NLP for Health Ontologies

Aya Zirikly
09/03/2025

UMLS

- Metathesaurus
  - A comprehensive database that includes millions of biomedical and health-related concepts from source vocabularies:SNOMED CT, MeSH, LOINC, and ICD-10
- Semantic Network
  - Consistent categorization of UMLS concepts into semantic types and relationships
- SPECIALIST Lexicon:
  - A large lexical resource that includes biomedical and general English words, along with the lexical tools that aid natural language processing tasks
    - Lexical Entries: Each entry includes details about the word or term, its syntactic properties, and usage examples
    - Tools:
      - Tokenization
      - Stemming
      - Spelling correction

UMLS

Concepts

Definition: Concepts in UMLS represent unique meanings or ideas related to biomedical and health-related information

CUI (Concept Unique Identifier): Each concept is assigned a unique identifier, known as a CUI, which helps unify terms from various vocabularies that have the same meaning.

Concept Names

Synonyms: UMLS includes synonymous names and terms that refer to the same concept.

Preferred Names: Among the synonyms, a preferred name is designated for consistency across sources.

Source Vocabularies: Concept names come from multiple source vocabularies included in UMLS, such as SNOMED CT, LOINC, and MeSH.

UMLS Representation

Concept:

- Heart Attack
- CUI: C0018799

Concept Names

- Preferred Name: Myocardial Infarction
- Synonyms:
  - Heart Attack
  - MI
  - Cardiac Infarction
  - Acute Myocardial Infarction

Source Vocabularies

These names are integrated from various medical vocabularies. For example:

- SNOMED CT: Includes terms like "Acute Myocardial Infarction"
- MeSH: May use "Myocardial Infarction"
- ICD-10: Uses codes like "I21" for "Acute Myocardial Infarction"

What are these sources

SNOMED CT

MeSH

ICD-10

SNOMED CT

Systematized Nomenclature of Medicine

Key Features

- Comprehensive Coverage: wide range of medical terms covering diseases, procedures, symptoms, findings, pharmaceuticals, body structures, …
- Structured Format
- Interoperability: Facilitates the exchange of clinical health information across different healthcare systems by providing standardized terms
- Support for Decision Making: Enhances clinical decision support systems with standardized terminology, aiding in accurate diagnosis and treatment
- Multilingual healthcare terminology

SNOMED CT

Use in Healthcare

- Electronic Health Records (EHRs): it promotes consistent and precise recording of patient data
- Research and Data Analysis: Supports clinical research and healthcare data analysis by providing a standardized language for clinical concepts

Maintenance

- International Health Terminology Standards Development Organisation (IHTSDO)
    - Manages and updates SNOMED CT with latest medical knowledge and practices

Recent example papers

- Using Snomed to recognize and index chemical and drug mentions, 2019
  - https://aclanthology.org/D19-5718.pdf

- Clinical Text Classification to SNOMED CT Codes Using Transformers Trained on Linked Open Medical Ontologies, 2023
  - https://aclanthology.org/2023.ranlp-1.57/

- Translating SNOMED CT Terminology into a Minor Language
  - https://aclanthology.org/W14-1106.pdf

SNOMED & LLMs

- Incorporating SNOMED CT into LLM inputs
  - Using concept descriptions to expand training corpora

- Integrating SNOMED CT into additional fusion modules
- SNOMED CT as an external knowledge retriever during inference

MeSH

Medical Subject Headings, is a comprehensive controlled vocabulary used for indexing and cataloging biomedical literature

Key Features

- Hierarchy: MeSH terms are organized in a hierarchical structure, allowing for broad-to-narrow topic searches
- Controlled Vocabulary: Provides consistent terminology to index articles, facilitating efficient information retrieval
- Categories: Covers a wide range of biomedical topics, including diseases, chemicals, procedures, …

Uses

- Indexing: Primarily used by the National Library of Medicine (NLM) to index articles in PubMed
- Searching: Helps users perform precise literature searches by standardizing terms
- Cataloging

MeSH

Structure

- Descriptors: Main headings used in indexing and searching
- Qualifiers: Allow for more specificity, modifying descriptors to refine searches
- Entry Terms: Synonyms or alternative phrases that guide users to preferred descriptors

Maintenance

- National Library of Medicine (NLP), National Institutes of Health

MeSH

Term: Diabetes Mellitus

Descriptor: Diabetes Mellitus
Unique ID: D003920
Entry Terms:
- Diabetes
- Diabetes Disease

Tree Numbers: Indicate the position of the term in the MeSH hierarchy, e.g., C18.452.394.750 for endocrine system diseases.
Qualifiers: Can include subheadings like "diagnosis," "therapy," or "genetics" to refine searches.

MeSH & LLMs

- Include MeSH as Retrieval-Augmented Generation (RAG) system to reduce hallucination

- Me-LLaMA 13B and Me-LLaMA 70B
  - Encode comprehensive medical knowledge, along with their chat-optimized variants Me-LLaMA-13/70B-chat
  - Task: assign MeSH terms to a given piece of biomedical literature given title and abstract of an article
  - https://pmc.ncbi.nlm.nih.gov/articles/PMC11142305/

ICD-10

ICD-10, the International Classification of Diseases, 10th Revision, is a coding system used worldwide for various healthcare-related purposes.

Purpose

- Disease Classification: Provides a standardized way to code and classify diseases and health conditions
- Statistical Analysis
- Billing and Reimbursement: Used by healthcare providers for billing purposes and health insurance claims

ICD-10

Structure

- Codes: Alphanumeric codes, usually 3 to 7 characters, represent specific diagnoses or conditions (e.g., E11 for Type 2 Diabetes Mellitus).
- Chapters: Organized into chapters based on different body systems or conditions.

Uses

- Epidemiology: Supports tracking and analyzing disease prevalence and outcomes.
- Clinical: Helps in clinical documentation and communication.
- Research: provides consistent disease classification.

Maintenance

- WHO: World Health Organization
  - Global standardization and periodic updates

How they differ

SNOMED	MeSH	ICD-10
Comprehensive clinical terminology for EHRs	Indexing and cataloging biomedical literature	Classification of diseases and health conditions
Detailed clinical information	Biomedical and health-related topics	Diagnosis coding
Clinical settings, EHR documentation	Research databases like PubMed	Healthcare billing, epidemiology, statistics
Interoperability, clinical decision support	Literature searching, cataloging	Health data analysis, insurance claims

Summary of tasks related to Ontologies

- Medical Concept Normalization
- Entity Extraction
- Entity Classification

Next Week ToDos

- Add your suggested papers for next week to the sheet (extra points)
  - Ideally relevant to the topic, but can be different
  - Please add your name under Notes and any other special notes you have
- We will create a poll by Friday morning/Thursday evening on slack
- Results Friday evening/Saturday morning
- Updates for project and project teams?

Computing Resources

- What do we think about GW HPC?
- Hugging Face Transformers and Models
- Google T5 or Flan-T5
- Colab
  - Free 1 yeay subscription https://blog.google/outreach-initiatives/education/colab-higher-education/#:~:text=Google%20Colab%20has%20new%20features,features%20at%20the%20notebook%20level.
- Free LLM API
  - Google Gemini API
  - Mistral
  - Hugging Face Inference API
  - MAI-DS-R1 variant of DeepSeek
    - https://openrouter.ai/microsoft/mai-ds-r1:free
    - openRouter has other models to access
  - Llama models
  - Agentic AI
    - LangGraph
    - AutoGen
- Benchmark
  - https://huggingface.co/spaces/m42-health/MEDIC-Benchmark