MACHINE LEARNING FOR HEALTH

Biomedical NLP: Unlocking Intelligence in Clinical Text

Information extraction, contextual representation, and protection of the narrative

TECHNICAL HANDOUT · SESSION 09

Professor Flávio Luiz Seixas
Graduate Program in Computing

Academic teaching material based exclusively on the PDFs available in the project

Table of Contents

Introduction

Clinical texts record hypotheses, negation, temporality, relationships, and rationales that do not always fit into structured fields. Natural language processing transforms this narrative into computational representations or outputs for search, cohort identification, summarization, and decision support. [1]

This session organizes a clinical NLP pipeline and examines tasks, representations, evaluation, interoperability, and privacy. [1]

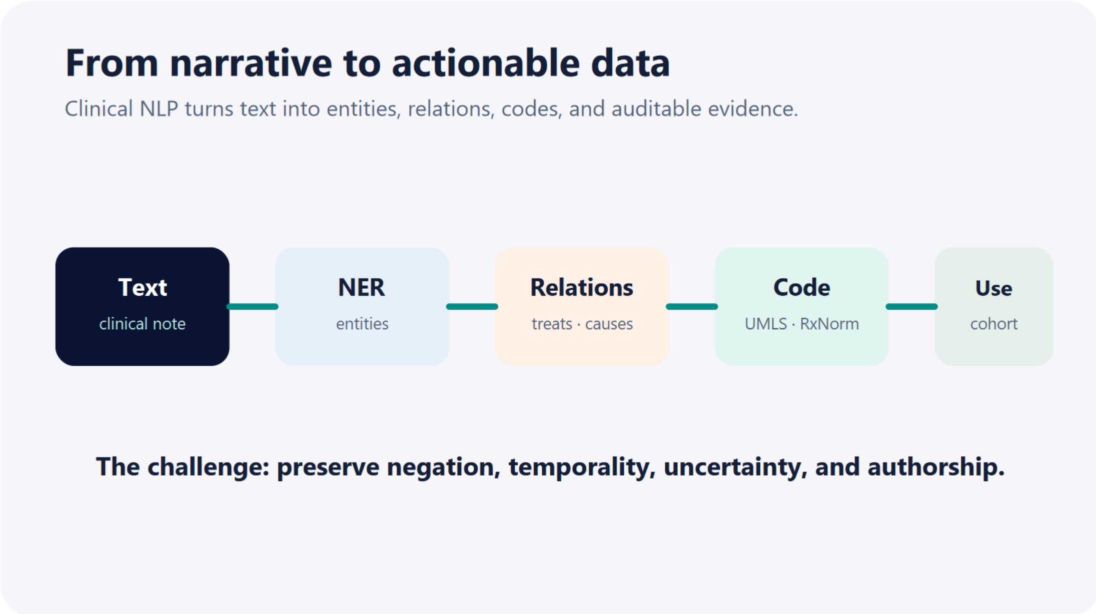

Figure 1 — Clinical NLP pipeline. Source: original illustration based on the course material and [1] and [2].

1. Texts and tasks

Biomedical sources include literature, reports, notes, and messages. The production context shapes vocabulary, abbreviations, structure, and errors. An application must define the unit of analysis and the output: document, sentence, entity, relationship, or template field [1].

Basic tasks include topic modeling, text classification, sequence labeling, relation extraction, and template filling. In a note, recognizing a medication is not enough if the application must also associate its dose, negation, temporality, or the patient's experience [1].

KEY CONCEPT
Biomedical sources include literature, reports, notes, and messages.

2. From text to representation

Classical pipelines combine segmentation, normalization, linguistic features, and classifiers. Distributed representations map textual units to vectors; contextual models change a representation according to neighboring words. The clinical domain requires handling specialized language and local usage [2].

BERT is an example of a Transformer-based model that uses bidirectional context in language-understanding tasks. Transfer reduces the amount of annotation needed but does not eliminate domain differences or the need for evaluation on the target dataset [2].

Before any model, a tokenizer converts text into units and identifiers. The vocabulary, tokenization method, and training corpus change how abbreviations, rare terms, and characters are decomposed. In clinical text, inspecting this boundary avoids assuming that visually similar words have equivalent representations [3].

3. From extraction to generative models

Clinical NLP is not synonymous with generation. Traditional systems can extract entities and relationships into predefined schemas, producing more constrained outputs; generative models accept flexible instructions and produce open-ended text. The choice should follow the task, tolerance for error, and need for verification rather than output fluency alone [4].

A practical architecture can combine components: extraction to locate facts, normalization to link them to concepts, retrieval to provide context, and generation to draft a synthesis. Each stage needs its own reference and metric because a plausible final answer may conceal an error produced early in the pipeline [1] [4].

4. Negation, temporality, and normalization

The same entity may appear as present, absent, hypothetical, past, or attributed to a family member. These modifiers change its interpretation. Normalization connects mentions to standardized concepts, enabling aggregation and links to terminology systems [1].

5. Evaluation, integration, and privacy

Appropriate performance depends on the purpose. Retrieving all candidates may prioritize sensitivity; automatically filling a field may require high precision. Reference annotation needs guidelines and agreement, and evaluation must reflect real documents and users [1].

NLP systems are integrated into larger applications and must comply with exchange standards. Notes also contain sensitive information that is difficult to remove completely; protection, access, and reidentification risk are part of the design [1].

Session summary

Texts and tasks: biomedical sources include literature, reports, notes, and messages.
From text to representation: classical pipelines combine segmentation, normalization, linguistic features, and classifiers.
From extraction to generative models: clinical nlp is not synonymous with generation.
Negation, temporality, and normalization: the same entity may appear as present, absent, hypothetical, past, or attributed to a family member.
Evaluation, integration, and privacy: appropriate performance depends on the purpose.

References

[1] Demner-Fushman, D.; Elhadad, N.; Friedman, C. Natural Language Processing for Health-Related Texts. In: Biomedical Informatics. Springer, 2021. External link

Local file: pdfs/2-Biomedical Informatics/VIII-Natural Language Processing for Health-Related Texts.pdf · PDF pages used: 1, 3, 8, 14, 20, 23.

[2] Xu, H.; Roberts, K. Natural Language Processing. In: Intelligent Systems in Medicine and Health. Springer, 2022. External link

Local file: pdfs/3-Intelligent Systems/VII-Natural Language Processing.pdf · PDF pages used: 1, 4, 8, 12, 16.

[3] Alammar, J.; Grootendorst, M. Hands-On Large Language Models: Language Understanding and Generation. O'Reilly Media, 2024.

Local file: pdfs/1-Machine Learning/Hands-On Large Language Models_ Language Understanding and -- Jay Alammar, Maarten Grootendorst -- 1, 2024 -- O'Reilly Media, Incorporated -- isbn13 9781098150969 -- 44bf8ae4fd6eb0c873190856f9ec6b35 -- Anna's Arch.pdf · PDF pages used: 40, 65, 66, 71, 128.

[4] Shaw, L.; Mahajan, S.; Upreti, K. (eds.). Natural Language Processing for Healthcare: The Rise of Intelligent Assistants. Academic Press, 2026.

Local file: pdfs/1-Machine Learning/Natural Language Processing for Healthcare.pdf · PDF pages used: 24, 30, 32, 33.