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 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)
- Tools (i.e. to access data repositories, e.g, Entrez tools to access 30 databases)
- Ontologies
- Language models / Semantic representations for medical text (BioBERT, ...)

Corpora of scientific literature
and relevant institutions

US National Library of Medicine (NLM)

- 1818: Initially a resource for military physicians (the Surgeon's General Library) since 1818; 2,300 volumes
- 1870s: + 120,000 volumes in the 1870s + introduced Index Medicus (index of the journal articles in the Library).
- 1956: it became the USA national library in the Public Health Service: **NLM** - National Library of Medicine. (John F. Kennedy)
- 1960s: specifications for computerized system for producing the monthly Index Medicus => MEDLARS (Medical Literature Analysis and Retrieval System) - provided individualized bibliographies through specialized "demand search" => keeping up with the demand eventually raised the need for automatic information retrieval

Corpora of scientific literature
and relevant institutions

=> NLM issued a policy that stated “the need for a national resource for information systems research and development relevant for human health” and “a clearing-house and coordinating agency for information systems R and D within the Public Health Service”
1988: **NCBI** (National Center for Biotechnology Information) - division of NLM

Link: https://www.nlm.nih.gov/bsd/mmshome.html

Corpora of scientific literature
and relevant institutions

MEDLINE database

Medical Literature Analysis and Retrieval System Online

Large bibliographic database maintained by NLM: authoritative and comprehensive source of references into peer reviewed fundamental biomedical studies, and clinical evidence
19 million references to to articles from 4.800 biomedical journals in 30 languages
1,500-3,000 references added everyday
Each citation includes: title, authors, journal, type, date etc; + link to full paper from original publisher, many of them free

Link: http://www.nlm.nih.gov/bsd/

Corpora of scientific literature
and relevant institutions

Medical Subject Headings (MeSH)

Additional metadata for each MEDLINE citation.

MeSH: the thesaurus of NLM controlled vocabulary
23,000 descriptors arranged in a hierarchical structure
151,000 Supplementary Concept Records (additional chemical substance names)
+ software assistive tools since 2002: MTI (Medical Text Indexer) that automatically suggests MeSH headings (Aronson et al 2004)

Link: https://www.nlm.nih.gov/mesh/meshhome.html - downloadable! ; browser: https://meshb.nlm.nih.gov/
Example

Corpora of scientific literature

and relevant institutions

Corpora of scientific literature
and relevant institutions

PubMed

MEDLINE is accessible on the web through PubMed, NLM's gateway
PubMed is a boolean search engine that indexes titles, abstracts and metadata separately + query augmentation for recognized terms

over 7.5 million full-text biomedical articles

+ **Entrez** Programming Utilities for batch retrieval of MEDLINE citations

Link: https://pubmed.ncbi.nlm.nih.gov/

Corpora of scientific literature
and relevant institutions

PubMed Central International

PubMed Central: digital archive of biomedical literature (full texts of articles), among other such archives
Part of PubMed Central International network, which includes US PubMed Central and UK PubMed Central, PubMed Canada - they archive articles from journals and manuscripts funded by NIH and Welcome Trust and Canadian Institutes of Health Research respectively ([where they have rights to the content])
A common format format is used (the Journal Archiving and Interchange Tag Suite), in view of development of the global network for exchange of biomedical information

Link: https://www.ncbi.nlm.nih.gov/labs/pmc/

Corpora of scientific literature
and relevant institutions

GENIA

GENome Information Acquisition project: enhancing the MEDLINE knowledge base for specific purposes: automatic extraction of biochemical information from journal papers and abstracts by means of language engineering (Collier et al 1999)

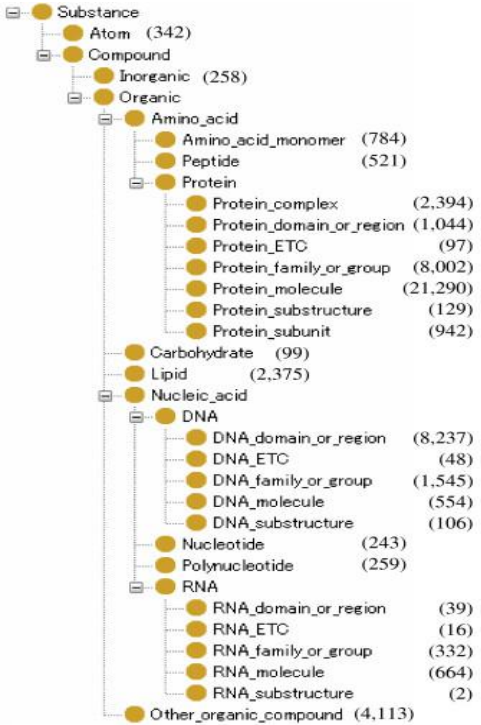
Developed eventually into a set of tools and annotated documents for development and use of BioNLP methods

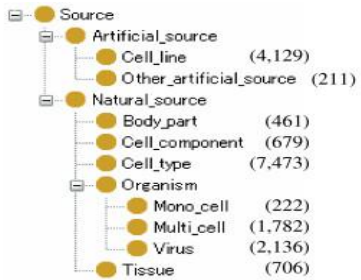
The corpus contains 1,999 MEDLINE selected abstracts, annotated with linguistic and semantic information from GENIA ontology (https://dl.acm.org/doi/pdf/10.5555/1289189.1289260)
Tools: POS tagger, shallow parsing and NER; annotation tool (XConc) = XML editor, concordancer, and ontology browser; tool for protein-protein interactions extraction

Corpora of scientific literature

and relevant institutions

GENIA





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

NLP useful for:

- 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

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/

Ontologies

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

Language models / semantic representations

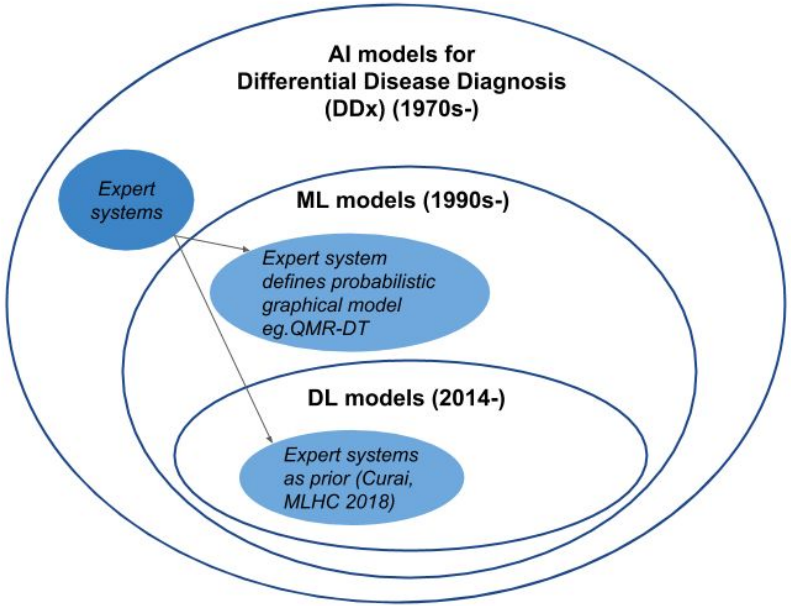
BioBERT, MedBERT, BioELECTRA, PubmedBERT, BLUEBert, ...

Why do we need them:

Example:

More tools

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



Other unstructured data

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

EHRs

Next