

APRENDIZADO DE MÁQUINA PARA SAÚDE

NLP Biomédico: desbloqueando a inteligência do texto clínico

Extração de informação, representação contextual e proteção da narrativa

APOSTILA TÉCNICA · ENCONTRO 09

Prof. Flávio Luiz Seixas
Pós-graduação em Computação

Sumário

Sumário	2
Introdução	3
1. Textos e tarefas.....	3
2. Do texto à representação	3
3. Do extrator ao modelo generativo	4
4. Negação, temporalidade e normalização	4
5. Avaliação, integração e privacidade.....	4
Síntese do encontro.....	4
Referências	5

Introdução

Textos clínicos registram hipóteses, negações, temporalidade, relações e justificativas que nem sempre cabem em campos estruturados. O processamento de linguagem natural transforma essa narrativa em representações ou saídas computacionais para busca, coorte, sumarização e apoio à decisão. [1]

O encontro organiza um pipeline de NLP clínico e examina tarefas, representações, avaliação, interoperabilidade e privacidade. [1]



Figura 1 — Pipeline de NLP clínico. Fonte: elaboração própria a partir do material da disciplina, com base em [1] e [2].

1. Textos e tarefas

Fontes biomédicas incluem literatura, relatórios, notas e mensagens. O contexto de produção condiciona vocabulário, abreviações, estrutura e erros. Uma aplicação deve definir a unidade de análise e a saída: documento, sentença, entidade, relação ou campo de um template [1].

Entre as tarefas básicas estão modelagem de tópicos, classificação de texto, rotulagem de sequência, extração de relações e preenchimento de templates. Em uma nota, reconhecer um medicamento não basta se a aplicação também precisa associar dose, negação, temporalidade ou experiência do paciente [1].

CONCEITO-CHAVE

Fontes biomédicas incluem literatura, relatórios, notas e mensagens.

2. Do texto à representação

Pipelines clássicos combinam segmentação, normalização, características linguísticas e classificadores. Representações distribuídas mapeiam unidades textuais para vetores; modelos contextuais alteram a representação conforme as palavras vizinhas. O domínio clínico requer lidar com linguagem especializada e uso local [2].

BERT é um exemplo de modelo baseado em Transformer que usa contexto bidirecional em tarefas de compreensão. A transferência reduz a quantidade de anotação necessária, mas não elimina diferença de domínio nem a necessidade de avaliação no conjunto-alvo [2].

Antes de qualquer modelo, um tokenizador converte o texto em unidades e identificadores. Vocabulário, método de tokenização e corpus de treinamento mudam como abreviações, termos raros e caracteres são decompostos. Em texto clínico, inspecionar essa fronteira evita supor que palavras visualmente semelhantes tenham representação equivalente [3].

3. Do extrator ao modelo generativo

PLN clínico não é sinônimo de geração. Sistemas tradicionais podem extrair entidades e relações em esquemas predefinidos, oferecendo saídas mais restritas; modelos generativos aceitam instruções flexíveis e produzem texto aberto. A escolha deve seguir a tarefa, a tolerância a erro e a necessidade de verificação, e não apenas a fluência da saída [4].

Uma arquitetura prática pode combinar componentes: extração para localizar fatos, normalização para ligá-los a conceitos, recuperação para trazer contexto e geração para redigir uma síntese. Cada etapa precisa de referência e métrica próprias, pois uma resposta final plausível pode ocultar erro produzido no início do pipeline [1] [4].

4. Negação, temporalidade e normalização

A mesma entidade pode aparecer como presente, ausente, hipotética, passada ou atribuída a familiar. Esses modificadores mudam a interpretação. Normalização associa menções a conceitos padronizados, permitindo agregação e conexão com sistemas terminológicos [1].

5. Avaliação, integração e privacidade

O desempenho adequado depende da finalidade. Recuperar todos os candidatos pode priorizar sensibilidade; preencher automaticamente um campo pode exigir alta precisão. Anotação de referência precisa de diretrizes e concordância, e a avaliação deve refletir documentos e usuários reais [1].

Sistemas de NLP integram aplicações maiores e devem respeitar padrões de intercâmbio. Notas também carregam informação sensível difícil de remover completamente; proteção, acesso e risco de reidentificação fazem parte do desenho [1].

Síntese do encontro

Textos e tarefas: fontes biomédicas incluem literatura, relatórios, notas e mensagens.

Do texto à representação: pipelines clássicos combinam segmentação, normalização, características linguísticas e classificadores.

Do extrator ao modelo generativo: pln clínico não é sinônimo de geração.

Negação, temporalidade e normalização: a mesma entidade pode aparecer como presente, ausente, hipotética, passada ou atribuída a familiar.

Avaliação, integração e privacidade: o desempenho adequado depende da finalidade.

Referências

- [1] Demner-Fushman, D.; Elhadad, N.; Friedman, C. Natural Language Processing for Health-Related Texts. In: Biomedical Informatics. Springer, 2021.
- [2] Xu, H.; Roberts, K. Natural Language Processing. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [3] Alamm, J.; Grootendorst, M. Hands-On Large Language Models: Language Understanding and Generation. O'Reilly Media, 2024.
- [4] Shaw, L.; Mahajan, S.; Upreti, K. (eds.). Natural Language Processing for Healthcare: The Rise of Intelligent Assistants. Academic Press, 2026.