

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE**

LUCAS BREHM RONNAU

**MapClin: MAPEAMENTO AUTOMÁTICO DE TERMOS CLÍNICOS EM
PORTUGUÊS PARA A SNOMED CT**

CURITIBA

2019

LUCAS BREHM RONNAU

**MapClin: MAPEAMENTO AUTOMÁTICO ENTRE TERMOS CLÍNICOS EM
PORTUGUÊS E A SNOMED CT**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Tecnologia em Saúde da Escola Politécnica da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Tecnologia em Saúde.

Linha de Pesquisa: Informática em Saúde

Orientadora: Prof.^a Dr.^a Claudia M^a C. Moro Barra

Coorientadora: Prof.^a Dr.^a Marcia Regina Cubas

CURITIBA

2019

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Edilene de Oliveira dos Santos CRB 9 /1636

R773m
2019 Ronnau, Lucas Brehm
MapClin : mapeamento automático entre termos clínicos em português e a
Snomed CT / Lucas Brehm Ronnau ; orientadora, Claudia M^a C. Moro Barra ;
coorientadora, Marcia Regina Cubas. -- 2019
90 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná,
Curitiba, 2019
Bibliografia: f.66-69

1. Processamento de linguagem natural. 2. Medicina - Terminologia. 3.
Medicina – Linguagem. 4. Sistemas de informação em saúde. 5. Documentos
eletrônicos. I. Barra, Claudia Maria Cabral Moro, 1969-. II. Cubas, Marcia Regina.
III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação
em Informática. IV. Título.

CDD 20. ed. – 004



Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós Graduação em Tecnologia em Saúde

**ATA DE DEFESA DE DISSERTAÇÃO DE Mestrado
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE**

DEFESA DE DISSERTAÇÃO Nº 266

ÁREA DE CONCENTRAÇÃO: TECNOLOGIA EM SAÚDE

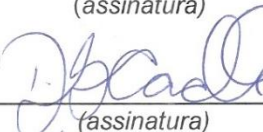
Aos vinte e oito dias do mês de fevereiro de 2019 às 12:00h no Auditório, Guglielmo Marconi, Térreo-Bloco 8 (Parque Tecnológico - Bloco Elétrica), realizou-se a sessão pública de Defesa da Dissertação: **“MapClin: MAPEAMENTO AUTOMÁTICO ENTRE TERMOS CLÍNICOS EM PORTUGUÊS E A SNOMED CT, APOIADO NA UMLS”** apresentado pelo aluno Lucas Brehm Ronnau sob orientação da Prof. Dr. Claudia Maria Cabral Moro Barra e coorientação da Prof. Dr. Profª. Drª. Marcia Regina Cubas como requisito parcial para a obtenção do título de Mestre em Tecnologia em Saúde, perante uma Banca Examinadora composta pelos seguintes membros:

Prof. Dr. Claudia Maria Cabral Moro Barra
PUCPR (Presidente)


(assinatura)

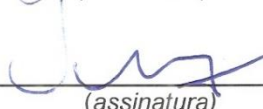

(Aprov/Reprov.)

Prof. Dr. Deborah Ribeiro Carvalho
PUCPR (Examinador)


(assinatura)


(Aprov/Reprov.)

Prof. Dr. Stefan Schulz
Medical University of Graz (Examinador)


(assinatura)


(Aprov/Reprov.)


Início: 12:00 Término: 13:45

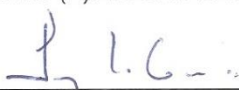
Conforme as normas regimentais do PPGTS e da PUCPR, o trabalho apresentado foi considerado aprovado (aprovado/reprovado), segundo avaliação da maioria dos membros desta Banca Examinadora.

Observações: _____

O(a) aluno(a) está ciente que a homologação deste resultado está condicionada: (I) ao cumprimento integral das solicitações da Banca Examinadora, que determina um prazo de 60 dias para o cumprimento dos requisitos; (II) entrega da dissertação em conformidade com as normas especificadas no Regulamento do PPGTS/PUCPR; (III) entrega da documentação necessária para elaboração do Diploma.

ALUNO (A): LUCAS BREHM RONNAU


(assinatura)


Prof. Dr. Percy Nohama,
Coordenador do PPGTS PUCPR



RESUMO

Introdução: Para o compartilhamento das informações contidas nos textos das narrativas dos registros eletrônico em saúde (RES) é fundamental que estas sejam representadas por uma terminologia padronizada. A SNOMED CT (SCT) é considerada a principal e mais abrangente terminologia clínica, e foi definida pelo Governo do Brasil como padrão para compartilhamento de informações entre RES. Porém, ainda não está disponível em português e não existe um método de mapeamento para este idioma. **Objetivo:** Propor um método de mapeamento para a SCT de termos clínicos escritos em português brasileiro na forma de linguagem natural extraídos de narrativas de registros eletrônicos em saúde. **Método:** A Unified Medical Language System (UMLS), por conter relações com outras terminologias e conceitos em português, foi utilizada como meio para relacionar os termos clínicos a SCT. Para a associação dos termos aos conceitos UMLS foram definidas regras de comparação utilizando abordagens de processamento de linguagem natural (PLN). Geralmente não é possível uma comparação direta com os conceitos UMLS pois os termos apresentam variações ortográficas. Para a criação das regras de comparação e mapeamento foi aplicado um método incremental em espiral, composto por 4 fases: proposta ou melhora de uma regra, elaboração ou adaptação da regra, teste com a base de treinamento e avaliação dos resultados do mapeamento. Após a identificação do conceito UMLS, a associação com a SCT foi realizada a partir da relação UMLS-SCT existente. A base de treinamento foi composta por todos os 216 termos clínicos (136 simples e 80 compostos) presentes em 5 narrativas clínicas selecionadas aleatoriamente de 1.030 textos do RES de três hospitais que compõem um corpus com anotação semântica. O método proposto, MapClin, foi aplicado em 200 termos clínicos (100 simples e 100 compostos) mais frequentes deste corpus, que representam 35% do total de termos simples e 41% dos compostos. Os resultados do mapeamento foram validados por três especialistas da área da saúde. **Resultados:** O MapClin é composto por uma etapa de pré-processamento, uma etapa constituída por sete regras de comparação termo-UMLS, e uma terceira etapa de associação do código de identificação da UMLS com a SCT. Após estas três etapas, o profissional de saúde seleciona quais conceitos SCT representam o termo clínico. O pré-processamento consiste na expansão dos acrônimos e abreviaturas dos termos clínicos, e posterior normalização (eliminação de acentuação e alteração para letras minúsculas). A seguir, é verificado se cada termo possui correspondente léxico igual na UMLS, primeira regra. Caso não seja identificado um correspondente, são aplicadas mais 6 regras de PLN que modificam lexicamente o termo. Estas regras incluem utilização do lema, radical e sinônimos dos termos, mapeamento para um conceito mais abrangente (pai) ou mais específico (filho) na UMLS; e tradução para o inglês. A partir do código do conceito UMLS identificado para o termo, são aplicadas 4 regras de associação (UMLS - SCT) para determinar um ou mais conceitos SCT; se o conceito SCT é obsoleto e existe uma relação de equivalência, ou se não existe conceito SCT associado. Dos 200 termos mais frequentes, 141 resultaram *exact match* na validação, apresentando 89,8% de *precision*, 90,8% de *recall* e 90,3 % de *F-score*. **Conclusão:** O MapClin possibilita o mapeamento automático entre termos clínicos em português e a SCT, apoiando e facilitando a tarefa do profissional de saúde.

Palavras-chave: *Unified Medical Language System. Systematized Nomenclature of Medicine CT. Processamento de Linguagem Natural. Interoperabilidade.*

ABSTRACT

MapClin - AUTOMATIC MAPPING BETWEEN CLINICAL TERMS IN PORTUGUESE AND SNOMED CT, BASED ON UMLS

Introduction: To share information contained in clinical narratives of Electronic Health Record (EHR), it is essential that they be represented by a standardized terminology. SNOMED CT (SCT) is considered the main and most comprehensive clinical terminology and was defined by the Brazilian government as one of interoperability standard for EHR. However, is not available in Portuguese and there is no mapping method for this language. **Objective:** Propose a mapping method between clinical terms of narratives in Portuguese and the SCT. **Method:** As the Unified Medical Language System (UMLS) has relations with other terminologies and concepts in Portuguese, was used as a strategy to relate clinical terms to SCT. Comparison rules for the association of terms with UMLS concepts were defined using Natural Language Processing (NLP) approaches. A direct comparison with UMLS concepts generally is not possible since the terms have orthographic variations. Rules of comparison and mapping were created applying an incremental spiral method with 4 phases: proposal or improvement of a rule, elaboration or adaptation of a rule, test with the training base and evaluation of mapping results. The association with the SCT was performed from the existing UMLS-SCT relation after the identification of the UMLS concept. The training base was composed of all 216 clinical terms (136 simple and 80 compounds) of 5 clinical narratives randomly selected from a corpus with semantic annotation of 1,030 EHR texts from three hospitals. The proposed method, MapClin, was applied in 200 clinical terms (100 simple and 100 compounds) more frequent in this corpus, representing 35% of the total of simple terms and 41% of the compounds. Mapping results were validated by three health experts. **Results:** MapClin consists of a preprocessing phase, one with seven UMLS-term comparison rules, and a third phase to associate the UMLS identification code to SCT. After these three phases, the health professional selects which SCT concepts represent the clinical term. Preprocessing consists of expanding the acronyms and abbreviations of clinical terms, and their normalization (elimination of accentuation and change to lower case letters). Next, it is verified if each term has an equal correspondent lexicon in UMLS, first rule. If a correspondent is not identified, others 6 NLP rules that modify lexically the term are applied. These rules include use of the lemma, stem and synonym of the terms; mapping to a more comprehensive (parent) or more specific (child) concept in the UMLS; and translation into English. From each UMLS concept code identified for the term, 4 association rules (UMLS - SCT) are applied to determine one or more SCT concepts; if the SCT concept is obsolete and it has an equivalence relation, or if there is no associated SCT concept. During the validation, 141 of most frequent terms were mapped with an exact match, resulting in 89.8% of accuracy, 90.8% of recall and F-score of 90.3%. **Conclusion:** MapClin enables automatic mapping between clinical terms in Portuguese and SCT, supporting and facilitating the health professional's task.

Keywords: Unified Medical Language System. Systematized Nomenclature of Medicine. Natural Language Processing. Health Information Interoperability.

LISTA DE FIGURAS

Figura 1 - Os 3 recursos de conhecimentos da UMLS.	16
Figura 2 - A abrangência da CUI, LUI, SUI e AUI.	18
Figura 3 - A abrangência da CUI, LUI, SUI e AUI.	19
Figura 4 - Terminologias de origem dos termos em português da UMLS.	22
Figura 5 - Hierarquias da SCT.	23
Figura 6 - Relações hierárquicas do conceito 'apendicectomia'.	Erro! Indicador não definido.
Figura 7 - Atributos que podem ser usados na hierarquia 'achados clínicos'.	Erro! Indicador não definido.
Figura 8 - Atributos que podem ser usados na hierarquia 'procedimentos'.	Erro! Indicador não definido.
Figura 9 - Execução do algoritmo de Levenshtein.	29
Figura 10 - Variantes para o termo "ocular".	31
Figura 11 - Exemplo de anotação feita nas narrativas clínicas.	34
Figura 12 - Extração de termo, com a expansão da abreviatura e a inserção de etiquetas semânticas.	34
Figura 13 - MapClin.	40
Figura 14 - Diagrama de casos de uso.	41
Figura 15 - diagrama de iteração.	42
Figura 16 - diagrama de atividades.	43
Figura 17 - Pré-processamento do termo "IAM".	44
Figura 18 - Representação da Regra Direta para o termo "torax".	46
Figura 19 - Representação da Regra Lematização para o termo "anormal".	46
Figura 20 - Representação da Regra Stemização para o termo "leito".	47
Figura 21 - Representação para Regra Sinônimos para o termo "face corada".	48
Figura 22 - Representação da Regra Pai para o termo "MSD".	49
Figura 23 - Representação da Regra Filho para o termo "Tibias".	50
Figura 24 - Representação da Regra Inglês para o termo "lucido".	51
Figura 25 - Procedimento quando houver apenas um conceito SCT associado.	51
Figura 26 - Procedimento quando houver mais de um conceito SCT associado.	52
Figura 27 - Procedimento quando houver conceitos SCT associado obsoletos ou desativados.	53
Figura 28 - Procedimento quando houver nenhum conceito SCT associado.	53
Figura 29 - Recall, Precision e F- score para "todos os termos", "termos simples" e "termos compostos".	55
Figura 30 - Frequência de termos simples escolhidos.	56
Figura 31 - Frequência de termos compostos escolhidos.	56
Figura 32 - Precision, recall e F- score para o experimento final.	58
Figura 33 - Frequência de termos simples escolhidos.	58
Figura 34 - Frequência de termos compostos escolhidos.	59

LISTA DE QUADROS

Quadro 1 – Descrição do conteúdo das colunas da tabela MRCONSO	17
Quadro 2 – Descrição do conteúdo das colunas da tabela MRREL	20
Quadro 3 – Descrição do conteúdo das colunas da tabela MRXW_POR.....	21
Quadro 4 – Valor de distância das variantes.....	30
Quadro 5 – Lista de abreviaturas e acrónimos com suas expansões.....	35
Quadro 6 – Número de termos validados por cada.....	39
Quadro 7 – Stop Words removidas dos termos de entrada.....	44
Quadro 8 – Exemplo de seleção	54
Quadro 9 - Comparação entre diferentes métodos de mapeamento.....	63

LISTA DE ABREVIATURAS E SIGLAS

API	<i>Application Programming Interface</i>
AUI	Atom Unique Identifiers
CID	Classificação Internacional de Doenças
CIPE	Classificação Internacional para a Prática de Enfermagem
CUI	<i>Unique Concept Identifier</i>
DescriptionID	<i>Description Identifier</i>
FSN	<i>Fully Specified Name</i>
ICPC	<i>International Classification of Primary Care</i>
IHTSDO	<i>International Health Terminology Standart Development Organization</i>
LUI	<i>Lexical (term) Unique Identifiers</i>
LOINC	<i>Logical Observation Identifiers Names and Codes</i>
MedRA	<i>Medical Dictionary for Regulatory Activities</i>
MeSH	<i>Medical Subject Headings</i>
NLM	<i>National Library of Medicine</i>
PUCPR	Pontifícia Universidade Católica do Paraná
Pt-BR	Português brasileiro
PLN	Processamento de Linguagem Natural
PPGTS	Programa de Pós-Graduação em Tecnologia em Saúde
RefSets	<i>Reference sets</i>
RES	Registro Eletrônico em Saúde
SUI	String Unique Identifiers
SCT	<i>SNOMED CT</i>
SCTID	<i>SNOMED CT concept identifier</i>
UMLS	<i>Unified Medical Languages System</i>
RUI	<i>Unique Relationship Identifier</i>
WHO	Word Health Organization terminology
TF-IDF	<i>Term Frequency / Inverse Document Frequency</i>

SUMÁRIO

1. INTRODUÇÃO	10
1.1 OBJETIVOS	13
1.1.1 Objetivo Geral	13
1.1.2 Objetivos Específicos	13
1.2 CONTRIBUIÇÕES	14
1.2.1 Contribuições Científicas	14
1.2.2 Contribuições Sociais	14
2. REFERENCIAL TEÓRICO	15
2.1 TERMINOLOGIAS CLÍNICAS	15
2.1.1 UNIFIED MEDICAL LANGUAGE SYSTEM (UMLS)	15
2.1.2 SNOMED CT	22
2.2 MAPEAMENTO TERMINOLÓGICO	27
2.3 METAMAP	30
2.4 MÉTRICAS	32
3. METODOLOGIA	33
3.1. LOCAL DO ESTUDO	33
3.2. PREPARAÇÃO DA BASE DE DADOS	33
3.2.1. ASPECTOS ÉTICOS	36
3.3. ENCAMINHAMENTO METODOLÓGICO	36
4. RESULTADOS	40
4.3. ETAPA 1 – PRÉ-PROCESSAMENTO	44
4.3.1. FASE 2 – MAPEAMENTO DA UMLS PARA A SNOMED CT	51
4.4. ETAPA 3 – SELEÇÃO DOS TERMOS	53
4.5. RESULTADO EXPERIMENTO PARCIAL	55
4.6. RESULTADOS DA VALIDAÇÃO	57
5. ANÁLISES DOS RESULTADOS E DISCUSSÃO	60
5.1. LIMITAÇÕES DO ESTUDO	63
5.2. TRABALHOS FUTUROS	64

6. CONCLUSÃO.....	65
7. REFERÊNCIAS.....	66
APÊNDICE A – TERMOS COMPOSTOS (MAPEAMENTO EXATO).....	70
APÊNDICE B – TERMOS COMPOSTOS (MAPEAMENTO PARCIAL)	74
APÊNDICE C – TERMOS COMPOSTOS (NÃO MAPEADOS)	77
APÊNDICE D – TERMOS COMPOSTOS (NÃO MAPEADOS) MAPEAMENTO MANUAL	77
APÊNDICE E – TERMOS SIMPLES (MAPEAMENTO EXATO)	79
APÊNDICE F – TERMOS SIMPLES (MAPEAMENTO PARCIAL)	86
APÊNDICE G – TERMOS SIMPLES (NÃO MAPEADOS).....	87
APÊNDICE H – TERMOS SIMPLES (NÃO MAPEADOS) MAPEAMENTO MANUAL	87
ANEXO A - PARECER CONSUBSTANCIADO COMITÊ DE ÉTICA EM PESQUISA.....	88

1. INTRODUÇÃO

Com a crescente implantação de Registros Eletrônicos em Saúde (RES), muitos dados clínicos vêm sendo produzidos. Os objetivos traçados para o RES são: facilitar a comunicação de informações sobre pacientes entre os profissionais, permitir identificação de doenças, diagnósticos, processos, tratamentos e diferentes situações de um paciente específico, possibilitar o melhor uso desses dados para que pesquisadores descubram fatos relevantes ou estudem uma doença específica, mas também serem usados para melhorar o gerenciamento do cuidado (MARTINEZ SORIANO; PEÑA, 2018).

O grande auxílio desses dados, para sistemas de informação em saúde e para pesquisas, não seria possível sem a ampla disponibilidade e cobertura de conteúdo de terminologias, sistemas terminológicos e codificação padronizada (ALLONES; MARTINEZ; TABOADA, 2014; SAIWAL *et al.*, 2012).

A maior parte dos dados do RES é escrita em linguagem natural na forma de texto livre, chamados de narrativas clínicas, no entanto, é necessário mapeá-los para uma terminologia padronizada. Para que a interoperabilidade semântica seja possível, os dados devem ser representados ou mapeados para um mesmo sistema terminológico. O mapeamento pode ser realizado manualmente, contudo é recomendável utilizar auxílio computacional.

Um sistema terminológico é um modelo de conceitos e relações juntamente com os termos que lhes pertencem (KEIZER; ABU-HANNA; ZWETSLOOT-SCHONK, 2000). Segundo a Organização Internacional de Normalização (ISO, 2016), o conceito é uma unidade de conhecimento criada a partir de uma combinação de características, as quais são representadas pelos termos que descrevem em linguagem natural um conceito. Os sistemas terminológicos possibilitam interoperabilidade semântica dos dados (HUSSAIN *et al.*, 2014; KIM, 2016), o que permite a comunicação entre diferentes sistemas sem ambiguidade.

O mapeamento terminológico pode ser feito de maneira lexical ou semântica, o léxico leva em conta as propriedades de escrita dos termos, de modo que ocorre primeiro a normalização dos termos antes de compará-los (DHOMBRES; BODENREIDER, 2016; SIERRA *et al.*, 2015) para poder abranger as diferentes variantes ortográficas destes. Já o mapeamento semântico é baseado no significado das palavras e considera iguais os termos que representam o mesmo significado. Por

exemplo, se compararmos os termos “algia” e “dor” com o mapeamento léxico, esses não são reconhecidos como iguais, pois são escritos de maneira diferente. Em contrapartida, no mapeamento semântico, são considerados iguais, uma vez que representam o mesmo conceito.

A *Unified Medical Language System* (UMLS) é um sistema de terminologia biomédica que integra diversos vocabulários, ontologias e terminologias comumente usados em bases de dados de RES (BECKER *et al.*, 2017). A UMLS é constituída por três recursos de conhecimento: o *Metathesaurus*, a rede semântica e o léxico especializado (JOUBERT *et al.*, 2009). O *Metathesaurus* é composto por mais de 130 terminologias da saúde, unidas em uma única base, com conceitos relacionados semanticamente a outros com o mesmo significado por meio de identificadores únicos (NLM, 2018), o que torna este uma excelente ferramenta para mapeamento de termos da saúde.

O *Metathesaurus* contém, na versão 2017ab (versão do segundo semestre de 2017), 338.847 termos em português, incluindo o brasileiro e o europeu (Portugal), que representam, aproximadamente, 3% do total que compreendem a UMLS. Em inglês, são 8.836.759, representando 69% do total de termos. Além disso, 163.620 termos, em português, são representados pela terminologia de origem LOINC, que não possui associação com outras terminologias. Isso limita a utilização na tarefa de mapeamento de termos em português, excluindo 49% do total de termos nessa língua.

A SNOMED CT (SCT) é uma terminologia clínica para o uso em RES, considerada a maior terminologia clínica multilíngue em uso atualmente (HE; GELLER; CHEN, 2015; SNOMED CT, 2017), com mais de 300.000 conceitos associados a termos (ALLONES; MARTINEZ; TABOADA, 2014; HE; GELLER; CHEN, 2015). Foi desenvolvida nos Estados Unidos e no Reino Unido pelo Colégio Americano de Patologia e o Serviço Nacional de Saúde do Reino Unido (SNOMED CT, 2017) com a união do SNOMED RT e do *Clinical Terms version 3* (AL-HABLANI, 2017). Em 2007, foi adquirida pela International Health Terminology Standards Development Organization (IHTSDO) (BÁNFAI; PORCIÓ; KOVÁCS, 2014; LEE *et al.*, 2014; RAFIEI *et al.*, 2014), a qual mudou de nome para SNOMED *International* e gerencia e atualiza a SCT atualmente.

Além da SCT ser considerada a mais abrangente entre as terminologias da saúde, ainda possibilita o uso de expressões entre seus conceitos para formar outros mais específicos, unindo dois ou mais para formar um novo. Tais expressões são

conhecidas como expressões pós-coordenadas (DHOMBRES; BODENREIDER, 2016; LEE *et al.*, 2014; SNOMED CT, 2017). Isso faz com que a SCT consiga abranger praticamente todos os conceitos na área da saúde, sendo considerada a terminologia biomédica mais completa. Em 2014, 19 países já consideravam a SCT como terminologia clínica referência para o uso em RES (LEE *et al.*, 2014). Atualmente conta com 38 países membros (SNOMED CT, 2017).

No Brasil, a Portaria nº 2.073, de 31 de agosto de 2011, regulamenta que os sistemas de informação em saúde devem utilizar padrões de interoperabilidade e de informação em saúde indicando o uso da SCT para a codificação de termos clínicos e mapeamento das terminologias nacionais e internacionais em uso no País (MINISTÉRIO DA SAÚDE, 2011).

Para o inglês, existem diversas ferramentas que auxiliam no mapeamento de termos, como o MetaMap, um software desenvolvido para mapear textos em linguagem natural para a UMLS. O algoritmo MetaMap consiste em extrair termos dos textos e gerar sinônimos, variantes derivacionais, siglas, combinações significativas e variantes de inflexão e ortografia (Bousquet *et al.*, 2012; Shivade *et al.*, 2015). Essas variantes geradas são comparadas aos conceitos no *Metathesaurus* para construir um conjunto candidato ordenado segundo a força de mapeamento, uma medida de similaridade utilizada pelo MetaMap.

Atualmente a SCT está disponível em inglês, dinamarquês, espanhol, sueco e holandês (SNOMED CT, 2017). Apesar de existirem algumas iniciativas em inglês e em outros idiomas, a SCT não foi traduzida ou mapeada para uma terminologia em português e não existe uma técnica ou ferramenta específica para a tarefa de mapeamento desta em português. Na revisão realizada por LEE *et al.* (2014), é apresentado apenas um artigo do Brasil, o qual somente descreve a terminologia, não citando versão para a língua materna do País. No estudo de Khorrami; Ahmadi; Sheikhtaheri (2018), que descreve a cobertura de conteúdo da SCT, também não é apresentado relação ao Pt-BR.

Durante o desenvolvimento deste projeto, foi apresentada uma iniciativa de pesquisadores do Hospital Italiano de Buenos Aires (RENATO *et al.*, 2018), em que é explorada a possibilidade de elaborar um conjunto SCT de termos em português, a partir da tradução de conceitos disponíveis em espanhol. Porém, o ideal, segundo a SNOMED *International*, é construir o conjunto de termos em um determinado idioma a partir da real utilização destes na prática clínica (SNOMED CT, 2017). Os termos

em dinamarquês foram traduzidos a partir do inglês sem o mapeamento, de maneira que dificulta sua aplicação na saúde (RANDORFF; ELBERG; KJ, 2014).

Verificando a importância do uso da SCT e de terminologias clínicas, e comparando com a situação desta no cenário nacional, é gerada a seguinte questão norteadora: qual o método capaz de realizar o mapeamento de termos em linguagem natural em português brasileiro (Pt-BR) para conceitos da SCT?

Quanto à hipótese deste estudo: é possível elaborar um método capaz de mapear termos clínicos extraídos de texto livre em Pt-BR para um equivalente semântico na SCT automaticamente?

A recuperação de informações é uma tarefa que permite identificar ou extrair dados de RES. O grupo de pesquisa de Recuperação de Informações do PPGTS/PUCPR, no qual este estudo está inserido, vem desenvolvendo projetos relacionados a processamento de linguagem natural (PLN). A identificação de termos/conceitos clínicos contidos em narrativas clínicas elaboradas em texto livre é uma das tarefas realizadas. Para tanto, é fundamental o mapeamento dos termos contidos nas narrativas para terminologias padronizadas. Essa tarefa de mapeamento está relacionada ao projeto *Fine-Grained Text Mining for Clinical Trials* (FIGTEM) e a outro projeto denominado *Natural Language Processing for Portuguese Clinical Documents* (NLPPCD). Este é realizado em parceria com a *Philips Healthcare®* e tem por objetivo a construção de um *corpus* a partir da anotação semântica de mil narrativas dos principais domínios clínicos como Cardiologia, Oncologia, Clínica Médica, Obstetrícia, Nefrologia, dentre outros.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Propor um método de mapeamento para a SCT de termos clínicos escritos em português brasileiro na forma de linguagem natural extraídos de narrativas de registros eletrônicos em saúde.

1.1.2 Objetivos Específicos

- a) Definir um método que utilize a processamento de linguagem natural e UMLS no mapeamento de termos clínicos para conceitos da SCT.

- b) Mapear, com o método proposto, um conjunto de termos extraídos de textos não estruturados de evoluções clínicas e sumários de alta hospitalar.

1.2 CONTRIBUIÇÕES

1.2.1 Contribuições Científicas

Desenvolvimento de um método de mapeamento de termos na língua portuguesa para a SCT.

Uma ferramenta de mapeamento de termos em português brasileiro para a UMLS e a SCT.

1.2.2 Contribuições Sociais

O estudo contribui socialmente de forma indireta, ao proporcionar um método para padronizar os dados, o qual facilita a utilização destes em diversas tarefas que melhoram o atendimento ao paciente.

2. REFERENCIAL TEÓRICO

Neste Capítulo, são apresentados os conhecimentos necessários para o desenvolvimento deste estudo, dentre quais: Terminologias Clínicas, a UMLS, a SCT, mapeamento terminológico, MetaMap e as Métricas.

2.1 TERMINOLOGIAS CLÍNICAS

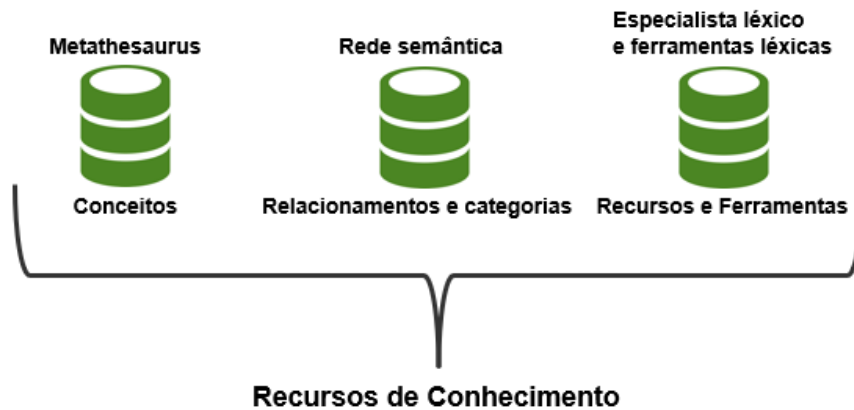
Um sistema terminológico é um modelo de conceitos e relações juntamente com os termos que lhes pertencem (KEIZER; ABU-HANNA; ZWETSLOOT-SCHONK, 2000). Uma terminologia clínica é formada por conceitos comumente usados na área da saúde.

Existem diversas terminologias na área da saúde, cada uma com foco em diferentes áreas, como o CID, relacionado à classificação de doenças, o MeSH, voltado para termos médicos, ou o *Logical Observation Identifiers Names and Codes* (LOINC), concernentes a termos associados aos exames laboratoriais, dentre outros. Entretanto, duas terminologias – SCT e UMLS – se sobressaem visto que abordam um grande número de conceitos e termos relacionados à saúde.

2.1.1 UNIFIED MEDICAL LANGUAGE SYSTEM (UMLS)

A UMLS, desenvolvida pela *National Library of Medicine* (NLM), tem por objetivo facilitar a integração e a interoperabilidade de terminologias clínicas em sistemas informáticos (BOUSQUET *et al.*, 2012). Contém 201 terminologias com seus conceitos unidos semanticamente por um identificador único de conceito (CUI) (BECKER *et al.*, 2017). A UMLS possui 3 recursos de conhecimento: o *Metathesaurus*, a rede semântica e o léxico especialista (NLM, 2018), conforme Figura 1.

Figura 1 – Os 3 recursos de conhecimentos da UMLS



Fonte: Traduzido de UMLS *Basics Tutorial*¹

O *Metathesaurus* é um banco de dados de vocabulário extenso, multiusuário e multilíngue que contém informações de conceitos biomédicos relacionados à saúde (NLM, 2018). Este possibilita ligar nomes e visualizações alternativas do mesmo conceito e identificar relações úteis entre diferentes conceitos, o que faz deste um ótimo recurso para mapeamento semântico entre terminologias (KIM, 2016; NLM, 2018; TRAN *et al.*, 2015). Possui diversas tabelas no seu escopo, das quais as mais importantes para este estudo são: MRCONSO, MRREL e MRXW_POR.

A tabela MRCONSO lista todos os conceitos incorporados a UMLS sem duplicação (JOUBERT *et al.*, 2009; NLM, 2018). Cada linha representa um termo UMLS, os conceitos são identificados pela CUI, por exemplo os termos “*pain*” e “*dor*” possuem a CUI “C0030193”, significa que os dois representam o mesmo conceito UMLS. Além da CUI, a tabela MRCONSO traz outras informações importantes sobre os termos, como: terminologia de origem, o código deste na terminologia de origem, representar se o termo é o preferido ou um sinônimo, a sua representação em linguagem natural e outras mais. No Quadro 1, são representadas todas as colunas existentes na tabela MRCONSO seguido do tipo de dados aceitos.

¹ Disponível em: <https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html>.

Quadro 1 – Descrição do conteúdo das colunas da tabela MRCONSO

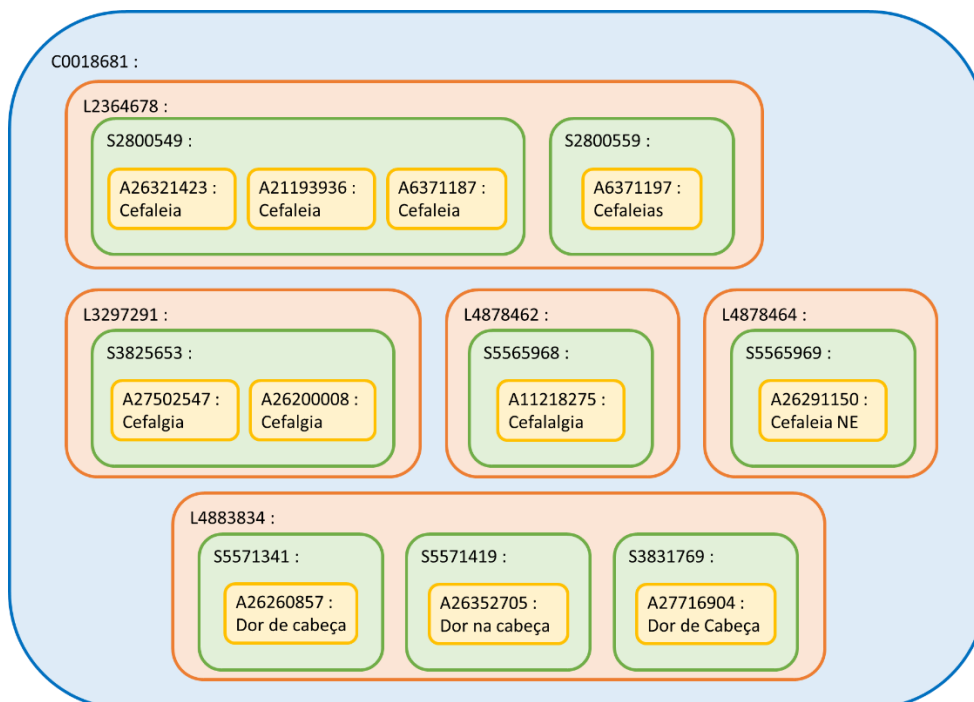
MRCONSO	
Coluna	Descrição
CUI	Identificador único de conceito
LAT	Idioma do termo
TS	Situação do termo
LUI	Identificador único do termo
STT	Tipo da <i>String</i>
SUI	Identificador único de <i>String</i>
ISPREF	Situação atômica – preferido (Y) ou não (N) para essa <i>String</i> dentro desse conceito.
AUI	Identificador único atômico
SAUI	Identificador único atômico da terminologia de origem [opcional]
SCUI	Identificador único de conceito da terminologia de origem [opcional]
SDUI	Identificador único de descritor da terminologia de origem [opcional]
SAB	Abreviação do nome da terminologia de origem
TTY	Abreviação para o tipo do termo no vocabulário de origem
CODE	Identificador na terminologia de origem mais usado ou um identificador de entrada da terminologia de origem criado pelo gerador – <i>Metathesaurus</i> .
STR	Termo em linguagem natural (<i>String</i>)
SRL	Restrição de nível na terminologia de origem
SUPPRESS	Valor flag supressivo
CVF	Flag de visualização de conteúdo. Campo de bits usado para sinalizar linhas incluídas na Exibição de Conteúdo. Esse campo é um campo varchar para maximizar o número de bits disponíveis para uso.

Fonte: *UMLS Reference Manual*.

Os termos UMLS possuem outros códigos além da CUI, o Identificador Único de Termo (LUI – Lexical (*term Unique Identifiers*), Identificador Único de *String* (SUI – *String Unique Identifiers*) e Identificador Único Atômico (AUI – *Atom Unique Identifiers*). O código LUI une os termos que possuem a mesma representação léxica incluindo suas variantes ortográficas, o código SUI identifica os termos escritos de maneira igual e o código AUI identifica aqueles de forma única. *String* é um conjunto de caracteres composto por letras e números.

Um exemplo desses códigos é apresentado na Figura 2 para o conceito “Dor de cabeça”, que possui a CUI “C0018681”, representado lexicalmente de 5 formas distintas identificadas na Figura 2 pelos códigos LUI “L2364678”, “L3297291”, “L4878464”, “L4883834” e “L4878462”. O primeiro código LUI possui uma variante ortográfica “Cefaleias” (plural de Cefaleia) fazendo com que este apresente dois códigos SUI para representar cada uma das variantes “S2800549” e “S2800559”. O primeiro SUI representa um termo que é usado por terminologias de origem diversas, distinguidos pelos códigos AUI “A6371187”, “A26321423” e “A21193936”.

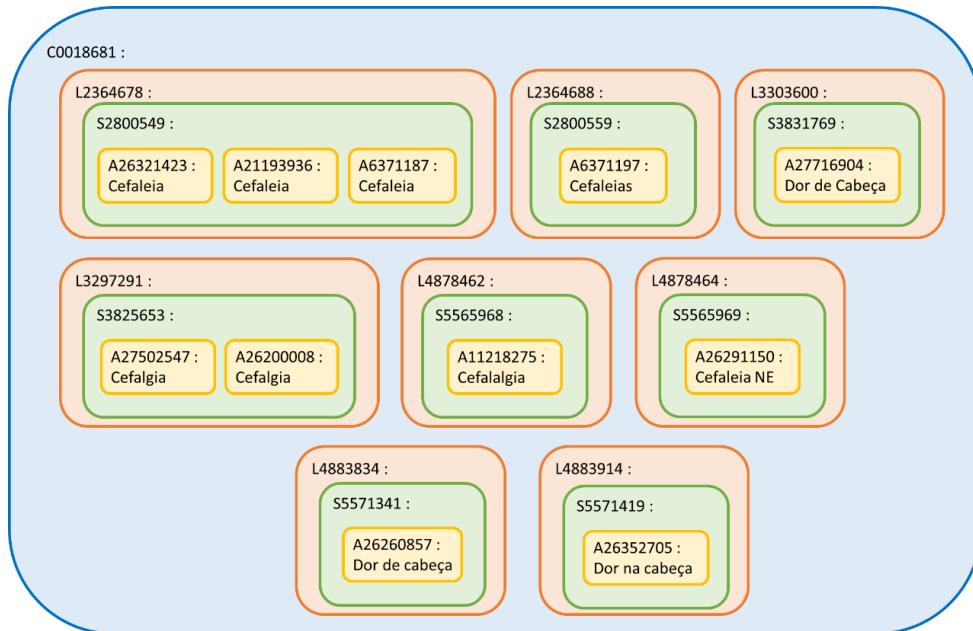
Figura 2 – A abrangência da CUI, LUI, SUI e AUI



Fonte: Elaborado pelo autor.

A Figura 2 representa um cenário ideal, considerando o uso de ferramentas léxicas para o Pt-BR, assim como é feito para o inglês, porem como estas existem apenas para o inglês os códigos LUI e SUI representam o mesmo conjunto de termos, como apresentado na Figura 3.

Figura 3 – A abrangência da CUI, LUI, SUI e AUI



Fonte: Elaborado pelo autor.

A tabela MRREL representa os relacionamentos, se houver, entre conceitos nas suas terminologias de origem e aqueles adicionados a estes pela UMLS (JOUBERT *et al.*, 2009; NLM, 2018). Cada relacionamento, presente no *Metathesaurus*, possui um identificador de relacionamento exclusivo (RUI). O objetivo dele é facilitar a detecção de mudanças nos relacionamentos nas versões futuras da UMLS. Além do RUI, a tabela MRREL tem outras informações como os CUI dos dois conceitos envolvidos no relacionamento, suas terminologias de origem, qual o tipo do relacionamento, entre outros. O tipo do relacionamento é mostrado em duas colunas, na coluna REL em que é descrito de forma mais básica como “Filho de” “Qualificador de”, “Pai de”, entre outros, e de maneira mais exata na coluna RELA como “É um”, “Componente de”, “Classificado como”, “Tradução de”, entre outros. No Quadro 2, são representadas todas as colunas existentes na tabela MRREL, seguida pelo tipo de dado que estas aceitam.

Quadro 2 – Descrição do conteúdo das colunas da tabela MRREL

MRREL	
Coluna	Descrição
CUI1	Identificador único do primeiro conceito
AUI1	Identificador atômico do primeiro conceito
STYPE1	Nome da coluna, na tabela MRCONSO, que contém o identificador usado para o primeiro elemento do relacionamento, ex: AUI, CODE, CUI, SCUI, SDUI.
REL	Relação do segundo conceito ou átomo para o primeiro conceito ou átomo.
CUI2	Identificador único do segundo conceito
AUI2	Identificador atômico do segundo conceito
STYPE2	Nome da coluna, na tabela MRCONSO, que contém o identificador usado para o segundo elemento do relacionamento, ex: AUI, CODE, CUI, SCUI, SDUI.
RELA	Adicional (mais específico) rótulo de relação [opcional]
RUI	Identificador único do relacionamento
SRUI	Identificador único de relacionamento na terminologia de origem, se presente.
SAB	Abreviação do nome da terminologia de origem
SL	Rótulo do relacionamento na terminologia de origem
RG	Grupo de relacionamento. Usado para indicar que um conjunto de relacionamentos deve ser analisado junto.
DIR	<i>Flag</i> de direcionamento na terminologia de origem. Y indica que esta é a direção do relacionamento em sua origem, N indica que não ser, espaço em branco indica não ser importante ou ainda não foi determinado.
SUPPRESS	Valor <i>flag</i> supressivo
CVF	<i>Flag</i> de visualização de conteúdo. Campo de <i>bits</i> usado para sinalizar linhas incluídas na Exibição de Conteúdo. Esse campo é um varchar que maximiza o número de <i>bits</i> disponíveis para uso.

Fonte: UMLS Reference Manual.

A tabela MRXW_POR contém uma lista de palavras em português que existem nos termos da UMLS, associando a cada uma ao CUI, LUI e SUI do termo em que essa existe. No Quadro 3, são representadas todas as colunas da tabela MRXW_POR, seguidas da descrição do conteúdo em cada coluna.

Quadro 3 – Descrição do conteúdo das colunas da tabela MRXW_POR

MRXW_POR	
Coluna	Descrição
LAT	Abreviação ou idioma da <i>String</i> na qual a palavra está presente.
WD	Palavra em minúsculo
CUI	Identificador único de conceito
LUI	Identificador único de termo
SUI	Identificador único de <i>String</i>

Fonte: *UMLS Reference Manual*.

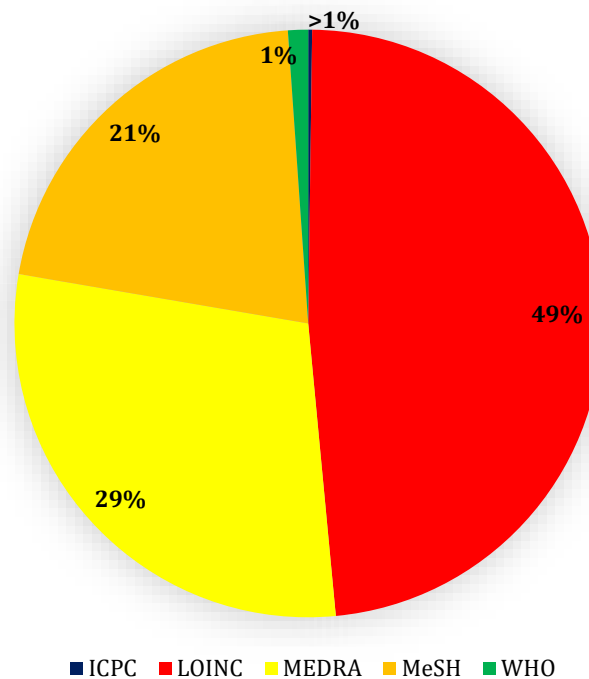
Outro recurso de conhecimento, disponível na UMLS, é a rede semântica, a qual ajuda na interpretação do significado de cada termo, fornecendo informações sobre o conjunto de tipos semânticos básicos ou categorias, os quais podem ser atribuídos a esses conceitos e definem o conjunto de relacionamentos que podem ocorrer entre os termos (NLM, 2018). A rede semântica contém 133 tipos semânticos (exemplo: sinal ou sintoma, conceito temporal, substância do corpo, alimento, animal e substância farmacológica) e 54 relacionamentos (exemplo: resultado de, exibidos, usos, diagnósticos, medidas e causas). Um ou mais tipos semânticos podem ser atribuídos a cada um dos conceitos do *Metathesaurus*, descrevendo a semântica do conceito e identificando sua categoria ou categorias (HE *et al.*, 2014).

O terceiro recurso de conhecimento disponível é o léxico especialista, que fornece informações léxicas necessárias para tarefas de PLN, para o inglês, com várias informações sintáticas, morfológicas e ortográficas sobre cada palavra (NLM, 2018). Ferramentas léxicas são construídas para abordar as diversas variâncias entre palavras, a qual a linguagem natural possui.

A UMLS, em sua versão 2017ab (versão do segundo semestre de 2017), conta com 12.809.371 termos, destes 8.836.759 estão em inglês (aproximadamente 69% de toda a UMLS), em português, 338.847 (aproximadamente 3% de toda a UMLS). Os

termos em português são divididos em 5 terminologias de origem: *International Classification of Primary Care (ICPC)*, *LOINC*, *Medical Dictionary for Regulatory Activities (MedDRA)*, *MeSH* e *World Health Organization terminology (WHO)*. Os números de termos que estas possuem, em português, respectivamente é 723, 163.620, 98.914, 71.840 e 3.750, na Figura 4, é possível observar a parcela de cada uma delas em relação ao total de termos UMLS disponíveis em português.

Figura 4 – Terminologias de origem dos termos em português da UMLS

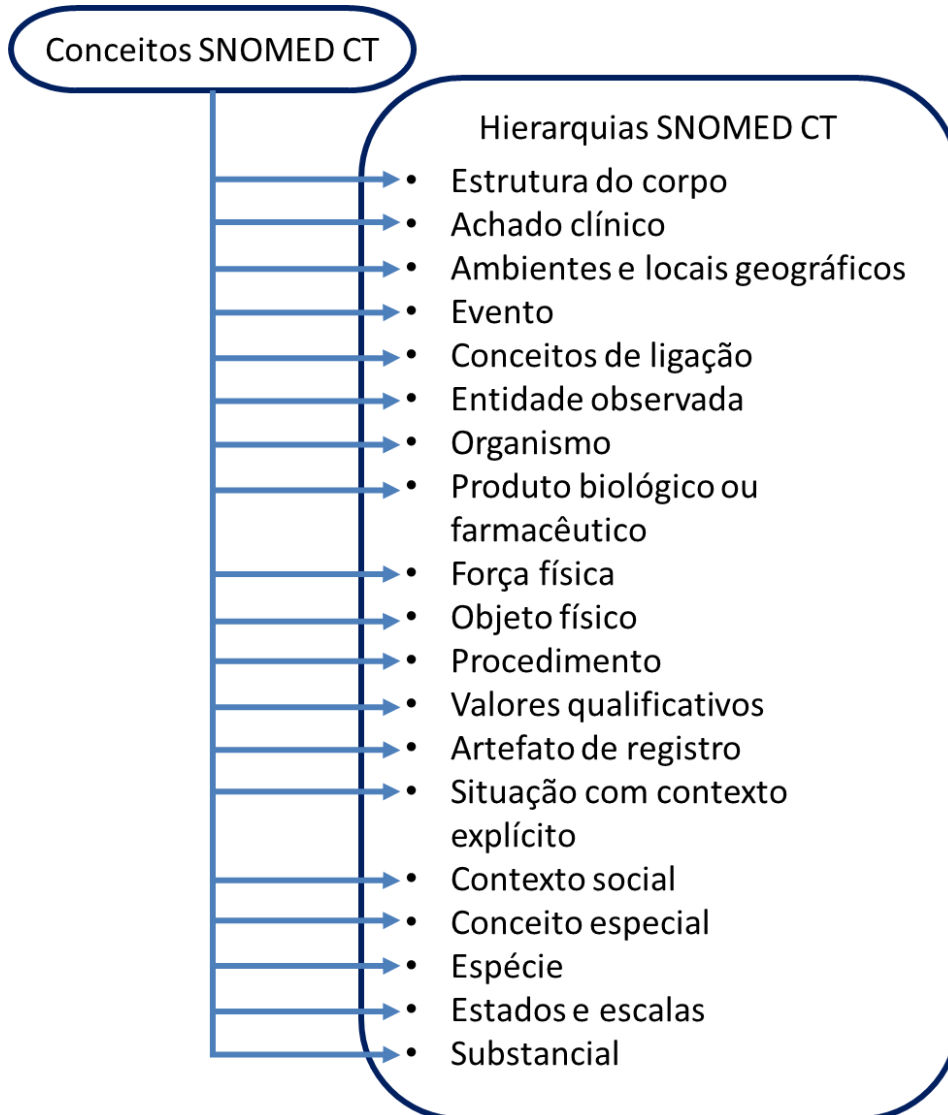


Fonte: Elaborado pelo autor.

2.1.2 SNOMED CT

A SCT é considerada a maior terminologia clínica multilíngue, dedicada a codificar todos os aspectos dos registros eletrônicos de saúde (ALLONES; MARTINEZ; TABOADA, 2014). Seu conteúdo está organizado em 19 hierarquias diferentes (KIM, 2016; SAIWAL *et al.*, 2012) apresentadas na Figura 5.

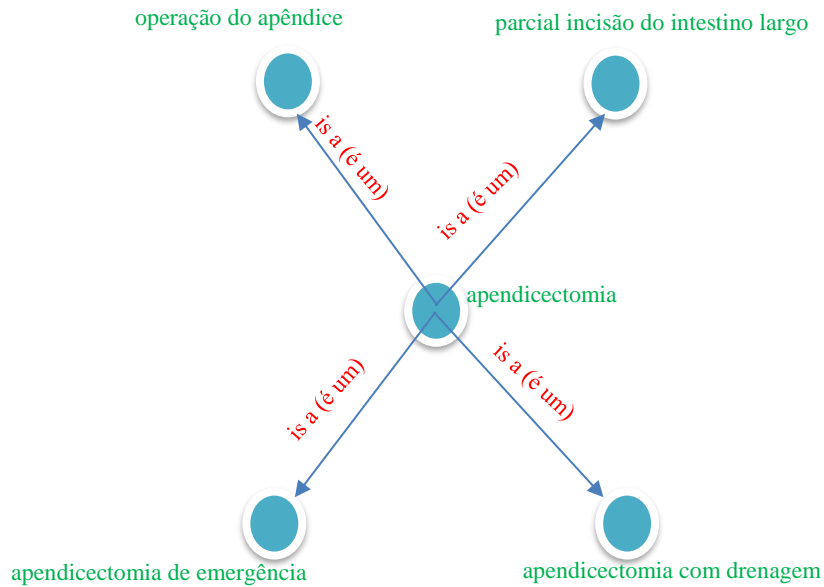
Figura 5 – Hierarquias da SCT



Fonte: Traduzido e adaptado do *starter guide* da SCT.

A SCT possui dois tipos de relacionamento entre conceitos, um hierárquico do tipo pai-filho, ligando os conceitos através do termo “*is a*” (é um). Cada pai pode ter diversos filhos, e cada filho pode ter diversos pais, por exemplo, na Figura 6, o conceito apendicectomia é filho dos conceitos “operação no apêndice” e do conceito “parcial incisão do intestino largo” e é pai dos conceitos “apendicectomia de emergência”, “apendicectomia com drenagem”, dentre outros.

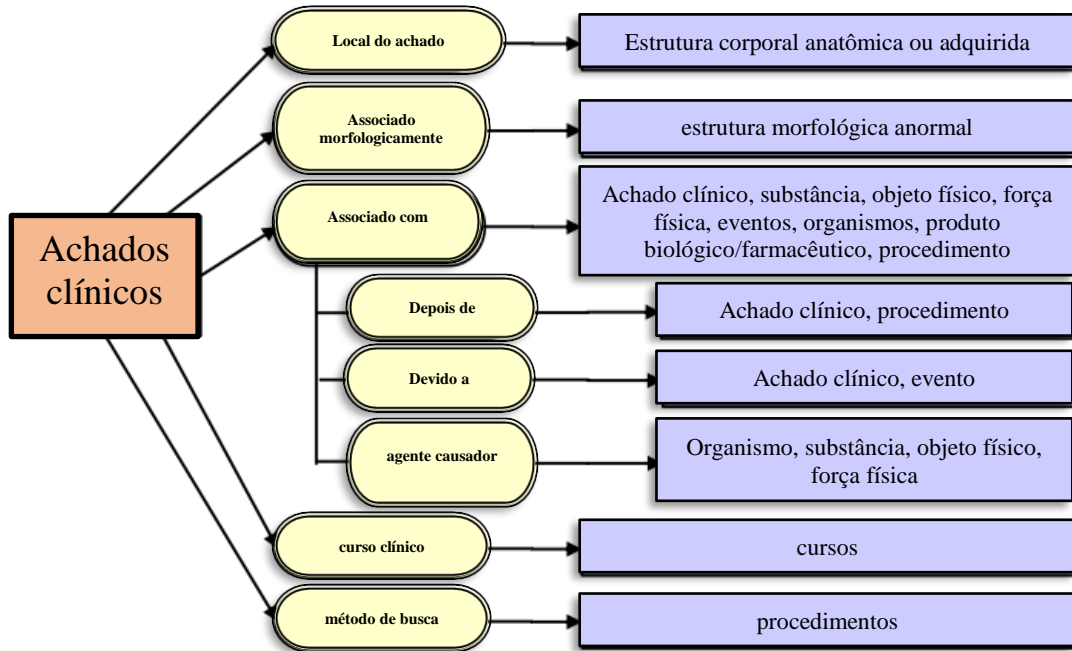
Figura 6 – Relações hierárquicas do conceito “apendicectomia”



Fonte: Traduzido do *starter guide* da SNOMED CT.

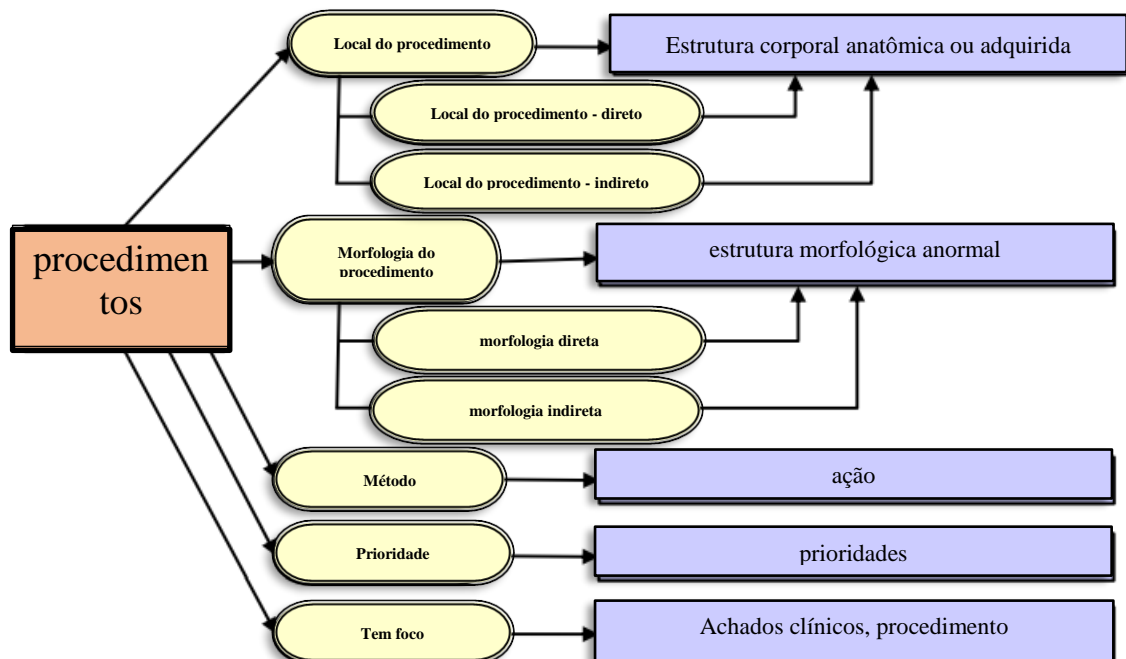
O outro relacionamento é chamado relacionamento por atributos, no qual os conceitos são ligados por diferentes atributos, definidos conforme a hierarquia a qual o conceito pertence, por exemplo, os conceitos ligados à hierarquia achado clínico possuem os atributos mostrados na Figura 7, diferentes dos atributos usados para a hierarquia procedimentos na Figura 8, cada um deles relaciona o conceito ao descrito em roxo nas Figuras 7 e 8, ou a um de seus filhos.

Figura 7 – Atributos que podem ser usados na hierarquia “achados clínicos”



Fonte: Traduzido do *starter guide* da SCT.

Figura 8 – Atributos que podem ser usados na hierarquia “procedimentos”



Fonte: Traduzido do *starter guide* da SCT

Cada conceito possui um identificador único numérico nomeado SCT *concept identifier* ou simplesmente SCTID (SNOMED CT, 2017). E um ou mais termos que os descrevem, dos quais um é do tipo *fully specified name* (FSN), que nomeia o conceito, atribuído apenas a ele, de modo a evitar ambiguidade, e os outros considerados sinônimos, podendo descrever diversos conceitos (BÁNFAI; PORCIÓ; KOVÁCS, 2014; SNOMED CT, 2017). Os conceitos SCT podem ser ditos definidos ou primitivos, o primeiro refere-se àqueles cujos relacionamentos estão ligados a este e descrevem-nos completamente, já segundo não pode ser identificado apenas pelos relacionamentos que possuem (JOUBERT *et al.*, 2009; SNOMED CT, 2017).

Os termos contêm identificadores numéricos únicos, *descriptionid*, e são identificados como preferidos ou aceitáveis. Os preferidos são os termos usados para representar o conceito que esses descrevem na língua em que está configurada a SCT, quanto aos termos aceitáveis, são as outras formas como o conceito pode ser escrito (SNOMED CT, 2017).

Um componente da SCT, de grande utilidade, é o *reference set* (Refsets), descrito como um padrão flexível que a SCT usa para suportar variedades de requerimentos para customização e melhora da SCT. Estes incluem representações de subconjuntos, preferências de linguagem de um determinado termo e mapeamento para outros sistemas de código. Todo Refset tem um identificador de conceito numérico único (Refsetid) (SNOMED CT, 2017).

Uma característica que faz da SCT abrangente é a possibilidade do uso de expressões para representar um termo clínico por meio da combinação de um ou mais conceitos (BÁNFAI; PORCIÓ; KOVÁCS, 2014; SNOMED CT, 2017). Existem dois tipos de expressões, as pré-coordenadas e as pós-coordenadas.

Uma expressão pré-coordenada é uma forma de representar um conceito da SCT e pode ser expressada pelo identificador do conceito, por exemplo: 31978002, ou por um junto com o termo que define o conceito separado por |, exemplo: 31978002 | fracture of tibia |.

A expressão pós-coordenada contém dois ou mais conceitos identificadores, nos quais o significado é representado pela combinação dos significados destes. Para representar o conceito fratura da tíbia esquerda, podemos usar a expressão:

- 31978002 | *fracture of tibia* | : 272741003 | *laterality* | = 7771000 | *left* |

Ou pode também ser expressada apenas pelos identificadores:

- 31978002 : 272741003 = 7771000

Então, é possível refinar o significado de um conceito aplicando valores mais específicos para um ou mais relacionamentos de definição, por exemplo, o Conceito 31978002 | *fracture of tibia* | tem o relacionamento de definição 116676008 | *associated morphology* | = 72704001 | *fracture* |, por isso, o termo clínico “fratura aberta da tíbia” pode ser representado por 31978002 | *fracture of tibia* |: 116676008 | *associated morphology* | = 52329006 | *fracture, open* |, ou simplesmente, 31978002: 116676008 = 52329006. Também pode-se refinar o significado de um Conceito aplicando valores a outros atributos permitidos pela SCT, por exemplo, 31978002 | *fracture of tibia* | é um subtipo de 404684003 | *clinical finding*|. A SCT permite que subtipos de 404684003 | *clinical finding*| usem o atributo 42752001 | *due to* |, assim, o termo clínico “fratura da tíbia por queda no gelo” pode ser representado por 31978002 | *fracture of tibia* |: 42752001 | *due to* | = 75354000 | *fall on ice* |, ou apenas 31978002 : 42752001 = 75354000.

2.2 MAPEAMENTO TERMINOLÓGICO

O mapeamento terminológico tem por objetivo encontrar conceitos correspondentes de termos em linguagem natural, para terminologias padronizadas, podendo ser do tipo semântico ou léxico (SIERRA *et al.*, 2015).

O mapeamento léxico busca termos da terminologia alvo que possuam a mesma característica léxica do termo em linguagem natural (DHOMBRES; BODENREIDER, 2016; KIM, 2016). Para se obter um melhor desempenho, todos os termos devem ser normalizados antes de ser comparados, eliminando letras maiúsculas e acentuações, e removendo *Stop Words* (BOUSQUET *et al.*, 2012; KIM, 2016; SUN; SUN, 2006), os quais são um conjunto de palavras que não possuem conteúdo semântico significativo no contexto no qual são inseridas, geralmente, são palavras conectivas como preposições e artigos.

O mapeamento semântico leva em consideração as características semânticas do termo, uma maneira de realizá-lo é levar em conta as relações semânticas que o termo possui na terminologia alvo e o tipo semântico que este possui (ALLONES; MARTINEZ; TABOADA, 2014; KIM; COENEN; HARDIKER, 2012).

Os textos livres apresentam termos com variações ortográficas, o que dificulta o mapeamento, a fim de contornar tais problemas, pode-se usar os algoritmos de stemização e lematização.

O algoritmo de stemização é um procedimento computacional que reduz todas as palavras com a mesma raiz para uma forma comum, geralmente, removendo de cada palavra seus sufixos derivacionais ou flexionais (LOVINS, 1968; WILLETT, 2006). Um exemplo são os termos “**infecção**”, “**infecções**” e “**infeccionou**”, os quais apresentam diferenças léxicas significativas entre si, no entanto, ao remover seus sufixos, apresentam a mesma raiz “**infec**” o que garante a similaridade entre eles. Para o Pt-BR, existem alguns algoritmos que realizam tal procedimento. O presente estudo optou por usar o Snowball (SNOWBALL, 2019), uma vez que este facilita para diferentes tipos de linguagens de programação e por ser atualizado constantemente.

A lematização é o processo de encontrar a forma mais simples (ou lema) de uma palavra, considerando suas formas flexionadas (LIU *et al.*, 2012; RODRIGUES; OLIVEIRA; GOMES, 2014). Para se realizar a lematização, é necessário um dicionário de consulta (ISMAILOV *et al.*, 2016). Um exemplo de lematização é transformar os termos “**infecção**”, “**infecções**” e “**infeccionar**” em sua forma mais básica que, no caso, é o termo “**infecção**”.

Para realizar o mapeamento, é necessário o uso de ferramentas que comparem *strings*. Os principais métodos para essa tarefa são a distância de edição, o TDIDF (*term frequency–inverse document frequency*) e a distância baseada em *tokens*.

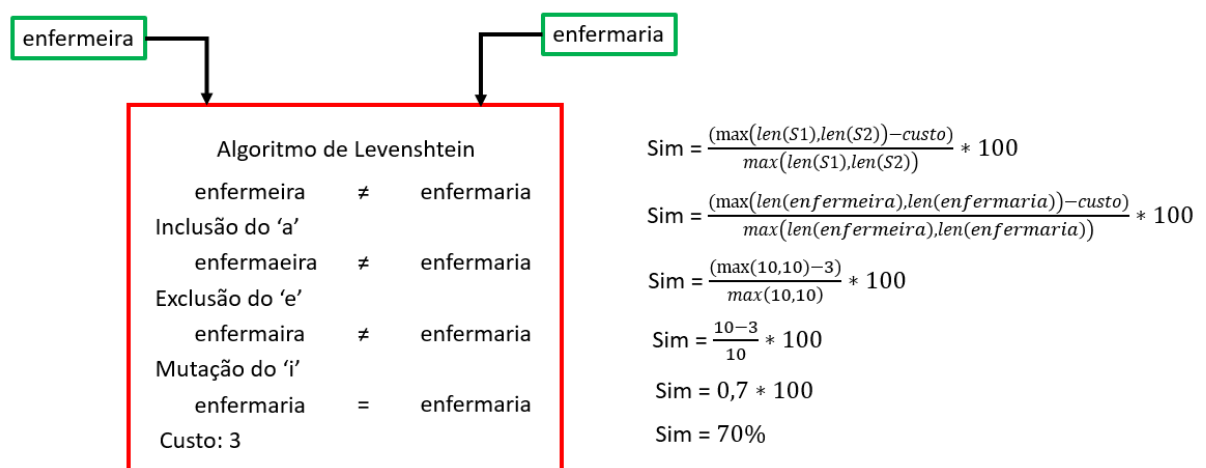
Os algoritmos de distância, baseados em *tokens*, consideram um termo como um conjunto de palavras e calcula o quão similar estas são por intermédio do número de palavras em comum (ALLONES; MARTINEZ; TABOADA, 2014). Este compara dois termos S e T para calcular a similaridade entre eles, dividindo o número de palavras de S que não estão presentes em T pelo número de palavras de S presentes em T, seguindo a fórmula $|S \cap T| / |S \cup T|$, em que quanto mais próximo de 0 for o resultado mais similar os termos S e T são. Por exemplo, ao comparar os termos “**respiração espontânea**” com “**respiração**”, verifica-se que possuem uma palavra em comum e uma diferente, ao realizar o cálculo de similaridade, tem-se “1” como resultado. Já ao comparar o termo “**membro superior direito**” com “**membro superior**”, estes possuem duas palavras em comum e uma diferente será encontrada, ao fazer o cálculo de similaridade, o resultado é 0,5.

O algoritmo TFIDF ou similaridade dos cossenos faz o cálculo a depender das palavras que os termos têm em comum, contudo, neste as palavras são ponderadas de acordo com sua frequência, em que quanto menos aparecerem na coleção de termos, maior o peso no cálculo atribuído a esses (COHEN; RAVIKUMAR; FIENBERG, 2003). Então, ao comparar dois termos como por exemplo os termos “cirurgia do coração” com “cirurgia de coração” às palavras “do” e “de”, teriam menos peso por serem preposições que possuem uma alta frequência, o que faria com que não influenciassem na similaridade entre os termos de forma significativa.

Os algoritmos de distância de edição calculam a similaridade entre os termos comparados verificando o número de passos necessários para transformar um termo em outro (ALLONES; MARTINEZ; TABOADA, 2014; PINHAS *et al.*, 2013). São permitidas as seguintes operações para esses passos: inserção de uma letra em alguma posição do termo, exclusão de uma letra e a mutação (troca) de posição com outra (PINHAS *et al.*, 2013).

Para este estudo, foi utilizada a distância de edição de *Levenshtein* na qual cada uma das operações tem o custo 1 (SINTCHENKO *et al.*, 2009). Logo os termos comparados, que obtiverem um menor custo, são candidatos a serem escolhidos como mais semelhantes ao termo buscado. Como mostra a Figura 9, ao comparar o termo “enfermeira” com “enfermaria”, o algoritmo realiza 3 operações para transformar um em outro, fazendo com que o custo entre os dois termos seja de 3.

Figura 9 – Execução do algoritmo de *Levenshtein*



Fonte: Elaborado pelo autor.

Para facilitar a comparação entre os termos, é possível transformar o valor de custo dado pelo algoritmo de *Levensthein* em um valor em porcentagem para representar a similaridade entre os termos. Esse cálculo pode ser feito como demonstrado na Equação 01, em que S1 é um termo e S2 o termo a ser comparado com aquele, len é o cálculo do número de letras que o termo contém, max é o cálculo para selecionar o maior len entre os dois termos e Sim é a similaridade apresentada em porcentagem. Usando o nosso exemplo da Figura 10 o termo “enfermeira” apresenta uma similaridade de 70% com o termo “enfermaria”.

$$Sim = (\max(\text{len}(S1), \text{len}(S2)) - \text{custo}) / \max(\text{len}(S1), \text{len}(S2)) * 100 \quad (1)$$

2.3 METAMAP

O MetaMap tem por objetivo mapear textos biomédicos para conceitos da UMLS (ARONSON, 2001). É um dos algoritmos mais usados para essa tarefa e está disponível apenas para o inglês. O MetaMap divide o texto em frases nominais e busca conceitos candidatos na UMLS para representá-las através da geração de variantes ortográficas e de um sistema de valor de distância, no qual cada tipo de variante possui uma pontuação diferente, mostradas no Quadro 4.

Quadro 4 – Valor de distância das variantes

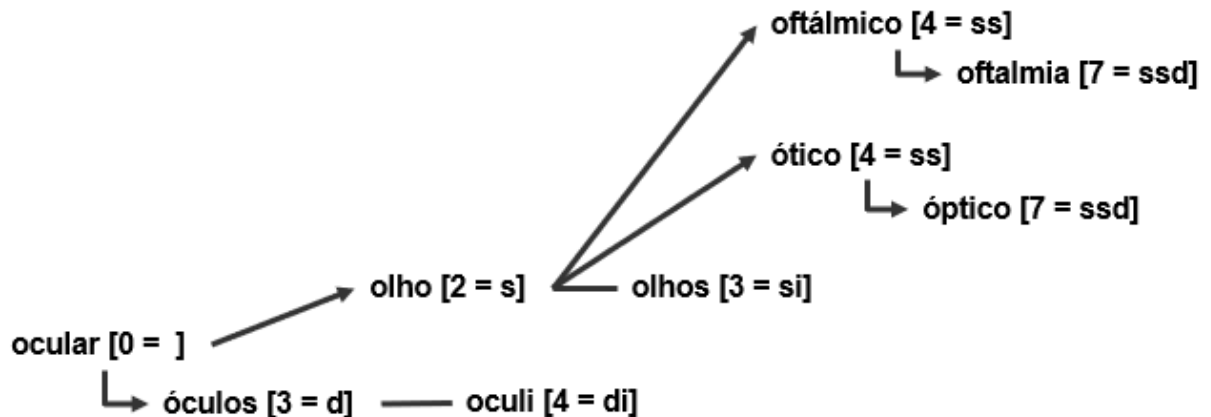
Tipo de variante ortográfica	Valor de Distância
Mesma ortografia	0
Variante flexional	1
Sinônimo ou acrônimo/abreviatura	2
Variante derivacional	3

Fonte: Traduzido do artigo de Aronson (2006).

Essa distância é calculada com o intuito de verificar quantas variantes são necessárias para chegar ao conceito UMLS. Na Figura 10, temos o exemplo do termo “ocular”, no qual “s” significa sinônimo, “i” variante flexional e “d” variante derivacional.

Caso exista o conceito UMLS “óptico”, este terá uma distância de 7, pois para chegar até este, foram usados dois sinônimos e uma variante derivacional.

Figura 10 – Variantes para o termo “ocular”



Fonte: Traduzido e adaptado da Figura 2 de Aronson (2001).

Para selecionar os conceitos candidatos, o MetaMap não usa apenas essa distância, também calcula a força do mapeamento usando uma função de avaliação baseada em princípios linguísticos, que consistem em uma média ponderada de quatro métricas: centralidade (envolvendo o escopo), variação (média dos escores de distância inversa) e cobertura e coesão (ARONSON, 2001). E, por fim, os candidatos são apresentados de acordo com a força de mapeamento, a qual varia de 0 a 1000.

Há algumas abordagens para mapeamento de termos clínicos em outros idiomas, como o estudo de Chiaramello *et al.* (2016), os quais adaptaram para o italiano a estrutura do MetaMap, um método proposto por Aronson (ARONSON, 2001), para extração e mapeamento para a UMLS de textos livres em inglês. Para esse caso, foi necessário substituir as etapas de *parsing* e geração de variantes por ferramentas específicas para o italiano com foco na área clínica. Todavia, as propostas de método existentes não podem ser aplicadas ou adaptadas diretamente para o português, uma vez que não estão disponíveis ferramentas para gerar as variantes ortográficas para o português.

2.4 MÉTRICAS

Para medir a evolução do estudo e a qualidade dos resultados, algumas métricas foram definidas. Utilizá-las requereu que os mapeamentos realizados fossem divididos em 3 grupos: mapeamento exato, mapeamento parcial e não mapeamento.

Mapeamento exato: quando o conceito na terminologia destino, definido para representar o termo de entrada (termo em linguagem natural extraídos dos RES usados para este estudo), equivale a este semanticamente de forma completa.

Mapeamento parcial: quando o conceito na terminologia destino, designado para representar o termo de entrada, pode ser considerado ou trata-se de um conceito pai ou filho.

Não mapeamento: quando não é encontrado um conceito na terminologia destino para representar o termo de entrada, ou é designado um conceito com um significado semântico diferente para representar o termo de entrada.

As métricas adotadas por este estudo foram *recall*, *precision* e *F-score*. Para o cálculo destas, foram usadas as Equações 1, 2 e 3 apresentadas abaixo.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{F-score} = 2 * \frac{\text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \quad (3)$$

Para adotar as métricas, é necessário identificar o número de mapeamentos verdadeiros positivos (TP) e negativos (TN), os falsos positivos (FP) e negativos (FN). Os TP representam os resultados não nulos e verdadeiramente válidos no mapeamento, estes foram compostos dos mapeamentos exatos somado com metade dos parciais. Os TN representam os resultados não nulos e não verdadeiros; no mapeamento, estes foram compostos por metade dos mapeamentos parciais. Os FP representam os resultados nulos que realmente não são possíveis; para o mapeamento, o cálculo deste foi o total de não mapeamentos, não encontrado manualmente também. Os FN representam os resultados nulos que não deveriam ser nulos; no mapeamento, estes representaram o número de não mapeamentos que foram mapeados manualmente.

3. METODOLOGIA

Este estudo caracteriza-se como de desenvolvimento transversal, tecnológico e experimental. Durante este trabalho, foi proposto um método, baseado nele foi elaborada uma ferramenta em formato de software, o MapClin.

3.1. LOCAL DO ESTUDO

O estudo foi realizado no Laboratório de Informática em Saúde (LAIS), do Programa de Pós-graduação em Tecnologia em Saúde (PPGTS) com o apoio do grupo de pesquisa.

3.2. PREPARAÇÃO DA BASE DE DADOS

Este estudo utilizou dos termos clínicos extraídos pelo projeto *NLPPCD* (OLIVEIRA *et al.*, 2017). Neste, os termos foram extraídos das narrativas clínicas sem alteração léxica, ou seja, foram mantidos da mesma forma com a qual foram escritos, ainda que incluíssem erros de ortografia e/ou digitação. Essa extração foi realizada com a inclusão de etiquetas semânticas, ação denominada marcação, nas narrativas clínicas. Dois especialistas anotaram individualmente cada uma das narrativas clínicas selecionadas, e um terceiro fez a adjudicação das etiquetas.

Na Figura 11, é apresentado um exemplo de como foi realizada a marcação e a estrutura da ferramenta utilizada nessa tarefa. São mostradas as marcações feitas pelos anotadores 1 e 2. Acima dessas, a ferramenta marca automaticamente os termos anotados do mesmo modo pelos dois anotadores, e abaixo o assistente de adjudicação apresenta todas anotações divergentes para que o adjudicador as aceite ou não. Por exemplo, “IMC 37” foi marcado pelos dois anotadores de formas diferentes, o adjudicador decide, então, qual das duas marcações é a mais adequada ou se nenhuma delas é válida.

Figura 11 – Exemplo de anotação feita nas narrativas clínicas

IMC 37 .
 # EM USO DE HCTZ .
 # DISCUTO CASO COM .
 CONFORME ORIENTADO : SOLICITO ENCAMINHAMENTO PARA CX BARIATRICA .

Anotador 1
 Tags: Quantitative Concept
 # IMC 37 .
 # EM USO DE HCTZ .
 # DISCUTO CASO COM .
 CONFORME ORIENTADO : SOLICITO ENCAMINHAMENTO PARA CX BARIATRICA .

Anotador 2
 # IMC 37 .
 # EM USO DE HCTZ .
 # DISCUTO CASO COM .
 CONFORME ORIENTADO : SOLICITO ENCAMINHAMENTO PARA CX BARIATRICA .

↓ Assistente de adjudicação 6

Deseja anotar o termo 37 como Quantitative Concept?
 Anotador: Anotador_1 Sim Não

Deseja anotar o termo USO como Quantitative Concept?
 Anotador: Anotador_2 Sim Não

Fonte: Elaborado pelo autor.

Os anotadores atribuíram etiquetas semânticas a todos os termos extraídos, conforme mostrado na Figura 12. Estas são baseadas nos tipos semânticos da UMLS. Foi adicionado também a etiqueta 'Abbreviation' junto as etiquetas semânticas para indicar quando este termo se encontrava de forma abreviada ou acrônimo.

Figura 12 – Extração de termo, com a expansão da abreviatura e a inserção de etiquetas semânticas

EVOLUÇÃO DIA 01/01/10 16:57h Veio do CC as 10h acordado, responsivo, respiração espontanea em ar ambiente, no momento em repouso no leito e com relato de algia em reg. incisao cirurgica abdominal com debito hematico e dreno de suctor a esquerda com debito hematico + dreno de torax a esquerda com 250ml de debito sanguinolento ate o momento. Apresenta acesso venoso periferico MSD salinizado e diurese por SVD presente em coletor amarelo citrico. Enfermeria Florence coren 9999



Fonte: Elaborado pelo autor.

Esses termos foram extraídos da base de 1030 narrativas clínicas utilizadas no projeto *Natural Language Processing for Portuguese Clinical Documents*. Esse

conjunto de narrativas contém evoluções médicas e de enfermagem e sumários de alta hospitalar oriundas de três hospitais do Grupo Marista de Curitiba – Paraná. Neste estudo, foram consideradas as evoluções médicas e os sumários de alta. O mapeamento de evoluções de enfermagem não é realizado, uma vez que este já foi objeto de pesquisa de outros projetos do PPGTS.

Considerando-se os termos sem repetição, ou repetição léxica, das evoluções médicas e sumários de alta, a base de termos extraídos é composta de 4603 termos simples (com apenas uma palavra) e 7905 termos compostos (com mais de uma palavra).

Para a seleção dos termos da base de treinamento deste estudo, foram selecionadas aleatoriamente 5 narrativas (dentre as 1030) e identificados os termos nelas anotados. Dessa forma, a base de treinamento foi composta por 216 termos clínicos distintos (136 termos simples e 80 termos compostos).

Com objetivo de abranger a maior incidência de termos, para a base de teste deste estudo, foram selecionados os 100 termos simples e os 100 compostos, com a maior frequência de anotação, totalizando 41% das anotações de termos compostos e 35% das anotações de termos simples.

Os anotadores também expandiram todos os acrônimos e as abreviaturas encontradas, como demonstrado no Quadro 5. Produzindo uma base de dados com essas expansões, que será utilizado pelo estudo nos termos extraídos dessa base e nos termos oriundos de outras bases. Foram registradas 1444 expansões, alguns exemplos são apresentados no Quadro 3.

Quadro 5 – Lista de abreviaturas e acrônimos com suas expansões

Abreviatura/acrônimo	Expansão
Pupilas iso/foto	pupilas isocóricas e fotorreagentes
CA Biliar	câncer biliar
PROC CIRÚRGICO	procedimento cirúrgico
SNE fechada	sonda nasoenteral fechada
SUBCLÁVIA D	subclávia direita
MV +	murmúrios vesiculares presentes
HID	Hidratado
MMIS	membros superiores e inferiores

Fonte: Elaborado pelo autor.

Também foi utilizado o *Metathesaurus* da UMLS de maneira local com todos os termos em português e inglês contidos na UMLS versão 2017ab, além do banco de termos e conceitos da SCT através de uma API disponibilizada pela própria SCT.

3.2.1. ASPECTOS ÉTICOS

O processo de anotação das narrativas clínicas, assim como este trabalho, tem o parecer ético nº 1.354.675 (07/12/2015) (ANEXO A) referente ao projeto para elaboração de algoritmos e métodos de processamento de documentos clínicos. Este estudo não utiliza diretamente as narrativas, somente o conjunto de termos delas extraídos.

3.3. ENCAMINHAMENTO METODOLÓGICO

Para realizar o mapeamento dos termos em linguagem natural anotados pelo projeto NLPPCD (termos de entrada) para a SCT, devido à falta de uma versão desta para o português, foi necessário encontrar alternativas para o mapeamento. A alternativa adotada, por este estudo, foi utilizar a UMLS como meio para se chegar a SCT. Esta possui um número considerável de termos em português e, através da CUI desses termos, é possível identificar um conceito SCT, mesmo que em inglês, associado a este. Por exemplo, o termo clínico “dor” possui na UMLS a CUI “C0030193”, essa CUI tem o conceito SCT “22253000|Pain (finding)” associado.

Entretanto, muitos termos de entrada não podem ser diretamente associados a um conceito UMLS, devido a formas diferentes de se escrever a mesma informação. Isso torna necessário o uso de diferentes regras do PLN para encontrar as variações ortográficas que tais termos podem assumir em um texto livre.

Para o identificar os diferentes padrões encontrados nos termos da base de treinamento e desenvolver as diferentes regras, foi adotado um método incremental em espiral, composto de 4 passos distintos: proposta ou melhora de uma regra, desenvolvimento da regra, teste com a base de treinamento e avaliação dos resultados do mapeamento.

O desenvolvimento foi, inicialmente, realizado com a adaptação das estratégias de mapeamento propostas por JL Allones, D.Martinez, M. Taboada, com o uso do

algoritmo de distância de *Levenshtein* para verificar a similaridade entre os termos, e o método de MetaMap, gerando variantes ortográficos para os termos de entrada. Para tanto, são usados os algoritmos de lematização e stemização para aumentar o número de possíveis variantes dos termos em português, e o método de T.Y. Kim, o qual usou a UMLS como meio para mapear conceitos CIPE para conceitos SCT.

A partir deles, foram propostas diferentes regras para mapeamento dos termos deste trabalho para a UMLS. Cada regra criada foi testada com a base de treinamento e os resultados foram avaliados para verificar se a regra encontrava conceitos candidatos aos termos em que foi projetada, ou se existia a necessidade de melhora para abranger mais termos, ou se esta não abrange termos diferentes das regras anteriores e não mapeia com menor número de falsos positivos que as regras já estabelecidas.

Durante o desenvolvimento, foi observada uma grande quantidade de conceitos SCT atribuídos a um mesmo termo de entrada pelas diversas regras. Para solucionar tal problema, foi adicionada uma fase de seleção, que apresenta os diversos conceitos ao usuário, e este seleciona o mais adequado a representar o termo de entrada.

Após o mapeamento dos termos em linguagem Natural para a UMLS, foi necessário encontrar o correspondente deste na SCT, para isso foi feito de forma manual, a associação de um grupo de conceitos UMLS para códigos SCT através dos relacionamentos apresentados na tabela MRREL e dos relacionamentos com os termos em inglês através da CUI na tabela MRCONSO.

A se alcançar metade do trabalho foi realizado um experimento parcial com as regras que tinham sido criadas até o momento, com a intenção de verificar de maneira geral como estas estavam se comportando e quais os grupos de termos que estavam mapeando ou deixando de mapear. Para auxiliar a tomada de decisões para as próximas regras a serem desenvolvidas e a melhora das regras até então desenvolvidas.

3.4. MAPEAMENTO DE TERMOS – VALIDAÇÃO

Para validar o método criado, foram mapeados os termos da base de teste e feita uma validação por pares, com a ajuda de três profissionais da área da saúde (A1, A2, A3), especialistas, com experiência na SCT. Destes três especialistas somente

um tinha participado do processo de anotação do projeto NLPPCD, e este também participou da avaliação durante o desenvolvimento do MapClin em que foram analisados os resultados no mapeamento da base de treinamento.

Todos os 200 termos da base de teste foram mapeados pelo MapClin usando uma ou mais das sete regras criadas. Foi elaborada uma tabela contendo para cada um dos termos de entrada os conceitos SCT candidatos relacionados a ele após a execução do MapClin. Os especialistas classificaram resultado do mapeamento em “exact match”, quando existia um conceito candidato que representava ao menos uma mesma informação que o termo de entrada representa; “partial match”, quando existiam apenas conceitos que representavam de forma parcial uma mesma informação; ou “not match”, se nenhum conceito candidato representava alguma informação do termo de entrada.

Cada especialista avaliou 50 termos simples e 50 compostos. Inicialmente cada um dos especialistas avaliou 50 termos, os especialistas A1 e A2 avaliaram o mesmo conjunto de 50 termos simples, correspondes aos de maior frequência. Os especialistas A1 e A3 avaliaram os 50 primeiros termos compostos. A partir destes primeiros resultados os pesquisadores analisaram a discordância entre os anotadores na validação conjunta. Como esta foi abaixo de 5% e todas relacionadas a discordâncias entre “exact match” e “partial match”, optou-se que o segundo conjunto de validações fosse realizada somente por um avaliador, evitando sobrecarga de trabalho que pode influenciar no resultado devido a cansaço e tempo dispendido. Os 50 termos compostos restantes foram avaliados por A2, e A3 avaliou os demais 50 termos simples. A distribuição dos termos validados por cada especialista é indicado no Quadro 6. O avaliador A3 possui maior experiência com os termos compostos da SCT, por isto foi atribuído a este a validação individual dos termos compostos, e o avaliador A2 é entre eles o com maior experiência no mapeamento para a SCT.

Após as validações dos especialistas, os pesquisadores em conjunto com os três anotadores realizaram adjudicação e discussão para definição do consenso nos resultados discordantes, decidindo entre “exact match” ou “partial match”. O objetivo foi possibilitar a inclusão dos discordantes nos cálculos das métricas.

Os termos classificados como “not match”, passaram por uma avaliação pelo autor para verificar se estes se tratavam de falsos negativos ou falsos positivos. Para cada um destes foi tentado o mapeamento manual para um conceito SNOMED CT,

se encontrado um conceito correspondente o mapeamento automático era considerado falso negativo, se não encontrado era considerado falso positivo.

Quadro 6 – Número de termos validados por cada

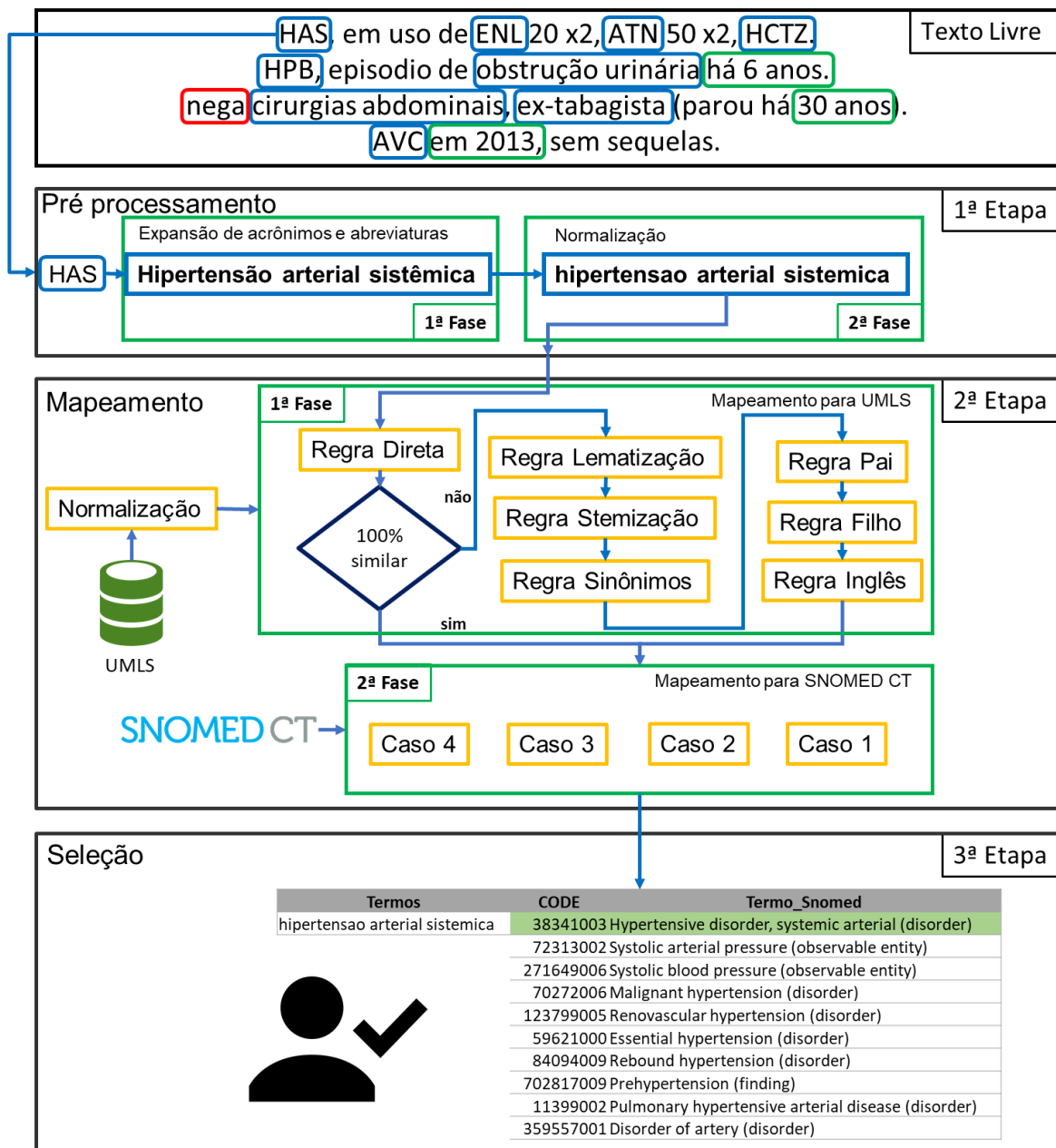
Termos	Termos simples	Termos compostos
50	A1 A3	A1 A2
50	A2	A3
	100	100

Fonte: Elaborado pelo autor.

4. RESULTADOS

O principal resultado deste estudo é o método proposto (MapClin), representado por completo na Figura 13. Um resultado secundário foi o mapeamento gerado pelo experimento final e pelo experimento parcial.

Figura 13 – MapClin

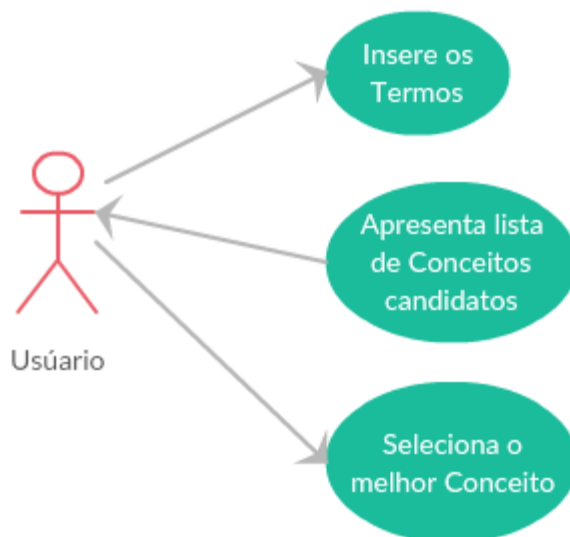


Fonte: Elaborado pelo autor.

O MapClin mapeia uma lista de termos clínicos para a SCT, os quais passam inicialmente por uma etapa de pré-processamento, nela, os acrônimos e as abreviaturas são expandidos e é realizada a normalização dos termos. A etapa seguinte mapeia os termos de entrada para conceitos UMLS, através de 7 regras desenvolvidas de acordo com os padrões observados na base de treinamento. A última etapa realiza o mapeamento dos conceitos UMLS para a SCT e, por fim, os conceitos SCT são apresentados ao usuário, o qual seleciona o melhor conceito para representar cada termo de entrada.

Para representar as funcionalidades do MapClin, é apresentado, na Figura 14, diagrama de casos de uso.

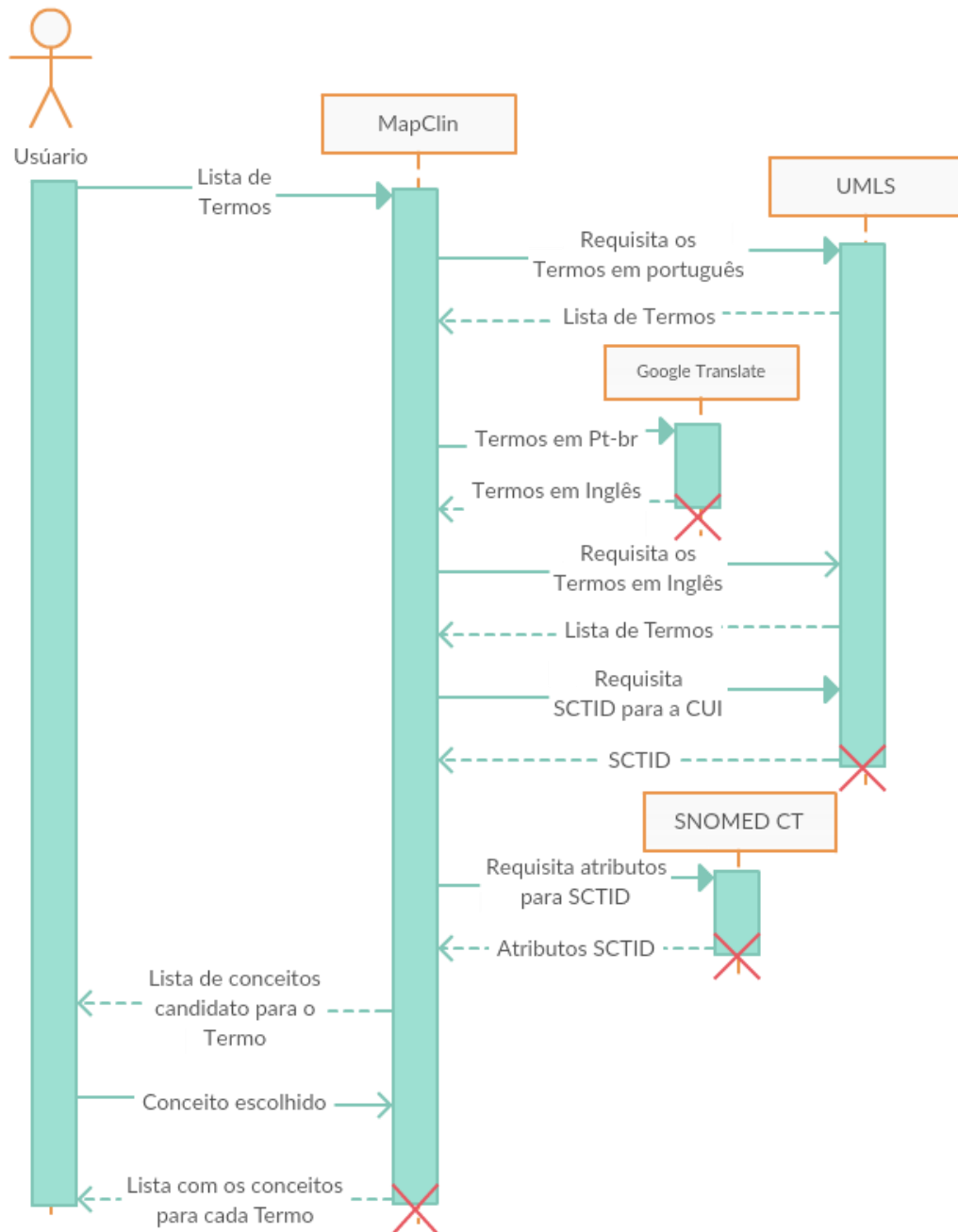
Figura 14 – Diagrama de casos de uso



Fonte: O autor.

A Figura 15 representa o diagrama de iteração do Mapclin, o qual demonstra as principais iterações deste com outros sistemas, bancos de dados e o usuário em uma sequência cronológica.

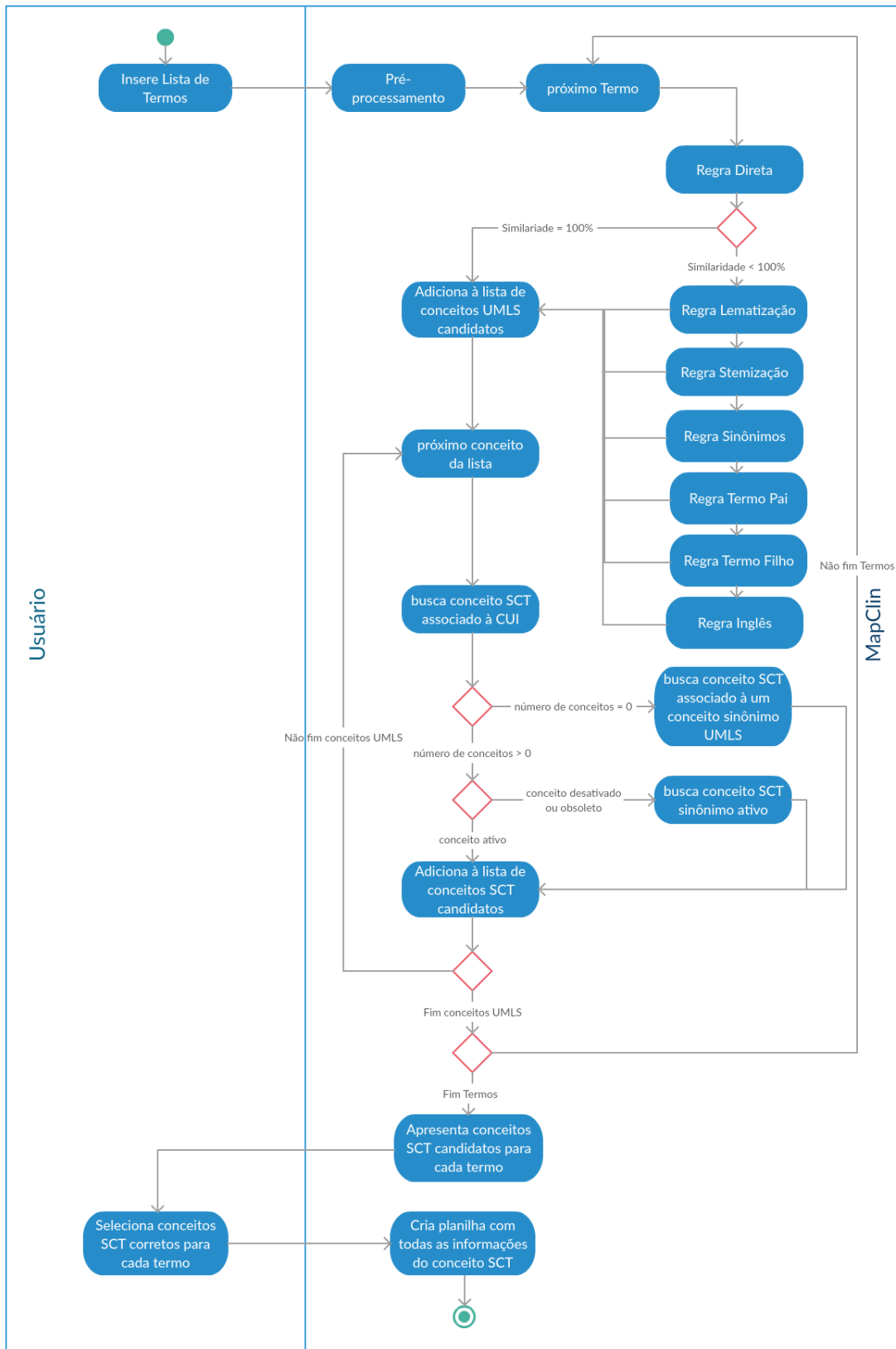
Figura 15 – Diagrama de iteração



Fonte: Elabora pelo autor.

Por último, é apresentado, na Figura 16, o diagrama de atividades do MapClin com as principais atividades desenvolvidas pelo sistema e pelo usuário. Nas seções seguintes deste capítulo, é explicada, com mais detalhes, cada uma das atividades desenvolvidas e, na seção 4.5, é apresentado o resultado do experimento parcial e na seção 4.6 o resultado obtido na validação do MapClin.

Figura 16 – Diagrama de atividades



Fonte: Elaborado pelo autor.

4.3. ETAPA 1 – PRÉ-PROCESSAMENTO

Na etapa de pré-processamento, são realizadas modificações nos termos de entrada e nos termos da UMLS. Essa etapa é composta por duas fases: expansão de acrônimos e abreviaturas e normalização dos termos. Na expansão, é utilizada a lista de acrônimos e abreviaturas para realizar o processo, por exemplo, o termo “IAM” é expandido para “Infarto Agudo do Miocárdio”. Essa fase, especificamente, é realizada apenas para os termos de entrada, os quais são oriundos das anotações realizadas nas narrativas clínicas.

A segunda fase realiza a normalização dos termos e é composta por 3 passos. O primeiro é a remoção de acentos e caracteres especiais, o segundo altera todas as letras em maiúsculo para minúsculo e, por último, é feita a remoção das *Stop Words*. O Quadro 7 mostra as *Stop Words* removidas. Um exemplo dessa normalização pode ser observado na Figura 17, na qual o termo “Infarto Agudo do Miocárdio” no fim da etapa é representado como “infarto agudo miocardio”.

Quadro 7 – *Stop Words* removidas dos termos de entrada

NE	de	do	Dos	da	das	uma	quem	Por
o	as	a	As	com	para	seu	uns	Umas
e	ou	pra	na	nas	no	pelo	como	Sua
nos	ao	aos	em	que	um	pela		

Fonte: Elaborado pelo autor.

Figura 17 – Pré-processamento do termo “IAM”

Pré-processamento	Ex:	IAM
expansão de acrônimos e abreviaturas		Infarto Agudo do Miocárdio 1ª etapa
normalização: 1. Remoção de acentuação e caracteres especiais; 2. Alterar letras em maiúsculo para minúsculo; 3. Remoção de stopwords;		Infarto Agudo do Miocardio infarto agudo do miocardio infarto agudo miocardio 2ª etapa

Fonte: Elaborado pelo autor.

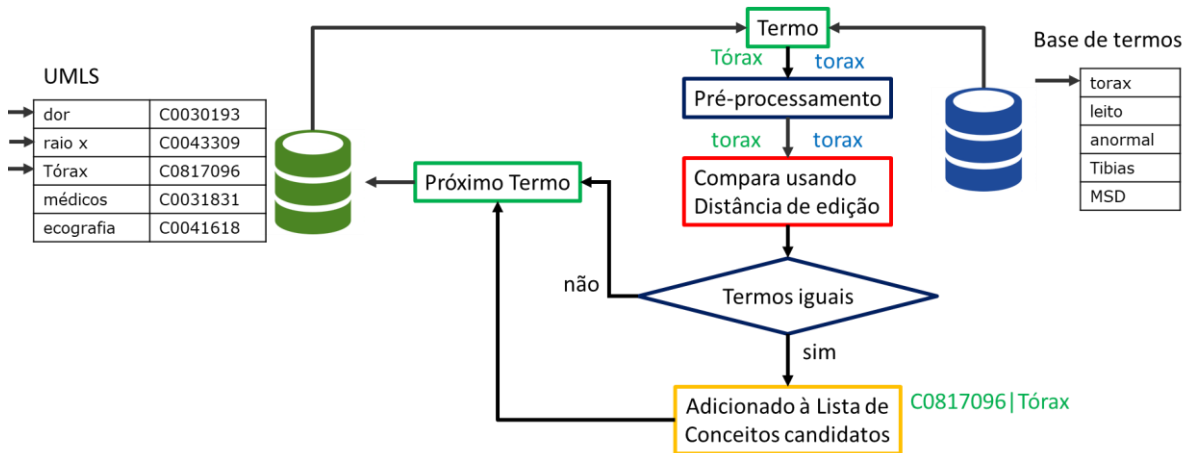
Regra Direta

Essa regra consiste em comparar o termo de entrada pré-processado com os termos da UMLS também pré-processados, usando o algoritmo de Distância de edição de *Levenshtein*. Se a distância entre o termo de entrada e um ou mais termos UMLS for igual a 0, os conceitos UMLS que eles representam são selecionados como conceitos candidatos e as demais regras não são executadas (critério de parada). Esse processo só será executado para essa regra, as demais são executadas de forma sequencial, sem um critério de parada ou não execução das próximas.

Quando não houver nenhum termo UMLS com uma distância de edição igual a 0 em relação ao termo de entrada, é realizado o cálculo de similaridade entre eles. Para os termos UMLS que a similaridade for maior ou igual a 90%, o conceito UMLS que o representa, é adicionado à lista de conceitos candidatos que podem representar o termo de entrada, e a comparação continua para o próximo termo UMLS.

A Figura 18 apresenta um exemplo do uso da Regra Direta, em que a “Base de termos” representa os termos de entrada extraídos das narrativas clínicas. No exemplo, o termo de entrada “torax” e o termo UMLS “Tórax” passam pelo pré-processamento e são comparados, apresentando uma distância de edição igual a 0, o que faz com que a ferramenta os considere como iguais, e o conceito “C0817096|Tórax” seja adicionado à lista de conceitos candidatos para representar o termo de entrada. Nesse exemplo, quando a regra terminar de percorrer todos os termos UMLS, não são executadas as próximas regras, visto que existe, ao menos, um termo com distância de edição igual a 0 para com o termo de entrada.

Figura 18 – Representação da Regra Direta para o termo “torax”



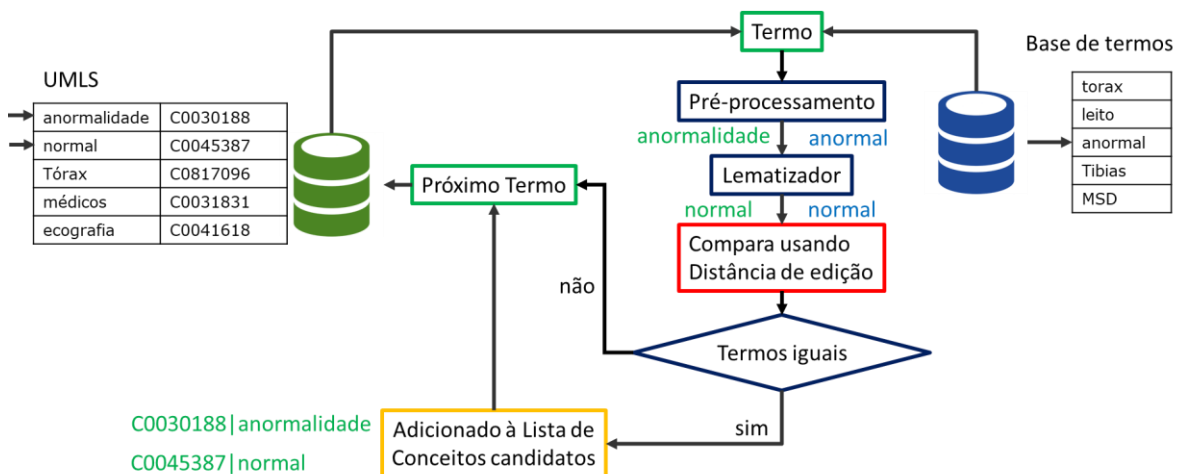
Fonte: Elaborado pelo autor.

Regra Lematização

A ideia, nessa regra, é comparar o lema do termo de entrada com o dos termos UMLS, aqueles que apresentarem uma similaridade maior que 90% são adicionados à lista de conceitos candidatos.

A Figura 19 apresenta um exemplo do uso da Regra Lematização, o termo de entrada “anormal” possui o lema “normal”, ao comparar com os da UMLS, é encontrado o mesmo lema para “anormalidade” e “normal”. Para tal caso, então, os dois conceitos “C00300188|anormalidade” e “C0045387|normal” são adicionados à lista de conceitos candidatos que podem representar o termo “anormal”.

Figura 19 – Representação da Regra Lematização para o termo “anormal”



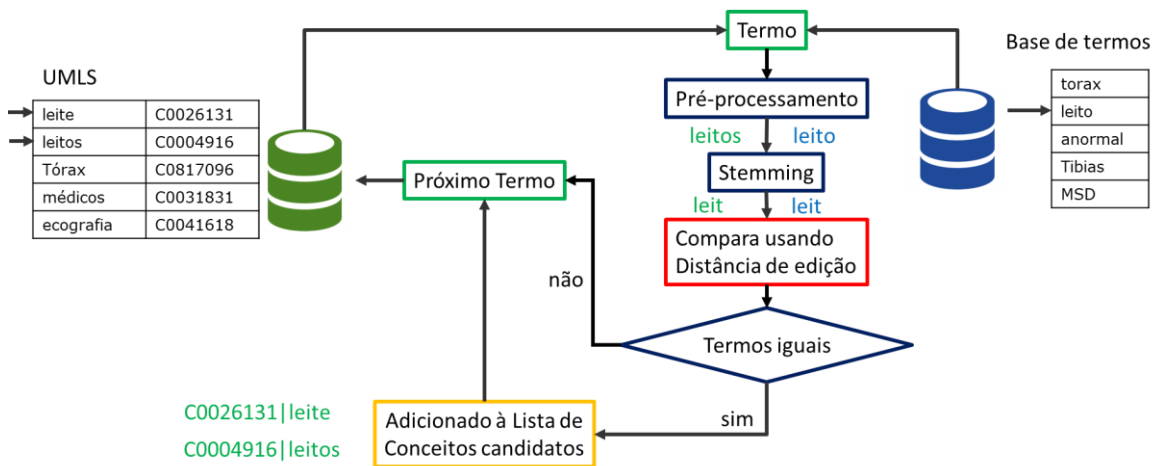
Fonte: Elaborado pelo autor.

Regra Stemização

O objetivo da regra é comparar as raízes ortográficas dos termos UMLS com a de entrada, aqueles que possuem uma similaridade maior que 90% são adicionados à lista de conceitos candidatos.

A Figura 20 apresenta um exemplo de uso da Regra Stemização, o termo de entrada “leito” possui a raiz ortográfica “leit”. Ao comparar com os termos da UMLS, são encontrados dois termos com essa mesma raiz, os termos “leite” e “leitos”. Para esse caso, então, os dois conceitos “C0026131|leite” e “C0004916|leitos” são adicionados à lista de conceitos candidatos que podem representar o termo “leito”.

Figura 20 – Representação da Regra Stemização para o termo “leito”



Fonte: Elaborado pelo autor.

Regra Sinônimos

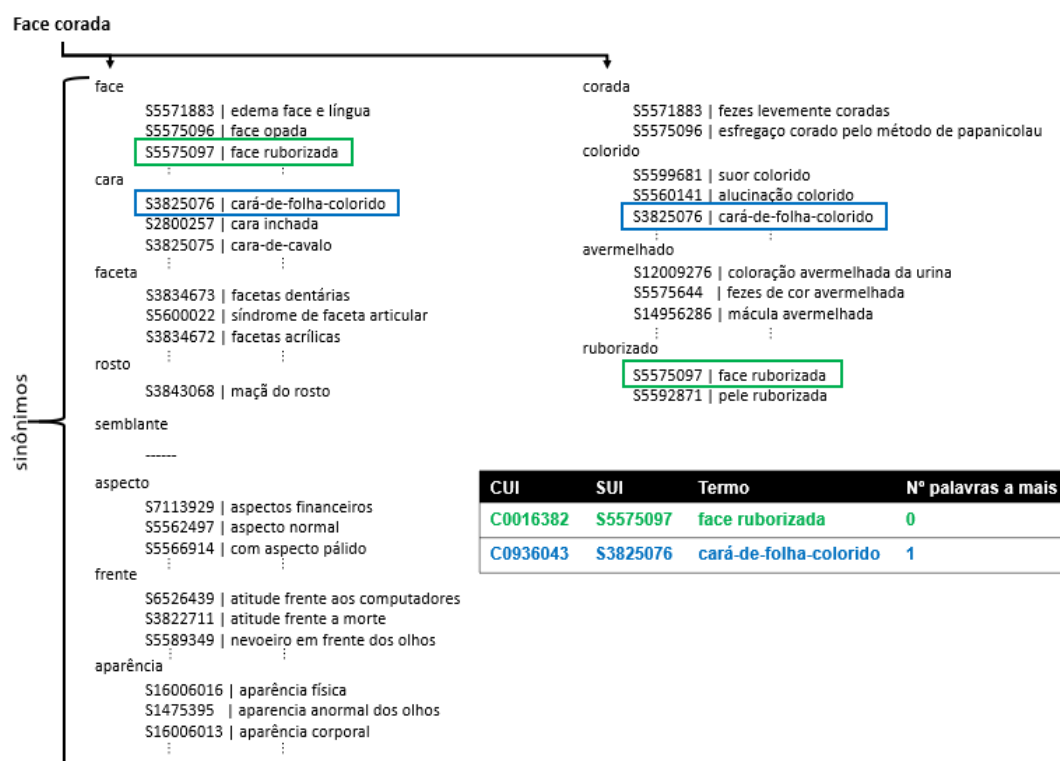
São gerados sinônimos ortográficos dos termos de entrada usando um dicionário de sinônimos para português brasileiro (dicio.com.br), formando um conjunto de sinônimos. Se o termo de entrada for único, o conjunto de sinônimos deste é comparado com os termos UMLS em português usando as 3 regras anteriores, gerando um conjunto de conceitos UMLS candidatos para representar o termo de entrada.

Se o termo de entrada for composto, é necessário dividi-lo em termos únicos, dado que o dicionário de sinônimos atua apenas para estes. São formados, assim, diferentes conjuntos de sinônimos para cada palavra do termo de entrada. Cada

conjunto de sinônimos gerado é comparado com a tabela MRXW_POR da UMLS, usando as 3 regras anteriores, sempre que um dos sinônimos for encontrado na tabela o código SUI do termo UMLS será salvo. No fim das comparações, tem-se um conjunto de códigos SUI para cada palavra do termo de entrada. É verificado se existe algum SUI que esteja presente em todos os conjuntos, em caso afirmativo, é comparada a quantidade de palavras do termo UMLS que tem o código SUI e o termo de entrada, caso possuam o mesmo número de palavras, o conceito que detém aquele SUI é adicionado à lista de conceitos candidatos para representar o termo de entrada.

Um exemplo da Regra Sinônimos é apresentado na Figura 21. O termo de entrada “face corada”, não encontrado pelas regras anteriores, é dividido em palavras, para cada palavra é gerada uma lista de sinônimos. Para cada sinônimo, buscam-se os termos UMLS em que ele aparece. Os termos UMLS, que aparecem na lista de sinônimos de todas as palavras, são adicionados como candidatos. É comparado o número de palavras a mais que o termo UMLS possui em relação ao termo de entrada, não levando em conta as *Stop Words*. Os conceitos dos termos, que apresentam a menor quantidade, são adicionados à lista de conceitos candidatos para representar o termo de entrada que, nesse caso, é o conceito “C0016382|face ruborizada”.

Figura 21 – Representação para Regra Sinônimos para o termo “face corada”



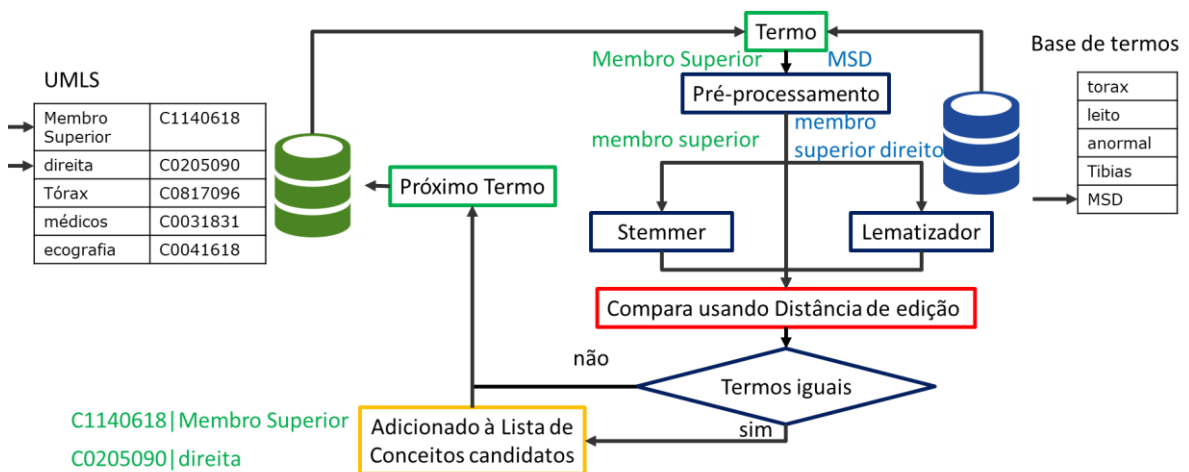
Fonte: Elaborado pelo autor.

Regra Pai

Essa regra tem por finalidade encontrar um termo UMLS com um significado semântico mais abrangente que o de entrada. Só é usada para termos de entrada compostos de duas ou mais palavras. Para tanto, o termo de entrada é dividido em palavras, por exemplo “membro superior direito” é dividido em “membro”, “superior” e “direito”. Usando, desse modo, as regras anteriores é buscado, na tabela mrwx_POR, o código SUI de cada termo que possui as palavras do termo de entrada. Os termos UMLS que tem seu código SUI encontrado o mesmo número de vezes em que a quantidade de palavras que estes contêm, é considerado termo pai do termo de entrada e esses são adicionados à lista de conceitos candidatos.

A Figura 22 apresenta um exemplo da Regra Pai, o termo “MSD”, abreviatura de “membro superior direito”, não possui um termo UMLS que o represente de forma exata, no entanto, possui termos contidos nesse como os termos “Membro Superior” e “direita” que, nessa regra, são adicionados à lista de termos candidatos para representar o termo “MSD”. Uma observação a ser feita nesse exemplo é a de que se existisse na UMLS o termo “membro direito” também seria adicionado à lista de termos candidatos, visto que a ordem não faz diferença, e o que é levado em consideração são as palavras que possuem em comum.

Figura 22 – Representação da Regra Pai para o termo “MSD”



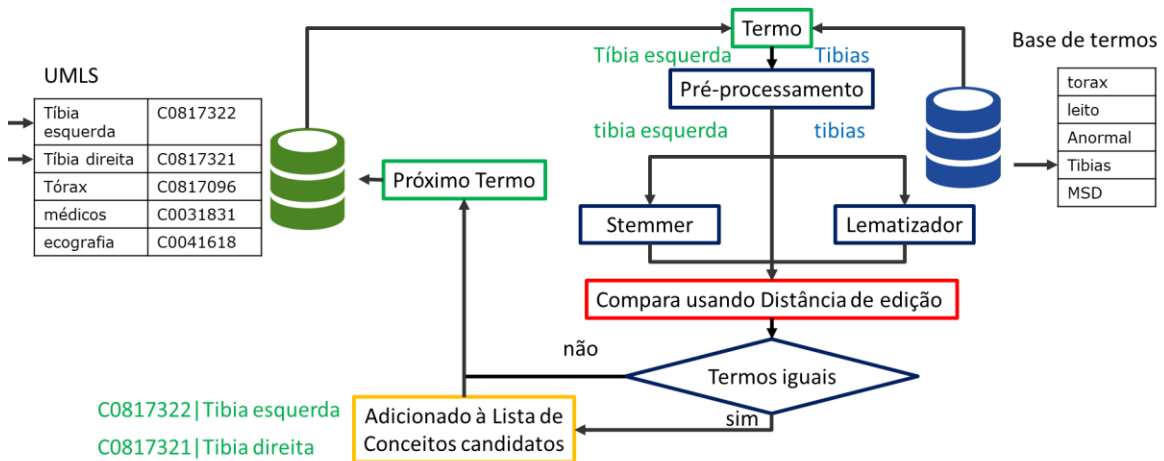
Fonte: Elaborado pelo autor.

Regra Filho

A regra tem por objetivo encontrar um termo UMLS com um significado semântico mais específico que o termo de entrada. Para isso, é usada a mesma estratégia da regra anterior, contudo, para ela são buscados termos UMLS que tenham sua SUI encontrada o mesmo número de vezes de palavras que o termo de entrada contém, então, os termos UMLS, que possuem o menor número de palavras em relação ao termo de entrada, são adicionados à lista de termos candidatos.

A Figura 23 apresenta um exemplo da Regra Filho, o termo “Tibias” não possui um termo UMLS que o represente de forma exata, mas, termos com um significado semântico mais específico que o seu.

Figura 23 – Representação da Regra Filho para o termo “Tibias”



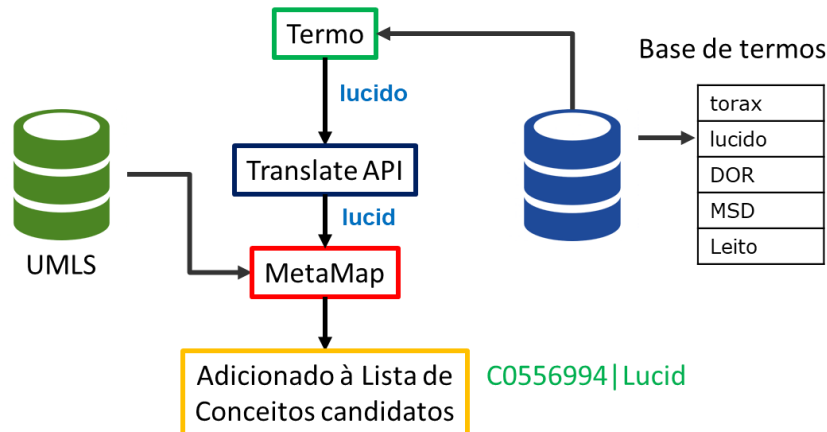
Fonte: Elaborado pelo autor.

Regra Inglês

Para essa regra, o termo em linguagem natural é traduzido para o inglês usando uma API do Google Translate. O termo em inglês é mapeado para a UMLS realizando a busca de forma direta no *Methatesaurus* e através do MetaMap, os conceitos resultantes são adicionados à lista de conceitos candidatos.

Na Figura 24, é apresentado um exemplo da Regra Inglês, o termo de entrada “lucido”, traduzido para o inglês como “lucid”, é mapeado pelo MetaMap para o conceito “C0556994|Lucid”, o qual é adicionado à lista de conceitos candidatos para representar “lucido”.

Figura 24 – Representação da Regra Inglês para o termo “lucido”.



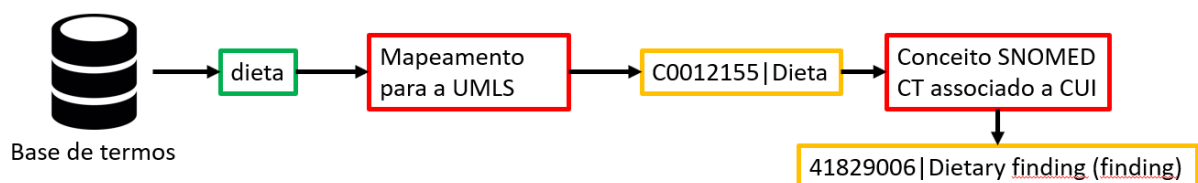
Fonte: Elaborado pelo autor.

4.3.1. FASE 2 – MAPEAMENTO DA UMLS PARA A SNOMED CT

Nessa fase, são mapeados os conceitos candidatos encontrados na UMLS para conceitos SCT. Para cada CUI dos conceitos candidatos, foi buscado no campo SAB, o qual se refere ao nome da terminologia de origem na UMLS, algum termo originário da SCT e, ao encontrar, foi usado o campo CODE, o qual guarda o código daquele conceito na terminologia de origem para identificar este conceito SCT. O que pode apresentar uma das quatro situações a seguir:

- a) Caso 1: Existe apenas um conceito SCT. Para isso, foi buscado o código dele conceito no SCT, representado pelo campo CODE e, em seguida, por meio de uma API da SCT encontradas as demais informações para o conceito em questão. Essa é a melhor situação possível para o mapeamento. Esse processo está representado na Figura 25.

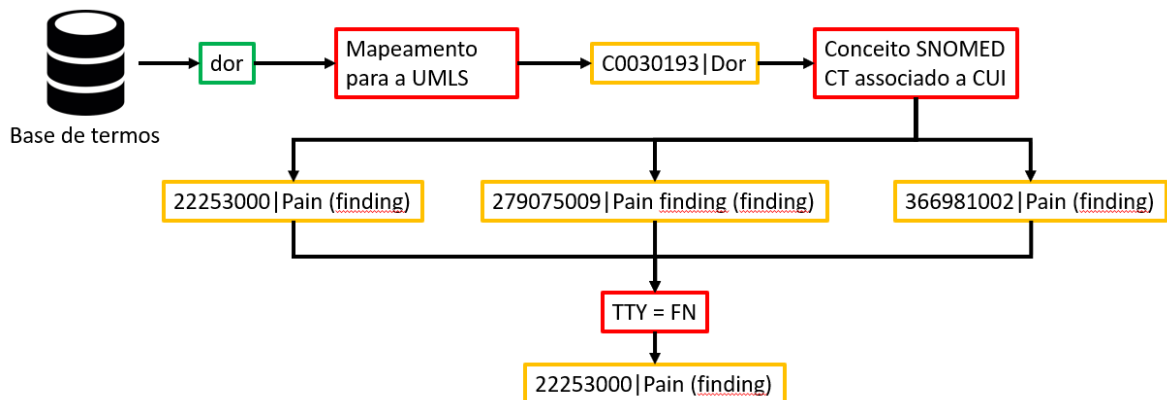
Figura 25 – Procedimento quando houver apenas um conceito SCT associado



Fonte: Elaborado pelo autor.

- b) Caso 2: Existe mais de um conceito SCT. Nessa situação, foi escolhido o conceito que na UMLS possui o valor “FN” no campo TTY, o qual representa que este é o termo *Fully Specified Name* (FSN) na SCT. Se este não possuir um termo com o valor “FN” e os termos não contenham os valores “OF”, “OAS”, “OAP” ou “IS” no campo TTY, os quais representam conceitos obsoletos ou conceitos que não são mais utilizados pela SCT, todos os conceitos restantes são apresentados como candidatos para representar o termo de entrada. Esse processo é representado na Figura 26.

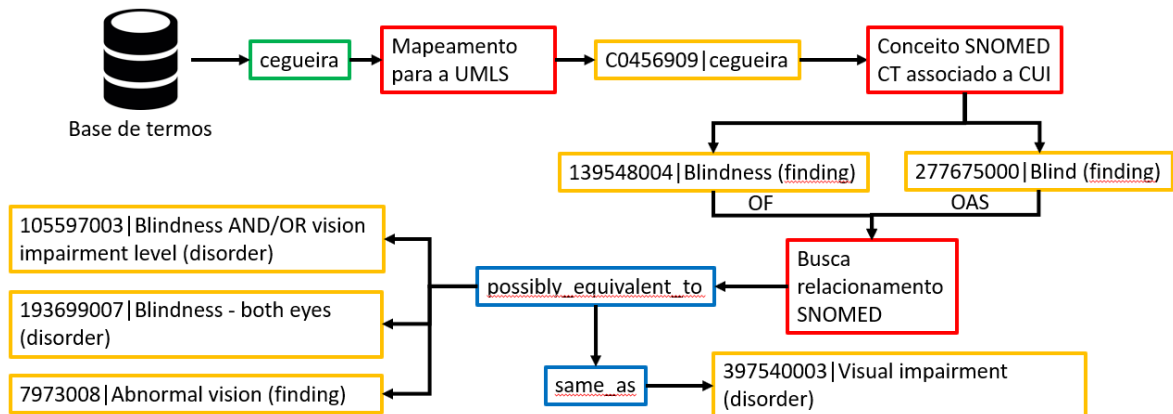
Figura 26 – Procedimento quando houver mais de um conceito SCT associado



Fonte: Elaborado pelo autor.

- c) Caso 3: Existe apenas conceitos SCT obsoletos ou desativados associados a CUI, neste caso verifica se se este possui na SCT um relacionamento do tipo “*possibly_equivalent_to*” ou “*same_as*” ou um relacionamento na UMLS do tipo “SY”. Se conter um desses relacionamentos, e o conceito relacionado conter um conceito SCT associado, este conceito é apresentado como conceito candidato. Este processo pode ser visualizado na Figura 27.

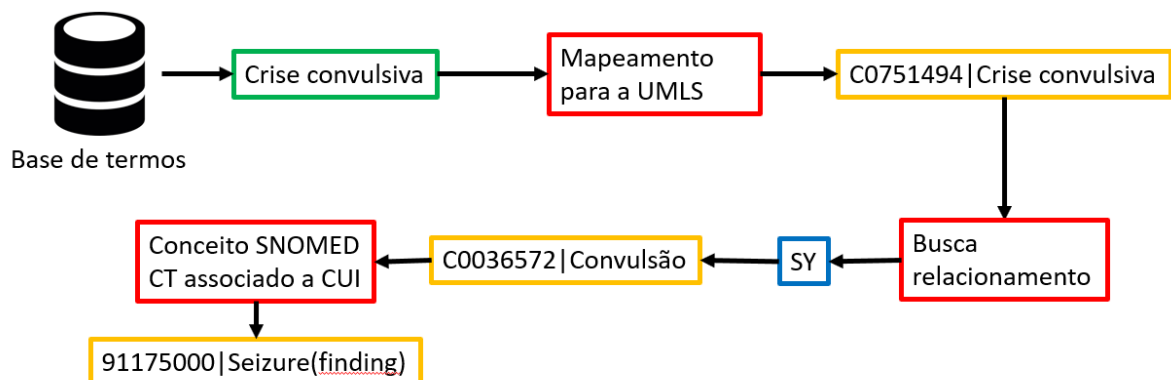
Figura 27 – Procedimento quando houver conceitos SCT associados obsoletos ou desativados



Fonte: Elaborado pelo autor.

- d) Caso 4: Não existe nenhum conceito SCT associado. Para essa situação foi buscado nos relacionamentos da UMLS um termo associado a CUI mapeada que contenha um conceito SCT associado e assim considerar esse termo como o correspondente na SCT. Essa etapa é demonstrada na Figura 28.

Figura 28 – Procedimento quando não houver nenhum conceito SCT associado



Fonte: Elaborado pelo autor.

4.4. ETAPA 3 – SELEÇÃO DOS TERMOS

A seleção é realizada com base nos conceitos candidatos encontrados para cada termo. É verificado se, dentre os conceitos candidatos, existe um ou mais conceitos que representem o termo de entrada de forma completa ou de forma parcial, ou se não existe nenhum conceito capaz de representar o termo de entrada.

Um exemplo da etapa de seleção, para o termo de entrada “Membro superior esquerdo”, é apresentado no Quadro 8. O método encontrou 22 conceitos SCT, que poderiam ser possíveis candidatos para representá-lo. O avaliador, dessa maneira, seleciona o conceito entre os 22 que melhor represente o termo de entrada, nesse caso, é o primeiro conceito que está apresentado em destaque no Quadro 8.

Quadro 8 – Exemplo de seleção

Termo Linguagem Natural	Código	Descrição conceito SCT
membro superior esquerdo	80768000	Structure of left upper limb (body structure)
	371195002	Bone structure of upper limb (body structure)
	53120007	Upper limb structure (body structure)
	41764006	Monoplegia of upper limb (disorder)
	62100001	Juvenile osteochondrosis of upper extremity (disorder)
	23406007	Fracture of upper limb (disorder)
	400136002	Amputation of upper limb (procedure)
	74170002	Radiography of upper limb (procedure)
	298747001	Deformity of upper limb (finding)
	102558002	Edema of the upper extremity (finding)
	61599003	Phlebitis (disorder)
	27550009	Disorder of blood vessel (disorder)
	49601007	Disorder of circulatory system
	155461000	Circulatory system disease NOS
	195646003	Circulatory system disease NOS
	155494002	Circulatory system disease NOS
	266275004	Circulatory system diseases
	155263000	Circulatory system diseases
	194707003	Circulatory system diseases
	449671007	Cellulitis of upper limb (disorder)
	95673003	Paresthesia of upper limb (finding)
	249944006	Monoparesis - arm (disorder)

Fonte: Elaborado pelo autor.

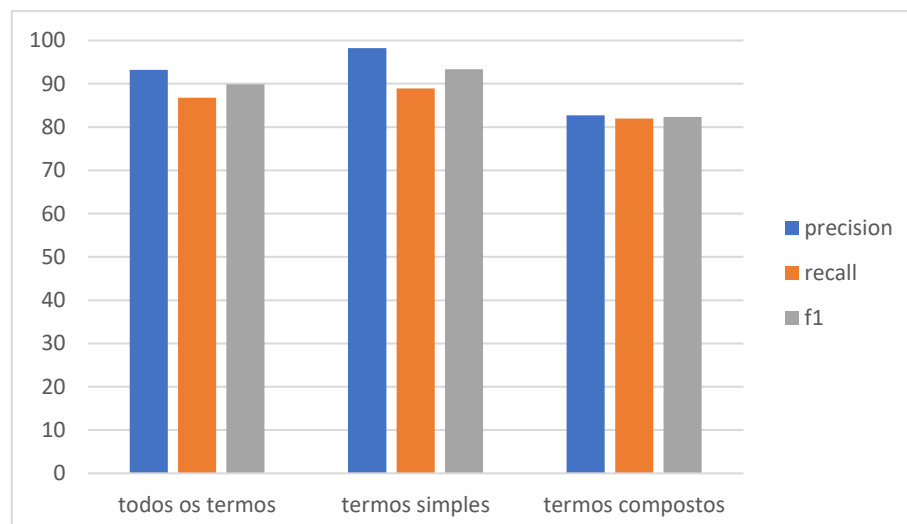
4.5. RESULTADO EXPERIMENTO PARCIAL

Com intuito de avaliar as regras desenvolvidas na metade do processo, foi realizado um experimento com as regras criadas até o momento (Regras Direta, Lematização, Stemização e tradução do termo para o inglês) com os termos da base de treinamento. O processo de avaliação dos resultados gerados pelo experimento foi feito pelo autor e por um profissional da área da saúde.

Dos 216 termos o algoritmo encontrou ao menos um conceito candidato ou mais para 207 termos: 146 termos foram considerados como mapeamento exato, 21 considerados como mapeamento parcial e 47 considerados como não mapeados. Dos termos que não foram mapeados 14 termos simples e 10 termos compostos quando mapeados de forma manual foram associados a um conceito SCT, os outros mesmo de forma manual não apresentaram um conceito SCT para representá-los.

Os termos analisados apresentaram um *precision* de 93,2%, *recall* de 86,8% e *F-score* de 89,8%. Analisando separadamente os termos simples e compostos, têm-se que 110 dos termos simples foram considerados como mapeamento exato, 4 como parcial e 22 como não mapeados, com *precision* de 98,25%, *recall* de 88,9% e *F-score* de 93,3%. Dos termos compostos, 36 foram considerados mapeamento exato, 19 parcial e 25 não mapeados, com uma *precision* de 82,7% um *recall* de 82% e um *F-score* de 82,35%. Na Figura 29 é apresentado o gráfico com estes resultados para todos os termos analisados.

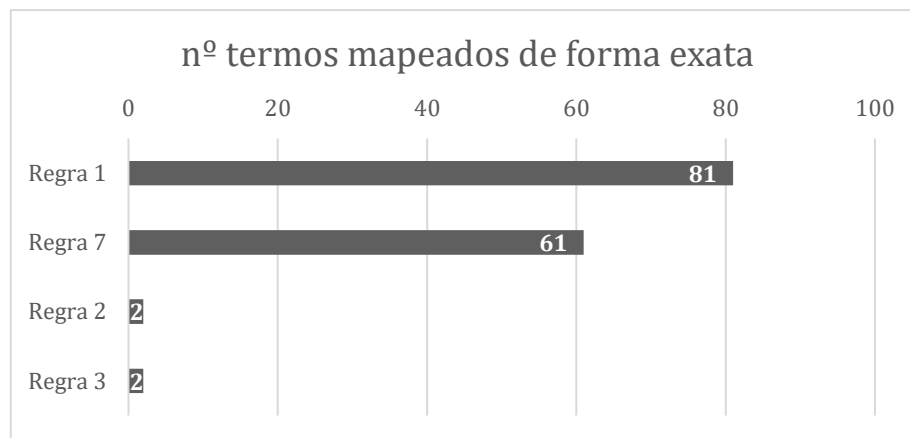
Figura 29 - Recall, Precision e F- score para “todos os termos”, “termos simples” e “termos compostos”.



Fonte: Elaborado pelo autor.

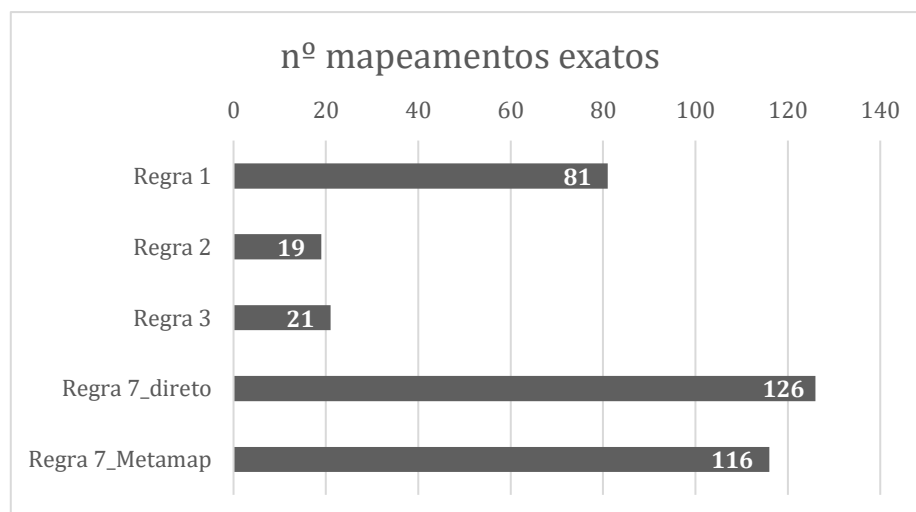
Para os 146 termos mapeados de maneira exata foi feita uma análise mais detalhada, a fim de verificar quais as regras que causavam maior impacto no resultado até então. As análises foram separadas em duas classificações, a primeira levou em conta a prioridade para o mapeamento que foi usada para cada regra, e a segunda fez a análise sem levar em conta a prioridade, verificando para cada regra quantos termos ela mapeou para o conceito considerado como mapeamento exato pelo especialista. Os números feitos por esta análise são apresentados nas Figuras 30 e 31. Por último também foi verificado que dos 61 termos mapeados em inglês, 9 foram mapeados de forma direta em inglês outros 9 foram mapeados pelo MetaMap e os outros 43 foram mapeados por ambos os métodos.

Figura 30 - Frequência de termos simples escolhidos.



Fonte: Elaborado pelo autor.

Figura 31 - Frequência de termos compostos escolhidos.



Fonte: Elaborado pelo autor.

4.6. RESULTADOS DA VALIDAÇÃO

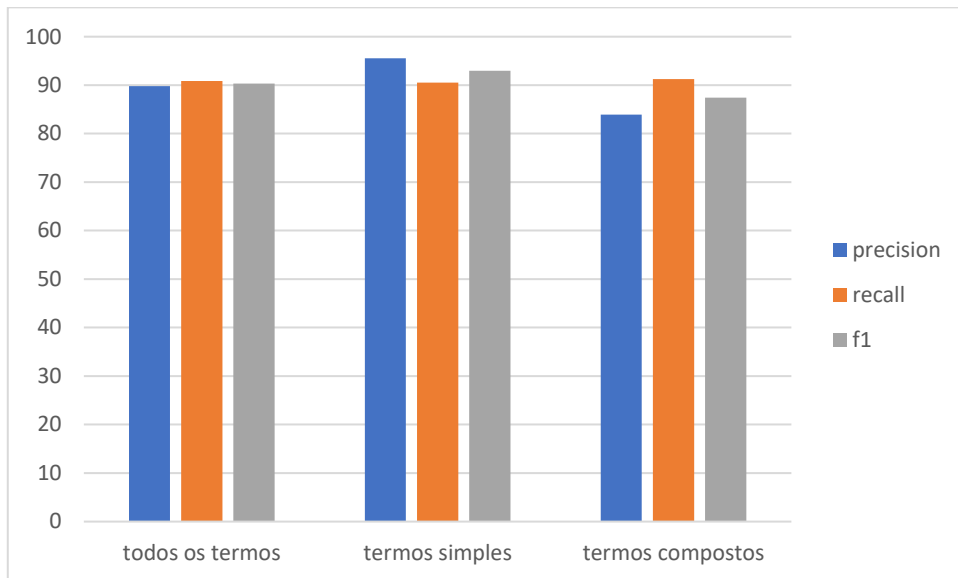
Na validação realizada por 2 anotadores para os 50 termos simples no experimento final, foram identificadas diferentes avaliações em 4 termos: “abdômen”, “abdome”, “limpo” e “orientado”; foram avaliados como mapeamento parcial por um anotador e como mapeamento exato por outro. Nos 50 termos compostos houve discordância na avaliação de 5 termos: “acesso venoso periférico”, “membro inferior esquerdo”, “membro inferior direito”, “angioplastia coronariana transluminal percutânea” e “tubo orotraqueal”, marcados como mapeamento parcial por um anotador e como mapeamento exato por outro. As discordâncias foram resolvidas através de uma conversa entre os avaliadores e o autor deste estudo, chegando-se a uma decisão final para cada um.

No experimento final, realizado para a validação do método, foram mapeados de forma exata 59 termos compostos (APÊNDICE A) e 82 termos simples (APÊNDICE E). O mapeamento de forma parcial ocorreu em 28 termos compostos (APÊNDICE B) e 8 termos simples (APÊNDICE F). Não foram mapeados 13 termos compostos (APÊNDICE C) e 10 termos simples (APÊNDICE G). Para estes últimos foi feito um mapeamento manual pelo autor deste estudo (APÊNDICE D e H), resultando em 7 termos compostos e 8 simples mapeados para a SCT, que são considerado falsos negativo; e 6 termos compostos e 2 simples não mapeados, verdadeiros negativos.

Os termos simples apresentaram 8 falsos negativos, em 7 o conceito SCT associado a este de forma manual não possuía um correspondente em Pt-BR na UMLS, e o outro caso é o termo “meticorten” que de forma manual pode ser mapeado para “prednisona” que é a principal substância do meticorten. Para os termos compostos houve 7 falsos negativos, em 6 o conceito SCT associado a este de forma manual não possuía na UMLS um correspondente em Pt-BR, o sétimo caso corresponde ao termo “Pronto Atendimento” mapeado manualmente para “Pronto Socorro” ou “Serviços Médicos de Emergência”, que são representados pelo CUI “C0013961”.

O experimento final obteve para os termos simples uma *precision* de 95,5% um *recall* de 90,5% e um F- score de 93%, para os termos compostos uma *precision* de 84% um *recall* de 91,25% e um F- score de 87,4%, e para todos os termos a *precision* ficou em 89,8%, o *recall* em 90,8% e o F- score em 90,3%, estes números podem ser observados na Figura 32.

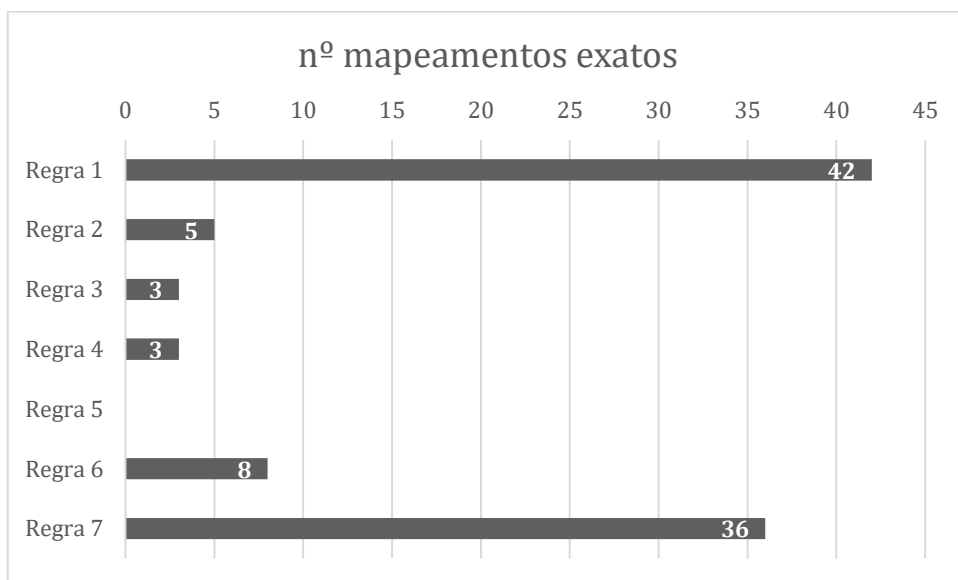
Figura 32 - Precision, recall e F- score para o experimento final.



Fonte: Elaborado pelo autor.

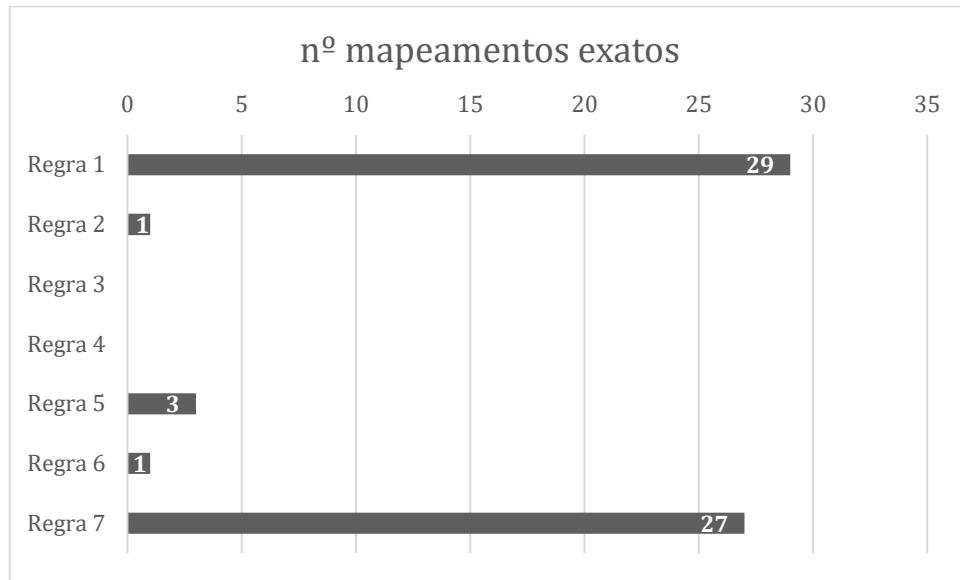
Para os 59 termos compostos e os 82 termos simples mapeados de forma exata, foram realizadas análises mais aprofundadas para verificar a influência de cada regra, projetada no resultado do experimento. Esta análise verificou quantos termos cada regra mapeou para o conceito SCT, que o anotador considerou como representante semântico. Os resultados destas análises são apresentados nas Figuras 33 e 34.

Figura 33 - Frequência de termos simples escolhidos.



Fonte: Elaborado pelo autor.

Figura 34 - Frequência de termos compostos escolhidos.



Fonte: Elaborado pelo autor.

5. ANÁLISES DOS RESULTADOS E DISCUSSÃO

A primeira etapa do método, o pré-processamento, é uma parte essencial, pois os termos extraídos são oriundos de texto livre de narrativas elaboradas em cenários reais, contendo alguns erros de ortografia. É importante que o método proposto seja capaz de processar textos com erros ortográficos para estar adequado a realidade dos RES. A retirada de acentuação minimiza parte do problema de diferenças ortográficas. É possível ainda realizar correção ortográfica dos textos antes de iniciar o processo de mapeamento. Além disto, os textos contêm diversas abreviaturas e acrônimos, que foram possíveis expandir. A lista de abreviaturas é um compilado de diversos conjuntos elaborados em outros estudos e foi complementada pelos anotadores a partir das narrativas dos hospitais do Grupo Marista, o que permitiu a expansão de todos os termos abreviados ou em acrônimo. Em alguns hospitais existe um siglário específico, que pode ser usado pelo MapClin para realizar a expansão das abreviaturas e acrônimos dos textos oriundos destes outros hospitais.

O mapeamento, segunda etapa do MapClin, apresentou os melhores resultados nos dois experimentos (parcial e final) tanto na Regra Direta quanto na Regra Inglês. O critério de parada de execução do método incluído na Regra Direta eficiente no experimento parcial, no final ocasionou em 2 mapeamentos com erros (aproximadamente 3% dos casos). Isto ocorreu quando no mapeamento UMLS- SCT chegou ao Caso 4, e o sinônimo UMLS relacionou conceitos com significado semântico diferente do conceito candidato. Uma solução para este problema é criar uma exceção neste caso, que execute as outras regras, para gerar outros conceitos candidatos. Neste Caso 4 o ideal é que o critério de parada seja ignorado, isto é, que o mapeamento continue sendo realizado pela aplicação das demais regras.

As regras 2 e 3 não mostraram grande influência em nenhum dos dois experimentos, porém ao analisar os termos mapeados por estas, nota-se o seu valor para termos com grau ou gênero diferentes do termo definidor do conceito na terminologia destino. Isto fez com que estas fossem mantidas no método proposto. O experimento final foi o que estas menos contribuíram, um possível acarretador, pode ser a forma de seleção dos termos, pois ao selecionar os termos com maior frequência, evitam-se termos mais problemáticos com variações ortográficas, para as quais estas regras foram projetadas.

A Regra Sinônimos criada para o experimento final, não influenciou na melhora dos resultados, podendo ser excluída do MapClin, sem afetar o mapeamento. A Regra Pai não possui influência para com os termos simples, já que o intuito desta é encontrar um termo com menos palavras que o termo de entrada, isso faz com que esta seja aproveitada apenas para os termos compostos. A Regra Filho gerou mapeamentos parciais na sua maioria, os quais por vezes possibilitam facilitar a tarefa do especialista, direcionando para um possível termo filho.

Considerando-se a quantidade de termos mapeados pela Regra Inglês, 135 dos 142 termos que foram mapeados de forma exata no experimento parcial para os conceitos SCT considerados como representantes do termo de entrada. Apesar da quantidade de acertos, esta regra resulta e uma alta taxa de erro, devido a tradução errada dos termos para o inglês. O mapeamento a partir do termo traduzido para o inglês foi feito por dois métodos no experimento parcial, uma utilizando MetaMap e a outra buscando o termo em inglês de forma direta no Methatesaurus. Verificou-se que é vantajoso utilizar os dois métodos juntos para fazer o mapeamento, já que os dois mapearam termos distintos.

Os mapeamentos realizados pelo MapClin que foram identificados como falsos negativos reforçam a necessidade da inclusão de um maior número de termos clínicos em Pt-BR nas diferentes terminologias existentes. Por exemplo, incluir os 13 termos que não possuíam na UMLS um correspondente em Pt-BR, bem como a relação de “meticorten “e prednisona em um mesmo conceito. O uso de um guia farmacêutico poderia minimizar esses erros, pois este traz o nome comercial do medicamento e associa ao seu nome genérico. Por outro lado, estes resultados podem orientar na indicação de termos as serem incluídos nas terminologias. Este é um ponto em que o MapClin pode contribuir, em especial com a criação de uma versão da SCT para o português.

O mapeamento dos termos compostos ainda é um desafio. Uma possibilidade para apoiar esta tarefa é o uso de pós coordenações da SCT, que possibilitaria unir conceitos distintos para formar um novo conceito SCT. Porém, é necessário muito cuidado ao usar as pós-coordenações, pois estas podem substituir termos que já existem na SCT, podendo criar ambiguidades durante a utilização dos sistemas computacionais.

Comparando os resultados obtidos no experimento parcial (com 4 regras) na Figura 29, e o experimento final (com 7 regras) na Figura 32 é possível perceber uma

melhora com a aplicação das 7 regras, principalmente no *recall* dos termos compostos, que refletiu no *F-score* referente a eles. Porém, para os termos simples, a utilização das 7 regras gera uma menor *precision*. Em relação ao *F-score* não houve diferenças significativas.

Desta forma, conclui-se que para os termos simples aplicar as 4 regras iniciais já é suficiente. Porém, o mapeamento dos termos compostos podem ser beneficiados com a utilização das regras Pai e Filho. A Regra Sinônimos, relacionada aos sinônimos, não acrescentou melhora no mapeamento, além de tornar este processo mais demorado. Talvez se for utilizado conjunto de sinônimos mais específicos da área de saúde, esta regra possa apoiar melhor o mapeamento.

Os resultados do MapClin no experimento final são similares, especialmente em relação à *precision*, aos de estudos que apresentam métodos de mapeamento de termos clínicos, conforme pode ser observado no Quadro 9. O *recall* e o *F-score* são influenciados nos trabalhos relacionados ao SAMT (ALLONES; MARTINEZ; TABOADA, 2014) e o cTAKES (SAVOVA et al., 2010) pelos resultados do reconhecimento de entidades nomeadas, por eles também realizados. O MapClin o método de KIM (KIM, 2016) realizam apenas o mapeamento. As entidades nomeadas (p. ex. diagnósticos, procedimentos, sinais e sintomas) neste estudo já estão identificadas a partir da anotação, caracterizadas pela lista de termos.

Nos quatros estudos, além do MapClin, apresentados no Quadro 6 foram realizados mapeamento de termos em inglês. O estudo de ALLONES; MARTINEZ; TABOADA (2014), elaborou um método (SAMT) automático para fazer o reconhecimento de termos no texto e mapeá-los para a SCT. O SAMT foi aplicado de diferentes formas, incluindo ou não alguns passos como a lematização. Foram escolhidos os Setting 3 e o Setting 4 por terem apresentado os melhores resultados, e mantidos os dois na comparação por um ter a *precision* maior e o outro o *F-score*. KIM (2016) apresenta um método automático para o mapeamento entre a CIPE e a SCT, também utilizando a UMLS como forma de gerar variantes léxicas. SAVOVA et al. (2010) apresentam o cTAKES amplamente utilizada para realizar o mapeamento de termos clínicos para a UMLS, mas não faz referência a SCT.

Quadro 9 - Comparação entre diferentes métodos de mapeamento.

	MapClin	SAMT:Setting 3	SAMT:Setting 4	Kim	cTAKES
Precision	89,8%	88%	91,3%	79%	88,9%
Recall	90,8%	51,4%	42%	65%	76,7%
F-Score	0,9	0,71	0,66	0,71	0,824

Fonte: Elaborado pelo autor.

Pode-se considerar que o resultado do MapClin tem a vantagem em relação aos demais estudos de ser aplicado ao pt-BR, o que até então não foi descrito na literatura, sendo uma primeira abordagem para mapeamento de termos de narrativas clínicas neste idioma para a SCT. Os demais métodos resultam de estudos realizados há mais de uma vez, já incluindo aprimoramentos e um idioma para o qual são realizadas diversas iniciativas.

Um ponto que vale ressaltar é que os termos que foram mapeados pelo método, podem auxiliar na tarefa de tradução dos conceitos SCT para Pt-BR, além de terem sido extraídos de narrativas clínicas reais. Muitos estudos utilizam tradução direta de uma lista de termos da SCT em inglês ou outro idioma (RENATO, et al., 2018), para a proposição do MapClin foram utilizados uma lista dos termos que mais aparecem nas narrativas clínicas. Isto pode facilitar a inclusão do contexto para desambiguação e seleção do conceito candidato, isto pode ser desenvolvido em estudos futuros.

5.1. LIMITAÇÕES DO ESTUDO

O método apresenta um grande número de conceitos candidatos para alguns termos, dificultando o trabalho de análise do conceito que melhor representa o termo clínico. A necessidade da análise de um profissional especialista na terminologia destino para reduzir o resultado final para o conceito que melhor define o termo de entrada, é também uma limitação do estudo.

A não utilização de um conjunto de sinônimos mais específicos é mais uma limitação deste estudo. Apenas um dicionário de sinônimos ortográficos para o Pt-Br foi encontrado. Como ele é genérico e não específico da área de saúde, gera uma lista extensa de sinônimos.

5.2. TRABALHOS FUTUROS

Além da inclusão de um dicionário de sinônimos mais específicos da área da saúde, como já descrito, buscar formas de automatizar a extração de termos dos textos clínicos também foi identificado como trabalho futuro. Isto vem sendo desenvolvida no *Natural Language Processing for Portuguese Clinical Documents*, através do reconhecimento de entidades nomeadas, usando as etiquetas semânticas que foram atribuídas pelos anotadores aos termos das narrativas clínicas.

Uma outra abordagem para gerar os sinônimos dos termos é utilizar *word embeddings*, pois este trabalha com a vetorização das palavras e consegue aproximar as que representam um mesmo escopo. Para esta técnica gerar bons resultados, é necessária uma maior quantidade de narrativas clínicas.

A aplicação de métodos de PLN com *Machine Learning* pode melhorar os resultados obtidos por este estudo. O mapeamento que vem sendo produzido pelo MapClin até o momento, pode ajudar, por exemplo, a treinar uma rede neural para melhorar os resultados.

Para minimizar a quantidade de conceitos candidatos, o MapClin pode incluir o armazenamento das escolhas dos usuários, utilizando nas decisões seguintes. Facilitando a escolha dos conceitos para os próximos usuários.

O uso da CID 10 junto com a UMLS é proposto para aumentar o número de termos em português abrangidos. Apesar da CID 10 não estar presente em português na UMLS, está em inglês. A CID 10 possui uma versão em português que pode ser usada para encontrar os códigos e depois mapear para a SNOMED CT através da versão em inglês mapeada na UMLS.

6. CONCLUSÃO

Concluiu-se que é possível o mapeamento automático de termos da área da saúde em Pt-BR para a SCT utilizando a UMLS. O método proposto acelera a tarefa de mapeamento, e diminui o escopo no qual o profissional especialista precisa procurar pelo conceito que representa o termo em linguagem natural. O uso da UMLS como meio para se chegar a SCT demonstrou ser uma boa escolha, mas possui limitações devido a UMLS conter atualmente poucos termos em Pt-BR.

O método proposto apresentou bons resultados tanto nos experimentos parciais como para o experimento final, mapeando de forma exata 287 termos somando o resultado dos dois experimentos.

7. REFERÊNCIAS

- AL-HABLANI, B. The Use of Automated SNOMED CT Clinical Coding in Clinical Decision Support Systems for Preventive Care. **Perspectives in health information management**, v. 14, n. Winter, p. 1f, 2017.
- ALLONES, J. L.; MARTINEZ, D.; TABOADA, M. Automated Mapping of Clinical Terms into SNOMED-CT. An Application to Codify Procedures in Pathology. **Journal of Medical Systems**, v. 38, n. 10, p. 134, 2 out. 2014.
- ARONSON, A. R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. **Proceedings of the AMIA Symposium**, p. 17–21, 2001.
- ARONSON, A. R. Metamap: Mapping text to the umls metathesaurus. **NLM**, p. 1–26, 2006.
- BÁNFAI, B.; PORCIÓ, R.; KOVÁCS, T. Implementing reusable software components for SNOMED CT diagram and expression concept representations. **Studies in health technology and informatics**, v. 205, p. 1028–32, 2014.
- BECKER, B. F. H. et al. CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project. **Pharmacoepidemiology and Drug Safety**, v. 26, n. 8, p. 998–1005, ago. 2017.
- BOUSQUET, C. et al. Method for mapping the French CCAM terminology to the UMLS metathesaurus. **Studies in Health Technology and Informatics**, v. 180, p. 164–168, 2012.
- CHIARAMELLO, E. et al. Use of “off-the-shelf” information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. **Journal of Biomedical Informatics**, v. 63, p. 22–32, 2016.
- COHEN, W. W.; RAVIKUMAR, P.; FIENBERG, S. E. A comparison of string metrics for matching names and records. **KDD Workshop on Data Cleaning and Object Consolidation**, v. 3, p. 73–78, 2003.
- DHOMBRES, F.; BODENREIDER, O. Interoperability between phenotypes in research and healthcare terminologies—Investigating partial mappings between HPO and SNOMED CT. **Journal of Biomedical Semantics**, v. 7, n. 1, p. 3, 9 dez. 2016.
- HE, Z. et al. Sculpting the UMLS Refined Semantic Network. **Online Journal of Public Health Informatics**, v. 6, n. 2, p. e181, 16 out. 2014.
- HE, Z.; GELLER, J.; CHEN, Y. A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization. **Artificial Intelligence in Medicine**, v. 64, n. 1, p. 29–40, maio 2015.
- HUSSAIN, S. et al. A framework for evaluating and utilizing medical terminology mappings. **Studies in health technology and informatics**, v. 205, n. ii, p. 594–8,

2014.

ISMAILOV, A. et al. A comparative study of stemming algorithms for use with the Uzbek language. **2016 3rd International Conference on Computer and Information Sciences, ICCOINS 2016 - Proceedings**, p. 7–12, 2016.

JOUBERT, M. et al. Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, v. 2009, p. 291–5, 14 nov. 2009.

KEIZER, N. F.; ABU-HANNA, A.; ZWETSLOOT-SCHONK, J. H. M. Understanding terminological system I: terminology and typology. **Method Inform Med**, v. 39, p. 16–21, 2000.

KHORRAMI, F.; AHMADI, M.; SHEIKHTAHERI, A. Evaluation of SNOMED CT content coverage: A systematic literature review. **Studies in Health Technology and Informatics**, v. 248, n. 6, p. 212–219, 2018.

KIM, T. Y. Automating lexical cross-mapping of ICNP to SNOMED CT. **Informatics for Health and Social Care**, v. 41, n. 1, p. 64–77, 2 jan. 2016.

KIM, T. Y.; COENEN, A.; HARDIKER, N. Semantic mappings and locality of nursing diagnostic concepts in UMLS. **Journal of Biomedical Informatics**, v. 45, n. 1, p. 93–100, 14 fev. 2012.

LEE, D. et al. Literature review of SNOMED CT use. **Journal of the American Medical Informatics Association**, v. 21, n. e1, p. e11–e19, 1 fev. 2014.

LIU, H. et al. BioLemmatizer: A lemmatization tool for morphological processing of biomedical text. **Journal of Biomedical Semantics**, v. 3, n. 1, p. 1–29, 2012.

LOVINS, J. B. Development of a Stemming Algorithm. **Mechanical Translation and Computational Linguistics**, v. 11, n. 4, p. 22–31, 1968.

MARTINEZ SORIANO, I.; PEÑA, J. L. C. STMC: Semantic Tag Medical Concept Using Word2Vec Representation. **Proceedings - IEEE Symposium on Computer-Based Medical Systems**, v. 2018–June, p. 393–398, 2018.

MINISTÉRIO DA SAÚDE. **PORTARIA Nº 2.073, DE 31 DE AGOSTO DE 2011**. [s.l.: s.n.]. Disponível em:

<http://bvsmms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073_31_08_2011.html>.

NLM. UMLS ® Reference Manual. n. September, 2018.

OLIVEIRA, L. E. S. et al. **A statistics and UMLS-based tool for assisted semantic annotation of Brazilian clinical documents**. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). **Anais...IEEE**, nov. 2017 Disponível em: <<http://ieeexplore.ieee.org/document/8217805/>>

PINHAS, T. et al. Efficient edit distance with duplications and contractions. **Algorithms for Molecular Biology**, v. 8, n. 1, p. 27, 2013.

RAFIEI, M. et al. Systemized Nomenclature of Medicine Clinical Terms for the structured expression of perioperative medication management recommendations. **American Journal of Health-System Pharmacy**, v. 71, n. 23, p. 2020–2027, 1 dez. 2014.

RANDORFF, A.; ELBERG, P. B.; KJ, S. SNOMED CT adoption in Denmark - why is it so hard ? **EHealth-For Continuity of Care: Proceedings of MIE2014**. p. 205–226, 2014.

RENATO, A. et al. A Machine Translation Approach for Medical Terms. **Healthinf**, v. 5, n. Biostec, p. 369–378, 2018.

RODRIGUES, R.; OLIVEIRA, H. G.; GOMES, P. LemPORT: a High-Accuracy Cross-Platform Lemmatizer for Portuguese. **Maria João Varanda Pereira José Paulo Leal**, p. 267, 2014.

SAITWAL, H. et al. Cross-terminology mapping challenges: A demonstration using medication terminological systems. **Journal of Biomedical Informatics**, v. 45, n. 4, p. 613–625, ago. 2012.

SAVOVA, G. 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.

SHIVADE, C. et al. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. **Journal of Biomedical Informatics**, v. 58, n. 10, p. S103–S110, dez. 2015.

SIERRA, H. et al. Confocal Imaging–Guided Laser Ablation of Basal Cell Carcinomas: An Ex Vivo Study. **Journal of Investigative Dermatology**, v. 135, n. 2, p. 612–615, fev. 2015.

SINTCHENKO, V. et al. Towards bioinformatics assisted infectious disease control. **BMC Bioinformatics**, v. 10, n. Suppl 2, p. S10, 2009.

SNOMED CT. SNOMED CT Starter Guide. **SNOMED International**, n. July, p. 1–56, 2017.

SNOWBALL. **No Title**.

SUN, J. Y.; SUN, Y. A System for Automated Lexical Mapping. **Journal of the American Medical Informatics Association**, v. 13, n. 3, p. 334–343, 1 maio 2006.

TRAN, L.-T. T. et al. Exploiting the UMLS Metathesaurus for extracting and categorizing concepts representing signs and symptoms to anatomically related organ systems. **Journal of Biomedical Informatics**, v. 58, n. 7594, p. 19–27, dez. 2015.

WILLETT, P. The Porter stemming algorithm: then and now. **Program: electronic**

library and information systems, v. 40, n. 3, p. 219–223, jul. 2006.

APÊNDICE A – TERMOS COMPOSTOS (MAPEAMENTO EXATO)

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	hipertensão arterial sistêmica	C0020538	ENG	Hypertensive disease	7	11,49	38341003	Hypertensive disorder, systemic arterial (disorder)
2	ácido acetilsalicílico	C0004057	S3818729	acido acetilsalicílico	1	100	7947003	Aspirin (product)
		C0004057	S3818729	acido acetilsalicílico	1	100	387458008	Aspirin (substance)
3	sonda vesical de demora	C0179802	ENG	Urinary cateter	7	10,06	20568009	Urinary catheter, device (physical object)
4	membros inferiores	C0023216	S2980405	membros inferiores	1	100	61685007	Lower limb structure (body structure)
5	transplante renal	C0022671	S3859006	transplante renal	1	100	70536003	Transplant of kidney (procedure)
6	frequência cardíaca	C0018810	S3836656	frecuencia cardíaca	1	100	364075005	Heart rate (observable entity)
7	membro superior esquerdo	C0230330	ENG	Left upper extremity	7	5,18	80768000	Structure of left upper limb (body structure)
8	membro superior direito	C0230329	ENG	Right upper extremity	7	5,18	6921000	Structure of right upper limb (body structure)
9	diabetes mellitus	C0011849	S3830001	diabetes mellitus	1	100	73211009	Diabetes mellitus (disorder)
10	membro superior	C1140618	S2980402	membro superior	1	100	53120007	Upper limb structure (body structure)
11	pos operatório	C0032790	ENG	Postoperative Period	7	17,8	262061000	Postoperative period (qualifier value)
		C0241311	ENG	post operative (finding)	7	5,18	133899007	Postoperative care (regime/therapy)
12	ventilacao mecânica	C0199470	S3860927	ventilacao mecânica	1	100	40617009	Mechanical ventilation
13	ausculta pulmonar	C0004339	ENG	Auscultation	7	13,14	129436005	Auscultation - action (qualifier value)
		C0004339	ENG	Auscultation	7	13,14	37931006	Auscultation (procedure)
		C0024109	ENG	Lung	7	6,66	39607008	Lung structure (body structure)
14	pressao expiratoria positiva final	C0032740	S12848772	pressao positiva expiratoria final	5	100	250854009	Positive end expiratory pressure (observable entity)
		C0419023	S14489538	terapia pressao expiratoria positiva	5	100	45851008	Positive end expiratory pressure ventilation therapy, initiation and management (procedure)
15	repouso no leito	C0004910	ENG	Bed rest	7	8,34	225316001	Bedrest (regime/therapy)
		C0277816	ENG	Lying in bed	7	5,18	17535004	Lying in bed (finding)
16	infarto agudo do miocardio	C0027051	S16012883	infarto agudo miocárdio	1	100	22298006	Myocardial infarction (disorder)

17	membro inferior direito	C0230415	ENG	Right lower extremity	7	5,18	62175007	Structure of right lower limb (body structure)
		C0023216	S2980405	membros inferiores	5	100	61685007	Lower limb structure (body structure)
18	unidade de terapia intensiva	C0021708	S6529217	unidade terapia intensiva	1	100	309904001	Intensive care unit (environment)
19	bomba infusora	C0021436	ENG	Infusion Pump	7	14,64	57118008	Perfusion pump, device (physical object)
		C0021436	ENG	Infusion Pump	7	14,64	430033006	Infusion pump (physical object)
20	membros superiores	C1140618	S2980406	membros superiores	1	100	53120007	Upper limb structure (body structure)
21	frequencia respiratoria	C0231832	S5577026	frequencia respiratória	1	100	86290005	Respiratory rate (observable entity)
22	raio x	C0043309	S11377986	Raio	1	100	52250000	X-ray electromagnetic radiation (physical force)
23	pos operatorio imediato	C0032790	ENG	Postoperative Period	7	16,37	262061000	Postoperative period (qualifier value)
		C0241311	ENG	post operative (finding)	7	3,75	133899007	Postoperative care (regime/therapy)
24	campos pleuro pulmonares	C0225759	ENG	Lung field	7	3,74	34922002	Lung field (body structure)
25	olho esquerdo	C0229090	ENG	Left eye structure	7	5,18	8966001	Left eye structure (body structure)
26	olho direito	C0229089	ENG	Right eye	7	5,18	18944008	Right eye structure (body structure)
27	protamina neutra hagedorn	C0033603	ENG	Protamines	7	12,94	350083007	Protamine (product)
		C0033603	ENG	Protamines	7	12,94	372630008	Protamine (substance)
28	pressao arterial media	C0428886	ENG	Mean blood pressure	7	5,18	6797001	Mean blood pressure (observable entity)
29	acesso venoso central	C0444466	ENG	Central venous	7	3,75	263968003	Central venous (qualifier value)
		C0007435	S5565944	cateterizacao venosa central	4	100	303728004	Venous catheter (physical object)
		C0007435	S5565944	cateterizacao venosa central	4	100	52124006	Central venous catheter, device (physical object)
30	lucido orientado tempo espaco	C1704322	ENG	Orientation (spatial)	7	3,47	311552005	Spatial orientation, function (observable entity)
31	diurese presente	C0012797	ENG	Diuresis	7	12,96	60309003	Diuresis, function (observable entity)
		C0012797	ENG	Diuresis	7	12,96	162182005	Diuresis (finding)
32	cateter venoso central	C1145640	ENG	Central venous catheter, device	7	14,64	52124006	Central venous catheter, device (physical object)
		C0007435	S3825545	cateterismo venoso central	2	100	52124006	Central venous catheter, device (physical object)

		C0007435	S3825545	cateterismo venoso central	2	100	303728004	Venous catheter (physical object)
33	diurese efetiva	C0012797	ENG	Diuresis	7	13,14	60309003	Diuresis, function (observable entity)
		C0012797	ENG	Diuresis	7	13,14	162182005	Diuresis (finding)
34	queixa principal	C0277786	ENG	Chief complaint (finding)	7	3,68	409586006	Complaint (finding)
35	pronto socorro	C0013961	S3850687	pronto socorro	1	100	409971007	Emergency medical services (qualifier value)
36	ausculta cardiaca	C0018793	ENG	Heart Auscultation	7	17,8	449263002	Auscultation of heart (procedure)
37	insuficiencia renal cronica	C0022661	S5582103	insuficiencia renal crônica	1	100	90688005	Chronic renal failure syndrome (disorder)
38	murmurio vesicular positivo	C0231857	ENG	Vesicular breathing	7	3,75	77047003	Vesicular breathing (finding)
		C1446409	ENG	Positive	7	3,48	10828004	Positive (qualifier value)
39	sonda nasogastrica	C0085678	ENG	Nasogastric tube	7	5,18	17102003	Nasogastric tube, device (physical object)
40	exame físico	C0031809	S3834521	exame físico	1	100	5880005	Physical examination procedure (procedure)
41	ventilacao controlada a pressão	C0564626	ENG	Pressure controlled ventilation	7	5,18	286812008	Pressure controlled ventilation (procedure)
42	doenca arterial coronariana	C1956346	S3831061	doenca arteria coronariana	1	96,15385	414024009	Disorder of coronary artery (disorder)
43	aparelho cardiovascular	C0007226	ENG	Cardiovascular system	7	5,18	113257007	Structure of cardiovascular system (body structure)
		C1269562	ENG	Entire cardiovascular system	7	5,18	278198007	Entire cardiovascular system (body structure)
44	infeccao trato urinario	C0042029	S5580827	infeccao tracto urinário	1	95,83333	68566005	Urinary tract infectious disease (disorder)
		C0042075	S3831611	doencas trato urinário	5	100	128606002	Disorder of the urinary system (disorder)
		C0042075	S3831611	doencas trato urinário	5	100	41368006	Disorder of urinary tract proper (disorder)
45	tala gessada	C0032159	ENG	Plaster Casts	7	17,8	34164001	Plaster cast, device (physical object)
46	traumatismos multiplos	C0026771	S3859447	traumatismo múltiplo	1	90	12835000	Multiple injuries (morphologic abnormality)
		C0026771	S3859447	traumatismo múltiplo	1	90	57028002	Multiple wounds (morphologic abnormality)
		C0026771	S3859447	traumatismo múltiplo	1	90	262519004	Multiple injuries (disorder)
47	oximetria de pulso	C0034108	S3847804	oximetria pulso	1	100	708065004	Pulse oximetry technique (qualifier value)
		C0034108	S3847804	oximetria pulso	1	100	252465000	Pulse oximetry (procedure)
48	pressao intraocular	C0021888	S13703070	pressao intraocular	1	100	41633001	Intraocular pressure (observable entity)

49	descolamento da retina	C0035305	S1488120	descolamento retina	1	100	42059000	Retinal detachment (disorder)
50	dispneia paroxistica noturna	C1956415	S5570444	dispneia paroxistica nocturna	1	96,55172	55442000	Paroxysmal nocturnal dyspnea (finding)
51	fibrilacao atrial	C0004238	S3835749	fibrilacao atrial	1	100	49436004	Atrial fibrillation (disorder)
52	idade gestacional	C0017504	S3839751	idade gestacional	1	100	57036006	Fetal gestational age (observable entity)
53	ulcera por pressao	C0011127	S10691650	ulcera por pressão	1	100	399912005	Pressure ulcer (disorder)
		C0011127	S10691650	ulcera por pressão	1	100	420226006	Pressure ulcer (morphologic abnormality)
54	cuidados de enfermagem	C0028682	S3828864	cuidados enfermagem	1	100	9632001	Nursing procedure (procedure)
55	indice de massa corporal	C0005893	S3840253	indice massa corporal	1	100	60621009	Body mass index (observable entity)
56	transplante cardiaco	C0018823	S3858962	transplante cardíaco	1	100	32413006	Transplantation of heart (procedure)
57	historia morbida familiar	C1705495	ENG	Family History Domain	7	5,18	416471007	Family history of clinical finding (situation)
		C1705495	ENG	Family History Domain	7	5,18	57177007	Family history with explicit context (situation)
58	arteria coronaria direita	C1261316	ENG	Right coronary artery structure	7	5,18	13647002	Right coronary artery structure (body structure)
59	hiperplasia prostatica benigna	C1704272	S3839162	hiperplasia prostatica benigna	1	100	266569009	Benign prostatic hyperplasia (disorder)

APÊNDICE B – TERMOS COMPOSTOS (MAPEAMENTO PARCIAL)

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	bom estado geral	C0348080	ENG	Condition	7	3,63	260905004	Condition (attribute)
		C0205170	ENG	Good	7	3,48	20572008	Good (qualifier value)
2	acesso venoso periferico	C0031127	S3825546	cateterismo venoso periferico	4	100	385757000	Venous catheter care (regime/therapy)
3	sonda nasoenteral	C0085590	ENG	catheter device	7	6,83	19923001	Catheter, device (physical object)
4	bulhas cardiacas normofoneticas	C0232187	ENG	Cardiac rhythm type	7	5,18	251149006	Cardiac rhythm type (observable entity)
5	membro inferior esquerdo	C0015385	ENG	Limb structure	7	6,83	66019005	Limb structure (body structure)
		C0023216	S2980405	membros inferiores	5	100	61685007	Lower limb structure (body structure)
6	murmurios vesiculares presentes	C0231857	ENG	Vesicular breathing	7	3,98	77047003	Vesicular breathing (finding)
7	angioplastia coronariana transluminal percutanea	C0162577	S3820996	angioplastia percutanea transluminal	5	100	418285008	Angioplasty of blood vessel (procedure)
		C2936666	S5560811	angioplastia percutanea transluminal	5	100	5431005	Percutaneous transluminal angioplasty (procedure)
		C2936667	S3820999	angioplastia transluminal	5	100	446878003	Transluminal angioplasty (procedure)
8	tubo orotraqueal	C0175730	ENG	biomedical tube device	7	3,68	83059008	Tube, device (physical object)
9	sonda nasogastrica aberta	C0085678	ENG	Nasogastric tube	7	5,18	17102003	Nasogastric tube, device (physical object)
		C0812428	ENG	Nasogastric tube procedures	7	5,18	17102003	Nasogastric tube, device (physical object)
10	pedido de consulta	C0184741	ENG	Patient referral for consultation	7	5,18	44383000	Patient referral for consultation (procedure)
11	diurese espontanea	C0012797	ENG	Diuresis	7	13,14	60309003	Diuresis, function (observable entity)
		C0012797	ENG	Diuresis	7	13,14	162182005	Diuresis (finding)
		C0205359	ENG	Spontaneous	7	3,5	5054005	Spontaneous (qualifier value)
12	pressao nao invasiva	C0033095	ENG	Pressure- physical agente	7	9,99	279046003	Pressure - physical agent (physical force)
		C0460139	ENG	Pressure (finding)	7	3,68	13543005	Pressure (finding)
		C0205303	ENG	Non-invasive	7	3,5	22762002	Non-invasive (qualifier value)

13	queixas álgicas	C0277786	ENG	Chief complaint (finding)	7	3,68	409586006	Complaint (finding)
		C0700624	ENG	Allergic	7	3,5	277054007	Allergen (attribute)
14	insuficiencia cardiaca congestiva	C0018801	S3840930	insuficiencia cardiaca congestiva	1	100	84114007	Heart failure (disorder)
15	doador compativel	C0040288	ENG	Tissue Donors	7	6,83	105468003	Tissue donor (person)
		C0013018	ENG	Donor person	7	3,68	105455006	Donor for medical or surgical procedure (person)
		C1524057	ENG	Compatible	7	3,5	7883008	Compatible with (attribute)
16	curativo oclusivo	C0028791	ENG	Occlusive Dressings	7	11,49	63995005	Bandage, device (physical object)
17	unidade basica de saude	C0033137	S16006379	atencao basica saúde	5	100	5351000124100	Primary care clinic (environment)
18	acesso periferico	C0444454	ENG	Access	7	3,68	260507000	Access (attribute)
		C0205100	ENG	Peripheral	7	3,5	14414005	Peripheral (qualifier value)
19	roncos difusos	C0037384	ENG	Snoring	7	17,8	72863001	Snoring (finding)
		C0037384	ENG	Snoring	7	17,8	162375000	Snoring symptoms (finding)
		C0205219	ENG	Diffuse	7	5,18	19648000	Diffuse (qualifier value)
20	arteria descendente anterior	C0003842	ENG	Arteries	7	9,94	51114001	Arterial structure (body structure)
21	ventilacao mecanica controlada	C0419012	ENG	Controlled mandatory ventilation	7	5,18	243148004	Controlled mandatory ventilation (procedure)
		C0199470	S3860927	ventilacao mecânica	5	100	40617009	Mechanical ventilation
22	pos operatorio tardio	C0032790	ENG	Postoperative Period	7	16,15	262061000	Postoperative period (qualifier value)
		C0205087	ENG	Late	7	3,63	260383002	Late (qualifier value)
		C0241311	ENG	post operative (finding)	7	3,53	133899007	Postoperative care (regime/therapy)
23	pos transplante renal	C0022671	ENG	Kidney Transplantation	7	13,22	70536003	Transplant of kidney (procedure)
		C0687676	ENG	Post	7	3,48	288563008	After values (qualifier value)
24	media quantidade	C1265611	ENG	Quantity	7	3,68	246205007	Quantity (attribute)
25	sonda vesical de demora aberta	C0179802	ENG	Urinary cateter	7	10,32	20568009	Urinary catheter, device (physical object)
26	condicoes e habitos de vida	C0018464	ENG	Behaviorial Habits	7	9,9	363898005	Habits (observable entity)
		C1705253	ENG	Logical Condition	7	3,98	260905004	Condition (attribute)

		C0037403	S7114213	condicoes vida	5	100	82996008	Social condition
27	pressao venosa central	C0199666	S5594797	pressao venosa central	1	100	54654001	Measurement of central venous pressure (procedure)
28	historia morbida atual	C0019665	ENG	Historical aspects qualifier	7	9,94	51042001	History of present illness
		C0019665	ENG	Historical aspects qualifier	7	9,94	392521001	History of (contextual qualifier) (qualifier value)
		C0262512	ENG	History of present illness	7	3,63	422625006	History of present illness section (record artifact)

APÊNDICE C – TERMOS COMPOSTOS (NÃO MAPEADOS)

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	pronto atendimento							
2	ruidos hidroaereos positivos							
3	ruidos hidroaereos presentes							
4	ruidos adventícios							
5	centro cirúrgico	C0038942	S16007689	centro cirurgico	1	100		
6	ambos os olhos							
7	campos pleuropulmonares livres							
8	historia morbida pregressa							
9	sem sopro							
10	pressao inspiratória							
11	nevoa úmida							
12	descendente anterior							
13	exame geral							

APÊNDICE D – TERMOS COMPOSTOS (NÃO MAPEADOS) MAPEAMENTO MANUAL

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	pronto atendimento						409971007	Emergency medical services (qualifier value)
							274409007	Special care unit (environment)
2	ruidos hidroaereos positivos							não encontrei
3	ruidos hidroaereos presentes							não encontrei
4	ruidos adventicios						53972003	Added respiratory sounds (finding)
5	centro cirúrgico	C0038942	S16007689	centro cirurgico	1	100	405607001	Ambulatory surgery center (environment)
6	ambos os olhos						40638003	Structure of both eyes (body structure)
7	campos pleuropulmonares livres							não encontrei
8	historia morbida pregressa							não encontrei
9	sem sopro						248551002	Cannot blow (finding)
							78064003	Respiratory function (observable entity)

10	pressao inspiratoria						251907000	Respiratory pressure (observable entity)
11	nevoa úmida							não encontrei
12	descendente anterior							não encontrei
13	exame geral						162673000	General examination of patient (procedure)

APÊNDICE E – TERMOS SIMPLES (MAPEAMENTO EXATO)

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	Cesáreas	C0007876	S3826080	Cesárea	2	100	11466000	Cesarean section (procedure)
2	Paciente	C0030705	S16016972	Paciente	1	100	116154003	Patient (person)
3	Edema	C0013604	S3832029	Edema	1	100	79654002	Edema (morphologic abnormality)
		C0013604	S3832029	Edema	1	100	267038008	Edema (finding)
		C0013604	S3832029	Edema	1	100	423666004	Edema (observable entity)
4	Tacrolimo	C0085149	S6155532	Tacrolimo	1	100	109129008	Tacrolimus (product)
		C0085149	S6155532	Tacrolimo	1	100	386975001	Tacrolimus (substance)
5	eletrocardiograma	C1623258	S2802807	Electrocardiograma	1	94,44444444	29303009	Electrocardiographic procedure (procedure)
		C0013798	ENG	Electrocardiogram	7	17,8	164845003	Electrocardiogram - general (procedure)
		C0013798	ENG	Electrocardiogram	7	17,8	164860000	Electrocardiogram - general - NOS (procedure)
		C0013798	ENG	Electrocardiogram	7	17,8	142008000	ECG – general
		C0013798	ENG	Electrocardiogram	7	17,8	271334005	ECG – general
		C0013798	ENG	Electrocardiogram	7	17,8	142020008	ECG - general – NOS
		C0013798	ENG	Electrocardiogram	7	17,8	29303009	Electrocardiographic procedure (procedure)
		C0013799	S5572237	electrocardiograma ambulatório	6	90	164850009	Ambulatory electrocardiogram (procedure)
6	Retorno	C0332156	ENG	Return to	7	5,18	7528007	Return to (contextual qualifier) (qualifier value)
7	Curativo	C0175723	ENG	Bands	7	3,68	77541009	Band, device (physical object)
		C0013119	S3828944	Curativos	2	100	333453004	Medical dressing (physical object)
		C0028791	S3828942	curativos oclusivos	6	100	63995005	Bandage, device (physical object)
		C0028791	S3828942	curativos oclusivos	6	100	350785008	Gauzes (physical object)
		C0677875	S14234182	curativos compressivos	6	100	417660008	Esmarch bandage (physical object)
		C0677875	S14234182	curativos compressivos	6	100	701721005	Compression bandaging kit (physical object)
		C0178388	S5561934	aplicacao penso ferida	4	100	182531007	Dressing of wound (procedure)
8	Queixas	C0277786	ENG	Chief complaint (finding)	7	5,18	409586006	Complaint (finding)
9	Consciente	C0234421	ENG	Conscious	7	14,64	106167005	Consciousness

		C0038535	S3856893	Subconsciente	2	100	89970005	Psyche structure
10	Carvedilol	C0054836	ENG	Carvedilol	7	14,64	108551001	Carvedilol (product)
		C0054836	ENG	Carvedilol	7	14,64	386870007	Carvedilol (substance)
11	Abdome	C0000726	S3818313	Abdome	1	100	113345001	Abdominal structure (body structure)
		C0000726	S3818313	Abdome	1	100	277112006	Abdominal (qualifier value)
12	Afebril	C0277797	ENG	Apyrexial	7	5,18	86699002	Apyrexial (situation)
13	Alta	C0205250	ENG	High	7	5,18	75540009	High (qualifier value)
		C0015967	S5600754	temperatura alta	6	100	386661006	Fever (finding)
		C0030685	S3820333	alta paciente	6	100	58000006	Patient discharge (procedure)
14	Cateterismo	C0007430	S3825541	Cateterismo	1	100	45211000	Catheterization (procedure)
15	Enalapril	C0014025	S3832408	Enalapril	1	100	15222008	Enalapril (product)
		C0014025	S3832408	Enalapril	1	100	372658000	Enalapril (substance)
16	Hidroclorotiazida	C0020261	S3838938	Hidroclorotiazida	1	100	91667005	Hydrochlorothiazide (product)
		C0020261	S3838938	Hidroclorotiazida	1	100	387525002	Hydrochlorothiazide (substance)
17	Sinvastatina	C0074554	S3856349	Sinvastatina	1	100	96304005	Simvastatin (product)
		C0074554	S3856349	Sinvastatina	1	100	387584000	Simvastatin (substance)
18	Enfermeira	C0028661	S16010463	Enfermeira	1	100	106293008	Nursing personnel (occupation)
		C0028661	S16010463	Enfermeira	1	100	106292003	Professional nurse (occupation)
19	Lantus	C0876064	S16013995	Lantus	1	100	126212009	Insulin glargine product (product)
		C0876064	S16013995	Lantus	1	100	411529005	Insulin glargine (substance)
20	Sirolimo	C0072980	S6155494	Sirolimo	1	100	116109004	Sirolimus (product)
		C0072980	S6155494	Sirolimo	1	100	387014003	Sirolimus (substance)
21	Doutor	C0031831	ENG	Doctor – Title	7	5,18	112247003	Medical doctor (occupation)
		C2348314	ENG	Physicians	7	14,64	309343006	Physician (occupation)
22	Prednisona	C0032952	S3850174	Prednisona	1	100	10312003	Prednisone preparation (product)
		C0032952	S3850174	Prednisona	1	100	116602009	Prednisone (substance)
23	Seco	C0205222	ENG	Dry	7	5,18	13880007	Dry (qualifier value)
24	Hemodiálise	C0019004	S3838483	Hemodiálise	1	100	302497006	Hemodialysis (procedure)
25	Omeprazol	C0028978	S3847246	Omeprazol	1	100	25673006	Omeprazole (product)

		C0028978	S3847246	Omeprazol	1	100	387137007	Omeprazole (substance)
26	Abdomen	C0000726	S12847494	Abdômen	1	100	113345001	Abdominal structure (body structure)
		C0000726	S12847494	Abdômen	1	100	277112006	Abdominal (qualifier value)
27	Pupilas	C0034121	ENG	Pupil	7	17,8	392406005	Pupil structure (body structure)
28	Atenolol	C0004147	S3822698	Atenolol	1	100	87652004	Atenolol (product)
		C0004147	S3822698	Atenolol	1	100	387506000	Atenolol (substance)
29	Ecocardiograma	C0013516	S2802646	Ecocardiograma	1	100	40701008	Echocardiography (procedure)
30	Dor	C0030193	S3831792	Dor	1	100	22253000	Pain (finding)
31	Losartana	C0126174	ENG	Losartan	7	24,11	96309000	Losartan (product)
		C0126174	ENG	Losartan	7	24,11	373567002	Losartan (substance)
32	Cirurgia	C0543467	S3826556	Cirurgia	1	100	387713003	Surgical procedure (procedure)
		C0543467	S3826556	Cirurgia	1	100	257556004	Surgery (qualifier value)
		C0543467	S3826556	Cirurgia	1	100	83578000	Surgical (qualifier value)
33	Laboratorial	C0011911	S3830032	diagnostico laboratorial	6	100	46159000	Laboratory diagnosis (contextual qualifier) (qualifier value)
		C0022885	S5561784	analise laboratorial	6	100	108252007	Laboratory procedure (procedure)
		C0022885	S5561784	analise laboratorial	6	100	15220000	Laboratory test (procedure)
		C0022885	S5561784	analise laboratorial	6	100	269814003	Laboratory procedures -general (situation)
		C0151749	S5561783	analise laboratorial interferência	6	100	108252007	Laboratory procedure (procedure)
		C0151749	S5561783	analise laboratorial interferência	6	100	15220000	Laboratory test (procedure)
		C0151749	S5561783	analise laboratorial interferência	6	100	269814003	Laboratory procedures -general (situation)
34	Oriento	C0029266	ENG	psychological orientation	7	11,49	43173001	Orientation, function (observable entity)
		C0025369	S3847449	Orientadores	2	100	171002009	Vocational counseling (procedure)
35	Orientado	C1704322	ENG	Orientation (spatial)	7	5,18	311552005	Spatial orientation, function (observable entity)
		C0029266	S3847446	Orientação	2	100	43173001	Orientation, function (observable entity)
36	Metformina	C0025598	S3844211	Metformina	1	100	109081006	Metformin (product)
		C0025598	S3844211	Metformina	1	100	372567009	Metformin (substance)
37	traqueostomizado	C0040590	ENG	Tracheostomy procedure	7	11,49	48387007	Tracheostomy

		C1700189	ENG	Tracheostomy Route of Drug Administration	7	5,18	265841002	Tracheostomy
		C1700189	ENG	Tracheostomy Route of Drug Administration	7	5,18	280361007	Tracheostomy stoma (morphologic abnormality)
		C1700189	ENG	Tracheostomy Route of Drug Administration	7	5,18	316076007	[V]Has tracheostomy
38	Procedimento	C1948041	ENG	Surgical and medical procedures	7	5,18	71388002	Procedure (procedure)
		C0087111	S14708278	procedimentos curativos	6	100	277132007	Therapeutic procedure (procedure)
		C0087111	S14708278	procedimentos curativos	6	100	276239002	Therapy (regime/therapy)
		C0543467	S3850405	procedimentos cirurgicos operatorios	6	100	387713003	Surgical procedure (procedure)
		C0199171	S16290409	procedimento medico	6	100	50731006	Medical procedure (procedure)
		C0204193	S5594915	procedimento ortodôntico	6	100	16177004	Orthodontic procedure (procedure)
		C0430022	S5594890	procedimento diagnostico	6	100	103693007	Diagnostic procedure (procedure)
39	Dia	C0332173	ENG	Daily	7	5,18	69620002	Daily (qualifier value)
		C0439228	ENG	Day	7	5,18	258703001	day (qualifier value)
		C0439505	ENG	per day	7	5,18	259032004	per day (qualifier value)
40	Medicações	C2598133	ENG	Medications:-:Point in time:^Patient:-	7	5,18	373873005	Pharmaceutical / biologic product (product)
		C2598133	ENG	Medications:-:Point in time:^Patient:-	7	5,18	105903003	General drug type (product)
		C2598133	ENG	Medications:-:Point in time:^Patient:-	7	5,18	410942007	Drug or medicament (substance)
41	Anticoagulante	C0003280	S16743443	Anticoagulante	1	100	81839001	Anticoagulant (product)
		C0003280	S16743443	Anticoagulante	1	100	372862008	Anticoagulant (substance)
42	Levotiroxina	C0040165	S3842316	Levotiroxina	1	100	38076006	Tetraiodothyronine preparation (product)
		C0040165	S3842316	Levotiroxina	1	100	73187006	Thyroxine (substance)
43	Dislipidemia	C0242339	S2802165	Dislipidemia	1	100	370992007	Dyslipidemia (disorder)
44	Evacuação	C1282573	ENG	Evacuation procedure	7	5,18	129292007	Evacuation - action (qualifier value)
		C1282573	ENG	Evacuation procedure	7	5,18	122461007	Evacuation procedure (procedure)

45	Edemas	C0013604	ENG	Edema	7	11,49	79654002	Edema (morphologic abnormality)
		C0013604	ENG	Edema	7	11,49	267038008	Edema (finding)
		C0013604	ENG	Edema	7	11,49	423666004	Edema (observable entity)
46	Comunicativa	C0009452	ENG	Communication	7	11,49	263536004	Communication (attribute)
47	acompanhamento	C0001758	S16744029	assistencia seguimento	4	100	413467001	Aftercare (regime/therapy)
48	Exame	C1261322	S5574638	Exame	1	100	129265001	Evaluation - action (qualifier value)
		C1261322	S5574638	Exame	1	100	386053000	Evaluation procedure (procedure)
49	Exames	C0582103	ENG	Medical Examination	7	3,98	225886003	Medical assessment (procedure)
		C1261322	S5574638	Exame	3	100	129265001	Evaluation - action (qualifier value)
		C1261322	S5574638	Exame	3	100	386053000	Evaluation procedure (procedure)
		C0031809	S3834521	exame físico	6	100	302199004	Examination - action (qualifier value)
50	Palpação	C0030247	S3847961	Palpação	1	100	129434008	Palpation - action (qualifier value)
		C0030247	S3847961	Palpação	1	100	113011001	Palpation (procedure)
51	Solicito	C1272683	ENG	Request – action	7	3,68	385644000	Requested (qualifier value)
		C0686900	ENG	Request for	7	3,68	103320006	Request for (contextual qualifier) (qualifier value)
52	Externamente	C0205101	ENG	Extrinsic	7	5,18	261074009	External (qualifier value)
53	Consulta	C0009818	S3828088	Consulta	1	100	223475005	Consulting with (procedure)
		C0009818	S3828088	Consulta	1	100	11429006	Consultation (procedure)
54	Insulina	C0021641	S3840957	Insulina	1	100	39487003	Insulin product (product)
		C0021641	S3840957	Insulina	1	100	67866001	Insulin (substance)
		C0021641	S3840957	Insulina	1	100	412222002	Regular insulin (substance)
55	Jejum	C0015663	S3841495	Jejum	1	100	16985007	Fasting (finding)
56	Stent	C0038257	ENG	Stent, device	7	11,49	65818007	Stent, device (physical object)
57	Orientada	C1704322	ENG	Orientation (spatial)	7	5,18	311552005	Spatial orientation, function (observable entity)
		C0029266	S3847446	Orientação	2	100	43173001	Orientation, function (observable entity)
58	Dieta	C0012155	S3830205	Dieta	1	100	41829006	Dietary finding (finding)
59	Anlodipino	C0051696	S16005778	Anlodipino	1	100	108537001	Amlodipine (product)

		C0051696	S16005778	Anlodipino	1	100	386864001	Amlodipine (substance)
60	Biomicroscopia	C0419360	S16745414	Biomicroscopia	1	100	55468007	Ocular slit lamp examination (procedure)
61	Furosemina	C0016860	S3836793	Furosemina	1	100	81609008	Furosemide (product)
		C0016860	S3836793	Furosemina	1	100	387475002	Furosemide (substance)
62	Noradrenalina	C0028351	S3846622	Noradrenalina	1	100	111130009	Norepinephrine preparation (product)
		C0028351	S3846622	Noradrenalina	1	100	45555007	Norepinephrine (substance)
63	Dispneia	C0013404	S2970889	Dispneia	1	100	230145002	Difficulty breathing (finding)
		C0013404	S2970889	Dispneia	1	100	267036007	Dyspnea (finding)
64	Ultrassonografia	C0041618	S6912115	Ultrassonografia	1	100	278292003	Ultrasound imaging - action (qualifier value)
		C0041618	S6912115	Ultrassonografia	1	100	16310003	Diagnostic ultrasonography (procedure)
		C0041618	S6912115	Ultrassonografia	1	100	359659005	Echography (procedure)
65	Clopidogrel	C0070166	ENG	Clopidogrel	7	14,64	108979001	Clopidogrel (product)
		C0070166	ENG	Clopidogrel	7	14,64	386952008	Clopidogrel (substance)
66	Alerta	C0239110	ENG	Consciousness clear	7	5,18	248221007	Consciousness clear (finding)
		C0003808	S3846554	nivel alerta	6	100	312012004	Cognitive function: awareness (observable entity)
		C0003808	S3846554	nivel alerta	6	100	27625002	Wakefulness (observable entity)
67	Fisioterapia	C0031818	S2499893	Fisioterapia	1	100	91251008	Physical therapy procedure (regime/therapy)
68	Avaliação	C0220825	ENG	Evaluation	7	14,64	129265001	Evaluation - action (qualifier value)
		C0220825	ENG	Evaluation	7	14,64	386053000	Evaluation procedure (procedure)
		C0036591	S13701246	autoavaliacao	2	100	225885004	Health assessment (procedure)
		C0013656	S3823028	avaliacao educacional	6	100	20135006	Screening procedure (procedure)
		C0015196	S3834257	estudos avaliação	6	100	360155005	Assessment - action
69	Lucido	C0556994	ENG	Lucid	7	5,18	285241002	Lucid (finding)
70	Espironolactona	C0037982	S3833790	Espironolactona	1	100	13929005	Spiroinolactone (product)
		C0037982	S3833790	Espironolactona	1	100	387078006	Spiroinolactone (substance)
71	Interna	C0205102	ENG	Internal	7	5,18	260521003	Internal (qualifier value)
72	Indolor	C0234226	ENG	Painless	7	5,18	255350008	Painless (qualifier value)
73	Fentanil	C0015846	ENG	Fentanyl	7	14,64	40648001	Fentanyl (product)

		C0015846	ENG	Fentanyl	7	14,64	373492002	Fentanyl (substance)
74	Infusão	C0574032	S5581840	Infusão	1	100	129330003	Infusion - action (qualifier value)
		C0574032	S5581840	Infusão	1	100	447826007	Infusion technique (qualifier value)
75	Hospital	C0019994	S14707350	Hospital	1	100	285201006	Hospital environment (environment)
76	Simétrico	C0332516	ENG	Symmetrical	7	5,18	255473004	Symmetrical (qualifier value)
		C0332516	ENG	Symmetrical	7	5,18	18772005	Symmetry (qualifier value)
77	Ticlopidina	C0040207	S3858351	Ticlopidina	1	100	108971003	Ticlopidine (product)
		C0040207	S3858351	Ticlopidina	1	100	386950000	Ticlopidine (substance)
78	Emergência	C2745965	S16010299	Emergência	1	100	182813001	Emergency treatment (procedure)
		C2745965	S16010299	Emergência	1	100	225728007	Accident and Emergency department (environment)
79	Bactrim	C0591139	ENG	Bactrim	7	24,11	703745000	Sulfamethoxazole + trimethoprim (substance)
		C0591139	ENG	Bactrim	7	24,11	398731002	Sulfamethoxazole + trimethoprim (product)
80	Diurese	C0012797	S3830864	Diurese	1	100	60309003	Diuresis, function (observable entity)
		C0012797	S3830864	Diurese	1	100	162182005	Diuresis (finding)
81	Tratamento	C0087111	S10691443	Tratamento	1	100	277132007	Therapeutic procedure (procedure)
		C0087111	S10691443	Tratamento	1	100	276239002	Therapy (regime/therapy)
82	Plano	C1301732	ENG	Planned	7	5,18	397943006	Planned (qualifier value)
		C0030678	S3849411	plano tratamento	6	100	413467001	Aftercare (regime/therapy)
		C0030678	S3849411	plano tratamento	6	100	386367000	Mutual goal setting (regime/therapy)
		C0030678	S3849411	plano tratamento	6	100	316254009	[V]Specified procedures and aftercare

APÊNDICE F – TERMOS SIMPLES (MAPEAMENTO PARCIAL)

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	Conduta	C0004927	S3827943	conduta	1	100	844005	Behavior finding (finding)
2	Peso	C1305866	S2810695	peso	1	100	39857003	Weighing patient (procedure)
3	Limpo	C1947930	ENG	Cleaning (activity)	7	5,18	228403004	Cleans
4	Intercorrências	C1171258	ENG	Complication Aspects	7	11,49	116223007	Complication (disorder)
5	Comunicativo	C0009452	ENG	Communication	7	11,49	263536004	Communication (attribute)
6	Corado	C0009393	S3828362	cor	3	100	703247007	Color (observable entity)
		C0009393	S3828362	cor	3	100	263714004	Colors (qualifier value)
7	Corada	C0009393	S3828362	cor	3	100	703247007	Color (observable entity)
8	Cliente	C0008942	ENG	Clients	7	8,34	406193000	Client satisfaction (observable entity)

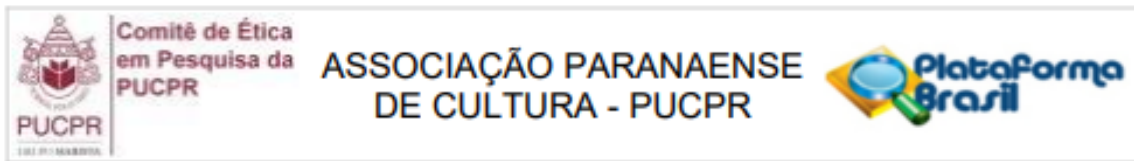
APÊNDICE G – TERMOS SIMPLES (NÃO MAPEADOS)

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	Meticorten							
2	Salinizado							
3	Cuidados	C0150499	S16747180	Cuidados	1	100		
4	Hidratada							
5	Após							
6	Globoso							
7	Eupneico							
8	Pos							
9	Hidratado							
10	Normal	C0853283	S13958999	Normal	1	100		

APÊNDICE H – TERMOS SIMPLES (NÃO MAPEADOS) MAPEAMENTO MANUAL

	Termos	CUI	SUI	STR_UMLS	Regra	num_lev	CODE	Termo_Snomed
1	Meticorten						116602009	Prednisone (substance)
2	Salinizado							não encontrei
3	Cuidados	C0150499	S16747180	Cuidados	1	100	161056001	In care (finding)
4	hidratada						405006006	Hydration status (observable entity)
5	apos						255234002	After (attribute)
6	globoso							não encontrei
7	eupneico						22803001	Normal respiratory function (finding)
8	Pos						288563008	After values (qualifier value)
9	hidratado						405006006	Hydration status (observable entity)
10	Normal	C0853283	S13958999	normal	1	100	17621005	Normal (qualifier value)

ANEXO A - PARECER CONSUBSTANCIADO COMITÊ DE ÉTICA EM PESQUISA



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: IRDischarge - PNL para Identificação de Informações em Narrativas Clínicas

Pesquisador: Claudia Maria Cabral Moro Barra

Área Temática:

Versão: 1

CAAE: 51376015.4.0000.0020

Instituição Proponente: Pontifícia Universidade Católica do Parana - PUCPR

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.354.675

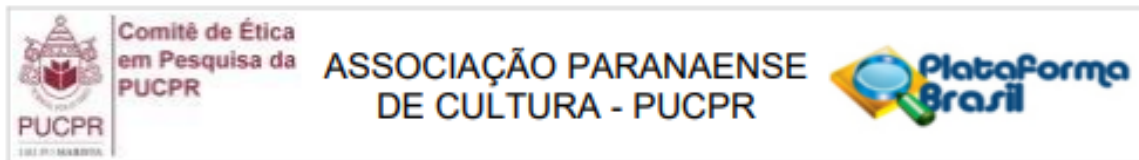
Apresentação do Projeto:

O IRDischarge é um sistema para apoio à identificação de informações em narrativas clínicas, desenvolvido pelo grupo de pesquisa de Recuperação de Informações em Saúde do PPGTS/PUCPR. Ele é baseado em algoritmos de processamento de linguagem natural (PLN) elaborados para: extração de conceitos clínicos, identificação da presença de conteúdo específicos, determinação de negações e desambiguações de abreviaturas utilizadas, incluindo também análises dos aspectos de abstrações de tempo. O IRDischarge pode ser acoplado à prontuários eletrônicos de saúde. A avaliação de sistemas de informação em saúde é necessária para garantir o sucesso da implantação dos mesmos. Sendo assim, o objetivo principal deste trabalho é a avaliar algoritmos de processamento de linguagem natural para identificação de informações em narrativas clínicas. Durante a aplicação das técnicas de PLN geralmente é necessário a utilização de um corpus (coleção de termos adicionada a definição morfossintática respectiva) anotado (corrigido por especialistas). São raros os corpora para textos com português, especialmente focados na área de saúde. Atualmente, o grupo de pesquisa deste projeto utiliza um corpus anotado específico da área de saúde, elaborado em português, construído em 2010. Porém, este corpus precisa ser complementado, o que também será realizado durante este projeto.

Objetivo da Pesquisa:

Objetivo Primário: Avaliar algoritmos de processamento de linguagem natural para identificação de

Endereço: Rua Imaculada Conceição 1155
Bairro: Prado Velho **CEP:** 80.215-901
UF: PR **Município:** CURITIBA
Telefone: (41)3271-2103 **Fax:** (41)3271-2103 **E-mail:** nep@pucpr.br



Continuação do Parecer: 1.354.675

informações em narrativas clínicas. Objetivo Secundário: avaliar algoritmos de PLN para identificação de: negações, desambiguação de abreviaturas, abstração temporal, presença de continuidade clínica, e identificação de conceitos clínicos. Analisar os diferentes métodos existentes para avaliação de SIS, considerando a recuperação de informações; e atualizar o corpus clínico anotado.

Avaliação dos Riscos e Benefícios:

Os riscos e benefícios apresentados estão adequados e de acordo com a resolução 466/2012.

Comentários e Considerações sobre a Pesquisa:

A metodologia e objetivos apresentados estão adequados e em acordo com a resolução 466/2012.

Considerações sobre os Termos de apresentação obrigatória:

Os termos apresentados estão adequados e em acordo com a resolução 466/2012.

Recomendações:

Ver Conclusões ou Pendências e Lista de Inadequações.

Conclusões ou Pendências e Lista de Inadequações:

Projeto aprovado.

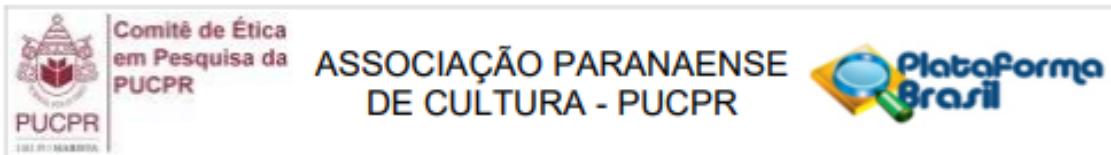
Considerações Finais a critério do CEP:

Lembramos aos senhores pesquisadores que, no cumprimento da Resolução 466/2012, o Comitê de Ética em Pesquisa (CEP) deverá receber relatórios anuais sobre o andamento do estudo, bem como a qualquer tempo e a critério do pesquisador nos casos de relevância, além do envio dos relatos de eventos adversos, para conhecimento deste Comitê. Salientamos ainda, a necessidade de relatório completo ao final do estudo. Eventuais modificações ou emendas ao protocolo devem ser apresentadas ao CEP PUCPR de forma clara e sucinta, identificando a parte do protocolo a ser modificado e as suas justificativas. Se a pesquisa, ou parte dela for realizada em outras instituições, cabe ao pesquisador não iniciá-la antes de receber a autorização formal para a sua realização. O documento que autoriza o início da pesquisa deve ser carimbado e assinado pelo responsável da instituição e deve ser mantido em poder do pesquisador responsável, podendo ser requerido por este CEP em qualquer tempo.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações	PB_INFORMAÇÕES_BÁSICAS_DO_P	27/11/2015		Aceito

Endereço: Rua Imaculada Conceição 1155
Bairro: Prado Velho **CEP:** 80.215-901
UF: PR **Município:** CURITIBA
Telefone: (41)3271-2103 **Fax:** (41)3271-2103 **E-mail:** nep@pucpr.br



Continuação do Parecer: 1.354.675

Básicas do Projeto	ETO_600951.pdf	13:52:55		Aceito
Folha de Rosto	IRDischargeFolhaRostoassinada.pdf	27/11/2015 13:51:23	Claudia Maria Cabral Moro Barra	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_QuestionariosAval.pdf	24/11/2015 23:38:26	Claudia Maria Cabral Moro Barra	Aceito
Outros	IRDischargeTCUD.pdf	24/11/2015 23:36:05	Claudia Maria Cabral Moro Barra	Aceito
Projeto Detalhado / Brochura Investigador	CEP_IRDischarge.pdf	24/11/2015 23:22:08	Claudia Maria Cabral Moro Barra	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

CURITIBA, 07 de Dezembro de 2015

Assinado por:
NAIM AKEL FILHO
(Coordenador)

Endereço: Rua Imaculada Conceição 1155
Bairro: Prado Velho **CEP:** 80.215-901
UF: PR **Município:** CURITIBA
Telefone: (41)3271-2103 **Fax:** (41)3271-2103 **E-mail:** nep@pucpr.br