Natural Language Processing and Large Language Models

What is Biomedical & Health Informatics?

William Hersh

Oregon Health & Science University

Natural language processing (NLP) and large language models (LLMs)

- Clinical NLP
- LLMs
- Future directions

Clinical NLP methods and results

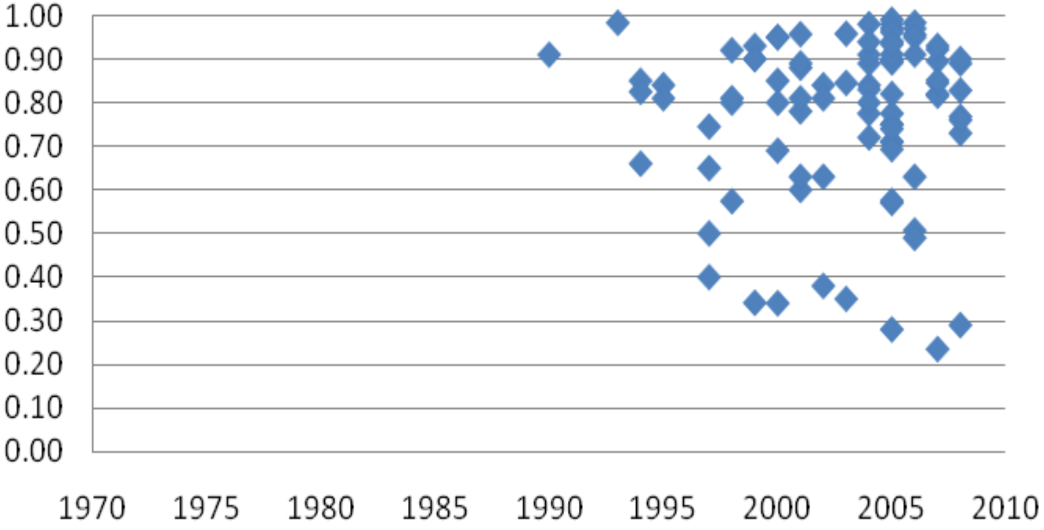
- Early approaches and systems
- Applications
- Systematic reviews
- Challenge evaluations

Early approaches and systems

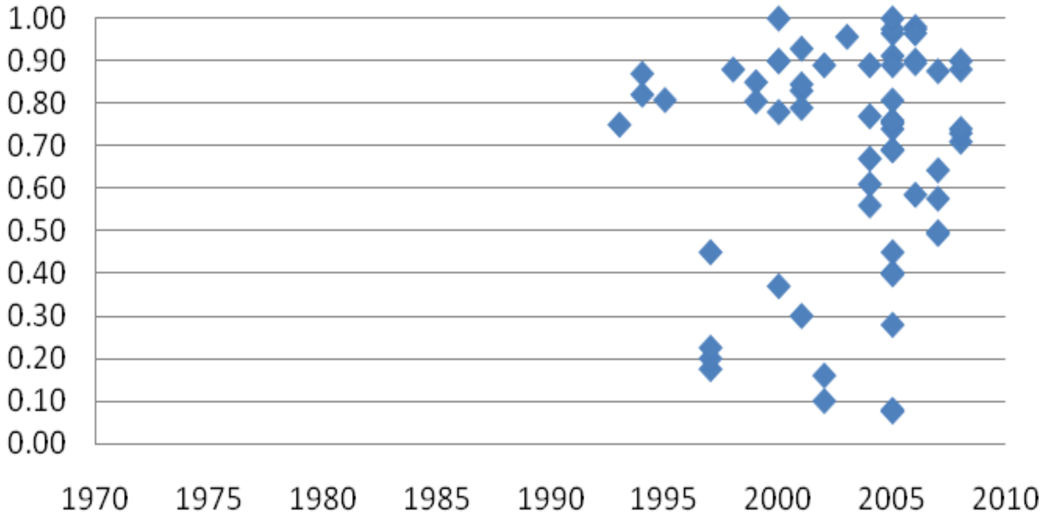
- Linguistic String Project (Sager, 1987)
  - Clinical notes were a “subgrammar” of larger human grammar
  - Most clinical narrative statements could be reduced to small number of information formats, e.g., medication, test and result, etc.
- Medical Language Extraction and Encoding System (MedLEE) (Friedman, 1994)
  - Core approach was “semantic grammar” that recognized terms and attributes but not syntax
  - Initially focused on radiology reports but expanded to other domains
  - Compared with human coders, fell within range of disagreement (Hripcsak, 1995)

Systematic review of early systems (Stanfill, 2010)

- Recall of coding and classification studies over time



- Precision of coding and classification evaluations over time



Application areas of clinical NLP

- Identifying patients and their attributes
  - Postoperative complications (Fitzhenry, 2013; Tien, 2015)
  - High-risk heart failure patients (Evans, 2016)
  - ICU risk of death and length of stay (Weissman, 2018)
  - Alcohol misuse (Afshar, 2019)
  - Geriatric syndromes (Chen, 2019)
  - Progression and mortality in cancer (Kehl, 2020)
  - Risk of nosocomial infection (Goodwin, 2020)
  - Social determinants of health (Feller, 2020)
  - COVID-19 patient advising and testing (Meystre, 2021)

Applications of clinical NLP (cont.)

- Improve processing or radiology images and reports
  - Terms from radiology reports generalized across institutions (Sugimoto, 2021)
  - Improving processing using standard ontologies (Filice, 2021)
  - Improved identification of findings in CXRs by processing text of reports (Zhou, 2022)
- Measuring healthcare quality
  - Determination of healthcare quality measures (Hazlehurst, 2005; Yetisgen, 2014; Kim, 2017, Meystre, 2017)
  - Implementation in practice settings (Garvin, 2018)
- Assisting patients
  - Linking EHR language to lay definitions (Chen, 2018)
- Conversational agents
  - Assist physicians with prescribing by linking to knowledge-based information (Preininger, 2020)

Applications of clinical NLP (cont.)

- Augmenting clinical research
  - Finding patients with congestive heart failure (Pakhomov, 2007)
  - Electronic Medical Records and Genomics (eMERGE) Network
    - https://emerge-network.org/
    - Aims to link phenotype (patient data) with genotype (genetic sequencing) (McCarty, 2011; eMERGE Consortium, 2021)
    - Early work replicating genome-wide association studies (Ritchie, 2010; Denny, 2013)
    - Recent focus on polygenic risk scores (Xu, 2021)
  - Case detection of diabetes (Zheng, 2016)
  - Association between androgen deprivation therapy and risk of dementia (Nead, 2017)
  - Extraction outcomes in cancer patients from radiology reports (Kehl, 2019) and pathology reports (Alawad, 2020)
  - Cohort selection for clinical studies (Wang, 2019; Chamberlin, 2020)
  - Classifying patients into phenotypes using deep learning (Si, 2021)

More recent applications

- Zero-shot prompting for patient from EHR text (Sivarajkumar, 2022)
- BERT model of de-identified clinical notes for diagnostic code assignment and therapeutic drug class inference (Sushil, 2022)
- Reading comprehension tasks from clinical practice guidelines (Mahbub, 2023)
- Combining human and machine intelligence for clinical trial eligibility querying (Fang, 2022)

Recent systematic reviews

- Measure and improve quality of diabetes care (Turchin, 2021)
- Use of SNOMED CT to represent processed text (Gaudet-Blavignac, 2021)
- A scoping review of publicly available language tasks in clinical NLP (Gao, 2022)
- Use of unstructured text in prognostic clinical prediction models (Seinen, 2022)
- Automatic documentation of professional health interactions (Falcetta, 2023)
- Using text mining in clinical decision support – only 20% of systems tested in actual clinical use (van de Burgt, 2023)

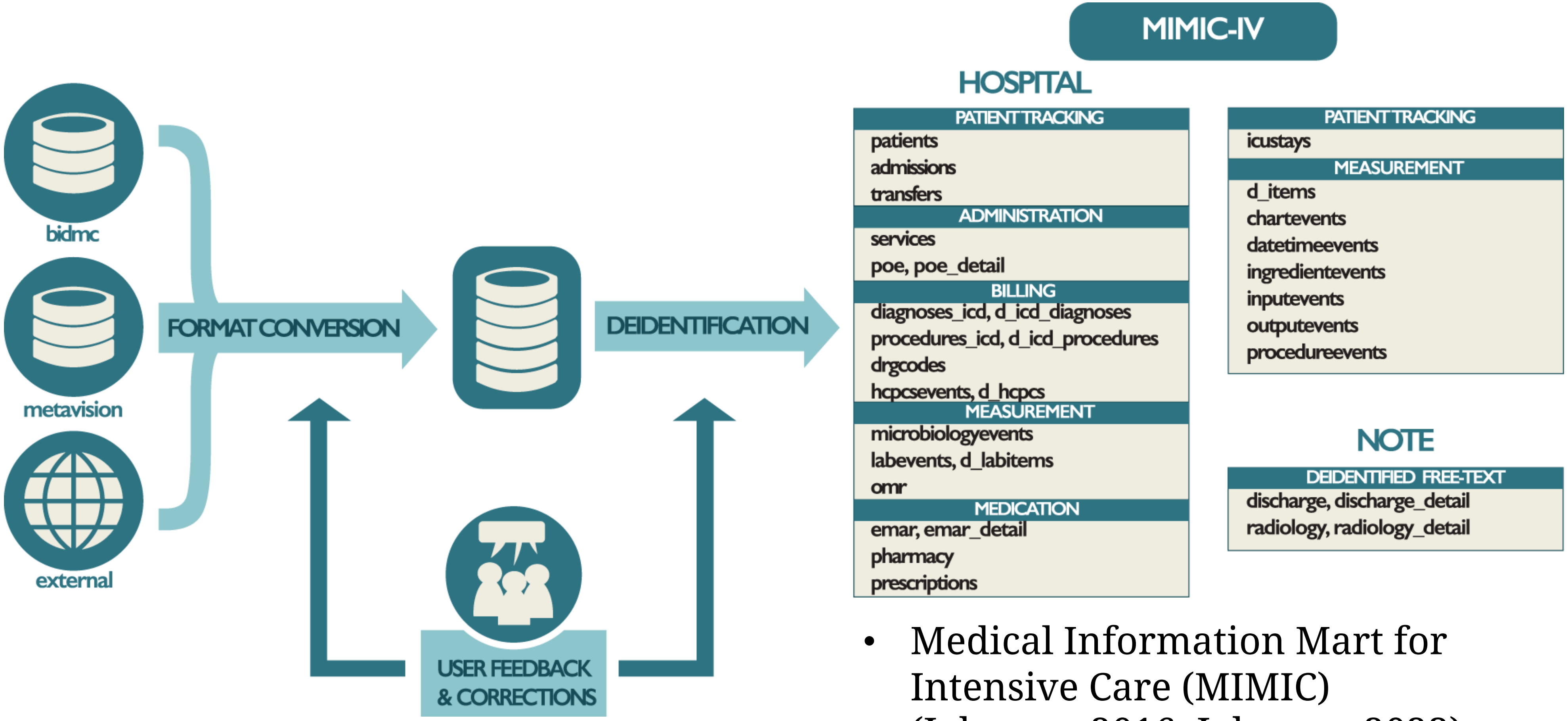
I2b2/n2c2 challenge evaluations

- Annual challenges with overview and system papers
  - https://www.i2b2.org/NLP/DataSets/Main.php
  - Automated de-identification of records (Uzuner, 2007)
  - Identification of smoking status (Uzuner, 2008)
  - Identification of obesity and its co-morbidities (Uzuner, 2009)
  - Extracting medication information (Uzuner, 2010)
  - Relationships between concepts in clinical text (Uzuner, 2011)
  - Coreference resolution and sentiment classification (Uzuner, 2012)
  - Temporal relations (Sun, 2013)
  - De-identification and cardiovascular risk factor detection (Stubbs, 2015)
- Recast as National NLP Clinical Challenges (n2c2)
  - https://n2c2.dbmi.hms.harvard.edu/
  - Cohort selection for clinical trials (Stubbs, 2019)
  - Adverse drug events and medication extraction in EHRs (Henry, 2020)
  - Concept normalization in clinical records (Henry, 2020)
  - Clinical semantic textual similarity (Wang, 2020)
  - Family history extraction (Shen, 2021)
  - Contextualized medication event extraction
  - Extracting social determinants of health (SDOH) (Lybarger, 2023)
  - Progress note understanding: assessment and plan reasoning

Additional tasks and data sets

- Question-answering – from document; MultiMedQA includes data sets emrQA (Pampari, 2018), MedQA (Jin, 2021), MedMCQA (Pal, 2022), PubMedQA (Jin, 2019), MMLU clinical topics (Hendrycks, 2020)
- Natural language inference (MedNLI) – conclusion inferred from sentence, EHR records annotated by clinicians (Romanov, 2018; Shivade, 2019)
- CLIP dataset for extracting action items for physicians from hospital discharge notes (Mullenbach, 2021)
- Discharge Summary Clinical Questions (DiSCQ) – 2,000+ questions paired with snippets of text (triggers) that prompt each question (Lehman, 2022)
- cpgQA question-answering dataset for clinical practice guidelines (Mahbub, 2023)

Another challenge – reliance on almost a single source of data



- Medical Information Mart for Intensive Care (MIMIC)
- (Johnson, 2016; Johnson, 2023)
- https://physionet.org/about/database/

LLMs

- Generate code (Li, 2022), solve college-level math (Drori, 2022), generate images (Ramesh, 2021) and video from text (Edwards, 2022)
- Answering clinical questions
  - PaLM (Singhal, 2022; Chowdhery, 2022) – basis of Google healthcare chatbot (Saha, 2022)
  - Smaller clinical models outperform larger general models (Lehman, 2023)
  - PubMedGPT performed well on clinical questions (Bolton, 2022)
- GatorTron large model with clinical tuning showed state-of-the-art results for concept and relationship extraction, textual similarity detection, natural language inference, and question-answering (Yang, 2022)
- Extracting breast cancer phenotypes from electronic health records (Zhou, 2022)
- Publicly available BERT embeddings better for extraction than de-identification tasks (Alsentzer, 2019)
- Longer-sequence models perform better with longer clinical texts (Li, 2023)

ChatGPT – https://chat.openai.com/

- Diagnostic and triage accuracy for 45 vignettes comparable to physicians (Levine, 2023)
- Answers to 21 of 25 questions about cardiovascular disease prevention deemed acceptable by cardiology clinicians for patient-facing information platform and as AI-generated draft responses to questions sent by patients for clinician review (Sarraju, 2023)
- Performed at or near passing for three levels of USMLE (Kung, 2023)
- Scientific abstracts undetectable by plagiarism checkers (Gao, 2022)
- Can create templates for discharge summaries (Patel, 2023)
- Need policies for “non-human” authors of scientific papers (Liebrenz, 2023; Flanagin, 2023; Nature, 2023)

Challenges for clinical NLP

- "Note bloat" and redundancy in clinical notes reduce NLP model performance (Liu, 2022)
  - Growing note length and redundancy over years (Rule, 2021)
  - Half of discharge summary content emanates from outside records of hospital stay – 39% from other records and 11% from no records at all (Ando, 2022)
- Benchmark data sets tend to focus on information task and not clinical need (Blagec, 2023)
- LLMs expensive to build and maintain – can only be done by large companies (Dickson, 2022)
- In voice recognition, non-lexical conversational sounds (*mm-hm*, *uh-huh*, etc.) degrade performance (Tran, 2023)

Ethical issues in NLP

- Language bias – training on large amounts of language “learns” biases inherent in text
  - Google searches for “professional” vs. “unprofessional” hair styles reveal racial differences (Alexander, 2016)
  - Occupations by gender and race (Sheng, 2019)
  - Treatment of pain by race; can be corrected for (Logé, 2021)
  - ChatGPT a blurry JPEG of Web? (Chiang, 2023)
- Privacy
  - Google, Apple, and others show large language models trained on public data expose personal information (Carlini, 2020; Wiggers, 2020)
  - May be compromised by need for data to be used to improve tools, e.g., Dragon voice recognition (Ross, 2022)
- GPT-3
  - Medical advice to agree with human decision to carry out suicide (Rousseau, 2020)
  - Use of racist language (Heaven, 2020)

Ethical issues (cont.)

- Galactica from Meta (Taylor, 2022) pulled when gave wrong answer for treating seizures due to failure to recognize negation (Birhane, 2022)
- At-home consumer devices used for medical purposes, e.g., should Alexa diagnose Alzheimer's? (Simon, 2022)
- Ethical considerations about potential to address and/or perpetuate bias (Fu, 2022; Bear Don't Walk, 2022)
  - Selecting metrics that interrogate bias in models
  - Opportunities and risks of identifying sensitive patient attributes
  - Best practices in reconciling individual autonomy, leveraging patient data, and inferring and manipulating sensitive information of subgroups

Academic and commercial NLP systems

- Academic
  - MetaMap – from NLM, maps to concepts of UMLS Metathesaurus (Aronson, 2010)
    - https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html
    - MetaMap Lite provides simpler and faster version (Demner-Fushman, 2017)
      - https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/run-locally/MetaMapLite.html
  - EMERSE – from University of Michigan (Hanauer, 2006; Hanauer, 2015)
    - https://project-emerse.org/
  - cTAKES – from Mayo Clinic (Savova, 2010)
    - https://ctakes.apache.org
  - Canary – from Brigham & Women's Hospital (Malmasi, 2017)
    - http://canary.bwh.harvard.edu/
  - CLAMP – from UT Houston (Soysal, 2018)
    - https://clamp.uth.edu/
- Commercial
  - Nuance – acquired by Microsoft
    - https://www.nuance.com/omni-channel-customer-engagement/technologies/natural-language-understanding.html
  - Lingumatics
    - https://www.linguamatics.com/
  - M*Modal – acquired by 3M
  - Discern nCode – acquired by Cerner
  - Health Fidelity – commercial version of MedLEE
    - https://healthfidelity.com/

Future directions for clinical NLP

- NLP must move focus from "tasks as decisions" to "tasks as needs" for clinical use (Lederman, 2022)
- For population health management and measurement – need (Tamang, 2023)
  - Readiness of data and compute resources
  - Organizational incentives to use and maintain systems
  - Feasibility of implementation and continued monitoring
- Priorities for ChatGPT research – human expertise, accountability, open systems, embrace benefits, widen debate (van Dis, 2023)