

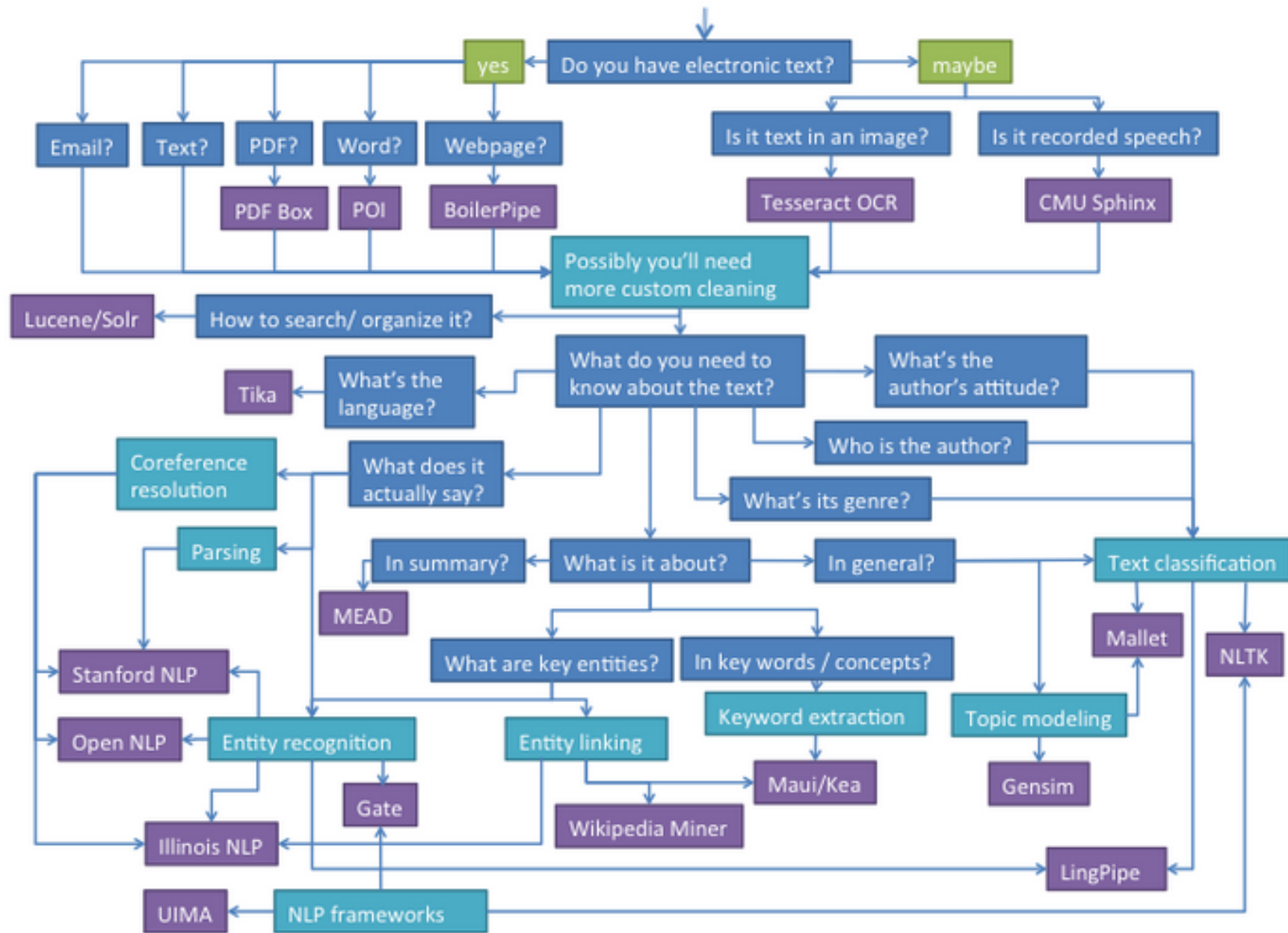
Semesterprojekt WS 17/18

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):

The screenshot displays the becas NER interface. At the top, the 'becas' logo and 'Annotate' button are visible. The right side of the header contains links for 'Help', 'API', 'Widget', 'About', and 'Contact'. On the left, a 'HIGHLIGHT' panel lists categories: Anatomy, Disorders, Chemicals, Genes and Proteins, Cellular Components, Molecular Functions, Biological Processes, and Ambiguous, each with a checked box. Below this is a link: 'New to becas? Take the tour >'. The main text area contains a paragraph about Duchenne muscular dystrophy (DMD) with entities highlighted in colored boxes. Below the text are 'Load text' and 'Export' buttons. The 'Export' button shows 'Annotated 46 concept occurrences in 0.404s.' Below the text is a 'Concept Tree' panel with expand/collapse controls. The tree shows a hierarchy: Anatomy (12) > Disorders (4) > DMD (1) > Duchenne muscular dystrophy (1) > Muscular Dystrophy, Duchenne (4) > NCI:C75482, NCim:C0013264, SNOMEDCT:76670001, omim.org:302045. Other categories like infiltration and inflammatory responses are also listed.

(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

Ken¹ Harris², bed three³, 71 yrs old⁴ under Dr Gregor^{6,1}, came in with arrhythmia⁷. He⁴ complained of chest pain¹ this am and ECG² was done¹ and was reviewed by the team¹. He was given some anginine¹ and morphine for the pain¹. Still tachycardic² and new meds¹ have been ordered in the medchart. still for pulse checks for one full minute¹. Still awaiting echo² this afternoon³. His BP is just normal² though he is scoring MEWS of 3 for the tachycardia². He is still for monitoring¹.

PATIENT INTRODUCTION:

1. GivenNames/Initials: Ken
2. LastName: harris
3. AgeInYears: 71 yrs old
4. Gender: He
5. CurrentBed: bed three
6. UnderDr: 6.1. LastName: Dr Gregor
7. AdmissionReason/Diagnosis: arrhythmia

MY SHIFT:

1. Status: chest pain
2. OtherObservation: tachycardic; BP is just normal; scoring MEWS of 3 for the tachycardia

APPOINTMENTS:

1. Status: was done; was reviewed by the team
2. Description: ECG; echo
3. Time: this afternoon

MEDICATION:

1. Medicine: anginine; morphine for the pain; new meds

FUTURE CARE:

1. Goal/TaskToBeCompleted/ExpectedOutcome: for pulse checks for one full minute; still for monitoring

These results suggest that reward-processing networks and oculomotor-control networks in the brain are connected in such a way (presumably with the dorsal striatum as a common nexus) that reward anticipation can facilitate oculomotor control and alleviate the deficiencies experienced by healthy seniors and PD patients.

Multi-tissue structure → Part-of → Organ
 Mttissue struct → Mttissue

- Event extraction

In this study, we have evaluated the inhibitory effect of UFT against RENCA cell-induced angiogenesis by a dorsal air sac assay.

Negative regulation → Ca → D/C → Theme → Cause → +Reg → Th → BVdev

(Sources: <https://medinform.jmir.org/2015/2/e19/>
<http://weaver.niplab.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):

The screenshot displays the PubMed search results page for the query "alternatives to animal experiments". The interface includes a search bar at the top with the query entered and a "Search" button. Below the search bar, there are options for "Format: Summary", "Sort by: Most Recent", and "Per page: 20". The search results are listed in a numbered format, with each entry including a checkbox, a title link, authors, journal information, and PMID. The results are: 1. "Antioxidant and Cholinesterase Inhibitory Activities of Ethyl Acetate Extract of Terminalia chebula: Cell-free In vitro and In silico Studies." by Rajmohamed MA, Natarajan S, Palanisamy P, Abdulkader AM, Govindaraju A. 2. "The ethical justification for the use of non-human primates in research: the Weatherall report revisited." by Amason G. 3. "A Novel 3D Skin Explant Model to Study Anaerobic Bacterial Infection." by Maboni G, Davenport R, Sessford K, Baiker K, Jensen TK, Blanchard AM, Wattegedera S, Entrican G, Töttemeyer S. 4. "Beyond mouse cancer models: Three-dimensional human-relevant in vitro and non-mammalian in vivo models for photodynamic therapy." by Kucinska M, Murias M, Nowak-Sliwinska P. 5. "Standardized mean differences cause funnel plot distortion in publication bias assessments." by Zwetsloot PP, Van Der Naald M, Sena ES, Howells DW, Int'Hout J, De Groot JA, Chamuleau SA, MacLeod MR, Wever KE. On the right side of the page, there are sections for "Results by year" (a bar chart), "Titles with your search terms" (listing "Alternatives to animal experiments" and "The potential of tissue engineering for developing at"), "Find related data" (with a database selection dropdown), "Search details" (showing the search query: "alternatives[All Fields] AND ('animal experimentation'[MeSH Terms] OR ('animal'[All Fields] AND 'experimentation'[All Fields]) OR 'animal experimentation'[All Fields] OR ..."), and "Recent Activity".

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

Question answering (QA):

The screenshot shows the Olelo search interface. At the top, there is a search bar with the text 'Search' and a magnifying glass icon. Below the search bar, a modal window is open with the title 'What are the diseases caused by the zika virus?'. The modal contains a list of diseases on the left and a detailed view of 'Zika Virus Infection' on the right. The list of diseases includes: DENGUE, RIFT VALLEY FEVER, MICROCEPHALY, ZIKA VIRUS INFECTION (highlighted in blue), YELLOW FEVER, INFECTION, BITES AND STINGS, ARTHRALGIA, EMERGENCIES, NERVOUS SYSTEM DISEASES, MALARIA, and CONJUNCTIVITIS. The detailed view for 'Zika Virus Infection' includes a description: 'A viral disease transmitted by the bite of Aedes mosquitoes infected with ZIKA VIRUS. Its mild DENGUE-like symptoms include fever, rash, headaches and ARTHRALGIA. The viral infection during pregnancy, however, may be associated with other neurological and autoimmune'. Below the description, there are several categories with counts: DISEASES 61, SPECIES 42, CHEMICALS AND DRUGS 40, GEOGRAPHICALS 36, ANATOMY 24, SYMPTOMS 13, and CLINICAL DIAGNOSTICS 9. At the bottom of the modal, there is a blue button that says 'SHOW 39 DOCUMENTS MATCHING YOUR SEARCH'.

(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:

Olelo 

 **How to treat [Guillain-Barre syndrome?](#)** 

A child with [Guillain-Barre syndrome](#) treated with [intravenous immune globulin \(IVIg\)](#) developed [neutropenia](#) (absolute neutrophil count = 390), which resolved 3 days after completion of the [therapy](#). This [report](#) deals with an [elderly lady](#) with [Guillain-Barre syndrome](#) (GBS), [who](#) presented with features of unusually severe hyponatraemia. [Gangliosides](#) are abundantly expressed in the [nervous system](#), and deregulated expression or activity of [Gangliosides](#) is associated with the progression of various disorders, including [lysosomal storage diseases](#), [Guillain-Barre syndrome](#) and [Alzheimer disease](#). As there is no specific drug for GBS, several drugs targeting the humoral and cellular components of the immune response have been used to treat EAN in the endeavour to find new [treatment](#) alternatives for GBS. At that [time](#), the mean improvement was 0.9 (SD 1.3) in the 121 PE-group [patients](#), 0.8 (1.3) in the 130 [IVIg-group patients](#), and 1.1 (1.4) in the 128 [patients who](#) received both [treatments](#) ([intention-to-treat analysis](#)).

[TRANSLATE](#) [CORRESPONDING DOCUMENTS](#)

(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

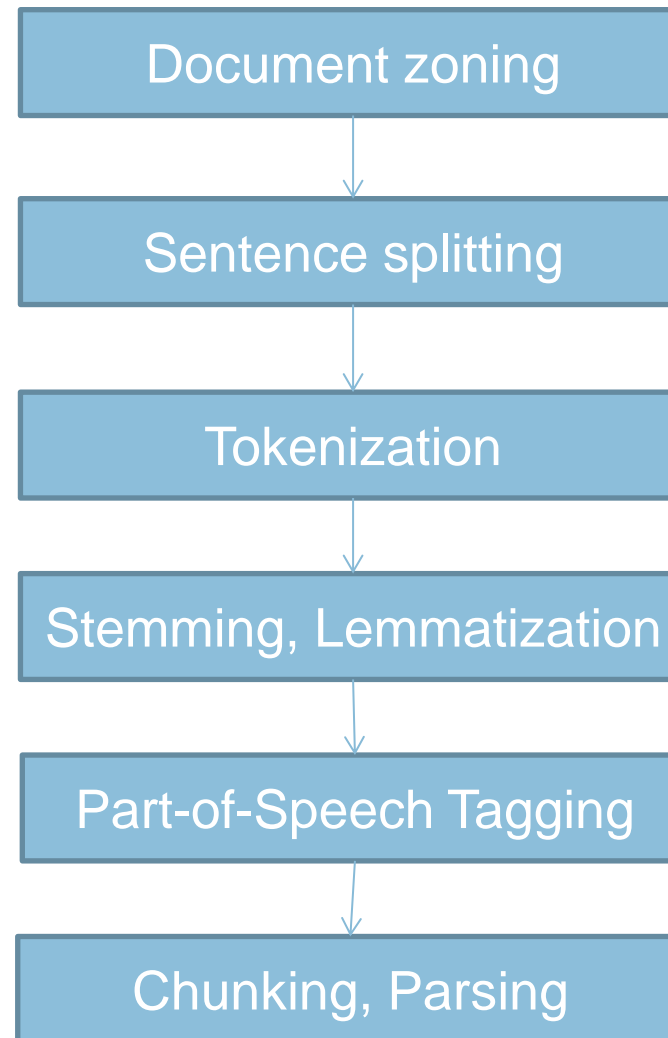
build passing

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 - heuristic method, [reference](#)
- Edmundson heuristic 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.”)!



NLTK 3.2.5

Stanford CoreNLP

spaCy

(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



Stanford CoreNLP

NLTK 3.2.5



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



Inhibition	Inhibition	NN	B-NP	0	
of	of	IN	B-PP	0	
NF-kappaB	NF-kappaB	NN	B-NP	0	B-protein
activation	activation	NN	I-NP	0	
reversed	reverse	VBD	B-VP	0	
the	the	DT	B-NP	0	
anti-apoptotic	anti-apoptotic	JJ	I-NP	0	
effect	effect	NN	I-NP	0	
of	of	IN	B-PP	0	
isochamaejasmin	isochamaejasmin	NN	B-NP	0	
.	.	.	0	0	

(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



Stanford CoreNLP

NLTK 3.2.5

spaCy



BioLemmatizer

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



Stanford CoreNLP

NLTK 3.2.5



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

Inhibition	Inhibition	NN	B-NP	0
of	of	IN	B-PP	0
NF-kappaB	NF-kappaB	NN	B-NP	B-protein
activation	activation	NN	I-NP	0
reversed	reverse	VBD	B-VP	0
the	the	DT	B-NP	0
anti-apoptotic	anti-apoptotic	JJ	I-NP	0
effect	effect	NN	I-NP	0
of	of	IN	B-PP	0
isochamaejasmin	isochamaejasmin	NN	B-NP	0
.	.		0	0



(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



Stanford CoreNLP

NLTK 3.2.5



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



Inhibition	Inhibition	NN	B-NP	0	
of	of	IN	B-PP	0	
NF-kappaB	NF-kappaB	NN	B-NP	B-protein	
activation	activation	NN	I-NP	0	
reversed	reverse	VBD	B-VP	0	
the	the	DT	B-NP	0	
anti-apoptotic	anti-apoptotic	JJ	I-NP	0	
effect	effect	NN	I-NP	0	
of	of	IN	B-PP	0	
isochamaejasmin	isochamaejasmin	NN	B-NP	0	
.	.	.	0	0	

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

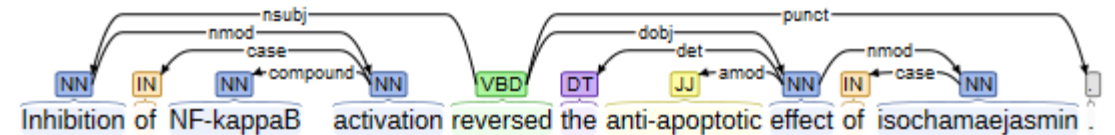
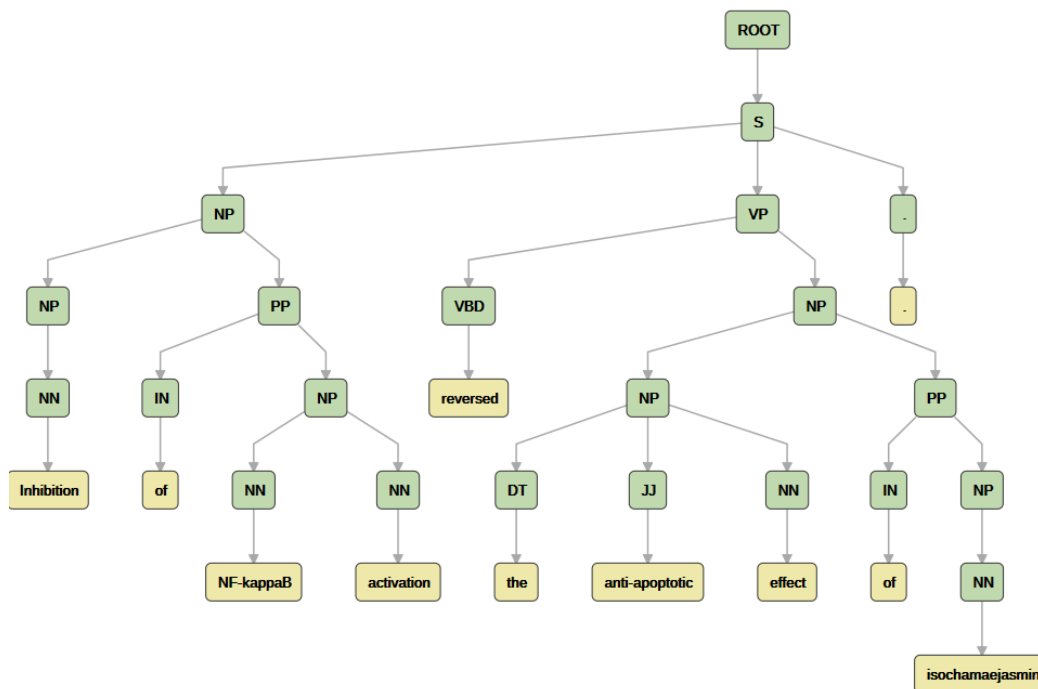
Parsing:

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



Stanford CoreNLP

NLTK 3.2.5



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

Biomedical resources:

- Documents

The screenshot shows a PubMed search results page. The search query is "alternatives to animal experiments". The results are sorted by "Most Recent" and show 1 to 20 of 3741 items. The first five results are listed:

- Antioxidant and Cholinesterase Inhibitory Activities of Ethyl Acetate Extract of *Terminalia chebula*, Cell-free *In vitro* and *In silico* Studies.**
Rajmohamed MA, Natarajan S, Palanisamy P, Abdulkader AM, Govindaraju A. *Pharmacogn Mag*. 2017 Oct;13(Suppl 3):S437-S445. doi: 10.4103/pm.pm_57_17. Epub 2017 Oct 11. PMID: 29142396
- The ethical justification for the use of non-human primates in research: the Weatherall report revisited.**
Amazon G. *J Med Ethics*. 2017 Oct 14. pii: medethics-2016-103827. doi: 10.1136/medethics-2016-103827. [Epub ahead of print] PMID: 29032368
- A Novel 3D Skin Explant Model to Study Anaerobic Bacterial Infection.**
Maboni G, Davenport R, Sessford K, Baiker K, Jensen TK, Blanchard AM, Wattegdera S, Entican G, Tolmeyer S. *Front Cell Infect Microbiol*. 2017 Sep 14;7:404. doi: 10.3389/fcimb.2017.00404. eCollection 2017. PMID: 28958685 Free PMC Article
- Beyond mouse cancer models: Three-dimensional human-relevant *in vitro* and non-mammalian *in vivo* models for photodynamic therapy.**
Kucinska M, Murias M, Nowak-Sliwinska P. *Mutat Res*. 2017 Jul;773:242-262. doi: 10.1016/j.mrev.2016.09.002. Epub 2016 Sep 12. Review. PMID: 28927532
- Standardized mean differences cause funnel plot distortion in publication bias assessments.**
Zwetsloot PP, Van Der Naald M, Sena ES, Howells DW, Int'Hout J, De Groot JA, Chamuleau SA, MacLeod MR, Wever KE. *Elife*. 2017 Sep 8;6. pii: e24260. doi: 10.7554/eLife.24260. PMID: 28844685 Free PMC Article



(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 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

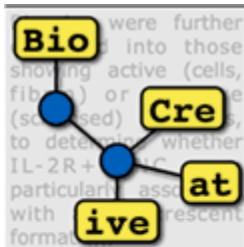
(Sources: <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html>
<https://bioportal.bioontology.org/>)

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)



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



DATABASE The Journal of Biological
Databases and Curation



**JOURNAL OF
BIOMEDICAL SEMANTICS**

Journal of Biomedical Informatics

Nucleic Acids Research



Biomedical resources:

- Conferences



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