PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

GIOVANI NÍCOLAS BETTONI

EXTRAÇÃO DE INFORMAÇÃO EM EVOLUÇÕES CLÍNICAS E INTEGRAÇÃO COM DADOS FARMACOGENÔMICOS

Porto Alegre
2022

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTRAÇÃO DE INFORMAÇÃO EM EVOLUÇÕES CLÍNICAS E INTEGRAÇÃO COM DADOS FARMACOGENÔMICOS

GIOVANI NÍCOLAS BETTONI

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Rafael Heitor Bordini

Porto Alegre
2022

Ficha Catalográfica

B565e Bettoni, Giovani Nícolas

Extração de informação em evoluções clínicas e integração com dados Farmacogenômicos / Giovani Nícolas Bettoni. – 2022.

94.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Rafael Heitor Bordini.

1. Reconhecimento de Entidades Nomeadas. 2. Modelos de Linguagem. 3. Interoperabilidade. I. Bordini, Rafael Heitor. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Loiva Duarte Novak CRB-10/2079

GIOVANI NÍCOLAS BETTONI

EXTRAÇÃO DE INFORMAÇÃO EM EVOLUÇÕES CLÍNICAS E INTEGRAÇÃO COM DADOS FARMACOGENÔMICOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 30 de Março de 2022.

BANCA EXAMINADORA:

Profª. Drª. Claudia Maria Cabral Moro Barra (PPGTS/PUCPR)

Profª. Drª. Isabel Harb Manssour (PPGCC/PUCRS)

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Este trabalho é dedicado àqueles que têm o privilégio de lutar pelo acesso igualitário à informação científica.

"*Aventurar-se causa ansiedade, mas deixar de arriscar-se é perder a si mesmo... E aventurar-se no sentido mais elevado é precisamente tomar consciência de si próprio.*"
(*Kierkegaard*)

AGRADECIMENTOS

Com o desenrolar de dois anos de trabalho, gostaria de agradecer, inicialmente a Profa. Dra. Renata Vieira, por ter me concedido a oportunidade de conhecer o meu orientador, Prof. Dr. Rafael H. Bordini, ao qual sou igualmente grato, pela confiança, ensinamentos, empenho e disponibilidade.

Nesse mesmo meio, não posso esquecer das contribuições dos meus colegas de curso, amigos e professores que contribuíram para a minha formação ao compartilharem seus conhecimentos. Agradeço a PUCRS e a CAPES/CNPQ por financiarem e assim permitirem a realização desta pesquisa.

Gostaria, também, de agradecer aos meus pais, pelo suporte, conselhos e palavras de conforto. E também, a Shaiane Rodrigues, pela troca de ideias e por ter me auxiliado nos momentos de dificuldade com muito zelo e paciência.

Por fim, gostaria de lembrar aqueles que já se foram, que marcam, e irão marcar a minha vida eternamente, em caráter e pensamentos principalmente.

Sem mais, sou eternamente grato à todos.

EXTRAÇÃO DE INFORMAÇÃO EM EVOLUÇÕES CLÍNICAS E INTEGRAÇÃO COM DADOS FARMACOGENÔMICOS

RESUMO

A Extração de Informação (EI) abrange uma série de tarefas de Processamento de Linguagem Natural (PLN). Entre elas, o Reconhecimento de Entidades Nomeadas (REN) é uma tarefa que busca identificar as Entidades Nomeadas de um texto, tais como nomes de pessoas, locais e organizações, classificado-as em um conjunto pré-definido de categorias. Nesta dissertação pretendemos utilizar técnicas e ferramentas de PLN para a tarefa de REN no domínio Biomédico em Português. Portanto, realizamos a construção de um corpus específico e propomos dois modelos baseados em redes neurais capazes de processar o texto incluído em evoluções clínicas: BERT e uma rede neural convolucional (CNN). Além disso, foi introduzido um novo mecanismo para incorporar conhecimento farmacogenômico que sirva como base para auxiliar na decisão clínica. Os resultados mostram uma melhoria das medidas do modelo BERT em comparação à CNN e demonstram que os modelos baseados em *Transformers* são promissores para o avanço do desempenho de métodos de extração de informação para entidades no domínio Farmacológico em Português. O Reconhecimento de Entidades Nomeadas em evoluções clínicas está ganhando popularidade por melhorar os projetos de extração clínica. Este estudo permitiu à comunidade que trabalha com PLN, no contexto clínico, obter uma análise formal dessa tarefa, incluindo as formas mais bem-sucedidas de realizá-la.

Palavras-Chave: Reconhecimento de Entidades Nomeadas, Modelos de Linguagem, Interoperabilidade.

EXTRACTION OF INFORMATION ON CLINICAL NOTES AND INTEGRATION WITH PHARMACOGENOMICS DATA

ABSTRACT

Information Extraction (IE) covers a number of Natural Language Processing (NLP) tasks. Named Entity Recognition (NER) is a task that seeks to identify the Named Entities of a text, such as names of people, places, and organizations, classifying them in a predefined set of categories. This dissertation intends to use NLP techniques and tools for the REN task in the Biomedical domain in Portuguese. Thus, we build a specific corpus and propose two models defined in neural networks able to process the text included in clinical evolutions: BERT and a convolutional neural network (CNN). In addition, a new mechanism has been introduced to incorporate pharmacogenomic knowledge that serves as a basis for aiding clinical decisions. The results show an improvement in the measures of the BERT model compared to CNN and demonstrate that Transformers-based models are promising for advancing the performance of information extraction methods for entities in the Pharmacologic domain in Portuguese. Recognition of Named Entities in clinical evolutions is gaining popularity for improving clinical extraction projects. This study allowed the community working with NLP, in the clinical context, to obtain a formal analysis of this task, including the most successful ways of performing it.

Keywords: Named Entity Recognition, Language Models, Interoperability.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANVISA – Agência Nacional de Vigilância Sanitária

API – Interface de Programação de Aplicações

AVC – Acidente Vascular Cerebral

BERT – *Bidirectional Encoder Representations from Transformers*

BILSTM – *Bidirectional Long Short-Term Memory*

BIO – *Begin-Inside-Outside*

BOW – *Bag-Of-Words*

CAS – *Chemical Abstracts Service*

CDS – Suporte à Decisão Clínica

CIAP – Classificação Internacional de Atenção Primária

CID – Classificação Internacional de Doenças

CLM – Modelagem de Linguagem Causal

CNN – Redes Neurais Convolucionais

CONLL – Conferência sobre Aprendizagem de Linguagem Natural Computacional

CRF – Campos Aleatórios Condicionais

DCB – Denominação Comum Brasileira

DCI – Denominação Comum Internacional

EAMS – Eventos Adversos Relacionados a Medicamentos

ELMO – *Embeddings from Language Models*

ETL – *Extract-Transform-Load*

EI – Extração de Informação

FNN – *Feedforward Neural Network*

HPO – *Human Phenotype Ontology*

JSON – *JavaScript Object Notation*

LOINC – *Logical Observation Identifiers Names and Codes*

LSTM – *Long Short-Term Memory*

MAE – Erro Médio Absoluto

MAPE – Erro Percentual Médio Absoluto

MESH – *Medical Subject Headings*

ML – *Machine Learning*

MLM – *Masked Language Modeling*

MLP – *Multi-Layer Perceptron*

MSE – Erro Médio Quadrático

NGS – Sequenciamento de Nova Geração

NLM – National Library of Medicine

NSP – *Next Sentence Prediction*

OOV – *Out-Of-Vocabulary*

PGX – Farmacogenômica

POS – *Part-Of-Speech*

PLN – Processamento de Linguagem Natural

RELU – Unidade Linear Retificada

REN – Reconhecimento de Entidades Nomeadas

RES – Registro Eletrônico em Saúde

RI – Recuperação de Informações

RL – Regressão Logística

RNS – Redes Neurais

RNAS – Redes Neurais Artificiais

RNR – Rede Neural Recorrente

RNRS – Redes Neurais Recorrentes

SGD – *Stochastic Gradient Descent*

SLP – *Single-Layer Perceptron*

SNOMED-CT – *Systematized Nomenclature of Medicine – Clinical Terms*

S-RES – Sistemas de Registro Eletrônico em Saúde

SVM – Máquinas de Vetores de Suporte

TAO – Ontologia TAMBIS

TLM – Modelagem de Linguagem de Tradução

UMLS – *Unified Medical Language System*

XML – *Extensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

Um dos principais objetivos da mineração de textos clínicos é a possibilidade de processar e analisar os grandes volumes de informações textuais contidas em prontuários clínicos. Na mineração de texto clínico, o Reconhecimento de Entidades Nomeadas (REN) de medicamentos e produtos químicos é uma tarefa importante no campo de Processamento de Linguagem Natural (PLN) usada para extrair conhecimento significativo sobre substâncias químicas e medicamentos das evoluções clínicas. O objetivo da REN Biomédica é identificar pedaços de texto que se referem a entidades específicas de interesse, como nomes de medicamentos, proteínas, sintomas e doenças, relatando aos especialistas uma grande quantidade de conhecimento incorporado nos dados textuais. Por meio desse tratamento das informações, procuramos responder a questões como: quais fármacos estão em uso por determinado paciente? Determinado paciente já utilizou algum fármaco com informação farmacogenômica? Essas questões podem parecer bastante simples para alguns profissionais médicos, mas se tornam extremamente complexas quando gerenciadas automaticamente por sistemas computacionais.

Embora existam informações valiosas transmitidas em textos biomédicos, clínicos e de saúde, elas não estão em um formato diretamente passível de processamento posterior. Esses textos são difíceis de processar de forma confiável devido às características inerentes e à variabilidade da linguagem [25]. Enquanto na maioria dos aplicativos automatizados, dados estruturados e padronizados estão prontamente disponíveis para processamento, há uma quantidade significativa de trabalho manual atualmente dedicado ao mapeamento de informações textuais para representações codificadas em biomedicina e assistência médica: codificadores profissionais atribuem códigos de cobrança correspondentes a diagnósticos e procedimentos para admissões hospitalares com base em resumos de alta e informações de admissão; indexadores da *National Library of Medicine* atribuem termos MeSH (Medical Subject Headings) para representar os principais tópicos de artigos científicos; e curadores de banco de dados extraem informações genômicas e fenotípicas sobre organismos da literatura. Devido à enorme quantidade de informações textuais nos domínios da saúde, o trabalho manual é custoso, demorado e impossível de manter atualizado. Um dos objetivos do Processamento de Linguagem Natural (PLN) é facilitar essas tarefas, permitindo o uso de métodos automatizados com alta validade e confiabilidade [85, 36].

Muitos pesquisadores na área de PLN se concentram na área de Extração de Informação (EI) no domínio Biomédico para enfrentar esses desafios. Os sistemas de EI usam textos semi-estruturados em linguagem natural como entrada e produzem informações estruturadas especificadas por determinados critérios e que são relevantes para uma aplicação específica. Dependendo das diferentes entradas dos sistemas de EI e das saídas esperadas, muitas subtarefas podem ser definidas, como o REN.

O termo Entidade Nomeada foi estabelecido em 1996, na 6ª Conferência de Entendimento de Mensagens (MUC-6) [41], para se referir a "identificadores únicos de entidades". Em linhas gerais, a tarefa de REN consiste em localizar e classificar partes do texto em categorias pré-definidas como lugares, pessoas, organizações, expressões de tempo e quantidades. No entanto, no domínio Biomédico, as entidades importantes incluídas nos documentos não se limitam às mencionadas acima. Nesse caso específico, é necessário reconhecer alguns tipos especiais de entidades nomeadas, como doenças, procedimentos, tratamentos, medicamentos, entre outros.

O REN não serve apenas como uma tarefa incluída na EI, mas também desempenha um papel essencial em uma variedade de aplicações PLN, como recuperação de informações, resumo automático de texto ou resposta a perguntas [66, 19]. Além disso, o Reconhecimento de Entidades Biomédicas em um texto pode ser um ponto de partida para a posterior extração de relações entre entidades, permitindo que esses conceitos sejam representados de alguma forma coerente e padronizada.

Atualmente, existem diversos métodos para recuperar informação em diferentes idiomas. Entretanto, neste trabalho, focamos na extração de informação de narrativas clínicas em Português, mais especificamente, na tarefa de REN. Mesmo que o Inglês seja hoje a língua mais falada, o mundo é multilíngue. A língua Portuguesa tem mais de 250 milhões de falantes nativos [1] e hoje em dia há um interesse mundial em processar textos médicos neste idioma e não há nada específico. Com este estudo, pretendemos avançar na tarefa de REN farmacológica nesta relevante linguagem e assim responder às questões acima referidas.

Para realizar este estudo, propomos uma metodologia baseada nas mais recentes técnicas de *Transfer Learning*. Com esta abordagem, pretendemos atingir o objetivo final desejado: reconhecer com precisão as entidades farmacológicas em evoluções clínicas.

Nas secções seguintes, mostramos as motivações que nos levaram a abordar a tarefa de REN na área Farmacológica. Descrevemos também os objetivos e metas que pretendemos alcançar com esta dissertação.

1.1 Motivação

A literatura biomédica é o principal meio que os pesquisadores usam para compartilhar suas descobertas, principalmente na forma de artigos, patentes e outros tipos de relatórios escritos [45]. Ao longo dos anos, o reconhecimento das entidades biomédicas motivou a comunidade científica a continuar desenvolvendo sistemas automáticos para fa-

[1] Língua Portuguesa na Wikipédia

cilitar a extração do conhecimento médico. O REN é uma tarefa difícil de resolver que pode ajudar em muitos outros sistemas médicos como os apresentados abaixo:

- **Apoio à decisão clínica:** o suporte à decisão clínica enfatiza a capacidade de produzir relatórios baseados em evidências sobre os serviços de saúde diários para auxiliar os especialistas em suas decisões e ações [24]. Essas informações podem ser usadas por métodos PLN que desenvolvem aplicativos baseados em evidências que detectam alertas precoces para o monitoramento de distúrbios e o desenvolvimento de atendimento personalizado ao paciente [125, 93]. Os sistemas automatizados de REN podem fornecer resultados em tempo real, o que significa que entidades como doenças podem ser detectadas imediatamente. Essa evidência pode ser usada para ajudar os profissionais a identificar problemas de saúde emergentes, por exemplo, para alertá-los sobre a presença de certos achados inesperados [135].

- **Base para outras tarefas de PLN:** o reconhecimento de entidades biomédicas serve como base para muitas outras áreas cruciais do gerenciamento de informações, como tarefas de classificação, resposta a perguntas, recuperação de informações e resumo de texto. Por exemplo, o uso do REN torna-se importante para análise do texto clínico e obtenção das tags mais relevantes em cada laudo, permitindo a classificação dos documentos. Relativamente à tarefa de responder a perguntas, é uma prática comum utilizar sistemas de REN para expandir a consulta utilizando sinônimos de entidades, descrições e siglas para obter melhores resultados. Outro importante desafio proposto pelo REN é atribuir a cada entidade um identificador único numa base de dados ou vocabulário controlado. Esse processo é conhecido como normalização de entidades em que, uma vez identificada uma entidade biomédica, ela pode ser compartilhada de forma padronizada com outros sistemas.

- **Representação da Entidade:** na tarefa REN, palavras diferentes podem ter significados semelhantes. Esse problema é causado pelas várias maneiras pelas quais uma entidade específica pode ser representada e escrita. Por exemplo, adriamicina, doxorrubicina ou hidroxildaunorrubicina referem-se ao mesmo medicamento amplamente utilizado na quimioterapia do câncer. Outra consideração é que as entidades aparecem como siglas ou suas descrições, e.g., "doença pulmonar obstrutiva crônica"e "DPOC"são referidas como a mesma doença (doença pulmonar inflamatória crônica que causa obstrução do fluxo de ar dos pulmões). Por outro lado, uma sigla nem sempre tem uma descrição única, pode ser interpretada como duas entidades diferentes dependendo do contexto. Por exemplo, em português, PCR pode ser referido como Parada Cardiorrespiratória ou Proteína C-Reativa. Finalmente, como vemos nos exemplos, as entidades biológicas também podem ter nomes com várias palavras, de modo que o problema é adicionalmente complicado pela necessidade de determinar os limites dos nomes e resolver a sobreposição de nomes candidatos.

- **Extrair informações estruturadas:** a tarefa de REN biomédica facilita os profissionais da saúde na estruturação de relatórios contribuindo para soluções como fornecer um resumo das condições do paciente ou servir como ferramenta para organizar a documentação do processo de tomada de decisão do corpo clínico de profissionais da saúde, desenvolvimento de planos e resultados do paciente.

Como podemos ver, são muitos os problemas e dificuldades que podem ser resolvidos indiretamente avançando na resolução da tarefa de REN no domínio da saúde. Portanto, a principal motivação desta dissertação é extrair conhecimento relevante de anotações clínicas que possam ser convenientes e úteis para problemas em aberto na área médica.

1.2 Objetivos

O principal objetivo desta dissertação centra-se no emprego de técnicas e ferramentas de PLN para a tarefa de REN no domínio Biomédico em Português. Este objetivo geral pode ser detalhado através dos seguintes objetivos específicos:

- Desenvolver um novo modelo usando modelos de linguagem pré-treinados;
- Avaliar o desempenho do método proposto no problema REN usando o domínio Farmacológico;
- Realizar uma análise de resultados comparando nosso sistema com o desempenho de outros trabalhos;
- Desenvolver um mecanismo de integração com dados Farmacogenômicos.

1.3 Esboço da Dissertação

Esta dissertação está organizada em 6 capítulos e um apêndice. Este primeiro capítulo contém uma introdução explicando a motivação e os objetivos que nos levaram a realizar o estudo. O restante desta dissertação está dividido em diferentes capítulos e está organizado da seguinte forma.

O Capítulo 2 apresenta uma visão geral sobre o contexto da interoperabilidade na saúde e das metodologias baseadas em ML comumente usadas na tarefa de REN e que são necessárias para entender as partes posteriores desta dissertação. Especificamente, este capítulo analisa as abordagens de ML realizadas para resolver o problema de REN no domínio Biomédico. Essas técnicas foram utilizadas ao longo desta dissertação e foram

divididas em duas categorias: métodos supervisionados e não supervisionados. Uma vez que o interesse desta dissertação também está na representação de palavras, este capítulo detalha a revisão dos métodos existentes de representação de palavras até o momento.

O Capítulo 3 resume o trabalho anterior sobre tarefas PLN baseadas em ML no domínio Biomédico e mostra uma revisão da literatura da tarefa REN no que diz respeito aos estudos atuais do estado da arte. Por fim, apresenta recursos de conhecimento amplamente utilizados pela comunidade de PLN no domínio Biomédico (BioNLP).

O Capítulo 4 descreve o modelo proposto para resolver o problema de extração de entidades biomédicas. Após uma extensa revisão de metodologias aplicadas anteriormente, propomos, inicialmente, uma abordagem baseada em uma rede neural convolucional (CNN). Seguindo essa ideia, nossa abordagem propõe a avaliação do BERT em nossos corpora anotados. Por fim, este capítulo mostra como podemos combinar os dados do paciente com bases de conhecimento farmacogenômicas.

O Capítulo 5 apresenta a experimentação realizada com a abordagem proposta anteriormente.

O Capítulo 6 contém nossa conclusão, onde resumimos nossas descobertas e principais contribuições. Além disso, este capítulo fornece uma perspectiva para o futuro.

2. PRESSUPOSTOS TEÓRICOS

Este capítulo fornece conhecimentos básicos para estabelecer a base para os capítulos seguintes. Faremos uma breve introdução à interoperabilidade e seus princípios gerais no contexto da saúde, em seguida, detalharemos as abordagens mais frequentes usadas em aprendizado de máquina no âmbito da inteligência artificial, como representamos palavras computacionalmente, o que é o BERT e como avaliamos a tarefa de REN e, por fim, quais os recursos de conhecimento biomédico existentes.

2.1 Interoperabilidade no Contexto da Saúde

O processo de atendimento ao paciente, que pode ser variado e complicado, também inclui inúmeros processos que podem ser melhorados com padronização. Um sistema de admissão hospitalar registra que um paciente tem o diagnóstico de diabetes mellitus, um sistema de farmácia registra que o paciente recebeu Gentamicina, um sistema de laboratório registra que o paciente teve certos resultados em testes de função renal e um sistema de radiologia registra que um médico solicitou um exame de raio-X para o paciente que requer corante de iodo intravenoso [58, 56]. Iniciamos, usando o termo no contexto da definição fornecida pelo IEEE 1990 como:

> *"A interoperabilidade é alcançada quando dois ou mais sistemas ou componentes dos sistemas podem trocar dados/informações e utilizar os dados/informações sendo trocados."* [34]

Atualmente, essa definição fornecida pela IEEE não é suficiente para abranger o papel da interoperabilidade. Como as tarefas descritas requerem coordenação de sistemas, são necessários métodos para transferir informações de um sistema para outro. Tais transferências eram tradicionalmente realizadas por meio de interfaces ponto a ponto personalizadas, mas essa técnica tornou-se impraticável à medida que o número de sistemas e as permutações resultantes de conexões necessárias aumentaram [44]. Uma abordagem atual para resolver o problema de múltiplas interfaces é através do desenvolvimento de padrões de mensagens. Tais mensagens devem depender da preexistência de padrões para identificação do paciente e codificação de dados.

Padrões são geralmente necessários quando a diversidade excessiva cria ineficiências ou impede a eficácia. Tradicionalmente, o ambiente de saúde consiste em um conjunto de unidades organizacionalmente independentes, pouco conectadas [145]. Os pacientes recebem atendimento em ambientes de atenção primária, secundária e terciária, com pouca comunicação bidirecional e coordenação entre os serviços. Os pacientes

são atendidos por um ou mais médicos de primeira linha, bem como por especialistas. Há pouca coordenação e compartilhamento de dados entre o atendimento hospitalar e o atendimento ambulatorial. Exemplificando no contexto da saúde, existem algumas modalidades para rotular interoperabilidade [12], tais como:

- **Sintática**, e.g., uso de um padrão na codificação e representação das informações;
- **Funcional**, e.g., é a capacidade de trocar informações de forma confiável e sem erros; ou
- **Semântica**, e.g., traduzindo o significado das informações de diferentes origens, através da representação de conhecimento sobre o mundo ou alguma parte deste.

Em outras palavras, a definição incorpora tanto a interoperabilidade funcional, ou seja, o intercâmbio de informações, quanto a interoperabilidade semântica, que é a capacidade de utilizar as informações que estão sendo trocadas.

Nos últimos anos, a crescente solicitação de Sistemas de Registro Eletrônico em Saúde (S-RES) levou os diferentes provedores de saúde a adotarem rapidamente várias soluções. Por outro lado, a adoção de diversos padrões em diferentes momentos, com diferentes requisitos de leis regionais ou nacionais, as necessidades específicas dos provedores de saúde e outros motivos, tornaram esses sistemas incapazes de interoperar [151, 18]. A falta de interoperabilidade entre esses sistemas resulta na diminuição da qualidade do atendimento ao paciente e no desperdício de recursos financeiros.

Para garantir a interoperabilidade dos S-RES e explorar suas informações, muitas abordagens foram propostas e implementadas [22, 23, 38, 92], definindo arquiteturas gerais ou específicas capazes de resolver os problemas relacionados. Com foco no atual cenário brasileiro, a definição de infraestrutura para a interoperabilidade dos S-RES não apenas resolveu o problema de troca de dados entre diferentes regiões, mas também forneceu aos médicos, pacientes, pesquisadores e formuladores de políticas uma enorme fonte de informações, formada pelos dados e informações que é possível extrair dos S-RES agregados. Para poder explorar ainda mais este tipo de dados, que incluem dados médicos, clínicos e sociais, é necessário definir instrumentos inovadores que permitam a Extração de Informação (EI).

Essa tarefa não é trivial, não apenas pelos problemas relacionados à dispersão, heterogeneidade e tamanho dos dados, mas pela necessidade de métodos avançados de EI, incluindo Processamento de Linguagem Natural (PLN), capazes de identificar e classificar os dados entre os documentos de texto livre não estruturados ou semiestruturados.

A Farmacogenômica (PGx) é um campo amplo que abrange dezenas de áreas terapêuticas, milhares de variantes genéticas e centenas de medicamentos [146, 17]. Avanços recentes em métodos para medir moléculas orgânicas e fenótipos, descrever estados clínicos e raciocinar em dados federados oferecem um conjunto cada vez mais preciso de tecnologias para descoberta da farmacogenômica como uma tradução clínica.

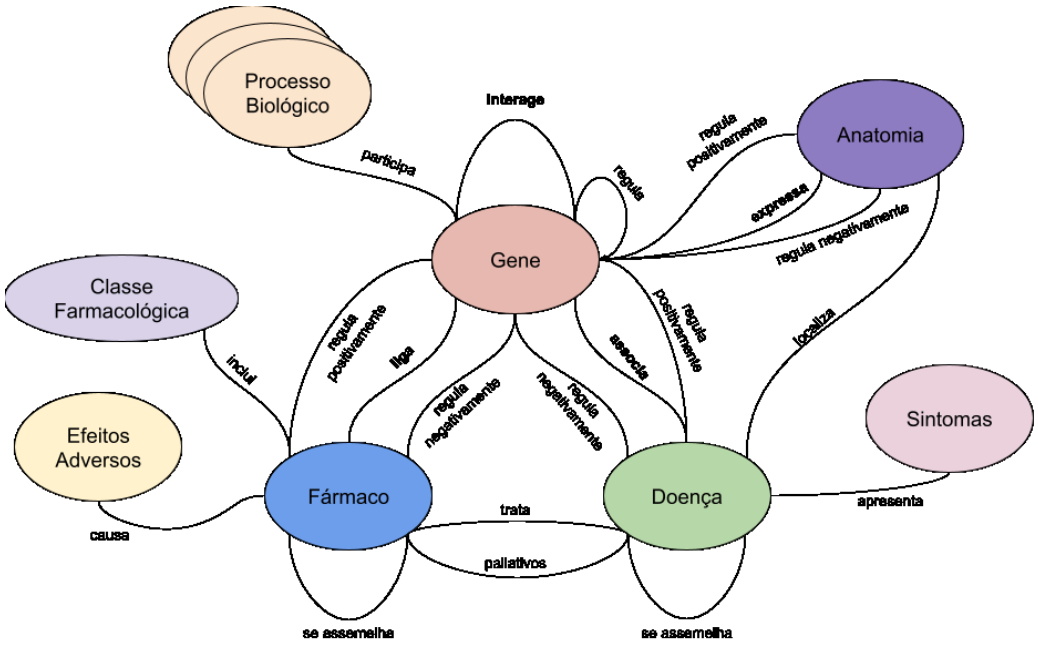

Figura 2.1 – Esquema onde entidades biomédicas da tríade Farmacogenômica são representadas como nós e as relações entre elas são representadas como bordas. Adaptado de Nicholson et al. [108]

O estudo dos efeitos adversos a medicamentos em pacientes devido à variação genética [68] é central para um melhor entendimento dos sistemas biológicos. Diariamente, profissionais da saúde tratam pacientes que respondem de maneiras diferentes a um mesmo medicamento. A Farmacogenômica, e mais amplamente medicina de precisão, compreende tecnologias para agrupar a população de pacientes e para aliviar a carga de reações adversas a medicamentos (hipoteticamente retratadas na Figura 2.2). Diz-se que até 95% das respostas à medicação são atribuídas à própria composição genética [65]. Com isso em mente, tem havido um grande interesse investido no estabelecimento de associações farmacogenômicas concretas e no desenvolvimento de ferramentas preditivas baseadas na Farmacogenômica [170].

Para entender as condições e pré-condições em que um medicamento pode causar reações adversas, os Sistemas de Registro Eletrônico em Saúde (S-RES) oferecem uma maneira indiscutivelmente mais eficiente de armazenar e compartilhar dados médicos para uso, principalmente, pelas equipes médicas.

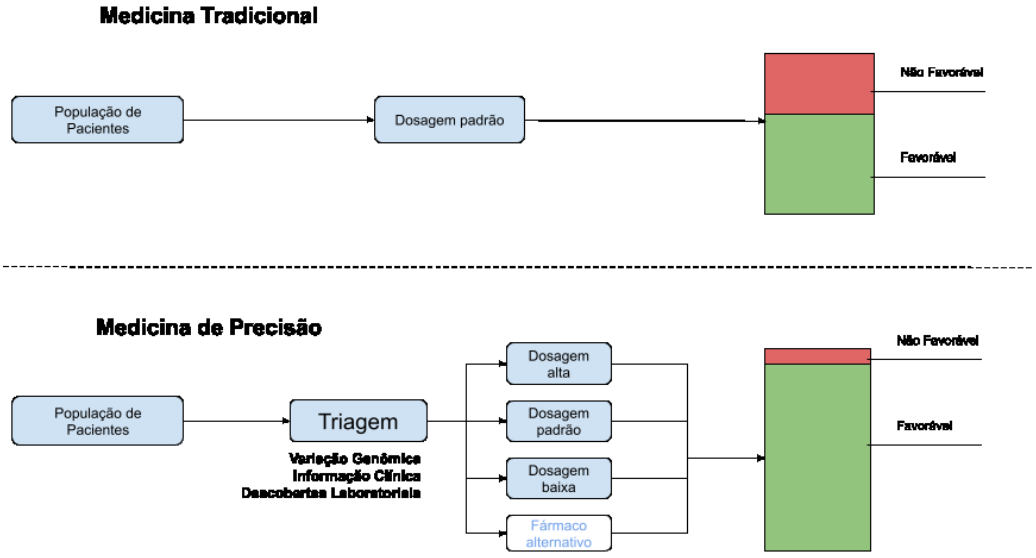

Figura 2.2 – Comparação entre a resposta de um paciente à dosagem de um mesmo fármaco usando a abordagem tradicional e a Medicina de Precisão. Adaptado de Alessandrini et al. [1]

Com algumas limitações relacionadas à interoperabilidade, os S-RES permitem que os dados genômicos sejam incorporados à continuidade do cuidado, à medida que os pacientes fazem a transição entre os ambientes de cuidado nas diferentes organizações de saúde [47]. No entanto, os S-RES são de natureza não estruturada e exigem processamento adicional para extrair informações estruturadas, como entidades nomeadas de interesse.

Na Figura 2.3 podemos observar com detalhes, um ciclo de implementação da Medicina Genômica. No canto esquerdo, observamos um S-RES que armazena dados demográficos, dados de admissão/alta, notas clínicas (evoluções) dos profissionais de saúde e informações inseridas sobre o paciente, tais como, procedimentos, medicamentos, resultados laboratoriais, relatórios de histopatologia, relatórios de radiologia, entre outros.

O processo *Extract-Transform-Load* (ETL) agrega e transforma dados para armazenamento lendo os dados S-RES desejados, convertendo-os em um formato utilizável e, em seguida, gravando-os em um banco de dados relacional. Para permitir a integração, compartilhamento e recuperação de tais dados, padrões como o *Unified Medical Language System* (UMLS), *Logical Observation Identifiers Names and Codes* (LOINC) [102], *Systematized Nomenclature of Medicine – Clinical Terms* (SNOMED-CT) [28]. No cenário brasileiro, já observamos o uso de alguns vocabulários e terminologias clínicas, como a Classificação Internacional de Doenças v.10 (CID-10), a Classificação Internacional de Atenção Primária v.2 (CIAP-2) [134]. A SNOMED-CT está em fase de tradução para incorporação, com a possibilidade de ser utilizada de forma mais abrangente. Contudo, esse processo pode ser um processo demorado [99].

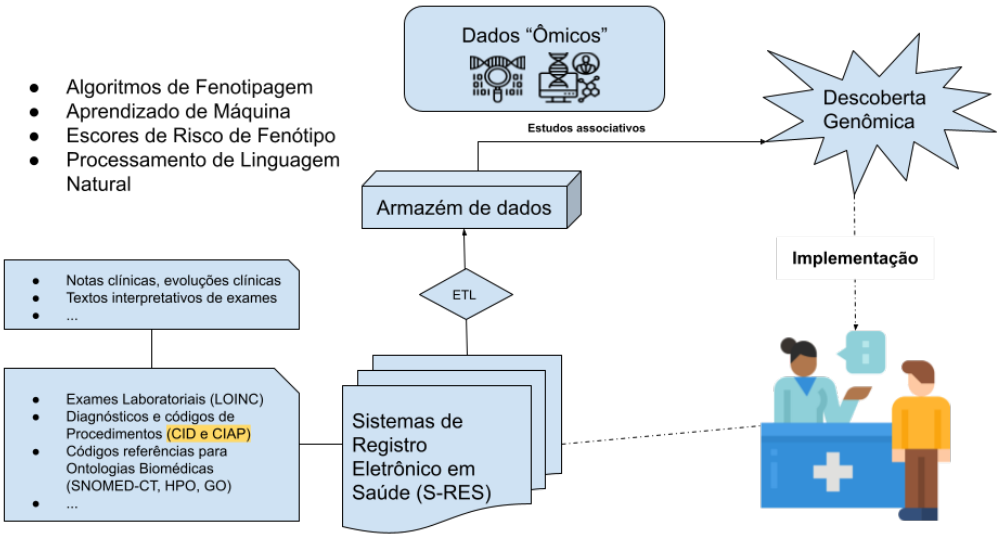

Figura 2.3 – Usando os dados de S-RES para descoberta genômica. Os dados de S-RES podem ser usados para estudar a base genética de doenças comuns e raras, identificar subfenótipos de doenças, avaliar a patogenicidade de novas variantes genômicas, investigar a pleiotropia e montar rapidamente coortes para ensaios clínicos de medicina genômica. Adaptada de Safarova et al. [130]

As ferramentas e padrões para o compartilhamento de informações computacionais têm avançado em conjunto com a tecnologia para medir e descrever informações farmacogenômicas. Nos últimos anos, o debate relacionado à aplicação da bioinformática na rotina clínica tem sido cada vez mais frequente. A integração de dados oriundos de Sequenciamento de Nova Geração (NGS) em conjunto com os mais diversos dados clínicos descritos nos S-RES podem auxiliar diretamente as decisões clínicas [143].

O conhecimento codificado em ontologias biomédicas desempenha um papel vital no desenvolvimento de sistemas de redes neurais profundas, fornecendo informações semânticas e de ancestralidade para entidades, como genes, proteínas, fenótipos e doenças. Pode-se também adicionar medidas de similaridade semântica como uma camada de informação adicional [6]. O *Human Phenotype Ontology* (HPO) foi desenvolvido para descrever anormalidades associadas a doenças humanas [72, 127]. Assim, rapidamente se tornou um padrão usado por vários consórcios de genética clínica, e ferramentas foram desenvolvidas para permitir que médicos e pacientes registrem dados fenotípicos de maneira eficiente, integrem-se às informações genômicas e gerem um diagnóstico diferencial.

Depois que os dados Farmacogenômicos forem estruturados de uma maneira que facilite a compreensão concisa por médicos, cientistas e pacientes, um método de acesso a esses dados é fundamental para sua utilidade. Neste momento, verificamos que a Interface de Programação de Aplicações (API) pode ser usada para padronizar a(s) entrada(s) e saída(s) esperadas para qualquer recurso baseado em conhecimento. Uma API apresenta

uma camada de abstração para dados e/ou funções que um serviço computacional gostaria de expor.

2.2 Abordagens de Aprendizado de Máquina

Segundo McCarthy [101], podemos definir a Inteligência Artificial da seguinte forma:

> *"Inteligência Artificial (IA) é a ciência e engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes. Ela está relacionada com a tarefa semelhante de usar computadores para entender a inteligência humana."*

Atualmente, a IA se tornou mais popular graças ao aumento do volume de dados, algoritmos avançados e melhorias no poder computacional e de armazenamento. Especificamente, a IA na área da saúde levou a um rápido avanço na medicina digital em várias especialidades clínicas, incluindo oncologia [106, 31], radiologia [117, 174] e cardiologia [167, 161]. Como uma quantidade substancial de informações clínicas mais relevantes está incorporada em dados não estruturados [141], o PLN desempenha um papel essencial na extração de informações valiosas que podem beneficiar a tomada de decisões, estruturação de relatórios, classificação de relatórios e reconhecimento de entidades, entre outros.

Para aplicações mais complexas de PLN, os sistemas são baseados em modelos de ML para melhorar sua compreensão da linguagem humana.

Nas últimas décadas, muitos sistemas automáticos de REN foram desenvolvidos e usados para identificar e categorizar entidades biomédicas usando abordagens de ML. Essas abordagens podem ser organizadas em diferentes categorias. Por exemplo, considerando se o algoritmo é treinado com dados rotulados ou não, podemos classificar tais algoritmos em aprendizado supervisionado e aprendizado não supervisionado.

No decorrer desta dissertação, estudamos algumas das abordagens incluídas nas categorias anteriores e as resumimos na Figura 2.4. Por um lado, dentro dos métodos não supervisionados, investigamos métodos baseados em regras e métodos baseados em dicionários. Por outro lado, em métodos de aprendizado supervisionado, fizemos uma distinção entre algoritmos tradicionais, redes neurais e modelos baseados em transformadores.

2.2.1 Aprendizado não supervisionado

Nesta seção, pretendemos fazer um levantamento de duas das técnicas de aprendizado não supervisionado usadas para tarefas de REN no domínio biomédico. O aprendi-

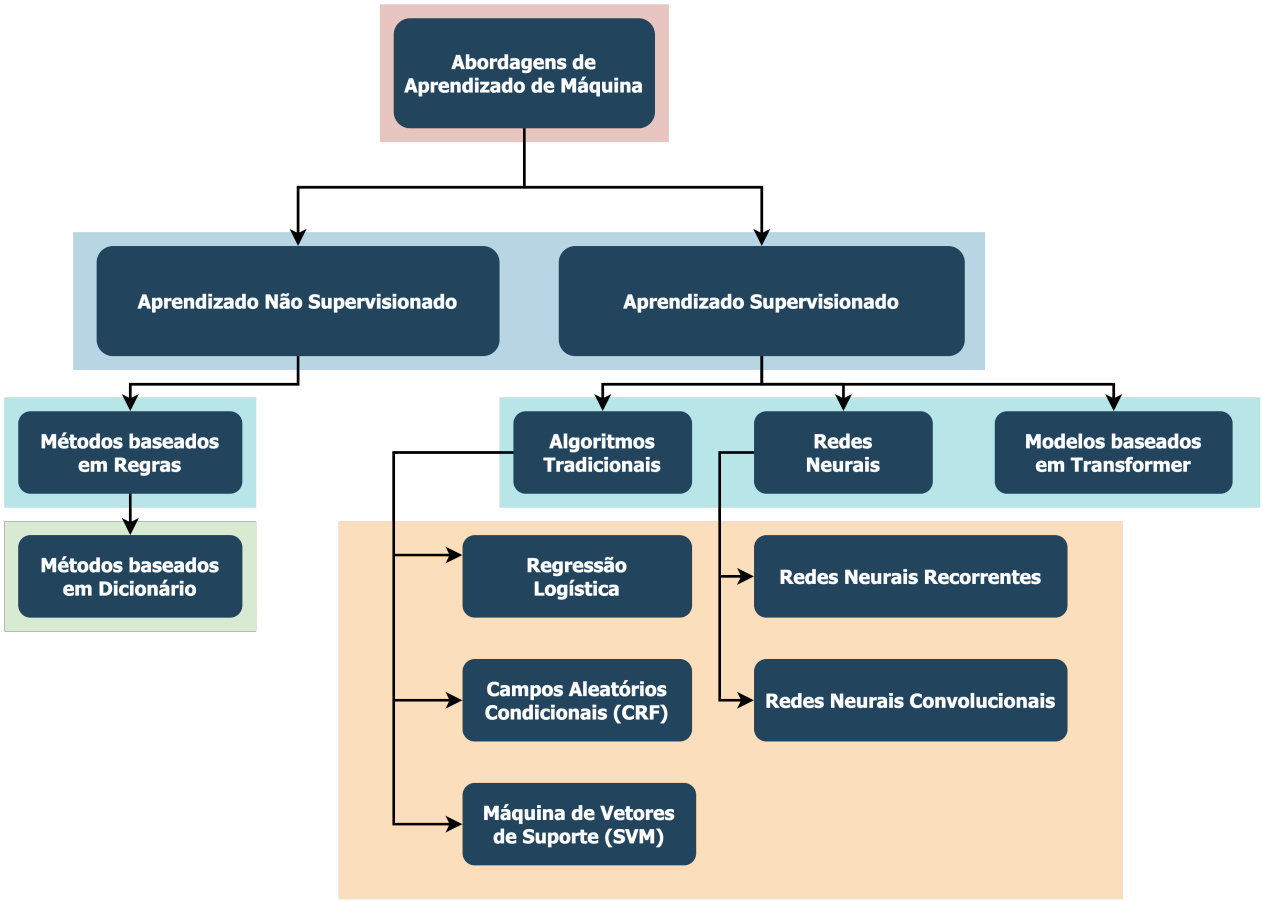

Figura 2.4 – Classificação dos métodos de aprendizado de máquina estudados ao longo desta pesquisa

zado não supervisionado é uma técnica de aprendizado de máquina na qual não é necessário supervisionar o modelo, ou seja, ele aprende padrões a partir de dados não rotulados.

Métodos baseados em regras

Os sistemas baseados em regras são um dos métodos não supervisionados mais usados em ML. Esses modelos são muito apropriados em situações em que o conhecimento a ser representado vem com uma estrutura de regras.

Os modelos artesanais são sistemas construídos à mão que dependem principalmente da intuição de designers humanos. Eles geralmente incorporam o conhecimento humano na forma de padrões ou regras. Normalmente, os padrões usam recursos gramaticais (por exemplo, partes do discurso), sintáticos (por exemplo, precedência de palavras) e outros recursos de linguagem para fazer uma identificação mais precisa.

Em sistemas baseados em regras, dois tipos de regras são normalmente usados:

- **Regras baseadas em padrões:** essas regras são baseadas na verificação de uma determinada sequência de tokens quanto à presença dos constituintes de algum pa-

drão. Para entender esse tipo de abordagem, apresentamos alguns exemplos de recursos de texto a serem considerados pelo designer de regras. Nesses exemplos, o objetivo seria reconhecer substâncias químicas e drogas em sentenças.

  - Podem ser siglas como "AAS"para "Ácido Acetilsalicílico";
  - Pode conter um prefixo como "amino-", por exemplo, "aminoácidos"ou "aminoglicosídeo";
  - Pode conter um sufixo como -nitrila"incluído em "propanonitrila"e "butanodinitrila"; e
  - Pode ser composto por uma fórmula molecular incluindo caracteres alfanuméricos em letras maiúsculas como "C6N4"referido como "tetracianoetileno"e "CD34"que é uma proteína fosfoglicoproteína transmembrana.

  As regras descritas acima não são suficientes para identificar todas as ocorrências de entidades em um documento. Frequentemente, as próprias regras são incompletas e não cobrem muitos exemplos.

  Os padrões em uma frase são frequentemente descritos usando expressões regulares (regexp) e correspondências. Uma expressão regular consiste exclusivamente em caracteres normais (como "abc") ou uma combinação de caracteres normais e metacaracteres (como "ab*c). Metacaracteres descrevem certas construções ou padrões de caracteres, por exemplo, se um caractere deve estar no início da linha ou se um caractere deve aparecer apenas uma vez [8].

- **Regras baseadas em contexto:** normalmente, as informações relevantes sobre as entidades nomeadas são fornecidas no contexto de suas menções. Analisar uma menção por humanos ou máquinas é uma tarefa difícil sem nenhuma informação contextual onde podemos encontrar o significado correto de uma palavra através de uma sequência. Por exemplo, se o termo "Apple"ocorrer sozinho, não é possível identificar a que esse termo se refere. Pode referir-se à fruta, uma pessoa, uma empresa ou um lugar. A resolução dessas ambiguidades costuma ser chamada de Desambiguação de Entidade Nomeada (DEN), que é um aspecto altamente desafiador de uma tarefa de extração de entidade. As regras baseadas em contexto estabelecem um nível mais alto de relacionamento entre os tokens e os recursos extraídos, por exemplo. tamanho das janelas em uma frase e o uso de conjunções para conectar palavras, frases, orações ou sentenças.

Os sistemas baseados em regras funcionam muito bem quando o léxico da linguagem não é diverso. As vantagens desses sistemas são que eles são relativamente fáceis de entender e a relação causa-efeito é transparente para que um especialista do domínio possa verificar a base de regras e fazer ajustes, se necessário. Devido a regras de domínio específicas e dicionários incompletos, muitas vezes observam-se alta precisão e baixa

revocação de tais sistemas. As principais desvantagens desses métodos são várias: i) a construção manual das regras, que pode ser uma tarefa demorada dependendo do domínio, ii) como as regras são criadas para um cenário muito específico, não é possível transferir este sistema para outro domínio diferente, e iii) não tratam muito bem informações incompletas ou incorretas, ou seja, dados que não tenham uma regra associada serão ignorados.

Métodos baseados em dicionário

As abordagens baseadas em dicionário são outras técnicas populares em ML não supervisionado. Essas técnicas utilizam recursos linguísticos como dicionários, glossários, listas de palavras vazias, taxonomias e tesauros para analisar os diferentes níveis da linguagem: fonético, lexical, semântico ou pragmático.

Este tipo de método é muito útil em campos onde as entidades a serem reconhecidas podem estar contidas em listas de palavras. No entanto, estas técnicas nem sempre são úteis, por exemplo, se a entidade a identificar for um nome próprio e apelido de uma pessoa, devido à sua natureza e diversidade, será difícil encontrá-las em qualquer recurso, seja num dicionário ou uma lista de palavras. Abordagens baseadas em dicionário requerem exploração de variações na ortografia da entidade para realizar o processo de correspondência [30]. Alguns exemplos de variações que podem ser consideradas pelo processo de correspondência automática de padrões são:

- Caracteres especiais como hífen, barra e colchetes são usados como separadores em diferentes combinações por diferentes autores. Por exemplo, a entidade "Ki-67"e "Ki67"refere-se à mesma proteína, mas contém um traço entre os caracteres, o que dificulta a correspondência correta;
- Partes dos nomes podem ser escritas em maiúsculas por alguns autores e minúsculas por outros. Cada autor pode escrever seus textos sem seguir um guia de formatação para que seja possível encontrar letras maiúsculas e minúsculas mistas como "Beta-HCG"e "beta-hcg"; e
- Existe uma grande variedade de siglas na linguagem humana, podemos ter problemas com a correspondência de descrições e siglas como "Ag"(antígeno).

Muitas vezes, os dicionários usados em PLN podem conter uma grande quantidade de informações úteis para resolver os problemas mencionados acima. Por um lado, é possível encontrá-los com uma simples lista de palavras que nosso sistema deve corresponder diretamente, mas por outro lado, eles podem conter sinônimos, antônimos, descrições, siglas, entre outros, pelo que o sistema precisa realizar uma pesquisa menos abrangente.

Como um dicionário pode ter muitos significados de uma entidade, o sistema precisa determinar qual significado é usado no contexto de um documento. Nesse processo,

observe que é importante desambiguar cada palavra no contexto correto para realizar uma correspondência correta. Por exemplo, se levarmos em conta a palavra "alegra"(em Português pode estar relacionada ao fármaco Fexofenadina ou a manifestação de alegria), a entidade pode ser incluída em um dicionário farmacológico, mas em uma frase como "A mangueira alegra nossas tardes", a palavra "alegra"não deve ser rotulada como fármaco, pois se refere ao humor.

2.2.2 Aprendizado supervisionado

O principal objetivo do aprendizado supervisionado é construir um modelo conciso da distribuição de rótulos de classe em termos de recursos preditivos usando um conjunto de dados de treinamento rotulado. O algoritmo resultante é usado para atribuir rótulos de classe a instâncias de teste. Para uma descrição desses métodos, realizamos uma divisão incluindo algoritmos tradicionais, redes neurais e modelos baseados em transformadores.

Algoritmos tradicionais

No contexto de ML, algoritmos tradicionais significam as coisas que fazemos há anos e geralmente são a base para ML mais avançado. Na seção a seguir, revisamos quatro algoritmos que são considerados métodos tradicionais de aprendizado de máquina e são amplamente utilizados por pesquisadores interessados na tarefa REN. Esses algoritmos incluem Regressão Logística, Campos Aleatórios Condicionais (CRF) e Máquinas de Vetores de Suporte (SVM). Embora existam mais algoritmos incluídos no aprendizado de máquina tradicional, nesta seção, queremos destacar aqueles comumente usados em REN e especificamente aqueles usados nos estudos iniciais desta tese.

Regressão Logística

A Regressão Logística (RL) é outra técnica que o ML adotou do campo estatístico [48]. Especificamente, o algoritmo RL é um modelo discriminativo que descreve a probabilidade condicional como:

$$P(y|X) = \frac{exp(\sum_{m=1}^{M} \lambda_m f_m(y, X))}{\sum_{y'} exp(\sum_{m=1}^{M} \lambda_m f_m(y', X)))} \tag{2.1}$$

A regressão logística é um método linear, mas as previsões são transformadas usando a função logística. Esta função também é chamada de sigmóide que descreve o peso $\lambda_m f_m$ dos recursos f_m definido em relação a y e X para gerar uma previsão de classe. Além disso, os recursos são definidos para pares de observação de estado $f_m(y, X)$ [9].

Campos Aleatórios Condicionais (CRFs)

Os Campos Aleatórios Condicionais (CRFs) são um tipo importante de modelos de ML motivados pelo princípio do Modelo de Markov de Entropia Máxima (MEMMs) [100] e usados para rotulagem de sequências. Lafferty, McCallum e Pereira [80] propuseram CRFs como modelos probabilísticos para segmentar e marcar dados de sequência para herdar as vantagens dos modelos anteriores, superar suas deficiências e aumentar sua eficiência. Segundo os autores, existem duas diferenças principais entre CRF e MEMMs são duas: MEMMs usam modelos de estado exponencial para as probabilidades condicionais dos próximos estados dado o estado atual, e o algoritmo CRF tem um único modelo exponencial para a probabilidade conjunta do estado atual. sequência de rótulo inteira dada a sequência de observação. Assim, os pesos das diferentes características em diferentes estados podem ser trocados entre si. O princípio básico do CRF é definir a distribuição de probabilidade condicional sobre as sequências de rótulos em uma dada observação [166]. Mais especificamente, a probabilidade condicional de uma sequência de rótulos y dada uma sequência de palavra X é mostrada na Equação 2.2.

$$P(y|X) = \frac{exp(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(y_{i-1}, y_i, X))}{\sum_y exp(\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_j f_j(y_{i-1}, y_i, X))} \tag{2.2}$$

Onde o denominador da equação é um fator de normalização de todas as sequências de estado. $f_j(y_{i-1}, y_i, X))$ é uma função m que descreve um recurso específico e λ_j é um peso aprendido para cada função de recurso.

Usando modelos CRF as sequências podem ser representadas por características linguísticas. Recursos típicos para CRFs podem ser generalizados, como a palavra anterior, a palavra atual, a próxima palavra e a tag *Part-Of-Speech* (POS) para fornecer contexto ao modelo. Além disso, outras características respondem à sintaxe da palavra como é pequena, é maiúscula, é um número, entre outros.

Máquinas de Vetores de Suporte (SVMs)

Máquinas de Vetores de Suporte (SVMs) são modelos de aprendizado supervisionado com algoritmos de aprendizado associados para classificação de dados. Os SVMs foram desenvolvidos na década de 1990, dentro do campo da ciência da computação. Embora tenham sido originalmente desenvolvidos como um método de classificação binária, sua aplicação se estendeu a múltiplas classificações e problemas de regressão. Os SVMs provaram ser um dos melhores classificadores para uma ampla gama de situações e, portanto, são considerados um dos benchmarks no campo da estatística e aprendizado de máquina [15].

O modelo SVM permite a expansão do espaço através de *kernels* [148]. Embora não entremos em detalhes, existem muitos *kernels* diferentes, sendo alguns dos mais usados: linear, polinomial, função de base radial gaussiana (RBF) e tangente hiperbólica ou sigmóide.

Redes Neurais

Nos últimos anos, as redes neurais profundas revolucionaram muitos domínios de aplicação de ML. As redes neurais profundas fazem parte de uma família mais ampla de métodos de aprendizado de máquina baseados em Redes Neurais Artificiais (RNAs). Uma RNA emprega uma hierarquia de camadas em que cada camada considera informações de uma camada anterior e então passa sua saída para outras camadas [40]. Embora os algoritmos de ML tradicionais sejam geralmente lineares, os algoritmos de aprendizado profundo são empilhados em uma hierarquia de complexidade e abstração crescentes.

As RNAs, geralmente chamadas de redes neurais, são inspiradas nas redes neurais biológicas que constituem os cérebros. Uma RNA é baseada em um conjunto de unidades conectadas ou nós chamados neurônios artificiais ou simplesmente neurônios, que tentam modelar os neurônios em um cérebro biológico. Cada neurônio tem entradas e produz uma única saída que pode ser enviada para vários outros neurônios. As entradas podem ser os valores característicos de uma amostra de dados externos, como imagens ou documentos. Além disso, essas características podem ser as saídas de outros neurônios. As saídas finais da rede neural cumprem a tarefa, por exemplo, a identificação de uma entidade no texto. Normalmente, os neurônios são organizados em múltiplas camadas como podemos ver na Figura 2.5. Nesta figura[1], a estrutura da rede neural consiste em camadas de entrada, ocultas e de saída. Observe que os neurônios em uma camada se conectam apenas aos neurônios nas camadas imediatamente anteriores e seguintes.

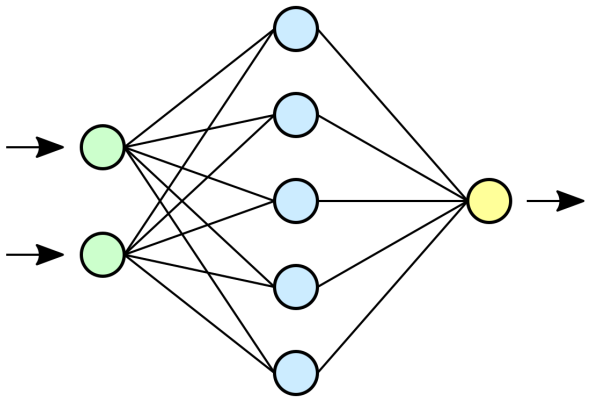

Figura 2.5 – Arquitetura básica de uma Rede Neural Artificial. Da esquerda para a direita, podemos observar as camadas de entrada, ocultas e de saída. Extraído da Wikipédia

[1] Rede Neural Artificial na Wikipédia

A *Feedforward Neural Network* (FNN) foi a primeiro e mais simples tipo de RNA [138]. Nesta rede em particular, as informações se movem em apenas uma direção: para a frente dos nós de entrada, através dos nós ocultos (se houver) e para os nós de saída, para que não haja ciclos ou loops na rede. Os FNNs podem ser divididos em dois grupos: *Single-Layer Perceptron* (SLP) e *Multi-Layer Perceptron* (MLP). Por um lado, o SLP consiste em uma única camada de nós de saída, em que as entradas são alimentadas diretamente às saídas por meio de uma série de pesos. Por outro lado, o MLP consiste em várias camadas de unidades computacionais, geralmente interconectadas de forma feed-forward. Cada neurônio em uma camada tem conexões diretas com os neurônios da camada subsequente [53].

O processo de aprendizagem de uma rede neural pega as entradas e saídas desejadas e atualiza o estado interno de acordo para que a saída calculada fique o mais próximo possível da saída desejada [124]. O processo de previsão recebe uma entrada e, posteriormente, gera (usando o estado interno) o resultado mais provável de acordo com sua experiência anterior. Para conseguir isso, discutiremos brevemente o processo de aprendizagem em várias etapas:

1. **Inicialização:** a inicialização do modelo refere-se à primeira hipótese que o processo pretende iniciar. Assim como os algoritmos genéticos e a teoria da evolução, as redes neurais podem começar de qualquer lugar. Portanto, uma inicialização aleatória do modelo é uma prática comum;

2. **Propagação para frente:** esta etapa se preocupa em propagar os cálculos de todos os neurônios dentro de todas as camadas que se movem da esquerda para a direita. Isso começa na camada de entrada e termina com a previsão final. Os cálculos diretos ocorrem durante o treinamento para avaliar o alvo e a função de perda sob as configurações atuais dos parâmetros de rede em cada iteração, bem como durante a previsão quando aplicados a dados novos e não vistos anteriormente;

3. **Função de perda:** neste passo, por um lado, temos a saída real da nossa rede neural inicializada aleatoriamente. Por outro lado, temos a saída desejada que queremos que a rede aprenda. A função de perda é uma métrica de desempenho de quão bem a rede neural atinge seu objetivo de gerar resultados o mais próximo possível dos valores desejados. Por esse motivo, os algoritmos de ML visam minimizar a função de perda. Em modelos de classificação, as funções de perda mais comuns usadas são entropia cruzada binária, entropia cruzada categórica e similaridade de cosseno, entre outras. Para problemas baseados em regressão, também temos funções como erro médio quadrático (MSE), erro percentual médio absoluto (MAPE) e erro médio absoluto (MAE);

4. **Diferenciação:** em matemática, a diferenciação é a etapa que pode ajudar a rede a otimizar os pesos. É a derivada parcial da função de perda. Portanto, ao encontrar a

derivada parcial da função de perda, sabemos quanto (e em qual direção) devemos ajustar nossos pesos e vieses para diminuir a perda;

5. **Retropropagação:** na rede neural, qualquer camada pode encaminhar seus resultados para muitas outras camadas, neste caso, para realizar a retropropagação. Essencialmente, esta etapa avalia a expressão para a derivada da função de perda como um produto das derivadas entre cada camada da esquerda para a direita usando o gradiente dos pesos; e

6. **Atualização de peso:** as atualizações dos pesos dos neurônios refletirão a importância do erro propagado para trás após a conclusão de uma passagem para frente. Os métodos de atualização de pesos são chamados de otimizadores.

7. **Iterar até a convergência:** como atualizamos os pesos com um pequeno passo, a rede precisará de várias iterações para aprender a saída ideal.

Em uma RNA, as funções de ativação determinam a saída de um modelo de aprendizado profundo, sua precisão e também a eficiência computacional do treinamento de um modelo. As funções *Softmax* e *sigmoid* são funções comuns usadas na camada final ou de saída de uma rede neural para obter uma distribuição categórica e de Bernoulli respectivamente (função para classificação binária) [59]. Para as camadas ocultas, as funções de ativação mais populares são a Unidade Linear Retificada (ReLU) e uma tangente hiperbólica ou função *tanh* [129]. Adicionalmente, algumas funções aplicadas à rede neural possuem parâmetros de configuração diferentes. Por exemplo, a taxa de aprendizado é um parâmetro de ajuste na função de otimização que determina o tamanho do passo em cada iteração enquanto se move em direção a uma função de perda mínima.

Como mencionado acima, as redes neurais costumam usar otimizadores para reduzir as perdas. São algoritmos ou métodos usados para alterar os atributos de uma determinada rede neural, como pesos e taxa de aprendizado. Alguns exemplos dos otimizadores mais usados são Adam, *Stochastic Gradient Descent* (SGD), Adadelta, RM-Sprop, Adamax e Adagrad [70, 29, 177].

As RNAs têm limitações para lembrar sequências quando são grandes. Por exemplo, suponha uma frase de 90 palavras na qual a penúltima palavra se refere ao início da frase. As RNAs tendem a esquecer informações sobre etapas de tempo que estão muito atrasadas. Para resolver este problema, surgiu o mecanismo de atenção para lidar com dados variáveis no tempo (sequências) [21]. A atenção é considerada uma das ideias mais influentes na comunidade de aprendizagem profunda porque mantém informações relevantes ao longo do tempo. Com este mecanismo, cada palavra da sentença contém um estado oculto com valores passados que serão levados em consideração em cada iteração.

Em resumo, o aprendizado profundo representa um conjunto de técnicas baseadas em RNAs. As duas principais arquiteturas projetadas para classificação são as Redes

Neurais Recorrentes (RNRs) e as Redes Neurais Convolucionais (CNNs) [138]. Esses dois modelos diferem no tipo de entrada que eles preveem: RNRs são projetados para classificar sinais temporais e CNNs são projetados para classificar sinais espaciais. As particularidades desses tipos de RNAs são descritas a seguir.

- **Redes Neurais Recorrentes (RNRs):** as RNRs foram estudados pela primeira vez em 1986 [54] e são uma classe de RNAs em que as conexões entre os nós formam um grafo direcionado ao longo de uma sequência de tempo. RNRs são comumente usados para problemas ordinais ou temporais, como Processamento de Linguagem Natual (PLN), Reconhecimento de Fala e Rotulagem de Imagens; eles também são incorporados em aplicativos populares como "Siri da Apple" e o "*Google Voice Search*".

  As RNRs usam uma célula de memória que recebe informações da entrada anterior para influenciar a entrada e a saída atuais. Enquanto as redes neurais profundas tradicionais assumem que as entradas e saídas são independentes umas das outras (sem feedback), a saída dos RNRs depende dos elementos acima dentro da sequência. A Figura 2.6 ilustra como RNR (à direita da figura) tem uma conexão recorrente no estado oculto e FNN não contém realimentação (à esquerda da figura). Essa restrição de loop garante que a informação sequencial seja capturada nos dados de entrada [21].

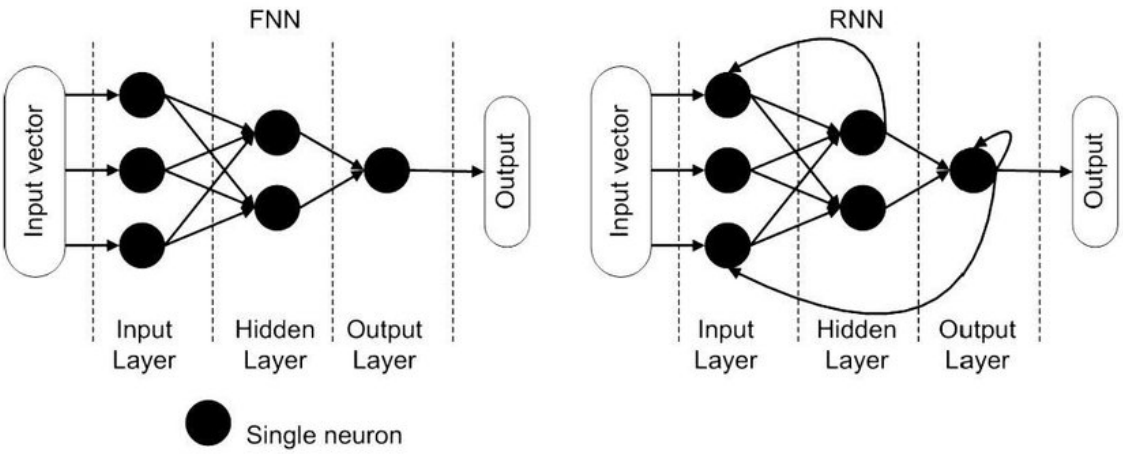

Figura 2.6 – Comparação entre FNNs (esquerda) e RNRs (direita). Extraída de Kranker et al. [76]

- **Redes Neurais Convolucionais (CNNs):** nos casos em que as entradas são grandes, os RNRs envolvem um grande número de parâmetros de treinamento. A ideia principal para superar este problema é pegar a representação local que descreve toda a entrada ao invés da representação global. A CNN usa camadas com filtros de convolução que são aplicados a recursos locais para representar informações locais [57].

Nesse tipo de rede neural, as conexões entre os nós não formam um loop, mas usam uma variação do MLP projetada para exigir o mínimo de pré-processamento.

Originalmente inventados para visão computacional, os modelos CNN demonstraram ser eficazes para PLN e alcançaram excelentes resultados na classificação de texto [58]. CNNs também podem ser aplicadas a tarefas PLN usando dados textuais porque as entradas são a representação vetorial de cada palavra em uma frase.

Três tipos de camadas compõem a CNN: camadas convolucionais, camadas de pool e camadas totalmente conectadas. Como mostramos na Figura 2.7, a convolucional é a primeira camada que é usada para extrair as várias características da entrada. Nesta camada, a operação matemática de convolução é realizada entre a entrada e um filtro de um determinado tamanho *MxM*. Na maioria dos casos, a camada convolucional é seguida pela camada pool. O objetivo principal desta camada é diminuir o tamanho do mapa de feições convolucionais para reduzir os custos de computação. Finalmente, a camada totalmente conectada consiste nos pesos e desvios juntamente com os neurônios e é usada para conectar os neurônios entre duas camadas diferentes. Estas camadas normalmente são colocadas antes da camada de saída e formam as últimas camadas da arquitetura CNN.

Modelos baseados em *Transformers*

Transformers é um modelo de aprendizado profundo introduzido em 2017, usado principalmente no campo da PLN [160]. Semelhante as RNRs, *Transformers* foi projetada para lidar com dados sequenciais, como linguagem natural, para tarefas como REN e classificação de texto. Ao contrário das RNRs, a *Transformer* não exige que os dados sequenciais sejam processados em ordem, portanto, não precisamos processar o início de uma frase antes de processar o final. Devido a este aspecto, esta técnica permite muito mais paralelização do que RNRs e, portanto, reduz os tempos de treinamento.

A parte mais importante da arquitetura *Transformer* é o **mecanismo de atenção**. O mecanismo de atenção representa a importância que outros tokens de uma entrada têm para a codificação de um determinado token. Em outras palavras, o mecanismo de atenção permite que o *Transformer* se concentre em certas palavras à esquerda e à direita para tratar a palavra atual de acordo com a tarefa de PLN que estamos abordando.

Outra vantagem desta arquitetura é que o aprendizado em um idioma pode ser transferido para outros idiomas por meio do aprendizado por transferência, ou *Transfer Learning*. Em termos gerais, a aprendizagem por transferência é a ideia de pegar o conhecimento adquirido ao realizar uma tarefa e aplicá-lo a uma tarefa diferente. A *Transformer* conta com essa técnica para obter resultados "estado da arte"[113].

Há uma grande diferença entre a abordagem tradicional de construção e treinamento de modelos de ML e o uso de uma metodologia que segue os princípios do aprendi-

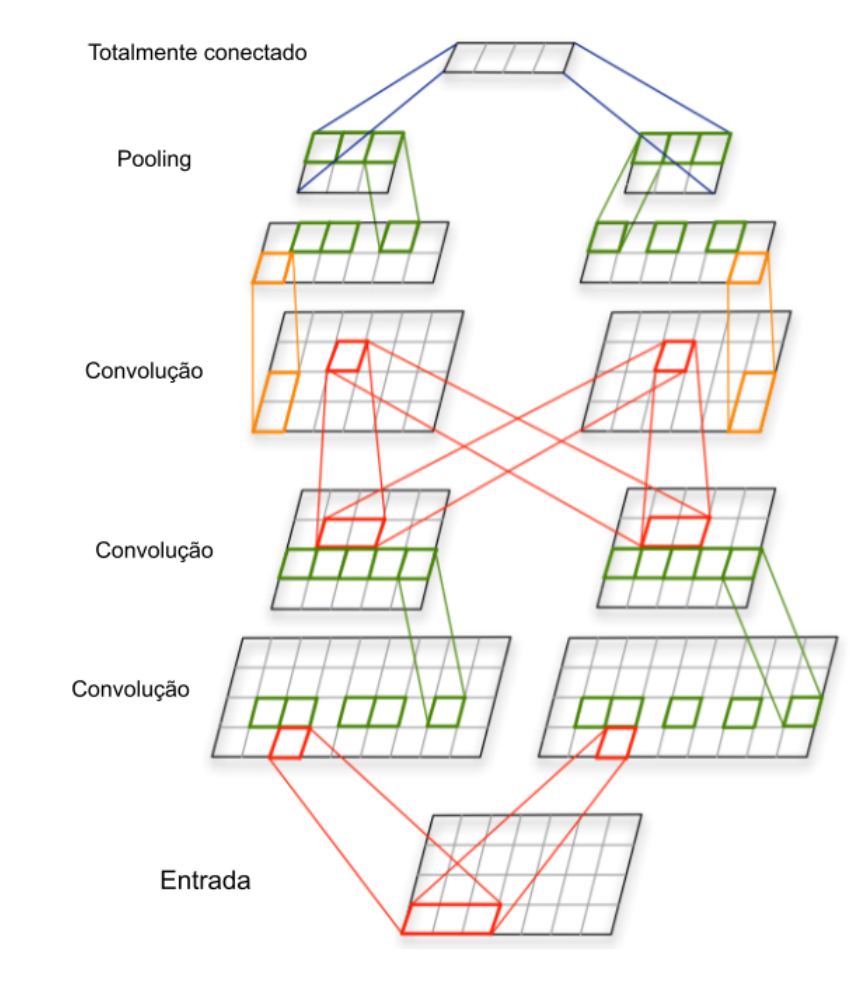

Figura 2.7 – Arquitetura básica de uma CNN. Adaptada de Kalchbrenner et al. [64]

zado por transferência. A Figura 2.8 ilustra a diferença entre o ML tradicional e a nova ideia baseada no aprendizado por transferência. Nesta figura, podemos ver que o ML tradicional (à esquerda da figura) é isolado e ocorre estritamente para tarefas específicas (tarefa 1 e tarefa 2) e conjunto de dados (conjunto de dados 1 e conjunto de dados 2). Portanto, os modelos de treinamento são independentes entre si e não retêm nenhum conhecimento que possa ser transferido de um modelo para outro. Ao contrário, usando modelos de aprendizagem por transferência (à direita da figura) é possível aproveitar o conhecimento (características, pesos, etc.) dos modelos previamente treinados para treinar novos sistemas.

Como podemos observar na Figura 2.8, os modelos *Transformers* são primeiro treinados em grandes quantidades de texto (conjunto de dados 1) em uma etapa chamada pré-treinamento. Durante esta etapa, espera-se que os modelos aprendam as palavras, estrutura, morfologia, gramática e outras características linguísticas da língua. Nesta etapa, o texto é representado por tokens por meio de um processo de tokenização convertendo o texto real em uma representação numérica que pode ser usada com modelos de redes neurais.

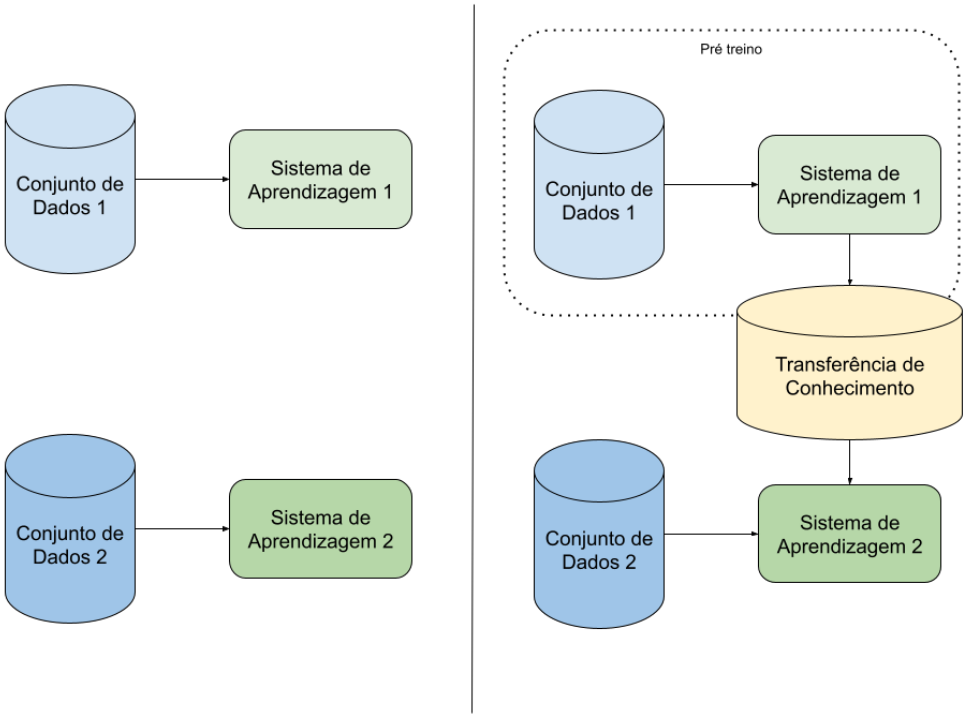

Figura 2.8 – Modelos tradicionais de aprendizado de máquina versus aprendizado por transferência

Depois que o texto é convertido para um formato compatível, podemos treinar o modelo para entender o idioma. *Masked Language Modeling* (MLM) é uma técnica utilizada para realizar esta tarefa [26]. Com essa técnica, uma certa porcentagem dos tokens em uma sequência é substituída por um token de máscara e o modelo é solicitado a prever o token que estava lá anteriormente. Quando treinado em uma tarefa usando essa técnica, um modelo é capaz de aprender boas representações dos vários tokens de vocabulário. Outras técnicas para aprender a representação de texto são a Modelagem de Linguagem Causal (CLM), que prevê a probabilidade de uma palavra dadas as palavras anteriores em uma frase, e a Modelagem de Linguagem de Tradução (TLM) projetada para dados entre idiomas [83].

O resultado desse processo de treinamento prévio é um modelo capaz de modelar uma língua com precisão, compreendendo as diferentes características e regras linguísticas da língua [119].

O aprendizado de transferência é popular no aprendizado profundo, devido aos enormes recursos necessários para treinar modelos de aprendizado profundo e os conjuntos de dados grandes e desafiadores nos quais os modelos de aprendizado profundo são treinados. Felizmente, muitos modelos pré-treinados já estão disponíveis para reutilização e servem como ponto de partida para diferentes tarefas. HuggingFace é a biblioteca Python mais popular que contém esses modelos. Entre os modelos pré-treinados mais conhecidos, podemos encontrar T5, GPT-3, GPT-2, BERT, XLNet e RoBERTa, que demonstram a capa-

cidade dos *Transformers* de realizar uma ampla variedade de tarefas relacionadas à PLN e têm o potencial para encontrar aplicativos do mundo real [82, 86, 11, 26, 91, 176, 20].

Em essência, a *Transformers* mudou a área de PLN oferecendo os seguintes benefícios:

- Introduziu um mecanismo revolucionário de atenção que substitui as arquiteturas convolucionais ou recorrentes;
- Produziu uma mudança no aprendizado de transferência do pré-treinamento (vetores de palavras) para extração de características para o treinamento de modelos de linguagem genéricos (modelos pré-treinados);
- Ele forneceu ajustes onde o modelo pode precisar ser adaptado para a tarefa de interesse;
- Isso resultou em um crescimento exponencial no tamanho dos modelos de linguagem pré-treinados, o que levou a um alto desempenho em uma série de tarefas de PLN envolvendo compreensão da linguagem; e
- Ela forneceu modelos pré-treinados para tarefas em que não temos grandes conjuntos de dados para treinar.

2.3 Representação de Palavras

No campo da PLN, os pesquisadores precisavam encontrar uma maneira de representar dados textuais como entrada em sistemas de ML. Esse processo consiste em transformar um conjunto de características categóricas do texto bruto (palavras, letras, tags de parte do discurso, posição das palavras, ordem das palavras, entre outros) em uma série de vetores.

Primeiramente, além de converter palavras em uma representação numérica, fazemos as seguintes perguntas: o que nos interessa saber sobre o texto ao realizar essa codificação de dados? E mais especificamente, o que exatamente queremos codificar? Nesta seção, discutimos as opções mais usadas pela comunidade de PLN para representar palavras que os sistemas de ML podem entender.

A primeira abordagem surgiu como um método simples no qual cada palavra do vocabulário recebia um identificador único. Um vocabulário em um corpus ou texto consiste nas palavras únicas incluídas nele. Os métodos de **pesquisa em dicionário** são uma maneira simples de representar texto verificando se uma string de entrada aparece em um dicionário. Caso contrário, se a palavra não aparecer, a string será marcada como uma palavra incorreta ou *Out-Of-Vocabulary* (OOV) [77].

A abordagem baseada em dicionário armazena o maior número possível de entidades nomeadas em uma lista chamada de dicionário geográfico, que oferece alta precisão na identificação correta dessas entidades [61]. Atualmente, tais métodos estão desatualizados para representação de palavras simples porque têm limitações, mas ainda são usados como recursos adicionais para representação de palavras em redes neurais [97, 149]. As principais deficiências dessas técnicas podem ser resumidas da seguinte forma: *i*) um dicionário curto pode não ser suficiente para encontrar a palavra no contexto, *ii*) um dicionário pode não conter todas as palavras que queremos representar e *iii*) um dicionário grande pode aumenta o custo da pesquisa.

A segunda abordagem para realizar a representação de palavras que estudamos é chamada de **one-hot encoding**. Essa técnica usa uma representação de variáveis categóricas como vetores binários. A ideia principal é criar um vetor de tamanho de vocabulário preenchido com todos os zeros, exceto uma posição. Então, para uma palavra, apenas a coluna correspondente é preenchida com o valor 1 e o restante tem valor zero. Além disso, este método usa uma posição vetorial para indicar que a palavra é OOV [37]. Para melhor compreensão deste tipo de representação, a Tabela 2.1(Figura 3.1) mostra um exemplo onde a frase "anomalias do sistema cardiovascular"é codificada através de zeros e uns. Como podemos ver, as palavras codificadas consistem em um vetor de dimensão $N + 1$, onde N é o tamanho do vocabulário e o 1 extra é adicionado para palavras OOV.

	anomalias	do	sistema	cardiovascular				OOV
anomalias	[1,	0,	0,	0,	0,	0,	...,	0]
do	[0,	1,	0,	0,	0,	0,	...,	0]
sistema	[0,	0,	1,	0,	0,	0,	...,	0]
cardiovascular	[0,	0,	0,	1,	0,	0,	...,	0]

Tabela 2.1 – Exemplo de *one-hot encoding*

Os vetores *one-hot* são frequentemente usados como representações de palavras para tarefas REN. Por exemplo, Kuru, Can e Yuret [79] desenvolveram um modelo baseado em redes neurais que tinha como entrada uma representação de vetores *one-hot* para codificar os caracteres de entrada.

O conceito de similaridade de palavras também é difícil de extrair, pois os vetores de palavras mencionados até agora são estatisticamente ortogonais. Por exemplo, os pares de palavras "tumor"e "tumores", ou "fármaco"e "medicamento", são semelhantes, mas são representados de maneiras diferentes. Portanto, precisamos de uma abordagem mais robusta para lidar com a descoberta de semelhanças entre palavras. Para resolver os problemas iniciais de similaridade de palavras, surgiram abordagens distribucionais para a representação de palavras.

Um dos métodos mais utilizados na família de **representações distribucionais** é denominado Tf-Idf (*Term frequency – Inverse document frequency*). O Tf-Idf é um modelo de ponderação *Bag-Of-Words* (BOW) [144] usado para dar pesos às palavras em uma co-

leção de documentos, medindo a frequência com que uma palavra é encontrada em um documento (Tf), compensada pela frequência com que a palavra é encontrado em toda a coleção (Idf) [131]. Muitas vezes, os modelos de ML também usam o Tf ponderado para atribuir o número de ocorrências de uma palavra em um documento. A ideia principal por trás dessa abordagem é que as palavras que normalmente aparecem em um contexto e documento semelhantes teriam um significado semelhante. Tf, Idf e Tf-Idf foram comparados em estudos relacionados ao reconhecimento de entidades biomédicas como pesos de palavras. Zhang e Elhadad [178] mostraram que os pesos Idf e Tf-Idf fornecem melhorias em relação ao uso apenas de frequência de termo como peso.

Uma das principais desvantagens das codificações de palavras anteriores mostradas é a falta de representação de significado. Com abordagens como *one-hot* ou distribucional, representamos a presença e ausência de palavras em um determinado texto, porém, não podemos determinar nenhum significado a partir da simples presença/ausência dessas palavras. Parte desse problema é que perdemos as relações posicionais entre as palavras e a ordem das palavras. Essa ordem na sequência das palavras acaba sendo fundamental na representação do significado das palavras e é discutida a seguir.

Word embeddings são uma família de técnicas de PLN que se concentram no mapeamento do significado semântico de uma palavra em um espaço geométrico. Para isso, um vetor numérico é associado a cada palavra do vocabulário, de modo que a distância entre quaisquer dois vetores captura parte da relação semântica entre as duas palavras associadas. Além disso, os *word embeddings* desempenham um papel fundamental no *Transfer Learning*, pois são treinados em grandes quantidades do corpus usando redes neurais [35]. A Figura 2.9 traz uma representação visual das dez palavras mais similares a "*medication*".

Na Figura 2.9, as linhas indicam associações de similaridade com a palavra "*medication*". No grafo, a palavra mais similar a "*medication*" (medicamento) é "*antidepressant*" (antidepressivo). *Word embeddings* foram popularizados pela Word2Vec em 2013 [103]. Posteriormente, Pennington, Socher e Manning [116] criaram o algoritmo GloVe que visa realizar explicitamente o procedimento de incorporação de significados do Word2Vec. Embora o vocabulário de um espaço de incorporação de palavras seja grande, podemos encontrar situações onde uma palavra é OOV. FastText foi projetado para resolver esta situação, melhorando Word2Vec [14].

Recentemente, embeddings de palavras contextuais, como Embeddings from Language Models (ELMo) e BERT, surgiram. Essas técnicas geram embeddings para uma palavra de acordo com o contexto em que a palavra aparece, gerando assim embeddings ligeiramente diferentes para cada ocorrência da palavra [157]. Por um lado, o ELMo é derivado de um LSTM bidirecional que é treinado com um objetivo de modelo de linguagem acoplado em um grande corpus de texto [136], dessa forma, o ELMo analisa a frase inteira antes de atribuir um vetor a cada palavra. Por outro lado, as representações BERT são

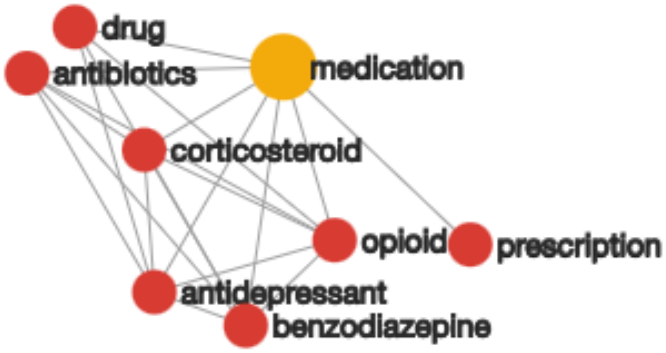

Figura 2.9 – Representação de similaridade da palavra "*medication*" obtida por um modelo Word2Vec [103] no corpus Google News. A visualização foi gerada a partir do site WebVectors: word embeddings online

condicionadas conjuntamente no contexto esquerdo e direito e usam o Transformer [160], uma arquitetura de rede neural baseada em um mecanismo de atenção (Seção 2.4).

2.4 BERT

Bidirectional Encoder Representations from Transformers [26], abreviado como BERT, é uma rede neural que gera representações bidirecionais de palavras, ou seja, baseado no contexto à esquerda e à direita. O BERT é composto por múltiplas camadas de codificadores *Transformer* [160]. O BERT apresenta duas versões, distintas pelo número de camadas *Transformer*, número de núcleos de atenção e tamanho de camada oculta. O BERT-base conta com 12 camadas, 12 núcleos de atenção e uma camada oculta com 768 neurônios totalizando 110M de parâmetros, o BERT-large conta 24 camadas, 12 núcleos de atenção e uma camada oculta de tamanho 1024, totalizando 340M de parâmetros.

Representação da Entrada

O BERT pode receber como entrada uma sentença ou um par de sentenças concatenados em uma sequência. As palavras da entrada são tokenizadas em *WordPieces* e convertidas em *word embeddings*. A tokenização WordPiece recebe esse nome porque

quebra as palavras ausentes no vocabulário em tokens, inserindo o símbolo ## nas partes não iniciais, por exemplo, “ga” e “##to”.

Além disso, no início da sequência é inserido um token especial [CLS], usado na etapa de classificação para concatenar as representações individuais de palavras de uma sentença. Para lidar com entradas compostas por um par de sentenças um outro token especial ([SEP]) delimita a fronteira das sentenças. Em adição a isso, treina-se um vetor de segmentos, indicando se a palavra pertence à sentença A ou à B. Por fim, assim como a *Transformer*, o BERT não captura a ordem de palavras, portanto necessita de um vetor de posições, nesse caso, computando a posição absoluta da palavra na sentença. A entrada final do BERT é dada pela somatória dos vetores de *word embedding*, de posição e de segmento, como pode ser observado na Figura 2.10.

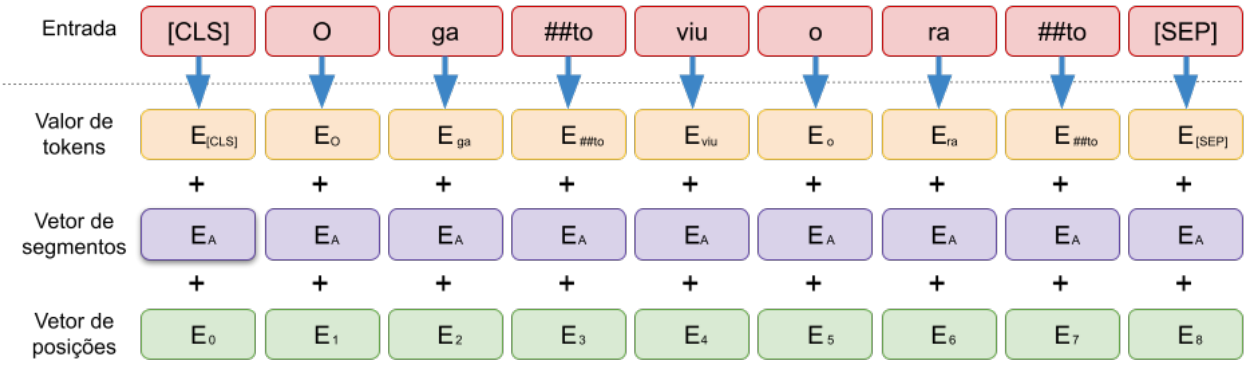

Figura 2.10 – Representação da entrada do BERT

Treinamento do BERT

São necessárias duas etapas de treinamento no BERT: o pré-treinamento e o refinamento (do inglês, *fine-tuning*). O pré-treinamento é feito em corpora não anotados em duas tarefas distintas: *Masked Language Model* (MLM) e *Next Sentence Prediction* (NSP). Por sua vez, o refinamento é treinado em corpora anotados para uma tarefa específica, como REN, tradução automática, etc. A Figura 2.11 traz uma representação das etapas de pré-treinamento e de refinamento do modelo.

Aplicações do BERT

Como já mencionado, o BERT pré-treinado pode ser refinado em uma tarefa específica. Nessa abordagem, treina-se o modelo em um corpus anotado, otimizando os pesos

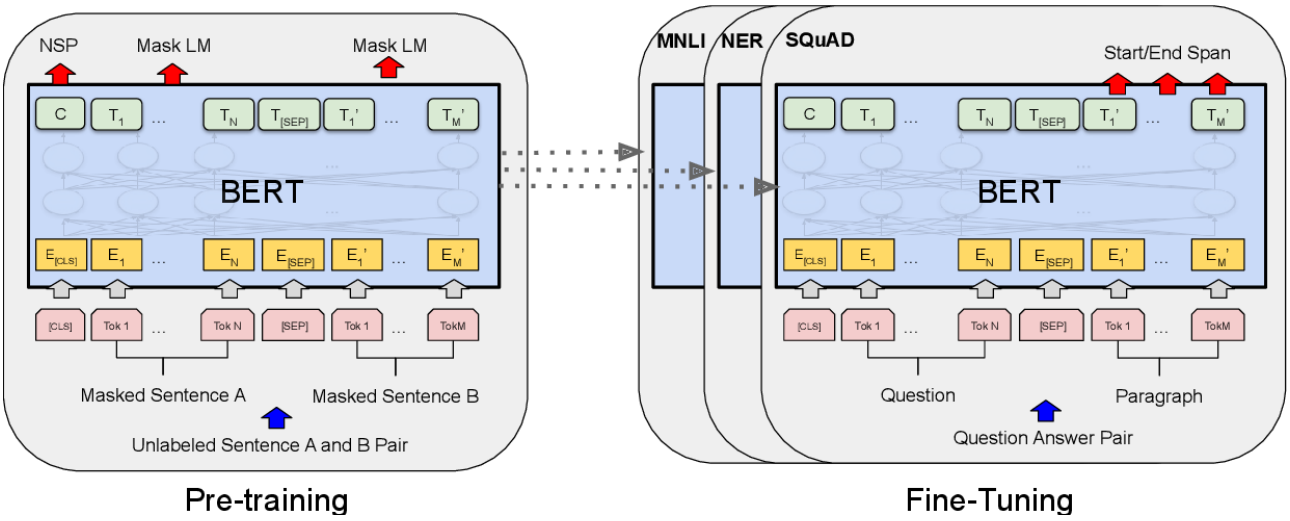

Figura 2.11 – Representação da etapa de treinamento extraído de Devlin et al.[26]. À esquerda, a etapa de pré-treinamento e à direita, o refinamento

aprendidos. As predições são obtidas por uma camada de saída para classificação na tarefa em questão, que recebem como entrada as saídas do BERT. O refinamento é menos custoso em termos computacionais do que o pré-treinamento, já que os pesos do BERT já passaram por treinamento, necessitando somente de ajustes para se adaptar na tarefa.

Além da opção acima, é possível usar o BERT como *word embeddings* em outros modelos, extraindo as representações de palavras geradas na etapa de pré-treinamento. Essa abordagem é conhecida como *feature-based*, uma técnica de *Transfer Learning*. Em vez de refinar o BERT em uma tarefa específica, retreinando seus pesos, os traços do modelo pré-treinado são extraídos e podem alimentar outro modelo de classificação, mantendo os pesos congelados, sem refinamento.

2.5 Avaliação da tarefa de REN

A avaliação dos sistemas de REN é feitas através da comparação do resultado obtido por um modelo e um corpus anotado com as respostas esperadas. Existem dois tipos de corpus anotado: dourado e prateado. Um corpus dourado é aquele contendo anotação manual, ou seja, feito e/ou revisado por humanos, em geral, especialistas na área de conhecimento da tarefa a ser resolvida. Já um corpus prateado é anotado automaticamente através de técnicas computacionais.

Para fazer a anotação de ENs em um texto, é preciso seguir diversos procedimentos, entre os quais está o tipo de etiquetagem. Uma abordagem muito utilizada nos primeiros trabalhos de REN foi a anotação por XML, sendo adotada em inúmeras competições como o MUC e o Harem. O XML é uma linguagem de marcação que estrutura a informação através de etiquetas. Tradicionalmente, somente a palavra ou palavras referen-

tes a uma entidade eram etiquetadas no texto, delimitadas por < e >, como apresentado no exemplo abaixo. A etiqueta “EN” indica as fronteiras da entidade nomeada e “TIPO” se refere a etiqueta atribuída a ela.

> A <EN TIPO“LOCAL”>Fundação Iberê</EN> é uma entidade cultural que tem como objetivos a preservação, o estudo e a divulgação da obra do pintor gaúcho <EN TIPO=“PESSOA”>Iberê Camargo</EN>.

Atualmente, é mais comum que os corpora de REN sejam anotados no formato BIO (*Begin-Inside-Outside*). Nele, todas as palavras do corpus recebem uma etiqueta indicando sua categoria, sendo que as letras **B** e **I** são usadas para marcar, respectivamente, o início de uma entidade e uma ou mais palavras dentro de uma entidade. Essas letras são adicionadas antes da categoria da EN, como “B-PESSOA”. Por sua vez, O é reservado para as palavras que não são ENs.

Giovani	B-PESSOA
é	O
natural	O
de	O
Bento	B-LOCAL
Gonçalves	I-LOCAL
.	O

Tabela 2.2 – Exemplo de anotação no estilo BIO

Na Tabela 2.2, podemos ver um exemplo de anotação no esquema BIO. Observe como “Bento” tem o prefixo “B-” porque segue imediatamente outra entidade LOCAL. Semelhante, mas mais detalhado que BIO, o esquema BILOU codifica o início, o interior e o última entidade de pedaços de várias entidades enquanto os diferencia dos pedaços de comprimento de unidade. Isso consiste nos prefixos **B**, **I**, **L**, **U** ou O, onde U é usado para representar um pedaço contendo uma única entidade. Os pedaços de comprimento maior ou igual a dois sempre começam com o prefixo “B-” e terminam com o prefixo “L-”. Observe na Tabela, como a mesma frase é anotada de forma diferente em BILOU.

Giovani	U-PESSOA
é	O
natural	O
de	O
Bento	B-LOCAL
Gonçalves	L-LOCAL
.	O

Tabela 2.3 – Exemplo de anotação no estilo BILOU

Os corpora mais recentes tendem a aplicar o esquema BIO de anotação, uma vez que ele é baseado em unidades de palavra (*token*) em vez de sequências de palavras (*chunk*), como no exemplo do XML.

Ao longo dos anos, foram propostas diferentes métricas de avaliação para o REN. A avaliação por entidade é feita com base na correspondência por *chunk*, que pode ser composta por uma ou mais palavras. Como medida de desempenho da tarefa, adotamos o protocolo de avaliação do CoNLL [133], em virtude de que outros trabalhos para REN em Português também se utilizam deste, o que torna possível a comparação das abordagens. As métricas de avaliação propostas nessa conferência são baseadas nas medidas de precisão, *recall* e medida-F. No CoNLL, a classificação de uma entidade é considerada correta somente se é uma correspondência idêntica daquela anotada no corpus. Assim, a precisão (p) é dada pela porcentagem de entidades classificadas corretamente pelo sistema (Equação 2.5) e o *recall* (r) é a porcentagem de entidades presentes no corpus dourado encontradas pelo sistema (Equação 2.6). O cálculo de medida-F (F) é apresentado nas Equações 2.3 e 2.4, em que $\beta = 1$.

$$F_\beta = \frac{(\beta^2 + 1) * p * r}{\beta^2 * (p + r)} \tag{2.3}$$

$$F_1 = \frac{2 * p * r}{p + r} \tag{2.4}$$

A Tabela 2.4 traz um exemplo de como é feita a avaliação de REN com uma anotação BIO.

$$p = \frac{\text{Número de EN encontradas corretamente pelo sistema}}{\text{Número de EN encontradas pelo sistema}} \tag{2.5}$$

$$r = \frac{\text{Número de EN encontradas corretamente pelo sistema}}{\text{Número de EN no corpus dourado}} \tag{2.6}$$

Texto	**Anotação Dourada**	**Predição**
Medx	O	O
em	O	O
uso	O	O
:	O	O
AAS	B-FAR	B-FAR
,	O	O
Paracetamol	B-FAR	B-FAR
e	O	O
Omeprazol	B-FAR	B-FAR
.	O	O

Tabela 2.4 – Exemplo de avaliação por entidade usando BIO

2.6 Recursos de Conhecimento Biomédico

Atualmente, o uso de recursos de conhecimento apropriados é essencial para o desenvolvimento de sistemas de PLN. Nesta seção, revisamos os recursos de terminologia mais populares no campo do BioNLP e ferramentas de detecção automática de entidades. Por um lado, os recursos terminológicos incluem ontologias, vocabulários controlados e léxicos que estão disponíveis aos pesquisadores para melhor representar o conhecimento por meio de conceitos, estruturas e relações entre eles. Por outro lado, *frameworks* computacionais foram desenvolvidos para construir rapidamente ferramentas para tarefas de extração de entidades biomédicas.

Recursos terminológicos

A identificação de entidades na literatura biomédica é um dos tópicos de pesquisa mais desafiadores dos últimos anos, tanto no PLN quanto nas comunidades biomédicas. Felizmente, existem vários recursos terminológicos corrigidos manualmente e com curadoria disponíveis para a comunidade científica, onde os pesquisadores podem encontrar entidades biomédicas relevantes.

Recursos de terminologia com curadoria fornecem a linguagem médica comum necessária para interoperabilidade e troca eficiente de dados clínicos. Para maximizar o valor da informação em saúde, esses recursos devem ser usados adequadamente de acordo com sua finalidade, domínio e design. Eles são projetados para atender a uma variedade de propósitos com os seguintes benefícios: (*i*) o conhecimento em saúde é incluído em um recurso facilmente acessível, (*ii*) os conceitos são frequentemente categorizados para pesquisa, (*iii*) facilitam a normalização de dados e (*iv*) permitem a interoperabilidade entre sistemas por meio de identificadores de conceito únicos.

Entre as dificuldades para identificar com sucesso os termos estão as amplas variações lexicais, que impedem que alguns termos sejam reconhecidos no texto biomédico, a sinonímia de termos e a homonímia de termos (quando um termo tem vários significados), que criam incerteza quanto à exata identidade do termo [75]. Para resolver este problema, diferentes ontologias, vocabulários controlados, terminologias e dicionários contendo uma variedade de termos foram projetados. Mostrando as questões e desafios em aberto, Rector et al. [123] fornecem um levantamento sobre a origem da palavra “ontologia” em sistemas de informação e discutem as lições e as implicações de projetar sistemas de raciocínio de mundo aberto baseados em lógicas de descrição para representação do conhecimento biomédico. Uma ontologia é uma especificação explícita de conceituações [43]. O termo é emprestado da filosofia, onde a ontologia é uma descrição sistemática da existência. No

entanto, no campo da ciência da computação, a visão de ontologia é um pouco mais restrita. Uma definição de ontologia foi dada por Uschold et al. [159] descrevendo a ontologia assim:

> *"Uma ontologia pode assumir uma variedade de formas, mas necessariamente incluirá um vocabulário de termos e alguma especificação de seu significado. Isso inclui definições e uma indicação de como os conceitos estão inter-relacionados que coletivamente impõem uma estrutura no domínio e restringem as possíveis interpretações dos termos."*

Em ontologias, as características associadas aos nomes das entidades (por exemplo, descrições, relações com outros objetos, funções, entre outros) descrevem mais especificamente o significado de cada uma delas. Além disso, os relacionamentos entre entidades tornam as ontologias bem estruturadas [81, 154].

Atualmente, nenhuma ontologia captura toda a gama de conceitos no domínio Biomédico. No entanto, apesar das preocupações mencionadas, existem várias ontologias biomédicas bem projetadas, como a UMLS (*Unified Medical Language System*) [13], a Ontologia TAMBIS (TaO) [10] e a *Gene Ontology* (GO) [7]. UMLS e GO são as ontologias biomédicas mais populares na comunidade BioNLP, uma vez que atualmente envolvem o maior número de conceitos.

Ferramentas de mapeamento

Conforme descrito acima, o alcance e a diversidade de ontologias, terminologias e léxico aumentaram dramaticamente ao longo dos anos. A demanda por mapeamento de dados textuais com esses recursos terminológicos levou à criação de sistemas e ferramentas automáticas. Cada um desses sistemas tem características comuns, todos os quais empregam um ou mais dos seguintes recursos: análise lexical, muitas vezes usando um léxico especializado; análise sintática, procedimento de mapeamento que leva em conta a correspondência parcial; e o uso de fontes de conhecimento para fazer a correspondência.

No domínio Biomédico, podemos encontrar algumas ferramentas interessantes para o inglês como MicroMeSH [96], CHARTLINE [104] e EDGAR [126]. A maioria deles usa a ontologia UMLS para combinar entidades reconhecidas nos dados textuais com o maior Metathesaurus conhecido até agora. Entre as ferramentas mais utilizadas pela comunidade BioNLP, podemos incluir o cTAKES e o MetaMap.

- **cTAKES:** é um sistema popular que visa construir e avaliar um sistema PLN de código aberto para a extração de informações do RES textual escrito em inglês. Este

sistema fornece mapeamentos para o UMLS usando diferentes componentes: tokenizer cTAKES, normalizador, tagger *Part-Of-Speech* (POS) e anotador REN. Estudos relevantes avaliaram o cTAKES como um sistema REN para descoberta de casos de doença arterial periférica [137], identificação de distúrbios [115, 172] e extração de conhecimento diagnóstico [122].

- **MetaMap:** é um aplicativo altamente configurável desenvolvido pela *National Library of Medicine* (NLM) para mapear texto biomédico para o Metathesaurus UMLS ou, equivalentemente, para identificar conceitos do Metathesaurus referidos em um texto em inglês. Essa ferramenta emprega uma abordagem de conhecimento intensivo, métodos PLN e técnicas linguísticas computacionais para identificar conceitos com mais precisão. O MetaMap é usado em muitos estudos como referência. Por exemplo, Jimeno et al. [61] compararam três soluções. Por um lado, usaram um modelo baseado em dicionário, por outro, um modelo estatístico e, finalmente, a ferramenta de mapeamento MetaMap. O estudo mostrou que as buscas em dicionários já proporcionam resultados competitivos em relação aos demais métodos.

Como o MetaMap e o cTAKES são as ferramentas mais utilizadas, há pesquisas comparando os dois sistemas [115, 122, 128], enquanto outros os combinam para obter melhor precisão [172].

3. TRABALHOS RELACIONADOS

Este capítulo resume o trabalho anterior cobrindo tarefas de PLN no domínio biomédico. Especificamente, as tentativas atuais de abordar a tarefa de REN serão apresentadas levando em consideração as arquiteturas e metodologias seguidas pelos autores. Para uma revisão abrangente, dividimos este capítulo em várias seções, incluindo aspectos importantes da literatura atual, como o domínio biomédico e abordagens aplicadas à tarefa REN.

3.1 Domínio Biomédico

BioNLP refere-se aos métodos e estudos de como a mineração de texto pode ser aplicada a textos e literatura da área biomédica e outros subdomínios mais específicos, como radiologia, oncologia e farmacologia [84]. Além disso, o BioNLP é frequentemente usado pelos serviços de saúde, pois traz benefícios como reduzir a incerteza, apoiar a tomada de decisões baseada em evidências e oferecer interoperabilidade com os sistemas de saúde. Todos esses potenciais benefícios são brevemente descritos abaixo para mostrar alguns métodos e possíveis aplicações em o campo biomédico.

Muitos estudos envolvidos com o desenvolvimento de abordagens BioNLP foram dedicados à detecção de incertezas. Por exemplo, em tarefas de Recuperação de Informações (RI), a detecção de incertezas melhora os resultados da extração de informações de relatórios de radiologia [171].

Seguindo essa ideia, Vincze et al. [162] criaram o corpus BioScope, que é um recurso de acesso aberto para pesquisas sobre gerenciamento de incertezas em textos biomédicos. O corpus é composto por três partes, a saber, textos livres médicos, artigos biológicos completos e resumos científicos biológicos. Devido à sua prevalência e alto nível de incerteza biomédica, o câncer de mama [155] ou a pneumonia [62] são casos importantes para analisar o impacto da biomedicina na identidade da doença. No que diz respeito ao suporte à tomada de decisão baseada em evidências, refere-se a um sistema de tecnologia da informação em saúde projetado para fornecer aos médicos e outros profissionais de saúde o Suporte à Decisão Clínica (CDS), ou seja, assistência nas tarefas de tomada de decisão clínica. Os pesquisadores de ML e PLN podem desempenhar um papel fundamental em tornar as evidências mais transparentes, por exemplo, facilitando a busca e extração de descobertas relatadas [165].

Peiffer-Smadja et al. [114] focaram-se na avaliação do uso de sistemas de apoio à decisão em várias técnicas de ML, na avaliação dos resultados e nas implicações desses

sistemas de apoio à decisão em nível clínico em tempo real para o diagnóstico de problemas cardíacos.

Atualmente, o SARS-CoV-2 (COVID-19) está criando uma ameaça importante e urgente à saúde global. Dessa forma, muitos esforços estão sendo focados no desenvolvimento de soluções automatizadas para apoiar médicos especialistas na detecção precoce da doença com base em imagens e textos médicos [112, 4]. Modelos de previsão que combinam variáveis ou recursos para estimar o risco de pessoas serem infectadas estão ajudando os médicos a lidar com o surto de COVID-19 [5].

O último benefício mencionado é a interoperabilidade. Nesse caso, os modelos BioNLP podem contribuir para a solução dos problemas de interoperabilidade semântica e reutilização de conhecimento em sistemas de informação clínica. Seguindo a definição de Miranda et al. [105]:

> *"Interoperabilidade é a capacidade de sistemas independentes trocarem informações significativas e iniciarem ações uns dos outros, a fim de operarem juntos para benefício mútuo."*

A interoperabilidade é atualmente uma questão importante dentro da comunidade científica porque os S-RES usados nas organizações de saúde se desenvolveram de forma independente com ferramentas, métodos, processos e procedimentos que resultam em um grande número de modelos exclusivos e proprietários que representam e registram as informações do paciente [57]. Para garantir a interoperabilidade semântica entre os sistemas de saúde, é necessário o uso de padrões que permitam a troca de dados, bem como o uso de vocabulários normalizados e curados, que unifiquem os dados utilizados em diferentes instituições resultando na correta troca de informações. Alguns dos vocabulários padronizados são detalhados na Seção 2.6 anterior. Nesse contexto, a tarefa REN é uma das mais utilizadas, pois permite a extração de conhecimento que pode ser compartilhado de forma padronizada e compreensível [110].

Esta dissertação se concentra na extração de entidades nomeadas usando textos biomédicos como fonte de informação. O REN se adapta a qualquer situação em que uma visão geral de alto nível de uma grande quantidade de texto seja útil. Além disso, a tarefa de REN pode ser aplicada a uma variedade de sistemas de saúde para realizar a extração automática de conhecimento.

3.2 Reconhecimento de Entidades Biomédicas

Métodos sofisticados de processamento de informações são necessários para a aquisição e integração eficientes de dados de um corpus de literatura biomédica. A identificação efetiva dos termos é fundamental para acessar as informações armazenadas, pois

são os termos que representam o conhecimento nos textos. Devido à complexidade da terminologia biomédica que muda dinamicamente, a identificação de termos tem sido reconhecida como um desafio na mineração de texto e, como consequência, tornou-se um importante tópico de pesquisa tanto no PLN quanto nas comunidades biomédicas.

Como atualmente há um grande crescimento na demanda por compreensão e extração de informações de textos médicos, a comunidade de PLN organizou uma série de desafios abertos focados na extração de entidades biomédicas. Esses desafios costumam ter várias vantagens: fornecem um corpus disponível em diferentes idiomas; propõem um sistema básico de experimentação; eles fornecem aos participantes um método de avaliação e oferecem o estado da arte em uma tarefa e conjunto de dados específicos. Alguns dos desafios mais populares focados em tarefas do REN são descritos brevemente abaixo.

Por um lado e com foco em inglês, o **DDIExtraction** [140] foi apresentado no SemEval 2013. A tarefa dizia respeito ao reconhecimento de medicamentos e à extração de Interações Medicamentosas (DDI) incluídas na literatura biomédica. Este desafio foi dividido em duas subtarefas: o reconhecimento e classificação de substâncias farmacológicas e a extração de DDI onde os participantes poderiam submeter seus sistemas. O **N2C2 - National NLP Clinical Challenges Shared Task** [46] foi focado na extração de Eventos Adversos Relacionados a Medicamentos (EAMs) de prontuários clínicos e três subtarefas foram avaliadas: extração de conceito, classificação de relação e sistemas de ponta a ponta. Outras oficinas também foram propostas no passado para abordar a tarefa de EAMs em outros textos que não relacionados a biomedicina, mais especificamente usando tweets [169]. Em 2015, o desafio **CHEMDNER** [74] foi organizado pela BioCreative e promoveu o desenvolvimento de novos, competitivos e acessíveis sistemas de mineração de texto químico. A trilha **PharmaCoNER** (*Pharmacological Substances, Compounds and proteins Named Entity Recognition*) também foi proposta como uma tarefa de reconhecimento de entidades no domínio farmacológico. O objetivo principal foi encontrar menções de produtos químicos e medicamentos em casos clínicos. O desafio foi composto por duas subtarefas: i) compensação de REN e classificação de entidades, e ii) indexação de conceitos usando SNOMED-CT como vocabulário [39]. O **Cantemist** (CANcer TExt Mining Shared Task) foi a primeira tarefa focada no reconhecimento de entidades no domínio da oncologia [106]. Os participantes desta tarefa poderiam submeter sistemas às três subtarefas propostas pelos organizadores denominadas cantemist-NER, cantemist-NORM e cantemist-CODING. Outros desafios relacionados ao domínio Biomédico, como o **eHeatlh-KD** [94] (eHealth Knowledge Discovery), ao invés de utilizar entidades específicas da área médica, utilizam entidades de propósito geral.

Pesquisadores interessados em tarefas de extração de entidades exploraram uma variedade de abordagens de ML. Como descrevemos no Capítulo 2, as abordagens de ML formulam a tarefa clínica de REN como um problema de rotulagem de sequência que visa encontrar a melhor sequência de rotulagem a partir do texto clínico. Muitos estudos ante-

riores aplicaram o método CRF [80] com o objetivo de realizar a identificação e posterior classificação das entidades. O CRF é a solução mais popular entre os algoritmos de ML convencionais. Um modelo CRF típico geralmente usa recursos de diferentes níveis linguísticos, incluindo ontologias, léxicos, informações sintáticas ou *embeddings* de palavras [158]. SVM é outro algoritmo útil usado em ML tradicional para identificar entidades biomédicas [63, 150, 173]. Por exemplo, Takeuchi e Collier [156] focaram na identificação de entidades do domínio da biologia molecular. Eles usaram uma coleção de textos de resumos MEDLINE [1] para realizar o experimento. Além disso, eles adicionam ao sistema um conjunto de recursos linguísticos em nível de palavra, incluindo formas de superfície de palavras, tags de parte da fala e recursos ortográficos. O estudo mostrou que a combinação de algumas características atinge resultados elevados (cerca de 74% F1-score) neste domínio específico.

Os primeiros estudos da tarefa REN visavam principalmente as RNRs para produzir resultados promissores. Esses estudos demonstraram a grande eficácia da RNR aplicada à extração de entidades biomédicas usando arquiteturas de rede complexas [88, 158, 89]. Em comparação com os métodos tradicionais de ML, as RNRs geralmente usam uma camada de incorporação como entrada para aprender a representação vetorial das palavras [35, 116, 60, 69].

Em 2015, Huang, Xu e Yu [55] propuseram uma variedade de modelos LSTM para rotulagem de sequências, incluindo redes LSTM, redes BiLSTM, LSTM com uma camada CRF (LSTM-CRF) e BiLSTM com uma camada CRF (BiLSTM-CRF). Sua pesquisa descobriu que o modelo BiLSTM-CRF fez uso eficaz de recursos de entrada passados e futuros. Os modelos apresentados produzem precisão de última geração em rotulagem *Part-Of-Speech*, chunking e conjuntos de dados REN. Mais recentemente, Hong e Lee [50] introduziram o DTranNER, um novo framework baseado em CRF que incorpora um modelo de transição rótulo-rótulo baseado em aprendizado profundo em tarefas biomédicas de REN. Eles realizaram experimentos em cinco corpos de referência para comparar os métodos de ponta em cada um deles. O modelo DTraNER atinge o melhor F1-score em quatro conjuntos de dados, incluindo BC2GM [147], BC4CHEMD [74], BC5CDR [87] em conjuntos de dados químicos e de doenças, superando o popular modelo BioBERT [86] baseado em Transformers. No entanto, o BioBERT supera o DTraNER no corpus NCBI-Disease [87].

Embora as RNRs tenham obtido altos resultados e uma vasta literatura relacionada à tarefa REN nos últimos anos, o pré-treinamento de modelos de linguagem baseados em Transformers como o BERT [26] também levou a ganhos impressionantes em sistemas REN [67]. Alguns modelos pré-treinados baseados em BERT são até específicos para o domínio Biomédico como o BioBERT [86], que é pré-treinado em corpora biomédicos de larga escala e o ClinicalBERT, especializado em textos clínicos e apresentando melhora em algumas tarefas de PLN no domínio clínico [54, 2]. O SciBERT é um modelo de linguagem pré-

[1] MEDLINE

treinado em texto científico que demonstrou resultados semelhantes no campo Biomédico, mas melhora no domínio da ciência da computação em comparação com o BioBERT [11].

Dado o crescente número de modelos pré-treinados disponíveis, a literatura relacionada, assim como os resultados do estado da arte, está em constante mudança. Assim, todos os modelos descritos acima são frequentemente comparados, ajustando-os a diferentes domínios e corpora [42, 71, 11]. Os primeiros estudos da tarefa REN visavam principalmente RNRs para produzir resultados promissores. Esses estudos demonstraram a grande eficácia da RNR aplicado à extração de entidades biomédicas usando arquiteturas de rede complexas [88, 89, 158]. Em comparação com os métodos tradicionais de ML, as RNRs geralmente usam uma camada de incorporação como entrada para aprender a representação vetorial das palavras [35, 116, 60, 69].

No que se refere ao português, existem alguns modelos de pré-treinados para esta língua. Em 2020, o grupo de pesquisa *Health Artificial Intelligence Lab* (HAILab) da Pontifícia Universidade Católica do Paraná (PUCPR) desenvolveu o BioBERTPt, um modelo de linguagem, adaptando a metodologia do BioBERT [86], capaz de realizar REN em textos biomédicos no Português [139]. O modelo REN Farmacológico faz parte do projeto BioBERTpt, onde 13 modelos de entidades clínicas (compatíveis com UMLS) foram treinados. Todos os modelos REN da PUCPR foram treinados a partir do corpus clínico brasileiro SemClinBr [111], com 10 épocas e formato BILOU/IOBES, a partir do modelo BioBERTpt(all).

4. METODOLOGIA

A seguir, são descritos os métodos aplicados na presente abordagem. O presente trabalho foi desenvolvido em parceria com o grupo de pesquisa do Instituto de Avaliação de Tecnologia em Saúde (IATS), através do projeto aprovado pelo Comitê de Ética em Pesquisa do Hospital Moinhos de Vento e, intitulado, *Proposição de modelo de gestão de saúde baseada em valor para os sistemas de saúde pública e suplementar do Brasil*.

4.1 Descrição do Problema

O Reconhecimento de Entidades Nomeadas de medicamentos e produtos químicos é um passo fundamental para uma futura mineração de texto médico e tem recebido muita atenção recentemente. Esta tarefa visa detectar automaticamente menções a substâncias químicas e medicamentos na literatura biomédica e é um grande desafio para a comunidade científica por vários motivos: existem várias maneiras de se referir a uma mesma substância química ou medicamento, abreviaturas e siglas são comumente usadas, símbolos são frequentemente incluídos em evoluções clínicas e publicações científicas, além de que novos fármacos e produtos químicos são constantemente relatados [90]. Logo, para interoperar esses dados, é necessário identificar quais terminologias e padrões de comunicação em saúde são empregados para viabilizarmos um mecanismo para a integração com dados Farmacogenômicos. O PLN pode ser uma solução que proporciona uma detecção de conceito rápida, precisa e automatizada, que pode proporcionar avanços importantes para a comunidade científica de REN. De forma resumida, apresentamos na Figura 4.1, uma arquitetura cuja proposta está alinhada com a resolução do problema descrito acima.

Nesta arquitetura, cobrimos todo o processo para se produzir modelos de domínio específico na saúde. Iniciamos com a etapa de construção de um corpus específico para o domínio da saúde, e com o auxílio de especialistas no domínio, realizamos a anotação manual para então, com os dados anotados, realizar o refinamento do modelo de linguagem pré-treinado para resolver a tarefa de REN.

4.2 Corpus

Em parceria com o grupo de pesquisadoras do Instituto de Avaliação de Tecnologia em Saúde (IATS), recebemos uma amostra de um conjunto de dados contendo evoluções clínicas de pacientes, com histórico de doenças Cardiovasculares. A Tabela 4.1 apresenta um pequeno resumo de cada um dos conjuntos. Os valores únicos apresentam o total de

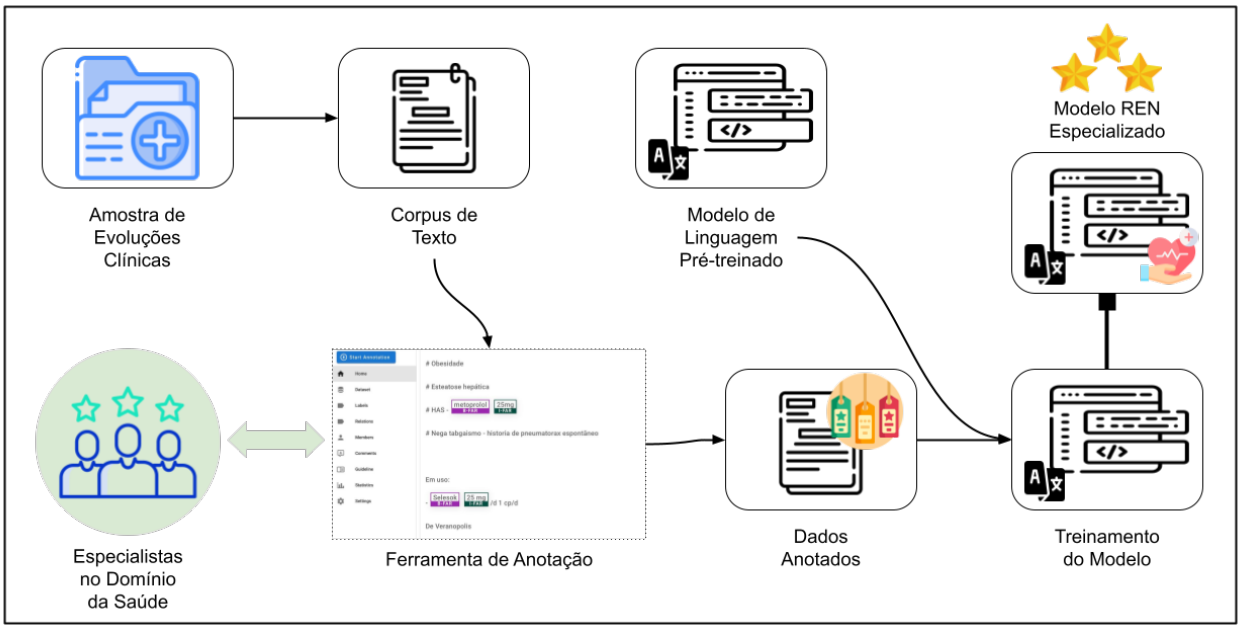

Figura 4.1 – Visão geral da arquitetura proposta

pacientes em cada um dos conjuntos, mas não reflete o número de evoluções clínicas de cada paciente.

	Total de Evoluções	Sentenças	Número de pacientes	Número de colunas	Colunas selecionadas	Possui Data?	Data de Recebimento
Conjunto 1	12083	251664	100	5	2	Não	30/10/2020
Conjunto 2	37773	748767	241	5	2	Não	13/11/2020
Conjunto 3	1558	47474	168	45	3	Sim	21/06/2021

Tabela 4.1 – Resumo numérico dos conjuntos recebidos apresentando número de evoluções clínicas por conjunto e número total de pacientes únicos.

Os dados disponíveis para esse estudo são os textos de evoluções clínicas de 341 pacientes hospitalizados no hospital-base para tratamento de AVC. Estes dados datam da primeira internação no dia 01/01/18 até a última alta no dia 02/04/2020.

4.2.1 Pré-processamento do Corpus

O pré-processamento de texto geralmente consiste em várias etapas que dependem de uma tarefa específica e do tipo de texto a ser tratado. Em tarefas textuais de PLN, isso significa que qualquer texto bruto precisa ser cuidadosamente pré-processado antes que o algoritmo possa processá-lo. Em nosso caso particular, trabalhamos com textos escritos em português, sem regras gramaticais e relacionados a diferentes subdomínios médicos. O pré-processamento realizado em todos os textos é o seguinte:

- Seleção de *features*: remoção de características que não são de interesse para o estudo;
- Separação de frases: esse processo consiste em dividir o texto em frases individuais. Para este procedimento, um algoritmo realizou a separação em sentenças através da detecção de caracteres de quebra de linha. A Figura 4.2 traz uma ilustração desta etapa;
- Alterar codificação: verificar se o texto estava codificado em UTF-8;

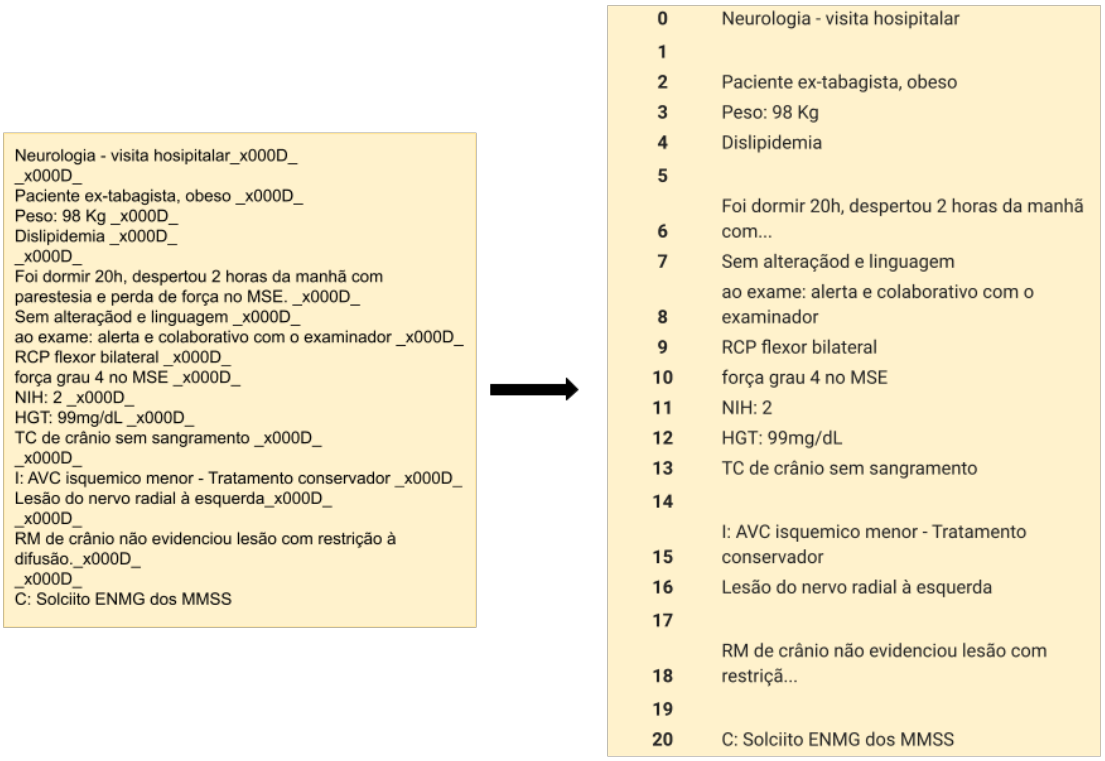

Figura 4.2 – Exemplo da obtenção de sentenças a partir do texto da evolução

4.3 Processo de Anotação Manual

Surpreendentemente, existem muitos corpora de texto anotados manualmente que são distribuídos juntamente sem as diretrizes que descrevem como as anotações foram geradas. Esses corpora de caixa preta têm a desvantagem de não poderem ser estendidos, afinal, é impossível compará-los de forma significativa com outros corpora pois não saberíamos lidar com possíveis inconsistências e erros de anotação. As diretrizes de anotação devem especificar as instruções necessárias para identificar os elementos de texto que devem ser marcados (e aqueles que não devem ser marcados) e como atribuí-los à classe de

entidade correspondente. Em um nível geral, eles representam as instruções sobre como o esquema de anotação deve ser aplicado aos dados de texto reais que serão rotulados.

Três etapas importantes precisam ser abordadas nas diretrizes de anotação: (*i*) o que rotular, (*ii*) os limites de menção desses rótulos e (*iii*) como classificar essas menções em categorias de medicamentos. A criação de diretrizes de alta qualidade que se encaixam na tarefa de anotação exigiu um processo iterativo de várias etapas: começando de um rascunho de diretriz inicial até que diretrizes claras e refinadas fossem obtidas. As definições aqui fornecidas (Apêndice A) são uma tentativa de exemplificar o que está sendo anotado e como cenários ambíguos devem ser tratados. Seguindo uma metodologia padrão para anotação de dados, construímos anotações médicas de alta qualidade.

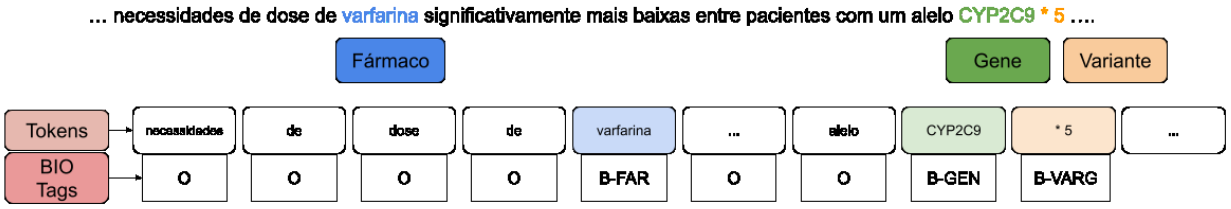

Figura 4.3 – Utilização do esquema BIO para anotação dos Fármacos

Escolher o esquema de anotação ideal é um problema complexo [73] e o impacto do uso de diferentes esquemas de anotação no desempenho do REN [3] precisa ser considerado. Nesta solução, inicialmente, optamos pela anotação somente de **Fármacos** no esquema BIO (Figura 4.3). Esse esquema também é referido na literatura como IOB e foi adotado pela Conferência sobre Aprendizagem de Linguagem Natural Computacional (CoNLL) [132]. Como discutimos anteriormente na Seção 2.5, atribuímos uma tag a cada palavra no texto, determinando se é o início (B) de uma entidade nomeada conhecida, o interior (I) dela ou se estamos nos referindo a uma palavra fora de qualquer entidade nomeada conhecida (O). Por último, é importante destacar que o rótulo FAR pode ser mapeado para o conceito de *Pharmacologic Substance* da UMLS [13].

4.4 Modelos implementados

Nesta pesquisa, foram testados dois modelos para o REN relacionadas a medicamentos: uma CNN e o BERT na abordagem *fine-tuning*. A CNN foi implementada em Python usando a biblioteca de código aberto spaCy [51]. Para o modelo BERT utilizou-se a arquitetura BERT_{BASE}, com o modelo pré-treinado BERT-Português [152], disponíveis na biblioteca Transformers, desenvolvida pelo HuggingFace, também, para a linguagem Python.

A spaCy fornece uma variedade de ferramentas práticas para processamento de texto em vários idiomas. Seus modelos surgiram como o padrão de fato para PLN prático devido à otimização da sua velocidade em CPUs, robustez e desempenho próximo ao estado da arte [52]. Como os modelos spaCy são populares e a API spaCy é amplamente conhecida por muitos usuários em potencial, optamos por construir sobre a biblioteca spaCy para criar um pipeline de processamento de texto Biomédico.

A arquitetura da SpaCy é baseada em CNNs com tokens representados como *embeddings Bloom em hash* [142] de prefixo, sufixo e lematização de palavras individuais aumentadas com um modelo de agrupamento baseado em transição [82]. Especificamente para REN é usada uma estrutura de quatro etapas: vetorização, codificação, atenção e predição. Primeiro as palavras são alteradas para representações vetoriais, e, dada uma sequência de vetores dessas palavras, a etapa de codificação calcula uma matriz de frase, levando em consideração o contexto, utilizando uma CNN para codificação. A camada de atenção da CNN reduz a representação da matriz produzida pela etapa de codificação para um único vetor, que é passado para uma rede MLP+Softmax para predição.

O BERT-Português foi pré-treinado no brWac [164], um corpus de português brasileiro contendo 2.68 bilhões de tokens extraídos de documentos da internet. O seu vocabulário *WordPiece* contém 230 mil unidades.

Com a popularização do BERT, várias bibliotecas Python forneceram implementações prontas o treinamento desses modelos em tarefas específicas. Na abordagem *fine-tuning*, adotamos a implementação do BERT para classificação de palavras da HuggingFace na versão baseada em Pytorch. Essa arquitetura é composta de um modelo pré-treinado BERT seguido de uma camada linear com ativação *Softmax*. Para cada token, a predição final é dada pela etiqueta cujo modelo atribui maior probabilidade, chamada de função *argmax*. O BERT foi testado com o modelo BERT-PT. A Figura 4.4 traz uma ilustração do modelo para a tarefa REN.

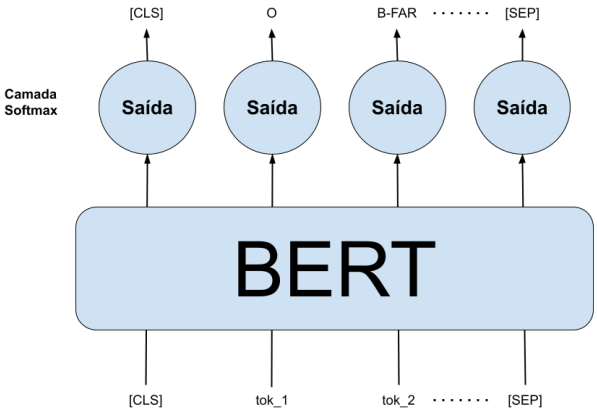

Figura 4.4 – Representação BERT para REN.

4.5 Preparação das entradas para as RNs

Seguindo a metodologia de Souza, Nogueira e Lotufo [152], está sendo considerada como uma sequência de entrada das redes neurais uma evolução clínica do corpus. Ao utilizar uma evolução, as redes neurais tem um contexto extenso para extrair informações para a representação da palavra.

Entrada da CNN

O modelo CNN da spaCy recebe como entrada uma lista contendo tuplas que possuem o texto e outra lista de tuplas com as informações sobre as classificações deste texto. Essas classificações indicam as posições inicial e final da parte classificada no texto, bem como qual a classe foi atribuída. O tipo de entrada de dados no modelo pode ser visualizado da seguinte forma: array(texto, 'entities': [(posição inicial, posição final, nome da entidade)]). Na Tabela 4.2 encontramos um exemplo do formato descrito acima.

id	text	tokens
0	Evolução emergência/ CN 3\\Paracetamol\\(...)	[{'text': 'Evolução', 'start': 0, 'end': 8, 'id': 0, 'ws': True}, {'text': 'emergência/', 'start': 9, 'end': 20, 'id': 1, 'ws': True}, {'text': 'CN', 'start': 21, 'end': 23, 'id': 2, 'ws': True}, {'text': '3', 'start': 24, 'end': 25, 'id': 3, 'ws': False}, {'text': 'Paracetamol', 'start': 27, 'end': 38, 'id': 5, 'label': 'FAR'}, (...)]

Tabela 4.2 – Exemplo dos dados exportados do anotador

Entrada do BERT

Cada evolução do corpus pré-processado é composta de tuplas de um token e o rótulo correspondente (Tabela 4.3).

id	tokens	ner_tags
0	[Evolução, emergência/, CN, 3, Enfermage...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, ...

Tabela 4.3 – Corpus tokens e respectiva ner_tag.

Antes de alimentar os tokens como entrada do BERT, é preciso pré-processá-los. O primeiro passo foi converter os tokens em *WordPieces* com o tokenizador da biblioteca Transformers. Além disso, no início e no final de cada sequência foram inseridos, nessa ordem, os tokens especiais [CLS] e [SEP], discutidos no Capítulo 2.

Isso gera um desalinhamento entre os tokens e as etiquetas, assim um pré-processamento semelhante foi aplicado à sequência de etiquetas para realinhá-los. Para isso, utilizou-se uma copia do rótulo do primeiro WordPiece de cada palavra para mapear os tokens iniciados por ##, ausentes antes da tokenização .Também foram inseridas três etiquetas “CLS”, “SEP” e “PAD” para mapear os tokens especiais [CLS], [SEP] e [PAD]. O token [PAD] é usado para preencher posições na normalização do tamanho da entrada.

Tamanho da Entrada

Em redes neurais como o BERT, é esperado que as sequências de entrada sejam normalizadas para o mesmo tamanho. Desse modo, foi definido um hiperparâmetro para o modelo: o tamanho máximo *S* da sequência de entrada. Após o pré-processamento, todas as entradas foram limitadas ao tamanho máximo estabelecido. Sequências menores que *S* foram estendidas até o tamanho máximo pela inserção do token especial [PAD]. Já os maiores que *S*, foram quebrados em períodos de tamanho *S* contando um passo *P* para indicar o início da sequência seguinte.

4.6 Hiperparâmetros dos modelos

Na CNN, utilizamos como hiperparâmetros otimização Adam com $\beta_1 = 0,9$ e $\beta_2 = 0,999$ e taxa de aprendizado inicial de 1e-5. No BERT, adotamos alguns hiperparâmetros indicados por Souza, Nogueira e Lotufo [152]: 15 épocas de treinamento com *batch* tamanho 3, otimização AdamW [95] com $\beta_1 = 0,9$ e $\beta_2 = 0,999$, *weigth decay* de 0,01 e taxa de aprendizado inicial de 1e-5. O tamanho máximo de $S = 512$ com passo $P = 128$.

4.7 Treinamento e Avaliação

Durante o treinamento, dividiu-se o corpus em duas partes: treino (75%) e teste(25%). Somente o modelo BERT-PT no esquema BILOU foi testado em três diferentes cenário: o cenário total (980 evoluções), considerando todo o conjunto de dados anotados, o cenário meio (490 evoluções), contendo metade do conjunto de dados anotados e o cenário intermediário, onde usamos 75% do conjunto de dados anotados (735 evoluções). Na Tabela 4.4 podemos observar o cenário total separado pela quantidade de anotações por tag, e o conjunto que será utilizado para validação, com um total de 196 evoluções.

Para a validação, utilizamos *5-fold-validation*, uma técnica usada para evitar um resultado enviesado pelo treinamento. Esse método consiste em treinar o mesmo modelo

Label	Treino	Teste	Total	Validação
O	99432	33151	132583	34354
B-FAR	742	244	986	200
I-FAR	84	25	109	14

Tabela 4.4 – Cenário completo do conjunto de treino e teste separados por tag anotada

por 5 rodadas diferentes, embaralhando aleatoriamente o corpus de treinamento a cada rodada. Assim, cada uma das rodadas receberá um conjunto de treino distinto, evitando o viés dos dados e fornecendo um resultado mais confiável. O desempenho é obtido pela média das 5 rodadas testadas. As métricas de avaliação de desempenho são aquelas propostas por COnLL: precisão, *recall* e medida-F, como descritas no Capítulo 2. A plataforma *Google Colaboratory* (Colab) serviu como ambiente de desenvolvimento para a implementação das redes neurais usando uma GPU Tesla T4. O Colab é uma ferramenta que permite escrever e rodar código Python.

4.8 Amarrando o Conhecimento

Uma dos grandes desafios desta dissertação estava no fato de unir a informação contida nas evoluções clínicas com o conhecimento de bases farmacogenômicas. Vários aspectos da farmacogenômica devem ser padronizados para fornecer semântica comum entre sistemas distintos [78]. Os elementos-chave de dados incluem quais variantes genéticas devem ser coletadas, termos de fenótipo e medicamentos envolvidos na interação gene-medicamento.

Também são necessários padrões para representar o conhecimento farmacogenômico, incluindo tabelas de tradução, recomendações clínicas e níveis de evidência, bem como a evolução desse conhecimento ao longo do tempo. Portanto, para manter a relevância e acompanhar a rápida geração de evidências, os recursos de conhecimento atualmente armazenados, mantidos e acessados localmente, precisarão migrar para arquiteturas mais orientadas a serviços [153, 120, 98].

Observando o andar científico, Hoffman et al. teorizaram um fluxo que agrupa dados do paciente através de um mecanismo de busca. Após a coleta, esses dados são combinados com as bases de conhecimento para fornecer uma interface para os S-RES (Figura 4.5).

Observando a imagem, poderíamos afirmar que uma possibilidade de *Query Engine* (à esquerda) para extrair listas de medicamentos através de uma detecção de conceito rápida e precisa, estaria no mecanismo de REN. Após estabelecido que a lista de medicamentos será extraída através de REN, passamos para a etapa de combinar recursos

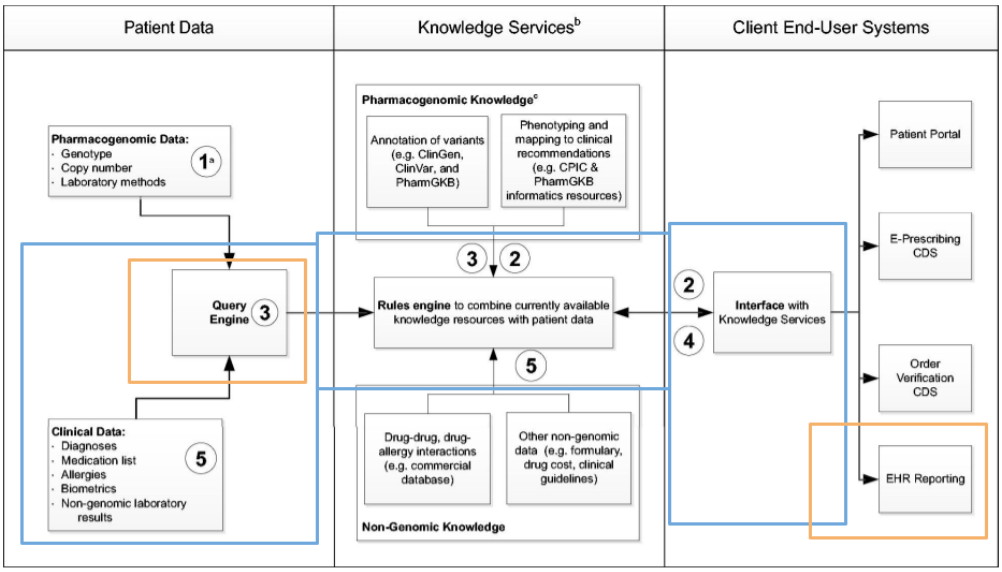

Figura 4.5 – Fluxo de informações idealizado para consulta de recursos de conhecimento farmacogenômico e retorno a um S-RES com suporte à decisão clínica. Adaptado de Hoffman et al. [49]

de conhecimento atualmente disponíveis com dados do paciente (bloco central). Ao analisarmos as principais bases de dados farmacogenômicas, encontramos que para realizar consultas, precisamos que os nomes dos medicamentos estejam no nome oficial genérico e não comercial de uma substância farmacológica, também conhecido como Denominação Comum Internacional (DCI).

Como forma de viabilizar a interoperabilidade semântica, uma abordagem bastante utilizada é o desenvolvimento e a aplicação de mapeamentos entre terminologias e vocabulários. A utilização de mapeamentos entre terminologias, por exemplo, em um S-RES, pode auxiliar tanto na identificação de termos novos, quanto como meio de proposição de termos de fontes diferentes. Portanto, é necessário uma etapa intermediária para a construção de uma ferramenta que permita o mapeamento (Figura 4.6) entre a Denominação Comum Internacional (DCI) e a Denominação Comum Brasileira (DCB).

DCI (INN em inglês):	Paracetamol
BAN - British Approved Names (Nomes Aprovados no Reino Unido):	Paracetamol
USAN - United States Adopted Name (Nome Adotado nos Estados Unidos):	Acetaminophen
DCB - Denominação Comum Brasileira (Nome Adotado no Brasil):	Paracetamol
Outros nomes genéricos:	N-acetil-p-aminofenol, APAP, p-Acetamidofenol, Acetamol, ...
Nomes comerciais:	Tylenol®, Gelocatil®, Panadol®, Panamax®, Perdolan®, Calpol®, Doliprane®, Tachipirina®, ben-u-ron®,Atasol®,...
Nomenclatura IUPAC:	N-(4-hidroxifenil)etanamida

Figura 4.6 – Exemplo de nomeações do *Paracetamol* em diferentes contextos. Fonte: [33]

Dessa forma, é possível verificar as complementaridades entre estas terminologias e indicar novos códigos para um ou mais vocabulários.

5. EXPERIMENTOS

Neste capítulo, serão apresentados os resultados obtidos durante os experimentos de acordo com a metodologia detalhada no Capítulo 4. Durante a Seção 5.1 são apresentados os resultados separados do modelo CNN, na Seção 5.2 contém os resultados do obtidos pelo modelo BERT. Ao final, fornecemos um quadro comparativo de trabalhos no REN do Português na área Farmacológica.

5.1 Resultados da CNN

Nesta seção, será discutido o resultado da CNN. Como podemos ser visto na Tabela 5.1, em geral, o modelo obteve bons resultados nesta etapa de avaliação, alcançando uma precisão geral e medida-F acima de 80%. Uma observação a se fazer é que os resultados também são bons quando tratamos da classe alvo, ou seja, quando queremos encontrar o rótulo B-FAR.

Métrica	**B-FAR**	**I-FAR**	**Geral**
Precisão	0.9313	0.8461	0.9565
Recall	0.9313	0.6875	0.7586
Medida-F	0.9313	0.7586	0.8461

Tabela 5.1 – Precisão, *Recall* e Medida-F calculadas para cada classe e também para o modelo

Apesar destes resultados, vemos que o modelo desenvolvido foi capaz de reconhecer padrões e indicar quando existiam fármacos em uma frase do domínio Biomédico. À respeito do *Recall*, o experimento indicou que o modelo proposto obteve um desempenho satisfatório quando tratamos da classe positiva (B-FAR). Ou seja, quando realmente pertence à classe positiva, o modelo é capaz de identificar corretamente em, aproximadamente, 70% dos casos.

5.2 Resultados do BERT

Nesta seção, serão discutidos os resultados do BERT na abordagem *fine-tuning*. Todos os resultados foram avaliado em relação ao cenário total (980 evoluções). O desempenho é calculado pela média dos resultados obtidos na etapa de validação em 5 rodadas. Na Tabela 5.2, seguem os resultados do BERT-PT.

Esquema de Anotação	Cenário	Precisão	Recall	Medida-F
BIO	total (100%)	0.9221	0.9787	0.9496
BILOU	total (100%)	0.8785	0.8913	0.8848
BILOU	intermediário (75%)	0.9527	0.9029	0.9272
BILOU	meio (50%)	0.7619	0.8421	0.8

Tabela 5.2 – Medidas de desempenho do BERT

Dispersão dos dados de Teste

Quando uma amostra de dados não é muito grande, como neste caso, a média torna-se uma medida de baixa segurança, pois é facilmente enviesada por valores extremos. Para fornecer confiabilidade nos resultados obtidos, avaliamos as medidas de dispersão para as 5 rodadas de validação em cada cenário. Os gráficos boxplot na Figura 5.1 representam a precisão, *recall* e a medida-F obtidas na validação do BERT-PT para o cenário total. A linha cortando o retângulo é a mediana, o triângulo preto demarca a média e os pontos mínimo e máximo são indicados pelas hastes.

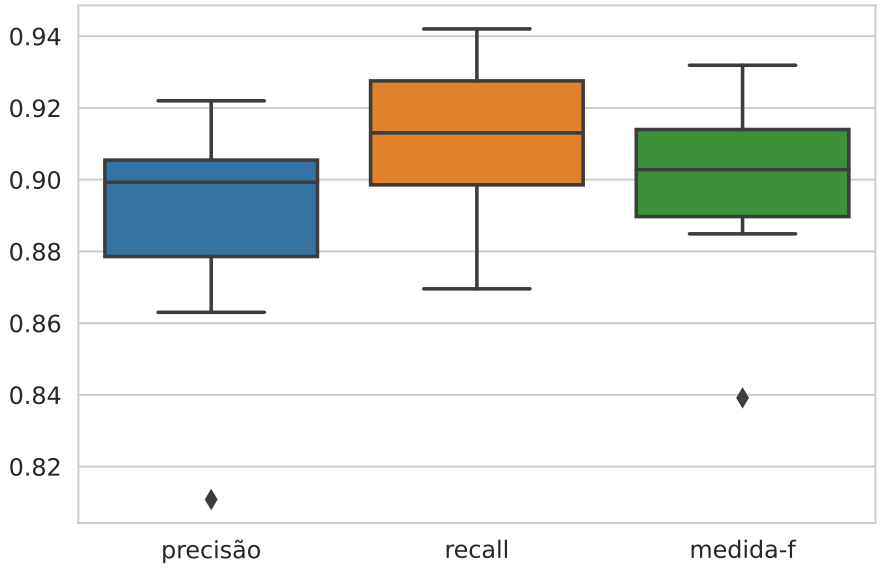

Figura 5.1 – Dispersão das medidas de desempenho do BERT-PT no esquema BILOU e no cenário total

5.3 Avaliação dos Modelos Implementados

Após apresentar os resultados de cada um dos modelos testados, dedicamos esta seção à comparação com pesquisas em redes neurais já discutidas no Capítulo 3. Para calcular as métricas utilizamos a biblioteca de avaliação proposta por Nakayama [121, 107].

Sistema	Cenário	Precisão	Recall	Medida-F
Clinicalnerpt-Pharmacologic BILOU (BioBERTpt) [139]	-	0.6938	0.6785	0.6938
CNN	100%	0.6744	0.5771	0.6219
BERT-PT BIO	100%	0.7142	0.6989	0.7142
BERT-PT BILOU	100%	0.7066	0.7066	0.7066
BERT-PT BILOU	75%	0.7091	0.7091	0.7091
BERT-PT BILOU	50%	0.7091	0.7091	0.7091

Tabela 5.3 – Comparação de desempenhos no cenário total

A partir desses resultados, é possível dizer que o BERT-PT BIO, atualmente, é o melhor modelo para o REN testado em um corpus manual anotado no domínio Farmacológico a partir de notas clínicas. Mesmo com a arquitetura BASE e uma camada de saída *Softmax*, o modelo chega a resultados promissores. Finalmente trouxemos um quadro geral de trabalhos recentes no REN do Português, comparando as arquiteturas mais adotadas e os resultados considerados o estado-da-arte com aqueles reportados em outras pesquisas, incluindo esta.

Um ponto de interesse desta pesquisa era tentar mensurar quanto dado deveria ser anotado para que as predições fossem robustas. Ao olhar na Tabela 5.3, observamos que ao reduzir o conjunto de treino, temos poucas oscilações nos valores enquanto que na etapa de avaliação, essa oscilação ocorreu de maneira abrupta. Um outro ponto é que o esquema utilizado, impactou levemente, onde os melhores resultados foram obtidos no BERT-PT BIO. Isso ocorreu em função da redução de labels do BIO para BILOU, e dado o estilo destes dados, é uma variável importante a ser considerada. Um outro ponto destacado por Schneider et al. [139] é que ao avaliar o BioBERTpt, eles descobriram que o domínio pode influenciar o desempenho de modelos baseados em BERT, particularmente para domínios com características únicas, como o médico. Entretanto, ao observarmos a Tabela acima, não encontramos diferenças significativas no treinamento quando partimos de um domínio mais genérico, como é o caso do BERT-PT.

5.4 Integração Farmacogenômica

Conforme apresentado na Tabela 5.4, a Agência Nacional de Vigilância Sanitária (Anvisa) oferece uma conjunto de dados que relaciona a DCB com a DCI através do identificador *Chemical Abstracts Service (CAS)*. Durante o decorrer desta pesquisa, solicitamos acesso à *CAS Common Chemistry API*, uma interface que fornece a capacidade de consultar o conjunto de dados *Common Chemistry* via HTTP.

id_dcb	dcb	id_cas
1	abacavir	136470-78-5
2	sulfato de abacavir	188062-50-2
3	abamectina	65195-55-3
4	abanoquila	90402-40-7
6	abaperidona	183849-43-6
	(...)	(...)

Tabela 5.4 – Conjunto de dados disponibilizados pela Anvisa

O retorno no formato *JavaScript Object Notation* (JSON), contém um dicionário, onde a chave **'name'** é capaz de entregar o nome do medicamento. Em conjunto com os dados da Tabela 5.4, podemos agregar a coluna **dci**, e assim termos um mecanismo de mapeamento da DCB para DCI, como apresentado na Tabela 5.5.

id_dcb	dcb	id_cas	dci
4766	ibuprofeno	15687-27-1	ibuprofen

Tabela 5.5 – Agrupamento realizado após retorno da chamada GET

Uma vez que tenhamos a DCI correspondente da DCB, podemos mesclar esse dado com as informações locais de bases farmacogenômicas. Como exemplo, solicitamos o conjunto de dados da base PharmGKB contendo informações de medicamentos. Essa base possui 24 chaves, sendo uma delas a correspondente do nome advindo da DCI. Portanto, como está apresentado na Figura 5.2 através da execução do nosso script, o **Ibuprofeno** é mapeado para o **id_cas** que solicita ao serviço do CAS a **dci** que ao ser buscada em bases farmacogenômicas, podemos encontrar conhecimento para construir relatórios que sirvam como base para auxiliar na decisão clínica.

Na Figura 5.3, podemos observar uma visão geral da arquitetura produzida.

5.5 Discussão

O reconhecimento de nomes de medicamentos e produtos químicos visa reconhecer tipos de menções em textos médicos não estruturados e classificá-los em categorias

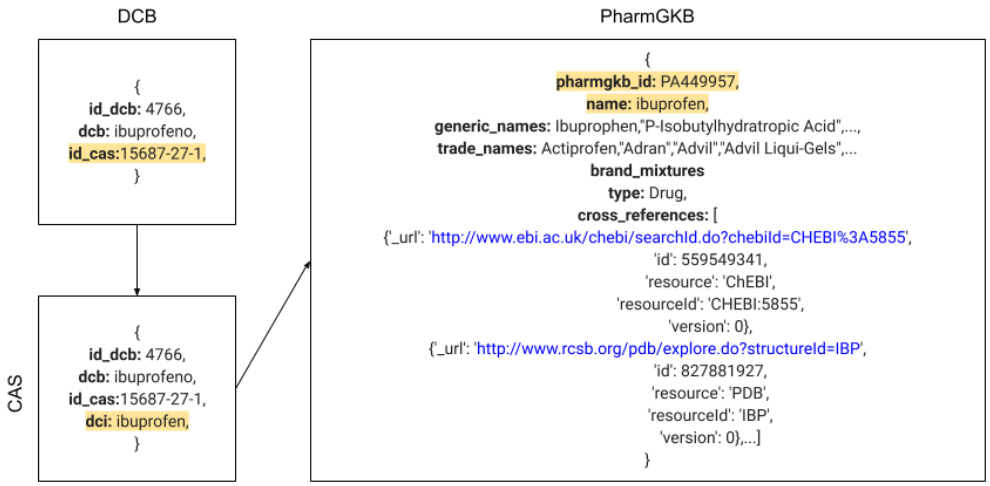

Figura 5.2 – Exemplo de mapeamento

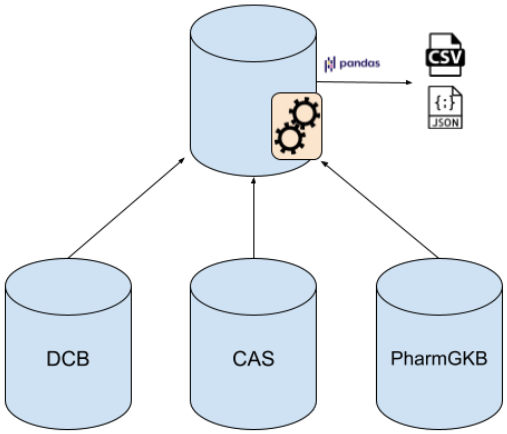

Figura 5.3 – Arquitetura resultante da estratégia de mapeamento

pré-definidas. Esses tipos de tarefas são fundamentais para a extração de informações médicas e sistemas de extração de relações médicas [168, 109, 135].

Dado o grande crescimento da comunidade científica pesquisando o subdomínio Farmacológico, a comunidade clínica de PLN organizou uma série de desafios abertos com foco na identificação de entidades químicas e medicamentosas a partir de notas clínicas narrativas. Estas oficinas são muito úteis porque os participantes utilizam sistemas inovadores e atualizados, oferecendo uma abordagem de última geração para as tarefas. Nosso modelo mostra diferentes modelos fazendo a mesma tarefa: encontrar medicamentos em anotações clínicas. Os resultados são promissores, pois obtemos 95.27% de precisão, 90.20% de *recall* e 92.29% de Medida-F. Ressalta-se que os resultados obtidos no corpus já foram elevados, portanto, alcançar melhorias permite novas pesquisas neste campo. No reconhecimento de medicamentos e produtos químicos em português, alcançamos resultados de ponta. Especificamente, mostramos uma melhoria de 2,04% no Medida-F e 2,04% de precisão em comparação com a melhor pesquisa até agora [139].

A análise dos resultados é um passo importante em nosso estudo. Com esta análise, podemos considerar as futuras melhorias do nosso sistema. Observamos vários casos de erro em que nosso sistema não foi rotulado corretamente. Nossa análise sugeriu que poderíamos melhorar a tokenização de nossos textos, pois às vezes não separava os tokens de forma mais eficaz. Também poderíamos tratar entidades marcadas como consecutivas, mas identificadas independentemente pelo nosso sistema ou o contrário, por exemplo, nosso sistema reconhece "levolisinato"e "ibuprofeno", porém a entidade correta seria "levolisinato de ibuprofeno".

6. CONCLUSÕES

O enorme crescimento de dados Biomédicos armazenados eletronicamente tornou a extração de conhecimento uma tarefa importante neste domínio. Os documentos de saúde podem incluir evidências relevantes, como descobertas, doenças e tratamentos, que podem ajudar os profissionais de saúde na tomada de decisões. No entanto, essas informações são difíceis de processar manualmente pelos profissionais devido ao tempo e custo envolvidos, sendo necessária a geração de recursos automáticos. Um dos objetivos do PLN é facilitar essas tarefas ao possibilitar o uso de métodos automatizados que extraem conhecimento de um texto com alta validade e confiabilidade. Especificamente, a tarefa REN aplicada ao domínio Biomédico visa extrair e identificar entidades de interesse que possam ser utilizadas pelos profissionais de saúde.

O Reconhecimento de Entidades Biomédicas é uma tarefa importante que ainda não foi resolvida, mas pode ajudar em outros sistemas relacionados à medicina. Por exemplo, o REN pode identificar achados importantes que são essenciais para cuidados de saúde seguros e eficazes. Além disso, essa tarefa pode ser aplicada a outras tarefas PLN, como classificação de texto, servindo como ponto de suporte. Por fim, o REN pode ser aplicado a vários subdomínios da saúde,como farmacologia, identificando medicamentos, ou eventos adversos e oncologia, reconhecendo achados relacionados ao câncer.

Os avanços na aprendizagem profunda nos motivaram a aplicá-los na tarefa REN para o domínio Biomédico em Português. A maior parte da pesquisa sobre a tarefa REN é realizada em Inglês. Portanto, esta dissertação visa avançar no estudo do reconhecimento de entidades em Português, a 5.ª língua mais falada no mundo, a 3.ª mais falada no hemisfério ocidental e a mais falada no hemisfério sul do planeta[1].

6.1 Contribuições

Esta pesquisa realizou uma série de estudos, análises e desenvolvimento de abordagens de PLN destinadas a abordar a tarefa de REN em textos Biomédicos no Português. Isso resultou em algumas contribuições:

1. Criação de corpora para REN no domínio da Farmacológico;
2. Disponibilização de diretrizes em Português para anotação manual de textos em Saúde;
3. Geração de novos Modelos de Linguagem para o Português:
    - CNN;

[1] Língua Portuguesa na Wikipédia

- FarBrBERT$_{BASE}$ no esquema BIO;
- FarBrBERT$_{BASE}$ no esquema BILOU;

4. Disponibilização de scripts para treino e avaliação para reprodução dos experimentos;
5. Disponibilização dos modelo FarBrBERT$_{BASE}$ no HuggingFace Hub;
6. Construção de ferramenta de mapeamento entre DCB e DCI permitindo a integração de dados Farmacogenômicos;
7. Artigo publicado:
    - Marques, F. B.; Leal, G. F.; Bettoni, G. N.; Souza, O. N. D. (2021). Integration of Bioinformatics and Clinical Data to Personalized Precision Medicine. In ITNG 2021 18th International Conference on Information Technology-New Generations (pp. 179-184). Springer, Cham.

6.2 Limitações

Uma possível limitação deste estudo é que ele foi elaborado sob a análise de apenas um conjunto de dados proveniente de um hospital parceiro. Uma vez que a escrita das evoluções podem variar entre profissionais e instituições, seria interessante validar a abrangência do modelo desenvolvido.

6.3 Trabalhos Futuros

Apesar dos resultados alcançados, ainda há potenciais trabalhos a serem desenvolvidos. Listaremos os principais trabalhos a serem realizados:

1. Como apresentamos na Figura 4.1 e na Tabela 4.4, a anotação dos fármacos ficou restrita a categoria **Fármaco**, faltando os modificadores de força, unidades e dose. Existem algumas outras abordagens que merecem ser investigadas e estudas, por exemplo:
    - *Medical Entity Linking (MEL)*: é a tarefa de identificar e padronizar menções em um texto médico não estruturado e vincular as menções às identidades únicas em uma determinada base de conhecimento médico [175];
    - *Named Entity Disambiguation (NED)*: é a tarefa de eliminar a ambiguidade no texto e vinculá-la ao conceito correto nas Bases de Conhecimentos [163];

2. Conforme apresentado pro Qin et al. [118], é difícil integrar os resultados farmacogenômicos aos S-RES porque os atuais S-RES comerciais não são projetados para armazenar informações genômicas em um formato adequado para uso a longo prazo, e os resultados farmacogenômicos geralmente são fornecidos como textos não estruturados em um arquivo PDF pelo laboratório de testes genômicos [16]. Na Visão Computacional, podemos escanear o documento, identificar a localização do texto e, finalmente, extrair o texto da imagem. Então, com técnicas de PLN, podemos extrair as entidades do texto e fazer a limpeza de texto necessária e analisar as entidades do texto para construir novos corpora [32].

3. Alguns estudos recentes mostram que usar metodologias também chamadas de *workflows* ou *pipelines* para criar corpus anotados adequadamente, através da utilização de estratégias como *Data Version Control* (DVC) tem melhorado os resultados para tarefas de REN em domínios específicos [27].

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Alessandrini, M.; Chaudhry, M.; Dodgen, T. M.; Pepper, M. S. “Pharmacogenomics and global precision medicine in the context of adverse drug reactions: Top 10 opportunities and challenges for the next decade”, *Omics: a journal of integrative biology*, vol. 20–10, 10 2016, pp. 593–603.

[2] Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; McDermott, M. “Publicly available clinical bert embeddings”, *arXiv preprint*, vol. 3, 06 2019, pp. 07.

[3] Alshammari, N.; Alanazi, S. “The impact of using different annotation schemes on named entity recognition”, *Egyptian Informatics Journal*, vol. 1, 11 2020, pp. 08.

[4] Apostolopoulos, I. D.; Mpesiana, T. A. “Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks”, *Physical and engineering sciences in medicine*, vol. 43–2, 04 2020, pp. 635–640.

[5] Arabi, Y. M.; Murthy, S.; Webb, S. “Covid-19: a novel coronavirus and a novel challenge for critical care”, *Intensive care medicine*, vol. 46–5, 05 2020, pp. 833–836.

[6] Aroyo, L.; Welty, C. “Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation”, *AI Magazine*, vol. 36–1, 03 2015, pp. 15–24.

[7] Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; et al.. “Gene ontology: tool for the unification of biology”, *Nature genetics*, vol. 25–1, 05 2000, pp. 25–29.

[8] Auger, A.; Barrière, C. “Pattern-based approaches to semantic relation extraction: A state-of-the-art”, *Terminology*, vol. 14–1, 06 2008, pp. 1–19.

[9] Bagley, S. C.; White, H.; Golomb, B. A. “Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain”, *Journal of clinical epidemiology*, vol. 54–10, 10 2001, pp. 979–985.

[10] Baker, P. G.; Brass, A.; Bechhofer, S.; Goble, C. A.; Paton, N. W.; Stevens, R.; et al.. “Tambis: Transparent access to multiple bioinformatics information sources.”, *Bioinformatics*, vol. 16, 2000, pp. 184–186.

[11] Beltagy, I.; Lo, K.; Cohan, A. “Scibert: A pretrained language model for scientific text”, *arXiv preprint*, vol. 3, 09 2019, pp. 06.

[12] Benson, T.; Grieve, G. “Principles of health interoperability: SNOMED CT, HL7 and FHIR”. Springer, 2016, 451p.

[13] Bodenreider, O. “The unified medical language system (umls): integrating biomedical terminology”, *Nucleic acids research*, vol. 32–01, 04 2004, pp. 267–270.

[14] Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. “Enriching word vectors with subword information”, *Transactions of the association for computational linguistics*, vol. 5, 06 2017, pp. 135–146.

[15] Boser, B. E.; Guyon, I. M.; Vapnik, V. N. “A training algorithm for optimal margin classifiers”. In: Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[16] Caraballo, P. J.; Hodge, L. S.; Bielinski, S. J.; Stewart, A. K.; Farrugia, G.; Schultz, C. G.; Rohrer-Vitek, C. R.; Olson, J. E.; Sauver, J. L. S.; Roger, V. L.; et al.. “Multidisciplinary model to implement pharmacogenomics at the point of care”, *Genetics in Medicine*, vol. 19–4, 09 2017, pp. 421–429.

[17] Carvalho, I. S. d. C. “O papel da farmacogenómica na investigação clínica”, Tese de Doutorado, Universidade de Coimbra, 2018, 73p.

[18] Ciampi, M.; Esposito, A.; Guarasci, R.; De Pietro, G. “Towards interoperability of ehr systems: The case of italy.” In: Proceedings of the International Conference on Information and Communication Technologies for Ageing Well (ICT4AgeingWell) and e-Health, 2016, pp. 133–138.

[19] Cohen, A. M.; Hersh, W. R. “A survey of current work in biomedical text mining”, *Briefings in bioinformatics*, vol. 6–1, 03 2005, pp. 57–71.

[20] Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. “Unsupervised cross-lingual representation learning at scale”, *arXiv preprint*, vol. 1, 11 2019, pp. 12.

[21] Cui, Y.; Chen, Z.; Wei, S.; Wang, S.; Liu, T.; Hu, G. “Attention-over-attention neural networks for reading comprehension”, *arXiv preprint*, vol. 3, 08 2016, pp. 10.

[22] da Saúde Brasil, M. “Portaria 2073”. Capturado em: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073_31_08_2011.html, 14 jan 2021.

[23] da Saúde Brasil, M. “Estratégia de saúde digital para o brasil 2020-2028”. Capturado em: https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf, 14 jan 2021.

[24] Demner-Fushman, D.; Chapman, W. W.; McDonald, C. J. “What can natural language processing do for clinical decision support?”, *Journal of biomedical informatics*, vol. 42–5, 11 2009, pp. 760–772.

[25] Demner-Fushman, D.; Elhadad, N.; Friedman, C. “Natural language processing for health-related texts”. In: *Biomedical Informatics*, Springer, 2021, pp. 241–272.

[26] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint*, vol. 1, 05 2018, pp. 16.

[27] Digan, W.; Névéol, A.; Neuraz, A.; Wack, M.; Baudoin, D.; Burgun, A.; Rance, B. “Can reproducibility be improved in clinical natural language processing? a study of 7 clinical nlp suites”, *Journal of the American Medical Informatics Association*, vol. 28–3, 01 2021, pp. 504–515.

[28] Donnelly, K.; et al.. “Snomed-ct: The advanced terminology and coding system for ehealth”, *Studies in health technology and informatics*, vol. 121, 07 2006, pp. 279.

[29] Duchi, J.; Hazan, E.; Singer, Y. “Adaptive subgradient methods for online learning and stochastic optimization.”, *Journal of machine learning research*, vol. 12–7, 07 2011, pp. 39.

[30] Egorov, S.; Yuryev, A.; Daraselia, N. “A simple and practical dictionary-based approach for identification of proteins in medline abstracts”, *Journal of the American Medical Informatics Association*, vol. 11–3, 06 2004, pp. 174–178.

[31] Enshaei, A.; Robson, C.; Edmondson, R. “Artificial intelligence systems as prognostic and predictive tools in ovarian cancer”, *Annals of surgical oncology*, vol. 22–12, 03 2015, pp. 3970–3975.

[32] Esteva, A.; Chou, K.; Yeung, S.; Naik, N.; Madani, A.; Mottaghi, A.; Liu, Y.; Topol, E.; Dean, J.; Socher, R. “Deep learning-enabled medical computer vision”, *NPJ digital medicine*, vol. 4–1, 01 2021, pp. 1–9.

[33] Foundation, W. “Denominação comum internacional”. Capturado em: https://pt.wikipedia.org/wiki/Denomina%C3%A7%C3%A3o_Comum_Internacional, 16 jan 2021.

[34] Geraci, A. “IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries”. IEEE Press, 1991, 217p.

[35] Giorgi, J. M.; Bader, G. D. “Towards reliable named entity recognition in the biomedical domain”, *Bioinformatics*, vol. 36–1, 01 2020, pp. 280–286.

[36] Gobinda, G. C. “Natural language processing”, *Annual Review of Information Science and Technology*, vol. 37, 01 2003, pp. 51–89.

[37] Goldberg, Y. “Neural network methods for natural language processing”, *Synthesis lectures on human language technologies*, vol. 10–1, 03 2017, pp. 1–309.

[38] Gomes, F.; Freitas, R.; Ribeiro, M.; Moura, C.; Andrade, O.; Oliveira, M. “Girls, a gateway for interoperability of electronic health record in low-cost system:* interoperability beetween fhir and openehr standards”. In: Proceedings of the 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM), 2019, pp. 1–6.

[39] Gonzalez-Agirre, A.; Marimon, M.; Intxaurrondo, A.; Rabal, O.; Villegas, M.; Krallinger, M. “Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track”. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 1–10.

[40] Goodfellow, I.; Bengio, Y.; Courville, A. “Deep learning”. MIT press, 2016, 775p.

[41] Grishman, R.; Sundheim, B. M. “Message understanding conference-6: A brief history”. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 1, 1996, pp. 06.

[42] Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. “Domain-specific language model pretraining for biomedical natural language processing”, *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3–1, 09 2021, pp. 1–23.

[43] Guarino, N.; Oberle, D.; Staab, S. “What is an ontology?” In: *Handbook on ontologies*, Springer, 2009, pp. 1–17.

[44] Hammond, W. E.; Jaffe, C.; Cimino, J. J.; Huff, S. M. “Standards in biomedical informatics”. In: *Biomedical informatics*, Springer, 2014, pp. 211–253.

[45] Hearst, M. A. “Untangling text data mining”. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 3–10.

[46] Henry, S.; Buchan, K.; Filannino, M.; Stubbs, A.; Uzuner, O. “2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records”, *Journal of the American Medical Informatics Association*, vol. 27–1, 01 2020, pp. 3–12.

[47] Hicks, J. K.; Aquilante, C. L.; Dunnenberger, H. M.; Gammal, R. S.; Funk, R. S.; Aitken, S. L.; Bright, D. R.; Coons, J. C.; Dotson, K. M.; Elder, C. T.; et al.. “Precision pharmacotherapy: integrating pharmacogenomics into clinical pharmacy practice”, *Journal of the American College of Clinical Pharmacy*, vol. 2–3, 06 2019, pp. 303–313.

[48] Hilbe, J. M. “Logistic regression models”. Chapman and hall/CRC, 2009, 03p.

[49] Hoffman, J. M.; Dunnenberger, H. M.; Kevin Hicks, J.; Caudle, K. E.; Whirl Carrillo, M.; Freimuth, R. R.; Williams, M. S.; Klein, T. E.; Peterson, J. F. “Developing knowledge resources to support precision medicine: principles from the clinical pharmacogenetics implementation consortium (cpic)”, *Journal of the American Medical Informatics Association*, vol. 23–4, 07 2016, pp. 796–801.

[50] Hong, S.; Lee, J.-G. “Dtranner: biomedical named entity recognition with deep learning-based label-label transition model”, *BMC bioinformatics*, vol. 21–1, 02 2020, pp. 1–11.

[51] Honnibal, M.; Montani, I. “spacy: Industrial-strength natural language processing in python”. Capturado em: https://spacy.io, 11 jul 2020.

[52] Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. “spacy: Industrial-strength natural language processing in python”. Capturado em: https://zenodo.org/record/3701227#.YsJKpXbMI_4, 26 jun 2020.

[53] Hornik, K. “Approximation capabilities of multilayer feedforward networks”, *Neural networks*, vol. 4–2, 10 1991, pp. 251–257.

[54] Huang, K.; Altosaar, J.; Ranganath, R. “Clinicalbert: Modeling clinical notes and predicting hospital readmission”, *arXiv preprint*, vol. 2, 03 2019, pp. 19.

[55] Huang, Z.; Xu, W.; Yu, K. “Bidirectional lstm-crf models for sequence tagging”, *arXiv preprint*, vol. 1, 08 2015, pp. 10.

[56] Information, H.; Society, M. S. “HIMSS dictionary of health information technology terms, acronyms, and organizations”. CRC Press, 2017, 426p.

[57] Iroju, O.; Soriyan, A.; Gambo, I.; Olaleke, J.; et al.. “Interoperability in healthcare: benefits, challenges and resolutions”, *International Journal of Innovation and Applied Studies*, vol. 3–1, 04 2013, pp. 262–270.

[58] Jaffe, C.; Nguyen, V.; Kubick, W. R.; Cooper, T.; Leftwich, R. B.; Hammond, W. E. “Standards in biomedical informatics”. In: *Biomedical Informatics*, Springer, 2021, pp. 205–240.

[59] JANg, E.; Gu, S.; Poole, B. “Categorical reparameterization with gumbel-softmax”, *arXiv preprint*, vol. 1, 11 2016, pp. 13.

[60] Jiang, J.; Wang, H.; Xie, J.; Guo, X.; Guan, Y.; Yu, Q. “Medical knowledge embedding based on recursive neural network for multi-disease diagnosis”, *Artificial Intelligence in Medicine*, vol. 103, 01 2020, pp. 101772.

[61] Jimeno, A.; Jimenez-Ruiz, E.; Lee, V.; Gaudan, S.; Berlanga, R.; Rebholz-Schuhmann, D. “Assessment of disease named entity recognition on a corpus of annotated sentences”. In: BMC bioinformatics, 2008, pp. 1–10.

[62] Jones, B.; South, B. R.; Shao, Y.; Lu, C.; Leng, J.; Sauer, B. C.; Gundlapalli, A. V.; Samore, M. H.; Zeng, Q. “Development and validation of a natural language processing tool to identify patients treated for pneumonia across va emergency departments”, *Applied clinical informatics*, vol. 9–01, 01 2018, pp. 122–128.

[63] Ju, Z.; Wang, J.; Zhu, F. “Named entity recognition from biomedical text using svm”. In: 5th international conference on bioinformatics and biomedical engineering, 2011, pp. 1–4.

[64] Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. “A convolutional neural network for modelling sentences”. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 655–665.

[65] Kalow, W. “Unusual responses to drugs in some hereditary conditions”, *Canadian Anaesthetists’ Society journal*, vol. 8–1, 01 1961, pp. 43–52.

[66] Kalyan, K. S.; Rajasekharan, A.; Sangeetha, S. “Ammu: A survey of transformer-based biomedical pretrained language models”, *Journal of biomedical informatics*, vol. 2, 09 2021, pp. 103982.

[67] Kalyan, K. S.; Sangeetha, S. “Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network”, *Artificial Intelligence in Medicine*, vol. 112, 02 2021, pp. 102008.

[68] Karczewski, K. J.; Daneshjou, R.; Altman, R. B. “Pharmacogenomics”, *PLoS computational biology*, vol. 8–12, 12 2012, pp. 19.

[69] Khattak, F. K.; Jeblee, S.; Pou-Prom, C.; Abdalla, M.; Meaney, C.; Rudzicz, F. “A survey of word embeddings for clinical text”, *Journal of Biomedical Informatics*, vol. 100, 12 2019, pp. 100057.

[70] Kingma, D. P.; Ba, J. “Adam: A method for stochastic optimization”, *arXiv preprint*, vol. 1, 12 2014, pp. 15.

[71] Kocaman, V.; Talby, D. “Biomedical named entity recognition at scale”. In: Proceedings of the 25th International Conference on Pattern Recognition Workshops (ICPR 2020), 2021, pp. 635–646.

[72] Köhler, S.; Doelken, S. C.; Mungall, C. J.; Bauer, S.; Firth, H. V.; Bailleul-Forestier, I.; Black, G. C. M.; Brown, D. L.; Brudno, M.; Campbell, J.; FitzPatrick, D. R.; Eppig, J. T.; Jackson, A. P.; Freson, K.; Girdea, M.; Helbig, I.; Hurst, J. A.; Jähn, J.; Jackson,

L. G.; Kelly, A. M.; Ledbetter, D. H.; Mansour, S.; Martin, C. L.; Moss, C.; Mumford, A.; Ouwehand, W. H.; Park, S.-M.; Riggs, E. R.; Scott, R. H.; Sisodiya, S.; Vooren, S. V.; Wapner, R. J.; Wilkie, A. O. M.; Wright, C. F.; Vulto-van Silfhout, A. T.; de Leeuw, N.; de Vries, B. B. A.; Washingthon, N. L.; Smith, C. L.; Westerfield, M.; Schofield, P.; Ruef, B. J.; Gkoutos, G. V.; Haendel, M.; Smedley, D.; Lewis, S. E.; Robinson, P. N. "The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data", *Nucleic Acids Research*, vol. 42–D1, 11 2013, pp. D966–D974.

[73] Konkol, M.; Konopík, M. "Segment representations in named entity recognition". In: Proceedings of the 18th International Conference on Text, Speech, and Dialogue - Volume 9302, 2015, pp. 61–70.

[74] Krallinger, M.; Leitner, F.; Rabal, O.; Vazquez, M.; Oyarzabal, J.; Valencia, A. "Chemdner: The drugs and chemical names extraction challenge", *Journal of cheminformatics*, vol. 7–1, 01 2015, pp. 1–11.

[75] Krauthammer, M.; Nenadic, G. "Term identification in the biomedical literature", *Journal of biomedical informatics*, vol. 37–6, 12 2004, pp. 512–526.

[76] Krenker, A.; Bešter, J.; Kos, A. "Introduction to the artificial neural networks", *Artificial Neural Networks: Methodological Advances and Biomedical Applications. InTech*, vol. 1, 03 2011, pp. 1–18.

[77] Kukich, K. "Techniques for automatically correcting words in text", *Acm Computing Surveys (CSUR)*, vol. 24–4, 12 1992, pp. 377–439.

[78] Kuperman, G. J.; Reichley, R. M.; Bailey, T. C. "Using commercial knowledge bases for clinical decision support: opportunities, hurdles, and recommendations", *Journal of the American Medical Informatics Association*, vol. 13–4, 07 2006, pp. 369–371.

[79] Kuru, O.; Can, O. A.; Yuret, D. "CharNER: Character-level named entity recognition". In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 911–921.

[80] Lafferty, J. D.; McCallum, A.; Pereira, F. C. N. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 282–289.

[81] Lago Martínez, P. A.; et al.. "Modeling and learning context enriched behavior patterns in ambient assisted living", *Repositorio Institucional Séneca*, vol. 1, 06 2017, pp. 279.

[82] Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. "Neural architectures for named entity recognition", *arXiv preprint*, vol. 3, 04 2016, pp. 11.

[83] Lample, G.; Conneau, A. “Cross-lingual language model pretraining”, *arXiv preprint*, vol. 1, 01 2019, pp. 10.

[84] Lamurias, A.; Couto, F. “Text mining for bioinformatics using biomedical literature”, vol. 1–76, 01 2019, pp. 602–611.

[85] Lavanya, P.; Sasikala, E. “Deep learning techniques on text classification using natural language processing (nlp) in social healthcare network: A comprehensive survey”. In: Proceedings of the 3rd International Conference on Signal Processing and Communication (ICPSC), 2021, pp. 603–609.

[86] Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. “Biobert: a pre-trained biomedical language representation model for biomedical text mining”, *Bioinformatics*, vol. 36–4, 10 2020, pp. 1234–1240.

[87] Li, J.; Sun, Y.; Johnson, R.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; Lu, Z. “Annotating chemicals, diseases, and their interactions in biomedical literature”. In: Proceedings of the fifth BioCreative challenge evaluation workshop, 2015, pp. 173–182.

[88] Li, L.; Guo, Y. “Biomedical named entity recognition with cnn-blstm-crf”, *Journal of chinese information processing*, vol. 32–1, 06 2018, pp. 116–122.

[89] Lin, J. C.-W.; Shao, Y.; Djenouri, Y.; Yun, U. “Asrnn: a recurrent neural network with an attention model for sequence labeling”, *Knowledge-Based Systems*, vol. 212, 12 2020, pp. 106548.

[90] Liu, S.; Tang, B.; Chen, Q.; Wang, X. “Drug name recognition: approaches and resources”, *Information*, vol. 6–4, 11 2015, pp. 790–810.

[91] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. “Roberta: A robustly optimized bert pretraining approach”, *arXiv preprint*, vol. 1, 07 2019, pp. 13.

[92] Lobo, T. C.; Bettoni, G. N.; da Silva, F. S.; Caregnato, R. C.; Flores, C. D. “Enabling communication among ehr systems with microsservices and hl7 fhir”. In: Actas de SABI, 2020, pp. 247.

[93] López-Úbeda, P.; Díaz-Galiano, M. C.; Martín-Noguerol, T.; Luna, A.; Ureña-López, L. A.; Martín-Valdivia, M. T. “Automatic medical protocol classification using machine learning approaches”, *Computer Methods and Programs in Biomedicine*, vol. 200, 01 2021, pp. 105939.

[94] López-Úbedaa, P.; Perea-Ortegab, J. M.; Díaz-Galianoa, M. C.; Martín-Valdiviaa, M. T.; Ure

textasciitilde na-Lópeza, L. A. “Sinai at ehealth-kd challenge 2020: Combining word embeddings for named entity recognition in spanish medical records”, *CEUR Workshop Proceedings*, vol. 1, 09 2020, pp. 10.

[95] Loshchilov, I.; Hutter, F. “Decoupled weight decay regularization”, *arXiv preprint*, vol. 1, 01 2019, pp. 19.

[96] Lowe, H. J.; Barnett, G. O. “Micromesh: a microcomputer system for searching and exploring the national library of medicine’s medical subject headings (mesh) vocabulary”. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, 1987, pp. 717.

[97] Magnolini, S.; Piccioni, V.; Balaraman, V.; Guerini, M.; Magnini, B. “How to use gazetteers for entity recognition with neural models”. In: Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5), 2019, pp. 40–49.

[98] Manolio, T. A.; Chisholm, R. L.; Ozenberger, B.; Roden, D. M.; Williams, M. S.; Wilson, R.; Bick, D.; Bottinger, E. P.; Brilliant, M. H.; Eng, C.; et al.. “Implementing genomic medicine in the clinic: the future is here”, *Genetics in Medicine*, vol. 15–4, 03 2013, pp. 258–267.

[99] Maria Cristiane Barbosa Galvão, I. L. M. R. “A snomed ct e os sistemas de informação em saúde.” Capturado em: http://www.ofaj.com.br/colunas_conteudo.php?cod=757, 15 jan 2021.

[100] McCallum, A.; Freitag, D.; Pereira, F. C. “Maximum entropy markov models for information extraction and segmentation.” In: Icml, 2000, pp. 591–598.

[101] McCarthy, J. “What is artificial intelligence? cogprints”. Capturado em: https://web-archive.southampton.ac.uk/cogprints.org/412/2/whatisai.ps, 07 out 2020.

[102] McDonald, C. J.; Huff, S. M.; Suico, J. G.; Hill, G.; Leavelle, D.; Aller, R.; Forrey, A.; Mercer, K.; DeMoor, G.; Hook, J.; et al.. “Loinc, a universal standard for identifying laboratory observations: a 5-year update”, *Clinical chemistry*, vol. 49–4, 04 2003, pp. 624–633.

[103] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. “Distributed representations of words and phrases and their compositionality”. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, 2013, pp. 3111–3119.

[104] Miller, R. A.; Gieszczykiewicz, F. M.; Vries, J. K.; Cooper, G. F. “Chartline: providing bibliographic references relevant to patient charts using the umls metathesaurus knowledge sources.” In: Proceedings of the Annual Symposium on Computer Application in Medical Care, 1992, pp. 86.

[105] Miranda, M.; Duarte, J.; Abelha, A.; Machado, J. M.; Neves, J. “Interoperability and healthcare”, *Eurosis*, vol. 1, 10 2009, pp. 12.

[106] Miranda-Escalada, A.; Farré, E.; Krallinger, M. “Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results.” In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), 2020, pp. 303–323.

[107] Nakayama, H. “seqeval: A python framework for sequence labeling evaluation”. Software available from https://github.com/chakki-works/seqeval, Capturado em: https://github.com/chakki-works/seqeval, 21 jan 2022.

[108] Nicholson, D. N.; Himmelstein, D. S.; Greene, C. S. “Reusing label functions to extract multiple types of biomedical relationships from biomedical abstracts at scale”, *bioRxiv*, vol. 1, 08 2019, pp. 730085.

[109] Nunes, R. O.; Soares, J. E.; dos Santos, H. D. P.; Vieira, R. “Meshx-notes: Web-system for clinical notes”. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (Artificial Intelligence in Health), 2019, pp. 5–12.

[110] Oemig, F.; Blobel, B. “Natural language processing supporting interoperability in healthcare”. In: *Text mining*, Springer, 2014, pp. 137–156.

[111] Oliveira, L.; Peters, A.; Silva, A.; Gebeluca, C.; Gumiel, Y.; Cintho, L.; Carvalho, D.; Hasan, S.; Moro, C. “Semclinbr – a multi institutional and multi specialty semantically annotated corpus for portuguese clinical nlp tasks”, *Journal of Biomedical Semantics*, vol. 1, 01 2020, pp. 19.

[112] Ozturk, T.; Talo, M.; Yildirim, E. A.; Baloglu, U. B.; Yildirim, O.; Acharya, U. R. “Automated detection of covid-19 cases using deep neural networks with x-ray images”, *Computers in biology and medicine*, vol. 121, 04 2020, pp. 103792.

[113] Pan, S. J.; Yang, Q. “A survey on transfer learning”, *IEEE Transactions on knowledge and data engineering*, vol. 22–10, 10 2009, pp. 1345–1359.

[114] Peiffer-Smadja, N.; Rawson, T. M.; Ahmad, R.; Buchard, A.; Georgiou, P.; Lescure, F.-X.; Birgand, G.; Holmes, A. H. “Machine learning for clinical decision support in infectious diseases: a narrative review of current applications”, *Clinical Microbiology and Infection*, vol. 26–5, 05 2020, pp. 584–595.

[115] Peng, J.; Zhao, M.; Havrilla, J.; Liu, C.; Weng, C.; Guthrie, W.; Schultz, R.; Wang, K.; Zhou, Y. “Natural language processing (NLP) tools in extracting biomedical concepts

from research articles: a case study on autism spectrum disorder", *BMC Med. Inform. Decis. Mak.*, vol. 20–Suppl 11, December 2020, pp. 322.

[116] Pennington, J.; Socher, R.; Manning, C. D. "Glove: Global vectors for word representation". In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[117] Pérez-Díez, I.; Pérez-Moraga, R.; López-Cerdán, A.; Salinas-Serrano, J.-M.; de la Iglesia-Vayá, M. "De-identifying spanish medical texts-named entity recognition applied to radiology reports", *Journal of Biomedical Semantics*, vol. 12–1, 03 2021, pp. 1–13.

[118] Qin, W.; Du, Z.; Xiao, J.; Duan, H.; Shu, Q.; Li, H. "Evaluation of clinical impact of pharmacogenomics knowledge involved in CPIC guidelines on Chinese pediatric patients", *Pharmacogenomics*, vol. 21–3, 02 2020, pp. 209–219.

[119] Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. "Pre-trained models for natural language processing: A survey", *Science China Technological Sciences*, vol. 63–10, 04 2020, pp. 1872–1897.

[120] Ramos, E. M.; Din-Lovinescu, C.; Berg, J. S.; Brooks, L. D.; Duncanson, A.; Dunn, M.; Good, P.; Hubbard, T. J.; Jarvik, G. P.; O'Donnell, C.; Sherry, S. T.; Aronson, N.; Biesecker, L. G.; Blumberg, B.; Calonge, N.; Colhoun, H. M.; Epstein, R. S.; Flicek, P.; Gordon, E. S.; Green, E. D.; Green, R. C.; Hurles, M.; Kawamoto, K.; Knaus, W.; Ledbetter, D. H.; Levy, H. P.; Lyon, E.; Maglott, D.; McLeod, H. L.; Rahman, N.; Randhawa, G.; Wicklund, C.; Manolio, T. A.; Chisholm, R. L.; Williams, M. S. "Characterizing genetic variants for clinical action", *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 166–1, 2014, pp. 93–104.

[121] Ramshaw, L.; Marcus, M. "Text chunking using transformation-based learning". In: Proceedings of the Third Workshop on Very Large Corpora, 1995, pp. 13.

[122] Reátegui, R.; Ratté, S. "Comparison of MetaMap and cTAKES for entity extraction in clinical notes", *BMC Med. Inform. Decis. Mak.*, vol. 18–Suppl 3, September 2018, pp. 74.

[123] Rector, A.; Schulz, S.; Rodrigues, J. M.; Chute, C. G.; Solbrig, H. "On beyond gruber:"ontologies" in today's biomedical information systems and the limits of owl", *Journal of Biomedical Informatics*, vol. 100, 03 2019, pp. 100002.

[124] Reed, R.; MarksII, R. J. "Neural smithing: supervised learning in feedforward artificial neural networks". Mit Press, 1999, 360p.

[125] Ribeiro, J. K.; dos Santos, H. D.; Barletta, F.; da Silva, M. C.; Vieira, R.; Morales, H. M.; da Costa Rocha, C. “A machine learning early warning system: Multicenter validation in brazilian hospitals”. In: Proceedings of the 33rd IEEE Symposium on Computer-Based Medical Systems (CBMS), 2020, pp. 321–326.

[126] Rindflesch, T. C.; Tanabe, L.; Weinstein, J. N.; Hunter, L. “EDGAR: extraction of drugs, genes and relations from the biomedical literature”, *Pac Symp Biocomput*, vol. 1, 07 2000, pp. 517–528.

[127] Robinson, P. N.; Mundlos, S. “The Human Phenotype Ontology”, *Clinical Genetics*, vol. 77–6, 01 2010, pp. 525–534.

[128] Rodríguez-González, A.; Costumero, R.; Martinez-Romero, M.; Wilkinson, M. D.; Menasalvas-Ruiz, E. “Extracting diagnostic knowledge from medline plus: a comparison between metamap and ctakes approaches”, *Current Bioinformatics*, vol. 13–6, 06 2018, pp. 573–582.

[129] Ruder, S. “An overview of gradient descent optimization algorithms”, *arXiv preprint*, vol. 1, 09 2016, pp. 14.

[130] Safarova, M. S.; Kullo, I. J. “Using the electronic health record for genomics research”, *Current opinion in lipidology*, vol. 31–2, 04 2020, pp. 85–93.

[131] Salton, G.; Buckley, C. “Term-weighting approaches in automatic text retrieval”, *Information processing & management*, vol. 24–5, 11 1988, pp. 513–523.

[132] Sang, E. F.; Buchholz, S. “Introduction to the conll-2000 shared task: Chunking”, *arXiv preprint cs/0009008*, vol. 1, 09 2000, pp. 06.

[133] Sang, E. F.; De Meulder, F. “Introduction to the conll-2003 shared task: Language-independent named entity recognition”, *arXiv preprint*, vol. 1, 06 2003, pp. 06.

[134] Santos, B. G. T. d.; Bettoni, G. N.; Silva, F. S. d. “Uma ferramenta para aplicação de mapeamentos entre termos snomed ct, cid-10 e ciap-2 e enriquecimento terminológico em segundas-opiniões formativas sobre hipertensão e diabetes”, *Resdite*, vol. 5, 04 2020, pp. 13.

[135] Santos, J.; dos Santos, H. D.; Vieira, R. “Fall detection in clinical notes using language models and token classifier”. In: Proceedings of the 33rd IEEE Symposium on Computer-Based Medical Systems (CBMS), 2020, pp. 283–288.

[136] Sarzynska-Wawer, J.; Wawer, A.; Pawlak, A.; Szymanowska, J.; Stefaniak, I.; Jarkiewicz, M.; Okruszek, L. “Detecting formal thought disorder by deep contextualized word representations”, *Psychiatry Research*, vol. 304, 10 2021, pp. 114135.

[137] Savova, G. K.; Fan, J.; Ye, Z.; Murphy, S. P.; Zheng, J.; Chute, C. G.; Kullo, I. J. “Discovering peripheral arterial disease cases from radiology notes using natural language processing”, *AMIA Annu. Symp. Proc.*, vol. 2010, November 2010, pp. 722–726.

[138] Schmidhuber, J. “Deep learning in neural networks: An overview”, *Neural networks*, vol. 61, 10 2015, pp. 85–117.

[139] Schneider, E. T. R.; de Souza, J. V. A.; Knafou, J.; Oliveira, L. E. S. e.; Copara, J.; Gumiel, Y. B.; Oliveira, L. F. A. d.; Paraiso, E. C.; Teodoro, D.; Barra, C. M. C. M. “BioBERTpt - a Portuguese neural language model for clinical named entity recognition”. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, 2020, pp. 65–72.

[140] Segura-Bedmar, I.; Martínez Fernández, P.; Herrero Zazo, M. “Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)”. In: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval), Second Joint Conference on Lexical and Computational Semantics (*SEM), 2013, pp. 10.

[141] Senthilkumar, S.; Rai, B. K.; Meshram, A. A.; Gunasekaran, A.; Chandrakumarmangalam, S. “Big data in healthcare management: a review of literature”, *American Journal of Theoretical and Applied Business*, vol. 4–2, 07 2018, pp. 57–69.

[142] Serrà, J.; Karatzoglou, A. “Getting deep recommenders fit: Bloom embeddings for sparse binary input/output networks”. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, 2017, pp. 279–287.

[143] Shameer, K.; Badgeley, M. A.; Miotto, R.; Glicksberg, B. S.; Morgan, J. W.; Dudley, J. T. “Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams”, *Briefings in Bioinformatics*, vol. 18–1, 02 2016, pp. 105–124, https://academic.oup.com/bib/article-pdf/18/1/105/25408390/bbv118.pdf.

[144] Shao, Y.; Taylor, S.; Marshall, N.; Morioka, C.; Zeng-Treitler, Q. “Clinical text classification with word embedding features vs. bag-of-words features”. In: Proceedings of the IEEE International Conference on Big Data (Big Data), 2018, pp. 2874–2878.

[145] Shortliffe, E. H.; Cimino, J. J. “Biomedical informatics: computer applications in health care and biomedicine”. Springer, 2021, 965p.

[146] Sissung, T. M.; McKeeby, J. W.; Patel, J.; Lertora, J. J.; Kumar, P.; Flegel, W. A.; Adams, S. D.; Eckes, E. J.; Mickey, F.; Plona, T. M.; et al.. “Pharmacogenomics

implementation at the national institutes of health clinical center", *The Journal of Clinical Pharmacology*, vol. 57, 09 2017, pp. S67–S77.

[147] Smith, L.; Tanabe, L. K.; Kuo, C.-J.; Chung, I.; Hsu, C.-N.; Lin, Y.-S.; Klinger, R.; Friedrich, C. M.; Ganchev, K.; Torii, M.; et al.. "Overview of biocreative ii gene mention recognition", *Genome biology*, vol. 9–2, 09 2008, pp. 1–19.

[148] Smola, A. J.; Bartlett, P. J.; Schuurmans, D.; Schölkopf, B.; et al.. "Advances in large margin classifiers". MIT press, 2000, 422p.

[149] Song, C. H.; Lawrie, D.; Finin, T.; Mayfield, J. "Improving neural named entity recognition with gazetteers", *arXiv preprint*, vol. 1, 03 2020, pp. 08.

[150] Song, H.-J.; Jo, B.-C.; Park, C.-Y.; Kim, J.-D.; Kim, Y.-S. "Comparison of named entity recognition methodologies in biomedical documents", *Biomedical engineering online*, vol. 17–2, 11 2018, pp. 1–14.

[151] Souza, A.; de Medeiros, A. P.; Martins, C. B. "Technical interoperability among ehr systems in brazilian public health organizations", *Rev Brasil Comput Aplicada*, vol. 11, 07 2019, pp. 42–55.

[152] Souza, F.; Nogueira, R.; Lotufo, R. "BERTimbau: pretrained BERT models for Brazilian Portuguese". In: Proceedings of the Intelligent Systems: 9th Brazilian Conference (BRACIS 2020), 2020, pp. 14.

[153] Starren, J.; Williams, M. S.; Bottinger, E. P. "Crossing the omic chasm: a time for omic ancillary systems", *Jama*, vol. 309–12, 03 2013, pp. 1237–1238.

[154] Stevens, R.; Goble, C. A.; Bechhofer, S. "Ontology-based knowledge representation for bioinformatics", *Briefings in bioinformatics*, vol. 1–4, 11 2000, pp. 398–414.

[155] Sulik, G. A. "Managing biomedical uncertainty: the technoscientific illness identity", *Sociology of Health & Illness*, vol. 31–7, 11 2009, pp. 1059–1076.

[156] Takeuchi, K.; Collier, N. "Bio-medical entity extraction using support vector machines", *Artificial Intelligence in Medicine*, vol. 33–2, 01 2005, pp. 125–137.

[157] Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Van Durme, B.; Bowman, S. R.; Das, D.; et al.. "What do you learn from context? probing for sentence structure in contextualized word representations", *arXiv preprint*, vol. 1, 05 2019, pp. 17.

[158] Úbeda, P. L.; Díaz-Galiano, M. C.; López, L. A. U.; Martín-Valdivia, M. T. "Using snomed to recognize and index chemical and drug mentions." In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 115–120.

[159] Uschold, M.; King, M.; Moralee, S.; Zorgios, Y. “The enterprise ontology”, *The knowledge engineering review*, vol. 13–1, 08 1998, pp. 31–89.

[160] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. “Attention is all you need”, *arXiv preprint*, vol. 5, 12 2017, pp. 15.

[161] Viani, N.; Miller, T. A.; Napolitano, C.; Priori, S. G.; Savova, G. K.; Bellazzi, R.; Sacchi, L. “Supervised methods to extract clinical events from cardiology reports in italian”, *Journal of biomedical informatics*, vol. 95, 06 2019, pp. 103219.

[162] Vincze, V.; Szarvas, G.; Farkas, R.; Móra, G.; Csirik, J. “The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes”, *BMC bioinformatics*, vol. 9–11, 11 2008, pp. 1–9.

[163] Vretinaris, A.; Lei, C.; Efthymiou, V.; Qin, X.; Özcan, F. “Medical entity disambiguation using graph neural networks”. In: Proceedings of the International Conference on Management of Data, 2021, pp. 2310–2318.

[164] Wagner Filho, J. A.; Wilkens, R.; Idiart, M.; Villavicencio, A. “The brwac corpus: A new open resource for brazilian portuguese”. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC), 2018, pp. 06.

[165] Wallace, B. C. “Automating biomedical evidence synthesis: Recent work and directions forward”. In: Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), 2018, pp. 4.

[166] Wallach, H. M. “Conditional random fields: An introduction”, *Technical Reports (CIS)*, vol. 1, 02 2004, pp. 22.

[167] Wang, Z.; Qu, Y.; Chen, L.; Shen, J.; Zhang, W.; Zhang, S.; Gao, Y.; Gu, G.; Chen, K.; Yu, Y. “Label-aware double transfer learning for cross-specialty medical named entity recognition”, *arXiv preprint*, vol. 2, 04 2018, pp. 15.

[168] Warrer, P.; Hansen, E. H.; Juhl-Jensen, L.; Aagaard, L. “Using text-mining techniques in electronic patient records to identify adrs from medicine use”, *British journal of clinical pharmacology*, vol. 73–5, 05 2012, pp. 674–684.

[169] Weissenbacher, D.; Sarker, A.; Klein, A.; O’Connor, K.; Magge, A.; Gonzalez-Hernandez, G. “Deep neural networks ensemble for detecting medication mentions in tweets”, *Journal of the American Medical Informatics Association*, vol. 26–12, 12 2019, pp. 1618–1626.

[170] Weng, L.; Zhang, L.; Peng, Y.; Huang, R. S. “Pharmacogenetics and pharmacogenomics: a bridge to individualized cancer therapy”, *Pharmacogenomics*, vol. 14–3, 02 2013, pp. 315–324.

[171] Wu, A. S.; Do, B. H.; Kim, J.; Rubin, D. L. “Evaluation of negation and uncertainty detection and its impact on precision and recall in search”, *Journal of digital imaging*, vol. 24–2, 04 2011, pp. 234–242.

[172] Xia, Y.; Zhong, X.; Liu, P.; Tan, C.; Na, S.; Hu, Q.; Huang, Y. “Combining metamap and ctakes in disorder recognition: THCIB at CLEF ehealth lab 2013 task 1”. In: Working Notes for CLEF Conference , Valencia, Spain, Forner, P.; Navigli, R.; Tufis, D.; Ferro, N. (Editores), 2013, pp. 5.

[173] Xie, X.-Y. “A review on support vector machines for biomedical ner”, *Data Science for NLP*, vol. 1, 06 2020, pp. 7.

[174] Xu, Y.; Tsujii, J.; Chang, E. I.-C. “Named entity recognition of follow-up and time information in 20 000 radiology reports”, *Journal of the American Medical Informatics Association*, vol. 19–5, 07 2012, pp. 792–799.

[175] Yan, C.; Zhang, Y.; Liu, K.; Zhao, J.; Shi, Y.; Liu, S. “Enhancing unsupervised medical entity linking with multi-instance learning”, *BMC medical informatics and decision making*, vol. 21–9, 11 2021, pp. 1–10.

[176] Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; Le, Q. V. “Xlnet: Generalized autoregressive pretraining for language understanding”, *Advances in neural information processing systems*, vol. 32, 06 2019, pp. 18.

[177] Zeiler, M. D. “Adadelta: an adaptive learning rate method”, *arXiv preprint*, vol. 1, 12 2012, pp. 6.

[178] Zhang, S.; Elhadad, N. “Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts”, *Journal of biomedical informatics*, vol. 46–6, 02 2013, pp. 1088–1098.

APÊNDICE A – DIRETRIZES DE ANOTAÇÕES PARA TEXTO DE SAÚDE

A seguir, estão disponibilizados integralmente as diretrizes derivados da implementação deste trabalho.

Categoria: Fármaco

Definição Uma droga ou outra preparação para o tratamento ou prevenção de doenças. Quaisquer produtos sanguíneos usados em transfusão também devem ser considerados na categoria “Fármaco”. Medicamentos genéricos como medicamentos para hipertensão e medicamentos para colesterol também devem ser anotados. Classes de medicamentos como antibióticos, anti-histamínicos, antidepressivos e assim por diante devem ser consideradas Fármacos.

Exemplos e regras de anotação

1. “A paciente havia consultado seus cuidados primários há alguns dias, foi retirada de seu lisinopril e digoxina.” a. “lisinopril” e “digoxina” são medicamentos.
2. Outros exemplos de medicamentos são “Omeprazol”, “Hidralazina”, “Prilosec” e assim por diante.
3. Considere “FFP”, “RBC Embalado” e outros hemoderivados transfusionais como medicação. “O2” administrado para tratamento também deve ser considerado como medicação. Produtos sanguíneos como “plasma fresco congelado” e “glóbulos vermelhos embalados” devem ser considerados sob medicação.
4. Via dos medicamentos não deve ser anotada em conjunto. a. Por exemplo: No caso de “Aspirina IV”, apenas aspirina será anotada como Fármaco.
5. Medicamentos genéricos como medicamentos para hipertensão e medicamentos para colesterol também devem ser anotados em Medicação.
6. Classes de medicamentos como antibióticos, anti-histamínicos, antidepressivos e assim por diante também devem ser consideradas sob Fármaco.
7. Medicamentos combinados devem ser anotados como uma única entidade. a. Por exemplo: em “Fluticasona/vilanterol” é uma droga combinada e seu nome comercial é Breo Ellipta. Anote “Fluticasona/vilanterol” como uma única entidade.
8. Medicamentos listados em seções de alergia. a. Por exemplo, “Alergias: Penicilinas / Sulfonamidas”. Neste caso, a medicação não é solicitada para o paciente nem faz

parte da lista de medicamentos contínuos do paciente. Apenas é mencionado que o paciente tem alergia a esses medicamentos. Devemos anotar: * medicação: "penicilinas", assunto: "paciente", status: "outro - alergia"* medicação: "Sulfonamidas", assunto: "paciente", status: "outro - alergia"

9. No exemplo: “Ele continuará a tomar aspirina e é aconselhado a interromper seus antibióticos”, devemos anotar: a. medicação: “aspirina”, status: “atual”, avaliação de certeza: “provável” b. medicação: “antibióticos”, avaliação temporal: “próxima”, avaliação de certeza: “um pouco improvável”

Modificador: Força

Definição A dosagem da droga indica a quantidade de ingrediente ativo em cada dosagem.

Exemplos e regras de anotação

1. "O paciente toma Aspirina 50 mg 2 comprimidos por via oral todos os dias". a. "50"é a dosagem do medicamento "Aspirina".

Modificador: Unidades

Definição

As forças geralmente são quantificadas em unidades de medidas.

Exemplos e regras de anotação

1. "O paciente toma Aspirina 50 mg 2 comprimidos por via oral todos os dias". a. "mg"é a unidade de medida da dosagem da medicação.

Modificador: Dose

Definição

Uma dose de um medicamento ou medicamento é uma quantidade medida dele que se destina a ser tomada de uma só vez. É a quantidade do medicamento prescrito para ser tomado de uma só vez.

Exemplos e regras de anotação

1. "O paciente toma Aspirina 50 mg 2 comprimidos por via oral todos os dias"a. "2"é a dose do medicamento.

Modificador: Formulário

Definição

A forma de medicação indica características físicas do medicamento específico.

Exemplos e regras de anotação

1. "O paciente toma Aspirina 50 mg 2 comprimidos por via oral todos os dias". a. "comprimidos"é a forma do medicamento.

Modificador: Formulário

Definição

A frequência de um medicamento refere-se à frequência com que é tomado.

Exemplos e regras de anotação

1. "O paciente toma Aspirina 50 mg 2 comprimidos por via oral todos os dias". a. "todos os dias"é a frequência da administração da medicação.

Modificador: Rota

Definição

Uma via de administração é a via pela qual a droga é levada para o corpo. A via geralmente é classificada pelo local em que o medicamento é administrado.

Exemplos e regras de anotação

1. "O paciente toma Aspirina 50 mg 2 comprimidos por via oral todos os dias". a. "por via oral"é a via do medicamento.

Modificador: Duração

Definição

O tempo até o qual o paciente deve tomar a medicação é conhecido como duração do medicamento.

Exemplos e regras de anotação

1. "O paciente deve tomar Aspirina 50 mg 2 comprimidos por via oral todos os dias durante 2 meses". a. "por 2 meses"é a duração durante a qual o paciente deve tomar a medicação.

Modificador: Status

Definição

O medicamento tem status diferente, como “aumento”, “diminuição”, “início”, “parada”, “continuação”, “descontinuação” e assim por diante, e tais informações devem ser capturadas sob o rótulo Status.

Exemplos e regras de anotação

1. “A aspirina foi interrompida.” a. “interrompida” representa o status do medicamento.

Regras adicionais para anotar atributos de fármacos

1. O limite do relacionamento é de uma frase. Não vincule entidades que estão relacionadas além de uma frase.
2. Se houver várias menções a uma droga em uma frase, todas as menções devem estar relacionadas aos seus atributos. * Por exemplo, “abrandador de fezes (colace) 2mg”. Neste caso, “2” como força e “mg” como unidade devem estar relacionados tanto ao “abrandador de fezes” quanto ao “colace”.
3. Os atributos não são independentes e devem estar relacionados à medicação.
4. Se o status de um medicamento for “alterado”, “trocado”, “parado”, “aumentar”, “diminuir”, ele deve ser anotado sob o atributo Status.

5. Se o status de um medicamento for “alterado”, “trocado”, “parado”, “aumentar”, “diminuir”, ele deve ser anotado sob o atributo Status.

6. Se o status de um medicamento for “alterado”, “trocado”, “parado”, “aumentar”, “diminuir”, ele deve ser anotado sob o atributo Status.

7. Para duração, anote a preposição para em “por 2 semanas” juntos, em vez de anotar apenas “2 semanas”.

8. Se a hora estiver escrita como “na hora de dormir”, anote também a preposição.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br