

BioMedical NLP

Class 2 - Tools & Resources
NLP Master's Programme, University of Bucharest

Lect. Dr. Ana Sabina Uban
auban@fmi.unibuc.ro





Types of data

The **biomedical** domain is generally rich in public resources.
The **clinical** domain, not as much; + there are ethical and privacy issues with data in this domain.
Where resources are scarce, there is research using user-generated data (such as social media data)



Types of resources

- Research corpora (scientific papers)
- Clinical corpora (medical records, histories, ...)
- Medical/bioinformatical resources/knowledge bases e.g. repositories of macromolecular structural data (Protein Data Bank)
- Software tools (i.e. to access data repositories, e.g, Entrez tools to access 30 databases)
- Ontologies
- Pretrained language models / semantic representations for medical text (BioBERT,)
- Libraries and frameworks for medical data analysis for programmers: scispacy, SparkNLP



Structured data





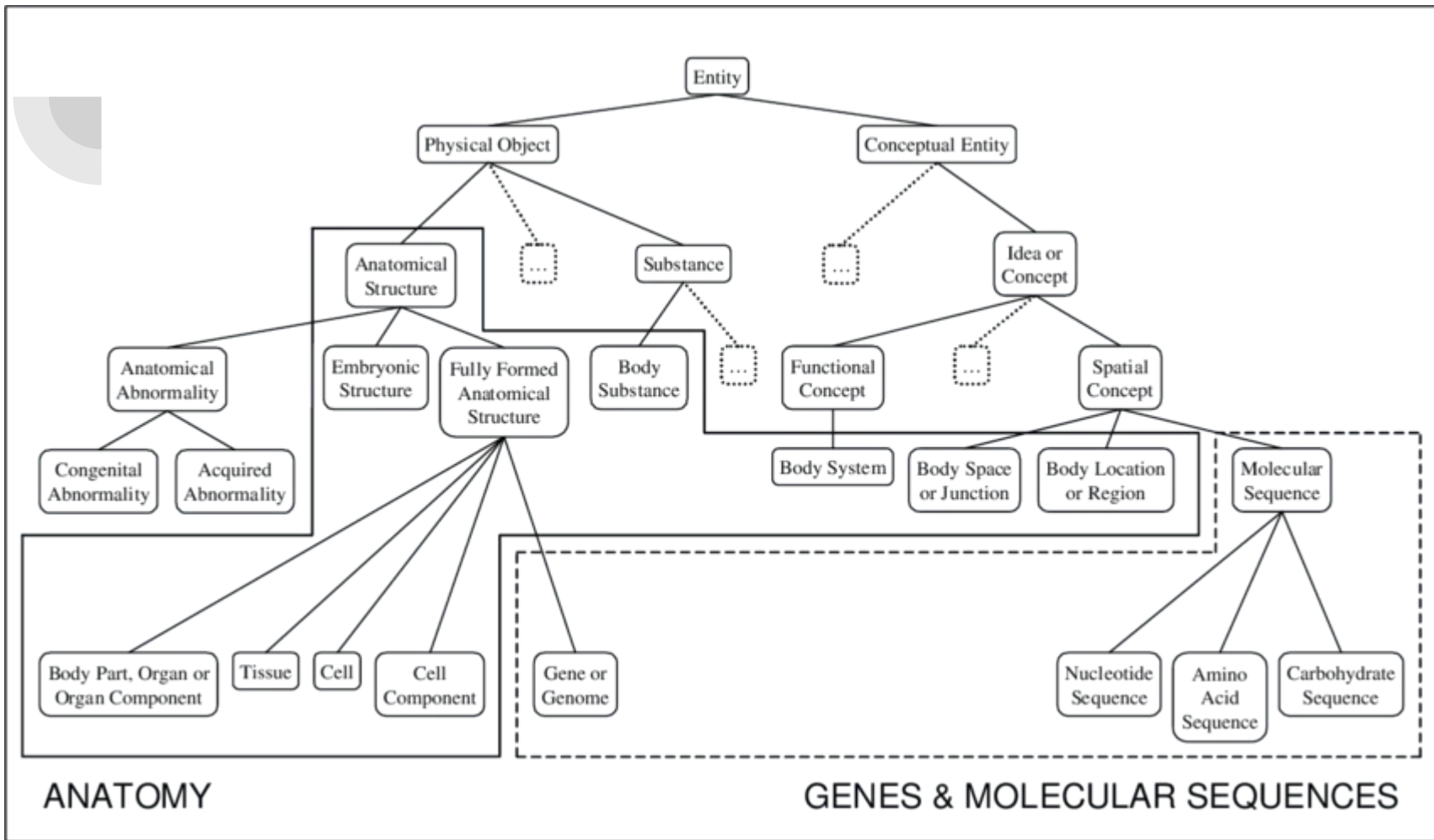
Ontologies

Ontology = concepts (+identifiers, synonyms, definitions) and relations

Special structure and language for encoding and querying (e.g. OWL)

Essential for the recent development of genome science

Major contribution of NLP in the medical field and more: recognition of concepts from ontologies when they occur in free text, entity linking





Ontologies

Used in combination with NLP for information extraction.

NLP useful for automatically building, improving and enriching ontologies:

- identifying ontology concepts in free text: very irregular, because of morphology, syntax, tokenization etc
- quality assurance for ontologies: violations of univocality for same concept (e.g. [Vespoor et al \(2009\)](#)) - avoid redundant terms
- mapping, aligning, and linking of different ontologies



UMLS (Unified Medical Language System)

Contains biomedical concepts, and links between them, in a unified format (from many different source vocabularies)

Consists of:

- Metathesaurus (information from original source vocabularies: names, attributes, relationships, plus unification of terms in source vocabularies): >2 million concepts linking >8 mil concept names/terms, from 150 source vocabularies
- Semantic Network: categorizing concepts into semantic types (groupings: organisms, anatomical structures, biological function, chemicals, events, physical objects, concepts/ideas) and relationships (133 semantic types and 54 relationships, both hierarchical and non-hierarchical e.g. IS-A, physically related to, spatially related to, functionally related to...)
- the SPECIALIST lexicon: contains syntactic, morphological and orthographic information for each biomedical term

Diagram of *Lou Gherig's disease* concept in the book (page 79)

Link: <http://www.nlm.nih.gov/research/umls/> (needs license to access)

SPECIALIST lexicons (open): <https://lhncbc.nlm.nih.gov/LSG/>



The Gene Ontology (GO)

NLP has facilitated the construction of genomic databases (through the automatic recognition of concepts in free text)

The Gene Ontology (built in 2000/2001 (The Gene Ontology Consortium 2000, 2001))

Became necessary in order to cross-reference genes across organisms (“does this gene in Organism A have the same function as this gene in Organism B?”),

They only had genomic databases per species and there were terminological differences; then joined the three different databases - initially developed only by the scientists, then the linguists came in

Contains three structured vocabularies describing 3 types of concepts associated to different genes:

- biological processes (more than one step, unlike molecular functions)
- cellular components (e.g. cell nucleus)
- molecular functions

Link: <http://geneontology.org/> + downloadable, browsable; tools <https://github.com/geneontology/>
example: <http://amigo.geneontology.org/amigo/search/annotation>



Databases of drugs and associated information

[Drugbank](#)

[Drug Repurposing Hub](#) [\[paper\]](#)

Drug-target Network [\[paper\]](#): built a bipartite graph composed of FDA-approved drugs and proteins linked by drug–target binary associations, based on DrugBank and other DBs.

[Orphanet Rare Diseases](#) - Rare diseases database and alignment with ICD-10, ICD-11, OMIM, UMLS, MeSH, MedDRA and GARD.

[Orphanet Rare Diseases Classification](#) - Rare diseases classified by medical specialties



Other resources

- [Medi-Span](#) (“encodes information about known usages, while Drugbank encodes information about drug targets and mechanisms of action”)
- [UCSC Xena](#) (genomic DBs, e.g. [cancer mutations](#))
- [ChemBank](#) (no longer maintained)
- [LINCS L1000](#) (gene expression profiles for thousands of perturbagens at a variety of time points, doses, and cell lines)
- DMAP (database of directed drug-to-protein effects and effect scores)
[\[paper\]](#)
- RareDis: a corpus annotated with rare diseases, their signs and symptoms
[\[paper\]](#)



Other resources

Dedicated database for mechanisms of action: [MANTRA](#)

“Mode of Action by NeTwoRk Analysis (MANTRA) is a computational tool for the analysis of the Mode of Action (MoA) of novel drugs and the identification of known and approved candidates for “drug repositioning”. It is based on network theory and non-parametric statistics on gene expression data.



Information extraction from structured data





Case study: Drug repurposing & Mechanism of Action

The **Mechanism of Action** = a term used to describe how a drug or other substance produces an effect in the body. For example, a drug's mechanism of action could be how it affects a specific target in a cell, such as an enzyme, or a cell function, such as cell growth.

For some drugs the mechanism of action is not clear.

Explicitly extracting the mechanism of action (known or unknown) for existing treatments could help find new applications for them. NLP and graph-based methods (structured data) can be used for extracting such information.



Mechanism of Action

Structured information on medical articles

Medical articles in MEDLINE/PubMed are published along with meta-data with relevant [MeSH terms](#). [MeSH information includes headings and subheadings, and qualifiers. \(notable e.g. “Pharmacological Action”\)](#)
[Mechanism-of-action \(MoA\) is not a single category directly linked to published articles through metadata on MEDLINE/PubMed.](#)

Discovering new mechanisms of action for known drugs or finding connections between mechanisms of action of certain treatments and their potential to be used for certain diseases is researched in the context of drug repurposing.

Then, connecting mechanism-of-action *referred to in a scientific article* in relation to a treatment or disease can only be inferred from this type of structured data + the unstructured text for the article.



Drug & Mechanism of Action discovery using ontologies (network analysis)

Discovering mechanisms of action for given substances or inferring potential new ones can be done based on existing ontologies on drugs and treatments, ...

In this approach, structured knowledge in the bio-medical domain comes in the form of ontologies or graphs, and new knowledge can be inferred by analyzing these graphs.

Some approaches in drug repurposing research are based on target-mechanism of action relations: these “take into account known mechanism of action and target role. Here, the data available on *signaling pathways, protein interactions and omics data are integrated* to identify the potential mechanism of action (MoA)” [\[paper\]](#) (and many others cited here...)



Drug & Mechanism of Action discovery using ontologies (network analysis)

- **Finding indirect links between a drug and treatment in a knowledge graph.** In particular between *a peptide and a rare disease* in [\[paper\]](#). Here they collect a corpus of 3M PubMed articles, use entity recognition to extract concepts and build a graph of their interactions, then mine indirect connections (paths in the graph) to find hidden mechanisms of action. As motivation: “Important pieces of information regarding chemical substances, biological processes and pathway interactions are scattered between publications from different communities of scientists, who are not always mutually aware of their findings”



Drug & Mechanism of Action discovery using ontologies (network analysis)

- A knowledge-based approach for drug discovery that can also provide potential MoA [\[paper\]](#). The authors extract predications (subject-relation-object triples) using SemRep from PubMed abstracts. The relation extracted from the UMLS Semantic Network are used for building the knowledge graph. A logistic regression model is trained on groundtruth and false sets of *drug* – *target* – *disease* and is further used for discovering candidate drugs for each disease. The model can predict the potential drugs from a given disease and the corresponding targets simultaneously. In addition to predict novel drugs, the method can also provide potential mechanism of action for the drugs.



Drug & Mechanism of Action discovery using ontologies (network analysis)

- EDGAR (Extraction of Drugs, Genes and Relations) - information extraction from MEDLINE database on the relationships between genes, drugs and cells (here for discovering cancer treatments). They use syntactic analysis based on the SPECIALIST Lexicon and rule-based extraction for parsing the text, plus structured information from Unified Medical Language System (UMLS) [\[paper\]](#).



Drug & Mechanism of Action discovery using ontologies (network analysis)

- **Topology-driven methods for discovering treatments for autism** [\[paper\]](#). Here they collect their own corpus of ~20K PubMed articles on autism build a graph of associated [MeSH](#) terms directly. They use “various network analyses including co-word network, clique percolation, weak component, pathfinding-based analysis of therapeutic groups, and detection of important drug interaction within a clique”. ... Specifically for autism, they discover four drugs (such as Tocilizumab, Tacrolimus, Prednisone, and Sulfisoxazole) are worthy of further study, “providing psychologists, physicians, and researchers with *data-based scientific hypotheses in autism-drug discovery*”.




Drug & Mechanism of Action discovery using ontologies (network analysis)

- Evaluating similarities between gene expressions: “The methodologies for evaluating similarities between gene expression profiles of different perturbagens are the key to understanding mechanisms of actions (MoAs) of unknown compounds and finding new indications for existing drugs - based on LINCS L1000 data“ [\[paper\]](#) and others.




Drug & Mechanism of Action discovery using ontologies (network analysis)

- Predicting drug-target interactions based on the *pharmacological effect similarity* : “drug-target interactions are more correlated with pharmacological effect similarity than with chemical structure similarity. We then develop a new method to predict unknown drug-target interactions from chemical, genomic and pharmacological data on a large scale. The proposed method consists of two steps: (i) prediction of pharmacological effects from chemical structures of given compounds and (ii) inference of unknown drug-target interactions based on the pharmacological effect similarity in the framework of supervised bipartite graph inference” [\[paper\]](#)



Corpora & other resources for NLP practitioners





Language models / semantic representations

Pre-trained transformer based models for BioNLP [\[survey\]](#):

Table 5

Summary of scientific literature-based BPLMs. NER - Named Entity Recognition, RE - Relation Extraction, IR - Information Retrieval, QA - Question Answering.

Model	Type	Pretrained from	Corpus	Publicly Available	Evaluation
BioBERT [16]	Scientific Literature	General BERT	PubMed and PMC	Yes	Biomedical NER, RE, and QA.
RoBERTa-base-PM [43]	Scientific Literature	General RoBERTa	PubMed and PMC	Yes	Sequence Labelling and Text Classification
RoBERTa-base-PM-Voc [43]	Scientific Literature	Scratch	PubMed and PMC	Yes	Sequence Labelling and Text Classification
BioALBERT [96]	Scientific Literature	General ALBERT	PubMed and PMC	Yes	Biomedical Concept Extraction
BioBERTpt-bio [97]	Scientific Literature	Multilingual BERT	Brazilian Biomedical corpus	No	Clinical Concept Extraction
PubMedBERT [20]	Scientific Literature	Scratch	PubMed and PMC	Yes	BLURB
BioELECTRA [98]	Scientific Literature	Scratch	PubMed	Yes	Biomedical NER, QA and RE.
BioELECTRA ++ [98]	Scientific Literature	BioELECTRA	PMC	Yes	Biomedical NER, QA and RE.
BioMegatron [99]	Scientific Literature	Scratch	PubMed and PMC	No	Biomedical NER, RE and QA.
OuBioBERT [21]	Scientific Literature	Scratch	PubMed	Yes	BLUE
BlueBERT-PM [18]	Scientific Literature	General BERT	PubMed	Yes	BLUE
BioMedBERT [100]	Scientific Literature	General BERT	BREATHE 1.0	No	Biomedical NER, IR and QA.
ELECTRAMed [101]	Scientific Literature	Scratch	PubMed	Yes	Biomedical NER, RE and QA
BioELECTRA-P [102]	Scientific Literature	Scratch	PubMed	Yes	BLURB, BLUE
BioELECTRA-PM [102]	Scientific Literature	Scratch	PubMed, PMC	Yes	BLURB, BLUE
BioALBERT-P [103]	Scientific Literature	ALBERT	PubMed	Yes	BLURB
BioALBERT-PM [103]	Scientific Literature	ALBERT	PubMed, PMC	Yes	BLURB



Language models / semantic representations

Pre-trained transformer based models for BioNLP [\[survey\]](#):

Table 2
Summary of EHR-based T-BPLMs.

Model	Type	Pretrained from	Corpus	Publicly Available	Evaluation
ClinicalBERT [17]	EHR	BioBERT	MIMIC-III Clinical Notes	Yes	MedNLI and Clinical Concept Extraction
ClinicalBERT (discharge) [17]	EHR	BioBERT	MIMIC-III Discharge summaries	Yes	MedNLI and Clinical Concept Extraction
MIMIC-BERT [40]	EHR	General BERT	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
ClinicalXLNet (nursing) [19]	EHR	General XLNet	MIMIC-III Nursing notes	Yes	Prolonged Mechanical Ventilation Prediction problem
ClinicalXLNet (discharge) [19]	EHR	General XLNet	MIMIC-III Discharge notes	Yes	Prolonged Mechanical Ventilation Prediction problem
BERT-MIMIC [39]	EHR	General BERT	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
ELECTRA-MIMIC [39]	EHR	General ELECTRA	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
XLNet-MIMIC [39]	EHR	General XLNet	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
RoBERTa-MIMIC [39]	EHR	General RoBERTa	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
ALBERT-MIMIC [39]	EHR	General ALBERTa	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
DeBERTa-MIMIC [39]	EHR	General DeBERTa	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
Longformer-MIMIC [39]	EHR	General Longformer	MIMIC-III Clinical Notes	Yes	Clinical Concept Extraction
MedBERT [31]	EHR	Scratch	Private EHR	No	Disease Prediction
BEHRT [27]	EHR	Scratch	Private EHR	No	Disease Prediction
BERT-EHR [32]	EHR	Scratch	Private EHR	No	Disease Prediction
AlphaBERT [75]	EHR	Scratch	Private EHR	No	Text Summarization



Tools & corpora

- [scispacy](#) - NLP library for scientific text
- [SparkNLP](#) for medical text (not free)
- [Lexical Systems Group](#) - suite of lexicons and tools for processing scientific medical data from National Library of Medicine
- [NegEx](#) - lexicon for negation: locates trigger terms indicating a clinical condition is *negated* or possible
- [Medical Data for Machine Learning](#) - medical datasets

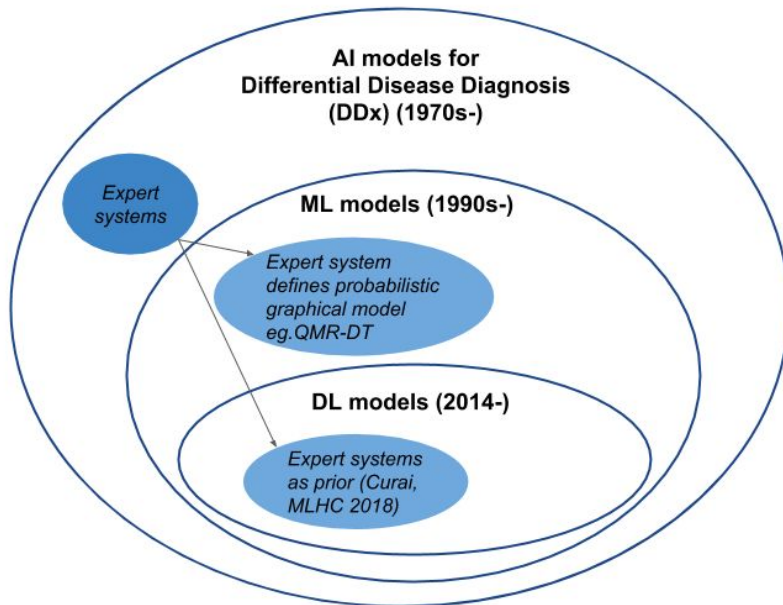
Datasets from shared tasks:

- [2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text - PMC](#) [\[paper\]](#)
- SemEval 2013 on drug-drug interactions (DDIExtraction): [source1](#), [source2](#)
- A Dataset for N-ary Relation Extraction of Drug Combinations [\[paper\]](#).
- RareDis: a corpus annotated with rare diseases, their signs and symptoms [\[paper\]](#)



More tools

Expert systems for medical diagnosis: <http://www.mghlcs.org/projects/dxplain>





Mental health

Annotated data

- Official diagnosis: EHRs? (see paper with FB data)
- Questionnaires? (see [eRisk](#))
- Therapist session transcriptions ([DAIC-WOZ](#))
- Social media data (semi-automatic annotations: eRisk, CLPsych)

Tools

Therapy chatbots: [woebot](#)



User generated data

- e-health: smartphone and fitness watches monitoring data from different sensors
- social media data posted by the user - clues on their physical&mental health
- news data relevant for fake news on medical topics



EHRs

Corpora of clinical data from electronic health records:

Next