

6.5b Natural Language Processing

10x10 – Introduction to Biomedical Informatics & Artificial Intelligence
William Hersh, MD, FACMI, FAMIA, FIAHSI
Copyright 2026, All Rights Reserved



Natural language processing (NLP)

Definitions and overviews

Motivations and uses

Challenges for processing clinical narratives

Methods and results

Challenges for clinical NLP

Future directions

Methods of NLP have mirrored epochs of AI

Initial work in 20th century focused on development of lexicons, grammars, and rules for specific domains

Syntax done via parsing, which required grammar – the rules governing the syntax of language

- Most common way to express grammar was as set of rewrite rules
 - e.g., $S \rightarrow NP VP NP$, (e.g., *The patient has severe hypertension.*)
 - $NP \rightarrow DET NP$, $NP \rightarrow ADJ NP$, $NP \rightarrow NOUN$
 - Symbols that cannot be further decomposed (e.g., ADJ, NOUN) are called terminal symbols
 - Symbols that can be further decomposed (e.g., S, NP) are non-terminal symbols

Semantics done by mapping parts of speech into standardized terminology

- Most descriptive terminology for NLP is SNOMED CT

Early approaches and systems focused on EHR

Linguistic String Project (Sager, 1987)

- Clinical notes were “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)

Other applications focused on specific tasks

- Predictions, e.g., postoperative complications (Fitzhenry, 2013; Tien, 2015)
- Progression and mortality in cancer (Kehl, 2020)
- Determination of healthcare quality measures (Hazlehurst, 2005; Yetisgen, 2014; Kim, 2017, Meystre, 2017)

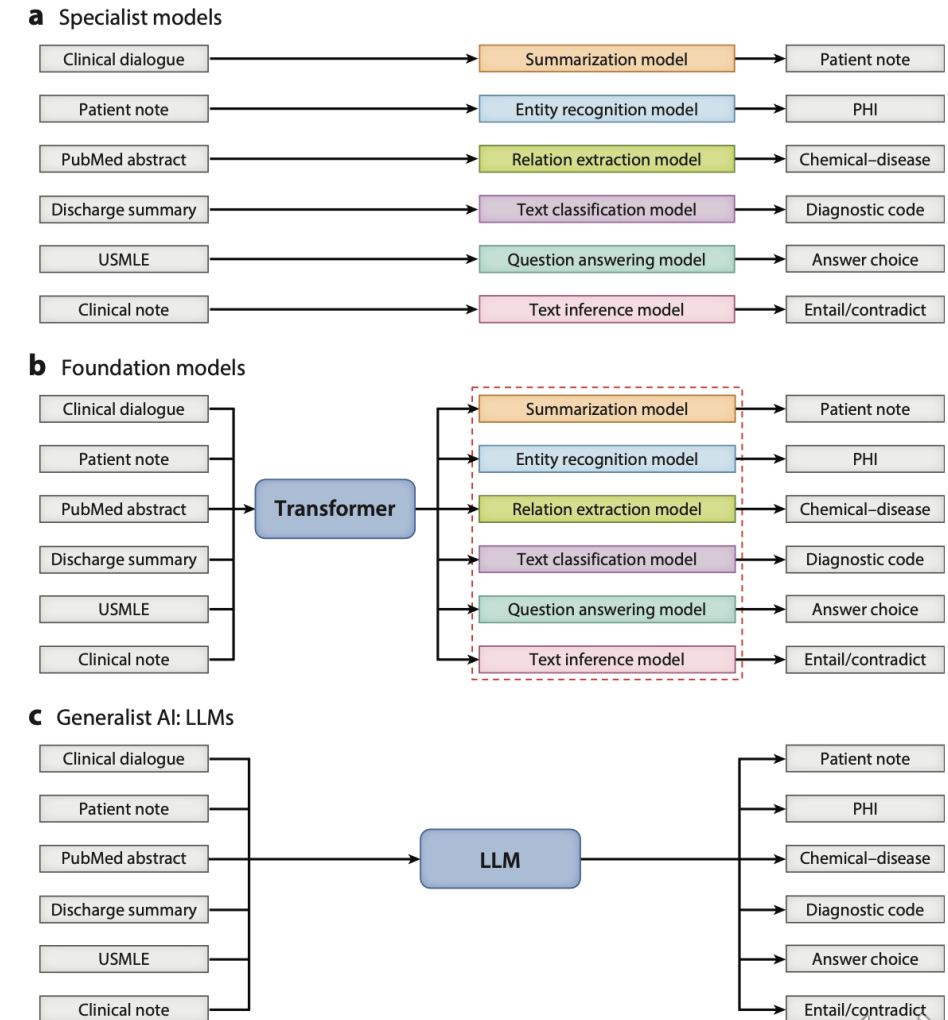
From rule-based to statistical-probabilistic to LLM methods

Initial statistical approaches followed advances in ML, e.g., supervised learning to deep learning, especially use of recurrent neural networks (Wu, 2020)

Advanced to use of embeddings

- Semantically meaningful mathematical representations of word, phrase, or other text, usually represented as vector so words and phrases with similar meanings have similar vectors (Percha, 2021)
- Allowed discovery of synonyms, mapping to concepts, identifying related concepts, etc.
- Gave rise to encoder models
 - Bidirectional Encoder Representations from Transformers (BERT) (Devlin, 2019)
 - Bidirectional and Auto-Regressive Transformers (BART) (Lewis, 2020)

And now, use of LLMs (Sarker, 2024; Usuyama, 2025)



Evaluation of NLP systems

Usually measured by

- Recall – proportion of correct assignments made (sensitivity)
 - e.g., 75 out of 100 concepts found = 75%
- Precision – proportion of correct assignments made (positive predictive value)
 - e.g., 150 concepts found, 75 correct = 50%
- F-Score = $2 * [(Precision \times Recall) / (Precision + Recall)]$

Often carried out in “challenge evaluations,” where multiple research groups compare results on the same task (Filannino, 2018)

- Largest clinical text challenge evaluation is [i2b2 NLP Shared Tasks](#)
- Renamed to [National NLP Clinical Challenges \(n2c2\)](#)

I2b2/n2c2 challenge evaluations

i2b2 challenges

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)

n2c2 challenges

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 (Mahajan, 2023)

Extracting social determinants of health (SDOH) (Lybarger, 2023)

Progress note understanding: assessment and plan reasoning (Gao, 2023)



Clinical NLP using LLMs

Studies assessing existing NLP tasks and datasets using LLMs with variable tuning and prompting

- GatorTronGPT training with clinical, biomedical and other text for 7 tasks (Peng, 2024)
- GPT and LLama variants on 6 tasks (Chen, 2025)
- Me-LLama foundational open-source model additionally trained with medical text on 6 tasks (Xie, 2025)

In general, perform comparably or somewhat better than non-LLM approaches, with additional concerns for LLMs in terms of hallucinations and higher cost

GatorTronGPT (Peng, 2024)

Methods and results

Clinical concept extraction

Clinical relation extraction

Clinical concept normalization

Clinical abbreviation disambiguation

Natural language inference (NLI)

Medication attribute filling

Progress note understanding

Tasks

Used unified model with fine-tuning and task-specific prompts

Achieved slightly better performance than previous models for first 5 tasks

Major limitation was hallucinations that were nonlogical, irrelevant, or interpretable

GPT and LLama variants (Chen, 2025)

Methods and results

Named entity recognition

Relation extraction

Multi-label document classification

Question answering

Text summarization

Text simplification

Tasks

Assessed GPT and LLama variants with zero-shot, few-shot, and fine-tuning with BERT and BART models

On 12 benchmarks across tasks, fine-tuning performed generally best but not always better than previous state of art

Limitations included inconsistent, missing, and hallucinated responses

Me-LLama (Xie, 2025)

Tasks

Question answering

Named entity recognition

Relation extraction

Classification

Summarization

Natural language inference

Methods and results

Me-LLaMA developed through continual pretraining and instruction tuning of LLaMA2 models using diverse biomedical and clinical data sources

Outperformed existing open medical LLMs in zero-shot and supervised settings and surpassed GPT-4 after task-specific instruction tuning for most text-analysis tasks

Some susceptibility to factual errors or biased information



GPT-4 processed free-text radiology reports into structured reports without error (Adams, 2023)

CLINICAL HISTORY: Evaluation for pancreatic cancer. TECHNIQUE: Non-contrast, arterial phase, and venous phase contrast-enhanced CT images of the abdomen and pelvis were obtained using a multidetector CT scanner. Multiplanar reformations were also reviewed. FINDINGS: Pancreatic tumor: There is a 3.5 x 3.2 x 2.8 cm hypodense mass in the pancreatic head, which demonstrates mild heterogeneous enhancement on arterial and venous phases. The mass causes mild upstream dilation of the pancreatic duct, measuring up to 4 mm in diameter. There is no pancreatic parenchymal atrophy. Metastatic disease: Multiple liver lesions are identified, the largest being in segment VIII, measuring 2.3 x 1.8 cm, with arterial enhancement and washout on the venous phase, consistent with metastatic deposits. No intrahepatic biliary dilatation is observed. There are enlarged peripancreatic, porta hepatis, and retroperitoneal lymph nodes, the largest measuring 1.8 x 1.3 cm. No pelvic or retrocrural lymphadenopathy is detected. No suspicious lung nodules or pleural effusions are noted. Vascular involvement: The pancreatic mass encases the superior mesenteric vein (SMV) and partially compresses it, with less than 50% narrowing of the vessel lumen. The main portal vein, splenic vein, and superior mesenteric artery (SMA) are patent and not encased by the tumor. No signs of cavernous transformation are observed. Thrombosis: No thrombus is identified within the portal vein, SMV, splenic vein, or inferior vena cava (IVC). Anatomy: The liver, spleen, adrenal glands, and kidneys demonstrate normal size, shape, and enhancement pattern. No renal or ureteral calculi are detected. The gallbladder, appendix, and bowel loops are unremarkable. No free fluid or air is observed in the abdominal or pelvic cavities. IMPRESSION: Hypodense mass in the pancreatic head, measuring 3.5 x 3.2 x 2.8 cm, consistent with a pancreatic neoplasm. Multiple liver lesions consistent with metastatic disease. Enlarged peripancreatic, porta hepatis, and retroperitoneal lymph nodes, suggestive of lymph node metastases. Partial encasement and

```
{
  "COMPARISON": {
    "ADENOPATHY_PRESENT": "yes",
    "ASCITES_PERIPANCREATIC_FLUID": "no",
    "METASTATIC_DISEASE": "yes",
    "PANCREATIC_TUMOR_TABLE": {
      "BILIARY_INVOLVEMENT": "no",
      "CONFINED_TO_PANCREAS_WITH_CLEAR_FAT_PLANE": "yes",
      "ENHANCEMENT_RELATIVE_TO_PANCREAS": "hypo",
      "LOCATION": "head",
      "PANCREATIC_TUMOR_PRESENT": "yes",
      "REMAINING_PANCREAS": "yes ductal dilation",
      "SIZE": "3.5 x 3.2 x 2.8 cm"
    },
    "PANCREATIC_VASCULAR_TABLE": {
      "ABERRANT_ANATOMY": {
        "MAJOR_ACCESSORY_OR_OTHER_REPLACED_ARTERIES_VEINS_COLLATERALS_DILATED_VESSELS": "no",
        "REPLACED_RIGHT_HEPATIC_ARTERY": "no"
      },
      "ATHEROSCLEROTIC_ORIGINS_OF_CELIAC_AXIS_SMA": "no",
      "CELIAC_INVOLVEMENT": "no",
      "DISTANCE_TO_SMV": "mm",
      "OTHER_VASCULAR_INVOLVEMENT": "no",
      "SMA_INVOLVEMENT": "no",
      "SMV_INVOLVEMENT": "50% /no",
      "THROMBOSIS_ANY_VESSEL": "no",
      "VASCULAR_TUMOR_INVOLVEMENT_AND_DEGREE": "360°"
    }
  },
  "IMPRESSION": "Hypodense mass in the pancreatic head, measuring 3.5 x 3.2 x 2.8 cm, consistent with a pancreatic neoplasm. Multiple liver lesions consistent with metastatic disease. Enlarged peripancreatic, porta hepatis, and retroperitoneal lymph nodes, suggestive of lymph node metastases. Partial encasement and compression of the superior mesenteric vein by the pancreatic mass, without thrombosis. No other vascular involvement or thrombosis identified. No additional intra-abdominal or pelvic pathologies detected.",
  "INDICATION": "Evaluation for pancreatic cancer",
  "TECHNIQUE": "Non-contrast, arterial phase, and venous phase contrast-enhanced CT images of the abdomen and pelvis were obtained using a multidetector CT scanner. RADIATION DOSE (DLP): [dose] mGy*cm."
}
```



Other applications and results

BRIDGE – 87 tasks from real-world clinical data sources across nine languages (Wu, 2025)

- Open-source LLMs achieved performance comparable to proprietary models, while medically fine-tuned LLMs based on older architectures often underperformed versus updated general-purpose models

Many use cases and data sets for omics research (Rannon, 2025)

Biomedical Retrieval-Augmented Generation Benchmark (BIORAB) evaluation framework across 5 biomedical NLP tasks and 11 datasets found RAG methods outperformed regular LLMs but still had limitations (Li, 2025)

LLM Python package (Hsu, 2025)

NLP datasets

[PubMedQA](#) (Jin, 2019)

[MedQA](#) (Jin, 2021)

- Based on USMLE exams

[MedMCQA](#) (Pal, 2022)

- Multiple-choice medical exams

[LiveQA](#) at TREC 2017 (Abacha, 2017)

- Consumer health question answering

[emrQA](#) (Pampari, 2018)

- Medication event extraction

[MedNLI](#) (Romanov, 2018; Shivade, 2019)

- Natural language inference

[MedicationQA](#) (Abacha, 2019)

- Consumer medication questions

[CLIP](#) (Mullenbach, 2021)

- Extracting action items for physicians from hospital discharge notes

[MMLU](#) (Hendrycks, 2021)

- Multi-domain language understanding, including clinical

[DiSCQ](#) (Lehman, 2022)

- Discharge summary clinical questions

[BioASQ](#) (Krithara, 2023)

- Biomedical semantic indexing and question-answering

[PLABA](#) (Attal, 2023)

- Plain language adaptation of biomedical abstracts

[cpgQA](#) (Mahbub, 2023)

- Machine reading comprehension tasks on clinical practice guidelines



NLP systems – academic and commercial

Academic

[MetaMap](#) – from NLM, maps to concepts of UMLS Metathesaurus (Aronson, 2010)

- [MetaMap Lite](#) provides simpler and faster version (Demner-Fushman, 2017)

[EMERSE](#) – from University of Michigan (Hanauer, 2006; Hanauer, 2015)

[cTAKES](#) – from Mayo Clinic (Savova, 2010)

[Canary](#) – from Brigham & Women's Hospital (Malmasi, 2017)

[CLAMP](#) – from UT Houston (Soysal, 2018)

Commercial

[Nuance](#) – acquired by Microsoft

[Lingumatics](#)

M*Modal – acquired by 3M

Discern nCode – acquired by Cerner

[John Snow Labs](#) (Kocaman, 2021)

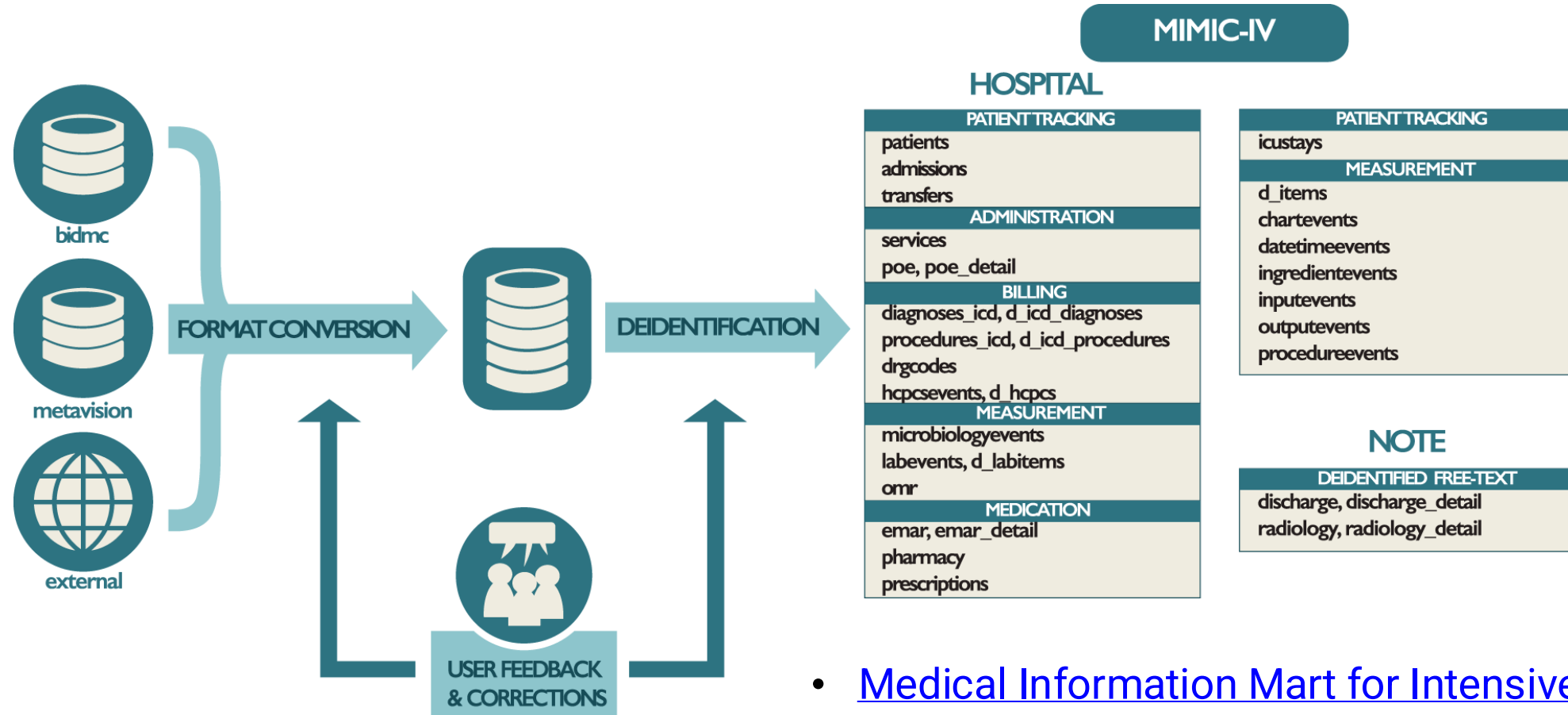
- [LangTest](#)

And of course LLMs from OpenAI, Google, Anthropic, Meta, Perplexity, etc.

- Enter PHI with care!



Another challenge – reliance on almost a single source of data



- [Medical Information Mart for Intensive Care \(MIMIC\)](#)
- (Johnson, 2016; Johnson, 2023)

Future directions for clinical NLP

NLP must move focus from “tasks as decisions” to “tasks as needs” for clinical use (Lederman, 2022)

For NLP in population health management and measurement, need (Tamang, 2023)

- Readiness of data and computing resources
- Organizational incentives to use and maintain systems
- Feasibility of implementation and continued monitoring

What will be role for LLMs, especially in light of

- Continual new models and updates of existing ones
- Proprietary models
- Focus on models over outcomes

