline
  - Organization
    - Health Care Related Organization
    - Professional Society
    - Self-help or Relief Organization
  - Group Attribute
  - Group
    - Professional or Occupational Group
    - Population Group
    - Family Group
    - Age Group
    - Patient or Disabled Group

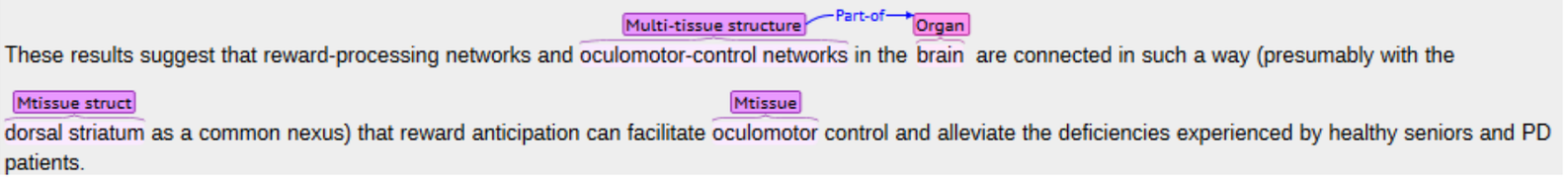
(Source: https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html)
(Download: https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml)

Information extraction (IE):

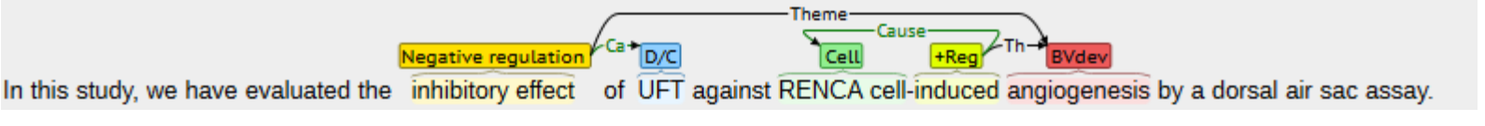
- NER as pre-processing step
- Slot filling



- Binary relationships

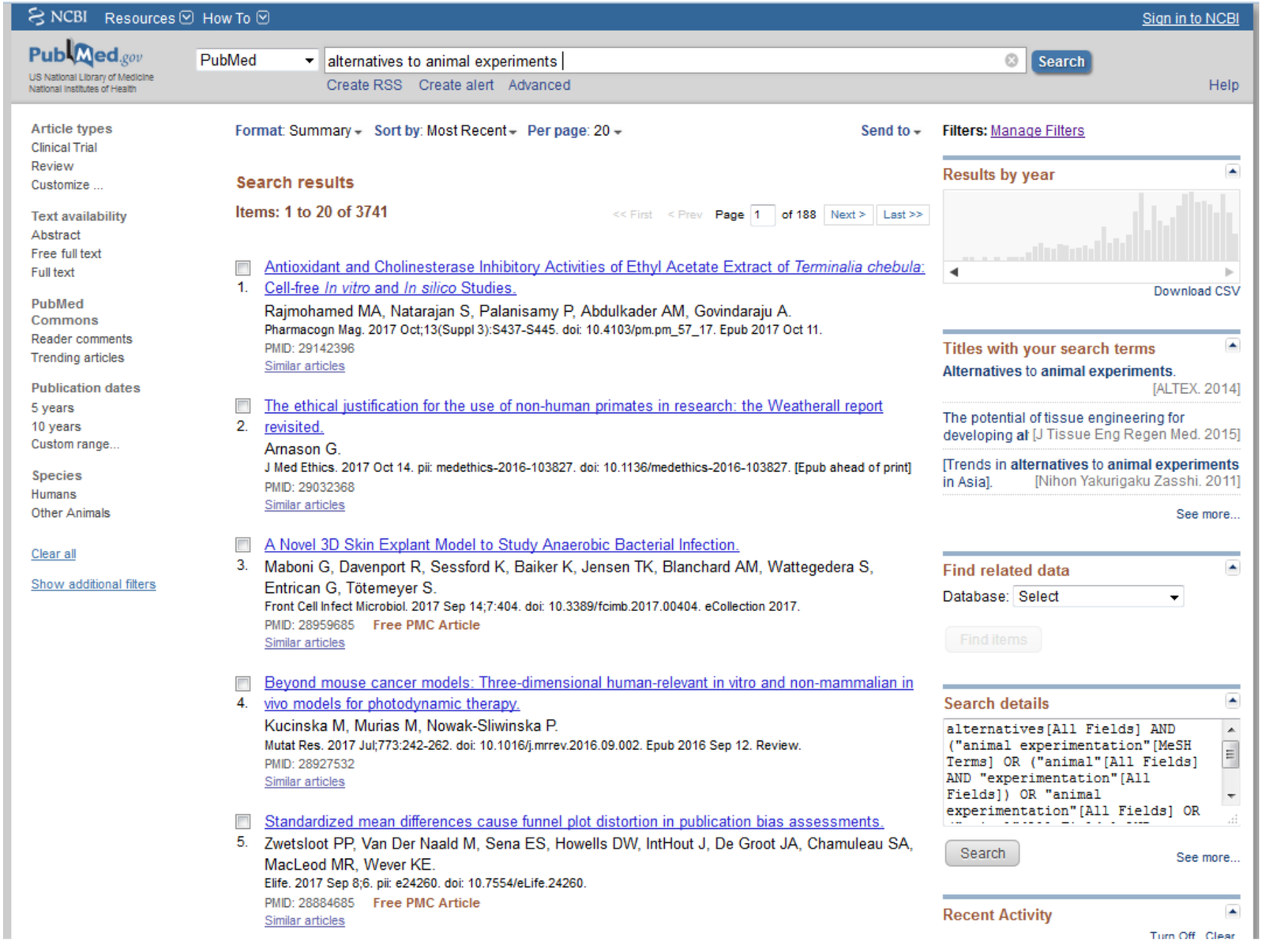


- Event extraction



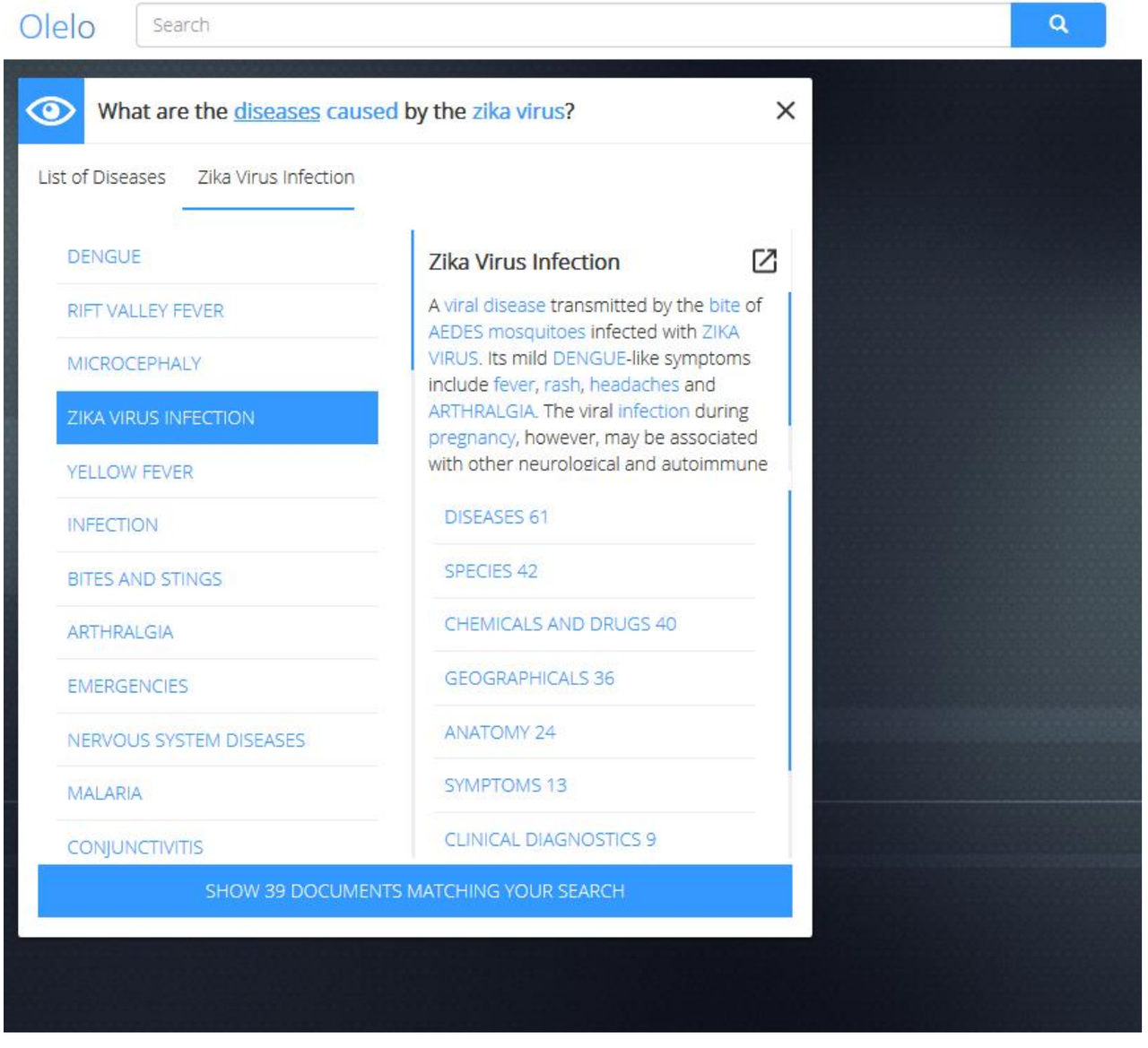
(Sources: https://medinform.jmir.org/2015/2/e19/
http://weaver.nlplab.org/~brat/demo/latest/#/not-editable/AnEM-1.0.4/PMC-2972690-sec-19
http://www.nactem.ac.uk/eccb2012/index.xhtml#/10473104)

Information retrieval (IR):



(Source: https://www.ncbi.nlm.nih.gov/pubmed/)

Question answering (QA):



(Source: http://hpi.de/plattner/olelo/)

Question answering (QA) Tool:

OAQA Biomedical Question Answering (BioASQ) System

The OAQA Biomedical Question Answering (BioASQ) System aims to identify relevant documents, concepts and passages (snippets) and automatically generate exact answer texts to arbitrary biomedical questions (factoid, list, yes/no). It won the best-performing system in the BioASQ QA Challenges in the factoid and list categories two years in a row in 2015 and 2016 (see official results).

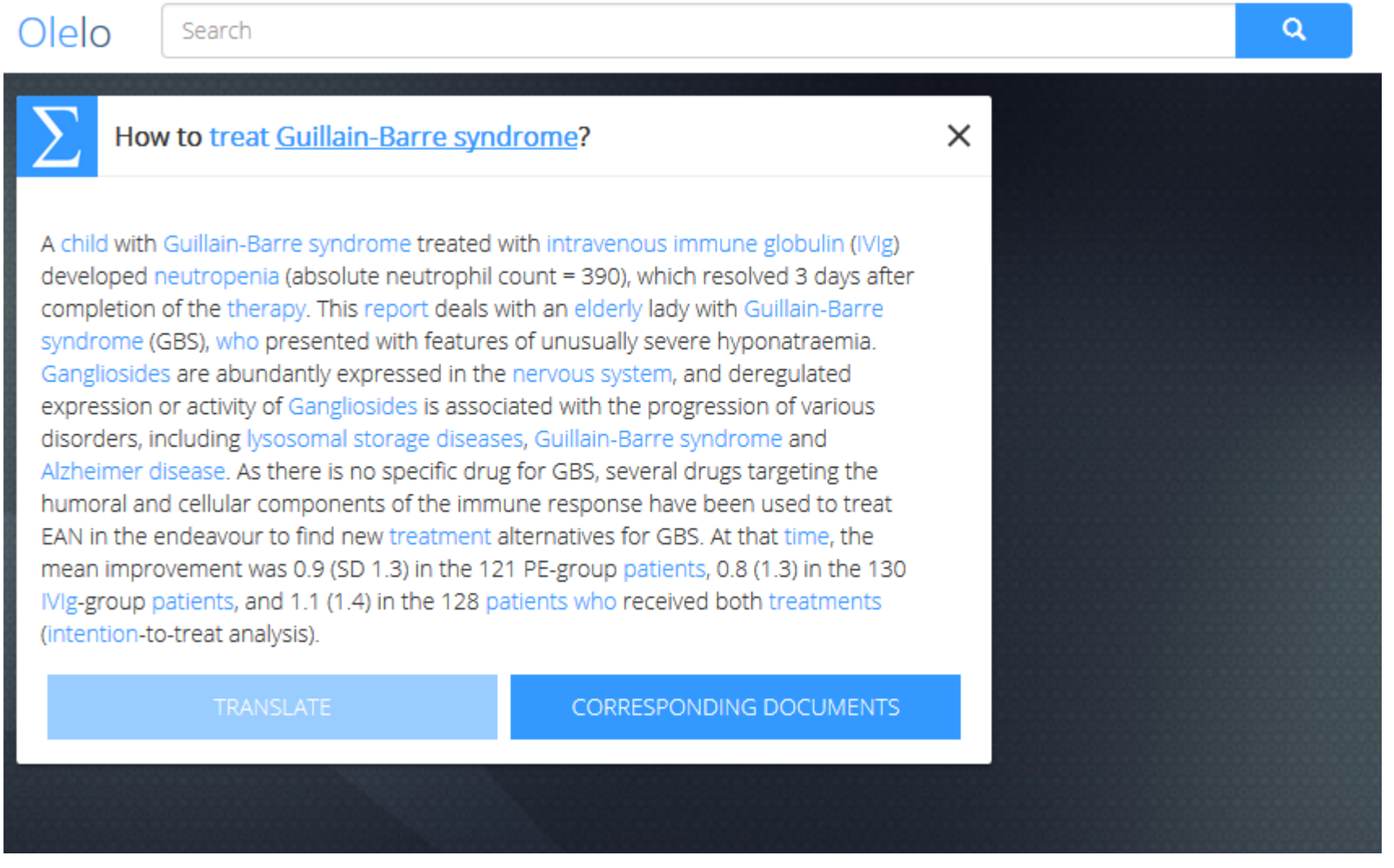
System description papers have the most details about the design and implementation of the architecture and the algorithms:

- Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, and Eric Nyberg. Learning to Answer Biomedical Factoid & List Questions: OAQA at BioASQ 3B. In Proceedings of *CLEF 2015 Evaluation Labs and Workshop, 2015*. [pdf]
- Zi Yang, Yue Zhou, and Eric Nyberg. Learning to Answer Biomedical Questions: OAQA at BioASQ 4B. In Proceedings of *Workshop on Biomedical Natural Language Processing, 2016*. [pdf]

Please contact Zi Yang if you have any questions or comments.

(Source: https://github.com/oaqa/bioasq)

Document summarization:



(Source: http://hpi.de/plattner/olelo/)

Document summarization Tools:

MEAD

MEAD is the most elaborate publicly available platform for multi-lingual summarization and evaluation.The platform implements multiple summarization algorithms such as position-based, centroid-based, largest common subsequence, and keywords. The methods for evaluating the quality of the summaries are both intrinsic and extrinsic. MEAD implements a battery of summarization algorithms, including baselines (lead-based and random) as well as centroid-based and query-based methods.

Download

- MEAD 3.12
- MEAD 3.11
- MEAD 3.10
- MEAD 3.09
- MEAD 3.07

Automatic text summarizer

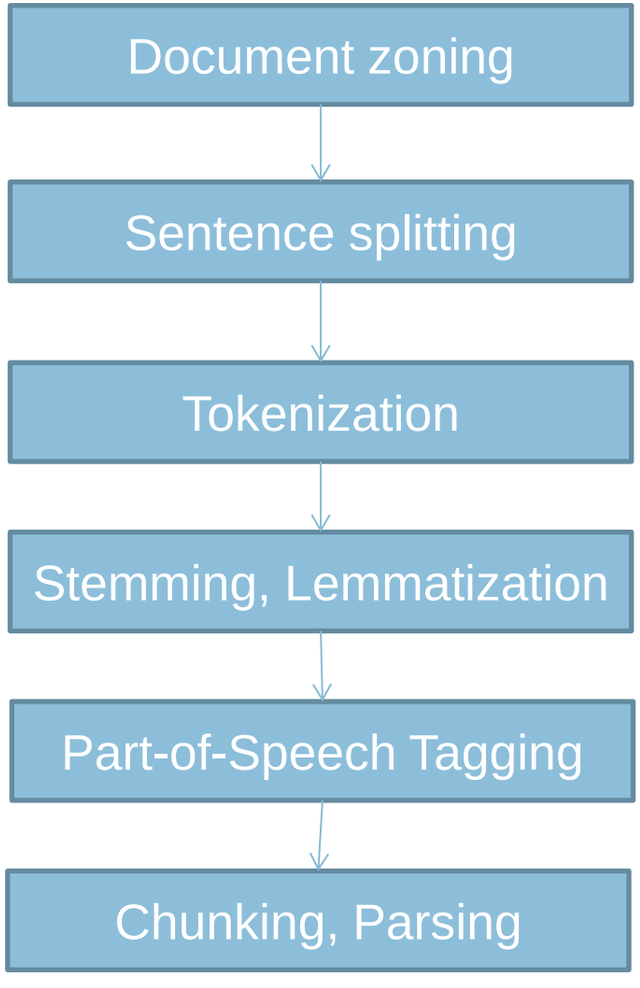
build passing

Simple library and command line utility for extracting summary from HTML pages or plain texts. The package also contains simple evaluation framework for text summaries. Implemented summarization methods:

- **Luhn** - heurestic method, reference
- **Edmundson** heurestic method with previous statistic research, reference
- **Latent Semantic Analysis, LSA** - one of the algorithm from http://scholar.google.com /citations?user=0fTuW_YAAAAJ&hl=en I think the author is using more advanced algorithms now. Steinberger, J. a Ježek, K. Using latent semantic an and summary evaluation. In In Proceedings ISIM '04. 2004. S. 93-100.
- **LexRank** - Unsupervised approach inspired by algorithms PageRank and HITS, reference
- **TextRank** - some sort of combination of a few resources that I found on the internet. I really don't remember the sources. Probably Wikipedia and some papers in 1st page of Google :)
- **SumBasic** - Method that is often used as a baseline in the literature. Source: Read about SumBasic
- **KL-Sum** - Method that greedily adds sentences to a summary so long as it decreases the KL Divergence. Source: Read about KL-Sum

(Sources: https://github.com/miso-belica/sumy http://www.summarization.com/mead/)

Linguistic pre-processing - workflow:



(Sources: http://hpi.de/plattner/olelo/)

Document zoning:

- Scientific literature:
  - Zones: abstract, introduction, methods, results, discussion, conclusions, captions, supplemental materials, references, etc.
  - Medline Structured Abstracts

Sentence splitting:

- Separate the sentences
- Not every dot is a separator (“Mr.”, “1.5”, “etc.”)!

(Sources: https://opennlp.apache.org/
http://www.nltk.org/
https://stanfordnlp.github.io/CoreNLP/
https://spacy.io/)

Tokenization:

- Separate the tokens (words)
- Not every only spaces are separators (also slash, hyphens, etc.)
- There are tools for the biomedical domain: Genia Tagger

OPEN NLP™ Stanford CoreNLP NLTK 3.2.5 spaCy

```
> echo "Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin."

Inhibition      Inhibition      NN      B-NP    O
of              of              IN      B-PP    O
NF-kappaB       NF-kappaB       NN      B-NP    B-protein
activation      activation      NN      I-NP    O
reversed        reverse         VBD     B-VP    O
the             the             DT      B-NP    O
anti-apoptotic  anti-apoptotic  JJ      I-NP    O
effect          effect          NN      I-NP    O
of              of              IN      B-PP    O
isochamaejasmin isochamaejasmin NN      B-NP    O
.               .               .       O       O
```

GENIA

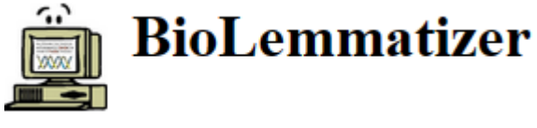
(Sources: https://opennlp.apache.org/ http://www.nltk.org/ https://stanfordnlp.github.io/CoreNLP/ https://spacy.io/ http://www.nactem.ac.uk/GENIA/tagger/)

Stemming, Lemmatization:

- Convert tokens to their stem or lemma:
  - Stem (based on rules): moving -> to move
  - Lemma (based on thesaurus): better -> good

- Classic stemmer: Porter stemmer (implementations for various programming languages)

- There are tools for the biomedical domain: BioLemmatizer



(Sources: https://opennlp.apache.org/
http://www.nltk.org/
https://stanfordnlp.github.io/CoreNLP/
https://spacy.io/
http://biolemmatizer.sourceforge.net/)

Part-of-Speech Tagging:

- Assign a part-of-speech (role) to every token, e.g., verb, noun, adjective, etc.

- There are tools for the biomedical domain: Genia Tagger

OpenNLP™ Stanford CoreNLP NLTK 3.2.5 spaCy

GENIA

```
> echo "Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin."

Inhibition      Inhibition      NN      B-NP    O
of              of              IN      B-PP    O
NF-kappaB       NF-kappaB       NN      B-NP    B-protein
activation      activation      NN      I-NP    O
reversed        reverse         VBD     B-VP    O
the             the             DT      B-NP    O
anti-apoptotic  anti-apoptotic  JJ      I-NP    O
effect          effect          NN      I-NP    O
of              of              IN      B-PP    O
isochamaejasmin isochamaejasmin NN      B-NP    O
.               .               .       O       O
```

(Sources: https://opennlp.apache.org/
http://www.nltk.org/
https://stanfordnlp.github.io/CoreNLP/
https://spacy.io/
http://www.nactem.ac.uk/GENIA/tagger/)

Chunking:

- Shallow parsing of sentence (sentence structural analysis) into “chunks”
- There are tools for the biomedical domain: Genia Tagger

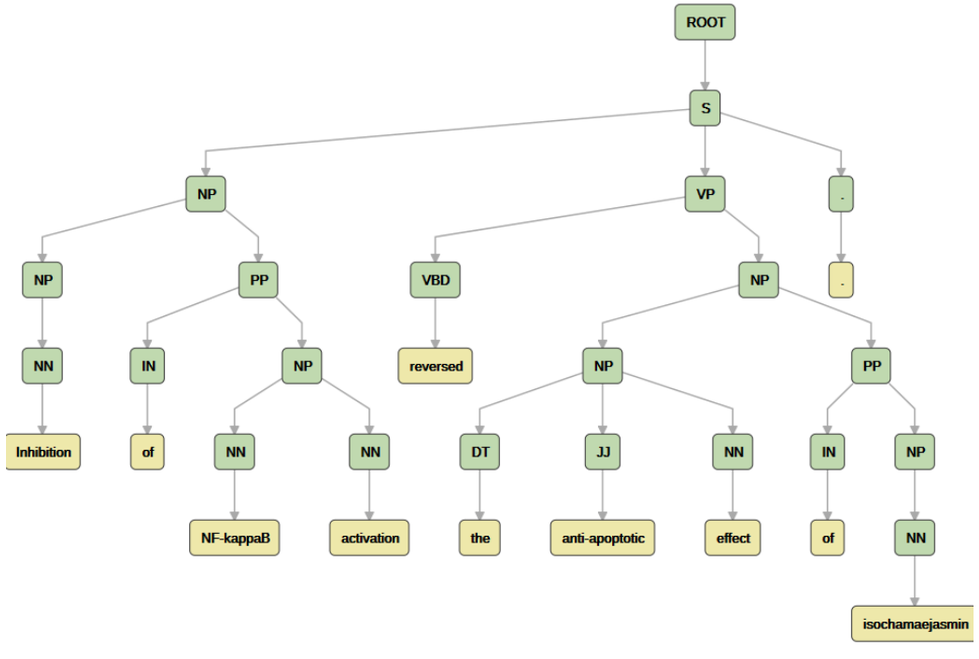
```
> echo "Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin."

Inhibition      Inhibition      NN      B-NP    O
of              of              IN      B-PP    O
NF-kappaB       NF-kappaB       NN      B-NP    B-protein
activation      activation      NN      I-NP    O
reversed        reverse         VBD     B-VP    O
the             the             DT      B-NP    O
anti-apoptotic  anti-apoptotic  JJ      I-NP    O
effect          effect          NN      I-NP    O
of              of              IN      B-PP    O
isochamaejasmin isochamaejasmin NN      B-NP    O
.               .               .       O       O
```

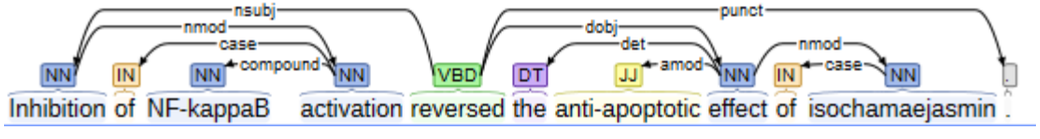
(Sources: https://opennlp.apache.org/ http://www.nltk.org/ https://stanfordnlp.github.io/CoreNLP/ https://spacy.io/ http://www.nactem.ac.uk/GENIA/tagger/)

Parsing:

- Full parsing of sentence (sentence structural analysis) into a parse tree
  - Dependency tree (cf. below, right)
  - Constituency tree (cf. below, left)

OpenNLP Stanford CoreNLP NLTK 3.2.5 spaCy

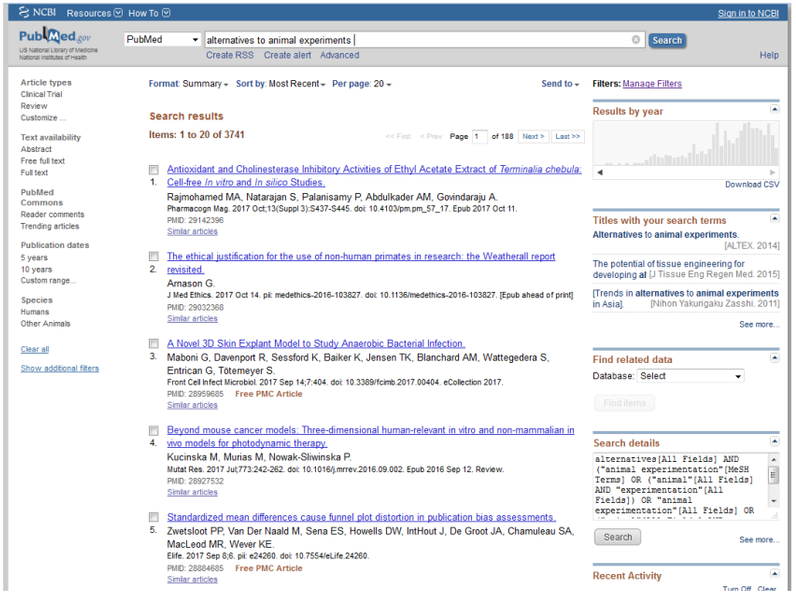




(Sources: https://opennlp.apache.org/
http://www.nltk.org/
https://stanfordnlp.github.io/CoreNLP/
https://spacy.io/
http://corenlp.run/)

Biomedical resources:

- Documents



SciELO
Scientific Electronic Library Online

bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

(Sources: https://www.ncbi.nlm.nih.gov/pubmed/
http://scielo.org
https://www.biorxiv.org/)

Biomedical resources:

- Databases

(Sources: http://www.uniprot.org/
http://www.genome.jp/kegg/
http://flybase.org/
http://www.informatics.jax.org/
http://rgd.mcw.edu/)

Biomedical resources:

- Terminologies and Ontologies: UMLS, MeSH terms, BioPortal

UMLS Metathesaurus Vocabulary Documentation

Choose a source by browsing one of the presentation tabs below

Alphabetical List | Restriction Categories* | Languages

Alphabetical List

Expand All Collapse All

A	5 sources
Source	**Last Updated**
AIR (AI/RHEUM)	1995AA
ALT (Alternative Billing Concepts)	2009AA
AOD (Alcohol and Other Drug Thesaurus)	2002AC
AOT (Authorized Osteopathic Thesaurus)	2006AD
ATC (Anatomical Therapeutic Chemical (ATC)-classification system)	2017AA
B	1 source
C	13 sources
D	6 sources
F	1 source
G	2 sources

BioPortal

BioPortal Statistics

Ontologies	662
Classes	8,169,345
Resources Indexed	48
Indexed Records	39,537,360
Direct Annotations	95,468,433,792
Direct Plus Expanded Annotations	144,789,582,932

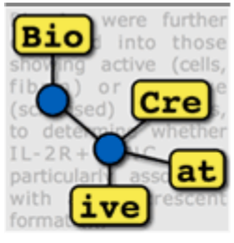
(Sources: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html https://bioportal.bioontology.org/)

Biomedical resources:

- Corpora:
  - Manually annotated collections of documents
  - Gold and Silver-standard
  - Training supervised learning methods
  - Support for system development
  - Evaluation and comparison of methods

Biomedical resources:

- Challenges:
  - Events organized by the BioNLP community to compare and/or boost performance of tools for a particular task



BioNLP-ST 2013

Text REtrieval Conference (TREC)

BioASQ

i2b2
Informatics for Integrating Biology & the Bedside

(Sources: http://www.biocreative.org/
http://2016.bionlp-st.org/
http://trec.nist.gov/
http://bioasq.org/
https://www.i2b2.org/NLP/DataSets/)

Biomedical resources:

- Journals

Bioinformatics

BMC Bioinformatics

DATABASE The Journal of Biological Databases and Curation

JAMIA

JOURNAL OF BIOMEDICAL SEMANTICS

Journal of Biomedical Informatics

Nucleic Acids Research

PLOS ONE

Biomedical resources:

- Conferences

BioNLP 2013

BUNDESINSTITUT
FÜR RISIKOBEWERTUNG

Bundesinstitut für Risikobewertung

DANKE FÜR IHRE AUFMERKSAMKEIT

Bundesinstitut für Risikobewertung

Max-Dohrn-Str. 8-10 ● 10589 Berlin

Tel. 0 30 - 184 12 - 0 ● Fax 0 30 - 184 12 - 47 41

bfr@bfr.bund.de ● www.bfr.bund.de