Programa de Pós-Graduação em Ciência da Computação

Reconhecimento de Entidades Nomeadas em Prontuários de Psiquiatria: Anotação de Corpus e Extração de Entidades Clínicas

LUIZ HENRIQUE PEREIRA NIERO

Rio Claro - SP
2023

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

LUIZ HENRIQUE PEREIRA NIERO

RECONHECIMENTO DE ENTIDADES NOMEADAS EM PRONTUÁRIOS DE PSIQUIATRIA: ANOTAÇÃO DE CORPUS E EXTRAÇÃO DE ENTIDADES CLÍNICAS

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Ivan Rizzo Guilherme

Coorientador: Prof. Dr. Lucas Emanuel Silva e Oliveira

Rio Claro - SP
2023

N676r

Niero, Luiz Henrique Pereira
Reconhecimento de entidades nomeadas em prontuários de psiquiatria : anotação de corpus e extração de entidades clínicas / Luiz Henrique Pereira Niero. -- Rio Claro, 2023
85 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Ivan Rizzo Guilherme
Coorientador: Lucas Emanuel Silva e Oliveira

1. Ciência da computação. 2. PLN. 3. Extração de informação. 4. Aprendizado de máquina. 5. Anotaçãodecorpus. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Geociências e Ciências Exatas, Rio Claro. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

LUIZ HENRIQUE PEREIRA NIERO

RECONHECIMENTO DE ENTIDADES NOMEADAS EM PRONTUÁRIOS DE PSIQUIATRIA: ANOTAÇÃO DE CORPUS E EXTRAÇÃO DE ENTIDADES CLÍNICAS

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Comissão Examinadora:

Prof. Dr. Ivan Rizzo Guilherme
IGCE/UNESP/Rio Claro (SP)

Prof. Dr. Arnaldo Cândido Junior
IBILCE/UNESP/São José Rio Preto (SP)

Prof. Dra. Claudia Maria Cabral Moro Barra
PUCPR/Curitiba (PR)

Conceito: Aprovado.

Rio Claro (SP), 30 de junho de 2023.

Dedicatória

Dedico este trabalho com profundo amor e gratidão:

À Deus, por me guiar através de todas as adversidades e me abençoar com força e orientação.

Aos meus filhos Sophia e Óliver, que são minha fonte constante de inspiração e alegria.

À minha querida esposa, Ingrid, cujo apoio inabalável torna cada passo desta jornada significativo.

Aos meus pais e irmã querida, cujo amor e encorajamento moldaram quem sou hoje.

À universidade pública, que proporcionou o ambiente propício para o desenvolvimento do conhecimento e da aprendizagem e que é, hoje, instituição de igualdade e justiça social no Brasil.

Que este trabalho seja um reflexo do amor e da dedicação presentes em cada momento compartilhado com aqueles que tornam a minha vida extraordinária.

Agradecimentos

Meus sinceros agradecimentos:

À Deus, por abrir as portas necessárias para ter chego até aqui.

Ao professor doutor Ivan, que além de dedicado orientador fez-se um grande amigo ao longo destes anos de aprendizado.

Ao professor doutor Lucas, que ofereceu valiosa coorientação nesta nobre área, tão pouco explorada no Brasil.

Ao professor doutor Gerardo, Leisa, Giovanna, Bianca, Loise e Gabriel, pela cooperação e horas de trabalho voluntário cedidas à elaboração e anotação do Corpus aqui apresentado.

Ao mestre Pedro Ivo, pelas orientações e companheirismo ao longo do desenvolvimento deste trabalho.

À CAPES e ao Hospital Dr. Adolfo Bezerra de Menezes, por me proporcionar as condições necessárias para apresentar o conteúdo deste trabalho em um congresso internacional.

Aos amigos, companheiros de jornada, e todos que contribuíram de alguma forma para a realização deste trabalho, direta e indiretamente, porém a memória não me agraciou com a lembrança enquanto escrevia esta nota agradecimento.

Epígrafe

Os lábios da sabedoria estão fechados,

Exceto aos ouvidos do entendimento.

O Caibalion.

Resumo

Grande parte das informações dos dados de pacientes presentes nos sistemas de prontuários eletrônicos encontram-se em formato de texto. Esses dados dizem respeito às informações diversas do paciente e são de extrema importância para acompanhar o tratamento dele. Porém, dados que estão no formato puramente textual não estão passíveis à aplicação de algoritmos computacionais para a extração de informações. Para resolver tal problema, existe uma área chamada Processamento de Linguagem Natural e uma subárea chamada de Reconhecimento de Entidade Nomeadas (REN), que tem por objetivo reconhecer unidades que representam algum sentido no texto – chamadas de Entidades. Os métodos mais atuais de REN utilizam-se de métodos de aprendizado de máquina em corpus de documentos anotados para atingir bons resultados, porém, de acordo com o levantamento realizado no presente trabalho, não foram encontrados corpus anotados para a língua portuguesa que pudessem ser treinados para o domínio da psiquiatria, que é o domínio explorado no presente trabalho. Portanto, neste trabalho foi realizada a anotação de um corpus de prontuários médicos de psiquiatria, o PSYclinBR. O corpus é composto por 300 prontuários de notas de admissão de um serviço de emergência psiquiatra e anotado em 9 categorias semânticas: comportamento autodestrutivo, diagnóstico, droga, fármaco, função psíquica, histórico familiar, histórico do paciente, observação e sintoma e queixa psíquica. Ao todo, são 2.480 sentenças e 5.822 entidades, com IAA de 0.6. O treinamento do modelo de REN baseado na arquitetura BERT, o psyBERTpt, foi feito com a utilização do corpus anotado e atingiu os valores médios de precisão 0,65, revocação 0,69 e f1-score 0,67, valores que estão próximos ao estado da arte para REN em dados clínicos da língua portuguesa. Tanto o Corpus como o modelo de REN foram disponibilizados publicamente.

Palavras-chave: Reconhecimento de Entidade Nomeada. Reconhecimento de Entidade Clínica. Anotação de corpus. Extração de Entidades Clínicas. Processamento de Linguagem Natural. BERT

Abstract

A significant portion of patient data within electronic medical record systems is represented in textual format. These data encompass diverse patient information and hold utmost importance in monitoring their treatment progress. However, data presented purely as text pose limitations in terms of applicability to computational algorithms for information extraction. To address this issue, a field known as Natural Language Processing (NLP) emerges, encompassing a subfield called Named Entity Recognition (NER). The primary goal of NER is to identify units within text that convey meaningful context, referred to as Entities. Contemporary NER methods employ machine learning techniques on annotated document corpora to achieve robust outcomes. Nonetheless, according to the investigation undertaken in this study, no annotated corpora for the Portuguese language, suitable for training in the domain of psychiatry, the domain explored in the present study, were found. As a result, this study undertook the annotation of a corpus derived from psychiatric medical records, termed as PSYclinBR. This corpus comprises 300 admission notes from a psychiatric emergency service, annotated across 9 semantic categories: self-destructive behavior, diagnosis, drug, medication, psychic function, family history, patient history, observation, and psychological symptom and complaint. The corpus encompasses a total of 2,480 sentences and 5,822 entities, yielding an Inter-Annotator Agreement (IAA) of 0.6. The training of the Named Entity Recognition model, psyBERTpt, which is based on the BERT architecture, was executed using the annotated corpus. This model attained average precision, recall, and f1-score values of 0.65, 0.69, and 0.67 respectively. These values are in proximity to the state-of-the-art results for NER on clinical data in the Portuguese language. Both the annotated corpus and the NER model have been made publicly available.

Keywords: Named Entity Recognition, Clinical Entity Recognition, Corpus Annotation, Clinical Entity Extraction, Natural Language Processing, BERT

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BERT	*Bidirectional Encoder Representations from Transformers*
CER	*Clinical Entity Recognition*
CFM	Conselho Federal de Medicina
CID	Código Internacional de Doenças
CLEF	*Clinical E-Science Framework*
CNN	*Convolutional Neural Network*
CoNLL	*Conference on Computational Natural Language Learning*
CRF	*Conditional Random Fields*
EN	Entidade Nomeada
EN	*English*
FN	Falsos Negativos
FP	Falsos Positivos
GloVe	*Global Vectors for Word Representation*
HAILab	*Health Artificial Intelligence Lab*
IAA	*Inter Annotator Agreement*
IE	*Information Extraction*
IO	*Inside Outside*
IOB	*Inside Outside Begin*
IOBES	*Inside Outside Begin End Single*
IOE	*Inside Outside End*
IREX	*Information Retrieval and Extraction Exercise*
LGPD	Lei Geral de Proteção de Dados
LSTM	*Long short-term memory*
MIMIC	*Medical Information Mart for Intensive Care*
MLM	*Masked Language Model*
NER	*Named Entity Recognition*
NSP	*Next Sentence Prediction*
PLN	Processamento de Linguagem Natural
POS	*Parts of Speech*
PT-BR	Português do Brasil
PT-EU	Português de Portugal
PUC-PR	Pontifica Universidade Católica do Paraná
QA	*Question Answering*
REN	Reconhecimento de Entidades Nomeadas
RNN	*Recurrent Neural Network*
SNOMED CT	*Systematized Nomenclature of Medicine - Clinical Terms*
SVM	*Support Vector Machines*
UIMA	*Unstructured Information Management Architecture*
UMLS	*Unified Medical Language System*
VP	Verdadeiros Positivos

Sumário

Capítulo 1 - Introdução

Em um estudo realizado por Dalianis, Hessel e Vepillai (2009) em um corpus de dados de registros clínicos de pacientes, verificou-se que aproximadamente quarenta por cento dos dados estavam no formato textual. Esses dados dizem respeito a resumos de admissão e alta em serviços de atendimento médico, laudos de exames, anotações da equipe multidisciplinar sobre a evolução do paciente no tratamento da doença, consultas de rotina, encaminhamentos médicos, entre outros, e têm valor inestimável para o entendimento do tratamento por parte dos profissionais envolvidos, além de servir para pesquisas e gestão de custos médicos.

Para obter informações a partir de dados textuais, existe uma área na computação chamada de Processamento de Linguagem Natural (PLN). Uma das principais tarefas em PLN é a extração de informações relevantes no texto e a transformação das mesmas em dados estruturados, conhecida como Reconhecimento de Entidade Nomeada (REN) ou NER (do inglês *Named Entity Recognition*) (SEKINE et al., 2018). É comum que o REN, quando aplicado a prontuários médicos, seja acrescido da palavra "Clínico", de forma que em inglês o REN aplicado à prontuários seja comumente chamado de *Clinical Named Entity Recognition* (*Clinical* NER) (TERUMI et al. ,2020).

As técnicas de REN modernas utilizam modelos de aprendizado supervisionado, onde o algoritmo aprende a partir de exemplos de Entidades Nomeadas (EN) anotadas fornecidos. Na área da saúde, é necessário que esses exemplos rotulados, que são corpus anotados, sejam construídos por especialistas, devido ao conhecimento técnico necessário para a interpretação de textos deste domínio, como a pluralidade de formas de representação de termos, o excesso de abreviações, e a falta de elementos sintáticos para conectar as orações.

Comunidades de pesquisadores de alguns países, como os Estados Unidos, empenharam-se na criação de ambientes favoráveis ao processamento de linguagem natural aplicado à saúde, como a criação de corpus de domínio público, como I2B2, CLEF, CER, MIMIC e a criação de um thesaurus de padronização, o UMLS. Esse ambiente favorável facilitou o desenvolvimento do NER aplicado à área da saúde para a língua inglesa.

No Brasil, por outro lado, não foram desenvolvidos muitos trabalhos de anotação de corpus e extração de informações em prontuários médicos. De acordo com os

levantamentos realizados no presente trabalho, apenas um corpus anotado com entidades clínicas foi disponibilizado publicamente até a presente data: o SemClinBr (OLIVEIRA et al., 2020). O fato de pesquisadores brasileiros não gozarem da pluralidade de corpus anotados com entidades clínicas de variadas formas, os trabalhos pertinentes á este domínio acabam ficando restritos à utilização do SemClinBr ou então carecem da anotação de um novo corpus, tarefa esta que envolve muito trabalho humano.

Considerando o contexto apresentado, neste trabalho foi reunido um conjunto de trezentos prontuários de admissão de pacientes em um serviço de emergência psiquiátrica. Foi elaborado um *guideline* específico para informações presentes em prontuários de psiquiatria e então o corpus formado pela reunião destes prontuários foi anotado de acordo com este *guideline*. O Corpus gerado, que foi batizado de PSYclinBR[1], por sua vez, foi utilizado para realizar o treinamento de um modelo de REN Clínico, o psyBERTpt[2]. Tanto o corpus anotado como o modelo de REN Clínico foram disponibilizados publicamente.

1.1. Motivação e Justificativa

Após o levantamento do estado da arte foi notado que há poucos corpora clínicos disponíveis para a aplicação de tarefas de Reconhecimento de Entidade Nomeada em dados clínicos para a língua portuguesa (OLIVEIRA et al., 2020). Semelhantemente, também há poucas ferramentas e modelos de linguagens disponíveis (TERUMI et al., 2020; DA ROCHA, 2021).

Em outros países, principalmente os de língua inglesa, por outro lado, pode-se encontrar diversos corpora disponíveis publicamente para a realização de tarefas de NLP (I2B2, 2018; JOHNSON et al., 2016, PATEL et al., 2018).

Nem em português nem em inglês encontrados corpora anotados com prontuários de psiquiatria ou que tivessem sido anotados com um *guideline* específico para a extração de informações importantes a este subdomínio da medicina.

A motivação deste trabalho é poder contribuir na diminuição desse déficit de corpora da língua portuguesa, a partir da criação, publicação e disponibilização de um corpus e um modelo de linguagem especializado em prontuários de psiquiatria.

1.2. Objetivos

Este trabalho possui dois objetivos principais e quatro secundários, conforme descrito a seguir.

Objetivos Principais

- Anotação e publicação de um corpus de prontuários médicos em português para a tarefa de REN: o PSYclinBR
- Treinamento e disponibilização de um modelo para REN a partir do corpus gerado: o psyBERTpt.

Tarefas Secundárias:

[1] https://github.com/luizniero/PSYclinBr

[2] https://huggingface.co/psybertpt/psyBERTpt

- Levantamento dos corpora disponíveis para REN em prontuários médicos em Português e outros idiomas.
- Levantamento das ferramentas capazes de realizar a tarefa de REN clínico na língua portuguesa.
- Estudo e apresentação do processo envolvido na criação de um corpus anotado de prontuários médicos.
- Estudo e apresentação das principais técnicas utilizadas para os recentes trabalhos de REN

1.3. Organização do Documento

O presente documento está organizado em mais sete capítulos, divididos da seguinte forma:

- Capítulo 2: Fundamentação teórica sobre processamento de linguagem natural. Neste capítulo, além de sólidos fundamentos necessários à compreensão do PLN, há também um panorama histórico de evolução, desde as técnicas utilizadas nos primórdios do PLN até as arquiteturas mais recentes
- Capítulo 3: O Reconhecimento de entidades nomeadas em prontuários médicos. Neste capítulo são apresentadas algumas ontologias de padronização utilizadas neste domínio, informações sobre o domínio, considerações em relação à legislação vigente no brasil e os tipos de informações que costumam ser extraídas.
- Capítulo 4: Anotação de Corpus. Neste capítulo é realizado um levantamento e uma sólida revisão teórica sobre todos os aspectos que devem ser considerados na anotação de um corpus.
- Capítulo 5: Trabalhos Relacionados. Neste capítulo é apresentado um extenso levantamento sobre trabalhos relacionados a presente dissertação. A relação entre os trabalhos considerou dois aspectos: a anotação de corpus e a aplicação de REN.
- Capítulo 6: Metodologia da geração do corpus. Neste capítulo é apresentada a metodologia utilizada para a geração, anotação do corpus e ajustes realizados para a criação do corpus final: o PSYclinBr, além de serem apresentadas informações sobre ele.
- Capítulo 7: Metodologia do reconhecimento de entidades nomeadas. Neste capítulo é apresentada a metodologia utilizada para o treinamento do modelo de entidades nomeadas baseada no corpus anotado do capítulo anterior, assim como os resultados obtidos pelo modelo.
- Capítulo 8: Conclusão e trabalhos futuros. Breve capítulo de conclusão da dissertação.

Capítulo 2 – Fundamentação Teórica

Extrair informações a partir de dados textuais tem sido objeto de estudo há décadas (SUNDHEIM, 1996). Uma das principais formas de se fazer isso é através do Reconhecimento de Entidade Nomeada. Neste capítulo, serão apresentados os principais conceitos e fundamentos teóricos relacionados ao REN. O capítulo tem por objetivo fornecer uma base sólida pra a compreensão dos métodos e técnicas utilizados em REN. Serão abordados não apenas os métodos considerados como estado da arte mas também os métodos que foram utilizados desde o surgimento do REN.

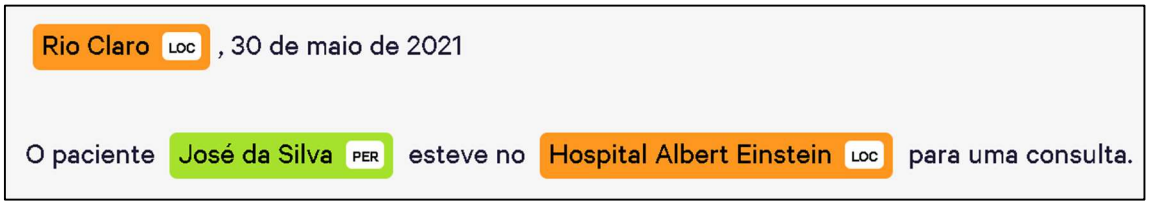
Na primeira seção, é apresentado um panorama histórico do REN, explorando suas origens e evolução ao longo do tempo, além de discutir as diferentes abordagens utilizadas para resolver esse desafio. A segunda seção trata do aprendizado de representações, destacando tipos de *embeddings*, como *word embeddings*, *character embeddings* e representações híbridas. Na terceira seção, é detalhado o modelo BERT, abordando as etapas de pré-treinamento e *fine-tuning*. A quarta seção aborda as métricas de avaliação de REN, apresentando a validação por correspondências relaxadas e a validação por correspondências exatas, que permitem mensurar a qualidade dos resultados obtidos. Por fim, na última seção, é feito um breve resumo do capítulo.

2.1. Reconhecimento de Entidade Nomeada

Reconhecimento de Entidade Nomeada é uma das tarefas primordiais dentro da área de Extração de Informações (SEYLER at al., 2018). A tarefa de REN consiste em detectar a existência de entidades nomeadas em um texto e atribuí-las a seu tipo semântico correspondente, como pessoas, organizações, locais, etc (SEYLER et al., 2018; SEKINE, 2004). As entidades são unidades de informação e podem ser formadas por uma ou mais palavras

A entrada de dados para um sistema de REN é o texto em forma livre e a saída é um conjunto de anotações que classificam e relacionam as entidades que foram reconhecidas naquele texto. Na Figura 1 é ilustrado um processo de REN.

Figura 1 – Reconhecimento de Entidade Nomeada



Fonte - Elaborada pelo Autor

As entidades reconhecidas no processo de REN mostrado na Figura 1 foram: Rio Claro, classificado como localização, José da Silva, classificado como pessoa e Hospital Albert Einstein, classificado como localização.

As entidades nomeadas em REN podem ser divididas em duas categorias: EN genéricas, que surgiram em grande parte com o início do REN, e são representadas por nomes de pessoas, organizações, locais, entre outros; e as EN de domínio específico, que representam nomes de uma determinada área de conhecimento ou domínio, como nomes de doenças, proteínas, genes e drogas, por exemplo, são ENs que pertencem ao domínio biomédico (LI et al, 2020).

Além de sua função auto contida, o REN também desempenha um papel essencial para várias aplicações de processamento de linguagem natural (PLN) como compreensão de texto (ZHANG et al., 2019; CHENG; ERK, 2020), recuperação de informação (GUO et al., 2009; PETKOVA; CROFT, 2007), resumo automático de texto (AONE et al., 1999), pergunta e entendimento (ALIOD et al., 2003), tradução automática (BABYCH; HARTLEY, 2003) e construção de bases de conhecimento (ETZIONI et al., 2005).

2.1.1. Panorama Histórico

O termo “Entidade Nomeada” teve sua origem nos Estados Unidos na década de noventa, durante as Conferências de Entendimento de Mensagens (BORTHWICK; STERLING; GRISHMAN, 1998). Naquela época os pesquisadores que participavam dessas conferências estavam interessados em tarefas de extração de informação (IE, do inglês *Information Extraction*) relacionadas a extrair dados estruturados a partir de textos não estruturados, como artigos de jornais. Ao longo do processo de desenvolvimento de sistemas que realizavam a estruturação destes dados textuais, percebeu-se que haviam algumas unidades de informação que tinham uma elevada importância no processo de detecção. Essas unidades foram: nomes de pessoas, organizações e locais, expressões numéricas, hora, data, moedas e porcentagens. Essas unidades de informação foram definidas como as categorias originais e passaram a ser chamadas de Entidades Nomeadas, e a extração delas ficou definida como uma das tarefas de IE: o REN (SEKINE, 2004).

No início da década seguinte surgiram vários projetos de REN ao redor do mundo. Dois dos grandes projetos que surgiram foram o *Information Retrieval and Extraction*

Exercise, para o idioma Japonês (SEKINE; ISAHARA, 2000) e as tarefas compartilhadas na *Conference on Computational Natural Language Learning* (CoNLL) em 2002 e 2003, para quatro idiomas: Inglês, Alemão, Holandês e Espanhol (CoNLL HP).

Os primeiros sistemas de REN utilizaram como técnica para extração dicionários ou regras (SEKINE, 2004). Porém, pouco tempo depois, foram surgindo novos modelos de aprendizado supervisionado e também sendo testados modelos já existentes. Por fim, a tarefa de REN acabou sendo adotada como uma boa tarefa para experimentos de modelos de aprendizagem supervisionada (SEKINE, 2004). Alguns dos modelos e algoritmos que foram utilizados na época foram: Árvore de decisão (SEKINE; GRISHMAN; SHINNOU, 1998), modelo oculto de Markov (BIKEL et al., 1997), modelo de máxima entropia (BORTHWICK et al., 1998), *voted perceptron* (COLLINS, 2002) e campos aleatórios condicionais (CRF, do inglês *Conditional Random Fields*) (MCCALLUM; Li, 2003).

Assim como os modelos utilizados para a tarefa de REN evoluíram, as categorias das EN também. Com o passar dos anos houve a necessidade de se incorporar novas categorias ao REN, principalmente quando o sistema ou a tarefa de REN tinha como objetivo realizar a extração de informações de dados de um domínio específico, como nomes de doenças. Em pouco tempo, as categorias iniciais não foram o suficiente para representar tudo o que a comunidade esperava que o REN fizesse, e novas categorias especializadas foram sendo criadas (SEKINE; SUDO; NOBATA, 2002).

2.1.2. Tipos de Abordagem

O tipo de abordagem é a estratégia computacional utilizada pelos algoritmos de um determinado sistema para a tarefa de REN. As abordagens para REN podem ser classificadas em três tipos: baseada em regras, baseadas em aprendizado de máquina e híbridas, que combinam ambos (NADEAU; SEKINE, 2007; YADAV; BETHARD, 2019; DEVLIN, 2018), detalhadas a seguir.

2.1.2.1. Abordagem baseadas em regras

Métodos de abordagem baseados em regras, também conhecido como abordagem baseada em conhecimento (AMARAL, 2013) utilizam características sintáticas e semânticas da linguagem previamente conhecidas para realizar o reconhecimento das EN.

As regras precisam ser construídas manualmente e podem envolver expressões, termos e padrões específicos relacionados ao domínio da aplicação, além de também poder fazer o uso de ferramentas linguísticas como expressões regulares, *gazetteers* e dicionários.

Uma das principais características das abordagens baseadas em regras é sua alta precisão. No entanto, essas abordagens também apresentam uma limitação em relação à sua revocação. O motivo disso é que as regras são construídas com base no conhecimento prévio do domínio ou tarefa, e pode ocorrer de determinadas palavras e expressões não serem totalmente cobertas. Além disso, palavras grafadas de forma inadequada ou sinônimos também não serão reconhecidos caso não façam partes das regras (LI et al., 2020) (DEVLIN, 2018).

2.1.2.2 Abordagens baseadas em Aprendizado de Máquina

Em REN, as abordagens baseadas em aprendizado de máquina estão subdivididas em aprendizado supervisionado baseado em recursos e aprendizado não supervisionado, conforme descritos a seguir:

Abordagem de Aprendizado Não Supervisionado

Métodos de abordagem por aprendizado não supervisionado extraem e agrupam as EN em clusters com base na similaridade dos contextos. Essa abordagem busca e explora padrões de linguagem presentes em grandes corpora, sem depender de anotações manuais (NADEAU; SEKINE, 2007; LI et al., 2020).

Ao aplicar métodos de aprendizado não supervisionado, o REN é capaz de descobrir padrões linguísticos e agrupamentos de entidades com base nas similaridades contextuais de um determinado domínio. De forma geral, essa abordagem é útil em situações onde não há anotações manuais disponíveis ou quando se lida com entidades desconhecida, ou então quando o objetivo é encontrar os clusters de agrupamento e as relações entre as palavras (NADEAU; SEKINE, 2007) (LI et al., 2020).

Abordagens de aprendizagem supervisionada baseados em recursos

Método de aprendizagem supervisionada utilizam algoritmos de aprendizado de máquina para aprender a partir de amostras de dados textuais com EN anotadas por humanos. Essas amostras compõem o que é chamado de corpus anotado. A partir do corpus, o algoritmo é treinado para aprender os padrões de EN. Posteriormente os modelos aprendidos podem ser utilizados para reconhecer e classificar as EN em uma tarefa de classificação de multiclasses em dados textuais não rotulados.

Nos modelos de aprendizagem supervisionada são utilizados vários recursos para representar a linguagem natural. Tais recursos podem representar matematicamente orações em vetores n-dimensionais, podem indicar a presença ou ausência de palavras em um determinado trecho de texto, a morfologia e classe gramatical de cada uma das palavras, ou analisar ainda cada subestrutura de uma determinada palavra, como caixa alta ou baixa. Resumindo, tratam-se de maneiras de representar matematicamente dados textuais (NADEAU; SEINE, 2007).

A partir desses recursos, aplica-se então algoritmos de aprendizado de máquina para aprender os padrões dos conjuntos de dados. Os algoritmos mais comuns utilizados para o aprendizado supervisionado são Máquinas de Vetor de Suporte (SVM, do inglês *Support Vector Machines*) (HEARST et al, 1998) e o CRF (LAFFERTY; MCCALLUM; PEREIRA, 2001; LI et al., 2020).

2.2. Aprendizado de Representações

Palavras são elementos de linguagem que possuem como função transmitir uma mensagem. Através do conhecimento da linguagem da mensagem, o receptor é capaz de compreender a informação que o emissor quis transmitir. Uma vez que a mensagem foi entendida pelo receptor, as palavras não são mais necessárias, pois a informação, que era o elemento central da mensagem, já foi transmitida e entendida. A segregação desses

conceitos e da importância de cada um dos elementos na linguagem é importante para compreendermos que as palavras carregam uma informação muito mais ampla que apenas a sua forma sintática ou sua capitalização (JURAFSKY; MARTIN, 2023).

Dada a complexidade de entendimento que uma palavra pode carregar, um modelo que represente uma palavra apenas pela sua forma sintática ou capitalização é um modelo insuficiente (CARLSON, 1977). Para deixar mais claro a insuficiência do modelo, observe o exemplo de pergunta e resposta abaixo (JURAFSKY; MARTIN, 2023):

Pergunta: Qual o significado da vida?

Resposta: VIDA

Está claro que a própria palavra, em sua forma, não é suficiente para representá-la. Há elementos que estão intrínsecos ao significado da palavra e que podem auxiliar na criação de um modelo mais representativo, como por exemplo, sinônimos, conotações positivas e negativas, entre outros. De forma geral, um modelo para a representação de uma linguagem suficientemente bom deve permitir que sejam feitas inferência sobre o significado da palavra (JURAFSKY; MARTIN, 2023), o que claramente não é possível no exemplo acima.

No ambiente computacional, é necessário gerar também modelos que sejam capazes de representar as palavras, para que os algoritmos consigam realizar tarefas sobre esses modelos. Para realizar essa representação, foram criados os vetores semânticos. A ideia desses vetores é representar uma palavra como um ponto em um espaço semântico multidimensional, derivado a partir da distribuição palavras vizinhas. Esses vetores semânticos ficaram conhecidos como *embeddings*, devido à semelhança que eles tinham com a operação matemática de mapeamento de um espaço ou estrutura para outro, chamada *embedding* (JURAFSKY; MARTIN, 2023).

Desde o início da exploração do *embedding* em PLN, vários modelos foram propostos, não somente para representar as palavras e suas relações, que foram chamados de *word embeddings*, mas também para representar os caracteres que compõe as palavras, que foram chamados de *character embeddings*, além de representações híbridas (LI et al., 2020).

2.2.1. Word Embeddings

Word embedding é a representação matemática de uma palavra dentro do contexto em que ela está inserida. Esses *embeddings* são vetores n-dimensionais, onde cada dimensão possui um valor numérico. A quantidade de dimensões dos vetores e os possíveis valores que podem ser assumidos dependem do algoritmo utilizado para a geração do *embedding*. Palavras que aparecem em contextos semelhantes tendem a ter *embeddings* semelhantes, enquanto palavras que não aparecem no mesmo contexto tendem a ter *embeddings* muito diferentes. Ou seja, é possível verificar a afinidade entre duas palavras a partir da comparação dos *embeddings* de ambas (WKY; MARTIN, 2023).

Existem dois tipos de algoritmos de *word embeddings*, os estáticos e os dinâmicos. Nos algoritmos de *embeddings* estáticos os *embeddings* de cada palavra têm sempre o mesmo valor para um mesmo vocabulário, enquanto nos algoritmos de *embeddings* dinâmicos os valores que variam de acordo com o contexto, a sequência de tokens, entre outras informações. São exemplos de algoritmos de *embeddings* estáticos o word2vec o fastText e o GloVe (JURAFSKY; MARTIN, 2023), e de *embedding* dinâmico o *Bidirectional Encoder Representations from Transformers* (BERT), o *Embeddings* from *Language Models* (ELMo) (PETERS et al., 2018) e o Generative Pre-training Transformer (GPT) (RADFORD et al., 2018).

Ambos os tipos de *word embeddings* são treinados por métodos chamados auto supervisionados. Esses métodos possuem esse nome porque os algoritmos são capazes de aprender, a partir do corpus, as relações entre as palavras na linguagem (JURAFSKY; MARTIN, 2023; LI et al., 2020).

2.2.2. Character Embeddings

O *character embedding* surgiu como uma adição à representação por *word embedding*, para potencializar a compreensão do modelo sobre o vocabulário (KURU; CAN; YURET, 2016) (TRAN; MACKINLAY; YEPES, 2017). Essa representação a nível de caracteres torna possível explorar informações que estão contidas no contexto da palavra, e, portanto, não são capturadas pelos *word embeddings*, como sufixo e prefixo, letras maiúsculas e minúsculas e outras características de vocabulário (LI et al, 2020).

Há dois algoritmos principais utilizados para gerar o *character embedding*: os baseados em redes neurais convolucionais (CNN, do inglês *Convolutional Neural Network*) e os baseados em redes neurais recorrentes (RNN, do inglês *Recurrent Neural Network*) (LI et al, 2020).

2.2.3. Representações Híbridas

As representações híbridas combinam os *embeddings* utilizados nas abordagens que utilizam aprendizado de máquina, como *word* e *character embedding*, com os recursos proveniente de fontes externas do conhecimento. Esses recursos são adicionados posteriormente aos *embeddings* de nível de palavra e caractere para incorporar informações adicionais à representação e podem melhorar a performance do REN, porém podem trazer prejuízos à generalização (LI et al., 2020).

É comum em abordagens híbridas, adicionar aos *embeddings* de *word* e *character gazetteers* (HUANG; XU; YU, 2015; LIU; YAO; LIN, 2019), similaridades léxicas (GHADDAR; LANGLAIS, 2018), dependências linguísticas (KURU; CAN; YURET, 2016) e grafos de conhecimento, sendo este último também chamado de *Knowledge Embedding* (WANG et al., 2021)

Um dos *embeddings* de abordagem híbrida é o *knowledge embedding*. O *knowledge embedding* codifica as entidades e suas relações em grafos de conhecimento como representações distribuídas. Grafos de conhecimento são grafos onde cada entidade é

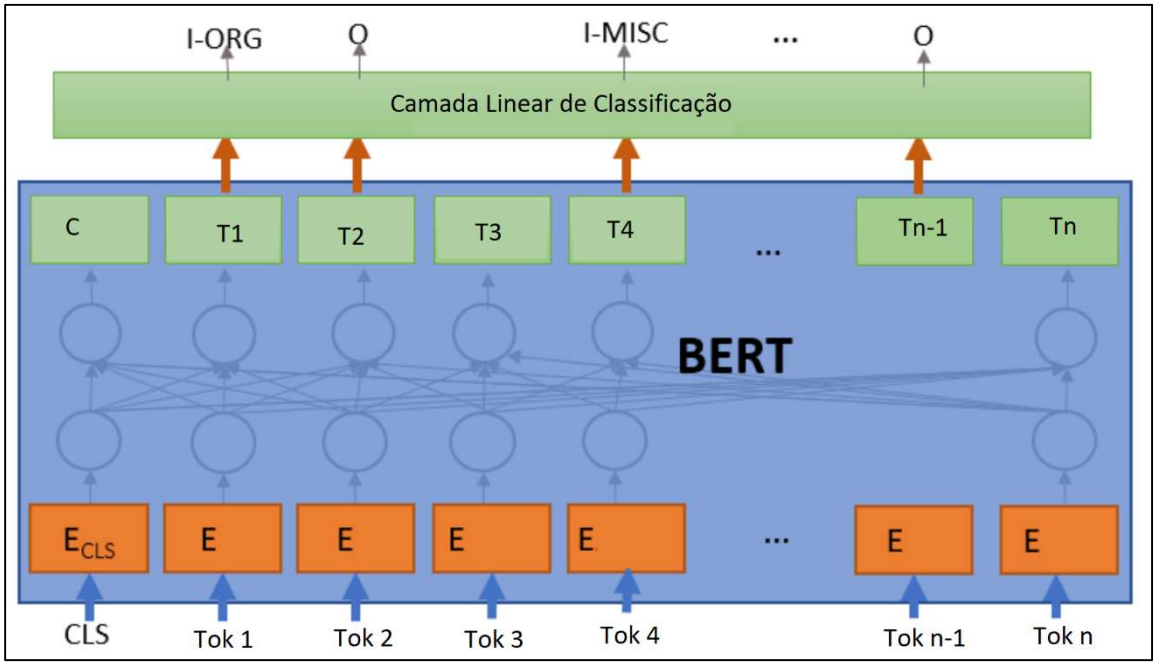
representada por um nó e as relações entre as entidades são arestas. Em modelos de *knowledge embedding* convencionais, é atribuído um vetor com n dimensões a cada entidade e relação, e é utilizada uma função de pontuação para treinar os *embeddings* e prever links. O *knowledge embedding* é capaz de beneficiar algumas tarefas de *downstream*, como predição de link e extração de relação (WANG et al., 2021). A adição de *knowledge embeddings* a modelos baseados em aprendizagem profunda, permite que o modelo incorpore conhecimento de domínio especializado estruturado em uma base de conhecimento (MICHALOPOULOS, 2020).

2.3. BERT

BERT é acrônimo de *Bidirectional Encoder Representations from Transformers*. BERT é um modelo de aprendizado profundo (*deep learning*) cuja arquitetura é composta pela camada de codificação dos *Transformers,* que podem ou não estar conectadas a uma camada final de um algoritmo de classificação, a depender da tarefa de PLN que esteja sendo executada pelo BERT.

Os *Transformers* são uma arquitetura baseada em redes neurais e mecanismos de atenção, projetadas para capturar relações entre palavras em um contexto bidirecional e são utilizados pelo BERT especificamente na camada de *encoding*, na qual o texto é transformado em representações vetoriais.

Figura 2 – Arquitetura do BERT para a tarefa de REN



Fonte: Adaptado de Moon et al. (2019)

As representações obtidas pelo BERT através do empilhamento de várias camadas de *encoding* são utilizadas para a realização das tarefas de PLN. Na Figura 2 está

representada a arquitetura do BERT utilizada para a tarefa de REN. Nela, cada círculo representa uma unidade de *transformer*, conectada à uma camada final de um algoritmo de classificação, que tem por objetivo classificar cada um dos tokens a partir das representações geradas pelos *transformers*.

O BERT atingiu o estado da arte em várias tarefas de processamento de linguagem natural utilizando uma arquitetura única para todas as tarefas de NLP, onde apenas a camada de saída precisa ser diferente, de acordo com a tarefa que está sendo executada. Isso possibilita que a etapa de pré-treinamento, que será explorada na sessão 2.3.1, seja a mesma para várias tarefas (DEVLIN, 2018).

Devido à sua capacidade de realizar o treinamento em duas etapas bem divididas: pré-treinamento e *fine-tunning*, o BERT é considerado um 'modelo de representação baseado em pré-treinamento', onde é possível utilizar modelos já treinados para servir como base em tarefas específicas de PLN.

As etapas de pré-treinamento e *fine-tunning* do BERT estão detalhadas nas sessões 2.3.1 e 2.3.2.

2.3.1. A etapa de Pré-Treinamento

A etapa de pré-treinamento do BERT utiliza transformers conectados bidirecionalmente para aprender a representação das palavras no contexto das frases e suas relações com as demais palavras tanto no sentido da direita-para-esquerda quanto no sentido da esquerda-para-direita. Para a versão do BERT em inglês, os *embeddings* foram gerados utilizando como repositório de dados a base Wikipedia em inglês (2.5 bilhões de palavras) e do BookCorpus (800 milhões de palavras). Para a sua versão em português, o BERTimbau, os *embeddings* foram gerados utilizando como repositório o BrWac (2.68 bilhões de palavras) (SOUZA; NOGUEIRA; LOTUFO, 2020). Pode-se dividir a etapa de pré-treinamento em duas sub-etapas: Modelo de Linguagem Mascarada (MLM, do inglês *Masked Language Model*) e Predição de Próxima Sentença (NSP, do inglês *Next Sentence Prediction*) (DEVLIN, 2018).

É importante ressaltar que, embora existam modelos do BERT já pré treinados e disponíveis para a utilização, também é possível pré-treinar um novo modelo do BERT com dados específicos de um domínio, desde que existam textos o suficiente para isso.

Modelo de Linguagem Mascarada

O MLM é uma etapa fundamental do processo de pré-treinamento do BERT. Nesta etapa uma parte do texto de entrada é aleatoriamente mascarada, substituindo algumas palavras por um token especial "[MASK]". O objetivo da substituição é fazer com que o modelo aprenda a prever a palavra mascarada com base no contexto fornecido pelas palavras adjacentes. Essa técnica de treinamento permite ao BERT desenvolver sua compreensão contextual e capturar relacionamentos entre as palavras de acordo com o contexto (DEVLIN, 2018).

Durante o pré-treinamento, o MLM deve conseguir prever corretamente as palavras que foram ocultadas. Dessa forma, ao mascarar e prever palavras ocultas, o BERT aprende a capturar informações contextuais e a fazer inferências baseadas em fragmentos do texto disponíveis (JURAFSKY; MARTIN, 2023). Isso é particularmente útil para tarefas de REN, onde o contexto em torno de uma entidade nomeada pode ser crucial para sua correta identificação e classificação, já que tokens iguais podem ter diferentes significados a depender do contexto (DEVLIN, 2018).

Uma vantagem do MLM é que ele não depende de anotações manuais, ou seja, seu aprendizado é auto supervisionado, tornando o pré-treinamento mais escalável e eficiente. As representações aprendidas pelo MLM no pré-treinadas são refinadas na etapa de *fine-tuning*, onde o modelo é ajustado para uma tarefa específica de REN (DEVLIN, 2018).

Predição de próxima sentença

Algumas tarefas de PLN necessitam que o algoritmo entenda a relação entre duas sentenças, como é o caso das respostas às perguntas (QA, do inglês *Question Answering*) e Inferência de Linguagem Natural. Porém, no algoritmo de aprendizado bidirecional utilizado na primeira subetapa do pré-treinamento, esse tipo de relação não é aprendido. Para dar suporte à essas tarefas de PLN, o BERT incorporou preditor binário de próximas sentenças (DEVLIN, 2018).

O objetivo final da etapa de pré-treinamento é transferir todos os parâmetros aprendidos com o processamento não supervisionado do modelo para a próxima etapa: de *fine tunning*, que será a etapa na qual de fato será aplicada alguma tarefa de PLN.

2.3.2. A etapa de *fine tunning*

A etapa de *fine tunning* herda os parâmetros aprendidos pela etapa de pré-treinamento. Nesta etapa, há a possibilidade de ajustar os parâmetros aprendidos a partir do fornecimento de dados não rotulados de um determinado domínio, como, por exemplo, prontuários médicos, fazendo que o BERT se torne mais eficiente na compreensão da linguagem utilizada nesse domínio, além da possibilidade de utilizar o modelo pré-treinado para a realização de uma tarefa de PLN específica, comumente chamada de tarefa de *downstream*.

Para as tarefas de *downstream* é necessário adicionar uma camada de classificação conectada completamente à camada de saída do BERT. Essa camada de classificação utilizará os parâmetros aprendidos no pré-treinamento e os dados rotulados fornecidos para poder realizar a tarefa de *downstream* especificada, como NER, *Question Answering* ou classificação de sentença (DEVLIN, 2018). Na Figura 2 é possível observar uma camada linear de classificação. Essa camada de classificação é composta por uma camada de conexão linear entre os parâmetros mais uma camada do algoritmo *Softmax*.

O algoritmo *softmax* ou função *softmax*, como costuma ser chamado, é um algoritmo de aprendizado supervisionado utilizado para gerar a distribuição probabilística para a predição dos tokens, de forma que a soma das probabilidades de cada classificação possível seja 1.A camada softmax é apenas uma camada de rede neural, treinada durante a etapa de *fine-tunning*, totalmente conectada com a função de ativação *softmax*

(MUNIKAR; SHAKYA; SHRESTHA, 2019). Na Figura 2, alguns dos tokens preditos pela função softmax são: *I-ORG*, *O* e *I-MISC*.

A possibilidade de realizar diferentes *fine-tunnings* utilizando o mesmo modelo pré-anotado do BERT possibilita que novos modelos de tarefas de *downstream* sejam treinados com pouco processamento de dados, já que a tarefa de *fine-tunning* precisa de muito menos poder de processamento que a sua tarefa anterior, de pré-treinamento (DEVLIN, 2018).

2.4. Métricas de Avaliação de REN

Sistemas de REN são geralmente avaliados através da comparação entre o resultado da saída gerada pela aplicação do REN com anotações humanas. A comparação pode ser feita considerando correspondências exatas ou correspondências relaxadas, porém é raro de encontrar trabalhos publicados que utilizam como estratégia de validação métodos de correspondências relaxadas, pois tais métodos dão margens a interpretações erradas (LI et al., 2020).

2.4.1. Validação por Correspondências Relaxadas

Os métodos de correspondência relaxadas foram os primeiros a serem criados para REN e atualmente são pouco utilizados. O mais antigo deles consiste em avaliar a entidade pela classificação e limite separadamente, da seguinte forma (NADEAU; SEKINE, 2007):

- Se a entidade for classificada corretamente e houver intersecção entre o limite correto e o limite reconhecido da entidade, então é creditado um tipo.
- Se os limites da entidade forem reconhecidos corretamente, então um limite é creditado.

O score final do método, *micro-averaged f-measure*, é a média entre a precisão e o revocação calculado entre os créditos de tipo e limite de todas as entidades (NADEAU; SEKINE, 2007).

2.4.2. Validação por Correspondências Exatas

Os métodos por correspondências exatas foram criados nas conferências CoNLL e são largamente utilizados para avaliação de REN atualmente. Neste método de avaliação, para que uma entidade seja considerada corretamente reconhecida é preciso garantir que ela tenha sido identificada corretamente em relação aos limites e ao tipo (PRADHAN et al., 2012) (DOGAN et al., 2019).

Nesse tipo de validação, são utilizados os números de falsos positivos (FP), falsos negativos (FN) e verdadeiros positivos (VP) para calcular a precisão, o revocação e o f-score. A seguir temos uma breve descrição de FP, FN e VP, precisão, revocação e F-score e as devidas equações de precisão, revocação e f-score.

- Falsos positivos (FP) são entidades que foram reconhecidas pelo sistema de REN, porém não deveriam ter sido reconhecidas pois não constam nas anotações manuais.
- Falsos negativos (FN) são entidades que não foram reconhecidas pelo sistema de REN, porém deveriam ter sido reconhecidas, pois constam nas anotações manuais.
- Verdadeiros positivos (VP) são entidades que foram reconhecidas pelo sistema de REN e que deveriam ter sido reconhecidas, ou seja, estão corretas.
- Precisão refere-se à percentagem dos resultados obtidos pelo sistema que foram reconhecidos corretamente.
- Revocação refere-se à percentagem total de entidades que foram corretamente reconhecidas.
- O F-β-*score* é uma medida que combina precisão e revocação. Geralmente utiliza-se β igual a 1, o que é chamado de F1-*score,* F-*score* ou apenas F1.

$$\text{Precisão} = \frac{VP}{(VP + FP)}$$

$$\text{Revocação} = \frac{VP}{(VP + FN)}$$

$$\text{F-beta-score} = 2 \times \beta \frac{Precisão \times Recall}{Precisão + Recall}$$

2.5. Considerações Finais

Neste capítulo foram apresentados os principais conceitos de REN, seu panorama histórico, abordagens e métricas de avaliação para REN independente do domínio de aplicação ou objetivo das tarefas. Foi possível observar que o REN está em constante mudança e que várias abordagens e métricas de validação foram sugeridas desde que foi proposta.

No capítulo seguinte serão abordados o REN e as suas particularidades para o domínio médico, fundamentais para o entendimento deste trabalho e de sua importância.

Capítulo 3 - Reconhecimento de Entidades Nomeadas em Prontuários Médicos

Grande parte dos registros clínicos presentes em prontuários médicos estão em formato textual. Nesses registros estão informações relevantes sobre sintomas, causas, hipóteses de diagnósticos de doenças do paciente, entre outras informações necessárias ao avanço do tratamento do paciente pela equipe multidisciplinar. Para transformar as informações textuais em dados que possam ser processados com mais facilidade, é necessário o uso de técnicas que permitem extrair informações relevantes de forma automatizada destes dados textuais e transformá-las em estruturados. Neste contexto é comum a utilização do REN clínico.

Neste capítulo é explorado o prontuário médico e algumas particularidades existentes na aplicação de REN neste domínio, como a padronização de terminologias, os tipos de aplicações mais comuns, como a detecção de informações pessoais e a detecção de fatos clínicos relevantes.

3.1. Padronização de Terminologias

No domínio da saúde existem várias formas de se expressar os mesmos conceitos. A falta de padronização para esses conceitos é uma das dificuldades encontradas no uso das técnicas de PLN envolvendo prontuários médicos (LINDBERG; HUMPHREYS; MCCRAY, 1993). O uso de ferramentas como tesauros e dicionários de terminologias médicas é uma das técnicas adotadas na resolução deste problema (DALIANIS, 2018). Entre os padrões mais comuns aplicados em notas clínicas, estão o Sistema unificado de Linguagem médica e o Código Internacional de Doenças.

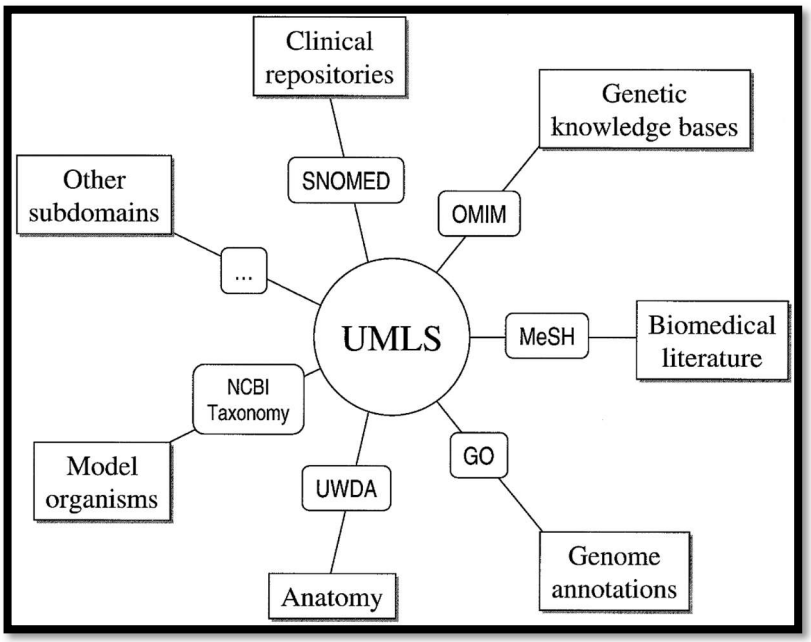
3.1.1. Sistema Unificado de Linguagem Médica

O Sistema Unificado de Linguagem Médica (UMLS, do inglês *Unified Medical Language System*) é um Metatesaurus que foi desenvolvido pela Biblioteca Nacional de Medicina do Estados Unidos como um esforço para superar duas grandes barreiras existentes na recuperação de informação (LINDBERG; HUMPHREYS; MCCRAY, 1993): a variedade de nomes usados para expressar o mesmo conceito; e, a ausência de um padrão para a distribuição de terminologias.

O UMLS concentrou seus esforços na unificação de terminologias presentes em vários vocabulários médicos em um único identificador de conceito, que é utilizado como identificador padrão de todos os termos sinônimos (CANTOR et al., 2003).

O UMLS na versão de 2021AA integra 157 famílias de vocabulários biomédicos e contém 16 milhões de nomes mapeados para 4.4 milhões de conceitos. Dos 16 milhões de nomes, 425 mil correspondem a nomes da língua Portuguesa, correspondendo a 2,64% dos nomes, enquanto os nomes da língua inglesa correspondem a 70,88% (UMLS, 2021). Na Figura 4 é possível observar o UMLS ao centro como tesaurus integrador de várias ontologias e taxonomias, algumas delas especificas para a área médica são descritas a seguir.

Figura 3 – Os Vários Subdomínios Integrados ao UMLS



Fonte: Extraído de Cantor et al., 2003.

3.1.2. Código Internacional de Doenças

O código Internacional de Doenças (CID) é uma terminologia padrão internacional muito utilizada na medicina para o mapeamento de doenças (DALIANIS, 2018). A primeira versão do CID, o CID-1 foi publicada em 1900 e usada até 1909. Atualmente, o CID está em sua décima versão, o CID-10, disponível em 42 idiomas e contendo 32 mil códigos de diagnósticos divididos em 22 grupos diferentes. Os códigos no CID são hierárquicos e vão do mais geral até o mais específico, conforme pode ser observado na Tabela 1 para o exemplo do CID-10 para F32 – "Episódios depressivos".

Tabela 1 – Hierarquia do CID-10 para o F32 - Episódios Depressivos

CID	Descrição
F32	Episódios depressivos
F32.0	Episódio depressivo leve
F32.1	Episódio depressivo moderado
F32.2	Episódio depressivo grave sem sintomas psicóticos
F32.3	Episódio depressivo grave com sintomas psicóticos
F32.8	Outros episódios depressivos
F32.9	Episódio depressivo não especificado

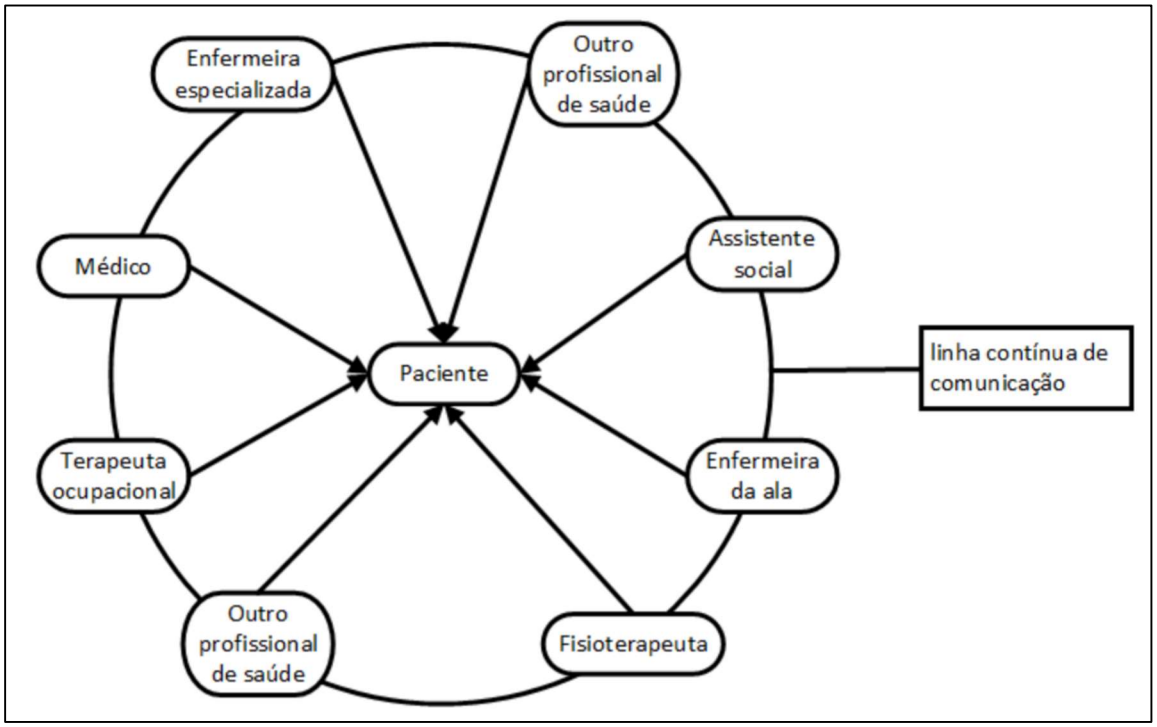
3.2. O Prontuário Médico

O Prontuário médico é um conjunto de documentos previsto pelo Conselho Federal de Medicina (CFM) como meio indispensável para aferir a assistência médica prestada e elemento valioso para o ensino e pesquisa, além de servir também como instrumento de defesa legal.

Prontuários médicos podem conter informações valiosas sobre sintomas, diagnósticos, tratamentos, e prescrições, geradas a partir de fatos, acontecimentos e situações sobre a saúde do paciente e a prestação do serviço médico ao mesmo, além de conter informações pessoais de pacientes como nome, telefone e endereços (DALIANIS, 2018) (CFM, 2002).

O prontuário médico é gerado pelos vários tipos de profissionais de saúde que fazem parte da equipe multidisciplinar de assistência ao paciente, como médicos, enfermeiros, fisioterapeutas, psicólogos, entre outros (DALIANIS, 2018) (CFM, 2002). Neste contexto de multidisciplinaridade, o prontuário não apenas possibilita a comunicação entre membros da equipe multiprofissional (DALIANIS, 2018) (CFM, 2002) mas atua também como uma ferramenta de Interdisciplinaridade, Multidisciplinaridade e Transdisciplinaridade (DE CARVALHO, 2017). Na Figura 5 estão ilustradas as múltiplas disciplinas envolvidas no tratamento de um paciente e uma linha de comunicação direta entre elas, que é o prontuário.

Figura 4 – Equipes multidisciplinares envolvidas no cuidado do paciente



Fonte: Thompson e Wright (2003), traduzida por de Carvalho (2017)

3.2.1. Características Linguísticas do Prontuário Médico

Prontuários médicos, também chamados de registros médicos, registros clínicos, registros do paciente (DALIANIS, 2018) costumam conter maior número de inconformidades gramaticais quando comparados a textos de outros gêneros, especialmente quando se trata de registros de uso interno, como registros de uma internação hospitalar (EHRENTRAUT et al., 2012).

Allvin et al. (2011) realizou um estudo onde comparou as anotações de enfermagem de hospitais da Finlândia e na Suíça. Verificou-se que em ambos os países existiam os mesmos vícios de anotação: As frases raramente estavam completas, verbos auxiliares quase nunca eram utilizados, os pacientes eram mencionados de forma abreviada ou então as frases eram de sujeito ausente, as frases eram formadas apenas pela aglutinação de vários sintomas do paciente, sem elementos de coesão textual, além de vários erros gramaticais. De acordo com Pakhomov, Pedersen e Chute (2006), trinta por cento dos registros clínicos trata-se de abreviações, acrônimos ou erros gramaticais, o que evidencia que há um grande ruído nos dados de prontuários médicos. Além disso, até trinta e três por cento das abreviações costumam ser ambíguas (DALIANIS, 2018)

Geralmente prontuários são escritos por profissionais da saúde especialistas que utilizam de padrões de escrita específicos de sua área de especialidade, como abreviações, jargões médicos ou mesmo formatações específicas, o que pode ser de difícil compreensão para profissionais de outras áreas de especialidade (DALIANIS, 2018), evidencializando mais uma característica específica deste domínio.

Parte das características supramencionadas podem ser atribuídas a diversos fatores, como a comum rotina sobrecarregada dos profissionais de saúde e a falta de tempo para a

realização desses registros, a falta de ferramentas corretoras gramaticais em sistemas de prontuários (DALIANIS, 2018) além da grande variedade de termos existentes para representar as mesmas informações médicas, como sinonímias, abreviações e siglas.

3.2.2. Disponibilidade e Acesso a Corpus de Prontuários

Não há muitos corpora de prontuários médicos disponíveis para pesquisa. A razão para isso é que os prontuários contêm informações confidenciais dos pacientes e costumam estar sujeitos à determinadas leis, a depender dos países. A maioria dos corpora de prontuários disponíveis para pesquisa está em inglês, porém há alguns corpora disponíveis para outros idiomas (DALIANIS, 2018).

No Brasil, o acesso a prontuários médicos para pesquisas deve obedecer à Resolução nº 196 de 10 de outubro de 1996 do Conselho Federal de Medicina (CFM, 1996), que sujeita o acesso aos dados para a pesquisa em prontuários médicos à aprovação do Comitê de Ética e Pesquisa da instituição responsável pela guarda dos dados, e também cumprir com os requisitos da Lei nº 13.709 de 2018 - Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018), que enquadra os dados de saúde como dados pessoais sensíveis e estabelece que quaisquer dados pessoais utilizados para pesquisa devem estar devidamente anonimizados.

3.3. As aplicações do Reconhecimento de Entidades Nomeadas em Prontuários Médicos

Em prontuários médicos, o REN tem duas aplicações primárias. A primeira delas diz respeito à identificação de dados pessoais dos pacientes para aplicação de anonimização, como dados pessoais de nome, telefone e endereço, que podem eventualmente estar presentes nestes documentos. A segunda delas diz respeito à identificação das informações clínicas relevantes presentes nos registros, como sintomas, diagnósticos, tratamentos, e outras informações relevantes que podem estar presentes no prontuário (DALIANIS, 2018).

3.3.1. Detecção de Dados Pessoais dos Pacientes

Além das informações sobre o tratamento do paciente, os prontuários médicos podem conter dados pessoais dos pacientes. Para que um prontuário possa ser utilizado para pesquisas é necessário que todos os dados pessoais contidos no documento sejam anonimizados. Portanto, é necessário reconhecer as entidades que dizem respeito a nomes, endereços e outros dados que tornem o paciente reconhecível afim de removê-las ou trocá-las por dados falsos, para então poder utilizar os documentos em pesquisas.

3.3.2. Detecção de Informações Clínicas Relevantes

Assim como existem vários agentes de interesse no prontuário, também existem vários agentes que podem estar interessados em extrair informações de prontuários médicos. Para cada tipo de agente, determinados tipos de EN são mais ou menos relevantes.

Médicos que estão responsáveis pelo tratamento do paciente estão interessados em eventos que indiquem o quadro de saúde do paciente e como ele pode estar relacionado com o uso de drogas ou doenças familiares ou anteriores. Médicos pesquisadores podem estar interessados em uma progressão temporal maior ou então em apenas alguns eventos do paciente. Gestores de hospitais podem estar interessados em uso de drogas genéricas, solicitações de exames e procedimentos médicos, e assim por diante. Assim sendo, o REN em prontuários médicos pode ter vários objetivos, a depender do agente de interesse dos dados.

Além das detecções das entidades, para que seja possível extrair as informações de contexto relacionadas aos fatos, em prontuários médicos também existe a necessidade de detectar algumas informações secundárias que complementam ou alteram o contexto das entidades, que são as negações e as incertezas ou fatualidades a respeito das mesmas.

Detecção de Negações

Detectar negações nos textos clínicos é uma tarefa importante porque muitas vezes a negação irá modificar a entidade que foi extraída pelo processo de REN, implicando um valor semântico contrário ao sugerido pela entidade extraída. A negação de uma entidade clínica, como um sintoma implica um contexto tão importante quanto o reconhecimento do mesmo para a correta compreensão dos dados dos prontuários. A negação pode ser considerada como um caso especial de detecção de fatualidade, onde a presença da negação junto ao fato invalida o mesmo (DALIANIS, 2018).

Detecção de Fatualidade ou Incerteza

Outras características encontradas nos textos clínicos são as expressões de níveis de fatualidade ou grau de incerteza. As expressões podem variar de totalmente especulativas, quando há apenas uma sugestão de que algo esteja ocorrendo com o paciente, por exemplo, até completamente afirmadas, quando uma hipótese é confirmada (DALIANIS, 2018). Detectar o nível da fatualidade que está associada a uma entidade é importante porque pode alterar a interpretação que o especialista deve dar à entidade.

Detecção de Históricos

Algumas doenças estão associadas a fatores hereditários. Por conta disso, é comum que no prontuário existam anotações de sintomas ou diagnósticos que estejam associadas a históricos familiares do paciente, chamados de histórico familiar, ou até mesmo a episódios relevantes que o paciente tenha tido no passado e que não estejam diretamente associados com o atual problema do paciente, mas que o especialista de saúde considerou importante registrar, chamados de histórico do paciente (DALIANIS, 2018). Ser capaz de identificar se as entidades reconhecidas estão relacionadas com históricos familiares ou históricos do paciente também é importante no processo de REN em prontuários médicos.

Detecção Temporal

Outra importante informação presente nos prontuários é o tempo em que os sintomas, doenças ou eventos ocorreram. Relatos de sintomas e doenças que ocorreram no passado não podem ser apresentados como se estivessem ocorrendo no presente, pois possuem interpretações diferentes para o diagnóstico médico. Dessa forma, a informação temporal associada à EN pode mudar a interpretação do especialista sobre ela no contexto de análise do prontuário. Além disso, a compreensão do tempo em que os eventos ocorreram é importante para conseguir entender a progressão de eventos no tratamento do paciente (DALIANIS, 2018).

3.4. Considerações Finais

Neste capítulo foram apresentadas as particularidades envolvidas na aplicação de REN no domínio de prontuários médicos, como a importância da detecção de informações secundárias relacionadas às EN, além das múltiplas formas em que os mesmos conceitos podem ser representados. A partir das particularidades pode-se verificar que a tarefa de REN em prontuários médicos está sujeita a particularidades intrínsecas a este domínio e que devem ser levadas em conta durante todo o processo de REN a depender das abordagens que forem escolhidas para tal.

No capítulo seguinte será apresentada a fundamentação teórica sobre anotação de corpus, fundamental para a aplicação das técnicas de REN, além de ser um dos objetos deste trabalho.

Capítulo 4 – Anotação de Corpus

Um corpus é um conjunto de documentos no formato textual computacionalmente processável (JURAFSKY; MARTIN, 2023). Um bom corpus deve ser capaz de representar as características presentes em seu domínio (MANNING; SCHUTZE, 1999). Em tarefas de PLN, os dados que serão trabalhados podem ser um ou vários corpora, além disso, o conjunto de exemplos a serem seguidos por técnicas de aprendizado de máquina também costumam ser corpus com características chamadas de anotação. A anotação de um corpus é o processo de enriquecer um corpus adicionando a ele informações de linguística ou domínios específicos, que são incorporados na forma de anotações. As anotações podem ser inseridas por pessoas ou algoritmos (HOVY; LAVID, 2010).

Uma vez que um corpus está devidamente anotado, as informações originais textuais do corpus e as informações anotadas podem ser utilizadas por algoritmos de PLN. As aplicações que costumam envolver corpus anotados são o REN, a normalização de entidades nomeadas, a extração de relações e a QA (HUANG et al., 2020).

Neste capítulo são apresentadas técnicas sobre a anotação de corpus, tanto para o domínio geral como para o domínio de prontuários médicos.

4.1. Tipos de Anotação

A tarefa de anotação de um corpus depende do objetivo desejado com os dados resultantes. Para o REN, é comum que as anotações adotem modelos conceituais disponíveis em ontologias, dicionários e regras de acordo com o domínio que está sendo anotado. Também é possível, dentro de um mesmo domínio, que haja diferentes estratégias de anotação, visto que as estratégias escolhidas serão de acordo com os objetivos definidos para a anotação daquele corpus.

Apesar das diversas formas de se anotar, há basicamente duas informações que podem ser anotadas em um texto: as informações relativas às funções das palavras como elementos gramaticais dentro de orações, chamada de anotação sintática, e as informações

relativas às funções das palavras em relação ao seu entendimento dentro do contexto das orações, chamada de anotação semântica, ambas descritas a seguir.

4.1.1. Anotação Morfossintática

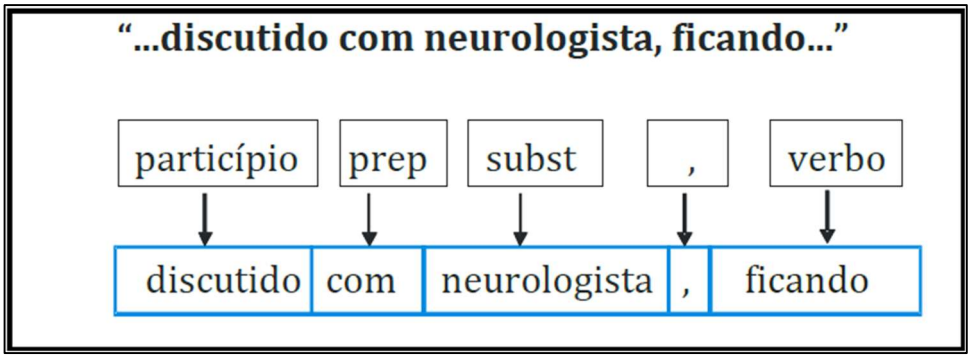
A anotação morfossintática é comumente encontrada na literatura como *POS tagging*. *POS* vem do inglês *parts of speech* e pode ser traduzido como partes do discurso, e *tagging* pode ser traduzido como anotação; é o procedimento que consiste em anotar todas as palavras presentes no texto de acordo com a função sintática de cada uma delas. Na anotação morfossintática, a função semântica das palavras não importa ao anotador, mas apenas a função gramatical que elas estão exercendo. São exemplos de funções gramaticais anotadas: sujeito, predicado, objeto direto, etc.

A anotação sintática está mais vinculada à linguagem do que ao domínio e, por conta disso, não necessita de especialistas de domínio para a anotação. Além disso, anotações sintáticas tendem a convergir bem para a maioria dos domínios da linguagem, desde que eles mantenham uma linguagem de comunicação comum, com a presença dos elementos sintáticos.

A razão de se anotar elementos sintáticos é que eles dão dicas úteis para estruturar o significado de uma oração. Saber se uma palavra é um substantivo ou um verbo ajuda a predizer sobre as possíveis palavras vizinhas, tornando o *POS tagging* um aspecto importante para a análise textual por algoritmos. Como exemplo da utilização da *POS tagging* pode-se citar na tarefa de desambiguação (JURAFSKY; MARTIN, 2023).

Apesar de sua importância na compreensão da linguagem e no auxílio da compreensão dos elementos linguístico, é incomum a anotação de *POS tagging* em corpus de domínios específicos, como o de prontuários médicos, sendo mais numerosas as anotações semânticas. Em Pakhomov, Coden e Chute (2006) os autores fizeram o uso do *POS tagging* em dados de prontuários médicos. Na Figura 6 é possível observar um trecho de texto com *POS tagging*.

Figura 5 – Texto de prontuário com POS *tagging*



Fonte: Extraído de Peters et al., 2010.

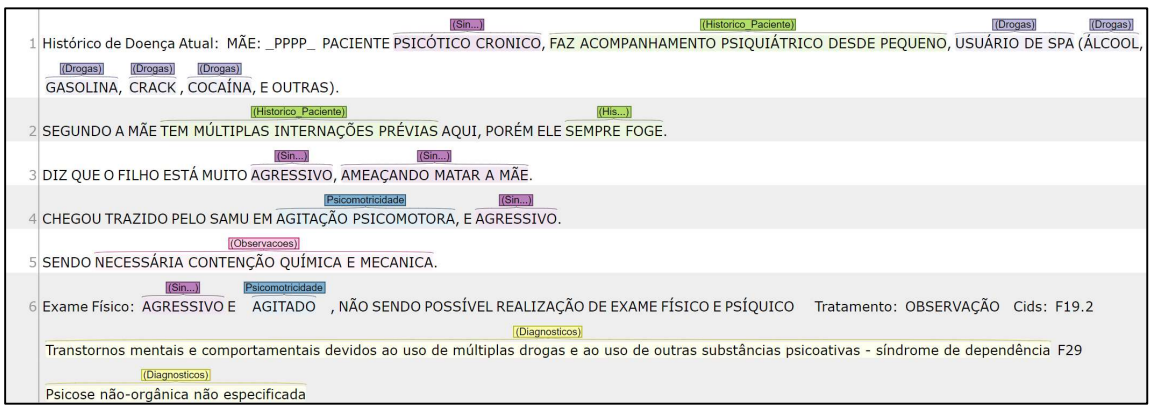
4.1.2. Anotação Semântica

A anotação semântica, é o processo que consiste em anotar algumas palavras presentes no texto de acordo com a função semântica que elas estão desempenhando. Originalmente, as anotações de EN consistiam apenas nos modelos genéricos: pessoas, organizações, locais, entre outros.

Após verificado o sucesso na extração de EN genéricas e a possibilidade de adaptar as categorias anotadas para diversas finalidades, vários domínios de conhecimento começaram a explorar as técnicas de REN, quando surgiram então alguns tipos de informações pré-determinados para anotações destes domínios, como ontologias, dicionários e tesauros. Por conta disso, além da complexidade e conhecimento necessário em cada domínio, as anotações semânticas de determinados domínios costumam ser feitas por especialistas do domínio, como é o caso do domínio de textos médicos, onde é necessário ao menos um profissional da saúde na equipe de anotação, para guiar a compreensão de termos e jargões que comumente estão presentes nos textos (ROBERTS et al., 2009).

Além das anotações dos elementos presentes no texto, nas anotações semânticas é comum a anotação das relações entre esses elementos. Dois ou mais elementos semânticos podem se relacionar e essa relação ser um terceiro tipo semântico, que também contém uma informação a ser adicionada ao documento. Em anotações de prontuários médicos, estabelecer critérios para a anotação de relações entre as entidades é uma tarefa difícil e pode gerar confusão até mesmo para especialistas do domínio (ROBERTS et al., 2009). Na Figura 7 é possível observar um documento anotado semanticamente.

Figura 6 – Prontuário de Psiquiatria anotado semanticamente



Fonte: Anotado pelo autor na ferramenta INCEpTION.

4.2. Formas de Representação

Quando um corpus é anotado, as entidades precisam ser representadas computacionalmente de alguma forma. Existem vários formatos presentes na literatura para a representação das entidades dentro de um corpus, apresentados nesta subseção.

Representação IO

A representação IO (do inglês *Inside-Outside*) é a forma mais simples de representação. Ela possui apenas duas informações: uma para indicar se a palavra está contida em uma entidade anotada (I) e outra para indicar o contrário (O). Ou seja, se uma palavra do texto faz parte de uma entidade, será marcada como I, e se não faz, como O (PRADHAN, 2015). A maior limitação desta representação é que ela não é capaz de diferenciar entidades consecutivas de um mesmo tipo (ALSHAMMARI; ALANAZI, 2020).

Representação IOB

A representação IOB (do inglês *Inside-Outside-Begin*) é uma forma levemente diferente da representação IO. A informação adicional nessa representação é o B que guarda a informação de todo token inicial de uma entidade (PRADHAN, 2015). Dessa forma, Entidades que possuem apenas um token terão *begin* igual *input* e *output*, ou seja, o *begin* tem serventia apenas quando a entidade possui mais de uma palavra. A representação IOB é considerada a forma de representação padrão por muitos pesquisadores e foi adotada na CoNLL (do inglês *Conference on Computational Natural Language Learning*) (SANG; BUCHHOLZ, 2000; JURAFSKY; MARTIN, 2023).

A representação IOB possui duas versões: a IOB ou IOB1 e a IOB2 (KRISHNAN; GANAPATHY, 2005). A diferença entre elas é que, na IOB1, a marcação 'B' é aplicada ao token que inicia um segmento que imediatamente sucede outro trecho da mesma entidade nomeada. Isso significa que, se houver duas entidades nomeadas consecutivas do mesmo tipo, o primeiro token da segunda entidade seria marcado como 'B', é marcado como ‘B’ o token que está no início de um trecho que está imediatamente após outro trecho da mesma entidade nomeada, enquanto na IOB2, é marcado como ‘B’ o token que está na posição inicial de qualquer EN (KRISHNAN; GANAPATHY, 2005).

Representação IOE

A representação IOE (do inglês *Inside-Outside-End*) é muito semelhante à IOB. A única diferença entre as duas é que, enquanto o parâmetro B da IOB marca se o token é o início da entidade, o token E da IOE marca se o token é o fim da entidade (PRADHAN, 2015).

Assim como a representação IOB, a representação IOE também possui duas versões: a IOE ou IOE1 e a IOE2 (KRISHNAN; GANAPATHY, 2005). A diferença entre elas é que, na IOE1, é marcado como ‘E’ o token que está no final de um trecho que está imediatamente antes de outro trecho da mesma entidade nomeada, enquanto na IOE2, é marcado como ‘E’ o token que está na posição final de qualquer EN (KRISHNAN; GANAPATHY, 2005).

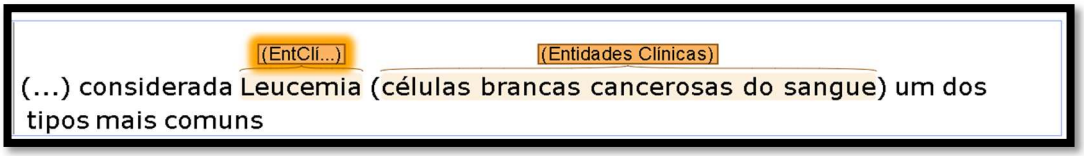
Representação IOBES

A representação IOBES (do inglês *Inside-Outside-Begin-End-Single*) é a união das representações IOB com a IOE com a adição de outra informação: o *single* (S), que marca entidades que são compostas por tokens únicos (ALSHAMMARI; ALANAZI, 2020).

Exemplos das Diferenças entre as Representações

Para deixar mais clara a diferença entre os padrões de representação apresentados nesta seção, um exemplo apresentado na Figura 8 e na Tabela 2 permitem observar a diferença entre os padrões de representação citados.

Figura 7 – Documento exemplo com anotações



Fonte: Texto traduzido de Alshammari e Slanazi (2020) e anotado pelo autor na ferramenta INCEpTION.

Tabela 2 – Tokens e suas representações

Palavra/*token*	**IO**	**IOB2**	**IOE2**	**IOBES**
considerada	O	O	O	O
Leucemia	I	B	E	S
(	O	O	O	O
células	I	B	I	B
brancas	I	I	I	I
cancerosas	I	I	I	I
do	I	I	I	I
sangue	I	I	E	E
)	O	O	O	O
um	O	O	O	O
dos	O	O	O	O
tipos	O	O	O	O
mais	O	O	O	O
comuns	O	O	O	O

Embora vários tipos de representação de anotação tenham sido utilizados na literatura, a escolha do padrão ideal de anotação para o projeto é um problema complexo e costuma ser feito de forma arbitrária (KONKOL; KONOPÍK, 2015), além de haver poucos estudos sobre o impacto de diferentes tipos de anotações para a qualidade do REN, até mesmo para o idioma inglês (ALSHAMMARI; SLANAZI, 2020).

Konkol e Konopík (2015) realizaram um experimento onde testaram diversos tipos de representações para a tarefa de REN para os idiomas tcheco, holandês, inglês e espanhol. Como conclusão, as representações IOE1 e IOE2 tiveram resultado superior que outras representações, porém os autores concluíram o estudo dizendo que não há diretrizes muito claras a serem seguidas para a escolha da representação para uma tarefa de anotação (KONKOL; KONOPÍK, 2015).

Alshammari e Slanizi (2020) realizaram um experimento semelhante à Konkol e Knopík (2015), porém para o idioma arábico. Nesse experimento, os autores anotaram um corpus composto por 27 artigos da área médica com os padrões IO, IOB, IOE, IOBES, BI, IE e BIES. Os autores utilizaram esses dados anotados para realizar a tarefa de REN com vários algoritmos de classificação e verificar os resultados obtidos. O padrão de anotação com melhor performance foi o IO, com F-score médio de 0.84. Os autores concluíram que um dos motivos desse resultado foi uma particularidade do corpus, que possuía poucas

entidades consecutivas (ALSHAMMARI; SLANIZI, 2020), corroborando para a ideia de Konkol e Konopík (2015), que a escolha do padrão pode variar de acordo com o corpus.

4.3. O Processo de anotação

Anotações de corpus são trabalho custosos, e geralmente anotações são criadas para finalidades específicas (HOVY; LAVID, 2010). As anotações de corpus podem ser feitas por pessoas, por algoritmos, ou de forma mista, a depender de vários fatores, como o tamanho do corpus, os dados que se pretende anotar, a existência de modelos de anotação pré-existentes com alta aceitação, entre outros. A depender do domínio, pode ser que a anotação seja realizada apenas por especialistas do domínio (ROBERTS et al., 2009).

De acordo com Hovy e Lavid (2010), um processo de anotação de um corpus realizada exclusivamente por anotadores humanos pode ser dividido em algumas etapas, como as seguintes:

1. Obtenção e preparação do corpus;
2. Definição do que será anotado e treinamento;
3. Anotação em caráter de treinamento, cálculo e determinação do índice de aceitação entre os anotadores e verificação de ajustes;
4. Anotação do corpus.

4.3.1. Obtenção e Preparação do Corpus

O corpus trata-se do conjunto de textos computacionalmente processáveis que serão utilizados para a realização da tarefa de anotação (HOVY; LAVID, 2010). A obtenção do corpus é uma tarefa cuja dificuldade pode variar de acordo com a natureza dos dados, domínio da aplicação e legislação do país em que o estudo está sendo conduzido. Durante a separação dos documentos que farão parte do corpus é necessário que seja considerado o potencial de representatividade do corpus para o subdomínio em questão. O ideal é que os fenômenos encontrados no corpus sejam capazes de representar os fenômenos encontrados no universo textual daquele domínio (HOVY; LAVID, 2010)

Uma vez o corpus estando preparado para a tarefa, é necessário a utilização de ferramentas que auxiliem no processo de anotação. Essas ferramentas devem possibilitar à equipe de anotação todos os recursos necessários para a execução das anotações, e muitas vezes trazem recursos extras que permitem gerenciar a tarefa de anotação como um projeto, como área integrada para comunicação entre os anotadores, *workflows*, processos de revisão automatizados, entre outros.

Há várias ferramentas disponíveis para a anotação de corpus. Andrade, Oliveira e Moro (2016) fizeram um levantamento das principais ferramentas de anotação disponíveis e suas principais características. No levantamento, a ferramenta chamada WebAnno[3] foi apontada como a mais adequada para o processo proposto no trabalho, por conter características importantes como extensa documentação, interface amigável e funcional,

3 WebAnno: https://webanno.github.io/webanno/

integração com ontologias, suporte à língua portuguesa, entre outros (ANDRADE; OLIVEIRA; MORO, 2016). Em 2018, o WebAnno deu lugar ao seu sucessor, o INCEpTION[4] (KLIE et al., 2018).

4.3.2. Definição das Anotações e Treinamento

Toda anotação é fundamentada por alguma teoria. A teoria provê a base para a criação e definição das categorias de anotação e para o desenvolvimento das diretrizes que deverão ser seguidas na anotação (HOVY; LAVID, 2010). A definição das categorias e das diretrizes da anotação é uma tarefa fundamental para o processo, porém não é trivial e sua complexidade é diretamente proporcional à complexidade do fenômeno que será anotado (HOVY; LAVID, 2010).

Entre as questões que são discutidas nas definições de anotação, o nível de detalhamento das categorias que serão anotadas é uma delas. Deve haver um meio termo entre o nível de sofisticação desejado das categorias e a possibilidade de seguir a sofisticação mantendo a coerência entre a equipe de anotadores. A solução mais comumente anotada pela comunidade de NLP é realizar a anotação com níveis não tão especializados de categorização (HOVY; LAVID, 2010).

Uma das partes fundamentais da definição das anotações é a criação do *guideline* de anotações. O *guideline* trata-se de um documento que reúne as definições das anotações e exemplos das categorias anotadas. É comum que o *guideline* não esteja pronto no início do processo de anotação, de forma que o mesmo vá sofrendo alguns ajustes durante a anotação do corpus, sendo necessário assim que os anotadores façam revisões periódicas (HOVY; LAVID, 2010).

O treinamento inicial dos anotadores consiste na apresentação da fundamentação utilizada como guia no processo de criação do *guideline*, estudo do *guideline* e anotações de exemplos iniciais, além da apresentação da ferramenta de anotação, métodos de comunicação com a equipe, entre outras informações importantes.

4.3.3 Avaliação das Anotações

É comum que a anotação do corpus se inicie por uma etapa de treinamento. Nessa etapa é realizada a anotação de uma pequena parte do corpus pelos anotadores. Essa etapa é importante porque é neste momento em que é calculado o índice de aceitação entre os anotadores (IAA) e são feitos os ajustes iniciais no *guideline*.

O IAA, que pode ser calculado pela equação abaixo, é uma medida que tem por objetivo demonstrar a concordância entre os anotadores para as mesmas entidades. Sendo assim, para que o IAA seja calculado, é necessário que ao menos dois anotadores anotem o mesmo texto. IAAs baixos mesmo após uma quantia razoável de treinamento podem indicar que o *guideline* está mal definido ou que a tarefa de anotação é difícil, e então devem ser realizados ajustes (HOVY; LAVID, 2010).

4INCEpTION: https://inception-project.github.io/

$$\text{IAA} = \frac{Acertos}{Acertos + Erros}$$

É importante que, ao longo da tarefa de anotação, o IAA seja calculado periodicamente, com o objetivo de verificar se as anotações estão mantendo determinada padronização. Variações do IAA ao longo do projeto podem significar problemas da equipe de anotação e requerem intervenções como reuniões para discussão ou melhorias no *guideline*.

Na etapa de treinamento também é comum que os anotadores discordem a respeito de como devem ser anotadas determinadas categorias de entidades, por isso é importante a realização de ajustes no *guideline* a partir das discussões feitas pelos próprios anotadores, aproximando o *guideline* do problema real encontrado por eles durante as anotações de treinamento. Dessa forma, busca-se que as dúvidas e discordâncias sejam resolvidas antes do início real da anotação do corpus. Para projetos de anotações complexas, onde o IAA é baixo, é comum a figura de um adjudicador: especialista responsável por tomar a decisão de qual a anotação mais correta quando duas anotações divergem, melhorando assim a concordância das anotações (HOVY; LAVID, 2010).

É comum que as anotações em caráter de treinamento ocorram juntamente com a verificação de IAA até um determinado período após a finalização das revisões sugeridas no *guideline* (caso houver), para que garanta-se que todos os anotadores iniciarão a anotação com as mesmas definições sobre a tarefa.

4.3.4. Anotação do Corpus

Para a anotação do corpus propriamente dita, os documentos utilizados durante as etapas de treinamento devem ser reconsiderados e distribuídos novamente a fim de não fragmentar o corpus.

Cada documento do corpus pode ser anotado por um único anotador (anotação simples) ou por dois ou mais (anotação dupla ou múltipla). A anotação dupla tem por objetivo possibilitar o cálculo do IAA nos documentos e pode ser aplicada em todo o corpus ou apenas em partes dele.

Em corpus onde realiza-se anotação simples, a anotação múltipla é aplicada em determinada percentagem das notas, periodicamente, com a única finalidade de medir o IAA entre os anotadores, que deve ocorrer várias vezes ao longo da etapa. Caso o IAA ultrapasse o limite mínimo definido para o projeto, é necessária uma intervenção.

Em corpus onde realiza-se anotação dupla ou múltipla, além do cálculo da IAA é necessário que um adjucador faça as devidas correções das inconformidades, para que os mesmos termos não sejam anotados de formas diferentes. A vantagem de um corpus completamente anotado desta forma é que se diminui o risco de inconsistência entre as anotações e também se obtém o IAA de forma síncrona, porém a desvantagem é que o custo humano envolvido no processo de anotação também é maior.

4.4. O Padrão Ouro

O termo padrão ouro refere-se as anotações que são realizadas de forma criteriosa afim de atingir com melhor resultado possível o objetivo da anotação daquele corpus (KILGARRIFF, 1998) e pode estar associado a um corpus completo ou apenas a um subconjunto dele (ROBERTS et al., 2009).

Por conta do crédito dado aos corpora ou parte de corpus padrão ouro, é comum que os algoritmos de aprendizado de máquina sejam treinados com esses corpora, afim de atingirem melhor performance.

De acordo com Kilgarriff (1998), para que um corpus seja considerado padrão ouro, sua anotação deve ser realizada de acordo com as seguintes regras:

- Os documentos devem ser, no mínimo, duplamente anotados
- Deve ser realizado o cálculo do IAA
- O IAA atingido deve ser alto o suficiente para o domínio.

De acordo com Roberts et al. (2009), atingir valores altos de IAA para determinados domínios é uma tarefa muito difícil, e em domínios como registros clínicos valores de IAA menores que 0.8 são aceitáveis para corpus padrão ouro.

4.5. Considerações Finais

Nesta seção foi apresentado o processo de criação e anotação de um corpus, os requisitos e os processos envolvidos nas etapas de anotação. Pode-se verificar que a anotação está relacionada com a tarefa específica que se deseja relacionar e ao seu domínio, e que os *guidelines* podem ser mais ou menos complexos de acordo com tais.

Um outro aspecto apresentado foi o alto custo humano que envolve projetos de anotação que utilizam essencialmente a anotação humana, já que, além de uma tarefa que requer atenção e volume, muitas das vezes também precisa ser feita por especialistas do domínio, o que encarece ainda mais o projeto.

Não foram abordadas nesta seção as anotações automatizadas ou semiautomatizadas, embora estas sejam comumente empregadas na criação de grandes corpora. O autor tomou a decisão de não abordar esses processos porque eles não foram utilizados nesta dissertação, porém tais processos podem ser encontrados em Hovy e Lavid (2010).

No próximo capítulo será exibido um amplo levantamento de trabalhos relacionados à presente dissertação, onde ficará ainda mais clara a importância e a escassez dos corpus anotados para o domínio de prontuários médicos na língua portuguesa.

Capítulo 5 - Trabalhos Relacionados

Neste capítulo são apresentados os trabalhos que estão relacionados com os objetivos deste trabalho. Para a apresentação, foi realizada uma extensa busca e levantamento do estado da arte em repositórios científicos em inglês e português. Os trabalhos relacionados estão divididos em duas categorias:

- Anotações de corpora no domínio da saúde.
- Reconhecimento de Entidades Nomeadas em Corpus de dados clínicos

5.1. Levantamento dos corpora anotados no domínio da saúde.

Nesta seção é são apresentados os corpora mais relevantes encontrados para o domínio de registros clínicos extraídos dos prontuários dos pacientes. No levantamento resumido na Tabela 3, foram considerados apenas dois idiomas: o inglês, representado na tabela pela sigla 'EN', e o português, que é apresentado como 'PT-EU' caso seja o português de Portugal, ou por 'PT-BR' caso seja o português do Brasil.

Em relação aos corpora no idioma inglês, é importante informar que há muitos outros corpora existentes que não estão presentes nesta tabela pois não fazem parte do recorte analítico da presente dissertação. Por outro lado, foi realizado um vasto levantamento dos corpora existentes para o idioma português, e estão listados na Tabela 3 todos os encontrados.

Tabela 3 – Corpus anotados de prontuários médicos

Corpus	**Fonte**	***Pos tagging***	**Anotação Semântica**	**Público**	**IAAA**	**Tamanho (nro. docs)**	**Idioma**
I2b2 2008	i2b2, 2018	Não	Sim	Sim	0,75	1.237	EN
MIMIC-III	Johnson et al., 2016	Não	Sim	Sim	-	61.293	EN
CER	Patel et al., 2018	Sim	Sim	Sim	0,96	5.160	EN
CLEF	Roberts et al., 2009	Não	Sim	Sim	0,73	20.234	EN
MedAlert	Ferreira; Teixeira; da Silva Cunha, 2010	Sim	Sim	Não	-	914	PT-EU
SemClinBr	Oliveira et al., 2020	Não	Sim	Não	0,77 0,92	1.000	PT-BR

ClinPt	Lopes; Teixeira; Oliveira, 2019	Sim	Sim	Sim	0,957	281	PT-EU
Não publicado	Da Rocha, 2021	Não	Sim	Não	-	1.200	PT-BR
Não publicado	Niero, 2022	Sim	Sim	Não	0,61	1.350	PT-BR

5.1.1. Descrição dos corpora

Quando um corpus é produzido, as anotações realizadas sobre os dados textuais costumam ter um determinado objetivo pré-definido, de acordo com o estudo que está sendo realizado durante a produção daquele corpus (HUANG et al., 2020). A seguir estão as descrições dos corpora que foram resumidos na Tabela 3.

I2b2 2008

O corpus I2b2 em sua versão de 2008 consiste de um conjunto composto de 1.237 resumos de altas de pacientes portadores de diabetes e obesidade (UZUNER, 2009). Diferentemente dos demais corpus, o índice de aceitação deste corpus foi medido pelo índice de aceitação Kappa (FRIEDMAN; HRIPCSAK; SHABLINSKY, 1998) e para as anotações textuais teve sua média de 0,755. Os documentos possuem dois diferentes tipos de anotações: textuais e intuitivas. As anotações textuais foram classificadas entre 15 categorias semânticas, sendo todas elas doenças associadas à obesidade:

- Asma
- Doença arterial coronária
- Insuficiência cardíaca congestiva
- Depressão
- Diabetes melito
- Cálculos biliares
- Doença do refluxo gastroesofágico
- Gota
- Hipercolesterolemia
- Hipertensão
- Hipertrigliceridemia
- Obesidade
- Apneia obstrutiva do sono
- Osteoartrite
- Doença vascular periférica
- Insuficiência venosa

Os autores então utilizaram as anotações para realizar uma classificação secundária na fatualidade dessas doenças. A classificação foi feita da seguinte forma:

- Presente: Indica que o paciente possui ou já possuiu a doença.

- Ausente: Indica que o paciente não possui e nem possuiu a doença.
- Questionável: Indica que o paciente pode ter a doença, porém há um grau de dúvida.
- Não mencionado: Não há informações sobre a doença no resumo de alta.

MIMIC-III

MIMIC é acrônimo de Centro de Informação Médica para Terapia Intensiva (do inglês *Medical Information Mart for Intensive Cares*), em sua terceira versão, o MIMIC-III (JOHNSON et al., 2016) foi gerado a partir de registros de um hospital no *Massachusetts*. O corpus é composto por dados de 53.423 admissões de pacientes adultos em leitos de unidades de tratamento intensivo (UTI) entre 2001 e 2012, 7.870 admissões de UTI neonatais. Cada prontuário possui uma média de 4.579 observações e 380 medidas laboratoriais. Os dados presentes no MIMIC-III estão divididos em nove categorias semânticas:

- Faturamento
- Descritivos
- Dicionários (como CID-10)
- Intervenções
- Laboratoriais
- Medicamentos
- Notas
- Fisiológicos
- Laudos

CER

CER é acrônimo de Reconhecimento de Entidades Clínicas (do inglês *Clinical Entity Recognition*). Esse corpus e foi construído a partir de 236.850 documentos de registros clínicos cedidos por 119 provedores (hospitais e clínicas) e que contém registros de médicos de quarenta diferentes especialidades. Desses registros, foi feito um trabalho para a criação de um corpus que fosse representativo, porém de menor tamanho, e então foram anotados 5.160 documentos (PATEL et al., 2018). Esses registros deram origem a onze categorias semânticas:

- Problemas
- Descobertas
- Procedimentos
- Estruturas Anatômicas
- Funções Corporais
- Dados laboratoriais
- Medidas Corporais
- Valores de Medidas
- Dispositivos Médicos
- Drogas
- Modificadores

Além disso, no CER também foi realizada a POS *tagging*, porém essa tarefa não foi realizada pela equipe de anotação, mas sim por uma ferramenta de anotação automática (PATEL et al., 2018).

CLEF

CLEF é acrônimo de Framework de Ciência Eletrônica Clínica (do inglês *Clinical E-Science Framework*). Esse corpus foi construído a partir de 565.000 documentos, dos quais foram anotados 20.234, compostos por registros clínicos, laudos e imagens que fazem parte dos registros de pacientes com câncer (ROBERTS et al., 2009).

Foram anotadas entidades e relações entre as entidades, e ambos tiveram os tipos extraídos do UMLS. Ao final da anotação, utilizou-se uma ferramenta para mapear cada entidade anotada para um conceito único UMLS, tornando o CLEF um corpus muito rico de informações clínicas.

As categorias semânticas anotadas no CLEF foram:

- Condições
- Intervenções
- Investigações
- Resultados
- Drogas ou Dispositivos
- Locais
- Sinais de negação
- Sinais de lateralidade
- Sinais de sub-localidades

As relações anotadas entre as entidades foram:

- Tem alvo
- Tem um achado
- Tem uma indicação
- Tem uma local
- Modificadores (negação, lateralidade e sub-local)
- Correferências.

Em Roberts et al. (2009), os autores detalharam todo o processo de anotação e da criação do CLEF, incluindo as dificuldades encontradas e as razões por terem tomado determinadas decisões, o que torna esse trabalho uma ótima fonte de conhecimento para anotação de corpus de registros clínicos.

MedAlert Corpus

O corpus do MedAlert foi construído a partir de anotações de resumos de alta de pacientes admitidos com hipertensão em um hospital em Portugal e foi o primeiro corpus de registros clínicos em língua portuguesa. O corpus consiste em um conjunto composto de 914 resumos de altas, que totalizam 51.695 *tokens* e 2.989 sentenças anotadas. Os dados podem estar relacionados às seguintes informações sobre os pacientes:

- Motivo da admissão
- Histórico clínico

- Exame físico
- Evolução
- Terapêutica
- Encaminhamento

O processo de anotação utilizou como padrão o UMLS em sua versão particular para a língua portuguesa, os Descritores em Ciências da Saúde (DeCS) e o CID-9. Devido ao alto custo envolvido nas anotações de dados clínicos, apenas 10 por cento do total de 900 documentos foram duplamente anotados, e, portanto, apenas 90 documentos geraram um padrão Ouro, e os outros 810 não geraram um padrão de comparação.

Embora o MedAlert não seja um corpus público e por isso seja pouco relevante para o ambiente acadêmico, ele está na tabela pelo fato de ser pioneiro para a língua portuguesa.

SemClinBr

O corpus SemClinBr foi construído a partir de 2.094.929 registros de um grupo de hospitais obtidos entre 2013 e 2018 e de 5.617 registros obtidos em um hospital universitário entre 2002 e 2007 e é o primeiro e único corpus de registros clínicos anotados em português do Brasil (OLIVEIRA et al., 2020).

Os registros dizem respeito a aproximadamente 553.952 pacientes e estão relacionados à diversas especialidades, destacando-se as especialidades de cardiologia, nefrologia e ortopedia. Além disso, as anotações foram geradas com dois principais objetivos: (i) Dar suporte ao desenvolvimento de algoritmos de REN e (ii) validar um algoritmo de busca semântica.

A fim de atingir um padrão ouro, a anotação seguiu rígido processo de homogeneidade e as entidades foram anotadas de acordo com os grupos e tipos semânticos do UMLS e deram origem a um total de 1.000 documentos duplamente anotados, sendo destes 613 documentos considerados padrão ouro e 387 considerados padrão platina. Dos 1.000 documentos, foram extraídas 65.129 entidades e 11.263 relações. As categorias semânticas anotadas neste corpus foram:

- Local ou região do corpo
- Parte do corpo, órgão ou componente do órgão
- Química orgânica
- Substância Farmacológica
- Conceito quantitativo
- Conceito qualitativo
- Conceito temporal
- Dispositivo médico
- Doença ou Síndrome
- Achado
- Prejuízo

O SemClinBr é o único corpus clínico anotado em PT-BR disponível publicamente no GitHub[5] do laboratório de inteligência artificial aplicado à saúde (HAILab, do inglês

[5] https://github.com/HAILab-PUCPR/SemClinBr

Health Artificial Intelligence Lab), que pertence à Pontifica Universidade Católica do Paraná (PUCPR).

ClinPt Corpus

O ClinPt corpus (LOPES; TEIXEIRA; OLIVEIRA, 2019) foi construído a partir de anotações de 281 documentos de casos clínicos retirados de publicações da sociedade portuguesa de Neurologia. Os casos clínicos, que se tratam de resumos dos pacientes, são um pouco dos prontuários médicos, porém ainda podem ser considerados como importantes devido à escassez de corpus de prontuários.

O ClinPt corpus foi pré-processado por uma ferramenta de PLN onde foi realizado POS tagging automático. Após isso, foram feitas as anotações semânticas para o domínio clínico. Não há informações de quais tipos de entidades foram utilizadas para a anotação, sendo necessário analisar todo o corpus para reunir essa informação, tarefa que não foi feita neste trabalho.

O ClinPt é o único corpus de dados de registros clínicos/casos clínicos disponível publicamente para a língua portuguesa. Tanto o corpus como um modelo pré-treinado do BERT com outros documentos da sociedade de portuguesa de Neurologia podem ser acessados na página do GitHub[6] do ClinPt.

Corpus de Rocha (2021)

O corpus criado no trabalho de Rocha (2021) não foi nomeado pela autora. Para a sua criação, a autora reuniu 30.000 documentos de prontuários médicos do Hospital das Clínicas de Botucatu, que trata especialidades diversas, e selecionou aleatoriamente 1.200 destes documentos para formar um corpus. Os documentos foram extraídos do campo "Evolução do paciente" presente nos prontuários dos pacientes.

A autora realizou a anotação semântica do corpus juntamente com estudantes do quarto ano de medicina, que criaram um corpus padrão ouro. As entidades do domínio foram anotadas nas seguintes categorias:

- Medicação
- Condição
- Tratamento
- Sintoma
- Exame
- Diagnóstico

A autora não forneceu valores de índice de aceitação e não informou se os documentos foram duplamente anotados, porém, como informou que o corpus gerado se trata de padrão ouro, subentende-se que o índice de aceitação e a anotação dupla fizeram parte do projeto. O corpus não está disponível publicamente.

[6] https://github.com/fabioacl/PortugueseClinicalNER

Corpus de Niero et al. (2022)

O corpus criado no trabalho de Niero et al. (2022)[7] também não foi nomeado pelos autores. Para a criação do mesmo, os autores reuniram 1.500 documentos de prontuários médicos de oncologia, sendo 350 da especialidade de câncer de mama e 1.150 da especialidade de câncer de próstata.

O corpus foi anotado semanticamente de forma simples (sem anotação dupla) por estudantes de medicina, de acordo com um *guideline* próprio dos autores, que foi baseado por sua vez no *guideline* ElilE, de Kang et al. (2021), de forma que as entidades foram anotadas nas seguintes categorias:

- Condições
- Medicações
- Medidas
- Localizações Anatômicas

Após a anotação semântica foi utilizado um anotador automático para a anotação sintática do corpus. A anotação sintática, porém, não foi avaliada em termos de qualidade.

Foi calculado o IAA apenas durante a etapa de treinamento dos anotadores, e o IAA atingido antes do início das anotações foi de 0,61. Os autores não disponibilizaram o corpus publicamente.

BRATECA

BRATECA (CONSOLI et al., 2022) é acrônimo de *Brazilian Tertiary Care Dataset* (Conjunto de dados terciários Brasileiros), e é o único corpus deste levantamento que não possui anotações. Apesar disso, o BRATECA destaca-se por ser o maior corpus de dados médicos da língua portuguesa, composto por 2,5 milhões de notas clínicas de pacientes admitidos em 10 diferentes hospitais do Brasil. Os documentos que compõe o BRATECA contêm informações sobre admissão, exames, notas clínicas e prescrições.

5.2. Metodologias de REN em corpus de registros clínicos

Além do levantamento dos corpora mencionados na seção anterior, foi feito um levantamento das abordagens de REN desenvolvidas e que utilizaram como fonte de dados os corpora. Seguindo a mesma abordagem, nem todos os trabalhos que realizaram REN nos corpora da língua inglesa foram incluídos na Tabela 4, fazendo parte dela apenas alguns dos mais relevantes, enquanto que, para os corpora da língua portuguesa, estão listados todos os trabalhos.

5.2.1. Tabela comparativa dos principais trabalhos encontrados

Nesta subseção estão duas tabelas com a comparação dos trabalhos que utilizaram como fonte de dados os corpora mencionados na seção 5.1. Na Tabela 4, é feita a

[7] Trata-se de outro trabalho do presente autor

comparação dos *scores* obtidos por cada um dos trabalhos. Na Tabela 5 são resumidos os métodos de aprendizado utilizado por eles.

Tabela 4 – Resultados dos trabalhos de REN que utilizaram os corpus apresentados na Tabela 3

Autores	Corpus	Precisão	Revocação	F-*score*
Neumann; Beltagy; Ammar, 2019	MedMentions	0,70	0,68	0,69
Reátegui; Ratté, 2018	I2b2 2008	0,89	0,88	0,88
Reátegui; Ratté, 2018	I2b2 2008	0,89	0,91	0,89
Hofer et al., 2018.	I2b2 2009	-	-	0,79
Ferreira; Teixeira; da Silva Cunha, 2010	MedAlert	0,95	-	-
Oliveira et al., 2020	SemClinBr	0,75	0,50	0,55
Singh et al., 2020	MIMIC-III	0,87	0,87	0,86
Lin et al., 2019	MIMIC-III	0,67	0,69	0,68
Alsentzer et al., 2019	I2b2 2006	-	-	0,95
Alsentzer et al., 2019	I2b2 2010	-	-	0,86
Alsentzer et al., 2019	I2b2 2012	-	-	0,79
Alsentzer et al., 2019	I2b2 2014	-	-	0,93
Ferreira; Teixeira; da Silva Cunha, 2010	MedAlert	0,95	-	-
de Souza et al.; 2019	SemClinBr	0,75	0,50	0,55
Patel et al., 2018	CER	0,92	0,91	0,92
Roberts et al., 2009	CLEF	0,81	0,63	0,71
Lopes; Teixeira; Oliveira, 2019	ClinPt	0,83	0,82	0,82
Rocha, 2021	Não publicado	0,73	-	0,64
Niero et al., 2022	Não publicado	-	-	0,63

*Para trabalhos que reportaram mais de um *score* para o mesmo corpus, foi considerado o maior valor

Tabela 5 – Metodologias de REN Utilizadas pelos Trabalhos que realizaram REN com os corpora da Tabela 3

Autores	Metodologia Utilizada
NEUMANN; BELTAGY; AMMAR, 2019	Scispacy: Stack-LSTM + parser de dependência baseado em transição
REÁTEGUI; RATTÉ, 2018	Metamap2015: Abordagem baseadas em regras
Reátegui; Ratté, 2018	Apache CTAKES v 3.2 e UIMA: Abordagem Híbrida: Mapeamento de dicionários
HOFER et al., 2018.	BERT *fine tunning* + *softmax*
SINGH et al., 2020	BERT *fine tunning* + *gelu*
LIN et al, 2019	BERT-MIMIC *fine tunning* + Camada de classificação não inforada
ALSENTZER et al., 2019	BERT *fine tunning* + Camada de classificação não informada
ALSENTZER et al., 2019	BioBert *fine tunning* + Camada de classificação não informada
ALSENTZER et al., 2019	ClinicalBert *fine tunning* + Camada de classificação não informada
FERREIRA; TEIXEIRA; DA SILVA CUNHA, 2010	Abordagem baseada em regras (ontologias + dicionários)
DE SOUZA et al., 2019	CRF
PATEL et al., 2018	Abordagem Híbrida: Regras, word2Vec, Dicionários + CRF
ROBERTS et al., 2009	GATE NLP: Abordagem Híbrida: Dicionários + SVM
LOPES; TEIXEIRA; OLIVEIRA, 2019	BiLSTM + CRF
DA ROCHA, 2021	Scispacy: Stack-LSTM + Parser de dependência baseado em transição

5.2.2. Descrição dos Trabalhos Relacionados

Nesta sessão estão descritos resumos sobre os trabalhos que foram introduzidos na Tabela 4.

Neumann; Beltagy; Ammar (2019)

Neste trabalho os autores apresentam uma nova biblioteca em Python para PLN de textos biomédicos e clínicos: o ScispaCy, que é baseado na spaCy. Na ScispaCy, os autores retreinaram o modelo de spaCy para POS *tagging*, análise de dependência, e REN utilizando corpus com textos biomédicos, além de melhorar o módulo de tokenização com regras adicionais. ScispaCy contém dois módulos: *en_core_sci_sm* e *en_core_sci_md*. Os módulos variam pelo tamanho do vocabulário, sendo o segundo deles mais completo e contendo *word vectors*.

Para o *POS tagging* e análise de dependência foi utilizado um arco baseado em transição, semelhante a Goldberg e Nivre (2012). Para a análise de dependências, foi utilizado o *treebank* de McClosky e Charniak (2008), que é baseado no corpus GENIA 1.0 (KIM et al., 2003). Também foram utilizados os artigos originais da PubMed que estavam relacionados aos resumos utilizados no GENIA 1.0. Posteriormente, o POS *tagging* e parser de dependência também foi treinado utilizando o OntoNotes 5.0.

O modelo de *embedding* utilizado tanto no spaCy como no ScispaCy é um modelo de *Chunking* baseado em transição (Stack LSTM + algoritmo de *Chunking*), que foi inspirado no modelo proposto por Lample et al. (2016). Para a associação de entidades, o ScispaCy utiliza as seções 0, 1, 2 e 9 do SNOMED do UMLS (BODENREIDER, 2004) versão AA de 2017. Esta versão contém 2.780 conceitos únicos e cobre 99% dos conceitos mencionados no corpus MedMentions, utilizado no trabalho (MURTY et al., 2018).

Por fim, os autores notaram que as abreviações eram responsáveis por grande parte da falha na busca dos candidatos a entidades associadas. Para resolver esse problema, foi implementado o algoritmo não supervisionado de detecção de abreviações, proposto por Schwartz e Hearst (2002).

O autor concluiu que o modelo utilizado pelo scyspaCy é rápido, de fácil utilização, escalável, e alcança desempenho próximo ao estado da arte, apresentando como resultados precisão de 0,70, revocação 0,68 e f-score de 0,69 para o treinamento realizado com o corpus mais completo, e resultados um pouco inferiores de precisão 0,70, revocação 0,67 e f-score de 0,68 para o treinamento realizado com o corpus reduzido.

Reátegui; Ratté (2018)

Neste trabalho, os autores comparam a ferramenta MetaMap (ARONSON; LANG, 2010), em sua versão de 2015 a ferramenta Apache cTakes (SAVOVA et al., 2010) em sua versão 3.2.

O sistema de análise de texto clínico e extração de conhecimento (cTAKES) combina técnicas de aprendizado de máquina e de aprendizado baseado em regras para extrair informações de textos clínicos. O cTAKES utiliza como forma de representação do conhecimento abordagens híbridas (SAVOVA et al., 2010). A parte da representação

baseada em regras utiliza mapeamento de dicionários, enquanto que a parte da representação que envolve o uso de métodos de aprendizado de máquina e foi implementada utilizando o framework Apache UIMA (do inglês *Unstructured Information Management Architecture)* (FERRUCCI; LALLY, 2004).

No trabalho de Neumann; Beltagy e Ammar, 2019, os autores utilizaram a ferramenta MetaMap e apache cTAKES para verificar a sua eficiência em relacionar as entidades extraídas do corpus I2b2 versão de 2008 com os respectivos conceitos únicos de identificação na UMLS.

O experimento foi dividido em duas partes complementares e como resultado final as ferramentas conseguem atingir os resultados de: precisão de 0,89, revocação de 0,88 e F-score de 0,88 para o MetaMap e precisão de 0,89, revocação de 0,91 e F-score de 0,89 para o cTakes.

Observa-se que o cTAKES obteve resultados levemente superiores ao MetaMap, porém o autor conclui o trabalho dizendo acreditar que a combinação entre as duas ferramentas deve gerar melhores resultados que os obtidos separadamente.

Hofer et al. (2018)

Neste trabalho, os autores apresentam como objetivo principal demonstrar que a quantidade de documentos de um corpus não é o fator mais relevante para obter bons resultados no processo de treinamento de um algoritmo de REN para dados clínicos.

O treinamento do algoritmo foi feito com a utilização de seis corpus, da seguinte forma: O corpus i2b2 2009, por ser o corpus utilizado para os testes, foi utilizado para o treinamento supervisionado. Os corpora i2b2 2010 e i2b2 2012 foram utilizados para o que o autor chamou de pré-treinamento de pesos. Os corpora BioNLP-2016, MIMIC-III e UK-CRIS foram utilizados para o treinamento não supervisionado de *word embeddings*. Por fim, o autor utilizou um corpus não médico para o que chamou de pré-treinamento supervisionado dos pesos. No corpus utilizado para treinamento supervisionado, i2b2 2009, foram utilizados apenas 10 documentos obtidos aleatoriamente do conjunto.

A arquitetura utilizada pelo autor foi baseado no proposto em Chiu e Nichols (2016). O modelo usa cada palavra em três níveis distintos: Nível de caractere, que dá origem aos *embedding* de caracteres; nível de palavra, que dá origem aos *embeddings* de palavras; e caixa alta e caixa baixa, gerando os *embeddings* de caixa alta/baixa (do inglês *casing embedding*), cada um dos quais codifica um aspecto diferente do texto. Depois de gerados os *embeddings*, estes passam por uma LSTM bidirecional e são classificados através da função softmax.

Como resultado, o autor alcançou F-score de 0,79, o que ele mesmo conclui que está muito abaixo dos *scores* obtidos pelos treinamentos que envolvem grandes conjuntos de dados, porém o objetivo do trabalho era demonstrar maneiras de obter bons resultados quando a quantidade de documentos disponíveis era muito pequena (apenas 10), e por isso os resultados obtidos podem ser considerados conservadores.

Alsentzer et al. (2019)

Neste trabalho, o objetivo foi criar e publicar um novo modelo de *embedding* baseado no BERT, porém com foco no *embedding* de registros médicos: o ClinicalBERT.

Os autores utilizam a mesma arquitetura do BERT (DEVLIN et al, 2018), porém o pré-treinamento do modelo foi feito com outras taxas de aprendizagem, que estão disponíveis no apêndice B do trabalho, além de um *fine tunning* adicional para a tarefa de REN.

Para realizar o treinamento do modelo proposto e dos demais modelos que viriam a ser comparados, os autores utilizaram o corpus MIMIC-III. Os autores não forneceram informações sobre qual foi o algoritmo utilizado para a classificação do NER.

Como método de validação, os autores executaram a tarefa de REN para os corpus i2b2 versões 2006, 2010, 2012 e 2014. Como resultado, os autores obtiveram F-scores de 0,92, 0,86, 0,79 e 0,93, respectivamente, enquanto os F-scores do BERT foram de 0,94, 0,84, 0,76 e 0,93 e os do BioBERT de 0,95, 0,87, 0,79 e 0,93, mostrando que o ClinicalBert não superou o modelo de *embedding* pré existente BioBERT.

Singh et al. (2020)

Neste trabalho, o objetivo foi identificar os diagnósticos e códigos de procedimentos a partir da aplicação de técnicas de PLN em registros de prontuários médicos.

Para o experimento foi utilizada uma arquitetura de BERT com conectada a uma camada de ativação com a função GELU (do inglês *Gaussian Error Linear Unit*). Os dados utilizados para o NER foram o corpus do MIMIC-III.

Como método de validação, os autores executaram a tarefa de REN também com os dados do MIMIC III, onde o corpus foi dividido em dois subconjuntos: top-10, contendo os 10 CIDs e códigos de procedimentos mais frequentes, e top-50, contendo os 50 CIDs e códigos de procedimentos mais frequentes. Como resultados, os autores obtiveram precisão de 0,8734, revocação de 0,8674 e f-score de 0,8582 para o conjunto top-10 e precisão de 0,93, revocação de 0,94 e f-score de 0,92 para o conjunto top-50.

Como conclusão, os autores constataram que conseguiram atingir resultados melhores que os trabalhos anteriores, elevando o estado da arte aos resultados atingidos.

Lin et al. (2019)

Neste trabalho, o objetivo foi criar um modelo baseado no BERT para extrair relações temporais de registros clínicos estejam elas dentro ou fora das frases. Embora as publicações dos últimos anos estivessem apontando que a extração de relações temporais de registros clínicos já apresentavam resultados expressivos, os autores acharam interessante testar o mesmo no BERT, já que este estava avançando o estado da arte para vários domínios de aplicação e que, além disso, o BERT é capaz de reconhecer as relações presentes dentro da sentença mas também a forma com que sentenças se relacionam, dada a sua arquitetura, possibilitando a criação de um modelo que não dependia do limite da frase, diferente das demais abordagens exploradas até o momento.

Foi utilizado BERT para realizar a adaptação do domínio utilizando todos os registros clínicos do corpus MIMIC-III, contendo 879 milhões de palavras relacionadas a registros de prontuários médicos. Esse pré-treinamento resultou em um BERT especialista no domínio de prontuários, que os autores chamaram de BERT-MIMIC. Depois os autores fizeram o *fine tunning* do BERT-MIMIC, BioBERT e BERT, para a tarefa de REN,

utilizando como auxílio um método baseado em janela de palavras do conjunto de dados provindos de um corpus chamado THYME, que contém anotações temporais em prontuários médicos (STYLER et al., 2014). Os autores não informaram qual foi o algoritmo utilizado na camada de classificação para o *fine tunning* de REN.

Como método de validação os autores fizeram as comparações de relações temporais em dados de câncer de cólon e câncer de cérebro para os três modelos e chegaram a um novo estado da arte para o BioBERT, atingindo F-score de 0,68 para o teste com os dados de câncer de cólon e de 0,57 para o teste com os dados de câncer de cérebro.

Como conclusão os autores disseram que o BERT-MIMIC ainda pode ser melhorado, porém são necessários mais dados para o pré-treinamento já que a quantidade de palavras presentes no corpus do MIMIC-III, com menos de 1 bilhão de palavras é muito inferior ao do BioBERT, por exemplo, que possui 18 bilhões de palavras.

Patel et al. (2018)

O principal objetivo neste trabalho era a criação do corpus CER, composto por 5.160 documentos de quarenta diferentes especialidades que, anotados, geraram 398.568 entidades. Ao final da tarefa realizaram o REN apenas como forma de validar o corpus. Embora os parâmetros utilizados para a tarefa de REN não estejam detalhados com riqueza, foi utilizado este trabalho para a tarefa de REN para o corpus do CER porque não foram encontrados outros trabalhos que tivessem realizado REN com o corpus criado por Patel et al. (2018).

Os autores informam que utilizaram algumas técnicas para realizar o REN: bag of words, prefixos, sufixos, POS *tagging*, inícios de frases, identificação de letras maiúsculas, dicionários e word2vec. Por fim, utilizaram o algoritmo CRF para a classificação das entidades.

Como resultado, os autores obtiveram 0,92 de precisão, 0,91 de revocação e 0,92 de F-score, valores consideravelmente bons considerando as técnicas utilizadas para as representaçõces.

Roberts et al. (2009)

O principal objetivo neste trabalho era a criação do corpus CLEF, composto por 20.234 documentos de pacientes que realizaram tratamento em um hospital especializado de câncer. Os 20.234 documentos foram extraídos de uma coleção de um total de 565.000 documentos, que estavam disponíveis para a seleção do corpus.

Os autores no final da tarefa de anotação realizaram a tarefa de REN, porém não deram muitas informações do processo utilizado, exceto que foi utilizado para a classificação o algoritmo de SVM implementado na ferramenta GATE, juntamente com algumas técnicas incrementais como *POS TAGGING* e dicionários. Como resultado, os autores obtiveram 0,81 de precisão, 0,63 de revocação e 0,71 de F-score.

Assim como em Patel et al. (2018), o trabalho de Roberts et al. (2009) não é muito citado pela tarefa de aplicação de REN, porém é muito citado na literatura pela sua contribuição às técnicas de anotação de corpus de registros clínicos.

Ferreira; Teixeira; da Silva Cunha (2010)

Neste trabalho foi gerado o primeiro corpus anotado de prontuários médicos em Português da Europa e foi construída uma ferramenta de extração de dados de textos clínicos, nomeada MedAlert, além de desenvolverem um método de detecção de abreviações em anotações clínicas.

Os autores não forneceram muitas informações a respeito das etapas do processamento e também dos algoritmos utilizados para a representação vetorial, bastando apenas a informação de que a ferramenta é baseada no framework Apache UIMA (FERRUCCI; LALLY, 2004), sendo o NER especificamente realizado em um módulo acoplado chamado REMMIX, que utilizar fontes externas como ontologias e terminologias para extrair informações.

Os autores informaram também que processo de pré-processamento conta com as etapas de tokenização, POS tagging e desabreviações.

A ferramenta construída foi treinada a partir do corpus que os próprios autores construíram, nomeada de MedAlert corpus, e que continha registros de resumos de altas de pacientes de um hospital de Portugal. O corpus é composto por 51.595 tokens, 2.989 sentenças e 914 documentos.

Por conta da não existência de um padrão de dados clínicos para o Português da Europa, o método de validação precisou ser medido apenas em precisão, que foi feito por comparações de Verdadeiros e Falsos Positivos, onde atingiu-se uma precisão média de 0,95. Os autores consideraram suficientes as comparações booleanas como forma de medida para dados de prontuários médicos.

de Souza et al. (2019)

Neste trabalho é apresentado a primeira tarefa de REN em dados de prontuários médicos na língua portuguesa do Brasil. Os autores utilizaram um corpus contendo 1.000 prontuários obtidos de três diferentes hospitais, sendo as especialidades predominantes nos documentos do corpus a Nefrologia, a cardiologia e a Endocrinologia.

O conjunto de dados utilizado continha registros de 27.111 transtornos, 16.338 procedimentos e 5.666 químicos e drogas, e 7.477 abreviações. Cada um dos registros foi dividido em quatro tipos semânticos. Os transtornos foram divididos em “encontrado”, “sinal ou sintoma”, “doença ou síndrome” e “outros”. Os procedimentos foram divididos em “procedimento terapêutico ou prev.”, “atividade de saúde”, “Procedimento de diagnóstico” e “outros”. Por fim, os químicos e drogas foram divididos em “Substância farmacológica”, “Química orgânica”, “hormônios” e “outros”. O grupo semântico de abreviação, diferente dos demais, não foi dividido.

A tarefa de REN foi realizada com a utilização de variações do algoritmo CRF, que obtiveram diferentes resultados para os *scores* que os autores levantaram: Abreviações, transtornos, procedimentos, químicos e drogas. Os melhores resultados obtidos pelos autores foram para as categorias de “abreviações”, com precisão 0,72, revocação 0,66 e F-score 0,69, “transtornos”, com precisão 0,73, revocação 0,68 e f-score 0,70,

“procedimentos”, com precisão 0,68, revocação 0,59 e f-score 0,63, e “químicos e drogas”, com precisão 0,86, revocação 0,39 e f-score 0,54.

Por fim os autores fizeram a comparação dos resultados obtidos com trabalhos que utilizaram o CRF para extração de REN em prontuários médicos da língua inglesa e que tiveram resultados superiores. Como conclusão os autores disseram que uma das causas dos *scores* baixos é que a língua portuguesa não conta com ferramentas linguísticas como a língua inglesa, como a possibilidade de unificação de termos médicos e também devido à homogeneidade do conjunto de prontuários, já que vieram de três diferentes hospitais.

Lopes; Teixeira; Oliveira (2019)

Neste trabalho os autores utilizaram o corpus que eles mesmo geraram, o ClinPt, para a tarefa de REN. Os autores utilizaram o algoritmo de BiLSTM para criar dois modelos de *embedding* para a etapa de pré-treinamento: Um modelo que chamaram de *‘Out-of-Domain'* e um modelo que chamaram de *'In-Domain’*.

No modelo *‘Out-of-domain'*, os *embeddings* foram baixados do site da ferramenta FastText, que foi treinada a partir de textos da Wikipedia e do Common Crawl (Grave et al., 2018), enquanto que no modelo *‘In-domain’* o *embedding* foi gerado a partir de 3.377 textos de casos clínicos.

Os autores então utilizaram os dois modelos gerados para a tarefa de REN, aplicando um *fine tunning* a partir dos 281 documentos anotados. Como algoritmo de aprendizado de máquina os autores utilizaram CRF.

Como resultado, o modelo *‘In-domain’* alcançou um resultado levemente melhor que o modelo *‘Out-of-domain*’. Os valores para *‘Out-of-domain'*: precisão, revocação e f-score 0,82, e para o *‘In-domain’*: precisão: 0,83, revocação e f-score: 0,82. Alguns aspectos importantes observados que diferenciam os dois modelos quanto à interpretação da linguagem, como, por exemplo, a incapacidade do *‘Out-of-domain'* interpretar ‘EGG’ de todas as formas corretas, como o modelo *‘In-domain’* foi capaz, já que foi treinado dentro do domínio específico.

Da Rocha (2021)

Neste trabalho é apresentada uma abordagem baseada em redes neurais artificiais para a extração de dados em prontuários médicos. Um novo corpus foi criado, composto de 1.200 documentos anotados, que foram retirados de uma coleção de 30.000 documentos.

Após anotar o corpus, foi utilizada a biblioteca spaCy para realizar o REN nos outros 30.000 documentos iniciais e verificar a eficácia do modelo e a qualidade do corpus anotado. No trabalho a autora não especificou a respeito dos métodos de *embedding* utilizados pela ferramenta spaCy para a realização de tal tarefa. Porém, segundo Neumann; Beltagy; Ammar (2019), a ferramenta spaCy utiliza como *embedding* um modelo de *Chunking* baseado em transição (Stack LSTM + algoritmo de *Chunking*), que foi inspirado no modelo proposto por Lample et al. (2016). Como resultado foram obtidos os valores médios de precisão de 0,73, revocação de 0,57 e F-score de 0,64.

Niero et al. (2022)

Neste trabalho é realizado um experimento de comparação entre os tamanhos dos corpora e o resultado obtido pelo modelo de REN treinados com os respectivos corpora. Para isso, os documentos do experimento são divididos em 10 lotes e para cada lote são treinados modelos, onde cada modelo foi treinado com os lotes anteriores e mais o seu lote respectivo, de forma que o primeiro modelo foi treinado com apenas 150 documentos e o último com 1.500 documentos.

Os modelos foram treinados à partir da aplicação do *fine tunning* do BERT. O modelo base utilizado foi o BioBERTpt e a biblioteca escolhida para a execução do treinamento foi a *SimpleTransformers*[8]. Foi utilizado um batch size de 14, e Learning Rate de 4e-05.

Os autores verificaram que a performance dos modelos não melhorava conforme o tamanho do corpus ia aumentando e, em alguns casos, a performance até diminuía. Os autores também aplicaram um pós-processamento na categoria "elementos anatômicos", a fim de verificar se conseguiriam melhorar a performance do respectivo modelo, e atingiram aproximadamente 57% de ganho. Os autores concluíram então que a qualidade da anotação impacta mais do que a quantidade de dados anotados para aquele conjunto específico de dados, uma vez que alguns modelos atingiram melhor performance com lotes de 300 documentos, porém em momentos em que o time de anotadores era mais cobrado pela qualidade de anotação.

A média final do resultado dos treinamentos foi de um f1-score de 0,63 para o conjunto de todas as entidades, e de 0,37 para elementos anatômicos antes do pós-processamentovvb e 0,57 após o pós-processamento, 0,51 para condições, 0,85 para medicamentos e 0,8 para medidas.

5.3. Considerações Finais

Neste capítulo foi realizado o levantamento do estado da arte em corpus e REN aplicados a prontuários médicos tanto na língua inglesa como em português do Brasil e de Portugal. Foi possível identificar nos trabalhos realizados a partir de 2018 uma predominância da utilização do BERT, o que demonstra que ele está sendo muito utilizado para estes fins.

Foi possível também identificar que apenas cinco corpora do domínio foram anotados para a língua portuguesa e que apenas um deles está disponível publicamente, o que explica outro ponto observado: A escassez de trabalhos que aplicam REN no domínio de prontuários médicos e registros clínicos na língua portuguesa, escassez essa que torna necessário que pesquisadores que queiram realizar experimentos de REN no domínio tenham a necessidade de criar o seu próprio corpus, como foi o caso de Rocha (2021).

Pode-se concluir, com base nessas duas observações e em tudo que foi visto até aqui, que a escassez de corpus clínicos públicos e anotados na língua portuguesa têm

8. *Simpletransformers*. Disponível em: https://simpletransformers.ai/

dificultado o treinamento de modelos de REN e, consequentemente, impossibilitado as valiosas aplicações que podem ser construídas a partir da extração destes dados.

Nos capítulos seguintes serão descritas as metodologias de anotação de corpus, treinamento e aplicação do REN, as duas principais contribuições deste trabalho. Em ambos os casos, foram utilizadas informações e técnicas presentes no estado da arte e referenciadas neste capítulo.

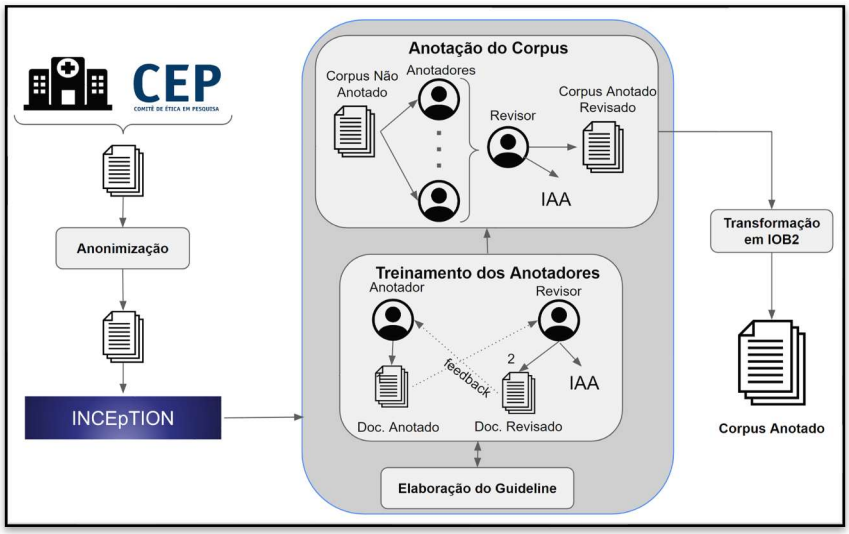
Capítulo 6 – Metodologia para a Geração do Corpus

A geração de um corpus anotado é um dos objetivos principais deste trabalho, e neste capítulo é descrito todo o processo utilizado para o cumprimento deste objetivo, desde a obtenção dos dados até a finalização da anotação e suas devidas estatísticas.

6.1. Resumo da Metodologia

A metodologia utilizada na geração do corpus está subdividida em várias etapas, que vão desde a obtenção do corpus até a preparação do corpus para servir de entrada de dados para um algoritmo. Essas etapas nem sempre acontecem de forma sequencial, podendo algumas delas ocorrer em paralelo. Todas as etapas realizadas para a geração do corpus estão descritas neste capítulo e representadas na Figura 9.

Figura 8 – Processo Completo de Criação do Corpus



Fonte: Elaborada pelo Autor

6.2. Obtenção dos registros clínicos

Existe um hospital referência em psiquiatria na cidade de São José do Rio Preto, estado de São Paulo, o Hospital Dr. Adolfo Bezerra de Menezes (HABM). Fundado em 1946, com 209 leitos de psiquiatria e com um serviço de emergência em psiquiatria que atende a 31 municípios, passou a utilizar prontuário eletrônico em 2018, e mostrou-se forte candidato a ser parceiro na realização deste trabalho.

Após procurada, a diretoria do hospital concordou em ceder os prontuários necessários para a realização deste estudo, desde que o mesmo fosse aprovado por comitê de ética em pesquisa (CEP) e os dados, antes de serem enviados para a realização do projeto, fossem desvinculados dos dados identificáveis dos pacientes.

Seguindo a legislação vigente, foi submetido um projeto de pesquisa junto à plataforma Brasil, e o CEP designado para cuidar do projeto foi o da Faculdade de Medicina de Rio Preto (FAMERP). O projeto de pesquisa foi aprovado sob o Certificado de Apresentação de Apreciação Ética (CAAE) número 54370721.0.0000.5415.

Foi solicitado ao hospital 300 documentos de notas clínicas de consultas de admissão do serviço de atendimento emergencial de psiquiatria. Segundo a diretoria clínica, as notas de admissão são mais ricas em conteúdo do que as demais notas presentes nos prontuários e, por isso, foram utilizadas para a criação do corpus.

Após recebidos os documentos, foi realizada uma etapa adicional de anonimização, onde um algoritmo de anonimização de dados textuais foi utilizado para remover possíveis dados pessoais que costumam estar presentes nas notas clínicas, como o nome do paciente, nome do acompanhante ou município de origem.

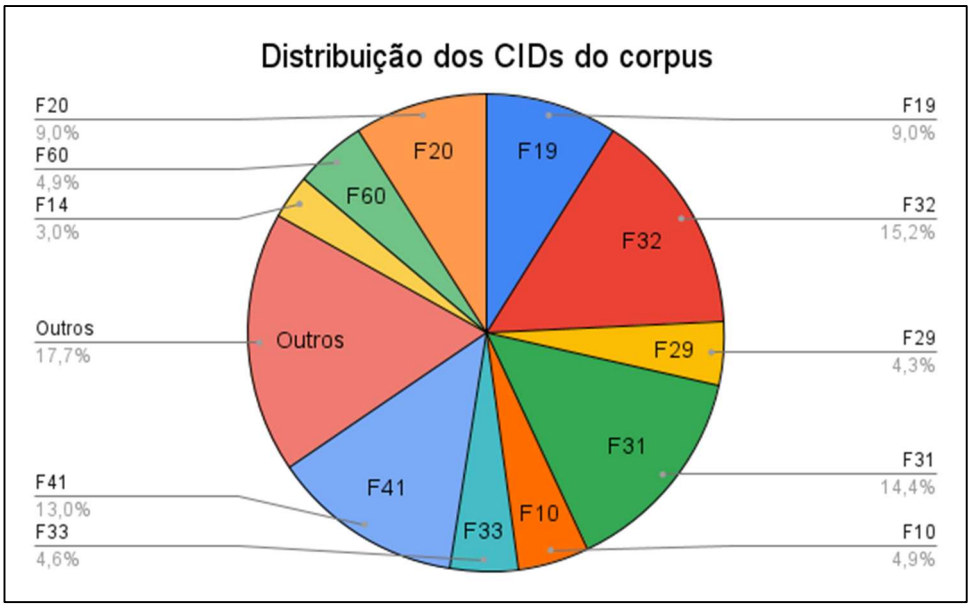
Cada um dos documentos contém um ou mais CIDs, que foram atribuídos ao paciente durante a realização do atendimento médico. A Tabela 6 mostra os CIDs com maior concentração no conjunto de documentos, e a Figura 10 mostra um gráfico da distribuição. Para obter uma melhor visualização, os CIDs foram agrupados no seu código principal. Exemplo: F31.2 e F31.3 foram agrupados em F31. Os CIDs que tiveram, mesmo após o agrupamento, menos de 10 ocorrências, foram contados como ‘outros’.

Tabela 6 – Distribuição de CIDs nos Documentos do Corpus

CID	Quantidade de ocorrências
F19	33
F32	56
F29	16
F31	39
F33	17
F41	48
F14	11
F60	18
F20	33
Outros	65

Após devidamente processados, os trezentos prontuários deram origem a um corpus de 52.891 tokens e 2.480 sentenças.

Figura 9 – Distribuição dos CIDs do Corpus



6.3 Anotação do Corpus

Por tratar-se de um subdomínio peculiar da medicina, para a anotação deste corpus foi formada uma equipe de 6 profissionais voluntários: um doutor em psiquiatria, uma psiquiatra, três residentes em psiquiatria e uma enfermeira especialista em revisão de prontuários e com experiência em psiquiatria. A anotação do corpus pode ser dividida em elaboração do *guideline*, treinamento, ferramenta de anotação, anotação do corpus e revisão das anotações, todas descritas a seguir.

6.3.1 O Processo de Elaboração do *Guideline*

Foram feitas diversas reuniões com a equipe afim de criar o *guideline* que seria utilizado para essa tarefa. Para não partir do zero, foi utilizado como guia base alguns dos tipos semânticas presentes no UMLS[9], conforme foi feito no SemClinBr (OLIVEIRA et al., 2020) e no CLEF (ROBERTS et al., 2009). As categorias utilizadas como base para a discussão do *guideline* estão presentes na Tabela 7.

Tabela 7 – Categorias das entidades e suas correspondentes UMLS

Categoria da Entidade	Correspondentes UMLS
Comportamento Social e Individual	T054 – Social Behavior T055 – Individual Behavior
Conceito Qualitativo	T080 – Qualitative Concept

[9] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/SemGroups_2018.txt

Conceito Quantitativo	T081 – Quantitative Concept
Conceito Temporal	T079 – Temporal Concept
Elemento Anatômico	T023 – Body Part, Organ, or Organ Component T029 – Body Location or Region
Medicamento	T195 – Antibiotic T200 – Clinical Drug
Negação	-
Procedimento	T059 – Laboratory Procedure T060 – Diagnostic Procedure T061 – Therapeutic or Preventive Procedure
Resultado de exame	T034 – Laboratory or Test Result
Sinal ou Sintoma	T184 – Sign or Symptom

Após discussões, chegou-se à conclusão de que algumas das categorias presentes no *guideline* base não faziam sentido para a psiquiatria, porque a informação era irrelevante para esta especialidade, ao passo que havia outras informações que não estavam sendo bem representadas por essas categorias. Foi decidido então que seriam criadas novas categorias independentes do UMLS.

Foi decidido também que não seriam anotadas categorias para Negação ou qualquer outra relação entre entidades. A decisão de não anotar as negações e as relações entre as entidades foi devido ao fato de que gostaríamos de ter a informação presente apenas na entidade, juntamente ao fato de que não tínhamos recursos humanos abundantes, e tais anotações tornariam a tarefa mais demorada. Dessa forma, foram anotadas apenas entidades.

As novas categorias inseridas no *guideline* tiveram como base, trabalhos de anotação realizados exclusivamente para a psiquiatria. Foram discutidos os *guidelines* de anotação utilizados em Mukherjee et al. (2020), Ellendford et al. (2016) e em Sridhar et al. (2021). As categorias finais utilizadas no *guideline* foram então criadas estritamente para a especialidade de psiquiatria, e podem não fazer sentido a aplicação delas em outras especialidades da medicina.

Na Tabela 8 estão as categorias das entidades e os seus especificadores. Os especificadores foram criados para as categorias que havia necessidade de uma especificidade maior da anotação, mas não grande o suficiente para serem divididos em diferentes categorias.

Tabela 8 – *Guideline* Final - Categorias das Entidades e Seus Especificadores

Categoria da Entidade	**Especificadores**
Comportamento Autodestrutivo	• Autolesão • Ideação suicida • Pensamento de morte ou ruína • Risco de suicídio • Tentativa de suicídio
Diagnóstico	-
Droga	-
Fármaco	-
Função Psíquica	• Afetividade • Apetite • Atenção

	• Consciência • Consciência do Eu • Inteligência • Linguagem • Memória • Orientação • Pensamento • Psicomotricidade • Sensopercepção • Vigilha • Vontade
Histórico Familiar	-
Histórico do Paciente	-
Observação	-
Sintoma e Queixa Psíquica	-

6.3.2 Informações sobre as Categorias Utilizadas no *Guideline*

A seguir estão descritas cada uma das categorias possíveis para as entidades, os motivos de cada uma delas existir, seus respectivos especificadores e quais informações cada uma destas devem conter.

Comportamento Autodestrutivo

Nas doenças que são relacionadas à psiquiatria, é comum que pacientes tenham comportamentos autodestrutivo. Além disso, o nível do comportamento deve ser considerado pelo médico psiquiatra na análise do tratamento que será dado ao paciente. A categoria comportamento autodestrutivo foi atribuída, juntamente com um especificador, para representar tal informação. Os especificadores utilizados para o comportamento autodestrutivo estão descritos abaixo.

Autolesão: Esse especificador foi utilizado quando as informações diziam respeito a lesões causadas pelo paciente nele mesmo. Exemplos: "cortou o antebraço", "se machucou com tesoura". Foram consideradas como autolesão as lesões que não tinham por intensão o suicídio. Por exemplo, 'cortou os pulsos', embora seja uma autolesão, é uma tentativa de suicídio, portanto não seria especificado como autolesão.

Ideação Suicida: Esse especificador foi utilizado quando as informações diziam respeito a pensamento ou planejamento de suicídio. Exemplos: "nega ideação suicida", "ideação suicida", "Planejamento suicida".

Pensamento de Morte ou Ruína: Esse especificador foi utilizado quando o paciente possuía vontade de morrer ou vontade de estar morto, porém não tinha planejamento e nem ideação de cometer o suicídio. Exemplo: "paciente diz que queria estar morto".

Risco de Suicídio: Esse especificador foi utilizado para informações que apresentam, de alguma forma, um determinado fator de risco de suicídio por parte do paciente, porém não

se encaixa em nenhum dos outros 4 especificadores. Exemplo: "Irmão se suicidou há 6 meses".

Tentativa de Suicídio: Esse especificador foi utilizado quando a informação diz respeito à uma tentativa de suicídio, seja um fato recente ou do passado do paciente. As informações podem ser mais claras, como por exemplo: "4 Tentativas prévias de suicídio", ou podem descrever as vias pela qual a tentativa ocorreu, como: "Ingeriu uma cartela de fluoxetina".

Diagnóstico

A categoria de diagnóstico foi atribuída a informações que relatavam, com determinado grau de certeza, uma patologia clínica ou psiquiátrica por parte do paciente. O enquadramento de patologias clínicas foi mais brando, e os relatos dos pacientes sobre qualquer uma destas foi considerado como diagnóstico, já para as psiquiátricas, o enquadramento foi mais criterioso, e apenas os diagnósticos descritos de forma mais formal pelo paciente, onde se havia a certeza de que esse diagnóstico de fato foi dado por um médico, que o mesmo foi considerado.

Exemplos de valores considerados para patologias clínicas: "Em uso de insulina para DM", "hipertensão", "HAS", "portador de HIV".

Exemplos de valores considerados para patologias psiquiátricas: "Paciente esquizofrênico de longa data", "relata ter sido diagnosticado com depressão desde a infância", "Transtorno afetivo bipolar"

Exemplos de valores não considerados para patologias psiquiátricas: "Paciente diz que está depressivo", "paciente diz ser muito ansioso".

Observa-se que "estar depressivo" e ter sido diagnosticado anteriormente com Depressão são informações diferentes, uma vez que o paciente pode fazer a associação de um sentimento de tristeza com o diagnóstico de depressão. Para que o modelo final não fizesse predições incorretas, foi necessário realizar essa distinção.

Droga

A categoria droga foi atribuída a qualquer informação que diz respeito a substâncias psicoativas (SPA), lícitas ou ilícitas, relatadas pelo paciente, e também à relação do paciente com essas substâncias. Exemplos: "nega uso de SPA", "relata uso de maconha e cocaína", "usou 20 pedras de crack", "etilista".

Fármaco

A categoria fármaco foi atribuída a medicamentos relatados pelo paciente. Foram considerados todos os fármacos, independente da relação deles com o paciente. Ou seja, fármacos de uso passado, em uso, prescritos pelo médico ou até mesmo suspenso foram todos considerados. Exemplos: "clonazepam (1-0-1)", ou "suspendo o uso da fluoxetina".

Função Psíquica

O quadro psíquico do paciente é, em grande parte, relatado pelos médicos psiquiatras no chamado Exame do Estado Mental (EEM). Neste exame, é comum ao médico inserir informações sobre as funções psíquicas do paciente

(DALGALARRONDO, 2018). Neste *guideline*, as funções psíquicas são utilizadas não apenas para as funções psíquicas observadas pelo médico e inseridas no EEM, mas também para trechos que tenham ligação com alguma função psíquica. Dessa forma, é possível relacionar entidades que relatam uma função psíquica de modo formal, como a classificação da função de afetividade do paciente com o fato de ele estar se apresentando choroso no momento da consulta.

Devido à importância do EEM para o entendimento do quadro do paciente, as funções psíquicas foram divididas em categorias. Essas categorias foram criadas de acordo com o Mapa Mental das Funções Psíquicas, tendo como base o trabalho de Dalgalarrondo (2018). Também foram adicionadas duas categorias ao mapa mental original: Apetite e vigília, pois a equipe de especialistas achou importante anotar essas informações como parte do exame do estado mental. As funções psíquicas utilizadas e uma breve explicação de cada uma delas são:

Afetividade: Diz respeito ao humor e ao afeto do paciente. Alguns exemplos encontrados no corpus são: "Chorosa", "humor eutímico", "afeto congruente com o humor", "chora por qualquer coisa", "irritada" e "humor irritável".

Apetite: Diz respeito ao apetite do paciente e à relação dele com o alimento e seu próprio peso. É uma das funções psíquicas que foram adicionadas ao *guideline* por conta da necessidade dos anotadores. Alguns exemplos encontrados no corpus são: "redução do apetite", "perda de peso", "não come", "hiporexia"

Atenção: Diz respeito à atenção e direção da consciência do paciente. Está presente geralmente no exame do estado mental, dificilmente é possível verificar a atenção em trechos diferentes do EEM no prontuário. Alguns exemplos são "hipoprosexia" e "dificuldade de concentração"

Consciência: Diz respeito à capacidade do paciente de contactar a realidade. Também está mais presente no trecho do EEM do que em outras partes do prontuário. Alguns exemplos são: "consciente" e "psicótico"

Consciência do Eu: Diz respeito à relação do paciente de com sua identidade e a oposição do eu em relação ao mundo. Também está mais presente no trecho do EEM do que em outras partes do prontuário. Alguns exemplos são: "pragmatismo e crítica preservados" e "ausência de crítica".

Inteligência: Diz respeito à inteligência do paciente, é testada através da capacidade dele de ter *insight* ou resolver problemas simples. Também está mais presente no trecho do EEM do que em outras partes do prontuário. Alguns exemplos são: "cognição reservada", "cognição prejudicada", "sem *insight*"

Linguagem: Esta função psíquica diz respeito à linguagem do paciente, é percebida pelo profissional ao longo da consulta, quando é percebido se o paciente está com problemas na verbalização. Utilizamos linguagem para descrever apenas a expressão do pensamento, sua elaboração e fluxo foram definidos como Pensamento. Um exemplo é "fala pouco".

Memória: Diz respeito à capacidade do paciente de evocar fatos passados, está mais presentes no trecho do EEM do que em outras partes do prontuário. Alguns exemplos são "Memória preservada" e "dificuldade de lembrar-se de fatos".

Orientação: Diz respeito à orientação do paciente em relação a si e ao espaço. Mais presente no trecho do EEM. Um exemplo é: "orientado no espaço e tempo"

Pensamento: Diz respeito à organização do fluxo de ideias do paciente, e pode apresentar-se dividido em curso, forma e conteúdo. Está presente ao longo do prontuário e não apenas no EEM, pois representa queixas comuns observadas pelo paciente ou pelo acompanhante do mesmo, como Delírios. Alguns exemplos são: "delírio" e "pensamento desorganizado"

Psicomotricidade: Diz respeito a execução de ato volitivo por parte do paciente. Está presente ao longo do prontuário e não apenas no EEM. Alguns exemplos são: "acelerado", "agitado" e "agitação psicomotora"

Sensopercepção: Diz respeito à consciência dos estímulos ambientais. Geralmente está associada ao paciente ver e ouvir coisas que não existem e, portanto, está presente ao longo do prontuário e não apenas no EEM. Alguns exemplos são: "alucinações auditivas e visuais" e "sem alteração da sensopercepção"

Vigilha: Diz respeito ao sono do paciente e a relação dele com o ato de dormir. É uma das funções psíquicas que foram adicionadas ao *guideline* por conta da necessidade dos anotadores. Alguns exemplos são: "Vigil", "não consegue dormir" e "insônia inicial".

Vontade: Diz respeito à vontade, expressão e execução dos desejos. Pode estar presente ao longo do prontuário e não apenas no EEM. Alguns exemplos são: "aparência descuidada", "desânimo" e "anedonia".

Histórico Familiar

É uma prática comum na medicina que o médico pergunte ao paciente sobre antecedentes familiares relacionados à patologia que está sendo tratada ou investigada. Na psiquiatria, geralmente os pacientes referem a parentes próximos que possuem doenças psiquiátricas, semelhantes ou não à dele. Para conseguir separar uma informação que se refere a um familiar da informação que se refere ao paciente utilizamos então esta categoria. Dentro dessa categoria não há distinção da informação: Diagnósticos, fármacos, autolesão, entre outras, são todas consideradas como Histórico Familiar se não estiverem relacionadas ao paciente. Também são anotadas, quando possível, a informação que descreve qual a pessoa ou nível de parentesco a informação principal está relacionada. Exemplos: "Mãe com esquizofrenia", "avó com Alzheimer".

Histórico do Paciente

Apesar do nome ser abrangente e dar a entender que deve conter as anotações que relatam ao histórico clínico do paciente, seja doenças anteriores, remédios já utilizados, ou qualquer informação que ocorreu no passado, essa categoria foi utilizada para as informações de históricos de tratamentos anteriores ou atuais, tratamentos interrompidos

e vida conturbada. Alguns exemplos são: “faz tratamento para depressão", “interrompeu a medicação por conta”, “quando criança fugiu de casa e foi encontrado em outra cidade”.

Observação
Algumas informações importantes sobre o paciente podem estar presentes no prontuário e não se encaixar em nenhuma das categorias anteriores. Nesse caso, foi utilizada a categoria Observação para a anotação dessas entidades. Em razão da sua função, esta categoria tende a ser a mais heterogênea do corpus. Alguns exemplos são: “separou-se recentemente do cônjuge”, “piora do quadro”, “está com problemas no trabalho”, “sofreu abuso sexual”.

Sintoma e Queixa Psíquica
Essa categoria foi criada para as informações que representam sintomas ou queixas relacionadas a patologias psiquiátricas, porém não são formais o suficiente para serem anotadas como Diagnósticos e também não se encaixam em uma função psíquica específica. Alguns exemplos são: “sente-se ansiosa”, “diz estar com depressão”, “relata sintomas depressivos”, “angústia”.

6.3.3 Treinamento dos Anotadores

Os profissionais envolvidos na anotação deste projeto nunca haviam trabalhado com anotação de corpus. Portanto, foi necessário realizar uma apresentação dos principais conceitos envolvidos na tarefa de anotação e realização do NER. Nessa apresentação foi explicado como os algoritmos funcionavam e como os exemplos deveriam ser fornecidos aos modelos para que eles pudessem aprender. Essa apresentação dos principais conceitos foi considerada como a primeira etapa do treinamento.

A segunda etapa do treinamento consistiu em uma tarefa de anotação conjunta, com uma versão quase final do *guideline*. Nesta tarefa, dois prontuários foram anotados na presença de todos, para que eles entendessem o funcionamento da ferramenta. Nesta etapa também surgiram dúvidas e sugestões, que foram consideradas para a criação da versão final do *guideline*.

A terceira etapa consistiu em selecionar aleatoriamente 10 documentos e pedir aos anotadores que os anotassem separadamente. Nesta etapa havia apenas três anotadores na equipe. Para calcular o IAA dos anotadores nesta etapa, foi utilizado um match restrito, onde as entidades consideradas como "anotadas corretamente" deveriam ser semelhantes em palavras e classificação para ambos os três anotadores. O valor de IAA obtido nesta etapa foi de 0,365.

Por conta do baixo IAA obtido, as anotações divergentes foram registradas em um documento e foi realizada uma discussão sobre as diferentes interpretações das anotações. Nessa discussão, todos os anotadores chegaram em consensos sobre as principais divergências. Nesta etapa alguns exemplos foram inseridos no *guideline*.

Foi realizada então uma quarta etapa, com um objetivo de verificar a consolidação das informações discutidas na etapa três e recalcular o IAA. Nessa etapa apenas 2 documentos foram anotados por cada um dos anotadores. O IAA obtido foi de 0,60. Esse IAA passou a ser o valor mínimo a ser atingido para os novos anotadores que integrassem

o projeto. A partir dessa etapa foram liberados os documentos do corpus final para que os anotadores pudessem anotar.

Posteriormente ao início das anotações, dois novos anotadores se juntaram à equipe. O treinamento desses novos anotadores foi realizado com base do *guideline* já atualizado. Os anotadores foram liberados a iniciar a anotação do corpus final após atingir o IAA mínimo.

6.3.4 O processo de Anotação do Corpus e os Índices de Aceitação Obtidos

Após os anotadores passarem pelas etapas de treinamento e serem considerados aptos à anotação do corpus, teve início a anotação do corpus final. Foi criado um canal de comunicação entre os anotadores e o presente mestrando, para que as dúvidas fossem tiradas à medida que elas pudessem surgir. Algumas dúvidas que surgiram ao longo da anotação transformaram-se em exemplos do *guideline*.

Cada anotador anotou aproximadamente 60 documentos. O processo de anotação durou em média quatro meses. Durante o processo, foi disponibilizado aos anotadores dois lotes contendo quatro documentos cada, que foram utilizados para medir o IAA, um aproximadamente na metade das anotações, que foi chamado de 'rodada 1' e o outro próximo ao final delas, que foi chamado de 'rodada 2'. Para que o IAA pudesse ser medido, todos os cinco anotadores anotaram os mesmos documentos. Além de serem utilizados para o IAA, as anotações desses documentos foram devolvidas com *feedbacks* aos anotadores, sobre as divergências encontradas nestas anotações.

O cálculo do IAA foi feito com base em duas abordagens: (a) Abordagem estrita, onde foram consideradas como corretas as anotações semelhantes em palavras (*span*) e em categoria, (b) Abordagem relaxada, onde foram consideradas como corretas as anotações que continuam um mesmo subconjunto de palavras e eram semelhantes em categoria. Essa técnica teve como base os cálculos de IAA presentes em Oliveira et al. (2020). O IAA obtido nas etapas calculadas durante o processo de anotação estão na Tabela 9.

Tabela 9 – IAA Calculado Durante as Etapas de Anotação

Etapa do processo	Quantidade de documentos	IAA Estrito	IAA Relaxado
Rodada de treinamento	10	0,365	-
Rodada pré-anotação	2	0,6	-
Rodada 1	4	0,34	0,49
Rodada 2	4	0,318	0,475
Média Geral entre as rodadas		0,406	0,48

Por conta do valor baixo obtido nos IAA aplicados durante a etapa de anotação, houve a necessidade de entender a distribuição dos erros ao longo das categorias. Foi feito então uma análise de IAA separada de acordo com cada categoria. Para as categorias que possuem especificadores, foi feita ainda uma segunda análise, para verificar se houve erro na anotação apenas do especificador. Os resultados estão presentes na Tabela 10.

Tabela 10 – Distribuição do IAA por Categoria

Categoria	IAA Estrito	IAA Relaxado
Comportamento Autodestrutivo (com especificação)	0,726	0,780
Comportamento Autodestrutivo (sem especificação)	0,768	0,863
Diagnóstico	0,259	0,778
Droga	0,645	0,694
Fármaco	0,897	0,913
Função Psíquica (com especificação)	0,437	0,609
Função Psíquica (sem especificação)	0,477	0,689
Histórico Familiar	0,896	1
Histórico do Paciente	0,289	0,466
Observação	0,257	0,284
Sintoma e Queixa Psíquica	0,362	0,456
Média Geral entre as Categorias	0,547	0,685

A ferramenta escolhida para a tarefa foi o INCEpTION[10] pois é sucessor do WebAnno, que teve avaliação positiva no levantamento realizado por Andrade, Oliveira e Moro (2016) e dá suporte a documentação interna e comunicação entre à equipe, além de ter uma interface amigável e ser acessível pelo navegador de Internet.

6.4. Revisão das Anotações

Devido aos valores baixos obtidos durante o cálculo do IAA, todos os documentos anotados passaram por um processo de revisão de um especialista em PLN. Neste processo, interpretações divergentes que os anotadores tiveram sobre determinados termos foram padronizados, com o objetivo de aumentar a homogeneidade das anotações do corpus.

Após a finalização das revisões das anotações foi gerado um novo corpus, contendo apenas os valores mantidos após a revisão, e os documentos não revisados foram deixados em um corpus à parte. Ambos os corpora, o não revisado e o totalmente revisado, foram então utilizados para treinar os modelos, a fim de verificar a importância dessa tarefa, através da diferença do F-score obtido em cada um dos corpora.

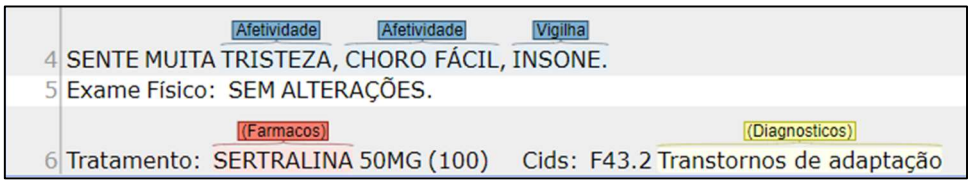
6.5. Preparação do Corpus para o Treinamento

Após criado o corpus, foi necessário realizar algumas adaptações para que ele pudesse servir como entrada de dados para modelos de REN. Entre as maneiras de representar os textos e suas anotações, foi escolhida a IOB2, pela simplicidade da representação. O corpus passou a ser representado pelos tokens e suas anotações. As sentenças foram separadas através de linhas em branco, e, os documentos, através de duas

[10] https://github.com/inception-project/inception

linhas em branco. Na Figura 11 há um exemplo de três sentenças anotadas na ferramenta de anotação, a na Tabela 11 a representação destas mesmas sentenças no formato IOB2.

Figura 10 – Exemplo de Sentenças Anotadas na Ferramenta de Anotação



Fonte: Anotado pelo autor na ferramenta INCEpTION.

Tabela 11 – Representação IOB2 dos Tokens das Sentenças

Token	**Representação**
SENTE	O
MUITA	O
TRISTEZA,	B-Funcao_psiquica:Afetividade
CHORO	B-Funcao_psiquica:Afetividade
FÁCIL,	I-Funcao_psiquica:Afetividade
INSONE	B-Funcao_psiquica:Vigilha
Exame	O
Físico:	O
SEM	O
ALTERAÇÕES.	O
Tratamento:	O
SERTRALINA	B-Farmacos
50MG	O
(100)	O
Cids:	O
F43.2	O
Transtornos	B-Diagnósticos
de	I-Diagnósticos
adaptação	I-Diagnósticos

Como as anotações originais continham tokens que podiam ser representados por mais de uma entidade, chamadas de *nested entities*, foi necessário definir qual abordagem seria feita para que essas entidades fossem corretamente representadas. Há duas formas de se tratar *nested entities* (MARCIŃCZUK, 2021): (a) Gerando um corpus para cada tipo de entidade e então treinando os modelos separadamente, (b): Criando categorias adicionais para as entidades, formadas pela aglutinação das categorias de cada token.

Como havia poucas *nested entities* no corpus, nenhuma das técnicas foi utilizada. Em vez disso, as *nested entities* foram removidas do texto. Para a remoção, quando não havia um critério mais óbvio no sentido clínico de qual entidade deveria ser mantida, foi considerado um grau de importância para cada categoria, conforme a Tabela 12, onde os valores menores de prioridade representam as categorias mais importantes, e os valores

maiores, as menos importantes. O grau de importância atribuído a cada categoria foi subjetivo, de acordo com a percepção do autor sobre o corpus.

Tabela 12 – Grau de Importância das Categorias para a Remoção de *Nested Entities*

Categoria da Entidade	Prioridade
Função Psíquica	0
Diagnóstico	1
Sintomas e Queixas Psíquicas	2
Histórico do Paciente	3
Drogas	4
Histórico Familiar	5
Fármacos	6
Observação	7

6.6. Corpus Final

A versão final do corpus ficou com um total de 5.822 entidades, representadas por 5.822 entidades do tipo B-Categoria e 8.611 entidades do tipo I-Categoria, formando um total de 14.433 tokens IOB2, além de 38.458 entidades do tipo 'O', distribuídas ao longo de 2.480 sentenças. A quantidade de entidades por categoria e a distribuição dos tokens pelo corpus estão nas Tabelas 13 e 14, respectivamente. Nas tabelas, a proporcionalidade entre as entidades não considerou o tipo 'O'.

Tabela 13 – Distribuição das Categorias no Corpus

Tipo de Entidade	Quantidade no corpus	Proporcionalidade
Comportamento Autodestrutivo	379	6,51%
Diagnóstico	819	14,07%
Droga	367	6,30%
Fármaco	853	14,65%
Função Psíquica	1.687	28,98%
Histórico Familiar	66	1,13%
Histórico do Paciente	395	6,78%
Observação	319	5,48%
Sintoma e Queixa Psíquica	711	12,21%

Tabela 14 – Distribuição de Tokens IOB2 de cada Categoria no Corpus

Token IOB2	Quantidade no corpus	Proporcionalidade
B-Comportamento_Autodestrutivo	379	2,63%
I-Comportamento_Autodestrutivo	819	5,67%
B-Diagnostico	1.045	7,24%
I-Diagnostico	2.167	15,01%
B-Droga	367	2,54%
I-Droga	963	6,67%
B-Farmaco	853	5,91%
I-Farmaco	78	0,54%
B-Funcao_Psiquica	1.687	11,69%

I-Funcao_Psiquica	1.749	12,12%
B-Historico Familiar	66	0,46%
I-Historico Familiar	186	1,29%
B-Historico_Paciente	395	2,74%
I-Historico_Paciente	1.093	7,57%
B-Observacao	319	2,21%
I-Observacao	811	5,62%
B-Sintomas_e_Queixas_Psíquicas	711	4,93%
I-Sintomas_e_Queixas_Psíquicas	748	5,18%

Além da distribuição entre *Begin* e *Inside* e quantidade de entidades no corpus, as entidades também se diferem de acordo com a variância dos seus termos. Nas tabelas de 15 a 23 estão os tokens com maior ocorrência em cada uma das categorias, acompanhados pela quantidade de vezes que eles ocorrem nas respectivas categorias e no corpus.

Tabela 15 – Tokens com maior frequência: Comportamento Autodestrutivo

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
suicida	134	135	em	17	446
de	125	2.276	tentativas	16	19
ideação	119	121	pensamento	16	73
nega	78	285	suicidio	13	16
morte	53	64	matar	11	29
pensamentos	51	65	intenção	10	15
tentativa	37	41	planejamento	10	11
suicídio	28	29	ou	10	111
se	24	128	tae	8	8
sem	24	387	.	8	2.192

Tabela 16 – Tokens com maior frequência: Diagnóstico

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
de	209	2.276	devidos	69	69
transtorno	150	156	comportamentais	68	68
e	127	1.124	dependência	59	60
ao	108	189	sintomas	54	185
uso	102	468	psicóticos	54	70
depressivo	102	123	não	50	326
episódio	96	112	bipolar	49	50
transtornos	83	84	atual	49	372
síndrome	71	77	afetivo	48	48
mentais	69	69	esquizofrenia	47	48

Tabela 17 – Tokens com maior frequência: Drogas

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
de	192	2.276	nega	26	285
uso	110	468	alcool	22	23
crack	51	52	usuário	19	22
drogas	38	77	por	19	198
maconha	35	35	.	17	2.192
álcool	33	50	anos	17	357
dia	28	92	e	16	1.124
há	28	439	pinga	13	13
spa	26	26	2	12	271
cocaína	26	37	faz	12	89

Tabela 18 – Tokens com maior frequência: Fármacos

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
sertralina	113	113	ácido	22	22
clonazepam	74	75	diazepam	22	22
risperidona	38	38	valpróico	21	21
prometazina	29	29	haldol	18	18
fluoxetina	28	28	carbamazepina	18	18
rivotril	25	25	valproico	16	16
haloperidol	25	25	paroxetina	15	15
escitalopram	25	25	acido	14	14
clorpromazina	24	24	amitriptilina	14	14
quetiapina	24	24	neozine	12	13

Tabela 19 – Tokens com maior frequência: Função Psíquica

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
de	125	2.276	choro	42	42
humor	97	100	insonia	42	42
tristeza	79	82	orientada	42	44
e	77	1.124	alucinações	41	43
sem	73	387	não	41	326
discurso	55	60	anedonia	39	39
pensamento	53	73	alterações	38	140
no	49	300	psicomotora	33	33
agitação	45	45	crítica	33	34
afeto	43	44	consciente	33	33

Tabela 20 – Tokens com maior frequência: Histórico Familiar

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
com	15	788	da	5	201
mãe	13	149	drogas	4	77
de	10	2.276	transtorno	4	156
pai	10	45	avó	3	13
mae	8	30	paterna	3	3
-	7	582	tab	3	20
depressão	7	36	quadro	3	118
e	5	1.124	suicidou-se	3	4
irmã	5	43	internação	3	90
tia	5	10	psiquiátrica	3	11

Tabela 21 – Tokens com maior frequência: Histórico do Paciente

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
tratamento	88	420	no	28	300
internações	61	64	caps	24	46
acompanhamento	56	73	não	24	326
de	44	2.276	há	23	439
internação	41	90	medicações	20	83
prévias	36	45	para	20	264
nega	35	285	uso	19	468
em	35	446	anos	16	357
psiquiátrico	33	42	fez	14	49
faz	29	89	a	14	572

Tabela 22 – Tokens com maior frequência: Observações

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
de	46	2.276	sintomas	14	185
do	30	264	sem	13	387
há	27	439	nega	13	285
uso	22	468	não	12	326
piora	20	40	da	12	201
após	20	104	com	11	788
quadro	17	118	que	9	746
samu	17	48	dias	9	184
pelo	16	76	anos	9	357
em	15	446	marido	9	66

Tabela 23 – Tokens com maior frequência: Sintomas e Queixas Psíquicas

Token	#Ocorrências na Categoria	#Ocorrências no Corpus	Token	#Ocorrências na Categoria	#Ocorrências no Corpus
de	109	2.276	nega	21	285
sintomas	78	185	sensação	20	29
depressivos	40	53	quadro	18	118
angústia	33	38	depressivo	16	123
crises	28	48	psicóticos	16	70
medo	27	35	pânico	15	28
ansiedade	26	52	agressiva	13	14
agressividade	25	25	heteroagressividade	13	14
agressivo	23	24	falta	12	20
comportamento	22	45	ar	12	12

A Tabela 24 resume a quantidade de tokens únicos, total de tokens, média de ocorrência dos 20 tokens mais constantes e médias de ocorrência entre todos os tokens em cada uma das categorias anotadas no Corpus.

Tabela 24 – Resumo dos Tokens por Categoria

Categoria	#Tokens únicos	#Total de Tokens	Média de ocorrência dos 20 tokens de maior ocorrência	Média de ocorrência entre todos os tokens
Comportamento Autodestrutivo	255	1.197	39,6	4,7
Diagnostico	358	3.211	83,2	9,0
Drogas	254	1.329	37,0	5,2
Fármacos	169	930	28,9	5,5
Funções Psíquicas	713	3.435	54,0	4,8
Histórico Familiar	126	248	6,0	2,0
Histórico Paciente	326	1.487	33,0	4,1
Observação	466	1.129	17,1	2,4
Sintomas e Queixas Psíquicas	456	1.458	28,35	3,2

6.7. Comparação com Demais Corpus Anotados

O Corpus anotado neste trabalho se diferencia dos demais corpus da literatura devido à duas características peculiares: (a) O próprio idioma, já que, como foi visto no capítulo de revisão bibliográfica, há pouquíssimos corpus clínicos anotados para REN na língua portuguesa e (b) As entidades voltadas para a área de psiquiatria.

Em relação às entidades voltadas para a área de psiquiatria, pode-se dizer que há poucos trabalhos que exploram a extração de informações psiquiátricas relevantes de notas clínicas de psiquiatria, sendo Mukherjee et al. (2020) um dos únicos exemplos. Este

trabalho se baseou nele para tentar extrair, em formas de entidades, todas as informações consideradas relevantes, do ponto de vista médico, para avaliar pacientes psiquiátricos. Essa adaptação resultou na criação das categorias de entidades "Funções Psíquicas" e "Sintomas e Queixas Psíquicas", além da categoria "Comportamento Autodestrutivo", sendo esta última mais conhecida na literatura por englobar relações com o suicídio.

6.8. Considerações Finais

Neste capítulo foi apresentado todo o processo de criação e anotação do corpus, assim como estatísticas gerais sobre ele e cada uma das categorias anotadas.

O processo de anotação de um corpus está longe de ser uma tarefa simples. Hovy e Lavid (2010) utilizam o termo "Ciência da anotação" no artigo que descreve importantes considerações a respeito de anotações de dados para REN. Entre as dificuldades encontradas no processo, manter o *agreement* entre os anotadores talvez seja a maior delas, dada à subjetividade envolvida em alguns tipos de anotação, principalmente quando as categorias possuem determinados níveis de intersecção, como é comum em corpus anotados para REN clínico.

O corpus gerado foi nomeado PSYclinBR, e traz uma série de contribuições. A primeira delas é que aumenta em 30% a quantidade de documentos anotados e disponibilizados publicamente para REN no domínio clínico. A segunda é que PSYclinBR traz um guideline de entidades que cobrem por completo a prática da psiquiatria, enquanto os demais corpus de psiquiatria cobriam apenas informações específicas.

O corpus gerado pela anotação está disponível no Github do autor[11].

No próximo capítulo será apresentada a metodologia utilizada em REN, que faz uso do corpus cuja anotação foi descrita neste capítulo e pode ser visto como uma conclusão do ciclo do trabalho.

[11] https://github.com/luizniero/PSYclinBr

Capítulo 7 – Metodologia do Reconhecimento de Entidades Nomeadas

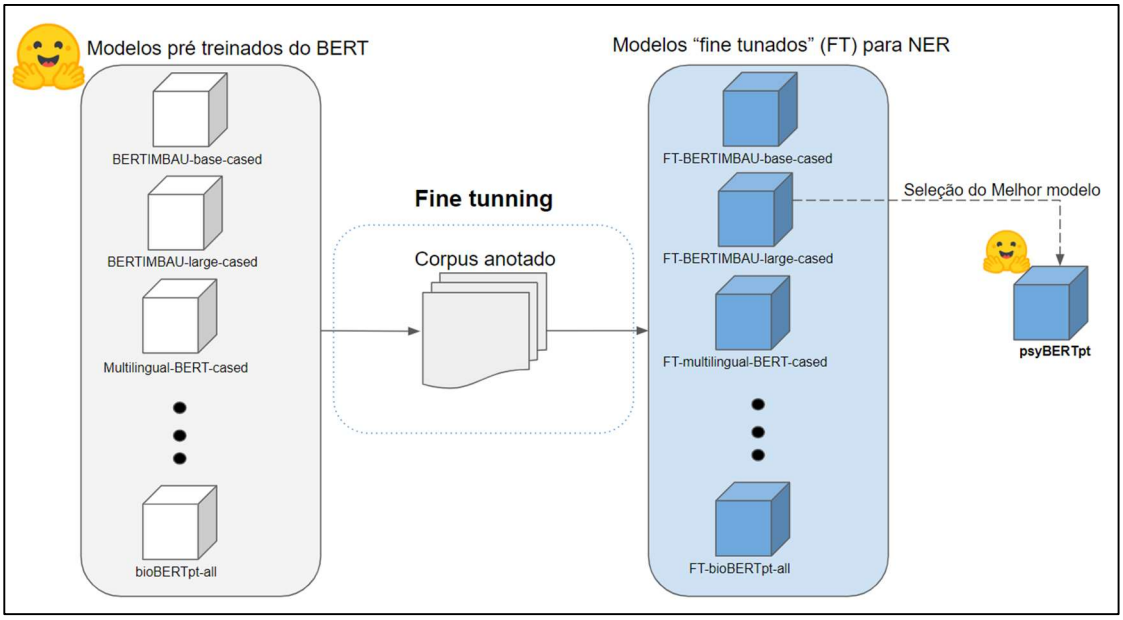
O treinamento de um modelo do BERT para a tarefa de REN é outro objetivo principal deste trabalho. Neste capítulo são apresentadas de forma detalhada as etapas envolvidas na metodologia utilizada nos experimentos dos modelos de REN baseados no BERT assim como os resultados obtidos.

7.1. Resumo da Metodologia

Após a anotação do corpus, foi realizado o treinamento de modelos de REN especializado para informações de prontuários de psiquiatria. Para o treinamento, foram utilizados como base os modelos de BERT disponíveis para a língua portuguesa acrescidos de uma camada do algoritmo *softmax*, de forma que o modelo com maior performance foi selecionado como modelo final, o qual recebeu o nome de psyBERTpt. A Figura 12 apresenta um resumo dos treinamentos e a geração do psyBERTpt. Tanto os modelos utilizados para os treinamentos como o modelo final gerado pelo treinamento[12] estão disponíveis no repositório do *HuggingFace*[13].

[12] https://huggingface.co/psybertpt/psyBERTpt
[13] https://huggingface.co/

Figura 11 – Resumo dos Treinamentos e a Geração do psyBERTpt



Fonte: Elaborada pelo autor.

7.2. Ambiente de Execução

O algoritmo escolhido para esta tarefa foi o BERT (DEVLIN, 2018), devido ao seu alto desempenho em tarefas de REN, conforme foi apresentado no Capítulo 2. Foi realizado o *fine tunning* do BERT para a tarefa de REN utilizando como dados de aprendizado o corpus anotado. Para a classificação dos Tokens, foi utilizada a função de ativação Softmax.

A fim de encontrar os melhores resultados com o mesmo conjunto de dados, foram utilizados modelos do BERT pré-treinados com dados genéricos da língua portuguesa: BERTIMBAU *large cased* e BERTIMBAU *base cased* (SOUZA; NOGUEIRA; LOTUFO, 2020), BERT *multilingual cased* e BERT *multilingual uncased* (DEVLIN, 2018), e também com modelos do BERT adaptados para textos clínicos, biomédicos e ambos: BioBERTpt-clin, BioBERTpt-bio e BioBERTpt-all *cased* (TERUMI et al, 2020). Todos os modelos referidos estão disponíveis para a utilização na API HuggingFace (WOLF et al., 2019).

Os modelos do bioBERTpt são os únicos da língua portuguesa (TERUMI et al, 2020) que passaram por uma adaptação de domínio para textos clínicos e biomédicos e por esse motivo são os únicos modelos com algum *fine tunning* prévio que estão entre a lista dos modelos usados.

Para a execução dos algoritmos de treinamento, foi utilizado o ambiente do Google Colab na versão Pro+. Os treinamentos (*fine tunnings*) foram realizados com os parâmetros descritos na sessão 7.2 e levaram, em média, 10 minutos cada.

7.3. Calibração dos Parâmetros para o Treinamento do Modelo

Para o treinamento dos modelos de REN, foi utilizada a biblioteca *SimpleTransformers*[14] *release* v4.31.0. Essa biblioteca possibilita que o treinamento dos modelos seja feito a partir do *fine tunning* de qualquer modelo pré-treinado de BERT. Para o treinamento, é possível alterar vários parâmetros do modelo, como taxa de aprendizagem, número de épocas, entre outros.

A fim de buscar um conjunto de parâmetros que resultaria em bons *scores* para o treinamento, foram feitos vários testes, tendo como base os valores estipulados para o *fine tunning* em Devlin (2018). Os resultados dos testes foram medidos de acordo com as principais métricas que indicam a performance de modelos de predição: *Training loss*, *validation loss*, precisão, revocação e f1-score. Os modelos que obtiveram os melhores resultados foram treinados com os parâmetros exibidos na Tabela 25.

Tabela 25 – Parâmetros Utilizados na Execução dos Treinamentos

Parâmetro	Valor	Significado
Max_seq_length	256	Tamanho máximo da sentença no BERT
Num_train_epochs	4	Quantidade de épocas do treinamento
Use_multiprocessing	True	Otimização do treinamento por multiprocessamento
Train_batch_size	16	Tamanho do batch do treinamento
Learning_rate	4e-5	Taxa inicial de aprendizado do modelo
Adam epsilon	1e-08	Valor de epsilon para a aplicação da otimização AdamW
Optimizer	AdamW	Algoritmo de Otimização de treinamento de redes neurais
Manual_seed	1	Inicia a rede neural com os mesmos pesos
Encoding	UTF-8	Encoding do corpus

A fim de avaliar se os parâmetros escolhidos estavam corretos e também se não havia problemas com a biblioteca, foi executado o *fine tunning* para REN com o corpus do CoNLL2003 utilizando os parâmetros selecionados. Os resultados obtidos estão na Tabela 26.

Tabela 26 – Resultados Obtidos pelo *Fine Tunning* com o Corpus do CoNLL2003

Resultados obtidos utilizando o corpus do CoNLL2003	Valor
Training loss	0,00017
Validation loss	0,1381
Precisão	0,9051
Revocação	0,9226
F1-Score	0,9138

[14] https://simpletransformers.ai/

Embora o resultado obtido pelo *fine tunning* do BERT-base-cased com a biblioteca utilizada esteja um pouco abaixo dos benchmarks do CoNLL2003, foi obtido um f1-score maior que 0,9. Considerando que a calibração do modelo foi feita com o conjunto de dados deste trabalho e não especificamente para o CoNLL2003, conclui-se que o treinamento dos modelos pelo *SimpleTransformers*, que correspondem ao *fine tunnig* dos demais modelos de BERT apresentados na literatura e, portanto, os parâmetros utilizados estão corretos.

7.4. Resultados Obtidos

Para avaliar o resultado final, foram medidos a média e o desvio padrão dos valores de precisão, revocação e f1-score obtido pelos 5 *folds* em cada um dos modelos onde o *fine tunning* foi aplicado. O melhor resultado foi obtido a partir do *fine tunning* do BERTIMBAU-large, com valores de precisão: 0,65 ± 0,02, revocação: 0,69 ± 0,03 e f1-score: 0,67 ± 0,02. Os valores após ± tratam-se do desvio padrão entre os *folds*. Os valores obtidos por cada um dos modelos está presente na Tabela 27, onde os melhores valores para cada um dos atributos estão destacados em negrito. O modelo criado a partir do *fine tunning* do BERTIMBAU large cased foi selecionado como melhor modelo porque atingiu *scores* gerais melhores que os demais modelos. Ao modelo que atingiu o melhor resultado foi dado o nome de psyBERTpt e o mesmo tornou-se disponível publicamente pela API da biblioteca *HuggingFace* em março de 2023.

Tabela 27 – Resultados Obtidos no Treinamento dos Modelos

Modelo Base	***Training loss***	***Eval loss***	**Precisão**	**Revocação**	**F1-Score**
BERTIMBAU base cased	0,23 ± 0,09	0,43 ± 0,02	0,62 ± 0,03	0,66 ±0,02	0,64 ±0,03
BERTIMBAU large cased	**0,13 ± 0,09**	**0,41 ± 0,03**	**0,65 ±0,02**	**0,69 ±0,03**	**0,67 ±0,02**
Multilingual BERT base cased	0,33 ± 0,13	0,47 ± 0,02	0,59 ±0,02	0,63 ±0,01	0,61 ±0,01
Multilingual BERT base uncased	0,74 ± 0,16	0,81 ± 0,02	0,62 ±0,02	0,41 ±0,13	0,45 ±0,01
BioBERTpt-clin	0,19 ± 0,08	**0,41 ± 0,03**	0,64 ±0,03	**0,69 ±0,03**	0,66 ±0,03
BioBERTpt-bio	0,30 ± 0,12	0,46 ± 0,02	0,61 ±0,03	0,64 ±0,02	0,62 ±0,03
BioBERTpt-all	0,19 ± 0,09	0,42 ± 0,03	0,63 ±0,03	**0,69 ±0,02**	0,66 ±0,03

A fim de analisar a performance obtida separadamente para cada uma das categorias, foi realizado o cálculo da precisão, revocação e f1-Score separadamente para cada uma delas, em cada um dos treinamentos. Os valores obtidos estão disponíveis na Tabela 26. O valor final obtido por cada uma das categorias, calculado pela média dos 5 respectivos *folds*, está resumido na Tabela 28.

Tabela 28 – Resultados Obtidos por Categoria

Categoria REN	**Precisão**	**Revocação**	**F1-Score**
Comportamento Autodestrutivo	0,54 ±0,05	0,61 ±0,08	0,57 ±0,06
Diagnóstico	0,85 ±0,04	0,85 ±0,04	0,85±0,03

Drogas	0,47 ±0,07	0,54 ±0,04	0,50 ±0,06
Fármacos	0,96 ±0,02	0,95 ±0,02	0,95 ±0,01
Função Psíquica	0,70 ±0,03	0,77 ±0,03	0,74 ±0,03
Histórico Familiar	0,39 ±0,08	0,34 ±0,09	0,35 ±0,06
Histórico Paciente	0,30 ±0,04	0,35 ±0,10	0,32 ±0,07
Observações	0,09 ±0,01	0,10 ±0,01	0,09 ±0,01
Sintomas e Queixas Psíquicas	0,53 ±0,06	0,59 ±0,07	0,56 ±0,06

A performance obtida pelo modelo treinado com os dados do corpus apresentado neste trabalho teve resultado semelhante, em termos de f1-score, a outros trabalhos de REN em dados de prontuários da língua portuguesa que realizaram o treinamento a partir do *fine tunning* do BERT. A Tabela 29 faz uma comparação dos resultados deste trabalho com trabalhos semelhantes em português do Brasil.

Tabela 29 – Resultados Obtidos neste Trabalho e em Trabalhos Relacionados

Trabalho	**Precisão**	**Revocação**	**F1-Score**
Lopes, Teixeira e Gonçalo, 2019	-	-	[0,63-0,7]
Da Rocha et al., 2022	0,73	-	0,64
Terumi et al., 2020	0,61	0,61	0,60
O método proposto	0,64	0,62	0,66

7.5. Análise dos Resultados

Embora o resultado geral do modelo tenha sido satisfatório quando comparado aos outros trabalhos com propostas semelhantes, é possível perceber a grande diferença entre os *scores* obtidos pelas diferentes categorias de entidades presentes no REN. Esta sessão propõe uma análise dos resultados destas categorias considerando três diferentes aspectos:

- Representatividade da categoria no corpus
- Separatividade da categoria
- Qualidade da anotação da categoria

7.5.1. Representatividade da Categoria no Corpus

A representatividade no corpus refere-se à faculdade da categoria da entidade de ser representada corretamente no corpus. A representatividade está relacionada a dois fatores: (a) A quantidade de vezes que as entidades da categoria ocorrem e são anotadas na categoria e (b) A quantidade de vezes que as entidades ocorrem no corpus, mas não na categoria. Categorias que possuem muitas stop words como tokens mais representativos são categorias que tendem a sofrer com a representatividade.

Uma categoria com excelente representatividade possui um alto valor para (a) e um baixo valor para (b), ou seja, possui muitas entidades e tokens anotadas exclusivamente para a sua categoria. As tabelas presentes na sessão 6.5 possuem informações sobre a

ocorrência das entidades nas categorias e no corpus, e podem auxiliar a análise das representatividades.

Categorias com boas representatividades no corpus são: Fármacos e Sintomas e Queixas Psíquicas, porque os tokens que compõe as entidades destas categorias quase não aparecem no Corpus como sendo de outras categorias.

Pode-se tomar como exemplos de alta representatividade de alguns tokens de Fármacos: sertralina, clonazepam e risperidona que, juntos, possuem apenas 1 token que não está anotado em 'Fármacos'.

Categorias com representatividades ruins no corpus são: Observações e Histórico familiar, porque os tokens presentes nas entidades nestas categorias estão presentes em volume muito maior no corpus do que foram anotados para esta categoria.

Pode-se tomar como exemplos de baixa representatividade alguns tokens do Histórico Familiar: "mãe" e "depressão"

O token "mãe" ocorre 13 vezes na categoria Histórico Familiar e 149 em todo o corpus, como um token do tipo 'O' ou de alguma categoria. Da mesma forma, o token "depressão" ocorre 7 vezes na categoria Histórico Familiar" e 36 vezes no corpus. Ou seja, tanto "mãe" como "depressão" são mais classificados como não sendo Histórico Familiar do que o oposto, afetando negativamente a representatividade da categoria.

7.5.2. Separatividade da Categoria

Refere-se ao quanto uma categoria pode ser distinguível das demais categorias, na amplitude dos entendimentos. A separatividade aplica-se desde a dificuldade que um anotador tenha para classificar determinada entidade em uma ou outra categoria até as confusões que serão aprendidas e criadas pelo modelo de linguagem para separar essas mesmas categorias.

Uma forma de evitar categorias com baixa separatividade é uma boa elaboração do *guideline*, que deixe bem claro o limite de cada categoria a ser anotado e dê exemplos minuciosos sobre cada um dos casos, e muitas vezes a percepção de um *guideline* com categorias que são pouco separadas deve ser percursora da modificação do *guideline*. São exemplos de situações em que as categorias possuem baixa representatividade:

Considere o trecho de prontuário:

"Paciente etilista de longa data, ingere etanol de posto de gasolina".

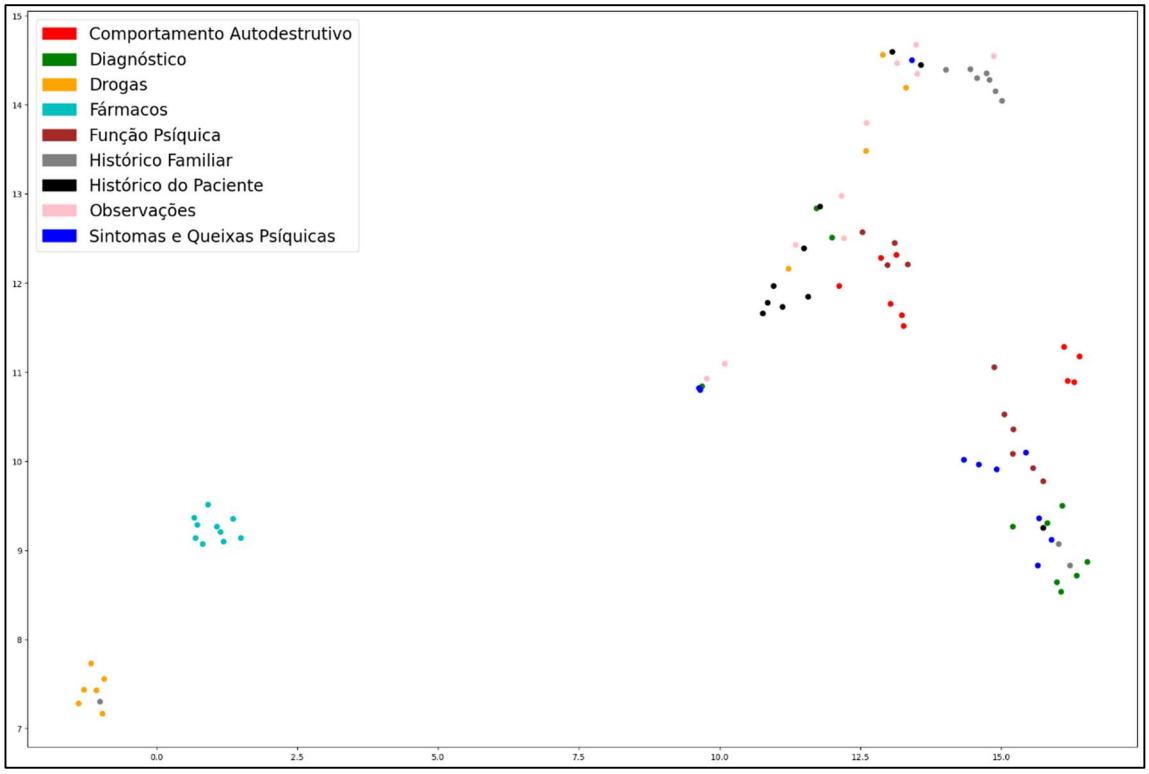
Considere que um anotador precisa anotar o trecho exibido e deve decidir se o termo "etilista" será marcado como "condição associada à doença" (Diagnóstico) ou "drogas e seu uso" (drogas). A depender do ponto de vista do anotador, pode ser as duas coisas, logo um *guideline* que não deixe claro essa percepção, diminui a separatividade destas categorias.

Considere agora que o anotador tenha decidido que o termo "etilista" se trata de um Diagnóstico. O modelo de linguagem tenderá a representar este termo com *embeddings* mais próximos da palavra "etanol" e "etilismo" do que a *embeddings* de outras doenças

como: “Transtorno de Personalidade”. Nesse caso, a elaboração do *guideline* deve considerar também que a separatividade das categorias em clusters com *embeddings* semelhantes poderá fazer com que os modelos treinados atinjam melhores *scores*.

Na Figura 12 há um exemplo da separatividade dos 10 tokens mais frequentes para cada uma das categorias. Para a plotagem foram considerados os *embeddings* da última camada, e foi utilizada a técnica UMAP para a redução de dimensionalidade para que os *embeddings* pudessem ser representados em 2 dimensões. Na figura 13 é possível verificar que há categorias que são mais facilmente separáveis, como Drogas (em laranja) e Fármacos (em ciano), enquanto outras aparecem de forma bem heterogênea e se misturam constantemente, como Sintomas e Queixas Psíquicas (azul), Função Psíquica (marrom) e Diagnóstico (Verde).

Figura 12 – Plotagem em 2 Dimensões dos *Embeddings* dos 10 Tokens Mais Frequentes de Cada Categoria do Corpus



Fonte: Elaborada pelo autor.

7.5.3. Qualidade da Anotação da Categoria

Se o IAA de uma categoria estiver baixo, serão necessários muito mais exemplos para que o modelo consiga aprender a classificar bem essa categoria, e talvez mesmo milhares de exemplos não sejam o suficiente. A qualidade da anotação interfere diretamente na qualidade do score da categoria, e por isso é necessário que, durante a tarefa de anotação, alguém fique responsável por verificar os IAA e tentar resolver problemas/padronizar as anotações.

Embora tenha-se visto que os *scores* finais são influenciados por vários fatores, é possível observar na Tabela 30 que o IAA, para algumas categorias, está diretamente relacionado com o f1-score, como o caso das categorias "Observações" e "Histórico do Paciente", onde o baixo IAA ocasionou em um f1-score igualmente ruim, enquanto que, para outras, essa relação não se aplica, como é o caso de "Histórico Familiar", onde mesmo com um IAA de 1 o F1-score ficou baixo, devido aos outros tipos de relação que interferem na qualidade. Na tabela, o valor do IAA corresponde ao IAA Relaxado presente na Tabela 30.

Tabela 30 – Comparação entre IAA e F1-Score de cada Categoria

Categoria REN	**IAA**	**F1-Score**
Comportamento Autodestrutivo	0,863	0,57 ± 0,06
Diagnostico	0,778	0,85 ± 0,03
Drogas	0,694	0,50 ± 0,06
Fármacos	0,913	0,95 ± 0,01
Função Psíquica	0,689	0,74 ± 0,03
Histórico Familiar	1	0,35 ± 0,06
Histórico do Paciente	0,466	0,32 ± 0,07
Observações	0,284	0,09 ± 0,01
Sintomas e Queixas Psíquicas	0,456	0,56 ± 0,06

Um dos motivos da IAA de algumas categorias terem sido ruins foi a definição destas categorias no *guideline*. Categorias como "Observações" acabaram ficando com o critério de análise mais subjetivo do que objetivo, o que foi um grande problema para a qualidade da anotação desta categoria. Ou seja, se o *guideline* tivesse deixado os limites entre as entidades mais bem definidos e diminuído a subjetividade para a anotação das entidades, possivelmente o IAA teria sido melhor e, consequentemente, os resultados obtidos pelo modelo também.

Capítulo 8 – Conclusão e Trabalhos Futuros

Neste trabalho foi reunido um conjunto de 300 documentos de prontuários de admissões de um serviço de psiquiatria, de modo a formar um corpus anotado e de acesso público: O segundo corpus anotado com entidades clínicas em português do Brasil e o primeiro especializado em psiquiatria. Foi elaborado um *guideline* para anotação de entidades com foco em extração de informações importantes ao domínio da psiquiatria e então esses documentos foram todos anotados, dando origem também ao primeiro corpus anotado para REN com dados psiquiátricos na língua portuguesa, que é um dos objetivos e contribuições do presente trabalho.

Também foi utilizado o corpus anotado para treinar modelos de REN clínico a partir da aplicação do *fine tunning* no BERT, o que resultou em um modelo de REN capaz de extrair informações sobre Comportamento Autodestrutivo, Diagnósticos, Drogas, Fármacos, Funções Psíquicas, Histórico Familiar, Observações e Sintomas e Queixas Psíquicas, com valores médios de precisão 0,64, revocação 0,62 e f1-Score 0,66. O modelo treinado foi disponibilizado publicamente na biblioteca *HuggingFace*, com a chave psybertpt/psyBERTpt[15]. O treinamento e disponibilização do modelo também é um dos objetivos e contribuições do presente trabalho.

Considerando que os objetivos relativos ao *guideline*, ao corpus e ao modelo, pode-se concluir que os objetivos pretendidos no presente trabalho foram alcançados.

Embora as anotações da categoria de Funções Psíquicas tenham sido feitas com especificadores, o modelo de REN foi treinado considerando apenas a categoria sem nenhuma especificação. Um possível ponto de melhoria a ser aplicado neste trabalho é treinar modelos de REN que também sejam capazes de classificar as entidades de acordo com os especificadores das funções psíquicas.

[15] https://huggingface.co/psybertpt/psyBERTpt

Outras possíveis melhorias são a respeito da qualidade das anotações. Para atingir melhores resultados no modelo, recomenda-se uma revisão e correção das anotações, principalmente as que tiveram o IAA relaxado abaixo de 0.5, como as Observações, Histórico do Paciente e Funções Psíquicas. Além disso, palavras que indiquem a negação de uma entidade, como "nega (...)", devem deixar de ser anotadas dentro da entidade, porque são comuns a várias categorias e isso faz que o modelo possa gerar confusão.

O psyBERTpt foi publicado no 36th International Symposium on Computer Based Medical Systems (CBMS) com o DOI 10.1109/CBMS58004.2023.00298 e pode ser acessado pelo link: https://ieeexplore.ieee.org/abstract/document/10178727

Referências

ALIOD MOLL, D. et al. Answer extraction from technical texts. **IEEE Intelligent Systems**, v. 18, n. 4, p. 12-17, 2003.

ALSENTZER, Emily et al. Publicly available clinical BERT embeddings. **CoRR**, abs/1904.03323, 2019. Disponível em: <http://arxiv.org/abs/1904.03323>.

ALSHAMMARI, Nasser; ALANAZI, Saad. The impact of using different annotation schemes on named entity recognition. **Egyptian Informatics Journal**, v. 22, n. 3, p. 295-302, 2021.

ALLVIN, Helen et al. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. In: **Journal of Biomedical Semantics**. BioMed Central, 2011. p. 1-11.

AMARAL, Daniela Oliveira Ferreira do. **O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa**. 2013. Dissertação de Mestrado. Pontifícia Universidade Católica do Rio Grande do Sul.

ANDRADE, Gabriel Herman Bernardim; OLIVEIRA, Lucas Emanuel Silva; MORO, Claudia Maria Cabral. Metodologias e ferramentas para anotação de narrativas clínicas. **J. health inform**, p. 1031-1040, 2016.

AONE, C. et al. A trainable summarizer with knowledge acquired from robust nlp techniques. Advances in Automatic Text Summarization. 1999.

ARONSON, Alan R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: **Proceedings of the AMIA Symposium**. American Medical Informatics Association, 2001. p. 17.

BABYCH, Bogdan; HARTLEY, Anthony. Improving machine translation quality with automatic named entity recognition. In: **Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003**. 2003.

BIKEL, Daniel M. et al. Nymble: a high-performance learning name-finder. **arXiv preprint cmp-lg/9803003**, 1998.

BODENREIDER, Olivier. The unified medical language system (UMLS): integrating biomedical terminology. **Nucleic acids research**, v. 32, n. suppl_1, p. D267-D270, 2004.

BORTHWICK, Andrew et al. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In: **Sixth Workshop on Very Large Corpora**. 1998.

BRASIL, Lei nº 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 18/07/2021.

CANTOR, Michael N. et al. An Evaluation of Hybrid Methods for Matching Biomedical Terminologies: Mapping the Gene Ontology to the UMLS®. **Studies in health technology and informatics**, v. 95, p. 62, 2003.

CARLSON, Gregory Norman. **Reference to kinds in English**. University of Massachusetts Amherst, 1977.

CHATURVEDI, Jaya et al. Development of a Corpus Annotated with Medications and their Attributes in Psychiatric Health Records. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. 2020. p. 2009-2016.

CHENG, Pengxiang; ERK, Katrin. Attending to entities for better text understanding. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. 2020. p. 7554-7561.

CHIU, Jason PC; NICHOLS, Eric. Named entity recognition with bidirectional LSTM-CNNs. **Transactions of the Association for Computational Linguistics**, v. 4, p. 357-370, 2016.

COLLINS, Michael. Ranking algorithms for named entity extraction: Boosting and the votedperceptron. In: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. 2002. p. 489-496.

Conselho Federal de Medicina (CFM – Brasil). Resolução nº 1638/2002. Disponível em https://sistemas.cfm.org.br/normas/visualizar/resolucoes/BR/2002/1638. Acesso em 18 de julho de 2021.

Conselho Federal de Medicina (CFM – Brasil). Resolução nº 196/1996. Disponível em https://bvsms.saude.gov.br/bvs/saudelegis/cns/1996/res0196_10_10_1996.html. Acesso em 18 de julho de 2021.

CONSOLI, Bernardo et al. BRATECA (Brazilian Tertiary Care Dataset): a Clinical Information Dataset for the Portuguese Language.

DA ROCHA, Naila Camila et al. Natural Language Processing to Extract Information from Portuguese-Language Medical Records. **Data**, v. 8, n. 1, p. 11, 2022.

DALGALARRONDO, Paulo. **Psicopatologia e semiologia dos transtornos mentais**. Artmed Editora, 2018.

DALIANIS, Hercules. **Clinical text mining: Secondary use of electronic patient records**. Springer Nature, 2018.

DALIANIS, Hercules; HASSEL, Martin; VELUPILLAI, Sumithra. The Stockholm EPR Corpus-characteristics and some initial findings. **Proceedings of ISHIMR**, p. 243-249, 2009.

DE CARVALHO, Ricardo. **Aplicação de técnicas de mineração de texto na recuperação de informação clínica em prontuário eletrônico do paciente.** 2017. Dissertação de Mestrado. Universidade Estadual Paulista "Júlio de Mesquita Filho" - Faculdade de Filosofia e Ciência.

DE SOUZA, João Vitor Andrioli et al. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. In: **Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. SBC, 2019. p. 318-323.

DEVLIN, Jacob et al. Bert: Pre-training of deep bidirectional transformers for language understanding, **CoRR**, abs/1810.04805, 2018. Disponível em: <http://arxiv.org/abs/1810.04805>.

DOGAN, Cihan et al. Fine-grained named entity recognition using elmo and wikidata. **CoRR**, abs/1904.10503, 2019. Disponível em: <http://arxiv.org/abs/1904.10503>.

Ellendorff, Tilia Renate; Foster, Simon; Rinaldi, Fabio (2016). *The PsyMine Corpus - A Corpus annotated with Psychiatric Disorders and their Etiological Factors.* In: **10th edition of the Language Resources and Evaluation Conference,** 2016. European Language Resources Association (ELRA), 3723-3729.

EHRENTRAUT, Claudia et al. Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. **A machine learning approach using Naïve Bayes, Support Vector Machines and C**, v. 4, p. 5, 2012.

ETZIONI, Oren et al. Unsupervised named-entity extraction from the web: An experimental study. **Artificial intelligence**, v. 165, n. 1, p. 91-134, 2005.

FERREIRA, Liliana; TEIXEIRA, António; DA SILVA CUNHA, Joao Paulo. Information extraction from Portuguese hospital discharge letters. **Evolution**, v. 8, n. 998, p. 506, 2010.

FERRUCCI, David; LALLY, Adam. UIMA: an architectural approach to unstructured information processing in the corporate research environment. **Natural Language Engineering**, v. 10, n. 3-4, p. 327-348, 2004.

Friedman C, Hripcsak G, Shablinsky I. An evaluation of natural language processing methodologies. AMIA Annu Symp Proc; 1998:855–9

GHADDAR, Abbas; LANGLAIS, Philippe. Robust lexical features for improved neural network named-entity recognition. **CoRR**, abs/1806.03489, 2018. Disponível em: <http://arxiv.org/abs/1806.03489>.

GOLDBERG, Yoav; NIVRE, Joakim. A dynamic oracle for arc-eager dependency parsing. In: **Proceedings of COLING 2012**. 2012. p. 959-976.

GRAVE, Edouard et al. Learning word vectors for 157 languages. **arXiv preprint arXiv:1802.06893**, 2018.

GUO, Jiafeng et al. Named entity recognition in query. In: **Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval**. 2009. p. 267-274.

HEARST, Marti A. et al. Support vector machines. **IEEE Intelligent Systems and their applications**, v. 13, n. 4, p. 18-28, 1998.

HOFER, Maximilian et al. Few-shot learning for named entity recognition in medical text. **CoRR**, abs/1811.05468, 2018. Disponível em: <http://arxiv.org/abs/1811.05468>.

HOVY, Eduard; LAVID, Julia. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. **International journal of translation**, v. 22, n. 1, p. 13-36, 2010.

HUANG, Ming-Siang et al. Biomedical named entity recognition and linking corpus: survey and our recent development. **Briefings in Bioinformatics**, v. 21, n. 6, p. 2219-2238, 2020.

HUANG, Zhiheng; XU, Wei; YU, Kai. Bidirectional LSTM-CRF models for sequence tagging. **CoRR**, abs/ 1508.01991, 2015. Disponível em: <http://arxiv.org/abs/1508.01991>.

I2b2: Informatics for Integrating Biology & the Bedside. Disponível em https://www.i2b2.org/. Acessado em 20/05/2021.

JOHNSON, Alistair EW et al. MIMIC-III, a freely accessible critical care database. **Scientific data**, v. 3, n. 1, p. 1-9, 2016.

JURAFSKY, Dan; MARTIN, James H. Speech and Language Processing (3rd ed. draft). 3rd ed. draft. 2023. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 06/08/2023.

KILGARRIFF, Adam. Gold standard datasets for evaluating word sense disambiguation programs. **Computer Speech & Language**, v. 12, n. 4, p. 453-472, 1998.

KIM, J.-D. et al. GENIA corpus—a semantically annotated corpus for bio-textmining. **Bioinformatics**, v. 19, n. suppl_1, p. i180-i182, 2003.

KLIE, Jan-Christoph et al. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: **Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations**. 2018. p. 5-9.

Konkol M, Konopík M, Segment representations in named entity recognition. In: **International conference on text, speech, and dialogue**. Springer; 2015. P. 61–70.

KRISHNAN, Vijay; GANAPATHY, Vignesh. Named entity recognition. **Stanford Lecture CS229**, 2005.

KURU, Onur; CAN, Ozan Arkan; YURET, Deniz. Charner: Character-level named entity recognition. In: **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. 2016. p. 911-921.

LAFFERTY, John; MCCALLUM, Andrew; PEREIRA, Fernando CN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

LAMPLE, Guillaume et al. Neural architectures for named entity recognition. **CoRR**, abs/1603.01360, 2016. Disponível em: <http://arxiv.org/abs/1603.01360>.

LI, Jiao et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. **Database**, v. 2016, 2016.

LI, Jing et al. A survey on deep learning for named entity recognition. **IEEE Transactions on Knowledge and Data Engineering**, 2020.

LIN, Chen et al. A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In: **Proceedings of the 2nd Clinical Natural Language Processing Workshop**. 2019. p. 65-71.

LINDBERG, Donald AB; HUMPHREYS, Betsy L.; MCCRAY, Alexa T. The unified medical language system. **Yearbook of Medical Informatics**, v. 2, n. 01, p. 41-51, 1993.

LIU, Tianyu; YAO, Jin-Ge; LIN, Chin-Yew. Towards improving neural named entity recognition with gazetteers. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. 2019. p. 5301-5307.

LOPES, Fábio; TEIXEIRA, César; OLIVEIRA, Hugo Gonçalo. Contributions to clinical named entity recognition in Portuguese. In: **Proceedings of the 18th BioNLP Workshop and Shared Task**. 2019. p. 223-233.

MAIOR, Marta da Cunha Lobo Souto; OSORIO-DE-CASTRO, Claudia Garcia Serpa; ANDRADE, Carla Lourenço Tavares de. Internações por intoxicações medicamentosas em crianças menores de cinco anos no Brasil, 2003-2012. **Epidemiologia e Serviços de Saúde**, v. 26, p. 771-782, 2017.

MANNING, Christopher; SCHUTZE, Hinrich. **Foundations of statistical natural language processing**. MIT press, 1999.

MCCALLUM, Andrew; LI, Wei. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. 2003.

MCCLOSKY, David; CHARNIAK, Eugene. Self-training for biomedical parsing. In: **Proceedings of ACL-08: HLT, Short Papers**. 2008. p. 101-104.

MICHALOPOULOS, George et al. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. **CoRR**, abs/2010.10391, 2020. Disponível em:

<http://arxiv.org/abs/2010.10391>.

MOON, Taesun et al. Towards lingua franca named entity recognition with bert. arXiv preprint **arXiv:1912.01389,** 2019.

MUKHERJEE, Sankha S. et al. Natural language Processing-Based Quantification of the mental state of psychiatric patients. **Computational Psychiatry**, v. 4, 2020.

MUNIKAR, Manish; SHAKYA, Sushil; SHRESTHA, Aakash. Fine-grained sentiment classification using BERT. In: **2019 Artificial Intelligence for Transforming Business and Society (AITB).** IEEE, 2019. p. 1-5.

MURTY, Shikhar et al. Hierarchical losses and new resources for fine-grained entity typing and linking. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).** 2018. p. 97-109.

NADEAU, David; SEKINE, Satoshi. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, v. 30, n. 1, p. 3-26, 2007.

NEUMANN, Mark et al. Scispacy: Fast and robust models for biomedical natural language processing. **CoRR**, abs/1902.07669, 2019. Disponível em: <http://arxiv.org/abs/1902.07669>

NIERO, Luiz et al. Challenges and Issues on Extracting named Entities from Oncology Clinical Notes. In: **Anais do XIX Congresso Brasileiro de Informática em Saúde. 2022.**

OLIVEIRA, L. E. S. et al. SemClinBr–a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks. **CoRR**, 2020.

PAKHOMOV, Serguei V.; CODEN, Anni; CHUTE, Christopher G. Developing a corpus of clinical notes manually annotated for part-of-speech. I**nternational journal of medical informatics**, v. 75, n. 6, p. 418-429, 2006.

PATEL, Pinal et al. Annotation of a large clinical entity corpus. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. 2018. p. 2033-2042.

PETERS, Ana Carolina et al. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta Hospitalar. In: **Anais do XII Congresso Brasileiro de Informática em Saúde**. 2010.

PETERS, Matthew E. et al. Deep contextualized word representations. CoRR abs/1802.05365 (2018). **arXiv preprint arXiv:1802.05365**, 1802.

PETKOVA, Desislava; CROFT, W. Bruce. Proximity-based document representation for named entity retrieval. In: **Proceedings of the sixteenth ACM conference on Conference on information and knowledge management**. 2007. p. 731-740.

PRADHAN, Sameer et al. CoNLL-2012 shared task: Modeling multilingual

unrestricted coreference in OntoNotes. In: **Joint Conference on EMNLP and CoNLL-Shared Task**. 2012. p. 1-40.

PRADHAN, Sameer et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. In: **Journal of the American Medical Informatics Association**, v. 22, n. 1, p. 143-154, 2015.

RADFORD, Alec et al. Improving language understanding by generative pre-training. 2018.

REÁTEGUI, Ruth; RATTÉ, Sylvie. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. **BMC medical informatics and decision making**, v. 18, n. 3, p. 13-19, 2018.

ROBERTS, Angus et al. Building a semantically annotated corpus of clinical texts. **Journal of biomedical informatics**, v. 42, n. 5, p. 950-966, 2009.

ROCHA, Naila Camila da. **Uso de redes neurais artificiais para extração de dados de prontuários médicos**. 2021. Dissertação de Mestrado. Universidade Estadual Paulista "Júlio de Mesquita Filho" - Instituto de Biociências de Botucatu.

SANG, Erik F.; BUCHHOLZ, Sabine. Introduction to the CoNLL-2000 shared task: Chunking. **arXiv preprint cs/0009008,** 2000.

SAVOVA, Guergana K. et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. **Journal of the American Medical Informatics Association**, v. 17, n. 5, p. 507-513, 2010.

SCHWARTZ, Ariel S.; HEARST, Marti A. A simple algorithm for identifying abbreviation definitions in biomedical text. In: **Biocomputing 2003**. 2002. p. 451-462.

SEKINE, Satoshi. Named entity: History and future. **Project notes, New York University**, p. 4, 2004.

SEKINE, Satoshi; GRISHMAN, Ralph; SHINNOU, Hiroyuki. A decision tree method for finding and classifying names in Japanese texts. In: **Sixth workshop on very large corpora**. 1998.

SEKINE, Satoshi; ISAHARA, Hitoshi. IREX: IR & IE Evaluation Project in Japanese. In: **LREC**. 2000. p. 1977-1980.

SEKINE, Satoshi; SUDO, Kiyoshi; NOBATA, Chikashi. Extended named entity hierarchy. In: **LREC**. 2002.

SEYLER, Dominic et al. A study of the importance of external knowledge in the named entity recognition task. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. 2018. p. 241-246.

SINGH, A. K. et al. Multi-label natural language processing to identify diagnosis and

procedure codes from MIMIC-III inpatient notes. **CoRR**, abs/2003.07507, 2020. Disponível em: <http://arxiv.org/abs/2003.07507>.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **Brazilian Conference on Intelligent Systems**. Springer, Cham, 2020. p. 403-417.

SRIDHAR, Srinivasan et al. Psychiatry Transcript Annotation: Process Study and Improvements. In: **Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care**. Sage CA: Los Angeles, CA: SAGE Publications, 2021. p. 71-75.

STYLER, William F. et al. Temporal annotation in the clinical domain. **Transactions of the association for computational linguistics**, v. 2, p. 143-154, 2014.

SUNDHEIM, Beth M. The message understanding conferences. In: **TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996**. 1996. p. 35-37.

TERUMI, Rubel Schneider Elisa et al. Biobertpt-a portuguese neural language model for clinical named entity recognition. In: **Proceedings of the 3rd Clinical Natural Language Processing Workshop**. Association for Computational Linguistics, 2020. p. 65-72.

THOMPSON, Deb; WRIGHT, Kim. **Developing a Unified Patient Record**: A Practical Guide. Oxon: Radcliffe Medical Press, 2003. 140 p. (Radcliffe Series).

TRAN, Quan; MACKINLAY, Andrew; YEPES, Antonio Jimeno. Named entity recognition with stack residual lstm and trainable bias decoding. **CoRR**, abs/1706.07598, 2017. Disponível em: <http://arxiv.org/abs/1706.07598>.

UMLS – Unified Medical Language System. Statistics –2021AA Release. Disponível em https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html. Acessado em 06/06/2021.

UZUNER, Özlem. Recognizing obesity and comorbidities in sparse data. **Journal of the American Medical Informatics Association**, v. 16, n. 4, p. 561-570, 2009.

WANG, Xiaozhi et al. KEPLER: A unified model for knowledge embedding and pre-trained language representation. **Transactions of the Association for Computational Linguistics**, v. 9, p. 176-194, 2021.

WOLF, Thomas et al. Huggingface's transformers: State-of-the-art natural language processing. **arXiv preprint arXiv:1910.03771**, 2019.

YADAV, Vikas; BETHARD, Steven. A survey on recent advances in named entity recognition from deep learning models. **CoRR,** abs/1910.11470, 2019. Disponível em: <http://arxiv.org/abs/1910.11470>.

ZHANG, Zhengyan et al. ERNIE: Enhanced language representation with informative entities. **CoRR**, abs/1905.07129, 2019. Disponível em: <http://arxiv.org/abs/1905.07129>.