BioMedical NLP

Class 1
NLP Master's Programme, University of Bucharest

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

Organization

Materials:

slides, book, conference/workshop proceedings

Workshops:

Biomedical Natural Language Processing Workshop (BioNLP)

Special Interest Group on Biomedical Natural Language Processing - ACL Anthology

Workshop on Computational Linguistics and Clinical Psychology - ACL Anthology

Organization
Topics we will discuss:

8. CONȚINUTURI
CONTENT

8.1. Curs
Course
Introduction, main problems, relationship with other disciplines / Introducere in domeniu, probleme, relatii cu alte discipline
Resources and frameworks for bio-medical NLP / Resurse si unelte pentru NLP în domeniul medical
Medical information retrieval and information extraction / Regasirea informatiei si extragerea informatiei din date medicale
Processing Electronic Health Records (EHRs) / Procesarea dosarelor medicale electronice
Medical Named Entity Recognition (NER) / NER medicale
Mental health problems. Computational approaches / Abordari computationale ale problemelor mentale
Depression. NLP based approaches / Abordari computationale ale depresiei
Physical and mental health in social media: information and misinformation / Sanatatea si social media: informatie si dezinformare
Ethical issues and privacy protection in processing medical data / Probleme de etica si confidentialitate in procesarea datelor medicale
Recent research topics (main conferences and workshops ACL, EMNLP, NAACL; BioNLP, CLPsych etc.) / Teme recente de cercetare in domeniu (articole recente din principalele forumuri si workshopuri asociate - ACL, EMNLP, NAACL; BioNLP, CLPsych etc)

BioMedical NLP

Universal end goal for using technology in biomedicine: All in all, general end goal: **improving quality of life through understanding life processes.**

Useful for: doctors, patients, other actors in the medical system (insurance, policy makers); researchers, maintainers of resources used in biology/medicine research

BioMedical NLP

BioMedical NLP = part of the wider area of Artificial Intelligence for Medicine / bio-informatics

BioMedical NLP = the use of **NLP methods** for helping biology and medicine research, clinical practice, and adjacent areas

Relations to other disciplines

- medicine...
- psychology
- genetics/genomics
- biology
- other NLP subdomains: information retrieval, NER, knowledge representation
- other types of AI/ML in medicine:
    - computer vision
    - robotics
    - .....

Relations to other disciplines

Why general NLP is not enough

      most NLP has focused on different types of data (news, ...)

      we do have large volumes of data in biological and medical domain

            problems around privacy and organizing data

            problems with annotating data - requires expertise

      general sense representations don't do a good job for medical terms; specific
      meanings, important distinctions

Relations to other disciplines

Why general NLP is not enough

Semantic class	Examples
Cell lines	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*
Cell types	*primary T lymphocytes, natural killer cells, NK cells*
Chemicals	*citric acid, 1,2-diiodopentane, C*
Drugs	*cyclosporin A, CDDP*
Genes/proteins	*white, HSP60, protein kinase C, L23A*
Malignancies	*carcinoma, breast neoplasms*
Medical/clinical concepts	*amyotrophic lateral sclerosis*
Mouse strains	*LAFT, AKR*
Mutations	*C10T, Ala64 → Gly*
Populations	*judo group*

Figure 22.31 A sample of the semantic classes of named entities that have been recognized in biomedical NLP. Note the surface similarities between many of the examples.

From Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing: Second Edition, Daniel Jurafsky & James H. Martin. C

Relations to other disciplines

Why general NLP is not enough

Biomedical Term	Category	BERT	SciBERT	PubMedBERT (Ours)
diabetes	disease	✓	✓	✓
leukemia	disease	✓	✓	✓
lithium	drug	✓	✓	✓
insulin	drug	✓	✓	✓
DNA	gene	✓	✓	✓
promoter	gene	✓	✓	✓
hypertension	disease		✓	✓
nephropathy	disease		✓	✓
lymphoma	disease		✓	✓
lidocaine	drug		✓	✓
oropharyngeal	organ			✓
cardiomyocyte	cell			✓
chloramphenicol	drug			✓
RecA	gene			✓
acetyltransferase	gene			✓
clonidine	drug			✓
naloxone	drug			✓

From Biomedical NLP Group - Microsoft Research

Relations to other disciplines

Applications in: "Everything from automated screening and diagnosis, adaptive clinical trials, operations research, global health, precision medicine, home health and wearables, genomic analysis, drug discovery and design, robotics, and many more."

Interdisciplinary - Healthcare ecosystem:
"This includes AI developers, tech companies, policymakers and regulators, healthcare system leadership, pharmaceutical and device industry, frontline conditions, ethicists, even patients and patient caregivers.

Today in NLP still relatively niche community (smaller than the MT subdomain for example), not all include mental health (but we will); a few workshops: ...

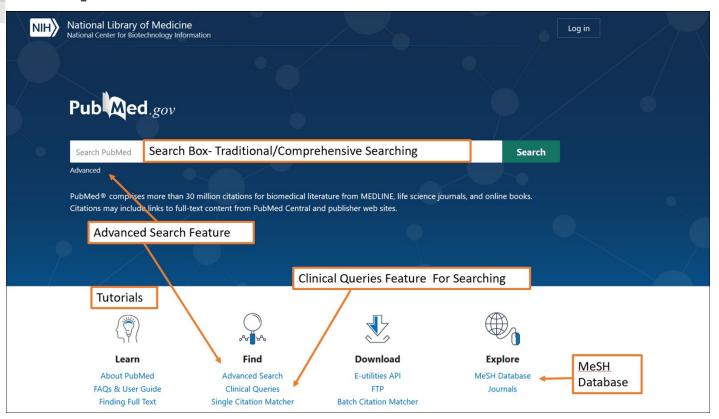# Relations to other disciplines

Holy trinity of the biological crossdisciplinary culture-"biology drives technology drives analytical tools revolutionizes biology." Biological breakthroughs require technological innovations which in turn necessitate new computing tools. New technological and computing tools, in turn, allow for the exploration of new biological frontiers.
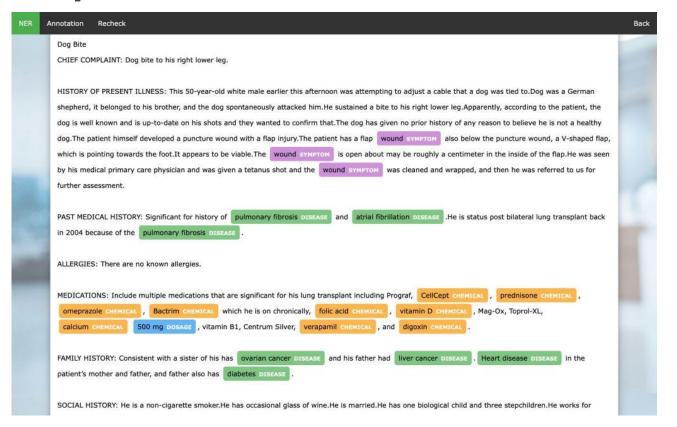*(Systems Biology and P4 Medicine: Past, Present, and Future)*

Importance of BioMedical NLP

Why is it important?...

- helping with medical decisions in rare cases (diagnosis, predicting side effects, effectiveness...) based on medical records databases
- advancing research through retrieving information in existing research
- health assistant, insights (health data collected from various sources, smartwatches, social media activity etc)

Importance

Importance

Importance

Some examples of AI for healthcare:

- someone has a rare side effect in some treatment, doctors look into other patient's medical records to make a decision
- detecting skin cancer with AI
- classifying tumors in radiology images
- the relevance of some genes for some diseases (discovered through analyzing occurrence in large amounts of research articles) leads to producing new targeted drugs

Importance

Importance

AI 'outperforms' doctors diagnosing breast cancer

Fergus Walsh
Medical correspondent
@BBCFergusWalsh

2 January 2020

R-CC   L-CC

Importance

Importance
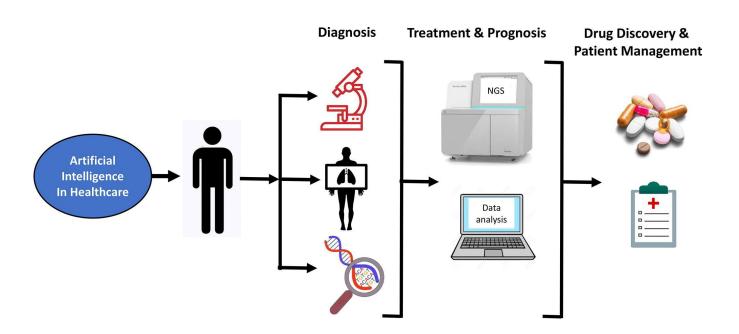## Innovations in Healthcare: the Role of Technology

½ of deaths preventable in Romania (⅓ in EU)

Large percentage of ineffective drugs across different diseases ("one size does not fit all")

Recent **paradigm shift** in healthcare: the age of **personalized medicine** - individual phenotypes & genotypes to determine predisposition to disease (P4 medicine, "Virtual Patient Model", computer-aided drug discovery)

Example innovation: targeted treatments for cancer based on genetic profile, increase effectiveness of treatment and life expectancy (e.g. small cell lung cancer)

Problem: the increase in data required for medical decisions > human cognitive capacity => technology to the rescue
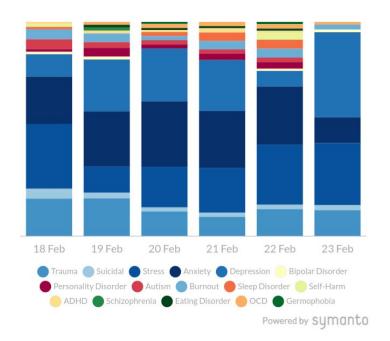
Importance

Importance

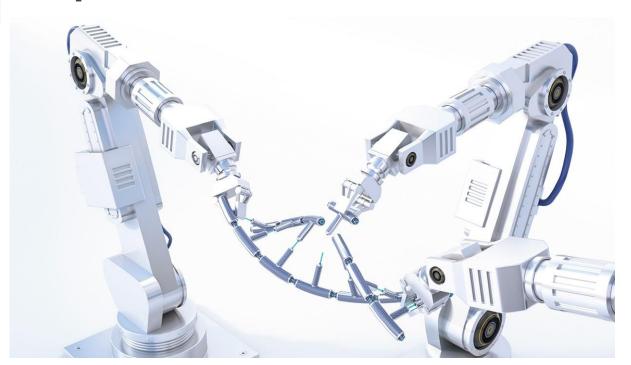# Importance

Importance

Example:

BioNTech a anunțat dezvoltarea unui nou instrument de monitorizarea a evoluției SARS-CoV2, Early Warning System, bazat pe inteligență artificială, care ar putea scurta timpul până la detectarea variantelor la risc înalt.

Prin modelarea datelor structurale legate de proteina spike și folosirea algoritmilor AI în cadrul platformei DeepChain, care prelucrează informațiile din bazele de date internaționale cu genomuri virale SARS-CoV-2, sistemul a permis detectarea cu două luni înainte a peste 90% din variantele raportate de OMS.

Importance

Most discussed mental health issues in the last 7 days

[Mental Health and Coronavirus | Live Impact Insights](#)

Importance

Importance

TABLE 1 | SELECTED RECENT FINANCINGS OF COMPANIES APPLYING AI IN DRUG DISCOVERY

Company	Date	Headline
Schrödinger	February 2020	Drug discovery software company closes $232 million IPO backed by Bill Gates and David Shaw.
Insitro	May 2020	Insitro raises $143 million in Series B funding, to help drive its machine learning-based drug discovery approaches further.
AbCellera	May 2020	AbCellera raises $105 million in Series B funding round to expand its antibody drug discovery platform.
Relay Therapeutics	July 2020	Relay Therapeutics, which focuses on understanding protein motion to design drug candidates, closes $400 million IPO.
Atomwise	August 2020	Sanabil Investments co-leads $123 million Series B funding round for Atomwise to support the development of its molecule identification softw...

[Tapping into the drug discovery potential of AI](#)

Types of data and users in bio-medical NLP

2 major areas:

- Mining clinical text <- improve healthcare outcomes
- Mining of biological publications <- advance knowledge of biology and medicine (discovery and understanding of fundamental physiological and pathological processes)

(borders disappearing in the case of personalized medical care based on fundamental biological studies)

3rd direction:

- User-generated data: social media etc; monitoring physical and mental health

Types of data and users in bio-medical NLP

Differences between subdomains and data types:

- types of text/data (heterogenous unstructured data in clinical domain vs papers in biological domain vs user-generated data);
- types and content of knowledge;
- available tools
  - scarcity of primary data in the case of the clinical domain vs papers in biological domain;
  - more lexical tools available for the clinical domain

Types of users

- clinicians: question-answering, document retrieval, ...
- coders: people assigning disease and procedure codes

Types of users

- model organisms database curators (model oraganisms e.g. mice, worm C elegans, zebra fish and genes in each): they are "crowd-produced", they need curators to triage and organize source information; need tools like document triaging, NER tools
- biologists doing experimental work: since mid-1990s, technical advancements allowed them to work on thousands and 30,000 genes at a time (as opposed to one) => need text mining to gather statistics about their behavior from previously published literature : text mining, information extraction, paper summarization; leveraging text mining on papers to perform data analysis such as clustering genes

History of BioMedical NLP

History of BioMedical NLP - clinical

1960s:

1966: **ELIZA** - system capable of detecting medical terms in free text using pattern matching (Weizenbaum 1966) https://web.njit.edu/~ronkowit/eliza.html one of the first chatbots in NLP (chatbot therapist);

1969: **APME** (automatic processing of medical English) - generate semantic interpretation of input medical text with the goal of identifying diagnoses (Pratt et al). Was rule-based and made use of SNOP lexicon (Systematized Nomenclature of Pathology) - terms in different semantic categories, primitive grammar

History of BioMedical NLP - clinical

1970s:

> first program for AI in medicine at Stanford (SUMEX-AIM) for clinical decision support
>
> Resulted in:
> - clinical applications, AI applications for diagnosis (infectious disease, diabetic retinopathy images), cancer drugs
> - AI Handbook Project - materials to explain AI for healthcare audience (to improve their practice)

Initial ideas started to get real results along with other applications in AI, along with increase in data volume and computational power

History of BioMedical NLP - clinical

1980s:
   identify biomedical concepts in patients' problem lists (e.g. SCAMP (Shapiro 1980));
   Linguistic String Project - LSP: mapped facts extracted from clinical text into a structured database (Friedman et al 1983)

present:
   more sophisticated systems being developed and used in clinical centers for extracting facts from clinical text, e.g. MedLEE (Medical Language Extraction and Encoding System),
   using lexicon, grammar, formal sentence representation, used for:
   - adverse event discovery,
   - risks of developing breast cancer,
   - association between diseases and drugs
   - ...

History of BioMedical NLP - biomedicine

1982:

    GenBank - earliest genomic database

1990s: changes in the way biological scientists worked, with advances in genomics and bioinformatics - now can analyze 10,000s genes in one experiment (instead of one).
 => need to identify patterns and organize data
      need to automatically analyze published papers (NLP) in order to identify new findings and populate gene databases
      building massive databases of all known genes in multiple organisms + gene function

History of BioMedical NLP - biomedicine

1998:

earliest paper on biological text mining *(Fukuda et al 1998, Blaschke et al 1999, Craven & Kumlien 1999)*

- named entity recognition,

- rule-based information extraction,

- machine learning based information extraction

History of BioMedical NLP - biomedicine

The first attempts were developed by biologists and not aligned with latest advances in NLP

"**Model organism databases**" - repositories of facts about specific genes for species of interest
- they were driving information extraction from biomedical literature
e.g.
- MedMiner (1999): literature exploration of gene-drug and gene-gene relationships (Tanabe et al 1999) (based on GeneCard + extracted from PubMed/MEDLINE);
- iHOP (2004) organized literature around gene/protein of interest (Hoffman & Valencia 2004);
- Textspresso (2004) linked gene names in papers to WormBase entries (Muller, Kenny & Sternberg 2004);
- ARROWSMITH application for information extraction (Smalheiser & Swanson 1999)

Open questions & challenges

Unresolved issues around using **AI in clinical medicine** (from the beginning):
-   workforce displacement
-   skill atrophy
-   algorithmic and user bias
-   patient privacy
-   medical-legal responsibility
-   oversight and regulations

Examples implemented systems

Open questions & challenges

TODO

- "is biomedical NLP efficient?" (see chapter 2)
- ethical problems
- fake news
- mental health
- current state and expected development
- most exciting recent developments, including biomedical AI? (See workshop proceedings)
-