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

YOHAN BONESCKI GUMIEL

EXTRAÇÃO DE RELAÇÕES TEMPORAIS DE TEXTOS CLÍNICOS NA LÍNGUA PORTUGUESA

CURITIBA
2020

YOHAN BONESCKI GUMIEL

EXTRAÇÃO DE RELAÇÕES TEMPORAIS DE TEXTOS CLÍNICOS NA LÍNGUA PORTUGUESA

Tese de Doutorado 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 doutor em Tecnologia em Saúde.

Orientadora: Profa. Dra. Deborah Ribeiro Carvalho

Coorientadora: Profa. Dra. Claudia Maria Cabral Moro Barra

CURITIBA

2020

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Pamela Travassos de Freitas – CRB 9/1960

Gumiel, Yohan Bonescki
G974c 2020 Extração de relações temporais de textos clínicos na língua portuguesa / Yohan Bonescki Gumiel ; orientadora: Deborah Ribeiro Carvalho ; coorientadora: Claudia Maria Cabral Moro Barra. – 2020.
267 f. : il. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2020
Bibliografia: f. 164-178

1. Registros médicos. 2. Banco de dados. 3. Documentos médicos. 4. Processamento de linguagem natural (Computação). 5. Registros eletrônicos de saúde. 6. Redes neurais (Computação). 7. Registros hospitalares. I. Carvalho, Deborah Ribeiro. II. Barra, Claudia Maria Cabral Moro. III. Pontifícia Universidade Católica do Paraná. Pós-Graduação em Tecnologia em Saúde. III. Título.

CDD 20. ed. – 615.82019

TERMO DE APROVAÇÃO DE TESE Nº 013

A Tese de Doutorado intitulada "**EXTRAÇÃO DE RELAÇÕES TEMPORAIS DE TEXTOS CLÍNICOS NA LÍNGUA PORTUGUESA"** defendida em sessão pública pelo(a) candidato(a) **Yohan Bonescki Gumiel** no dia **24 de setembro de 2020**, foi julgada para a obtenção do título de Doutor em Tecnologia em Saúde, e aprovada em sua forma final, pelo Programa de Pós-Graduação em Tecnologia em Saúde.

BANCA EXAMINADORA:

Prof. Dr. Deborah Ribeiro Carvalho – Orientador e Presidente – PUCPR

Prof. Dr. Renata Vieira - PUCRS

Prof. Dr. Thiago Alexandre Salgueiro Pardo - USP

Prof. Dr. Emerson Cabrera Paraiso – PUCPR

Prof. Dr. José Rocha Faria Neto - PUCPR

A via original deste documento encontra-se arquivada na Secretaria do Programa, contendo a assinatura da Coordenação após a entrega da versão corrigida do trabalho.

Curitiba, 23 de novembro de 2020.

Prof. Dr. Percy Nohama,

Coordenador do PPGTS PUCPR

AGRADECIMENTOS

Primeiramente, aos meus pais, Júlia e Emanoel, pelo apoio e incentivos em todos os momentos desta longa jornada.

À minha orientadora, Profa. Dra. Deborah Ribeiro Carvalho, por todo o apoio e dedicação nestes quatro anos. Muitas das boas práticas que adquiri vieram deste convívio, tendo sido um período de aprendizado maravilhoso.

À minha coorientadora, Profa. Dra. Claudia Maria Cabral Moro Barra, por todo o apoio e dedicação nestes quatro anos. Com você, aprendi muito sobre como transmitir as minhas ideias.

Aos professores que participaram da minha banca de PTD, Profa. Dra. Renata Vieira, Prof. Dr. Thiago Alexandre Salgueiro Pardo, Prof. Dr. Emerson Cabrera Paraiso e Prof. Dr. Guilherme Nunes Nogueira Neto, cujas contribuições foram essenciais para o resultado desta tese.

Aos meus companheiros dos grupos de pesquisa, Lucas Oliveira, Marcelo, Lilian, João Cubas, Everton, Cristiane, Elziane, Ana, Bárbara, Lorena, Georgia, Marcia, João Vitor, Lucas Ferro e Elisa. Todas as conversas que tivemos e toda a ajuda que me foi fornecida nestes quatro anos foram essenciais.

Ao apoio da minha aluna de Pibic, Carolina Montenegro, sempre me ajudando com a parte da medicina. Aos alunos Luisa, Eros, Caroline e Carolina Dorigon, que contribuíram com o projeto, além do meu aluno de Pibic Júnior, Marcos.

RESUMO

Aplicações voltadas à inteligência artificial podem ser desenvolvidas para potencializar o uso de dados presentes nos registros eletrônicos de saúde, em sua grande parte registrados em formato de texto livre, que oportuniza o Processamento de Linguagem Natural (PLN). Dentre as pesquisas em PLN, a extração de relações temporais permite inferir a ordem entre menções, sejam elas eventos clínicos ou expressões temporais. O aspecto temporal é especialmente relevante para o acompanhamento de doenças crônicas não transmissíveis, como na cardiologia, pois não apenas possibilita identificar os eventos cronologicamente, mas também as janelas de tempo. A partir de uma revisão sistemática sobre extração de relações temporais no contexto clínico, foram identificados alguns pontos interessantes para pesquisa: (i) extração de relações temporais no contexto clínico para o idioma português; (ii) *corpora* anotados com eventos, expressões temporais e relações temporais diferentes dos poucos disponibilizados em *shared tasks*; (iii) camadas de anotação de eventos, expressões temporais e relações para o domínio cardiológico; (iv) métodos de anotação para o domínio cardiológico em textos ambulatoriais. A partir disso, esta tese propõe a construção de um modelo para extração de relações temporais em narrativas clínicas em língua portuguesa. A pesquisa foi desenvolvida em quatro etapas: (i) criação dos *guidelines* para anotações de eventos, expressões temporais e relações temporais; (ii) processo de anotação de eventos, expressões temporais e relações temporais; (iii) extração das relações temporais anotadas por abordagens baseadas em aprendizado de máquina; (iv) avaliação das abordagens desenvolvidas e do processo de anotação. A partir dos *guidelines* criados na etapa 1, o processo de anotação ocorreu inicialmente para os eventos, adicionando expressões temporais e, por último, as relações temporais, com anotação dupla e adjudicação. Na etapa 3, foram desenvolvidas quatro abordagens distintas baseadas em *support vector machines*, para extração de relações temporais entre: (i) evento-evento em mesma sentença; (ii) evento-expressão temporal em mesma sentença; (iii) evento-evento em sentenças distintas; (iv) evento-data de criação do documento. Vale destacar que o desenvolvimento das abordagens priorizou a limitação da quantidade de pares, fato crucial na extração de relações temporais. A avaliação da anotação (etapa 4) ocorreu por meio do *Inter-Annotator Agreement* (IAA) e das quatro abordagens, por meio do F1-*score*. Os *guidelines* se mostraram efetivos pelos valores de IAA encontrados. Para anotação de eventos, foram obtidos valores de IAA de 0,9066 para marcação do evento no texto (*span*), 0,9117 para anotação do atributo Tipo, 0,9967 para Modalidade, 0,9891 para Polaridade e 0,9546 para RelTempDCD. Para anotação de expressões temporais, foram obtidos valores de IAA de 0,9497 para marcação da expressão temporal no texto (*span*), 0,974 para anotação do atributo Tipo, 0,9468 para Valor, 0,9834 para Cps, 0,9858 para Quant e 0,9913 para Mod. Para relações temporais, obteve-se IAA de 0,8835, assim como valores de IAA comparáveis ou superiores aos principais trabalhos do domínio clínico para anotação de eventos e expressões temporais. Para a anotação de relações temporais, o desenvolvimento de um novo esquema resultou em valores de IAA superiores aos principais trabalhos do domínio clínico. Na extração de relações temporais, as heurísticas propostas se mostraram efetivas na diminuição da quantidade de possíveis pares, aumentando a *performance* das abordagens baseadas em aprendizado de máquina. Foi obtido F1-*score* de 0,8213 para relações temporais presentes no texto e 0,9270 para o atributo RelTempDCD.

Palavras-chave: Processamento de linguagem natural. Extração de relações temporais. Textos clínicos.

ABSTRACT

Applications focusing on Artificial Intelligence (AI) can be developed to enhance the use of the data present in Electronic Health Records (EHR), mostly recorded in free text format. This format allows Natural Language Processing (NLP). Among NLP research, temporal relations extraction allows systems to infer order between mentions, clinical events, or temporal expressions. The temporal aspect is especially relevant for monitoring Non-Communicable Diseases (NCDs), such as cardiology, as it allows identifying events chronology and the time windows. From a systematic review on the extraction of temporal relation in the clinical context, involving 101 articles, some interesting points for research were identified: (1) extraction of temporal relations in the clinical context for the Portuguese language, (2) corpora annotated with events, temporal expressions and temporal relations different from the few available in shared-tasks, (3) layers of events, temporal expressions and temporal relations annotations for the cardiology domain, and (4) methods of annotation for ambulatory texts in cardiology domain. From these points, this thesis proposed constructing a model for the extraction of temporal relations in clinical narratives written in the Portuguese language. The research was developed in four-step: (1) creation of guidelines for annotating events, temporal expressions, and temporal relations, (2) annotation process of events, temporal expressions, and temporal relations, (3) extraction of the annotated relations by approaches based on machine learning (ML), and (4) evaluation of the developed approaches and the annotation process. From the guidelines created in step 1, the annotation process occurred initially for the events, adding temporal expressions and finally the temporal relations, with double annotations and adjudication. In step 3, four different approaches based on Support Vector Machines (SVM) were developed to extract temporal relations between (1) event-event in the same sentence, (2) event-time expression in the same sentence, (3) event-event in different sentences, and (4) event and document creation date (DCT). It is worth mentioning that the development of the approaches prioritized the number of pairs' limitations, which is crucial in the extraction of temporal relations. The annotation evaluation (step 4) occurred with the inter-annotator agreement (IAA) and four approaches with the f-measure. The guidelines were effective according to the IAA values. For event annotations, IAA values of 0.9066 were obtained to mark the event span, which was 0.9117 for annotating the Tipo attribute. At 0.9967 for annotation of the Modalidade, 0.9891 for annotation of the Polaridade attribute, and 0.9546 for annotating the RelTempDCD attribute. For temporal expressions annotations, IAA values of 0.9497 were obtained to mark the temporal expression span, which is at a rate of 0.974 for annotating the Tipo attribute, which is 0.9468 for normalizing the temporal expression, 0.9834 for annotating the Cps attribute, 0.9858 for notation of the Quant attribute and 0.9913 for annotation of the Mod attribute. For temporal relations, an IAA of 0.8835 was obtained. The IAA values were comparable or superior to the primary studies in the clinical domain regarding events and temporal expressions annotations. For the annotation of temporal relations, the development of a new scheme resulted in IAA values higher than the primary studies in the clinical domain. In the extraction of temporal relations, the proposed heuristics proved to be effective in decreasing the number of possible pairs, increasing ML-based approaches' performance. It was obtained an F1-score of 0.8213 for the temporal relation over the text and 0.9270 for the RelTempDCD attribute.

Keywords: Natural language processing. Temporal relation extraction. Clinical texts.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BERT	*Bidirectional Encoder Representations From Transformers*
BOW	*Bag-of-Words*
CID-10	Classificação Internacional de Doenças e Problemas Relacionados à Saúde
CLEF	*Clinical E-Science Framework*
CRF	*Conditional Random Fields*
DCD	Data de Criação do Documento
DCNT	Doenças Crônicas Não Transmissíveis
DICOM	*Digital Imaging and Communications in Medicine*
DRG	*Diagnosis Related Groups*
DT	*Decision Tree*
DTD	*Document Type Definition*
ET	Expressão Temporal
EVT	Evento
FN	Falso Negativo
FP	Falso Positivo
HAS	Hipertensão Arterial Sistêmica
i2b2	*Informatics for Integrating Biology and Bedside*
IA	Inteligência Artificial
IAA	*Inter-Annotator Agreement*
ILP	*Integer Linear Programming*
LR	*Logistic Regression*
LSTM	*Bidirectional Long Short-Term Memory*
MAMA	*Model-Annotate-Model-Annotate*
MEDLINE	Sistema Online de Busca e Análise de Literatura Médica
MERLOT	*Medical Entity and Relation LIMSI Annotated Text*
NYHA	New York Hearth Association
OMS	Organização Mundial da Saúde
PLN	Processamento de Linguagem Natural
POS	*Part-Of-Speech Tagging*
REN	Reconhecimento de Entidades Nomeadas
RES	Registro Eletrônico em Saúde

RF	*Random Forest*
RT	Relação Temporal
SVM	*Support Vector Machine*
THYME	*Temporal History of Your Medical Events*
TLINK	*Temporal Link*
UMLS	*Unified Medical Language System*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Há décadas, é comprovado que pesquisas sobre Inteligência Artificial (IA), com foco na área da saúde, trazem uma grande variedade de possíveis aplicações, extremamente valiosas no contexto médico (DUDCHENKO; GANZINGER; KOPANITSA, 2020), as quais podem ser desenvolvidas a partir de dados presentes nos Registros Eletrônicos de Saúde (RES), que são repositórios de informações do paciente que contêm dados estruturados (como medicamentos, exames laboratoriais e radiológicos) e não estruturados (textos clínicos) (JENSEN; JENSEN; BRUNAK, 2012). Grande parte das informações clínicas contidas nos RES é registrada em formato de texto livre (texto clínico), sendo, então, um tipo de dado oportuno para o Processamento de Linguagem Natural (PLN), ramo da IA relacionado a dados em formato textual.

Entre os tópicos de pesquisa em PLN, a extração de Relações Temporais (RTs) pode afetar positivamente o campo da saúde, devido à possibilidade de obter informações ricas e contextuais dos textos clínicos que não estão disponíveis em outras fontes, incluindo históricos sobre estados/condições atuais do paciente e até mesmo sobre seu passado (como um tratamento feito há muito tempo) (NIKFARJAM; EMADZADEH; GONZALEZ, 2013; STYLER *et al.*, 2014a). Além de conter informações sobre o passado e o presente, esses textos fornecem informações sobre o futuro (como intervenções e tratamentos previstos). Um recorte de texto clínico é fornecido na Figura 1.

Figura 1 – Recorte de texto clínico.

Data de Criação do Documento: 26/05/2014

HAS há 10 anos
IAM + ATC em 2009.
RVM em 2013.
Tabagista há 40 anos (10 cigarros/dia).

O # Sem edema em mmii

LAB 13/01/14: CT 119

P # Aumento selozok para 100mg /dia.

Fonte: O autor (2020).

A extração de RTs pode ser entendida como o fornecimento aos sistemas da capacidade de inferir ordem entre as menções, sejam elas Eventos (EVTs) ou

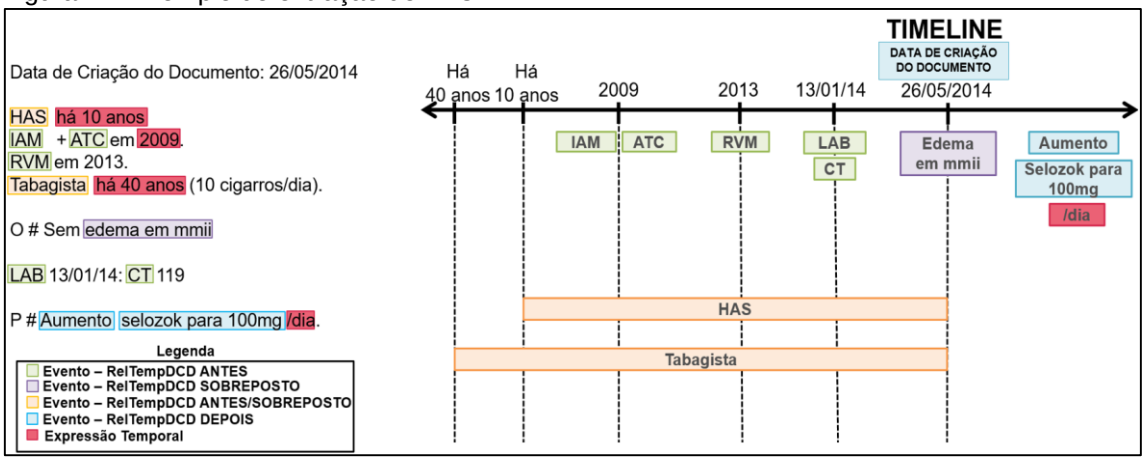
Expressões Temporais (ETs), por correlacionar repetidamente pares de menções em um texto, por meio de um conjunto de relações predefinidas, determinando, assim, como estão ligadas temporalmente.

No domínio geral, EVTs representam situações que ocorrem, tendo seu foco em ações, enquanto, no domínio clínico, significam qualquer questão importante relacionada ao paciente. Dada essa diferença, no domínio clínico, menções de EVTs são usualmente relacionadas a substantivos envolvendo questões como problemas médicos, tratamentos e testes. De forma distinta, menções de ETs são similares em ambos os domínios, geral e clínico, com diferença mais significativa nas marcações de frequência de uso de medicamento, podendo estas ser extremamente complexas, contemplando casos de diferentes números de comprimidos em distintos dias da semana, como “1/2 cp todos os dias (segunda a sábado) + 1 cp aos domingos”.

As RTs, no domínio clínico, podem ser complexas, uma vez que as marcações envolvem questões implícitas nos textos e questões dependentes de conhecimento do domínio (*expertise*). Os benefícios de sua extração podem ser observados na Figura 2, sendo verificado um maior detalhamento sobre a jornada do paciente. Ao contrário de simplesmente considerar que todos os EVTs ocorreram na mesma data, neste caso, a Data de Criação do Documento (DCD), as RTs trazem maior granularidade temporal acerca dos EVTs. Foram extraídos dois tipos de RT, os mesmos utilizados neste projeto, a saber: (i) RelTempDCD, envolvendo EVTs e DCD (tipo específico de ET); (ii) *Temporal Link* (TLINK), envolvendo EVTs e ETs presentes no texto. A RelTempDCD localiza o EVT em relação à DCD, neste caso, a data da consulta, trazendo certa informação temporal. Ainda, existem EVTs com marcações da RelTempDCD de antes, indicando EVTs passados que ocorreram antes da DCD, sendo: “IAM” (infarto agudo do miocárdio), “ATC” (angioplastia), “RVM” (revascularização do miocárdio), “LAB” (laboratório) e “CT” (colesterol total). Além deles, há: EVTs contínuos, com marcações de RelTempDCD de antes/sobreposto, representados por “HAS” (hipertensão arterial sistêmica) e “tabagista”; questões sobre o momento da consulta, como “edema de mmii” (edema em membros inferiores) encontrado durante o exame físico, com marcação de sobreposto; e questões relacionadas à conduta, como “aumento” e “selozok para 100mg”, com marcações de RelTempDCD de depois. A RelTempDCD já fornece alguma informação temporal, porém EVTs como “IAM”, “ATC”, “RVM”, “LAB” e “CT” ocorreram em momentos distintos no tempo, sendo necessárias TLINKs para inferir ordem. A mesma questão

ocorre com “HAS” e “tabagista”, só sendo possível diferenciar a duração deste EVT a partir das TLINKs. Sendo assim, com as TLINKs é possível inferir que: “IAM” e “ATC” ocorreram em “2009”, sendo que “IAM” veio antes de “ATC”; “RVM” ocorreu após esses EVTs, em “2013”; e “LAB” e “CT” ocorreram mais próximos da DCT, em “13/01/14”. Ademais, é possível inferir que o paciente é “tabagista” há 40 anos e apresenta “HAS” há dez anos.

Figura 2 – Exemplo de extração de RTs.



Fonte: O autor (2020).

O aspecto temporal é relevante para as Doenças Crônicas Não Transmissíveis (DCNT), pois, por sua natureza longitudinal, fornecem grandes e contínuos fluxos de dados, dos quais padrões significativos podem ser extraídos (SHEIKHALISHAHI *et al.*, 2019). Em textos clínicos, a progressão da doença e os EVTs são registrados cronologicamente e certos EVTs são relevantes apenas em determinada janela temporal (SHEIKHALISHAHI *et al.*, 2019), como, por exemplo, problemas médicos encontrados durante o exame físico em um encontro ambulatorial.

As DCNT são um tema de pesquisa importante, uma vez que são responsáveis por mais de 70% das mortes em todo o mundo (WHO *et al.*, 2013). Segundo a Organização Mundial da Saúde (OMS), a maioria das mortes prematuras por DCNT poderia ser evitada; uma das ações seria capacitar os sistemas de saúde para responder de forma mais equitativa e eficaz (WHO *et al.*, 2013). Assim, o fornecimento de ferramentas adicionais aos profissionais de saúde pode beneficiar a saúde e diminuir o número de mortes por DCNT.

Entre as doenças crônicas, as cardiovasculares têm forte efeito temporal. Uma mesma doença pode exigir diferentes intervenções de prevenção ou tratamento em

distintos momentos (JOHNSON *et al.*, 2017); por exemplo, se, durante o tratamento da Hipertensão Arterial Sistemática (HAS) com monoterapia para um paciente, não for possível obter controle adequado da pressão arterial, a tendência será implementar uma terapia combinada com dois ou três medicamentos. Na Figura 2, nota-se essa questão temporal, com o paciente tendo hipertensão arterial há determinado período e já tendo passado por diversos tratamentos, como angioplastia e revascularização do miocárdio no passado. Além disso, as doenças cardiovasculares, em conjunto com o câncer, foram responsáveis por mais da metade das mortes no Brasil, entre adultos de 45 a 64 anos, de 2000 a 2017 (LOTUFO, 2019). Sendo a causa mais comum de morte no Brasil, elas foram estabelecidas como tópico de pesquisa neste projeto.

Para levantar as lacunas da extração de RTs em textos clínicos, foi realizada uma revisão sistemática (na íntegra no Apêndice A). Constatou-se que, em 92 dos 101 artigos, os *corpora* anotados utilizados estavam em idioma inglês, não tendo sido encontrado nenhum trabalho para o idioma português. Isso evidencia uma lacuna de trabalhos em idiomas diferentes do inglês, principalmente o português.

Outro aspecto verificado na revisão foi a não existência de *corpora* além dos disponibilizados em *shared tasks* com camadas de anotações de EVTs, ETs e RTs, com exceção do *corpus* francês *Medical Entity and Relation LIMSI Annotated Text* (MERLOT) (CAMPILLOS *et al.*, 2018), o que demonstra tanto a complexidade de criar um *corpus* anotado com diversas camadas quanto a dificuldade da anotação de RTs, que usualmente apresentam valores de concordância inferiores às anotações de EVTs e ETs (independentemente do domínio do texto). Dessa forma, há uma lacuna envolvendo a falta de *corpora* anotados com camadas de anotações de EVTs, ETs e RTs no domínio clínico.

Ainda, além do trabalho de Viani *et al.* (2019), que contempla um tipo mais simples de RT, do EVT com a DCD, a não existência de nenhum estudo envolvendo a extração de RTs complexas (incluindo EVTs e ETs contidos no texto) para o domínio cardiológico demonstra uma lacuna da aplicação da extração de RTs para o domínio cardiológico.

Também há as dificuldades trazidas pelas próprias características dos textos clínicos, como formatação flexível, gramática atípica, taquigrafia abundante, erros ortográficos e vocabulário especialista, que tornam a extração de RTs em textos de domínio clínico desafiadora (JAGANNATHA; YU, 2016; KREIMEYER *et al.*, 2017; LEAMAN; KHARE; LU, 2015; MEYSTRE *et al.*, 2008).

Diante disso, este trabalho se propõe a lidar com as lacunas evidenciadas, criando um *corpus* anotado com EVTs, ETs e RTs, baseado em textos ambulatoriais cardiológicos em idioma português, tendo o objetivo final da extração de RTs. Para tanto, foi elaborada a seguinte pergunta de pesquisa: como pode ser feita a extração automática de RTs de narrativas clínicas para um *corpus* da área clínica anotado em português para auxiliar no sequenciamento de EVTs clínicos?

O objetivo é construir um modelo para extração de RTs em narrativas clínicas em língua portuguesa, sendo objetivos específicos:

Definir *guidelines* para anotação de EVTs, ETs e RTs para o domínio clínico.

Criar um *corpus* anotado com EVTs, ETs e RTs a partir de textos ambulatoriais de cardiologia.

Criar modelos especialistas para extração de RTs para sequenciar EVTs clínicos.

Avaliar o desempenho dos modelos desenvolvidos, assim como das heurísticas propostas, por meio de métricas de avaliação.

A contribuição científica está em trabalhar com um tema de estudo complexo, necessitando de diversas camadas de anotação, criando um *corpus* anotado, com intenção de torná-lo publicamente disponível, além da realização de um estudo que pode ser replicado para outros idiomas e especialidades médicas.

A contribuição social é a instrumentalização do profissional de saúde para o melhor entendimento da jornada do paciente no que se refere ao cuidado dispensado ao longo do tempo. No caso de pacientes com diversas consultas, ter uma visão longitudinal sobre os diversos problemas, exames e tratamentos pode trazer como benefícios um diagnóstico mais claro e uma conduta mais precisa. Ainda, esses dados com informações temporais, vindos dos textos, podem ser combinados com dados estruturados presentes nos RES para fornecer uma descrição longitudinal mais detalhada e precisa do paciente, com aplicações diretas em questões como codificação clínica (como Classificação Internacional de Doenças e Problemas Relacionados à Saúde – CID-10 e *Diagnosis Related Groups* – DRG), na avaliação de desfecho, na classificação de risco, na melhora da qualidade da informação e na identificação de EVTs adversos.

2 REFERENCIAL TEÓRICO

Em função da característica multidisciplinar desta pesquisa, este capítulo trata de assuntos relativos a diferentes áreas do conhecimento, como aspectos da área de saúde e computacionais, linguagem e ontologias temporais.

Na seção 2.1, aborda-se o tema RES, trazendo questões relativas à sua composição, assim como questões como o formato estruturado de documentação subjetivo, objetivo, avaliação e plano (SOAP) e as características encontradas em textos ambulatoriais. Na seção 2.2, é detalhado o processo de anotação, necessário para obtenção do *corpus* anotado, sendo detalhadas suas etapas e boas práticas. Na seção 2.3, trata-se do tema aprendizado de máquina, conceito fundamental para extração automática de RTs, trazendo questões sobre o tema em geral e um detalhamento do algoritmo utilizado neste projeto. Na seção 2.4, aborda-se o PLN, com alguns conceitos-base, tarefas de PLN para domínio geral e o conceito de PLN clínico. Na seção 2.5, o tema-foco são os padrões de anotação para EVTs, ETs e RTs, trazendo definições do domínio geral e as adaptações necessárias para o domínio clínico. Na seção 2.6, é fornecido um mapeamento entre o referencial teórico e os encaminhamentos metodológicos.

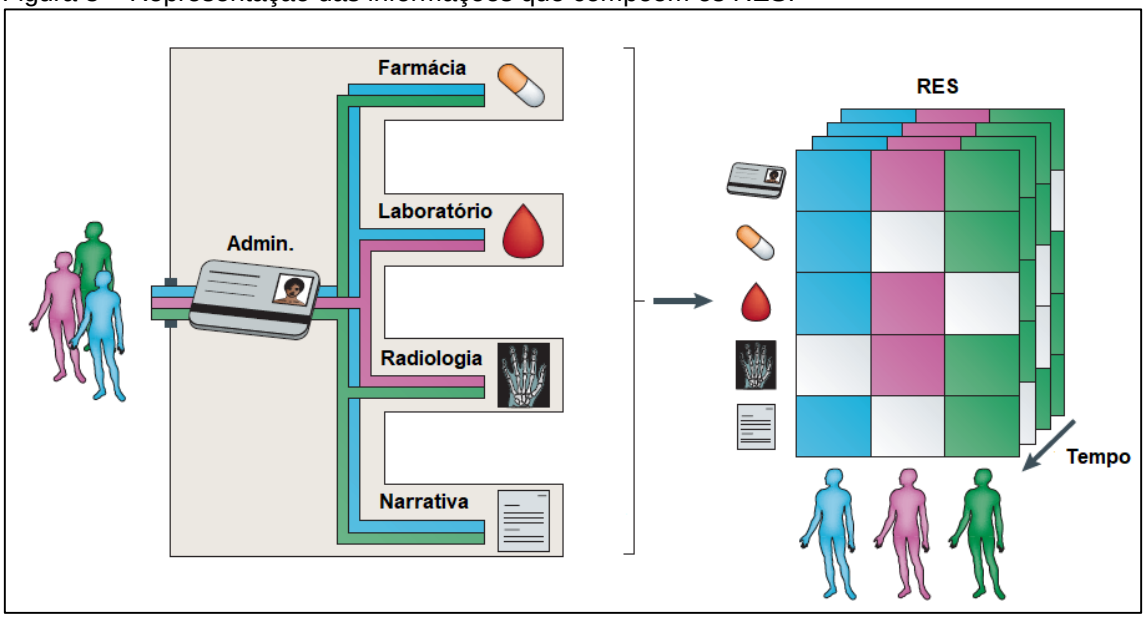
2.1 REGISTRO ELETRÔNICO EM SAÚDE

Os RES podem ser definidos como repositórios de informações sobre a condição de saúde e assistência prestada ao paciente, de forma que essas informações possam servir para múltiplos usos e usuários (MCDONALD; TANG; HRIPCSAK, 2014). Com sua ampla utilização, vem sendo criada uma disponibilidade crescente de dados clínicos coletados rotineiramente em formato eletrônico que podem ser usados para fins secundários, como, por exemplo, para pesquisa (CAPURRO *et al.*, 2014). Esse aumento no volume de RES disponíveis para fins de pesquisa fornece novas oportunidades para criar aplicações de assistência médica interoperáveis semanticamente e soluções para medicina baseadas em evidências (TAO *et al.*, 2010).

Algumas informações contidas nos RES estão em formato estruturado, como medicações, exames laboratoriais e de radiologia (Figura 3). Por exemplo, exames de imagem seguem o padrão de documentação *Digital Imaging and Communications in*

Medicine (DICOM), sendo esse o padrão internacional para qualquer atividade (como transmissão e armazenamento) envolvendo imagens médicas. No entanto, existem informações que estão em formato não estruturado, representadas pelos documentos clínicos redigidos em formato de texto livre. Em comparação com o formato estruturado, o texto livre é bem mais natural e expressivo para documentar EVTs clínicos, além de facilitar a comunicação da equipe de atendimento no ambiente da saúde (WANG *et al.*, 2018). Grande parte da informação clínica é armazenada em formato de texto livre (MEYSTRE *et al.*, 2017).

Figura 3 – Representação das informações que compõem os RES.



Fonte: Adaptado de Jensen, Jensen e Bunak (2012).

Existem diversos tipos de documento clínico contidos nos RES, podendo ser divididos em duas categorias principais: relatórios de diagnóstico e notas clínicas (WANG *et al.*, 2018).

Os relatórios de diagnóstico são fornecidos pelos serviços de diagnóstico, como relatórios de laboratório, radiologia e patologia. Por exemplo, relatórios de radiologia envolvem resultados de exames radiológicos e imagens de raios-X de diversas partes do corpo e órgãos específicos do paciente. Já relatórios de patologia contemplam exames patológicos de amostras de tecidos e tecidos de órgãos removidos durante os procedimentos cirúrgicos (WANG *et al.*, 2018).

As notas clínicas são usualmente geradas a cada interação entre pacientes e sistema de saúde, contendo registros de observações, impressões, planos e outras atividades decorrentes de episódios de atendimento ao paciente (ROSENBLOOM *et*

al., 2010). Entre elas, alguns exemplos de textos são os sumários de alta e textos ambulatoriais, fontes de dados para pesquisa. Os sumários de alta descrevem o resultado da internação, a condição de alta e as instruções de cuidado, enquanto textos ambulatoriais trazem os detalhes da consulta de pacientes atendidos ambulatoriamente, sem necessidade de internação (WANG *et al.*, 2018).

Textos de ambulatório são essenciais para entender a progressão da doença, justamente pelo fato de trazerem informações sobre todas as intervenções e tratamentos realizados pelo paciente ao longo de múltiplas consultas. Contudo, ao contrário dos sumários de alta, que são normalmente bem escritos e estruturados, uma vez que são redigidos para uma maior audiência, os textos ambulatoriais são redigidos somente para os profissionais de saúde da unidade clínica (DALIANIS, 2018). O padrão de documentação usualmente utilizado neles será detalhado na seção 2.1.1 e as características e dificuldades de trabalhar com esse tipo de texto, na seção 2.1.2.

2.1.1 Formato estruturado SOAP

Desde o começo dos estudos, médicos e outros profissionais da área da saúde são ensinados sobre a importância de não confiar na sua memória quando prestando cuidado a pacientes, devendo registrar suas observações, ações e respectivos motivos, para posterior comunicação com eles mesmos ou outros profissionais de saúde (SHORTLIFFE; BARNETT, 2014).

No domínio clínico, registros médicos são frequentemente escritos usando o formato estruturado SOAP, proposto por Weed (1964), que permite aos profissionais de saúde identificar, priorizar e acompanhar os problemas dos pacientes de forma que eles possam ser atendidos de forma pontual e organizada, além de fornecer uma avaliação contínua nos tratamentos propostos e na evolução do paciente (CAMERON; TURTLE-SONG, 2002).

A estrutura possibilita que os médicos revisem as anotações escritas por eles próprios ou por outros médicos, não existindo um padrão ideal para todos os médicos, especialidades e locais; dessa forma, o SOAP tem pequenas variações, representando quatro categorias principais de documentação: Subjetivo (S), Objetivo (O), Avaliação (A) e Planejamento/Plano (P) (PEARCE *et al.*, 2016). Um exemplo de registro médico escrito em SOAP é mostrado na Figura 4.

Figura 4 – Exemplo de texto ambulatorial de cardiologia escrito no padrão SOAP.

Fonte: O autor (2020).

A seção Subjetivo vem das experiências do paciente, em relação às informações contadas ao profissional de saúde (CAMERON; TURTLE-SONG, 2002; LENERT, 2016). Contempla histórico médico familiar, história médica anterior (com histórico de cirurgias e hospitalizações), comorbidades, alergias, hábitos (incluindo tabagismo), medicações em uso e sintomas experienciados (PEARCE *et al.*, 2016).

A seção Objetivo contém dados que podem ser diretamente verificados pela observação (PEARCE *et al.*, 2016). Existem dois tipos de dado objetivo: observações do profissional de saúde (achados do exame físico e visual) e material escrito externo (informações diagnósticas obtidas por testes de laboratório, radiologia ou outros) (CAMERON; TURTLE-SONG, 2002; LENERT, 2016; PEARCE *et al.*, 2016). O exame físico normalmente adota uma sequência da cabeça aos pés e começa com dados constitucionais (como altura, peso e sinais vitais) (PEARCE *et al.*, 2016).

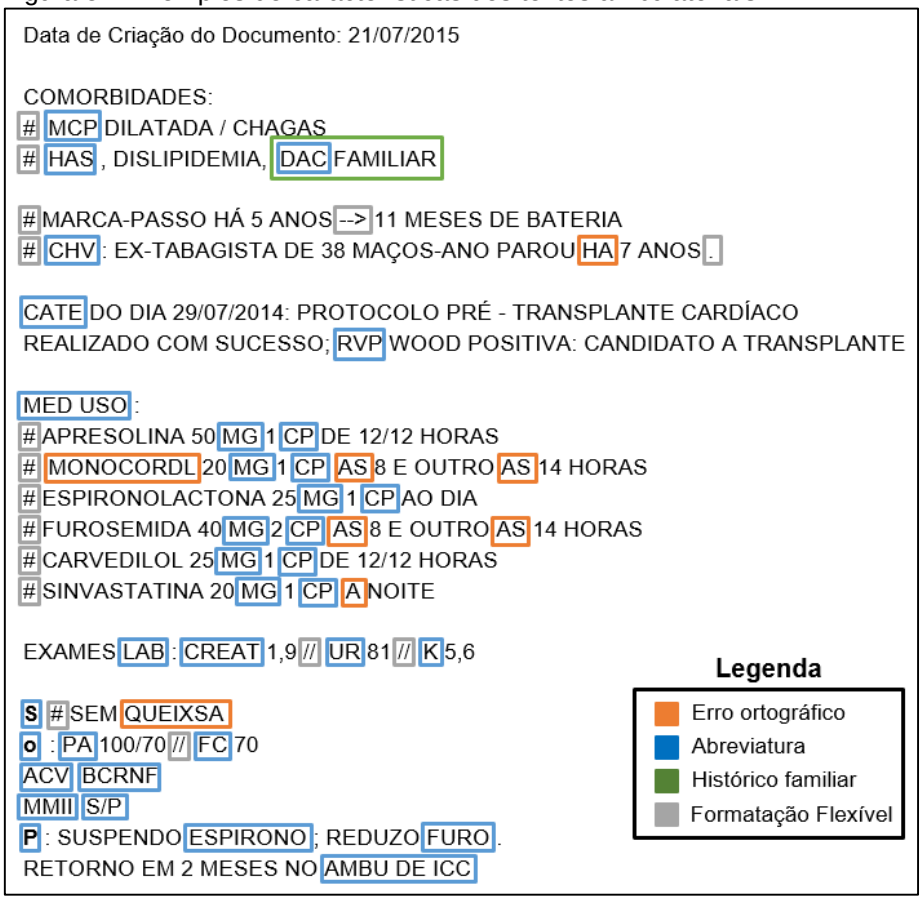
A seção Avaliação é a sumarização do pensamento do profissional de saúde sobre os problemas do paciente (CAMERON; TURTLE-SONG, 2002). Nela, as seções anteriores são sintetizadas e analisadas para chegar a um diagnóstico (LENERT, 2016).

Por fim, a seção Planejamento menciona as estratégias usadas para abordar o diagnóstico da seção Avaliação, fornecendo detalhes sobre testes adicionais, consulta com outros profissionais de saúde e os passos a serem tomados para tratar a doença do paciente (LENERT, 2016; PEARCE *et al.*, 2016).

2.1.2 Características dos textos ambulatoriais

Estes textos têm a característica de serem curtos e eficientes, além de escritos em estilo telegráfico, fazendo uso de abreviações e compactação de informações para reduzir seu tamanho (DALIANIS, 2018). Por serem escritos durante a consulta ambulatorial, existe a necessidade de serem textos práticos e rápidos, pois há um tempo limitado para expressar todos os detalhes relativos à consulta (LIN *et al.*, 2016a). Algumas características são mostradas na Figura 5, salientando-se que é um texto ambulatorial cardiológico, foco deste trabalho.

Figura 5 – Exemplos de características dos textos ambulatoriais.



Fonte: O autor (2020).

Usualmente, estes textos têm uma grande dependência da especialidade, visto que cada especialidade dentro da medicina usa seu próprio conjunto de termos, que podem ser incompreensíveis para outras (DALIANIS, 2018). São textos impregnados de acrônimos e abreviaturas, pela característica de tempo limitado de escrita.

Abreviaturas são formas incompletas de palavras, sendo eficientes para escrita, porém tornam lenta a leitura, uma vez que o leitor tem que as interpretar (DALIANIS, 2018). Um dos problemas do seu uso é o fato de não serem padronizadas, podendo haver diferentes significados de acordo com o contexto em que são aplicadas (SHORTLIFFE; BARNETT, 2014). Hospitais tentam estabelecer um padrão de abreviaturas aceitável, com seus respectivos significados, porém normalmente essas ações não têm sucesso (SHORTLIFFE; BARNETT, 2014). Por exemplo, o termo “gotas” pode ser abreviado como gt, gts ou GGT (vindo do latim *guttae*) (POZZOBON, 2011). Na Figura 5, diversas abreviações são utilizadas no texto, sendo algumas delas: med (medicamentos), cate (cateterismo cardíaco), lab (laboratoriais), espirono (espironolactona), furo (furosemida), creat (creatinina) e ambu (ambulatório).

Acrônimos são uma forma especializada de abreviação, normalmente usando a primeira letra de cada palavra na frase ou algum tipo de combinação de letras das palavras na frase (DALIANIS, 2018). Na Figura 5, estão diversos acrônimos, tais como: HAS (hipertensão arterial sistêmica), ACV (aparelho cardiovascular) e BCRNF (bulhas rítmicas normofonéticas).

Uma questão importante destes textos é a presença de erros ortográficos abundantes, especialmente em sistemas sem suporte (como suporte de ortografia/correção) (MEYSTRE *et al.*, 2008). Existe dificuldade de criar um sistema com suporte, devido ao vocabulário não padronizado utilizado (DALIANIS, 2018). Na Figura 5, são apresentados diversos erros ortográficos, alguns representados por: monocordl (erro ortográfico em monocordil) e queixsa (erro ortográfico em queixas). Além deles, há falta de acentuações, em casos como: ha (falta de acentuação em há) e as (falta de acentuação em às).

Outra característica é a formatação flexível, não existindo regras, com casos de falta de pontuação em sentenças ou expressões entre parênteses (como resultados de exames) (LEAMAN; KHARE; LU, 2015). Esse aspecto é representado na Figura 5 pelo uso de símbolos como: //, -->, #, pelo inconstante uso de pontuações e pela utilização de quebras de linha como delimitadores de sentença. Além disso, tem-se a gramática atípica, com a falta de palavras esperadas (como verbos, objetos ou artigos) (LEAMAN; KHARE; LU, 2015).

O contexto dos textos é dificultado pelo uso de muitas negações ou referências a diferentes assuntos, além de as avaliações poderem envolver incertezas e/ou tentativas de diagnóstico (JENSEN; JENSEN; BRUNAK, 2012). Negações são extremamente comuns em textos clínicos, uma vez que são usadas pelos médicos para excluir sintomas enquanto determinam o motivo da doença do paciente (DALIANIS, 2018). Não existem tantos casos de negação na Figura 5, porém, na Figura 4, há diversos elementos negados pelo uso das palavras “nega” e “sem”.

Nas Figura 4 e 5, tem-se outra questão importante dos textos de ambulatório: em diversos momentos, os problemas médicos citados são referências à família do paciente, fazendo parte da seção de histórico familiar. No entanto, o maior problema é a formatação do SOAP sobre as notas, não existindo um padrão específico seguido para todos os textos. A escrita sob a seção correta e até o próprio nome da seção podem variar de acordo com as unidades clínicas e hospitais (DALIANIS, 2018). Por exemplo, na Figura 5, não há seção Avaliação, pois o problema já foi diagnosticado

anteriormente e apenas os medicamentos foram ajustados. Ademais, somente os sintomas relatados e negados estão presentes na seção Subjetivo, fato recorrente nos textos. Inclusive, existem certos textos em que nenhuma seção do SOAP é marcada ou apenas seções como Avaliação o são.

2.2 ANOTAÇÃO DE TEXTOS CLÍNICOS

O processo de anotação consiste em adicionar informações linguísticas a um *corpus*, que pode ser definido como uma coleção de textos extraídos de qualquer fonte ou gênero, assim como transcrições de gravações de áudio, que são amostradas para representar questões de determinado idioma e/ou domínio linguístico (IDE, 2017; KUBLER; ZINSMEISTER, 2015; LU, 2014). Existem diversas informações linguísticas que podem ser adicionadas a *corpora*, como informações léxicas, morfológicas, sintáticas, semânticas e discursivas (LU, 2014). Algumas informações semânticas são entidades nomeadas, EVTs, ETs e RTs (IDE, 2017; KUBLER; ZINSMEISTER, 2015). Certos *corpora* contêm um ou dois tipos de anotação, enquanto outros incluem múltiplas camadas de anotação linguística (IDE, 2017). Nesta tese, existem três anotações, com EVTs e ETs anotados em primeiro plano e, em seguida, as RTs.

A anotação pode ser definida como o processo de adicionar rótulos a informações para melhorar a capacidade da máquina de realizar tarefas de PLN. No entanto, para ensinar a máquina efetivamente, é importante fornecer dados corretamente e na quantia necessária. Para certos domínios, pode existir dificuldade tanto no acesso aos dados quanto na obtenção de dados suficientes. Ainda, aspectos como tempo, dinheiro e recursos limitados podem limitar a quantidade de anotações a ser feita (PUSTEJOVSKY; STUBBS, 2012). Segundo Kubler e Zinsmeister (2015), como a anotação é uma ação difícil e consome muito tempo e mão de obra, é necessário o conhecimento de “boas práticas” antes de começar seu processo, as quais serão detalhadas adiante.

Uma das questões-chave é compreender a tarefa proposta e decidir quais perfis de anotadores serão necessários para prosseguir com o processo de anotação dos textos. Fazendo uma comparação entre textos biomédicos e clínicos, em documentos biomédicos, será mais fácil de o anotador identificar expressões genéticas em artigos científicos se já estiver familiarizado com os conceitos e vocabulário; já os documentos clínicos, como textos de ambulatório e sumários de

alta, podem ser mais complexos, existindo uma grande chance da necessitar de alguém da área da saúde, seja da medicina, seja da enfermagem, para interpretar os textos densos e cheios de jargões (PUSTEJOVSKY; STUBBS, 2012). Esse aspecto também é apontado por Roberts *et al.* (2009), mencionando que muitas informações dos textos clínicos podem ser entendidas, mesmo que de maneira simplista e de forma demorada, por alguém sem *expertise* da área da saúde com um dicionário médico, porém, para determinar relações entre entidades, é preciso um conhecimento maior.

O primeiro passo da anotação consiste na criação do *guideline*, em que o esquema é codificado, sendo definidas anotações corretas para cada fonte específica (ARSTEIN, 2017). O *guideline* é fornecido aos anotadores para que haja consistência na anotação, descrevendo em detalhes o que deve e não deve ser anotado, trazendo exemplos e mencionando como lidar com casos especiais (ROBERTS *et al.*, 2009). Exemplos podem ser adicionados para fins ilustrativos, porém não devem substituir a definição, pois é importante que os anotadores entendam a lógica por trás das anotações (FORT *et al.*, 2009). O *guideline* também fornece a sequência de passos, uma receita que os anotadores devem seguir quando trabalham com o documento (ROBERTS *et al.*, 2009).

Entretanto, escrever um *guideline* não é uma tarefa realizada em um momento único no começo do projeto, com pequenas modificações ao longo do tempo. Na verdade, ele evolui durante todo o processo de anotação, sendo essa evolução uma condição necessária para sua usabilidade como documentação que acompanha o *corpus* anotado (FORT, 2016). Inclusive, antes de começar a anotar os documentos reais, existe um processo de treinamento dos anotadores e refinamento incremental do *guideline*.

Usualmente, os anotadores são treinados para a tarefa, tanto na questão da própria anotação quanto no uso da ferramenta de anotação (FORT, 2016). Essas ferramentas tornam o processo mais prático, especialmente quando se usam linguagens de marcação, como *Extensible Markup Language* (XML), que se tornou a escolha padrão em projetos de anotação de *corpus*, assim como em ferramentas de anotação, por facilitar a adição de várias camadas de anotação que podem ser facilmente separadas do *corpus* original (sem anotações) (LU, 2014). Um exemplo de arquivo anotado no padrão XML, contendo algumas menções de EVTs (seguindo o padrão da tese), é mostrado na Figura 6. Mais detalhes sobre o padrão de anotação de EVTs podem ser encontrados na seção 4.2.

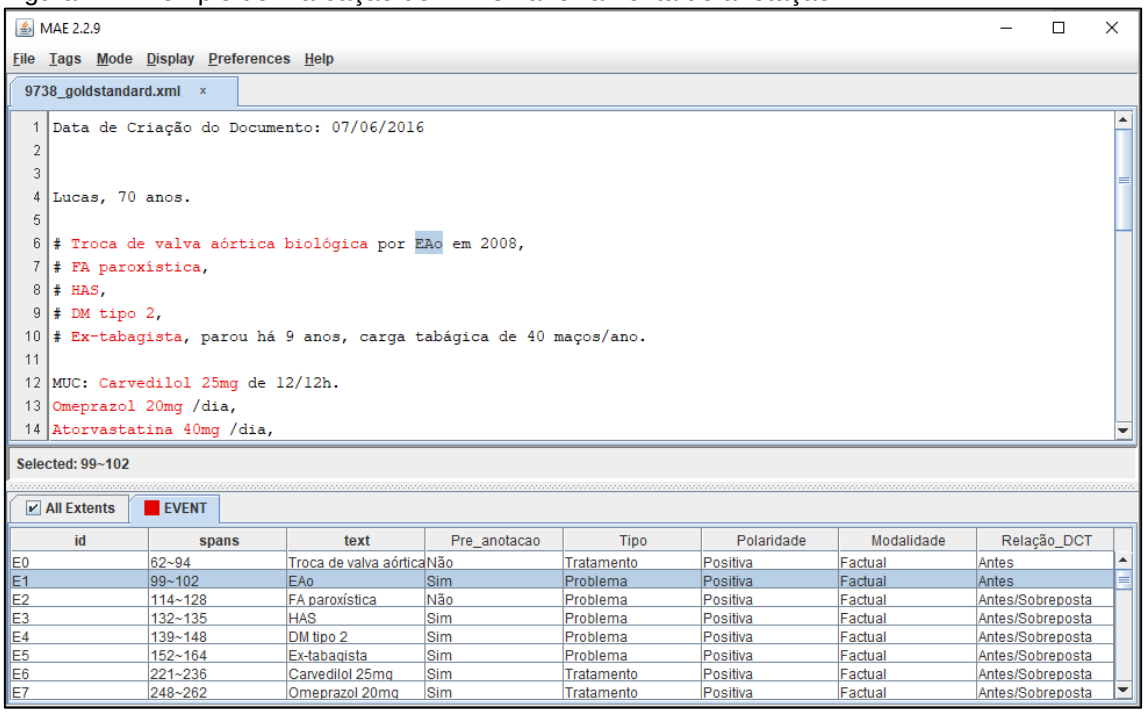
Figura 6 – Exemplo de marcação de EVTs seguindo o padrão XML.

```
<?xml version="1.0" encoding="UTF-8" ?>

<Annotation_v1>
<TEXT><![CDATA[Data de Criação do Documento: 14/08/2015

MCP  DILATADA
CINTILO  NEGATIVA  E CETS  SEM  CORONARIOPATIAS
13-7-15  HOLTER  ALTA  DENSIDADE DE ESV
NAO RETORNOU  AO AMB  IC ?
S   VERTIGEM  ESPORADICA - ASSOCIO  A  DOSE  OTIMIZADA  DE  CVD
P  MANETNHO CVD  12,5 12-12
MANTENHO AAS    SVT
ENCAMINHAO AO AMB  ARRITMIAS   VIA ROTA

 ]]></TEXT>
<TAGS>
<EVENT id="E0" spans="43~56" text="MCP  DILATADA" Tipo="Problema" Polaridade="Positiva" Modalidade="Factual"
DocTimeRel="Antes/Sobreposta" />
<EVENT id="E1" spans="59~66" text="CINTILO" Tipo="Teste" Polaridade="Positiva" Modalidade="Factual"
DocTimeRel="Antes" />
<EVENT id="E2" spans="80~84" text="CETS" Tipo="Teste" Polaridade="Positiva" Modalidade="Factual"
DocTimeRel="Antes" />
```

Fonte: O autor (2020).

As interfaces das ferramentas de anotação evitam o processo tedioso de escrever manualmente as marcações e os erros de digitação relacionados. Para anotadores muito competentes no seu domínio, mas não muito familiarizados com computadores, é importante achar a ferramenta mais apropriada, mesmo que seja menos eficiente (FORT, 2016). Escolher uma ferramenta de anotação que tenha uma curva de aprendizado muito acentuada ou uma que seja muito fácil de cometer erros pode causar erros no processo de anotação (PUSTEJOVSKY; STUBBS, 2012). É interessante buscar ferramentas de anotação didáticas e de fácil utilização, além de fornecer um *guideline* extra sobre seu uso, trazendo exemplos para todas as etapas. Na Figura 7, tem-se um exemplo da ferramenta de anotação MAE2 (RIM, 2016), que foi utilizada nesta tese, simulando uma anotação de EVTs.

Figura 7 – Exemplo de marcação de EVTs na ferramenta de anotação MAE2.



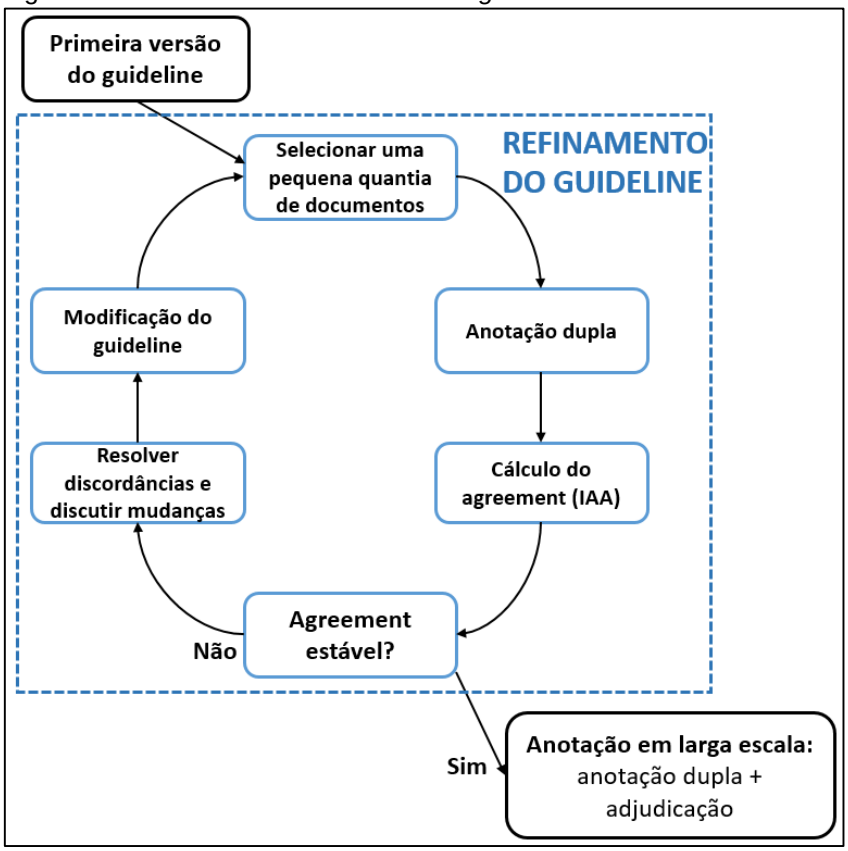
id	spans	text	Pre_anotacao	Tipo	Polaridade	Modalidade	Relação_DCT
E0	62~94	Troca de valva aórtica	Não	Tratamento	Positiva	Factual	Antes
E1	99~102	EAo	Sim	Problema	Positiva	Factual	Antes
E2	114~128	FA paroxística	Não	Problema	Positiva	Factual	Antes/Sobreposta
E3	132~135	HAS	Sim	Problema	Positiva	Factual	Antes/Sobreposta
E4	139~148	DM tipo 2	Sim	Problema	Positiva	Factual	Antes/Sobreposta
E5	152~164	Ex-tabagista	Sim	Problema	Positiva	Factual	Antes/Sobreposta
E6	221~236	Carvedilol 25mg	Sim	Tratamento	Positiva	Factual	Antes/Sobreposta
E7	248~262	Omeprazol 20mg	Sim	Tratamento	Positiva	Factual	Antes/Sobreposta

Fonte: O autor (2020).

A questão central da anotação de um *corpus* envolve a obtenção de anotações confiáveis, que sejam tanto informativas quanto consistentes (FORT, 2016). Para isso, o *guideline* é desenvolvido e refinado usando um processo iterativo, a fim de assegurar a consistência (ROBERTS *et al.*, 2009). Esse processo segue um padrão definido como ciclo *Model-Annotate-Model-Annotate* (MAMA), em que o modelo do *guideline* é refinado e o anotador é treinado para aquela tarefa até atingir resultados satisfatórios para uma anotação em larga escala (real) (PUSTEJOVSKY; STUBBS, 2012).

Na Figura 8, é mostrado o ciclo de refinamento proposto durante a anotação do *Clinical E-Science Framework* (CLEF) *corpus*, detalhado por Roberts *et al.* (2009). O primeiro passo consiste em elaborar a primeira versão do *guideline*, objetivando melhor definir suas marcações e trazendo alguns exemplos iniciais. O próximo passo é um ciclo iterativo de refinamento do processo, em que a cada rodada é fornecida uma pequena quantidade de documentos, feitas a anotação deles e uma consequente avaliação do desempenho. Esse processo pode ter dois desfechos: (i) uma anotação consistente e estável, podendo ser anotados os documentos finais/reais; (ii) uma anotação não consistente e/ou estável, sendo necessária outra rodada de refinamento. Não existe um valor aproximado do número de iterações necessárias, até porque tende a aumentar de acordo com a complexidade da tarefa.

Figura 8 – Desenvolvimento iterativo do *guideline*.



Fonte: Adaptado de Roberts *et al.* (2009).

Na Figura 8, são trazidos três aspectos fundamentais de um processo de anotação: a anotação dupla, a avaliação da anotação pelo *Inter-Annotator Agreement* (IAA) e o processo de adjudicação dos documentos, sendo os dois primeiros aspectos fundamentais tanto na etapa de refinamento quanto na anotação real e o último, ao final da anotação real para gerar o seu "padrão-ouro" (*gold standard*).

O primeiro aspecto a ser detalhado é a anotação dupla, sendo esta a estratégia selecionada devido aos diversos problemas associados a documentos anotados individualmente, como as peculiaridades e inconsistências de anotação de um anotador individual e o usual baixo desempenho de certos anotadores (ROBERTS *et al.*, 2009). Um exemplo prático de erro de um único anotador pode ser a correta marcação especificada no *guideline* de um problema médico como "dispneia ao esforço" *versus* a marcação incorreta de um anotador de somente "dispneia" ou uma marcação nula (nada é considerado), dado que, no contexto cardiológico, ter informação sobre a "dispneia" e, principalmente, sua característica de ser "em repouso" é algo importante de ser considerado. A questão-chave é que esse mesmo erro pode acabar se repetindo para todos os documentos anotados por esse único

anotador, impactando negativamente em todo o processo. Existem diversos esquemas de anotação alternativos desenvolvidos para superar isso, todos eles envolvendo mais tempo de anotação. A anotação dupla é uma alternativa amplamente utilizada, em que cada documento é anotado independentemente por dois anotadores e o conjunto de anotações é comparado para um consenso (ROBERTS *et al.*, 2009). Uma das suas vantagens é adicionar a possibilidade de medir a "qualidade" da anotação, a concordância entre os anotadores.

Quando se trabalha com anotações, não existe um conjunto de "anotações corretas" utilizadas para comparação, sendo somente possível medir a confiabilidade do processo, ou seja, quão consistentes foram os pares de anotadores durante o processo (ARSTEIN, 2017; FORT, 2016). A confiabilidade serve como uma condição necessária (mas não suficiente) para a exatidão; se o processo de anotação não é confiável, não é possível esperar que as anotações estejam corretas. Um processo de anotação como um todo é confiável se é reprodutível (ARSTEIN, 2017).

Para verificar a confiabilidade da anotação, é utilizado o IAA, que indica como anotadores individuais se comparam no processo de anotação; valores altos são indicações de que a tarefa está bem definida e da continuidade do trabalho, enquanto valores baixos revelam que o processo deve ser revisto. Ter um valor alto de IAA não é garantia de as anotações estarem corretas, somente significa que os anotadores estão interpretando suas instruções consistentemente do mesmo modo (PUSTEJOVSKY; STUBBS, 2012). Eles podem estar concordando em marcações incorretas e isso pode se dever tanto à questão de o *guideline* não estar claro o suficiente quanto ao tempo de treinamento insuficiente. Apesar de valores de IAA altos durante a etapa de refinamento do *guideline* e/ou anotação em larga escala (real), ainda assim é necessário realizar revisões manuais das anotações.

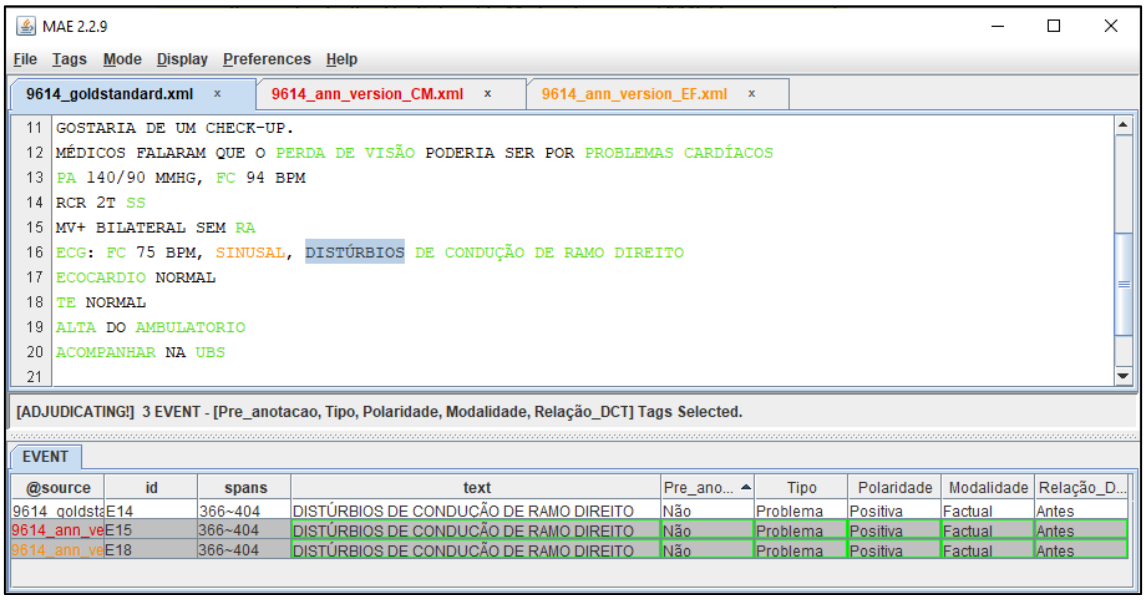
Outra questão extremamente importante é a definição "vaga" de valores altos e baixos de IAA, até porque existe uma relação direta com a dificuldade da tarefa de anotação. Por exemplo, durante estudos sobre o IAA no começo da anotação do TimeBank, foi evidenciado que as anotações das relações tiveram IAAs inferiores em comparação às demais categorias de anotação (EVTs e ETs), sendo a anotação das RTs a mais difícil (KUBLER; ZINSMEISTER, 2015).

Após todo o processo de refinamento do *guideline* e treinamento dos anotadores pelo ciclo iterativo, usualmente contando com diversas reuniões e ajustes do *guideline*, são obtidos valores de IAA confiáveis e estáveis (sem variações bruscas

de valores durante determinado período). Nesse cenário, a próxima etapa é a anotação em larga escala (anotação real), processo que também ocorre por anotação dupla, porém existindo um passo adicional de adjudicação.

A adjudicação é a etapa em que documentos duplamente anotados são corrigidos por um anotador experiente, chamado adjudicador, com o objetivo de criar um documento final (*gold standard*). A correção está ligada somente às *disagreements* (discordâncias) entre os anotadores, com o adjudicador não podendo contestar/anular anotações concordantes entre eles, tentando de alguma forma impor sua anotação individual de adjudicador (FORT, 2016; ROBERTS *et al.*, 2009). Contudo, esse conceito pode ser ampliado para correção do adjudicador de todas as anotações, um raro caso na anotação tradicional (FORT, 2016). Geralmente, é melhor ter adjudicadores que estiveram envolvidos em criar o *guideline* de anotação, pois terão o melhor entendimento do propósito da anotação (PUSTEJOVSKY; STUBBS, 2012). As ferramentas usualmente podem ter interfaces próprias para adjudicação dos textos. Um exemplo envolvendo a ferramenta MAE2 e a anotação de EVTs desta tese é mostrado na Figura 9, com textos de ambos os anotadores selecionados e um terceiro arquivo de texto gerado pela adjudicação.

Figura 9 – Exemplo de adjudicação de um texto da marcação de EVTs na ferramenta de anotação MAE2.



Fonte: O autor (2020).

Como visto neste capítulo, o processo de anotação tem diversas particularidades e influência direta no desempenho dos algoritmos, pois fornece os rótulos para o aprendizado. Existem “boas práticas” na criação de um *corpus*, todas aqui detalhadas.

2.3 APRENDIZADO DE MÁQUINA

Após o processo de anotação, ocorre a utilização de técnicas/algoritmos de aprendizado de máquina para predição. O objetivo é que seu *framework* consiga, por meio dos exemplos aprendidos com os dados provindos do processo de anotação, predizer anotações para novos dados.

Um sistema baseado em aprendizado de máquina é treinado, em vez de explicitamente programado, sendo fornecidos diversos exemplos da tarefa para que sejam achados padrões nos exemplos que permitam ao sistema inferir regras para automatizar a tarefa (CHOLLET, 2017). Essa questão pode ser visualizada na Figura 10.

Figura 10 – Esquema da programação clássica em comparação ao aprendizado de máquina.

Fonte: Adaptado de Chollet (2017).

Uma definição formal de aprendizado de máquina é dada por Alpaydin (2014), como sendo o ato de programar computadores para otimizar um critério de desempenho, usando dados como exemplos ou experiência passada. A experiência refere-se à informação passada que está disponível para aprendizado, usualmente assumindo a forma de dados eletrônicos coletados e destinados à análise (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012). O modelo pode ser preditivo, focando em fazer predições; descritivo, para ganhar conhecimento dos dados; ou ambos (ALPAYDIN, 2014).

Independentemente do caso, a qualidade e quantidade dos dados são cruciais para o sucesso das predições (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012). Os algoritmos de aprendizado de máquina precisam de quantidades significativas de dados, de preferência sem muito ruído, mas, com o aumento do tamanho do conjunto de dados, aumenta o custo computacional, que pode ser um fator limitante no seu conjunto de dados, como, por exemplo, em aplicações envolvendo vídeos (MARSLAND, 2015).

O tipo de aprendizado mais comum é o aprendizado supervisionado, termo originado da visão do rótulo (resposta correta), fornecido por um “instrutor” ou “professor” que mostra/ensina ao sistema de aprendizado da máquina o que fazer. No aprendizado não supervisionado, não existe “instrutor” ou “professor” e o algoritmo deve aprender com os dados sem esse guia (GOODFELLOW; BENGIO; COURVILLE, 2016). Nesta tese, será tratado somente o aprendizado supervisionado.

Para determinar os rótulos, pode ser necessário o envolvimento de *experts* no campo relevante e investimentos significativos de tempo (MARSLAND, 2015). Para o PLN, os rótulos podem ser determinados por meio do processo de anotação (descrito na seção 2.2). Dependendo do domínio em que se esteja trabalhando, a *expertise* é cara, podendo até limitar e/ou impossibilitar projetos. Por exemplo, se há marcações

específicas ao domínio clínico, é desejado que os anotadores tenham *expertise* na área da saúde.

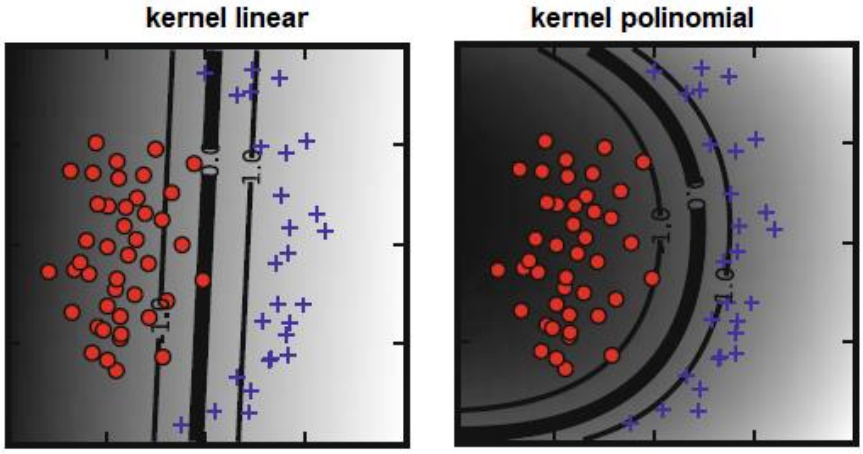
Problemas de aprendizado supervisionado podem ser agrupados em problemas de regressão e classificação. A classificação pode ser definida como um problema de aprendizado supervisionado com rótulos com valores discretos e a regressão, como um problema de aprendizado supervisionado com rótulos com valores contínuos (ZHU; GOLDBERG, 2009). Ainda, dentro do aprendizado de máquina, existe o subcampo “aprendizado profundo”, tendo como objetivo aprender representações de dados dando ênfase ao aprendizado de sucessivas camadas de representações progressivamente mais significativas. Nele, essas representações em camada são quase sempre aprendidas via modelos de redes neurais (CHOLLET, 2017). Neste projeto, o foco foi em aprendizado de máquina tradicional, utilizando o algoritmo *Support Vector Machine* (SVM) (BOSER; GUYON; VAPNIK, 1992), detalhado adiante.

O algoritmo SVM é um classificador com foco em achar bons limites de decisão entre dois conjuntos de pontos que pertencem a diferentes categorias. Um limite de decisão pode ser interpretado como uma linha ou superfície que separa os dados de treinamento em dois espaços, correspondentes às duas categorias de predição (CHOLLET, 2017). Seu treinamento produz uma função discriminante que minimiza o erro de treinamento, enquanto maximiza a margem, separando as classes (AYAT; CHERIET; SUEN, 2005). Ele pertence à categoria de métodos *kernel*, sendo, então, um algoritmo que depende dos dados apenas dos produtos escalares, os quais podem ser substituídos por funções de *kernel*, que computam produtos escalares em espaços de *features* de maior dimensão (BEN-HUR; WESTON, 2010). Nestes, um limite de decisão linear é construído com propriedades especiais que asseguram alta generalização para o algoritmo (CORTES; VAPNIK, 1995).

Conforme mencionado por Ben-Hur e Weston (2010), a escolha do *kernel* depende dos dados. Em muitas das aplicações de bioinformática, a flexibilidade dos *kernels* gaussiano e polinomial leva ao *overfitting* em dados de alta dimensionalidade com pequeno número de exemplos, de modo que o *kernel* linear acaba sendo a melhor escolha. Este não é tão eficaz quando existe uma relação não linear entre as *features*, aspecto verificado na Figura 11, em que os exemplos são mais bem separados pelo *kernel* polinomial (BEN-HUR; WESTON, 2010). No entanto, para casos em que o número de *features* é alto, o mapeamento dos dados para espaços

dimensionais maiores não melhora o resultado; assim, usar o *kernel* linear e somente buscar o valor do parâmetro C é suficiente (HSU; CHANG; LIN, 2003). O parâmetro C é o fator de penalidade, representando classificações incorretas, e diretamente impacta no limite de decisão; valores altos de C indicam penalidade alta aos erros e erros de margem (BEN-HUR; WESTON, 2010).

Figura 11 – Exemplo de aplicação de *kernels* linear e polinomial para o mesmo problema de classificação.



Fonte: Adaptado de Ben-Hur e Weston (2010).

Para extração de RTs entre EVTs e a DCD, denominadas RelTempDCD, o SVM é o algoritmo mais utilizado e apresenta os melhores resultados nos trabalhos analisados durante a revisão (Apêndice A). Para extração de RTs entre menções no texto, diversas abordagens envolveram uso de classificadores SVM especializados, fato constatado durante a revisão. Grande vantagem da utilização desses classificadores de forma especializada é ter maior compreensão sobre os diversos componentes relacionados à extração de RTs.

2.4 PROCESSAMENTO DE LINGUAGEM NATURAL

O PLN é um campo de pesquisa que investiga o uso de computadores para processar ou entender linguagem natural com o propósito de realizar tarefas, sendo um campo interdisciplinar que combina linguística computacional, ciência da computação, ciência cognitiva e IA (DENG; LIU, 2018). Os algoritmos e métodos desenvolvidos por PLN têm como entrada ou produzem como saída dados não estruturados (dados textuais) (GOLDBERG, 2017). Além de textos, dados referentes à “fala” podem ser utilizados com PLN, com a fala sendo considerada uma versão

ruidosa de texto, impondo passos adicionais para eliminação de ruído (DENG; LIU, 2018).

Para desenvolvimento de aplicações de PLN, necessita-se de *corpus* escrito ou falado como entrada, podendo, em casos específicos, existir múltiplos *corpora* associados como entrada (THANAKI, 2017). Vale salientar que usualmente *corpora* utilizados em pesquisa de PLN são anotados, fornecendo exemplos positivos para os algoritmos; sendo assim, o processo de anotação (detalhado na seção 2.2) se torna fundamental para gerar exemplos corretos e na quantidade necessária para o treinamento do algoritmo.

Uma tarefa de PLN pode ser composta por um ou mais níveis de análises linguísticas. Alguns desses tipos são sumarizados a seguir:

a) Fonológico: interpretação dos sons de fala dentro e entre as palavras, sendo crucial para compreensão na língua falada e em sistemas de reconhecimento de voz (FELDMAN, 1999; LIDDY, 2001).
b) Morfológico: análise da composição das palavras, focada na maneira como as palavras são formadas por morfemas (menor fragmento de palavra a carregar significado). Alguns exemplos são raiz, prefixos e sufixos (FELDMAN, 1999; KURDI, 2016).
c) Sintático: análise de palavras em uma sentença e identificação do papel de cada uma e sua relação com as outras. O objetivo final da sintaxe é trazer informações sobre a organização da sentença e dos princípios que governam as combinações e relações de dependências das palavras e sequências de palavras no texto (FELDMAN, 1999; KURDI, 2016).
d) Semântico: foco no significado literal da linguagem, determinando os possíveis significados de palavras e sentenças. Estabelece os possíveis significados de uma frase, concentrando-se nas interações entre significados em nível de palavra na sentença (LIDDY, 2001; KUBLER; ZINSMEISTER, 2015).
e) Pragmático e discurso: foco na questão da organização das informações no texto, questões implícitas, contexto de produção e veiculação do texto (PARDO, 2008).

Os níveis descritos refletem um aumento na unidade de análise, assim como na complexidade e dificuldade enquanto se move de cima para baixo na lista. Quanto

maior é a unidade de análise, menos preciso é o fenômeno apresentado e maiores são a escolha livre e a variabilidade (LIDDY, 1998).

Existe uma grande gama de tarefas de PLN; algumas das principais serão detalhadas, incluindo algumas menções de trabalhos.

Uma tarefa popular em PLN é o Reconhecimento de Entidades Nomeadas (REN), visando a identificar todas as ocorrências de tipos específicos de entidades nomeadas em documentos. Estas (como nomes de pessoas, organizações, localizações) são elementos semânticos básicos de um texto que carregam um tipo de significado limitado e específico (PALSHIKAR, 2013).

Há questões como as resoluções de anáforas e correferências. Nas primeiras, ocorre a identificação de um antecedente textual importante no discurso durante a interpretação de uma expressão; nas segundas, por sua vez, são identificadas menções a uma mesma entidade presente em um texto (COLLOVINI; GOULART; VIEIRA, 2004; FONSECA *et al.*, 2017). Um exemplo de aplicação é apresentado por Fonseca *et al.* (2017), com a resolução de correferências por meio de regras com a criação do modelo CORP.

A sumarização de textos objetiva reduzir a quantidade de texto de entrada em um sumário, de forma a permitir ao leitor descobrir mais rapidamente qual tipo de informação o documento apresenta (JAIN; KULKARNI; SHAH, 2018). Um exemplo de aplicação é citado por Jorge e Pardo (2010), com a sumarização de múltiplos documentos de mesmo tópico para o *corpus* jornalístico CSTNews (ALEIXO; PARDO, 2008).

Outra vertente importante de pesquisa em PLN é a análise de sentimento, também chamada mineração de opinião, área de estudo que analisa a opinião, sentimento, avaliações, elogios, emoções e atitudes de pessoas em relação a entidades como produtos, serviços, organizações e indivíduos (LIU, 2012; ALMEIDA *et al.*, 2018). Um exemplo de aplicação é o trabalho de Almeida *et al.* (2018), com a identificação de emoção em textos curtos com experimentos envolvendo dados de *sites* de notícias *on-line* e notícias de entretenimento extraídas do BuzzFeed.

Em relação ao domínio da saúde, a utilização dos RES tem produzido massivas quantidades de dados textuais sobre pacientes, despertando o interesse da comunidade de pesquisa no investimento de esforço substancial para fazer uso de textos clínicos via PLN (WU *et al.*, 2020). Por exemplo, para doenças crônicas, como textos clínicos são mais frequentes que dados estruturados, essa disponibilidade

textual cria oportunidades para o PLN extrair automaticamente informações relevantes que podem retardar ou prevenir o início de determinada doença (SHEIKHALISHAHI *et al.*, 2019).

A aplicação de técnicas de processamento de textos em documentos escritos por profissionais de saúde em ambiente clínico é denominada PLN clínico, o qual pode fornecer detalhes importantes sobre pacientes, que muitas vezes estão bloqueados em textos não estruturados e dispersos nos RES. É uma área de pesquisa movida pela necessidade de extração de informações precisas, em larga escala e em tempo real de dados provindos de RES, buscando apoio ao cuidado clínico, à prática de saúde pública e às pesquisas (incluindo a biomédica) (VELUPILLAI *et al.*, 2015a).

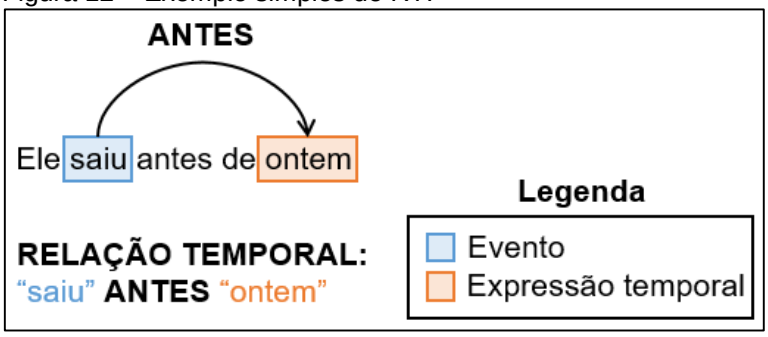
A pesquisa em PLN clínico é semelhante à pesquisa em PLN geral, envolvendo os mesmos tópicos de pesquisa. No entanto, durante uma revisão sobre trabalhos em PLN clínico para doenças crônicas, Sheikhalishahi *et al.* (2019) constataram que a maioria dos trabalhos descritos envolvia classificação de textos ou REN. Uma diferença entre ambas as linhas de pesquisa é a disponibilidade de *corpora* anotados. Para o domínio clínico, é essencial garantir que o *corpus* esteja em conformidade com regulamentos éticos e não revele nenhuma informação identificável do paciente (seja deidentificado), antes de torná-lo disponível (VELUPILLAI *et al.*, 2015a). Principalmente devido a essa questão, acaba sendo comum a não disponibilidade de textos clínicos para pesquisadores externos sem projetos colaborativos (DALIANIS, 2018; WANG *et al.*, 2018). Isso impacta a pesquisa na área clínica, limitando-a a trabalhos envolvendo *corpora* restritos ou disponíveis apenas por meio de *shared tasks*.

2.5 PADRÕES DE ANOTAÇÃO TEMPORAL

Uma das vertentes da pesquisa em PLN é a extração de relações entre elementos previamente definidos, que comumente envolvem entidades nomeadas. Alguns exemplos para o domínio da saúde seriam a extração de reações adversas a medicamento e da interação entre medicamentos. Um trabalho sobre reações adversas a medicamentos é o de Negi *et al.* (2019), classificando a relação entre o medicamento e a condição médica como causal, verificando se esta é um evento adverso daquele.

Usualmente, a extração de relação envolve a classificação de pares de menções em categorias predefinidas. Um dos casos específicos é a extração de RTs, em que menções, tanto de EVTs (entidades com “aspecto temporal”) quanto de ETs (datas, durações etc.), são relacionadas com o objetivo de trazer ordem entre elas (criar uma *timeline*). Um exemplo simples do conceito de RT é representado na Figura 12, trazendo uma relação entre o EVT “saiu”, uma ação, e a ET “ontem”, uma data. Devido ao contexto da frase, pode-se inferir que a ação de sair ocorreu “antes” da data referente ao dia de “ontem”.

Figura 12 – Exemplo simples de RT.



Fonte: O autor (2020).

No contexto clínico, EVTs são menções clinicamente relevantes que ocorrem nos textos (como tratamentos, problemas e testes), ETs aludem a menções temporais (como durações e datas) e RTs são as relações entre EVTs e ETs (MOHARASAN; HO, 2019). Os EVTs e ETs são especialmente importantes, porque as relações inferidas têm ligação direta com as marcações desses elementos.

Quando se mencionam RTs, existem dois conceitos fortemente associados: padrões de anotação e *shared tasks* (tarefas compartilhadas). Devido à escassez de *corpora* anotados publicamente disponíveis, são realizadas competições na área de informática, objetivando que diferentes sistemas sejam testados com o mesmo *corpus* (BETHARD *et al.*, 2016; VIANI *et al.*, 2019). Essas competições, em que pesquisadores se reúnem para trabalhar em prol de um problema específico de interesse, são chamadas *shared tasks*.

Pelas dificuldades já mencionadas concernentes à criação de *corpora* envolvendo RTs, *shared tasks* se tornam valiosas fontes de dados. Felizmente, várias delas foram organizadas para extração de RTs, fornecendo dados que a comunidade de pesquisa pode usar para desenvolver suas técnicas de extração temporal e

comparar seus resultados. As principais *shared tasks* contemplando extração de RTs são sinalizadas no Quadro 1.

Quadro 1 – Ano, domínio e padrão de anotação das *shared tasks* sobre extração de RTs.

Shared task	**Ano**	**Domínio**	**Padrão de anotação**
TempEval-1	2007	Jornalístico	TimeML
TempEval- 2	2010	Jornalístico	TimeML
i2b2 2012	2012	Clínico	Adaptação TimeML
TempEval- 3	2013	Jornalístico	TimeML
Clinical TempEval 2015	2015	Clínico	THYME-TimeML (adaptação TimeML)
Clinical TempEval 2016	2016	Clínico	THYME-TimeML (adaptação TimeML)
Clinical TempEval 2017	2017	Clínico	THYME-TimeML (adaptação TimeML)

Fonte: O autor (2020).

A partir do Quadro 1, é possível identificar as *shared tasks* que tiveram foco em extração de RTs, com informações sobre ano, domínio e padrão de anotação. *Shared tasks* para extração de RT no domínio jornalístico ocorrem desde 2007, com o TempEval-1 (VERHAGEN *et al.*, 2007), seguido do TempEval-2 (VERHAGEN *et al.*, 2010) e TempEval-3 (UZZAMAN *et al.*, 2013). Essas tarefas consistiam na identificação de EVTS, ETs e RTs em *corpora* jornalísticos anotados de acordo com o padrão ISO-TimeML (SAURÍ *et al.*, 2006; PUSTEJOVSKY *et al.*, 2010), desenvolvido por pesquisadores da comunidade de PLN com o objetivo de transformar informações temporais em textos livres em dados estruturados (CHENG *et al.*, 2013). O padrão ISO-TimeML é um esquema de anotação especificamente criado para anotação de EVTs, ETs e suas relações no texto (PUSTEJOVSKY *et al.*, 2010).

Desde 2012, têm sido propostos estudos de diferentes grupos para a mesma tarefa de extração de RTs para o domínio clínico pelas *shared tasks*. Ocorreram quatro cujo foco era a extração de RTs: *Informatics for Integrating Biology and Bedside* (i2b2) 2012 (SUN; RUMSHISKY; UZUNER, 2013a), Clinical TempEval 2015 (BETHARD *et al.*, 2015), Clinical TempEval 2016 (BETHARD *et al.*, 2016) e Clinical TempEval 2017 (BETHARD *et al.*, 2017). Contudo, o formato ISO-TimeML acabou sendo adaptado para o domínio clínico, sendo expandido para anotação do *Temporal History of Your Medical Events* (THYME) *corpus*, nomeado THYME-TimeML (STYLER *et al.*, 2014a). Esse *corpus* foi utilizado nas *shared tasks*: Clinical TempEval 2015, Clinical TempEval 2016 e Clinical TempEval 2017, tendo sido anotado a partir de notas clínicas e notas patológicas de pacientes da Mayo Clinic (BETHARD *et al.*, 2016).

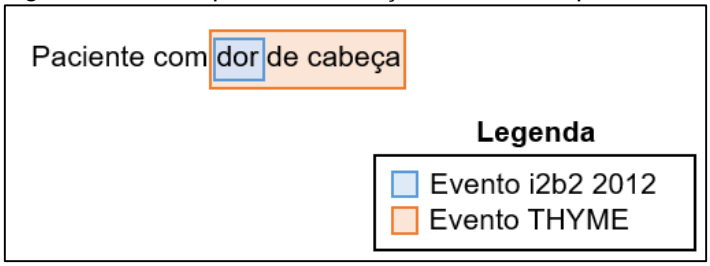
Para a anotação do i2b2 2012, o *guideline* de anotação também foi adaptado do ISO-TimeML para se adequar a dados clínicos, utilizando uma versão intermediária do *guideline* do projeto do THYME *corpus* como ponto de partida. O i2b2 2012 *corpus* foi anotado a partir de sumários de alta do Partners Health and Beth Israel Deaconess Medical Center (SUN; RUMSHISKY; UZUNER, 2013a).

Uma das etapas desta tese consiste na anotação de EVTs, ETs e suas consequentes RTs. Por esse motivo, a seguir serão detalhadas as anotações desses tipos de menção para i2b2 2012 e THYME-TimeML, trazendo questões importantes sobre o TimeML. Ressalta-se que, por questão de nomenclatura e devido a traduções aproximadas de termos, serão mantidos os padrões de anotação em inglês.

2.5.1 Eventos

O ISO-TimeML considera EVT um termo de cobertura para situações que acontecem, podendo ser pontuais ou durar determinado período. Predicados que descrevem estados ou circunstâncias em que algo é obtido ou permanece verdadeiro também são considerados EVTs (PUSTEJOVSKY *et al.*, 2003). Já as anotações do THYME e i2b2 2012 entendem os EVTs como quaisquer menções importantes na *timeline* do paciente (SUN; RUMSHISKY; UZUNER, 2013b; STYLER *et al.*, 2014a). O THYME considera apenas uma palavra para cada EVT (núcleo sintático), enquanto o i2b2 2012, diversas palavras. Essa questão é um *trade-off* entre detalhamento do EVT e complexidade de sua extração automática. Quanto mais longa e detalhada é uma menção de EVT, mais difícil é para um algoritmo identificar corretamente toda a delimitação, porém certos detalhamentos são essenciais de ser mantidos. Apesar de o THYME anotar somente uma palavra para cada EVT, eles são mais para frente expandidos automaticamente para capturar todo o conteúdo. Um exemplo de marcação do mesmo EVT para i2b2 2012 e THYME é mostrado na Figura 13.

Figura 13 – Exemplo da delimitação de um EVT para i2b2 2012 e THYME.



Fonte: O autor (2020).

Um EVT não é somente constituído por sua delimitação de texto, mas existem diversos atributos (características importantes) que acompanham a menção. Para o i2b2 2012, há três atributos: (i) *Type*, que categoriza os EVTs em grupos mais específicos; (ii) *Polarity*, que diferencia EVTs entre positivos e negativos; (iii) *Modality*, que diferencia EVTs que realmente ocorreram daqueles propostos, condicionais ou possíveis (SUN; RUMSHISKY; UZUNER, 2013a). Um resumo dos atributos e categorias para i2b2 2012 e THYME é mostrado no Quadro 2. Vale salientar que cada atributo só pode ter uma das possíveis marcações.

Quadro 2 – Atributos e categorias de EVTs para os padrões de anotação i2b2 2012 e THYME.

Padrão de anotação	**Atributo**	**Categorias**
i2b2 2012	*Type*	*Problem*, *Test*, *Treatment*, *Evidence*, *Clinical Department* e *Occurrence*
	Polarity	*Positive* e *Negative*
	Modality	*Factual*, *Hypothetical*, *Hedged* e *Conditional*
THYME	*Type*	*NA* (não disponível), *Aspectual* e *Evidential*
	Polarity	*Positive* e *Negative*
	Degree	*N/A*, *Most* e *Little*
	Modality	*Actual*, *Hypothetical*, *Hedged* e *Generic*
	DocTimeRel	*Before*, *Overlap*, *Before/Overlap* e *After*

Fonte: O autor (2020).

Para as anotações do THYME, os EVTs têm cinco atributos: (i) *Type*, que diferencia EVTs com informações aspectuais daqueles clínicos tradicionais; (ii) *Polarity*, que diferencia EVTs entre positivos e negativos; (iii) *Degree*, que adiciona categorias de quantificação do atributo anterior; (iv) *Modality*, que divide o EVT em categorias específicas, como factual e hipotético; (v) DocTimeRel, RT entre o EVT e a DCD (STYLER *et al.*, 2014a).

Uma das diferenças entre os padrões está no atributo *Type*. Fazendo um breve resumo das categorias de *Type* para o i2b2 2012, tem-se: *Problem* (como dor de cabeça, infarto e dispneia), *Test* (como cateterismo cardíaco e creatinina), *Treatment* (como angioplastia e sinvastatina 40 mg), *Evidence* (como relata, nega e mostrou), *Clinical Department* (como ambulatório de cardiologia e emergência) e *Occurrence* (como aumento, transferido e melhora) (SUN; RUMSHISKY; UZUNER, 2013b). Em comparação ao THYME, o i2b2 2012 tem categorias interessantes, pois subdividem

os EVTs em subcategorias de interesse, que possuem suas próprias características e regras de marcação.

Ambos têm o mesmo conceito de *Polarity*, que é um atributo voltado à ocorrência de um EVT. Um EVT com polaridade positiva indica que aconteceu em algum momento, está acontecendo no momento ou acontecerá. Uma marcação negativa indica o contrário. A diferença é que o THYME apresenta o atributo *Degree*, que consegue dar mais ênfase a quão positivo ou negativo um EVT é.

O THYME considera a DocTimeRel um atributo do próprio EVT, diferentemente do TimeML e i2b2 2012, em que as RTs com a DCD são consideradas entre dois elementos, sendo um deles a DCD. Nesses esquemas em que usualmente todo EVT tem relação direta com DCD, se torna interessante marcar como um atributo do próprio EVT, em vez de marcar relações separadas.

2.5.2 Expressões temporais

ETs são menções que envolvem representações relacionadas ao tempo de alguma forma. No ISO-TimeML, as ETs, nomeadas TIMEX3, podem ser de quatro tipos/categorias (SAURÍ *et al.*, 2006). De forma geral, podem ser: (i) uma expressão de data (*Date*), como 12/08/2008; (ii) uma hora específica do dia (*Time*), como 12:24; (iii) uma duração (*Duration*), como por 30 minutos; (iv) uma periodicidade de EVTs (*Set*), como toda semana (STRÖTGEN; GERTZ, 2016).

Apesar de existir certa similaridade entre ET no domínio geral e no domínio clínico, há algumas questões, como frequência de uso de medicamentos e menções de datas relacionadas a operatório, específicas do domínio. Os atributos e categorias de ETs para os padrões de anotação i2b2 2012 e THYME são mostrados no Quadro 3.

Quadro 3 – Atributos e categorias de ETs para os padrões de anotação i2b2 2012 e THYME.

Padrão de anotação	**Atributo**	**Categorias**
I2b2 2012	*Type*	*Date*, *Time*, *Duration* e *Frequency*
	Value	Valor normalizado conforme a ISO 8601
	Mod	*NA* (não disponível), *Approx*, *More*, *Less*, *Start*, *End* e *Middle*
THYME	*Class*	*Date*, *Time*, *Duration*, *Quantifier*, *Prepostexp* e *Set*

Fonte: O autor (2020).

O atributo *Type* do i2b2 2012 é similar ao do ISO-TimeML, com a categoria *Frequency* sendo igual ao *Set*. As diferenças do THYME em relação ao ISO-TimeML

e i2b2 2012 são as categorias adicionais específicas *Quantifier* (usada para marcar menções como duas vezes e quatro vezes) e *Prepostexp* (usada para marcar menções como pré-operatório, intraoperatório e pós-operatório) (STYLER *et al.*, 2014a). Em textos clínicos em que menções de tempo envolvendo pré, pós e intraoperatório são amplamente usadas, pode ser conveniente ter essa categoria adicional.

No i2b2 2012, houve a normalização do valor de TIMEX3, indicado pelo atributo *Value*. A normalização foi baseada na ISO 8601[1], que padroniza a quantificação da TIMEX3. Por exemplo, a normalização de menções de *Date* transforma a ET no padrão [YYYY-MM-DD], em que Y indica ano, M, mês e D, dia. O objetivo é transformar a menção em formato livre em um padrão estruturado que pode futuramente ser utilizado para sequenciamento. Um exemplo de normalização seria a transformação de 12 de janeiro de 1985 em 1985-01-12. No THYME, não houve a normalização das ETs.

Similarmente ao TimeML, o i2b2 2012 utiliza o atributo *Mod (modifier)*, que indica se um valor temporal é exato (marcação NA) ou não (demais marcações). No caso de não ser exato, existem diversas possibilidades de marcação para aproximar a menção: *More* (como mais de dois meses), *Approx* (como há cerca de dois meses), *Start* (como começo de setembro), *End* (como final de setembro) e *Middle* (como meio de setembro). Na maioria delas, o valor mais utilizado é o NA, tanto que é o valor padrão.

De forma geral, o padrão i2b2 2012 é mais robusto, até pela questão da normalização, porém as categorias do THYME podem ser interessantes em determinados conjuntos de texto. Uma diferença relevante das TIMEX3s do domínio geral para domínio clínico é relativa a marcações do tipo *Set*, pois existem menções de frequência de uso de medicamento que podem ser extremamente complexas. Um exemplo seriam frequências de uso do mesmo medicamento, porém com diferentes números de comprimidos conforme o dia da semana, tais como: 2 cp seg, quarta, sexta e 1 cp nos demais.

---

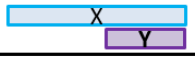
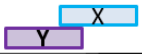
[1] Disponível em: https://www.iso.org/iso-8601-date-and-time-format.html. Acesso em: 15 nov. 2020.

2.5.3 Representação temporal

Um marco para extração de RTs foi a álgebra baseada em intervalos, proposta por Allen, em 1983. Diversos estudos adotaram sua representação, que rapidamente se tornou um padrão para modelagem temporal. Ela sustenta que, dados dois pontos no tempo ou intervalos no tempo, qualquer relação entre eles pode ser representada por sete tipos: *Before*, *Meet*, *Overlap*, *During*, *Starts*, *Finishes* e *Equal* (ALLEN, 1983). Considerando as relações inversas (*Equal* não possui relação inversa), existem 13 relações possíveis. As relações de Allen são mostradas na Figura no Quadro 4 (coluna Representação de Allen).

Em termos de relações, as maiores diferenças entre a representação de Allen (1983) e o padrão ISO-TimeML é que este não aborda relações do tipo *Overlap*, além de a relação *Equal* na representação de Allen se dar por quatro relações no ISO-TimeML: *Identify*, *Simultaneous*, *Hold* e *Held_By* (VERHAGEN, 2005). Essa mudança descarta relações do tipo *Overlap*, que são genéricas, pois, se duas menções têm qualquer tipo de sobreposição, já são consideradas *Overlap*. Relações como *Identify*, *Simultaneous*, *Hold* e *Held_By* são casos de sobreposição bem definidos. As relações do ISO-TimeML são mostradas no Quadro 4 (coluna ISO-TimeML).

Quadro 4 – Tipos de RT presentes na representação de Allen; padrão de anotação ISO-TimeML; i2b2 2012 e Clinical TempEval *corpora*; relações em cinza anotadas no *corpus*, porém não usadas durante a *shared task*.

REPRESENTAÇÃO	Representação de Allen	ISO-TimeML	I2b2 TLINK	I2b2 DocTimeRel	Clinical TempEval TLINK	Clinical TempEval DocTimeRel
1. 	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y
2. 	X MEETS Y	X I_BEFORE Y	X BEF_OVERLAP Y	X BEF_OVERLAP Y		X BEF_OVERLAP Y
3. 	X OVERLAPS Y		X OVERLAPS Y	X OVERLAPS Y	X OVERLAPS Y	X OVERLAPS Y
4. 	X DURING Y	X IS_INCLUDED Y	X DURING Y	X DURING Y		
5. 	X STARTS Y	X BEGINS Y			X BEGINS_ON Y	
6. 	X FINISHES Y	X ENDS Y				
7. 	X FINISHED_BY Y	X ENDS_BY Y	X ENDS_BY Y	X ENDS_BY Y	X ENDS_ON Y	
8. 	X AFTER Y	X AFTER Y	X AFTER Y	X AFTER Y		X AFTER Y
9. 	X MEET_BY Y	X I_AFTER Y				
10. 	X OVERLAPPED_BY Y					
11. 	X CONTAINS Y	X INCLUDES Y			X CONTAINS Y	
12. 	X STARTED_BY Y	X BEGINS_BY Y	X BEGINS_BY Y	X BEGINS_BY Y		
13. 	X EQUALS Y	X IDENTIFY Y X SIMULT. Y X HOLD Y X HELD_BY Y	X SIMULT. Y	X SIMULT. Y		

Fonte: O autor (2020).

Devido a particularidades da anotação do domínio clínico e para melhor representar a informação clínica, algumas adaptações foram feitas para o processo de anotação e o ISO-TimeML. A partir de agora, serão detalhadas algumas dessas adaptações para o i2b2 2012 e Clinical TempEval *corpora*. Todas as *shared tasks* do Clinical TempEval são baseadas no THYME *corpus*. Esses *corpora* são a fonte primária de pesquisas para extração de RTs em textos clínicos.

Para o i2b2 2012 *corpus*, as RTs anotadas foram: *Before*, *After*, *Overlap*, *Simultaneous*, *Before/Overlap*, *During*, *Begun_By* e *Ended_By* (SUN; RUMSHISKY; UZUNER, 2013b). Para a *shared task* relacionada, somente três tipos de relação foram utilizados (*Before*, *Overlap* e *After*). Houve um processo de simplificação devido aos baixos valores de IAA e baixo número de anotações para determinadas relações, ocorrendo a fusão de determinados tipos (como *Simultaneous*, *Overlap* e *During* em relações do tipo *Overlap*) (SUN; RUMSHISKY; UZUNER, 2013a). Os tipos de relação são detalhados no Quadro 4 (colunas i2b2 TLINK e i2b2 DocTimeRel). As RTs representadas na cor preta foram usadas durante a *shared task* (como *Before*) e as

em cinza foram fundidas (como *During*). A diferença no número de anotações para certos tipos de relação é mostrada por Sun, Rumshisky e Uzuner (2013b), exemplificada por 2,7% de anotações do tipo *Ended_By,* enquanto 66,6% eram anotações de *Overlap* e *Simultaneous*.

No Quadro 4, duas colunas representam as RTs do i2b2 2012, por essas RTs terem sido separadas em dois tipos distintos: entre menções (EVTs ou ETs) no texto, chamadas TLINKs (coluna i2b2 TLINK) e entre EVTs e data de sessões (Sectime), chamadas DocTimeRel (coluna i2b2 DocTimeRel). Datas de sessões são tipos de ETs relacionados à data de admissão ou de alta, dependendo de onde o evento estava localizado no texto; podem ser consideradas um tipo de DCD e, portanto, são denominadas DocTimeRel. O *corpus* utilizado para a *shared task* do i2b2 2012 foi composto por 310 sumários de alta anotados, com média de 86,6 EVTs, 12,4 ETs e 176 TLINKs por documento. Vale ressaltar o alto número de marcações de TLINKs no i2b2 2012, devido à DocTimeRel não considerar atributos de EVTs, mas, sim, TLINKs.

Na anotação do THYME *corpus*, cinco tipos diferentes de relação foram usados: *Before*, *Overlap*, *Begins_On*, *Ends_On* e *Contains* (STYLER *et al.*, 2014a). Diferentemente do *corpus* i2b2 2012, ele foi anotado com o conceito de *narrative container*, proposto por Pustejovsky e Stubbs (2011), devido à importância de um esquema de anotação que resultasse em máxima informação anotada, não dependendo de modelos muito difíceis de aplicar.

A escolha do uso de *narrative containers*, segundo Styler *et al.* (2014a), advém da dificuldade de captar todas as RTs possíveis e do aumento da discordância de quando os anotadores tentam fazê-lo. Ao usar essa opção, sempre que possível, as ETs e EVTs são conectados a um *narrative container* (EVT ou ET-âncora), que define seu intervalo temporal. Vários EVTs e ETs podem ser conectados à mesma âncora, que os contêm (linha *Contains* do Quadro 4). EVTs e ETs que estão no mesmo *narrative container* podem ser relacionados, como um único elemento, com outros (KHALIFA; VELUPILLAI; MEYSTRE, 2016). A maior vantagem dessa anotação é a redução na quantidade de anotações necessárias (KHALIFA; VELUPILLAI; MEYSTRE, 2016). Pelo uso da abordagem de *narrative containers* e pela eliminação de relações inversas confusas (por exemplo, *During* e *After*), o *agreement* para anotação de RTs melhorou em relação às anotações do i2b2 2012 (STYLER *et al.*, 2014a).

Nas *shared tasks* Clinical TempEval 2015, 2016 e 2017, baseadas no THYME *corpus*, havia dois tipos de RT. Existem duas colunas no Quadro 4 para representação do Clinical TempEval, pela separação em duas extrações de RTs (assim como no i2b2 2012). As relações foram: entre menções (EVTs ou ETs) no texto, chamadas TLINKs (coluna Clinical TempEval TLINK); e entre EVTs e DCD (considerada um atributo do EVT), chamada DocTimeRel (coluna Clinical TempEval DocTimeRel). Cada EVT tinha uma marcação de DocTimeRel com um dos seguintes tipos: *Before*, *After*, *Overlap*, *Before/Overlap* ou *After*. *Before/Overlap* representa que o evento ocorreu no passado e ainda ocorre durante a DCD, como uma doença crônica, por exemplo, que existe antes da criação do documento clínico e continua a existir durante sua escrita. Os TLINKs usados durante as *shared tasks* do Clinical TempEval (coluna Clinical TempEval TLINK) foram simplificados para considerar somente relações do tipo *Contains*. Na verdade, essas relações foram mais frequentemente anotadas (66% dos TLINKs anotados) no *corpus* THYME, seguidas de *Overlap*, com 14,6% (STYLER *et al.*, 2014a).

Em relação ao *corpus* das *shared tasks* do Clinical TempEval, a edição de 2015 era composta de 440 documentos, com média de 136,05 EVTs, 13,43 ETs e 37,43 TLINKs por documento. A edição de 2016 tinha 591 documentos, com média de 133,42 EVTs, 13,30 ETs e 39,33 TLINKs por documento. A edição de 2017 teve como objetivo a extração de relações entre domínios distintos, com diferentes tipos de texto para treinamento e teste. Na fase 1 (adaptação de domínio não supervisionada), a avaliação foi realizada em documentos de câncer de cérebro, sendo fornecidas para treinamento notas de câncer de cólon. Na fase 2 (adaptação de domínio supervisionada), uma pequena porção de documentos anotados de câncer de cérebro foi fornecida para treinamento. De acordo com Bethard *et al.* (2017), a *shared task* foi muito mais desafiadora que as anteriores (edições de 2015 e 2016), comprovado pela queda dos resultados em comparação à edição de 2016. A edição de 2017 continha 769 documentos, com média de 120,83 EVTs, 12,70 ETs e 33,28 TLINKs por documento.

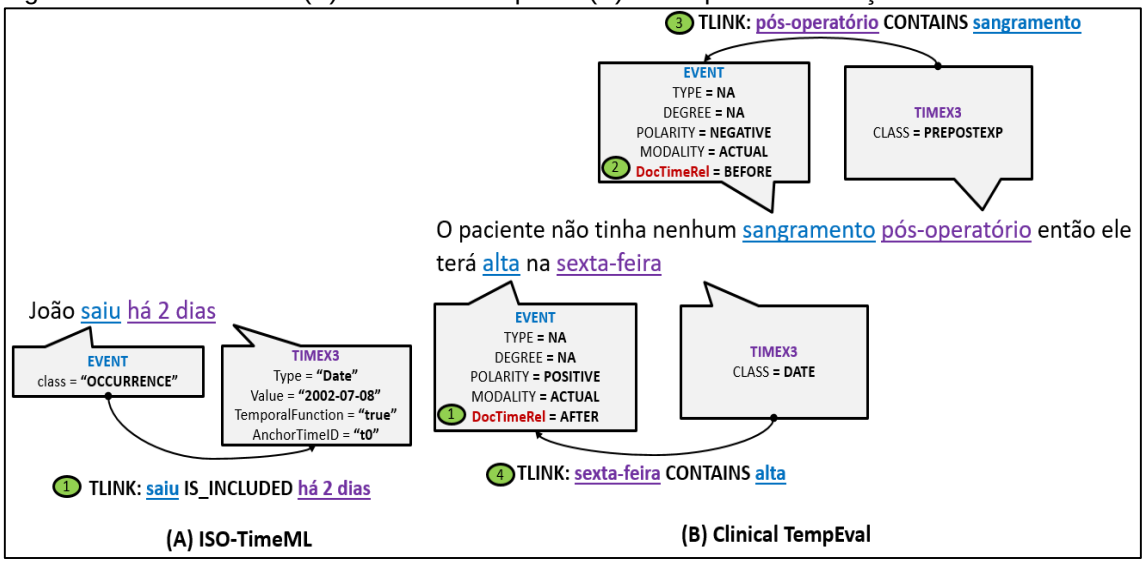
2.5.4 Exemplos de anotação temporal

Para exemplificar as diferenças de esquemas de anotação para domínio geral e clínico, são apresentados dois exemplos para discussão das características e

diferenças. A Figura 14A representa uma RT no padrão ISO-TimeML de uma frase simples “João saiu há 2 dias”, que é composta por um EVT, uma ET e uma RT. Esse exemplo foi baseado em um dos apresentados por Saurí *et al.* (2006).

A ET “há 2 dias” até relacionada à DCD (representando “T0”), mostrada pelo atributo *AnchorTimeID*. O termo “há” só faz sentido considerando a data em que o documento foi escrito (DCD). Como o valor normalizado pela ISO 8601 da DCD é 2002-07-10, a ET foi normalizada para 2002-07-8.

Figura 14 – ISO-TimeML (A) e Clinical TempEval (B): exemplos de anotações de RTs.



Fonte: O autor (2020).

O verbo “saiu” é um EVT da *Class Occurence*, que inclui todos os EVTs que não se enquadram nas demais categorias. Em uma marcação convencional pelo TimeML, cada EVT deveria ser conectado a pelo menos uma *tag MakeInstance*, que cria a “realização” do EVT e foi desenvolvida para resolver casos em que múltiplas instâncias estão relacionadas ao mesmo EVT (PUSTEJOVSKY *et al.*, 2005). A *tag MakeInstance* seria conectada à ET, neste caso. Para simplificar a explicação, a *tag* é ignorada e o EVT é diretamente conectado à ET.

Como “há 2 dias” se refere a um dia inteiro e “saiu” é um EVT que ocorreu em algum momento durante esse período, equivalente a um dia, pode-se inferir que o EVT “saiu” está totalmente contido (relação *Is_Included*) em um “dia inteiro”, que ocorreu faz dois dias. Em adição a essas *tags*, poderia ser marcada uma *tag* chamada *Signal*, que envolve palavras (como antes, depois, enquanto e em) que explicitamente denotam relações (PUSTEJOVSKY *et al.*, 2005). Nessa sentença, não existe

nenhuma *tag Signal*, mas, em uma sentença como “João saiu 2 dias antes do ataque”, o vocábulo “antes” seria marcado como um *Signal*. Detalhes adicionais sobre *MakeInstance*, *Signals* e exemplos completos podem ser encontrados em Saurí *et al.* (2016) e Pustejovsky *et al.* (2005).

Com o exemplo anterior, é possível ver as marcações que poderiam ser feitas em uma frase simples do domínio geral com ISO-TimeML. Para ilustrar as diferenças entre os domínios, foi adaptado um exemplo do THYME-TimeML (Figura 14B), mostrado por Bethard *et al.* (2016).

O THYME-TimeML não envolve *tags* como *Signals* e *MakeInstance*, pois são simplificadas. No *corpus* THYME, um EVT é qualquer menção considerada relevante para a construção de uma *timeline* clínica, o que envolve vários tipos de menção, como problemas médicos, tratamentos (medicamentos e procedimentos) e exames (diagnóstico por imagem ou exames físico e visual) (STYLER *et al.*, 2014a). Verbos como “negado”, “alta”, “continuado” e “mostrou” também são marcados como EVTs, mas a anotação é menos focada em verbos do que no ISO-TimeML. No exemplo, tanto “sangramento” quanto “alta” são marcados como EVTs.

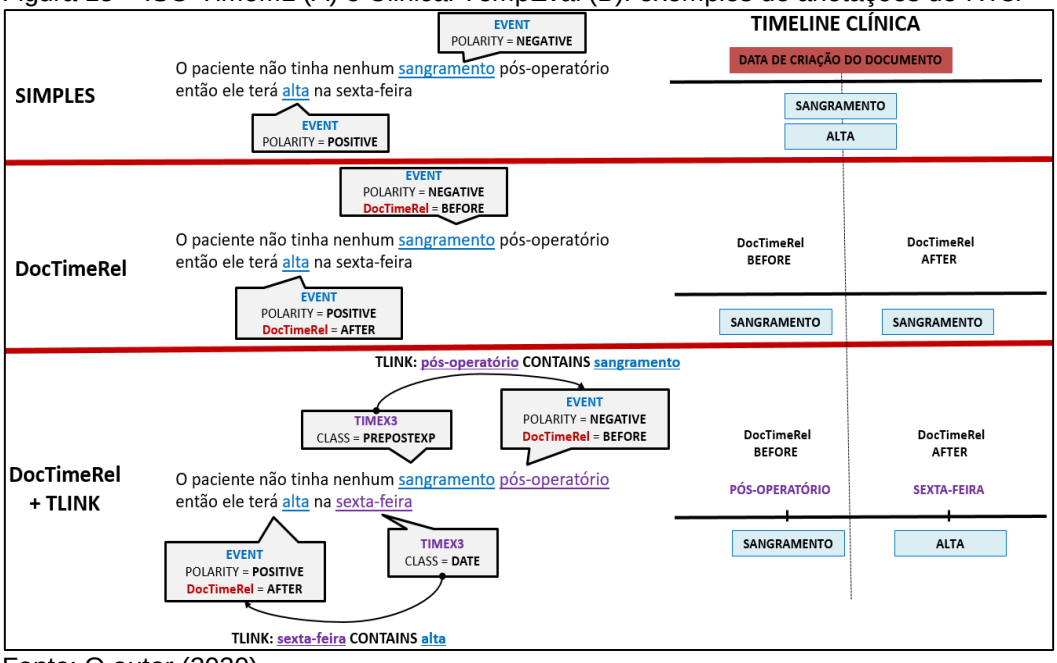
Para ambos os EVTs (sangramento e alta), o atributo *Type* é marcado como NA, devido a ser o valor padrão. Na prática, *Degree* era usado para inferir que algo é substancial ou ligeiramente verdadeiro (dor leve) (STYLER *et al.*, 2014a). Para ambos os EVTs, o atributo *Degree* é marcado como NA. No exemplo, para sangramento, o atributo *Polarity* foi marcado como *Negative*, porque o paciente não teve nenhum sangramento, enquanto o evento “alta” foi marcado como *Positive*, porque o paciente teria “alta” na “sexta-feira”. Para ambos, o atributo *Modality* foi marcado como *Actual*, pois representam EVTS que aconteceram no passado ou que irão acontecer. Se houvesse a uma menção à alta em uma frase, como “possível alta na sexta-feira”, seu atributo *Modality* seria alterado para *Hyphothetical*. A DocTimeRel é, como antes, *Before* para sangramento, porque é inferido pela construção textual que o sangramento (neste, ausência dele) ocorreu antes da DCD. Para o evento “alta”, a DocTimeRel foi marcada como *After*, devido ao uso do futuro simples em uma construção de frase em voz passiva.

Na Figura 14B, há duas ETs: pós-operatório e sexta-feira. É notável que, ao contrário do ISO-TimeML, não há normalização de valor. A menção “sexta-feira” é um típico caso de ET do tipo *Date*. Como informado, a menção “pós-operatório” é considerada uma ET do tipo *Prepostexp*.

Em relação aos TLINKs, nesse exemplo, há duas relações marcadas como *Contains*. Como o exemplo está relacionado à marcação utilizada para as *shared tasks* do Clinical TempEval, a relação *Contains* é o único tipo de relação entre menções (EVTs ou ETs). Como o pós-operatório não apresentou sangramento, isso indica que o sangramento (ausência dele) estava totalmente contido no pós-operatório. Da mesma forma, pode-se inferir que, como o paciente terá alta na sexta-feira, alta está totalmente contida (relação do tipo *Contains*) em um dia inteiro, que é sexta-feira.

O benefício de extrair tal temporalidade é mostrado na Figura 15. Usando uma abordagem “simples”, de apenas conectar todos os documentos à sua DCD, não se pode inferir qualquer ordem. Assim, nesse cenário, tanto o sangramento quanto a alta ocorreram ao mesmo tempo, o que não é verdade. Ao adicionar mais informações à anotação, com a DocTimeRel, pode-se diferenciar o sangramento, que é um EVT passado, da alta, que é um EVT futuro. O problema de usar apenas a DocTimeRel é a vinculação aos tipos de relação: no Clinical TempEval, existem apenas quatro categorias de DocTimeRel (*Before*, *After*, *Before/Overlap* e *Overlap*) em que EVTs podem ser divididos; além disso, relações do tipo DocTimeRel são muito genéricas para certos estudos de extração de RTs, uma vez que categorias como *Before* (antes da DCD) são muito extensas, porque não se referem a determinado período, mas a algo amplo.

Figura 15 – ISO-TimeML (A) e Clinical TempEval (B): exemplos de anotações de RTs.



Fonte: O autor (2020).

Adicionar TLINKs, como mostrado na Figura 15 (linha DocTimeRel), para ancorar EVTs a períodos específicos representados por ETs, melhora a representação da linha do tempo. Por exemplo, alta, além de pertencer ao amplo período *After* (após a DCD), que pode ser um dia após ou dez anos depois, pertence a um período dentro de um único dia (sexta-feira). Como nem todo EVT ou ET tem TLINKs associados, o uso de DocTimeRel permite algum tipo de ordenação entre EVTs, mas os TLINKs acabam fornecendo uma representação mais detalhada.

2.6 MAPEAMENTO ENTRE O REFERENCIAL TEÓRICO E OS ENCAMINHAMENTOS METODOLÓGICOS

Nesta seção, são apresentados dois quadros (Quadro 5 e Quadro 6) exibindo os mapeamentos entre os conceitos e suas relações com os encaminhamentos metodológicos, descritos adiante. Nesta tese, os encaminhamentos metodológicos foram divididos em dois capítulos, um deles referente ao processo de anotação (capítulo 0) e outro, à extração de RTs (capítulo 0); cada um desses possui seu próprio quadro.

O Quadro 5 apresenta o mapeamento em relação ao capítulo 0, referente à anotação, envolvendo os encaminhamentos metodológicos relacionados à criação dos *guidelines*, pré-processamento e anotação.

Quadro 5 – Mapeamento entre o referencial teórico e os encaminhamentos metodológicos para o processo de anotação.

Referências	**Conceito**	**Contribuição**	**Encaminhamento metodológico**
Pearce *et al.* (2016), Cameron e Turtle-Song (2002), Lenert (2016)	Formato SOAP	Estruturação do padrão SOAP, usado para construção de textos ambulatoriais.	Criação dos *guidelines*
Dalianis (2018), Meystre *et al.* (2008), Leaman, Khare e Lu (2015)	Características dos textos clínicos	Informações sobre as dificuldades que podem ser encontradas nos textos clínicos.	Criação dos *guidelines* e pré-processamento
Pustejovsky e Stubbs (2012), Roberts *et al.* (2009)	Criação dos *guidelines*	Boas práticas na criação de um *guideline* (passo a passo).	Criação dos *guidelines*
	Processo de anotação	Boas práticas no processo de anotação e adjudicação de textos.	Processo de anotação dos textos
Pustejovsky e Stubbs (2012), Arstein (2017)	Avaliação da anotação	Questões importantes sobre o IAA sobre avaliação da anotação.	Criação dos *guidelines* e processo de

			anotação dos textos
Styler *et al.* (2014a), Sun, Rumshisky e Uzuner (2013a)	Padrões de anotação de EVTs	Definições-base sobre anotações de EVTs.	Criação dos *guidelines*
Saurí *et al.* (2006), Styler *et al.* (2014a), Sun, Rumshisky e Uzuner (2013a)	Padrões de anotação de ETs	Definições-base sobre anotações de ETs.	Criação dos *guidelines*
Allen (1983), Saurí *et al.* (2006), Styler *et al.* (2014a), Sun, Rumshisky e Uzuner (2013a)	Padrões de anotação de RTs	Definições-base sobre anotações de RTs.	Criação dos *guidelines*

Fonte: O autor (2020).

O Quadro 6 apresenta o mapeamento em relação ao capítulo 0, referente à extração de RTs, envolvendo os encaminhamentos metodológicos relacionados.

Quadro 6 – Mapeamento entre o referencial teórico e os encaminhamentos metodológicos para o processo de extração de RTs.

Referências	**Conceito**	**Contribuição**	**Encaminhamento metodológico**
Chollet (2017), Alpaydin (2014), Marsland (2015)	Aprendizado de máquina	Definições gerais sobre aprendizado de máquina.	Extração de RTs
Goodfellow, Bengio e Courville (2016)	Aprendizado supervisionado	Definições gerais sobre aprendizado supervisionado.	Extração de RTs
Chollet (2017), Zhu e Goldberg (2009)	Classificação	Definições gerais e questões específicas da tarefa de classificação.	Extração de RTs
Boser, Guyon e Vapnik (1992), Ben-Hur e Weston (2010), Cortes e Vapnik (1995)	SVM	Conceito de SVM, assim como questões relevantes relacionados ao algoritmo.	Extração de RTs

Fonte: O autor (2020).

3 TRABALHOS RELACIONADOS

Neste capítulo, são apresentados alguns trabalhos relacionados que fundamentaram o modelo metodológico, como também auxiliaram na sustentação, justificativa e problematização. Trata-se de um recorte da revisão sistemática efetuada para o projeto, disponível em sua íntegra no Apêndice A. O objetivo da revisão foi apresentar uma revisão do estado da arte sobre a extração de RTs para textos clínicos, buscando responder à pergunta: qual é a efetividade de aprendizado de máquina e abordagens baseadas em regras na identificação de RTs em textos clínicos?

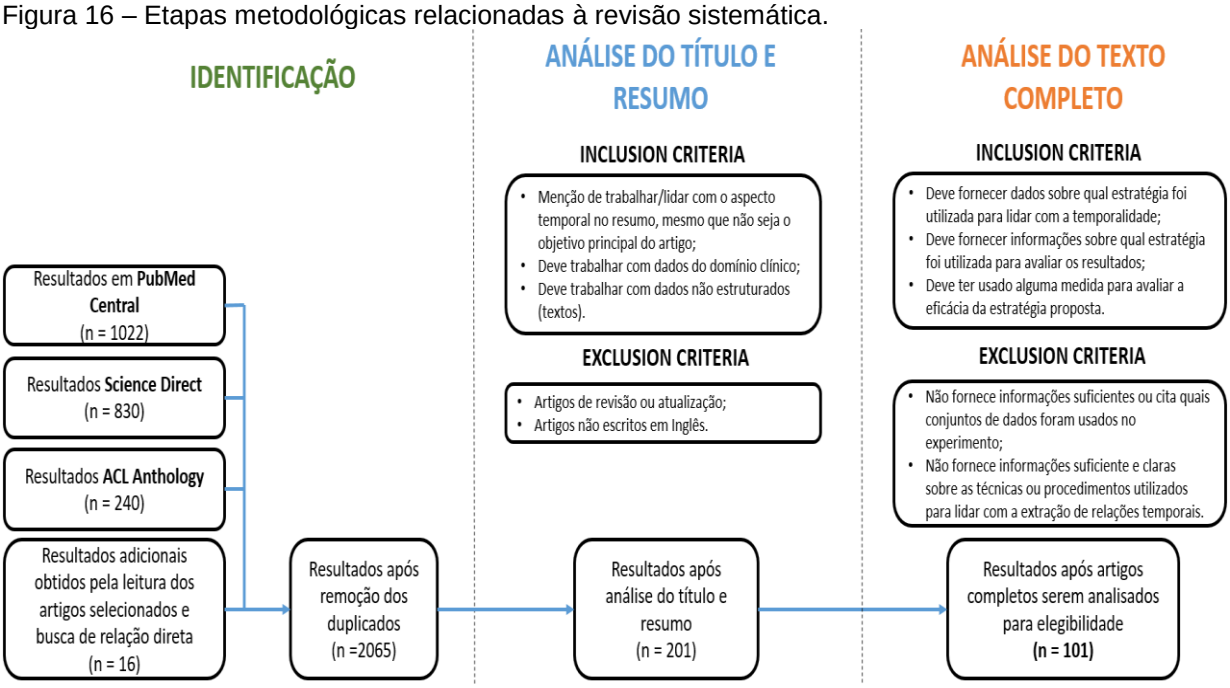
Como já mencionado, as *shared tasks* referentes à extração de RTs, tanto no domínio geral quanto clínico, focaram na divisão destas em relações entre menções (EVTs e ETs) no texto, chamadas TLINKs, e relações entre EVTs e a DCD, nomeadas DocTimeRel. Ambos os tipos de RT são abordados neste capítulo, visto que o projeto visou à sua extração.

3.1 METODOLOGIA DE BUSCA

As bases de dados utilizadas durante a revisão foram: Sistema Online de Busca e Análise de Literatura Médica (MEDLINE), Science Direct e ACL Anthology. Os descritores empregados nas buscas foram: (“*temporal relation*” OR “*temporal relations*” OR “*temporal extraction*” OR “*temporal information*” OR “*temporal relationship*” OR “*temporal relationships*” OR “*timeline*”) AND (“*clinical text*” OR “*clinical texts*” OR “*clinical narratives*” OR “*clinical narrative*” OR “*clinical reports*” OR “*clinical report*”).

Na Figura 16, são sumarizados os passos metodológicos. Foram identificados 2.092 artigos nas bases de dados e 16 artigos adicionais por leitura de estudos e busca de relações diretas entre os artigos adicionais e a revisão. Após a análise dos critérios de título e resumo, como também de texto completo, foram selecionados 101 artigos, avaliados de acordo com suas abordagens para lidar com a temporalidade e seus resultados quantitativos.

Figura 16 – Etapas metodológicas relacionadas à revisão sistemática.



Fonte: O autor (2020).

3.2 SUMARIZAÇÃO DAS PUBLICAÇÕES

Entre os 101 artigos, houve uma sumarização das abordagens em: (i) totalmente baseada em regras; (ii) baseada em aprendizado de máquina e híbrida (aprendizado de máquina e regras); (iii) aprendizado profundo. Cada tipo de RT (DocTimeRel e TLINK) apresentou essas divisões. As abordagens, em suas respectivas divisões, estão detalhadas na revisão sistemática presente no Apêndice A, contando, inclusive, com uma sumarização dos resultados e abordagens, assim como conclusões sobre a evolução das abordagens ao longo do tempo.

Para sustentar a metodologia da tese, foi feito um recorte dessa revisão, trazendo para esta seção somente as abordagens baseadas em aprendizado de máquina e híbridas (tanto para TLINK quanto para DocTimeRel), por serem o foco deste projeto. Na seção 3.3, serão detalhadas as abordagens para extração de DocTimeRel e, na seção 3.4, de TLINKs.

3.3 DOCTIMEREL

Os artigos que extraíram DocTimeRel com abordagens baseadas em aprendizado de máquina ou híbridas são listados no Quadro 7, que contém informação sobre os autores, o objetivo principal do artigo, a estratégia usada para extrair a temporalidade e o resultado, que é diretamente relacionado com a coluna

Sep DocTimeRel (avaliação de DocTimeRel separada). Se o artigo tivesse uma avaliação separada, os resultados seriam sobre a extração de RTs e, caso contrário, sobre o objetivo principal do sistema.

Quadro 7 – Artigos relacionados à extração de DocTimeRel que usaram abordagens baseadas em aprendizado de máquina ou híbridas.

Autores	**Objetivo**	**Melhor estratégia**	**Resultados**	**Sep DocTimeRel**
Henriksson *et al.* (2015)	EI de EAM	2 clfs RF	F1 0.243 para *Past* e 0,220 para *Future*	Sim
Mowery *et al.* (2009)	Comparação entre ConText e RIPPER	Clf RIPPER	Acc 0,971	Sim
Zhu, Yang e Yan (2017)	EIT de comunidades de saúde *on-line*	3 clfs SVM	F1 0,535	Sim
Raghavan, Fosler-Lussier e Lai (2012a)	Investigação do trabalho com *time bins*	Clf CRF	F1 0,84	Sim
Raghavan, Fosler-Lussier e Lai (2012b)	Resolução de correferências (*time bins* como *features*)	Clf CRF	F1 0,7574 AD, 0,9322 BA, 0,4610 OA, 0,5744 WBA e 0,9630 AA	Sim
Lin *et al.* (2015)	Identificação de EAM (*features* temporais)	Clf SVM	F1 of 0,829	Não
Torii *et al.* (2015)	UTHealth 2014 EFR	21 clfs RIPPER + votação	F1 0,9185	Não
Grouin, Moriceau e Zweigenbaum (2015)	UTHealth 2014 EFR	Clfs *OneRule*	F1 0,857	Não
Q. Chen *et al.* (2015)	UTHealth 2014 EFR	Estratégia *label-powerset* + clfs SVM	F1 0,9268	Não
Goodwin e Harabagiu (2015)	UTHealth 2014 EFR	*Markov networks* + regras	F1 0,9098	Não
Cormack *et al.* (2015)	UTHealth 2014 EFR	CART DT + df	F1 0,917	Não
Jonnagaddala *et al.* (2015)	UTHealth 2014 EFR	Clf NB + regras	F1 0,8302	Não
Roberts *et al.* (2015)	UTHealth 2014 EFR	3 clfs SVM + df + regras + refinamento anotação	F1 0,9277	Não
Cherry *et al.* (2013)	i2b2 2012 ERT	Clf SVM	F1 0,6954	Não
D'Souza e Ng (2013)	i2b2 2012 ERT	Clf CRF	F1 0,693	Não
Tang *et al.* (2013)	i2b2 2012 ERT	2 clfs SVM	F1 0,6932	Não
Xu *et al.* (2013)	i2b2 2012 ERT	2 clfs SVM	F1 0,6849	Não
Lin *et al.* (2016a)	i2b2 2012 ERT e CTE 2015 ERT	2 clfs SVM (i2b2) e clf SVM (CTE)	F1 0,695 (i2b2) e 0,807 (CTE)	Não
Nikfarjam, Emadzadeh e Gonzalez (2013)	i2b2 2012 ERT	Clf SVM + regras	F1 0,63	Não
Roberts, Rink e Harabagiu (2013)	i2b2 2012 ERT	Clf SVM + regras	F1 0,5594	Não
Styler *et al.* (2014a)	THYME corpus ERT	Clf SVM	F1 0,474	Sim

Autores	Objetivo	Melhor estratégia	Resultados	Sep DocTimeRel
Velupillai *et al.* (2015b)	CTE 2015 ERT	Clf CRF	F1 0,791	Sim
Fries (2016)	CTE 2016 ERT	Clf LR + regras	F1 0,743	Sim
Khalifa, Velupillai e Meystre (2016)	CTE 2016 ERT	Clf CRF	F1 0,844	Sim
Chikka (2016)	CTE 2016 ERT	Clf CRF	F1 0,714	Sim
Caselli e Morante (2016)	CTE 2016 ERT	Clf CRF	F1 0,712	Sim
Grouin e Moriceau (2016)	CTE 2016 ERT	Clf CRF	F1 0,687	Sim
Lee *et al.* (2016)	CTE 2016 ERT	Clf SVM	F1 0,835	Sim
Tourille *et al.* (2017a)	CTE 2016 + MERLOT ERT	Clf SVM (CTE) e clf SVM (MERLOT)	F1 0,87 (CTE) e 0,83 (MERLOT)	Sim
Tourille *et al.* (2016)	CTE 2016 ERT	Clf RF	F1 0,807	Sim
Cohan, Meurer e Goharian (2016)	CTE 2016 ERT	Clfs LR	F1 0,815	Sim
MacAvaney, Cohan e Goharian (2017)	CTE 2017 ERT	Clf CRF	F1 0,40 ANSD, 0,50 ASD	Sim
Sarath, Manikandan e Niwa (2017)	CTE 2017 ERT	Clf CRF	F1 0,45 ANSD, 0,52 ASD	Sim
Huang *et al.* (2017)	CTE 2017 ERT	Clf SVM	F1 0,49 ASD	Sim
Tourille *et al.* (2017b)	CTE 2017 ERT	Clf SVM	F1 0,519 UDA, 0,591 ASD	Sim
Leeuwenberg e Moens (2017a)	CTE 2016 ERT	*Structured perceptron* + ILP	F1 0,846	Sim
Leeuwenberg e Moens (2017b)	CTE 2017 ERT	*Structured perceptron* + ILP	F1 0,49 ANSD, 0,56 ASD	Sim
Viani *et al.* (2019)	ERT de textos cardiológicos	Clf SVM	F1 0,857 *Overlap*, 0,834 *Before* e 0,793 *After*	Sim

Fonte: O autor (2020).
Notas: EAM = eventos adversos a medicamentos; EI = extração de informação; RF = *random forest*; CRF = *conditional random fields*; SVM = *support vector machine*; LR = *logistic regression*; NB = Naive Bayes; DT = *decision trees*; AM = aprendizado de máquina; EIT = extração de informação temporal; EFR = extração de fatores de risco; ERT = extração de relações temporais; AD = *after discharge*; BA = *before admission*; OA = *on admission*; WBA = *way before admission*; AA = *after admission*; ANSD = adaptação não supervisionada do domínio; ASD = adaptação supervisionada do domínio; F1 = F1-*score*; clf = único classificador; clfs = mais de um classificador; df = *default value* (valor padrão); CTE = Clinical TempEval; ILP = *Integer Linear Programming*.

A maior parte desses artigos é relacionada a *shared tasks*, evidenciando a falta de *corpora* disponíveis para pesquisa. Entre os trabalhos não referentes a *shared tasks*, ressaltam-se: Henriksson *et al.* (2015), com o uso de um algoritmo *Random Forest* (RF) para classificação em duas relações; Mowery *et al.* (2009), testando diversos classificadores em relação ao ConText (modelo baseado em regras); Zhu, Wang e Yan (2017), usando um classificador SVM para cada tipo de relação; e

Raghavan, Fosler-Lussier e Lai (2012a, 2012b), com um modelo baseado em *Conditional Random Fields* (CRF) para classificar EVTs em uma de cinco categorias.

Grande parte dos trabalhos recuperados era relacionada às seguintes *shared tasks*: i2b2/UTHealth 2014 *challenge* (STUBBS *et al.*, 2015), i2b2 2012 e Clinical TempEval 2015, 2016 e 2017. Aquelas associadas ao Clinical TempEval e i2b2 2012 foram extensivamente detalhadas nas seções anteriores. Em relação à *shared task* i2b2/UTHealth 2014, houve menções a doenças cardíacas, com foco na descoberta de fatores de risco. Não existiu avaliação separada para DocTimeRel, devido à temporalidade não ser seu maior foco. Basicamente, cada fator de risco poderia ser conectado a uma DCD com uma ou mais relações (*Before*, *After* e *During*), a transformando em uma tarefa de classificação multirrótulo, aspecto que a torna diferente deste projeto, em que cada EVT só pode ter um rótulo.

As publicações referentes ao i2b2/UTHealth 2014 podem ser sumarizadas como: Torii *et al.* (2015), usando classificadores RIPPER e votação; Grouin, Moriceau e Zweigenbaum (2015), utilizando um classificador para cada fator de risco, testando classificadores SVM e *OneRule*; Goodwin e Harabagiu (2015), empregando *Markov Network*s e regras; Cormack *et al.* (2015), utilizando regras e *Decision Trees* (DTs); Jonnagaddala *et al.* (2015), usando Naive Bayes e regras; Q. Chen *et al.* (2015), aplicando uma estratégia baseada em *label-powerset* e classificadores SVM; e Roberts *et al.* (2015), utilizando regras e três classificadores SVM, um para cada tipo de relação.

Uma abordagem popular em trabalhos envolvendo esse *corpus* foi se basear no valor padrão da marcação, ou seja, verificar qual era a marcação mais presente no conjunto de treinamento para aquele fator de risco e o marcar automaticamente e/ou criar regras baseadas nesse aspecto. Roberts *et al.* (2015) exploraram o impacto de anotações adicionais no conjunto de dados, anotando e testando esses dois conjuntos. Os resultados foram melhores que a própria anotação original.

Os melhores resultados para *corpus* i2b2/UTHealth 2014 envolvem Roberts *et al.* (2015) e Q. Chen *et al.* (2015), ambos se baseando em diversos classificadores SVM, porém os primeiros obtiveram resultados superiores adicionando regras e refinamento da anotação.

Como já mencionado, no *corpus* i2b2 2012, existe DocTimeRel relacionada à data de admissão e à data de alta. A abordagem utilizada por Cherry *et al.* (2013) e D'Souza e Ng (2013) consistiu em utilizar um único classificador para ambas as

DocTimeRel, SVM e CRF, respectivamente. Tang *et al.* (2013), Xu *et al.* (2013) e Lin *et al.* (2016a) usaram dois classificadores SVM para cada DocTimeRel, sendo um relativo a relações com a data de admissão e outro, com a data de alta. O mesmo modelo desenvolvido por Lin *et al.* (2016a) foi aplicado para identificação de toxicidade hepática devido ao uso de metotrexato em pacientes com artrite reumatoide por Lin *et al.* (2015).

Outra abordagem empregada para o i2b2 2012 foi a híbrida, adicionando regras aos modelos. Nikfarjam, Emadzadeh e Gonzalez (2013) e Roberts, Rink e Harabagiu (2013) primeiramente usaram regras para conectar os EVTs a uma das DCDs e, em seguida, classificadores.

Pelo método de avaliação do i2b2 2012, não é possível diferenciar TLINKs de DocTimeRel, dificultando uma avaliação real dos resultados. No entanto, de forma geral, as abordagens propostas por Tang *et al.* (2013), Lin *et al.* (2016a), D'Souza e Ng (2013) e Cherry (2013) tiveram melhores resultados.

Em relação ao THYME *corpus*, houve um trabalho de Styler *et al.* (2014a) antes da divulgação do *corpus* concernente ao Clinical TempEval 2015. Foi utilizado o ClearTK-TimeML (BETHARD, 2013), um dos melhores classificadores para o *corpus* TempEval 2013, sendo evidenciada a questão de que a natureza do domínio clínico impactou negativamente nos resultados.

Entre os trabalhos relacionados ao Clinical TempEval (2015, 2016 e 2017), a maioria dos testes foi baseada somente em aprendizado de máquina, única exceção foi Fries (2016), combinando *Logistic Regression* (LR) e regras.

A abordagem mais popular foi a utilização de um classificador SVM ou um classificador CRF. Nesse contexto, CRF foi empregado pelos autores: Velupillai *et al.* (2015b), Khalifa, Velupillai e Meystre (2016), Chikka (2016), Casseli e Morante (2016), Grouin e Moriceau (2016), MacAvaney, Cohan e Goharian (2017) e Sarath, Manikandan e Niwa (2017). Já SVM foi usado pelos autores: Lin *et al.* (2016a), Lee *et al.* (2016), Tourille *et al.* (2017a, 2017b) e Huang *et al.* (2017).

Diferentemente das abordagens anteriores, Tourille *et al.* (2016) usaram um classificador RF e Cohan, Meurer e Goharian (2016), um classificador LR para cada tipo de relação. As publicações de Leeuwenberg e Moens (2017a, 2017b) tiveram foco em aprendizado de máquina estruturado, classificando juntamente DocTimeRel e TLINKs com um modelo baseado em *Structured Perceptron* e *Integer Linear Programming* (ILP).

Com *corpus* similar ao THYME, a publicação de Viani *et al.* (2019) trabalhou com textos cardiológicos em italiano usando os mesmos tipos de DocTimeRel. Para classificação, foi usado um classificador SVM.

Para o THYME *corpus*, as melhores abordagens foram baseadas em um único classificador, tanto SVM quanto CRF. A diferença residiu na engenharia de *features*. A abordagem baseada em *Structured Perceptron* e ILP de Leeuwenberg e Moens (2017a, 2017b) também esteve entre os melhores resultados.

3.3.1 Conclusões DocTimeRel

Dependendo do objetivo da aplicação, conectar o evento à sua DCD já pode fornecer informação temporal suficiente, com uma representação mais detalhada da relação em categorias predefinidas aumentando a dificuldade de extração.

Existe um *trade-off* entre adicionar mais categorias, trazendo uma temporalidade mais específica, e a dificuldade da extração. Por exemplo, melhores resultados foram obtidos por Henriksson *et al.* (2015) com duas categorias (*Before* e *After*) em comparação ao THYME *corpus*, que tem quatro categorias (*Before*, *Before/Overlap*, *Overlap* e *After*). Outra questão importante de comentar é que categorias como *Before* e *Before/Overlap* são mais difíceis de classificar devido ao fato de o classificador precisar diferenciar EVTs passados daqueles que também são do passado, porém continuaram ocorrendo até o momento da consulta. Em certos casos, essa diferenciação pode ser altamente dependente da interpretação e do contexto. Por isso, a classificação de DocTimeRel no THYME foi mais complexa que no trabalho de Henriksson *et al.* (2015).

No geral, os melhores resultados estiverem relacionados ao uso de aprendizado de máquina ou modelos híbridos baseados em SVM e CRF, com exceção de Leeuwenberg e Moens (2017a, 2017b). O conjunto de *features* utilizadas por essas abordagens envolveu aquelas relacionadas a informações dos *tokens*, como n-gramas, e *features* sintáticas, como *Part-Of-Speech Tagging* (POS), tanto em relação aos EVTs quanto ao contexto. Além disso, informações sobre verbos ao redor, como POS, sobre os atributos dos EVTs, seções do documento e ETs próximas foram efetivas. Alguns autores também se fundamentaram em *features* semânticas baseadas em dicionários, com destaque para a *Unified Medical Language System* (UMLS) (BODENREIDER, 2004).

3.4 TLINK

Os artigos que extraíram TLINKs com abordagens baseadas em aprendizado de máquina ou híbridas são listados no Quadro 8, que contém informações sobre os autores, o objetivo principal do artigo, a estratégia usada para extrair a temporalidade e o resultado, que é diretamente relacionado com a coluna Sep TLINK (avaliação de TLINK separada). Se o artigo tivesse uma avaliação separada, os resultados seriam sobre a extração de RTs e, caso contrário, sobre o objetivo principal do sistema.

Quadro 8 – Artigos relacionados à extração de TLINKs que usaram abordagens baseadas em aprendizado de máquina ou híbridas.

Autores	**Objetivo**	**Melhor estratégia**	**Seleção de pares candidatos**	**Resultados**	**Sep TLINK**
Luo *et al.* (2011)	Extração de restrições de critérios de elegibilidade	Clf CRF	-	F1 0,7981	Não
Bramsen *et al.* (2006a)	Ordenamento temporal por segmentos	BoosTexter	-	Acc 0,783	Sim
Bramsen *et al.* (2006b)	Ordenamento temporal por segmentos	BoosTexter + Perceptron + ILP	-	Acc 0,84	Sim
Raghavan *et al.* (2014)	Alinhamento de eventos	Clf ME	-	F1 0,673	Sim
Raghavan, Fosler-Lussier, e Lai (2012c)	Ordenamento de eventos *(pairwise* e *ranking*)	Clf SVM (*pairwise*), clf SVM (*ranking*)	-	Acc 0,8216 (ranking), 0,7133 (pairwise)	Sim
Chang *et al.* (2013)	i2b2 2012 ERT	2 clfs ME + regras	Regras	F1 0,5628	Não
Grouin *et al.* (2013)	i2b2 2012 ERT	9 clfs + regras	Produto cruzado	F1 0,6231	Não
Roberts, Rink e Harabagiu (2013)	i2b2 2012 ERT	2 clfs SVM	Regras	F1 0,5594	Não
Moharasan e Ho (2019)	i2b2 2012 ERT	SD: clf NB, MS: 2 clfs NB	Regras	F1 0,671	Não
Cherry *et al.* (2013)	i2b2 2012 ERT	MS: 2 clfs ME, SD: 1 clf ME + regras	MS: TPP, SD: regras	F1 0,6954	Não
Lin *et al.* (2016a)	i2b2 2012 e CTE 2015 ERT	MS: 2 clfs SVM, SD: 2 clfs SVM + regras (SD somente no i2b2 2012). CSL; expansão CT	MS: TPP, SD: regras	F1 0,695 (i2b2), 0,321 (CTE)	Não
Xu *et al.* (2013)	i2b2 2012 ERT	MS: 3 clfs SVM, SD: 3 clf SVM	MS: TPP, SD: regras	F1 0,6849	Não
Sohn *et al.* (2013)	i2b2 2012 ERT	Clf SVM + regras	MS: TPP, SD: regras	F1 0,537	Não
Cheng *et al.* (2013)	i2b2 2012 ERT	MS: clf ME + resolução de conflito, SD: regras	MS: regras	F1 0,43	Não

Autores	Objetivo	Melhor estratégia	Seleção de pares candidatos	Resultados	Sep TLINK
Nikfarjam, Emadzadeh e Gonzalez (2013)	i2b2 2012 ERT	MS: 2 clfs SVM + grafo temporal, SD: regras	MS: TPP, grafo	F1 0,63	Não
Tang *et al.* (203)	i2b2 2012 ERT	MS: 2 clfs SVM, CS: 2 clfs SVM	MS: regras, SD: regras	F1 0,6932	Não
D'Souza e Ng (2013)	i2b2 2012 ERT	Baseado em Tang *et al.* (2013) + regras	Baseado em Tang *et al.* (2013)	F1 0,693	Não
D'Souza e Ng (2014a)	i2b2 2012 ERT	Baseado em Tang *et al.* (2013) + regras	Baseado em Tang *et al.* (2013)	F1 0,702	Não
Lee *et al.* (2018)	Reanotação do i2b2 2012 para relações diretas (ERT)	Clfs SVM + regras + CSL	MS: TPP	F1 0,6377 (NTC)	Não
D'Souza e Ng (2014b)	Anotação de relações em SD faltantes no i2b2 2012 (ERT)	D'Souza e Ng (2013), D'Souza e Ng (2014a)	D'Souza e Ng (2013, 2014)	F1 0,341 (NTC)	Não
Miller *et al.* (2013)	THYME *corpus* ERT	Clf SVM	MS: TPP	F1 0,737 (NTC)	Sim
Lin *et al.* (2014)	THYME *corpus* ERT	Clf SVM	MS: TPP	F1 0,708 (NTC)	Sim
Styler *et al.* (2014a)	THYME *corpus* ERT	2 clfs SVM	MS: TPP	F1 0,204 (NTC)	Sim
Velupillai *et al.* (2015b)	CTE 2015 ERT	Clf CRF + regras	Regras	F1 0,181	Sim
Khalifa, Velupillai e Meystre (2016)	CTE 2016 ERT	MS: 2 clfs SVM, SD: 2 clfs SVM	MS: TPP, SD: regras	F1 0,511	Sim
Caselli e Morante (2016)	CTE 2016 ERT	MS: 2 clfs CRF	MS: TPP	F1 0,453	Sim
Barros *et al.* (2016)	CTE 2016 ERT	Clfs 4 CRF	MS: TPP, SD: regras	F1 0,264	Sim
Tourille *et al.* (2017a)	CTE 2016 + MERLOT ERT	WS: clf SVM. 3-classes	MS: TPP	F1 0,53 (CTE), 0,65 (MERLOT)	Sim
Tourille *et al.* (2016)	CTE 2016 ERT	MS: clf SVM, SD: clf SVM. Regras; 3-classes	MS: TPP, TPP: regras	F1 0,538	Sim
Lin *et al.* (2016b)	CTE 2016 ERT	MS: 2 clfs SVM. Expansão CT	MS: APP	F1 0,594	Sim
Lee *et al.* (2016)	CTE 2016 ERT	MS: 2 clfs SVM, CS: 4 clfs SVM. CSL	MS: TPP + filtragem de pares, SD: regras	F1 0,573	Sim
Chikka (2016)	CTE 2016 ERT	Clf CRF	-	F1 0,313	Sim
Leeuwenberg e Moens (2016)	CTE 2016 ERT	2 clfs SVM	MS: TPP + restrições, SD: regras	F1 0,551	Sim
Leeuwenberg e Moens (2017a)	CTE 2016 ERT	Structured Perceptron + ILP	TPP janela *token* + regras	F1 0,608	Sim
Sarath, Manikandan e Niwa (2017)	CTE 2017 ERT	MS: 2 ensembles de clfs, SD: 2 ensembles de clfs. CSL	MS: TPP, SD: regras. Filtragem de pares; regras	F1 0,23 ANSD, 0,15 ASD	Sim

Autores	Objetivo	Melhor estratégia	Seleção de pares candidatos	Resultados	Sep TLINK
MacAvaney, Cohan e Goharian (2017)	CTE 2017 ERT	MS: clf XGBoost	MS: TPP	F1 0,34 ANSD, 0,25 ASD	Sim
Leeuwenberg e Moens (2017b)	CTE 2017 ERT	Structured Perceptron + ILP	TPP em janela *token*	F1 0,32 ANSD, 0,28 ASD	Sim
Huang *et al.* (2017)	CTE 2017 ERT	WS: 2 clfs SVM	MS: TPP	F1 0,26 ASD	Sim

Fonte: O autor (2020).
Notas: CRF = *conditional random fields*; SVM = *support vector machine*; ME = *maximum entropy*; NB = Naive Bayes; ERT = extração de relações temporais; ANSD = adaptação não supervisionada do domínio; ASD = adaptação supervisionada do domínio; F1 = F1-*score*; clf = único classificador; clfs = mais de um classificador; CTE = Clinical TempEval; ILP = *Integer Linear Programming*; MS = TLINK em mesma sentença; SD = TLINK em sentença distinta; TPP = todos os possíveis pares; CT = conjunto de treinamento; 3-classes = transformar em um problema de classificação de três classes; Acc = *Accuracy*; CSL = *cost-sensitive learning*.

A maior parte desses artigos é relacionada a *corpora* disponibilizados por meio de *shared tasks*, evidenciando a falta de *corpora* disponíveis para pesquisa. Entre os trabalhos não referentes a *shared tasks*, ressaltam-se: Luo *et al.* (2011), com foco na extração de restrições de elegibilidade com CRF; Bramsen *et al.* (2006a, 2006b), com o objetivo de identificação de pares de segmentos temporais; Raghavan *et al.* (2014), com foco no alinhamento dos EVTs; e Raghavan, Fusler e Lai (2012c), com foco em *ranking* e classificação de pares de EVTs com SVM.

Os resultados para o i2b2 2012 foram sumarizados de acordo com três aspectos: técnica para seleção de pares candidatos, abordagem para extração de TLINKs em mesma sentença (pares de menções na mesma sentença) e TLINKs em sentenças distintas (pares de menções em sentenças diferentes, os quais têm grande impacto nos resultados da extração temporal). Como qualquer EVT ou ET pode gerar pares candidatos ao treinar um classificador, isso cria um alto número de exemplos negativos no conjunto de treinamento. A maioria dos autores diferenciou claramente as abordagens usadas para geração de pares para TLINKs em mesma sentença e em sentenças distintas, o que melhorou os resultados.

Para TLINKs em mesma sentença, Cherry *et al.* (2013), Lin *et al.* (2016a), Xu *et al.* (2013) e Sohn *et al.* (2013) consideraram todos os possíveis pares na sentença. Adicionalmente, Lin *et al.* (2016a) utilizaram *cost-sensitive learning* para reduzir a questão do desbalanceamento, além de uma técnica de expansão do conjunto de treinamento baseada em UMLS. Cheng *et al.* (2013) optaram por regras e Nikfarjam, Emadzadeh e Gonzalez (2013), por uma abordagem baseada em grafos. Tang *et al.*

(2013) consideraram todos os possíveis pares consecutivos de menções dentro uma sentença, assim como pares que apresentavam uma relação de dependência. Essa heurística foi utilizada como base nos trabalhos de D'Souza e Ng (2013, 2014a).

Para TLINKs em sentenças distintas, Tang *et al.* (2013) e Cherry *et al.* (2013) focaram em TLINKs entre EVTs durante a criação de pares candidatos. Tang *et al.* (2013) consideraram todos os possíveis pares entre os primeiros e os últimos EVTs em ambas as sentenças (sentenças consecutivas), adicionando heurísticas baseadas nos atributos dos EVTs e sintaxe para buscar casos de correferência. A mesma heurística foi usada como base nos trabalhos de D'Souza e Ng (2013, 2014a). Lin *et al.* (2016a) empregaram a abordagem de considerar somente pares entre os primeiros e os últimos EVTs em ambas as sentenças, anteriormente proposta por Tang *et al.* (2013). Cherry *et al.* (2013) criaram todos os possíveis pares de EVTs usando uma janela de cinco sentenças adjacentes, restringindo a EVTs de mesmos atributos. Já Xu *et al.* (2013) consideraram todos os pares em sentenças adjacentes.

De acordo com os resultados obtidos, nota-se que métodos para criação de pares candidatos para TLINKs em mesma sentença que consideraram todos os possíveis pares de menções, como Cherry *et al.* (2013), Lin *et al.* (2016a) e Xu *et al.* (2013), ou que definiram heurísticas específicas, como Tang *et al.* (2013), D'Souza e Ng (2013, 2014a), obtiveram melhores resultados. Para TLINKs em sentenças distintas, os melhores resultados foram obtidos por Tang *et al.* (2013), Cherry *et al.* (2013) e D'Souza e Ng (2013, 2014a), devido ao foco em definir heurísticas para selecionar somente certos tipos de TLINK entre EVTs, e pelas heurísticas propostas por Lin *et al.* (2016a) e Xu *et al.* (2013).

Em relação às abordagens utilizadas para extrair TLINKs, alguns autores empregaram a mesma para extração de TLINKs entre menções em mesma sentença e em sentenças distintas, a saber: Roberts *et al.* (2013), utilizando SVM; Grouin *et al.* (2013), separando a extração em 57 situações e desenvolvendo classificadores para nove; Chang *et al.* (2013), usando uma abordagem híbrida priorizando as regras; e Sohn *et al.* (2013), utilizando uma abordagem híbrida, com SVM e regras.

A maioria dos autores diferenciou entre TLINKs em mesma sentença e em diferentes sentenças. Para TLINKs em mesma sentença, Cherry *et al.* (2013), Tang *et al.* (2013), Xu *et al.* (2013), Lin *et al.* (2016a) e D'Souza e Ng (2013, 2014a) empregaram um classificador para TLINKs entre EVTs e outro classificador para TLINKs entre EVTs e ETs. D'Souza e Ng (2013, 2014a) primeiramente usaram regras

e, se não foram efetivas, recorreram aos classificadores. Tang *et al.* (2013), Lin *et al.* (2016a), D'Souza e Ng (2013, 2014a) e Xu *et al.* (2013) aplicaram classificadores SVM para ambos os TLINKs. Xu *et al.* (2013) adicionalmente utilizaram um classificador SVM para TLINKs entre ETs.

Para TLINKs em sentenças distintas, Tang *et al.* (2013), Xu *et al.* (2013), Lin *et al.* (2016a) e D'Souza e Ng (2013, 2014a) utilizaram um classificador SVM para EVTs em sentenças consecutivas e outro classificador SVM para EVTs que eram correferências. Já Cherry *et al.* (2013) treinaram um modelo para TLINK entre EVTs do tipo *Overlap* e usaram regras para relações do tipo *Before* e *After*. Xu *et al.* (2013) adicionalmente empregaram um classificador SVM para TLINKs entre ETs.

Para TLINKs em mesma sentença, os melhores resultados foram de Tang *et al.* (2013), Cherry *et al.* (2013), Xu *et al.* (2013), Lin *et al.* (2016a) e D'Souza e Ng (2013, 2014a), com a característica principal do comum uso de um classificador para TLINKs entre EVTs e outro classificador para TLINKs entre EVTs e ETs. Todos, com exceção de Cherry *et al.* (2013), utilizaram SVM.

Para TLINKs em sentenças distintas, classificadores especializados também obtiveram os melhores resultados. As abordagens de Tang *et al.* (2013), Lin *et al.* (2016a), Cherry *et al.* (2013) e D'Souza e Ng (2013, 2014a) foram efetivas para esse tipo. Lin *et al.* (2016a), Cherry *et al.* (2013) e D'Souza e Ng (2013, 2014a) adicionalmente consideraram regras.

Em estudo que teve como foco reanotar certos TLINKs, Lee *et al.* (2018) reanotaram *corpus* i2b2 2012 para considerar somente relações diretas, extraindo TLINKs entre EVTs e ETs, com o uso de *cost-sensitive learning*, SVM e regras. Os autores testaram sua abordagem em comparação à de Tang *et al.* (2013) no *corpus* refeito, atingindo melhores resultados.

O estudo de D'Souza e Ng (2014b) teve foco na anotação de TLINKs faltantes entre menções em diferentes sentenças no *corpus* i2b2 2012. Utilizou a mesma abordagem de D'Souza e Ng (2013, 2014a), atingindo resultados melhores com o *corpus* expandido do que com o original.

Entre os artigos para extração de TLINKs, a maioria envolveu o THYME *corpus*. Alguns estudos, como de Miller *et al.* (2013), Lin *et al.* (2014) e Styler *et al.* (2014a), extraíram informações de uma pequena parte do THYME *corpus* em períodos anteriores à disponibilização dele em *shared tasks*. Miller *et al.* (2013), Lin *et al.* (2014)

e Styler *et al.* (2014a) desenvolveram abordagens para TLINKs em mesma sentença com SVM, com os dois primeiros estudos utilizando *tree kernels*.

Em relação ao *corpus* usado para o Clinical TempEval 2015, houve baixo número de participantes, devido ao longo processo de autorização. Somente dois artigos, de Velupillai *et al.* (2015b) e Lin *et al.* (2016a), estão relacionados a ele. Velupillai *et al.* (2015b) consideraram pares entre menções, desenvolvendo heurísticas para limitar a quantidade de pares com base na distância de sentença e proximidade. Utilizaram CRF e regras para extração de TLINKs. Já Lin *et al.* (2016a) restringiram-se a TLINKs em mesma sentença, considerando todos os possíveis pares, usando um classificador SVM para TLINKs em mesma sentença e outro classificador SVM para TLINKs em sentenças distintas. Adicionalmente, aplicaram *cost-sensitive learning* e realizaram uma expansão do conjunto de treinamento por meio de UMLS. As abordagens propostas por Lin *et al.* (2016a) tiveram melhor resultado. Questões como reduzir a quantidade de pares candidatos, mesmo pelo descarte de TLINKs em sentenças distintas, e abordagens de aprendizado de máquina específicas para TLINKs entre EVTs e TLINKs entre EVTs e ETs melhoraram o resultado (como já visto durante a revisão dos artigos relacionados ao i2b2 2012).

As publicações sobre Clinical TempEval 2016 serão sumarizadas de acordo com três aspectos: técnica para seleção de pares candidatos, abordagem para extração de TLINKs em mesma sentença e abordagem para extração de TLINKs em sentenças distintas.

Para seleção de pares candidatos em TLINKs em mesma sentença, a estratégia mais comum foi considerar todos os possíveis pares entre menções, utilizada por: Khalifa, Velupillai e Meystre (2016), Caselli e Morante (2016), Barros *et al.* (2016), Tourille *et al.* (2016, 2017a) e Lin *et al.* (2016b). Lee *et al.* (2016) também consideram todos os possíveis pares, porém filtraram aqueles “não prováveis” de possuir TLINKs, com base em regras criadas a partir de observações no THYME *corpus*. As regras removiam um par se as menções não estivessem na mesma sentença, se a DocTimeRel de um EVT fosse *Before* enquanto do outro fosse *After* ou se um EVT fosse hipotético e outro, factual. Leeuwenberg e Moens (2016) também consideraram todos os possíveis, porém adicionaram restrições baseadas em novas linhas. Essa questão se torna interessante para esta tese, uma vez que em diversos casos existe a separação de sentenças por quebras de linha.

Para seleção de pares candidatos em TLINKs em sentenças distintas, a estratégia mais comum foi restringir a TLINKs em mesma sentença, descartando todos os possíveis pares envolvendo elementos em sentenças diferentes. Como observado por Tourille *et al.* (2016), aproximadamente 76% dos exemplos positivos de TLINKs ocorreram na mesma sentença no conjunto de treinamento. Os autores que focaram somente em extração de TLINKs em mesma sentença foram: Caselli e Morante (2016), Tourille *et al.* (2017a) e Lin *et al.* (2016b).

Diversos autores consideraram TLINKs em sentenças distintas pela criação de heurísticas para desenvolvimento de pares candidatos. Barros *et al.* (2016) consideraram apenas pares em até uma sentença adjacente, enquanto Tourille *et al.* (2016), até três sentenças adjacentes (cobrindo 89% dos exemplos positivos do conjunto de treinamento). Lee *et al.* (2016) também restringiram a sentenças adjacentes, porém adicionaram heurísticas específicas para omitir pares a mais de duas sentenças adjacentes. Khalifa, Velupillai e Meystre (2016) restringiram a sentenças adjacentes, adicionando heurísticas para criar pares entre as primeiras e as últimas menções das sentenças. Leeuwenberg e Moens (2016) somente consideraram pares em mesma linha e adicionaram heurísticas baseadas em vírgulas e pontos para considerar mais pares.

Leeuwenberg e Moens (2017a) desenvolveram uma abordagem agnóstica a sentenças, restringindo pares candidatos a pares que ocorriam em uma janela de 30 *tokens* e em mesmo parágrafo.

Uma abordagem popular para lidar com o desbalanceamento foi trabalhar o problema de classificação de duas classes (*Contains* e *No Relation*) em um problema de classificação de três classes (*Contains*, *No Relation* e *Is_Contained*) pela adição da classe *Is_Contained*, indicando que uma menção estava contida em outra. Todos os possíveis pares foram gerados da esquerda para a direita e, quando necessário, a relação *Contains* foi trocada para *Is_Contained*. Essa abordagem foi utilizada por Tourille *et al.* (2016, 2017a). Sua vantagem foi cortar o número de pares candidatos pela metade ao custo da adição de uma classe. O método foi, inclusive, amplamente utilizado para abordagens baseadas em aprendizado profundo.

Em relação às abordagens para extração de TLINKs, diversos estudos criaram abordagens para extração de TLINKs entre menções em mesma sentença e em sentenças distintas. Para TLINKs em mesma sentença, Khalifa, Velupillai e Meystre (2016) e Lee *et al.* (2016) usaram um classificador SVM para TLINKs entre EVTs e

outro classificador SVM para TLINKs entre EVTs e ETs. Khalifa, Velupillai e Meystre (2016) também utilizaram outros dois classificadores SVM, para TLINKs em sentenças distintas. Lee *et al.* (2016) adicionaram classificadores mais específicos para TLINKs em sentenças distintas, com um classificador SVM para TLINKs entre EVTs e outro para TLINKs entre EVTs e ETs envolvendo menções a até duas sentenças adjacentes, além de outros dois classificadores SVM (um para TLINKs entre EVTs e outro para TLINKs entre EVTs e ETs) para TLINKs mais distantes que duas sentenças adjacentes. Também empregaram *cost-sensitive learning* para reduzir o efeito do desbalanceamento. O uso dos classificadores especializados para TLINKs a mais de duas sentenças adjacentes foi o componente de melhor resultado na abordagem de Lee *et al.* (2016), devido à abordagem para seleção de pares candidatos restringir os pares somente àqueles que “provavelmente” teriam uma relação. Tourille *et al.* (2016) utilizaram um classificador SVM para TLINKs em mesma sentença e outro para TLINKs em sentenças distintas, considerando regras para alguns casos em específico.

Sobre trabalhos que desenvolveram abordagens para TLINKs na sentença (não focaram em TLINKs em sentenças distintas), envolveram: Caselli e Morante (2016), que utilizaram dois classificadores SVM, um para TLINKs entre EVTs e outro para TLINKs entre EVTs e ETs; Tourille *et al.* (2017a), com um classificador SVM para extração de TLINKs em mesma sentença, com o objetivo de comparar métodos de extração em idiomas distintos (comparação com o *corpus* francês MERLOT); Lin *et al.* (2016b), com o uso de dois classificadores SVM, um para TLINKs entre EVTs e outro para TLINKs entre EVTs e ETs.

Diversas abordagens usaram o mesmo sistema para classificar tanto TLINKs em mesma sentença quanto em sentenças distintas, a saber: Chikka (2016), com um classificador CRF; Barros *et al.* (2016), usando dois classificadores CRF (um para TLINKs entre EVTs e outro para TLINKs entre EVTs e ETs); Leeuwenberg e Moens (2016), empregando o sistema temporal disponível do cTAKES (LIN *et al.*, 2016a), com adição de novas *features*; Leeuwenberg e Moens (2017a), classificando TLINKs e DocTimeRel juntamente, com um modelo baseado em *Structured Perceptron* e ILP.

Entre as abordagens baseadas em aprendizado de máquina, os melhores resultados para o *corpus* Clinical TempEval 2016 foram de Leeuwenberg e Moens (2017a), seguidos de Lee *et al.* (2016). Ressalta-se que as abordagens propostas por Lee *et al.* (2016), de utilizar diversos classificadores especializados, definir heurísticas

para reduzir a quantidade de pares e utilizar *cost-sensitive learning*, já vinham sendo consideradas efetivas desde a análise prévia dos resultados envolvendo o *corpus* i2b2 2012.

Para o *corpus* Clinical TempEval 2017, os artigos serão sumarizados de acordo com três aspectos: técnica para seleção de pares candidatos, abordagem para extração de TLINKs em mesma sentença e abordagem para extração de TLINKs em sentenças distintas.

Para seleção de pares entre TLINKs em mesma sentença, a abordagem mais comum foi considerar todos os possíveis pares, como Sarath, Manikandan e Niwa (2017), MacAvaney, Cohan e Goharian (2017) e Huang *et al.* (2017). Já para seleção de pares candidatos em sentenças distintas, alguns autores restringiram-se a TLINKs entre elementos em mesma sentença, descartando TLINKs entre elementos em sentenças distintas. Essa abordagem foi usada por MacAvaney, Cohan e Goharian (2017) e Huang *et al.* (2017). Sarath, Manikandan e Niwa (2017) consideraram TLINKs em sentenças distintas, com pares em até duas janelas adjacentes.

A abordagem aplicada por Leeuwenberg e Moens (2017b) foi baseada em Leeuwenberg e Moens (2017a), envolvendo todos os pares possíveis em uma janela de 30 *tokens*, com as menções ocorrendo no mesmo parágrafo. Ainda, algumas abordagens adicionais foram utilizadas para reduzir o desbalanceamento no treinamento. Sarath, Manikandan e Niwa (2017) se basearam nas heurísticas propostas por Lee *et al.* (2016) e Khalifa, Velupillai e Meystre (2016), além de utilizar *cost-sensitive learning*.

As abordagens utilizadas para extração de TLINKs entre elementos em mesma sentença envolveram: Sarath, Manikandan e Niwa (2017), usando *ensembles* de classificadores (um para TLINKs entre EVTs e outro para TLINKs entre EVTs e ETs); MacAvaney, Cohan e Goharian (2017), utilizando um classificador XGBoost; e Huang *et al.* (2017), empregando dois classificadores SVM (um para TLINKs entre eventos e outro para TLINKs entre EVTs e ETs). Sarath, Manikandan e Niwa (2017) também usaram *ensembles* de classificadores (um para TLINKs entre EVTs e outro para TLINKs entre EVTs e ETs) para TLINKs em sentenças distintas.

Leeuwenberg e Moens (2017b) classificaram de maneira conjunta tanto TLINKs em mesma sentença quanto em sentenças distintas, por meio de um modelo baseado em *Structured Perceptron*, fundamentado em Leeuwenberg e Moens (2017a).

Entre as abordagens baseadas em aprendizado de máquina ou híbridas, os melhores resultados foram obtidos por MacAvaney, Cohan e Goharian (2017) e Leeuwenberg e Moens (2017b).

3.4.1 Conclusões TLINK

A maioria das publicações relacionadas a TLINKs envolveu *corpora* disponibilizados por *shared tasks*, o que se deve primariamente à dificuldade de desenvolver *corpora* com anotações de TLINKs. Para anotar TLINKs, é necessário ter camadas de EVTs e ETs previamente anotadas, o que adiciona dois novos processos de anotação, compostos de *rounds* de refinamento do *guideline*, treinamento do anotador, própria anotação dos textos e adjudicação. Além disso, a anotação de TLINKs é complexa, como mostrado pelo baixo IAA da anotação de TLINKs em relação às anotações de EVTs e ETs nos *corpora* relacionados ao 2i2b 2012 e Clinical TempEval *shared tasks*.

Para o *corpus* i2b2 2012, as melhores abordagens relacionadas a aprendizado de máquina são representadas por Tang *et al*. (2013), D'Souza e Ng (2013, 2014a), Cherry *et al.* (2013), Xu *et al.* (2013) e Lin *et al.* (2016a). Além do aprendizado de máquina, alguns utilizaram regras. Atingir o melhor resultado para esse *corpus* envolvia criar diversos classificadores especializados para TLINKs em mesma sentença e em sentenças distintas. Para TLINKs em sentenças distintas, resultados superiores foram obtidos pelo uso de heurísticas para reduzir a quantidade de pares candidatos.

De forma similar, os melhores resultados envolvendo trabalhos referentes ao THYME *corpus*, sem considerar aprendizado profundo, envolveram diferentes classificadores para TLINKs em mesma sentença e em sentenças distintas, além de diferenciação entre TLINKs entre EVTs e TLINKs entre EVTs e EVTs. Essas abordagens foram utilizadas por Lin *et al.* (2016a, 2016b) e Lee *et al.* (2016). Além disso, Lin *et al.* (2016a) e Lee *et al.* (2016) aplicaram *cost-sensitive learning*.

4 ENCAMINHAMENTOS METODOLÓGICOS – PROCESSO DE ANOTAÇÃO

Os encaminhamentos metodológicos foram divididos em dois capítulos: (i) processo de anotação (este capítulo); (ii) processo de extração de RTs (capítulo 0). Isso ocorreu devido às características desta tese, tendo diversos processos de anotação, além de variados componentes específicos para extração de RTs, sendo, assim, oportuno separar essas questões. Primeiramente, é caracterizada a pesquisa, seguida de uma explicação geral de todas as etapas da tese, para então serem detalhadas as etapas referentes ao processo de anotação.

Esta pesquisa pode ser caracterizada como de desenvolvimento tecnológico, tendo natureza aplicada, por objetivar produzir conhecimentos para aplicação prática, dirigida à solução de problemas específicos (SILVA, 2004). Sua abordagem é quantitativa, devido à avaliação dos resultados, sendo comparada mediante técnicas estatísticas, pela observação sistemática dos resultados (KAUARK; MANHÃES; MEDEIROS, 2010).

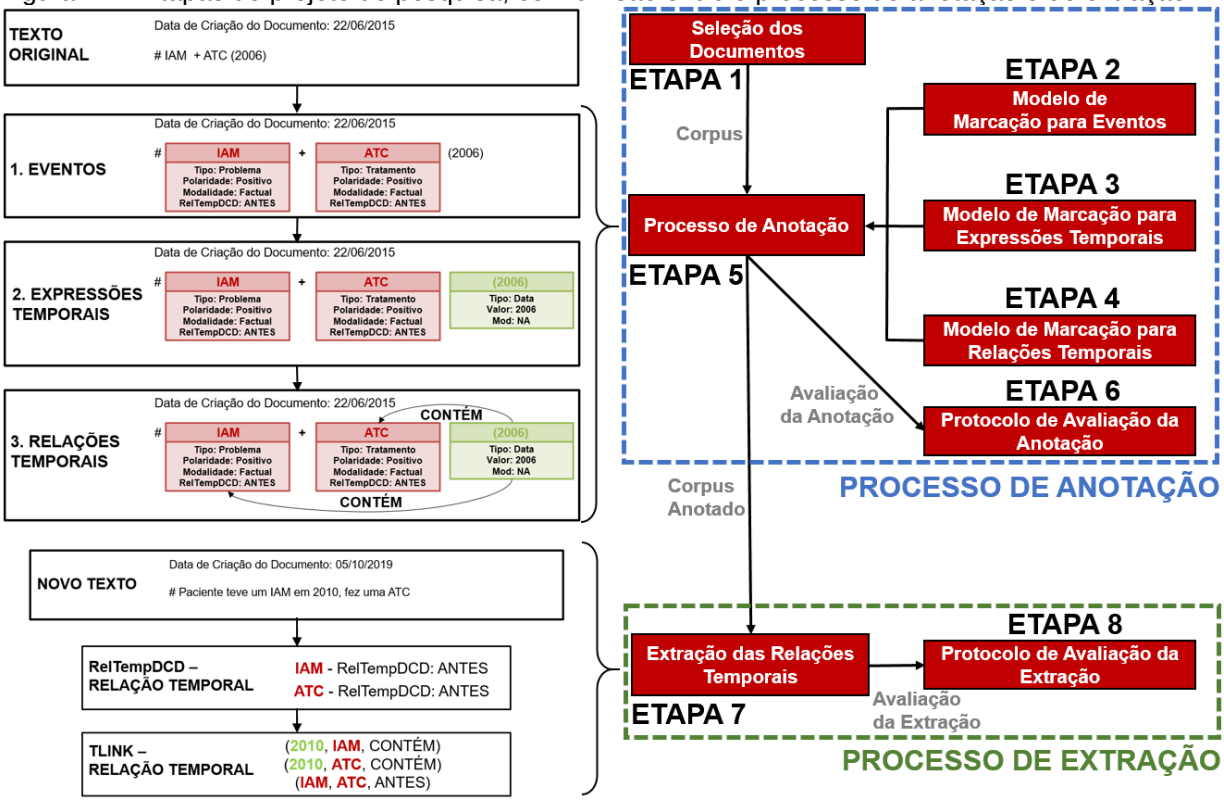
Os objetivos científicos podem ser classificados como exploratórios, visando a uma maior familiaridade com o problema, tornando-o explícito, ou à construção de hipóteses (GIL, 2002). Pode também ser classificada como uma pesquisa descritiva, expondo características de determinado fenômeno e estabelecendo correlações entre variáveis e definindo sua natureza. Os procedimentos adotados são bibliográficos e experimentais. Os primeiros são baseados em artigos publicados, livros editados, dissertações e teses apresentadas, enquanto os segundos são utilizados para determinar o resultado da extração de RTs, selecionando variáveis capazes de influenciá-las (SILVA, 2004).

Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Paraná, sob Parecer nº 1.354.675 (Anexo A). Consistiu em várias etapas, sumarizadas na Figura 17. As etapas 1 a 6 estão relacionadas ao processo de anotação (este capítulo) e as etapas 7 e 8 (próximo capítulo) são referentes ao processo de extração.

Das etapas envolvidas no processo de anotação, a etapa 1 (seção 4.1) esteve relacionada com a determinação do número de documentos para anotação, assim como especialidade médica e tipo. Um exemplo de texto sem marcações é trazido na Figura 17, na caixa Texto Original. As etapas 2 (seção 4.2), 3 (seção 4.3) e 4 (seção 4.4) envolveram criar modelos de marcação para os textos selecionados, tanto para

anotação de EVTs, quanto ETs e RTs. O *corpus* selecionado e os modelos gerados nessas etapas fazem parte da etapa 5 (seção 4.5), que engloba todo o processo de anotação, desde o refinamento incremental dos *guidelines* e anotadores até o processo de adjudicação. Na Figura 17, é possível observar o exemplo de um trecho de EVT anotado a partir do modelo criado na etapa 2, na caixa Eventos. Similarmente, na caixa Expressões Temporais, é mostrado um exemplo de trecho anotado a partir do modelo criado na etapa 3 e, na caixa Relações Temporais, de trecho anotado a partir do modelo gerado na etapa 4. A etapa 6 (seção 4.6) envolveu a avaliação das anotações previamente realizadas.

Figura 17 – Etapas do projeto de pesquisa, com divisão entre o processo de anotação e de extração.



Fonte: O autor (2020).

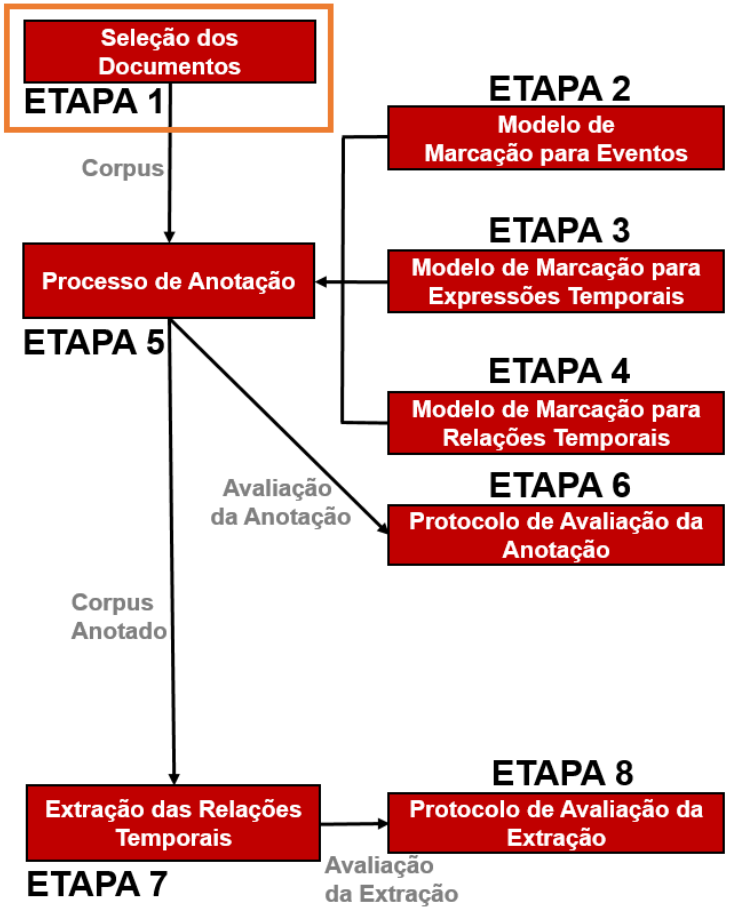
As etapas referentes ao processo de anotação serão detalhadas a seguir.

4.1 ETAPA 1 – SELEÇÃO DOS DOCUMENTOS

Esta etapa consistiu em determinar quais documentos seriam utilizados. Esse é um aspecto importante, pois, de acordo com a origem e material dos documentos, a extração de informação pode mudar de foco ou ter características distintas. Dentro

das etapas descritas anteriormente, a de seleção dos documentos está realçada na Figura 18.

Figura 18 – Etapas da pesquisa, realçando a etapa 1.



Fonte: O autor (2020).

O conjunto de dados total envolveu 2.094.929 entradas, provenientes de um conjunto de hospitais do Brasil, cada uma delas contendo dados estruturados e não estruturados; detalhes adicionados podem ser encontrados em Oliveira *et al.* (2020). Trata-se de um conjunto de dados que têm como característica conter textos de múltiplas especialidades médicas e tipos de texto clínico. Ressalta-se que, como mencionado por Oliveira *et al.* (2020), esses dados foram deidentificados previamente.

Nesta tese, foi utilizado um recorte desse conjunto de dados, de acordo com três critérios. O primeiro critério de seleção envolveu manter somente textos ambulatoriais, redigidos pelo médico durante a consulta com o paciente em ambiente ambulatorial. Diferentes tipos de texto têm características de construção, objetivos e temporalidade distintos. Por exemplo, os textos de evolução de enfermagem se referem a um período de 24 horas, sendo um dos objetivos reunir observações sobre

o estado do paciente e os procedimentos realizados nesse período (COFEN, 2016). Diferentemente deles, os sumários de alta visam à sumarização de aspectos do período de internamento. Por sua vez, os textos de ambulatório têm o objetivo de coletar informações relacionadas ao paciente, incluindo acontecimentos passados, medicações em uso, exames e sintomas relatados, para verificar o problema e planejar quais ações serão realizadas.

O segundo critério de seleção envolveu selecionar textos de uma única especialidade, a cardiológica. Como já mencionado, as doenças cardiovasculares têm forte efeito temporal; uma mesma doença pode exigir diferentes intervenções de prevenção ou tratamento em distintos momentos (JOHNSON *et al.*, 2017). Especialidades médicas têm menções de problemas, tratamentos e procedimentos de diagnósticos específicos. Um cenário de múltiplas especialidades dificulta a cobertura dos *guidelines*, traz dificuldades de anotação para casos específicos (não tão bem representados nos *guidelines*) e pode influenciar negativamente o desempenho dos algoritmos de extração de RTs, devido à baixa representatividade de certos tipos de anotação em determinadas especialidades.

O terceiro critério consistiu em selecionar somente textos com presença da DCD nos registros. Certos tipos de ET necessitam dela para normalização; em textos clínicos, é a data central no texto.

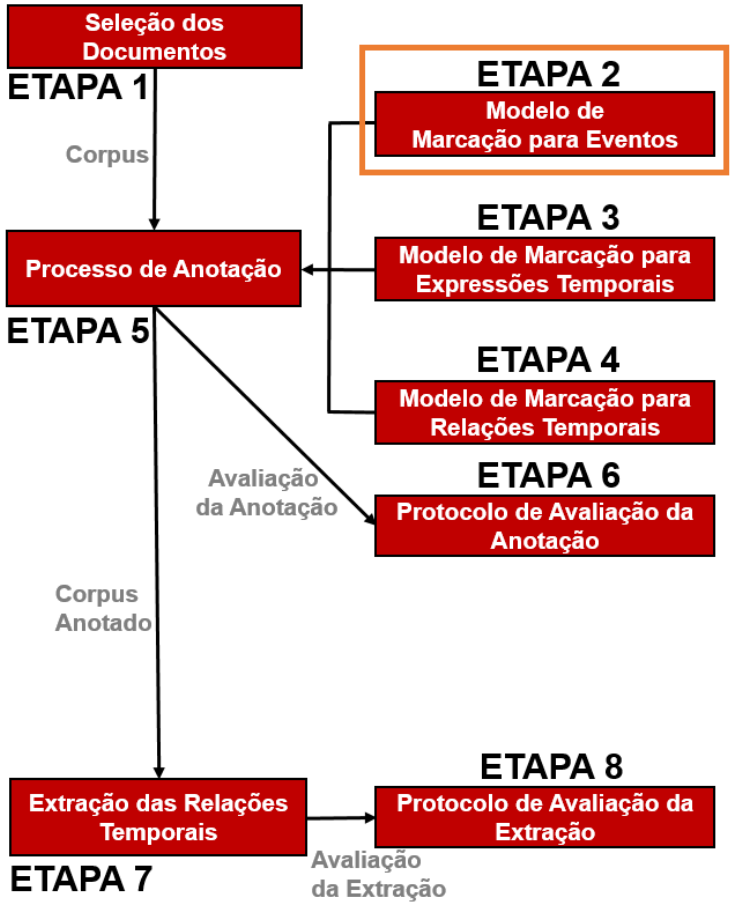
Ressalta-se que esses critérios limitaram a quantidade de textos, muitas vezes não eram satisfeitos ou não existiam formas de verificá-los, visto que informações importantes como especialidade e tipo de texto eram questões faltantes. A partir deles, foram filtrados 126 textos ambulatoriais de cardiologia, os quais formaram o *corpus* utilizado neste projeto. Esses textos selecionados representam 2.347 sentenças e 20.907 *tokens*.

Diferentes tipos de texto e especialidades têm distintas características, tanto na questão temporal quanto na questão do seu contexto de menções. Existem certos problemas médicos, tratamentos (como medicamentos) e testes (como de diagnóstico e de imagem) que são específicos para cada especialidade. Além disso, certas questões temporais são específicas da especialidade; por exemplo, no trecho “IAM + ATC”, presente em textos cardiológicos, pode-se inferir que o IAM (infarto agudo do miocárdio) veio antes de seu tratamento, a ATC (angioplastia).

4.2 ETAPA 2 – MODELO DE MARCAÇÃO DE EVENTOS

A base da anotação temporal consiste na marcação dos EVTs, que definem o que será ordenado temporalmente, além de serem menções mais numerosas nos textos em comparação às ETs. Dentro das etapas descritas anteriormente, esta de definição do modelo de marcação para EVTs está realçada na Figura 19.

Figura 19 – Etapas da pesquisa, realçando a etapa 2.



Fonte: O autor (2020).

O padrão de anotação ISO-TimeML considera EVTs como um termo de cobertura para situações que acontecem (PUSTEJOVSKY *et al.*, 2003). No domínio clínico, o EVT é definido como “qualquer menção relevante” para a *timeline* do paciente, existindo, assim, uma diferença de definição entre os dois domínios. Usualmente, “qualquer menção relevante” envolve anotações de diversos elementos, principalmente substantivos, trazendo menções como problemas, tratamentos e testes. Devido à diferente natureza dos elementos a ser marcados, padrões como o

THYME-TimeML e o utilizado no i2b2 2012 se tornam a opção ideal para usar como base de criação do *guideline*.

Essa anotação trouxe questões de ambos os padrões, fazendo adaptações para melhor atender à questão do domínio cardiológico e às características dos textos ambulatoriais. No Quadro 9, são feitas correlações dos termos utilizados na anotação, adaptados ao idioma português, a partir dos termos em inglês dos padrões THYME-TimeML e i2b2 2012.

Quadro 9 – Atributos dos EVTs em português, com seus termos originais em inglês e respectivas fontes (esquema adaptado) entre parênteses e definição para esta tese.

Atributo	**Termo original**	**Definição**
Tipo	*Type* (I2b2 2012)	Classifica o EVT em categorias mais específicas.
Polaridade	*Polarity* (I2b2 2012)	Diferencia EVTs positivos de negativos.
Modalidade	*Modality* (I2b2 2012)	Simplificado para somente diferenciar EVTs factuais de não factuais (hipotéticos ou incertos).
RelTempDCD	DocTimeRel (THYME-TimeML)	RT do EVT com a DCD.

Fonte: O autor (2020).

De forma geral, os *guidelines* do 2012 i2b2 e do THYME definem EVTs como acontecimentos importantes na *timeline* do paciente, porém essa definição é vaga, abrindo uma janela de interpretação. Nesta anotação, foi utilizado o atributo Tipo para tornar a definição mais específica, objetivando dividir os EVTs em categorias, tendo definições específicas para cada uma e, assim, melhor as conceituando. A vantagem de utilizar tipos específicos de EVT é a compreensão, com tipos intuitivos e de fácil entendimento. Além disso, diferentes tipos de EVTs têm características distintas de marcação; por exemplo, uma menção de problema pode envolver a marcação da localização, enquanto uma marcação de tratamento, a dosagem (no caso de medicamentos). Sendo assim, se torna essencial ter categorias menos abrangentes e estas, características específicas de anotação. Foram trazidas as categorias definidas pelo i2b2 2012, porém todas as definições foram adaptadas totalmente para este projeto. Para os tipos, foram trazidas definições da documentação clínica, estrutura SOAP e documentação de sintomas do paciente, assim como questões específicas para cardiologia.

Durante a criação do *guideline*, foi optado por marcações mais específicas para todos os EVTs, especialmente para menções de problemas, tratamentos e testes. No

entanto, para problemas, essa questão se torna mais complexa, principalmente quanto aos sintomas relatados pelo paciente durante a consulta. Os sintomas associados são relacionados à queixa principal e, durante a entrevista do profissional com o paciente, diversos aspectos podem ser capturados sobre eles (PEARCE *et al.*, 2016). Para dores e outros sintomas, entender características essenciais, como contexto, associações e cronologia, é essencial (BICKLEY; SZILAGYI, 2012). Elas podem ser sumarizadas como os sete atributos de um sintoma (BICKLEY; SZILAGYI, 2012; TALLEY; O'CONNOR, 2013; SWARTZ, 2020):

a) Localização: local do sintoma e padrão de irradiação (localizado ou difuso).
b) Característica: descrição do que o paciente entende pelo sintoma, sua caracterização (aguda, constante ou dor intermitente).
c) Gravidade ou severidade: descrição do quão severo ou intenso é o sintoma (para dor, pode ser utilizada uma escala de 1 a 10, por exemplo).
d) Duração: informações relacionadas ao tempo, como há quanto tempo começou, duração e frequência dos sintomas.
e) Início: cenário em que o sintoma ocorreu, incluindo fatores ambientais, atividades pessoais, reações emocionais ou outras circunstâncias que podem ter contribuído para a doença.
f) Fatores agravantes e de alívio: se existe algo que faz o sintoma melhorar (como medicamentos) ou piorar (como movimentação).
g) Manifestações associadas: quais fatores estão associados com o sintoma.

Contudo, trazer todas essas questões para a mesma marcação de EVT levaria a EVTs longos e muito específicos, podendo negativamente afetar a extração automática. Desses sete atributos, depois de discussões com especialistas, foram selecionados atributos-chave, quais sejam: localização, severidade e caracterização. Esse conceito foi estendido não só para sintomas, mas para os demais problemas, uma vez que problemas encontrados no exame físico também podem conter localização e gravidade, por exemplo. A duração foi considerada uma ET e fatores agravantes/de alívio ligados a medicamentos foram anotados separadamente, com o medicamento considerado um tratamento.

A delimitação das anotações de problemas por meio de atributos diariamente utilizados por profissionais da saúde é uma das contribuições únicas deste projeto. A caracterização de sintomas ligados à cardiologia, como a caracterização de uma

dispneia, com diferenciação entre dispneia em repouso e ao esforço, é essencial para entender o quadro do paciente.

Outro aspecto importante da anotação desta tese foi adicionar as classes funcionais definidas pelo sistema de classificação da New York Hearth Association (NYHA) às marcações, sendo essas classes específicas para textos cardiológicos. A NYHA foi escrita em 1928 e atualizada em 1994 como um método de avaliar efeitos de doenças cardíacas em pacientes durante as consultas. Varia de classe I (mais leve), de pacientes com doenças cardíacas sem limitação devido à atividade física, evoluindo a severidade até classe IV, de pacientes com doenças cardíacas com incapacidade de realizar qualquer atividade física sem desconforto (evidência de severa doença cardiovascular) (BENNETT *et al.*, 2002). Um exemplo seria a correta marcação de “IC isquêmica CF II”, indicando insuficiência cardíaca isquêmica com classe funcional II. As categorias utilizadas estão detalhadas no Quadro 10.

Quadro 10 – Categorias para cada atributo de EVT, sinalizando seu termo em inglês original (quando se aplica) e a definição criada/adaptada para este projeto.

Atributo	**Categoria**	**Termo original**	**Definição**
Tipo	Problema	*Problem*	Menções que diferem de condições normais esperadas. Exemplos: lesão, dor e HAS.
	Tratamento	*Treatment*	Menções relacionadas a qualquer procedimento ou intervenção usado para tratar problemas. Exemplos: marca-passo, angioplastia e Enalapril 10 mg.
	Teste	*Test*	Usado para detectar e avaliar problemas (como procedimentos de diagnóstico e exame físico). Exemplos: HDL, potássio, cateterismo cardíaco.
	Evidência	*Evidence*	Remete a palavras que conectam a fonte da informação àquela de interesse (geralmente uma menção de problema). Exemplo: “nega” na sentença “nega dispneia”.
	Ocorrência	*Occurrence*	Tipo abrangente, englobando diversos tipos de menção: situação do paciente, evolução do problema, conclusões sobre testes, menções relacionadas a consultas e mudanças de medicamentos. Exemplos: manter, retornar, reduzir, consulta e aumentou.
	Departamento Clínico	*Clinical Department*	Menções de departamentos clínicos referem-se a profissionais de saúde, instituições, locais, departamentos ou serviços. Exemplos: neurologista, ambulatório e médico.
Polaridade	Positiva	*Positive*	EVTs que aconteceram, estão acontecendo ou acontecerão.
	Negativa	*Negative*	EVTs que não aconteceram, não estão acontecendo ou não acontecerão.
Modalidade	Factual	-	EVTs em que existe total certeza sobre sua ocorrência ou não ocorrência.
	Não Factual	-	EVTs em que não existe total certeza sobre sua ocorrência ou não ocorrência.

RelTempDCD	Antes	*Before*	EVTs que ocorreram antes da DCD e não estão ocorrendo durante ela. Exemplos: cateterismo cardíaco, lesão, dispneia ao esforço e ataque cardíaco.
	Antes/ Sobreposto	*Before/ Overlap*	EVTs que ocorreram antes da DCD e certamente estão ocorrendo durante ela. Exemplos: DM II, ex-tabagista, marca-passo, insuficiência cardíaca e acompanhamento.
	Sobreposto	*Overlap*	EVTs que somente ocorreram durante a DCD, ou seja, aconteceram durante a consulta/encontro com o profissional de saúde. Exemplos: nega, relata, refere, edema nos membros inferiores, PA e FC.
	Depois	*After*	EVTs que irão acontecer, geralmente associados com o plano de ação do profissional de saúde. Exemplos: manter, retornar, suspender, aumentar, carvedilol 6,25 mg (medicamento prescrito) e ecg (para o próximo encontro).

Fonte: O autor (2020).

Os atributos Modalidade e Polaridade são importantes devido à diferenciação de EVTs de distintos contextos na *timeline*. A Polaridade é relevante por sinalizar se determinado EVT ocorreu, ocorre ou ocorrerá. Em uma sentença como “nega dor”, a informação que a dor “não existe” é essencial para o entendimento do texto e criação da *timeline*. Se não houvesse o atributo Polaridade, frases como “relata dor” e “nega dor” seriam marcadas com o mesmo contexto, apesar de refletirem questões distintas. Ambas as categorias do atributo Polaridade são definidas no Quadro 10.

O atributo Modalidade traz a questão da incerteza; nesta anotação, foi simplificada para diferenciação de EVTs factuais e não factuais (Quadro 10). Na anotação do i2b2 2012, por exemplo, Modalidade indica EVTs que aconteciam, eram meramente propostos, eram condicionais ou eram descritos como possíveis (SUN; RUMSHISKY; UZUNER, 2013b). Na anotação do THYME, existem categorias de Modalidade como genérica, usualmente anotadas quando o profissional de saúde justifica suas decisões e razões para mudar o cuidado (STYLER *et al.*, 2014a). As categorias de Modalidade foram simplificadas para esta anotação, sendo restritas a EVTs de cunho factual (certeza sobre a ocorrência) e não factual (EVTs cuja ocorrência está ligada à incerteza ou possibilidade). Em sentenças como “A # angina estável?”, existe certo grau de incerteza no diagnóstico, diferenciando este tipo de EVT dos demais.

O atributo RelTempDCD é de extrema importância para a anotação, por ser um tipo de RT, em que o EVT é relacionado à DCD. Em esquemas de anotação como o 2012 i2b2 e o ISO-TimeML, são criadas RTs diretas associando um EVT com uma ET

do tipo DCD. Relacionar diretamente esses itens no texto por meio de RTs é algo trabalhoso e propício a erros, até porque é necessário selecionar os EVTs desejados, um a um, e criar arcos específicos para delimitar as relações com a DCD. Em esquemas de anotação como THYME-TimeML e i2b2 2012, todo EVT tem uma relação direta com a DCD, sendo mais interessante trazer essa relação como um atributo do EVT. Em determinado contexto, o anotador pode se esquecer de criar o arco; trazer como um atributo melhora a anotação por tornar essa marcação obrigatória no EVT.

De forma geral, a RT representada pela RelTempDCD pode ser descrita como se existissem certas categorias envolvendo menções genéricas do tempo baseadas na DCD, como, por exemplo, passado (antes da DCD) e futuro (depois da DCD), nos quais os EVTs seriam “colocados”. As possíveis categorias para RelTempDCD são mostradas no Quadro 10, determinadas com base no THYME-TimeML, porém foram redefinidas para ter relação direta com o SOAP, objetivando criar um *guideline* simples de entender e que engloba as necessidades temporais cardiológicas por ser criado com base na maneira como profissionais de saúde registram seus encontros com os pacientes.

Apesar de utilizar as mesmas categorias de RelTempDCD que o THYME *corpus*, nesta tese todas as marcações são adaptadas para o contexto de textos ambulatoriais, visando a diferenciar temporalmente questões importantes relativas ao histórico do paciente, exame físico e visual, avaliação e prescrição, como diferenciar um medicamento que vem sendo usado (marcação de Antes/Sobreposto) de um medicamento futuro prescrito pelo profissional de saúde (marcação de Depois). Outro exemplo seria diferenciar problemas encontrados durante o exame físico (marcação de Sobreposto) de problemas localizados em exames de diagnóstico (marcação de Antes).

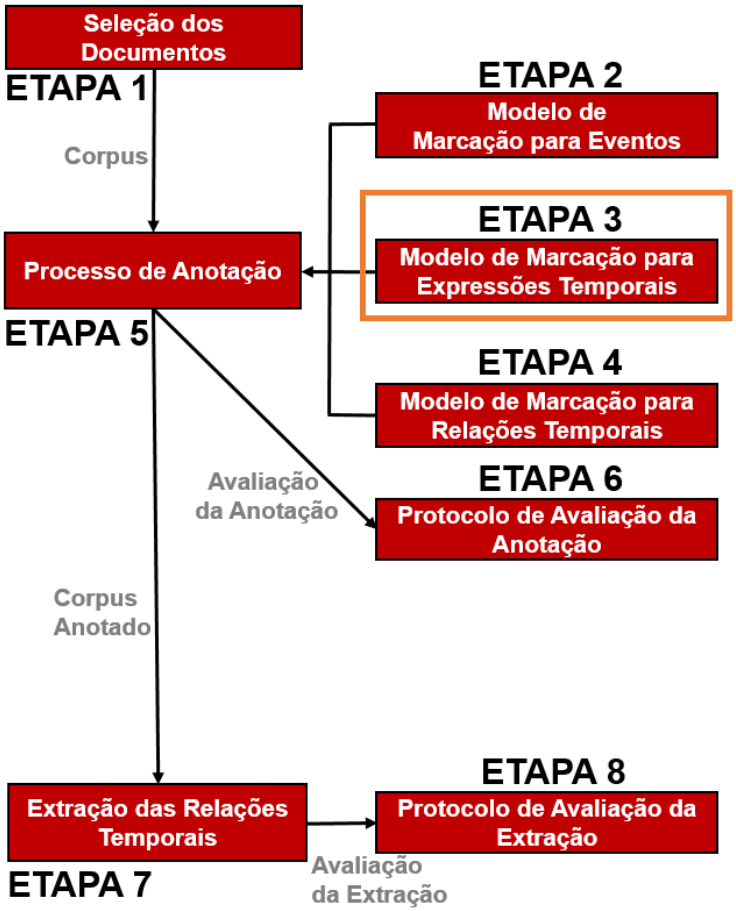
Definições e exemplos adicionais sobre a marcação de EVTs, com questões referentes à cardiologia e SOAP evidenciadas, são encontrados no Apêndice B, em um resumo do *guideline* para anotação de EVTs.

4.3 ETAPA 3 – MODELO DE MARCAÇÃO DE EXPRESSÕES TEMPORAIS

A anotação de ETs é de extrema importância na anotação de RTs, assim como de EVTs, porém, no contexto deste projeto, é necessário ter um cuidado adicional.

Essa anotação precisa considerar o *guideline* de EVTs, buscando criar marcações que façam sentido quando temporalmente relacionadas e, principalmente, que sigam um padrão de marcação conciso durante todo o processo. Dentro das etapas descritas anteriormente, esta de definição do modelo de marcação para ETs está realçada na Figura 20.

Figura 20 – Etapas da pesquisa, realçando a etapa 3.



Fonte: O autor (2020).

A definição de ET utilizada no projeto é a mesma de textos de domínio geral, determinada pelo ISO-TimeML. Uma ET pode ser tanto uma expressão de data (12/08/2008) quanto de hora específica do dia (12:24), assim como durações (por 30 minutos) e expressões se referindo à periodicidade de EVTs (2x ao dia) (STRÖTGEN; GERTZ, 2016).

Para criação do *guideline*, foram feitas adaptações no *guideline* proposto durante a anotação de Azevedo (2019), criado a partir dos padrões de anotação TimeML (SAURÍ *et al.*, 2006) e THYME-TimeML. O trabalho de Azevedo (2019) envolveu a extração e normalização de ETs de textos clínicos ruidosos escritos em

português brasileiro. ETs têm características similares no contexto do domínio geral e clínico; Styler *et al.* (2014a) observaram que as diferenças mais significativas ocorrem em casos de frequência de uso de medicamentos, com casos complexos e bem específicos, como “2 cps seg, quart, sexta e domingo; 1 cp terça, quinta e sábado”. Nesta tese, foi observado que ETs ligadas ao relato do paciente tendem a ser imprecisas, como a ET “há 10 anos” no trecho “IAM há 10 anos”, mas expressando um ponto específico no tempo, tendo um caráter incerteza quanto à data real. O conjunto de atributos resultantes, utilizado neste projeto e criado a partir de adaptações no *guideline* proposto por Azevedo (2019), é mostrado no Quadro 11.

Quadro 11 – Atributos de ETs anotados neste projeto, trazendo o termo original e sua fonte, quando necessário, assim como a definição criada/adaptada.

Atributo	Termo original	Definição
Tipo	*Type* (ISO-TimeML)	Classifica o EVT em categorias mais específicas.
Valor	*Value* (ISO-TimeML)	Valor normalizado de ET, para menções de horas e datas; é usada a ISO 8601 para normalização.
Mod	*Mod* (ISO-TimeML)	Usado quando a ET possui um modificador que altera ou melhora a interpretação do valor normalizado (FERRO *et al.*, 2001)
Quant	*Freq* (ISO-TimeML)	Valor inteiro e uma granularidade de tempo que represente a frequência com que uma ET normalmente ocorre (SAURÍ *et al.*, 2006). Representa a frequência com que uma ET regularmente ocorre dentro de um intervalo de tempo (duração), sendo utilizado somente em ETs do tipo Frequência.
Cps	-	Atributo adicionado nesta anotação para representar a quantidade de comprimidos em uma ET do tipo Frequência envolvendo medicamentos.

Fonte: O autor (2020).

Assim como na anotação de EVTs, Tipo serve para categorizar as ETs em conjunto menos abrangente, tendo definições e características distintas de anotação. Descrições gerais para cada Tipo, com definições derivadas de Strötgen e Gertz (2016), Styler *et al.* (2014a) e Saurí *et al.* (2006), são mostradas no Quadro 12, trazendo exemplos específicos para esta anotação. O termo em inglês *Set* foi adaptado para Frequência, baseando-se nos trabalhos de Baptista, Hagège e Mamede (2008) e Menezes Filho e Pardo (2011). Não foram adicionados tipos além dos padrões presentes no ISO-TimeML.

Poderiam ser considerados tipos de ET conforme outros trabalhos para o idioma português, como o padrão proposto por Baptista, Hagège e Mamede (2008) para o Segundo HAREM (MOTA; SANTOS, 2008). Esse padrão é baseado no

TimeML, trazendo diversos tipos e subtipos de ET, porém neste trabalho foi escolhido trazer um padrão já validado para o domínio clínico por meio de anotações em *shared tasks* (i2b2 2012 e Clinical TempEval 2015, 2016 e 2017), além do próprio trabalho de Azevedo (2019).

Os atributos Valor, Mod e Quant foram mantidos no padrão ISO-TimeML, conforme Quadro 11. O termo em inglês *Freq* foi adaptado para Quant, com base nos trabalhos de Baptista, Hagège e Mamede (2008) e Menezes Filho e Pardo (2011).

Quadro 12 – Categorias para cada atributo de ETs, sinalizando seu termo em inglês original (quando se aplica) e a definição criada/adaptada para este projeto.

Atributo	**Categoria**	**Termo original**	**Definição**
Tipo	Data	*Date*	Representa pontos no tempo com granularidade de dia ou maiores (mês ou ano). Exemplos: 2010, em 6 meses, 16/01/15, janeiro e desde janeiro.
	Tempo	*Time*	Refere-se a qualquer ponto no tempo com granularidade menor que dia, como alguma hora do dia. Exemplos: 12:45 e 3:30 12/05/2011.
	Duração	*Duration*	Representa informação relativa ao tamanho de um intervalo. Exemplos: há 5 anos, 40 dias, até um minuto, por 65 anos.
	Frequência	*Set*	Refere-se ao aspecto periódico de um evento, sendo relacionado com menções de frequência. Exemplos: 12/12h, /d, 1 cp ao dia, dias alternados, 2x / semana.
Valor	-	-	-
Mod	ND	NA	Não existe qualquer modificador na ET. É o valor padrão de marcação. Exemplo: há 5 meses.
	Mais	*More*	Exemplo: mais de 5 meses.
	Menos	*Less*	Exemplo: menos de 5 meses.
	Aprox.	*Approx*	Exemplo: há cerca de 5 meses.
	Começa	*Start*	Exemplo: começo de julho.
	Termina	*End*	Exemplo: final de julho.
	Metade	*Middle*	Exemplo: meio de julho.
Freq	-	-	
Cps	-	-	

Fonte: O autor (2020).

Foram realizadas algumas mudanças no *guideline* proposto por Azevedo (2019), de acordo com questões levantadas durante este trabalho, assim como características desta anotação. Essas questões serão sumarizadas na sequência.

A primeira modificação foi selecionar textos com a DCD presente, uma vez que esse aspecto impacta na questão da normalização. Menções de ETs do tipo Data e Tempo podem ser relativas, ou seja, precisam de uma referência no tempo para normalização, neste caso, a DCD. Por exemplo, na ET “há 10 anos” na frase “IAM há 10 anos”, é necessário subtrair dez anos da DCD para ter o valor normalizado. No trabalho de Azevedo (2019), houve grande contribuição na questão de criar

estratégias para definir uma DCD mais próxima possível da data real usando ETs explícitas – ET em que toda informação requerida para normalização está contida na própria menção (STRÖTGEN; GERTZ, 2016). No caso de Azevedo (2019), foi utilizada a ET do tipo Data mais recente como DCD, usualmente datas de exames de laboratórios e diagnóstico. No entanto, ao não utilizar a DCD real, existem algumas questões a serem ponderadas, tais como: (i) apesar de usar a menção de exame mais recente, não há garantia de que reflete a real DCD, com a possibilidade de o exame ter sido feito em outro período; (ii) existem textos em que todas as ETs são relativas, enquanto não há nenhuma ET explícita; (iii) menções específicas como “hoje”, amanhã” e “em 3 meses” podem acabar nunca refletindo a data real, pois sempre serão aproximadas. Por esses motivos, restringir a textos com DCD presente é importante no contexto de extração de RTs.

As demais modificações envolveram ajustes no *guideline* para melhor representar as ETs deste projeto, buscando concordância com as marcações de EVTs anteriores.

Nesta anotação, a miligramagem sempre estará marcada como parte do EVT relativo ao medicamento; já no trabalho de Azevedo (2019), a menção fazia parte da ET, devido às regras criadas para extração de frequências.

Nas marcações de ETs do tipo Frequência, foi adicionado um novo atributo para representar o número de comprimidos em uma menção de medicamento, atributo Cps (Quadro 11). Por exemplo, no trecho de texto “carvedilol 6,25 mg 2 cp 12/12”, o fato de tomar dois comprimidos (Cps) dobra a sua dosagem para 12,5 mg a cada 12 horas. É importante ter essa especificação quando esse tipo de ET é relacionado com o medicamento por meio de RTs.

Foram propostas algumas alternativas para solucionar erros de anotação devido à confusão de ETs dos tipos Data e Duração, aspecto observado nos trabalhos de Azevedo (2019) e Viani *et al.* (2019). Nas frases “infarto agudo do miocárdio há 2 anos” e “tabagismo há 2 anos”, existe a mesma ET, “há 2 anos”, utilizada com significados distintos. No primeiro caso, envolve uma Data e, no segundo, uma Duração, porém a expressão “há 2 anos” traz uma dúvida de interpretação, referente a “uma duração de 2 anos” ou “um ponto específico no tempo 2 anos atrás”. Isso se torna ainda mais grave quando se trabalha com RTs, pois, ao relacionar um EVT com uma ET, o tipo de relação inferida depende do tipo dessa expressão. Assim, foram definidas heurísticas baseadas na RelTempDCD, estabelecendo que, se o EVT

tivesse o atributo RelTempDCD com marcação Antes/Sobreposto, a ET seria marcada como Duração; se tivesse a marcação de Antes, seria Data. Como a marcação de RelTempDCD de Antes/Sobreposto indica uma continuidade, com o EVT ocorrendo desde o passado até a DCD, faz sentido marcar como Duração. De forma similar, como a marcação RelTempDCD de Antes indica um acontecimento no passado do paciente, faz sentido marcar como Data. Além de tornar clara a diferenciação de Data e Duração, a abordagem traz consistência com a RelTempDCD.

A última mudança envolveu a normalização de ETs relativas. Foram observadas certas divergências durante a anotação do *corpus* de Azevedo (2019) em relação à normalização; parte dos casos devia-se à DCD selecionada não ser completa (envolvendo dia, mês e ano), impactando, assim, na normalização. Nesta tese, toda normalização de ETs relativas foi sempre completa, contendo dia, mês e ano, visando a padronizar o processo de normalização.

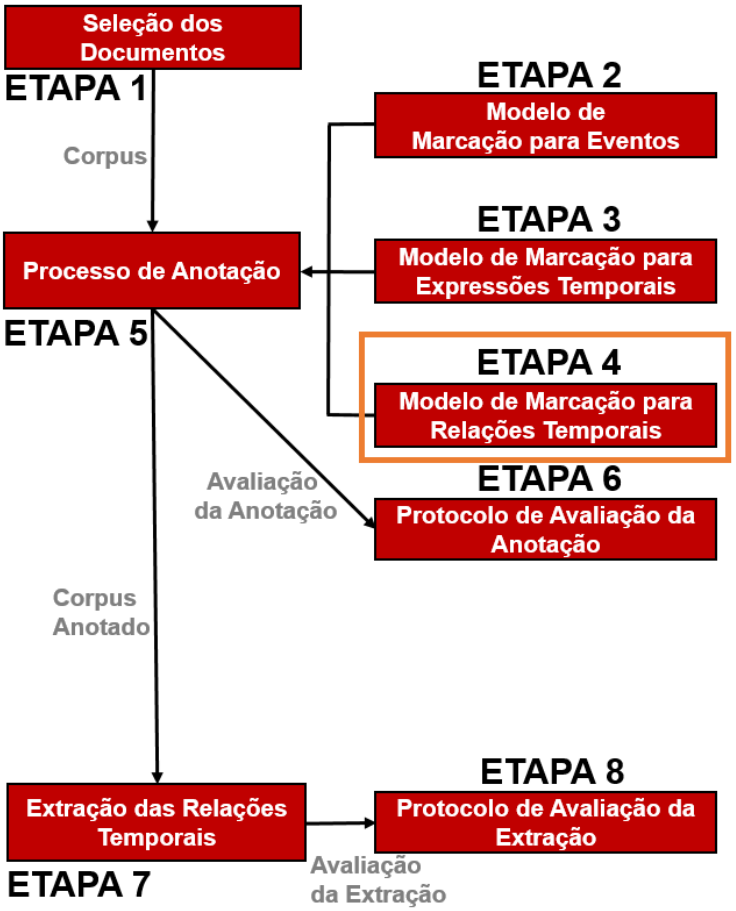
Ressalta-se que foram anotadas todas as ETs presentes no texto, indiferentemente de estarem relacionadas ou não a um EVT. Por exemplo, “2x por semana” foi marcado como ET do tipo Frequência no trecho “caminha 2x por semana”, apesar de caminhar não ser considerado um EVT.

Definições e exemplos adicionais sobre a marcação de ETs, com questões referentes à cardiologia evidenciadas, são encontrados no Apêndice C, em um resumo do *guideline* para anotação de ETs.

4.4 ETAPA 4 – MODELO DE MARCAÇÃO DE RELAÇÕES TEMPORAIS

A marcação de RTs é complexa por envolver anotações prévias, não sendo possível anotá-las sem marcações de EVTs e ETs. Um detalhe importante é a necessidade de essas anotações prévias serem consistentes e estarem de acordo com o objetivo da extração temporal. Dentro das etapas descritas anteriormente, esta de definição do modelo de marcação para RTs está realçada na Figura 21.

Figura 21 – Etapas da pesquisa, realçando a etapa 4.



Fonte: O autor (2020).

Quando se definem *guidelines* para marcações de EVTs e ETs, é preciso levar em conta quais relações se deseja expressar por essas menções. Por exemplo, no contexto da frase “tabagista há 40 anos”, pode-se definir o termo “há 40 anos” como uma ET do tipo Duração e marcar uma relação que indique a continuidade do tabagismo por um período de 40 anos. Da mesma forma, se o termo “há 40 anos” é definido como uma ET do tipo Data, precisa-se marcar uma relação que indique que o tabagismo ocorre desde determinada data no passado até o momento em questão. Dessa forma, a RT mudou totalmente a partir da marcação anterior da ET.

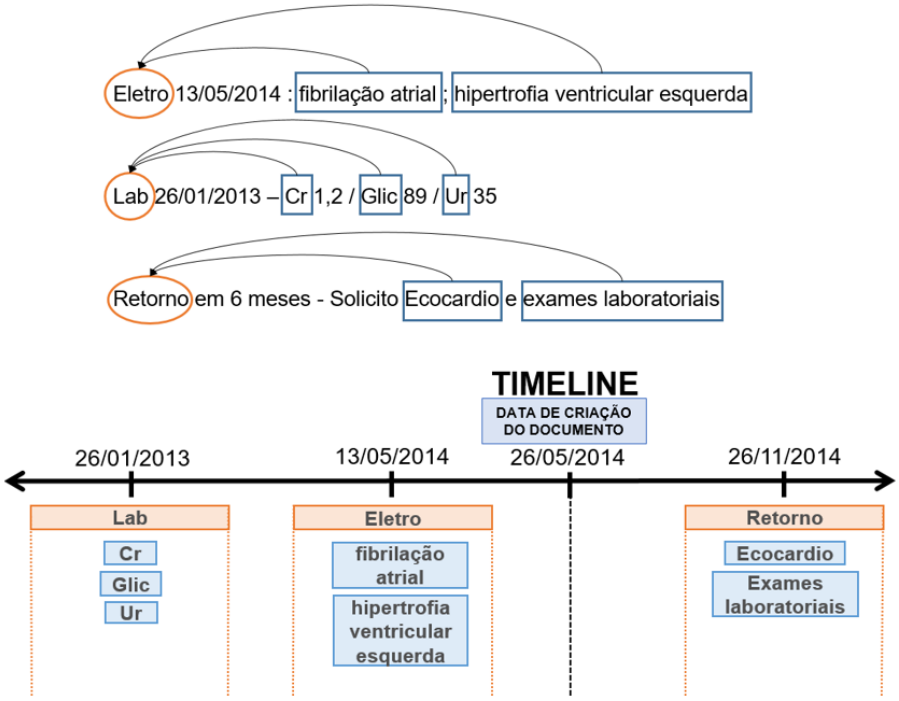
Outra questão é a não consistência das anotações. Se, em determinado contexto, como na frase “tabagista há 40 anos”, o termo “tabagista” não é marcado como um EVT, não existe RT. Entretanto, se em outro texto, em um contexto similar, o termo fosse marcado como um EVT, a RT existiria. Dessa forma, qualquer erro de anotação ou inconsistência em qualquer uma das duas anotações prévias impactaria diretamente na qualidade de anotação das RTs.

RTs do tipo RelTempDCD ajudam a representar eventos temporalmente, porém trazem consigo o problema de serem marcações envolvendo períodos amplos (aspecto detalhado nas seções 2.5.1 e 2.5.4). Por exemplo, em um contexto de “internamento há 20 anos” e “infarto agudo do miocárdio há 2 anos”, ambos os EVTs (internamento e infarto agudo do miocárdio) seriam colocados na mesma categoria de Antes da RelTempDCD, apesar de terem 18 anos de diferença.

Relacionamento entre EVTs ou de EVTs com ETs traz informações adicionais para tornar a representação mais específica. No contexto desta anotação, será utilizada a nomenclatura TLINK para esse tipo de RT, termo utilizado no padrão ISO-TimeML. Nesta tese, qualquer ET é válida para marcação de TLINKs, com exceção da DCD, que já é usada na RelTempDCD.

Para anotação de TLINKs, emprega-se a abordagem de *narrative containers* proposta por Pustejovsky e Stubbs (2011) e usada durante a anotação do THYME *corpus*. Detalhes dessa abordagem são fornecidos na seção 2.5.3. Apesar de muito da informação temporal estar presente na mesma sentença, usualmente curta, não tendo uma narrativa tão elaborada quanto textos no domínio geral, foi observado que existem certas menções centrais de EVTs e ETs em textos ambulatoriais, por isso um dos objetivos foi testar a abordagem para esse tipo de texto. Alguns desses casos são mostrados na Figura 22, sendo evidenciadas tendências de: (i) utilizar ETs do tipo Data e EVTs relacionados a menções gerais de exames de laboratório (como lab) como âncora para mencionar todos os exames laboratoriais pedidos, com seus respectivos resultados; (ii) utilizar ETs do tipo Data, em conjunto das menções desses exames de diagnóstico (como eletrocardiograma), como âncoras para mencionar todos os problemas encontrados durante esses exames; (iii) utilizar menções de retorno como âncoras para exames e demais acontecimentos referentes à próxima consulta. Ressalta-se que, no esquema de *narrative containers*, todas as menções estariam contidas (relação do tipo Contém) na menção central, fato representado na Figura 22.

Figura 22 – Exemplos de menções centrais na frase e marcações de RTs de acordo com a proposta de *narrative containers*.



Fonte: O autor (2020).

A anotação de TLINKs consiste em dois passos: o primeiro é identificar que há uma relação entre duas menções e o segundo é inferir qual tipo de RT existe. Além da mesma abordagem baseada em *narrative containers*, foram empregados os mesmos tipos de relação propostos durante a anotação do THYME *corpus*. Os tipos de marcação envolvendo TLINKs são sumarizados no Quadro 13, trazendo definições adaptadas de Styler *et al.* (2014b) para o texto deste projeto.

Quadro 13 – Tipos de TLINK, sinalizando seu termo em inglês original (quando se aplica) e a definição criada/adaptada para este projeto.

Tipo	**Termo Original**	**Definição**
Contém	*Contains*	Caso em que uma menção está totalmente contida dentro de outra. Se houver qualquer dúvida quanto a essa condição, o TLINK será marcado como Sobreposto.
Antes	*Before*	Caso em que uma menção ocorre antes de outra. Neste contexto, é necessário ter certeza de que uma das menções não ocorre mais quando a outra tem início.
Sobreposto	*Overlap*	Caso genérico representando todos os contextos em que duas menções podem ocorrer ao mesmo tempo. Se as menções tiverem qualquer tipo de conexão na *timeline*, mesmo que mínima, poderá ser marcado Sobreposto. TLINKs mais específicos, como Começa_Em, Termina_Em e Contém, podem ser generalizados para Sobreposto.
Começa_Em	*Begins_On*	Caso em que o EVT começa na menção a que está relacionado, porém essa menção não é específica o suficiente para marcar um TLINK do tipo Antes.

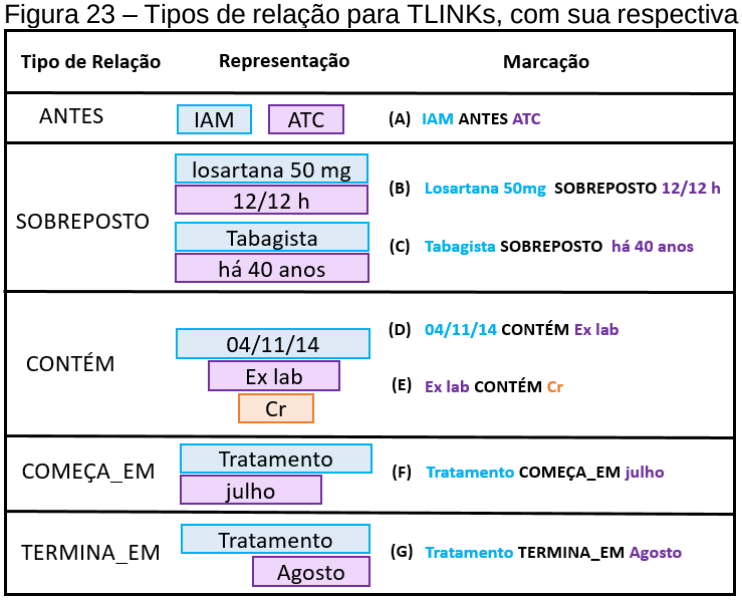
Termina_Em	*Ends_On*	Caso em que o EVT termina na menção a que está relacionado, porém essa menção não é específica o suficiente para marcar um TLINK do tipo Antes.

Fonte: O autor (2020).

Não foi feita nenhuma simplificação nos TLINKs da anotação do THYME *corpus*, mesmo tendo sido consideradas somente relações do tipo Contém para as *shared tasks*. O objetivo é anotar todos os tipos de TLINK, mesmo se existirem casos de baixo IAA ou baixo número de anotações.

Alguns exemplos de marcações de TLINKs são fornecidos na Figura 23. No exemplo A, tem-se uma relação do tipo Antes, indicando que o Problema, “IAM” (infarto agudo do miocárdio), ocorreu e foi realizado um Tratamento, “ATC” (angioplastia). Os exemplos B e C envolvem marcações do tipo Sobreposto. No exemplo B, tem-se uma ET do tipo Frequência relacionada com o medicamento (evento do tipo Tratamento). Essa marcação é muito utilizada nos textos de ambulatório, devido a pacientes terem diversas medicações em uso. No exemplo C, tem-se um EVT do tipo Problema (“tabagista”), relacionado com uma ET do tipo Duração (“há 40 anos”), indicando que o paciente é tabagista há 40 anos. Nos exemplos D e E, são mostradas marcações do tipo Contém. Neste caso, a menção geral dos exames de laboratório (“Ex lab”) está contida em determinada Data (“04/11/14”) e exames específicos, como “Cr” (creatinina), na menção geral de exame. No caso de diversos exames específicos, como nos textos de ambulatório, essa marcação se torna extremamente interessante. Nos exemplos F e G, há tipos de relação mais raros de ser usados, com os exemplos sendo relativos à frase “Tratamento de julho até agosto”. O exemplo F indica que o “tratamento” começou em “julho”, porém continua, enquanto o exemplo G indica que o “tratamento” começou antes, porém foi finalizado em “agosto”.

Figura 23 – Tipos de relação para TLINKs, com sua respectiva representação e exemplo de marcação.



Fonte: O autor (2020).

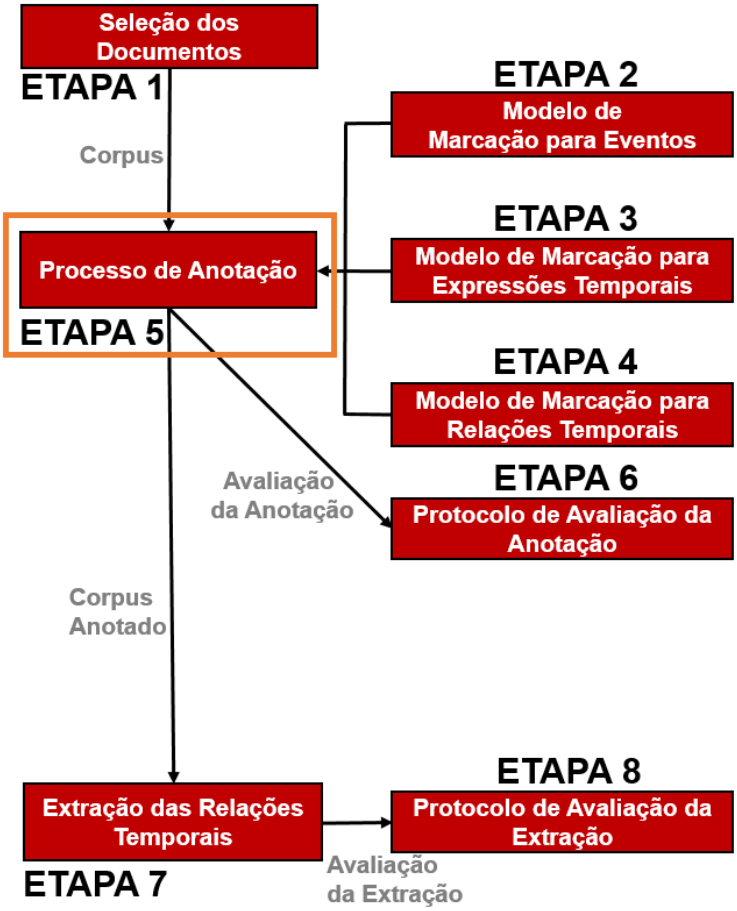
Uma das propostas do THYME, que é seguida nesta anotação, é marcar TLINKs somente quando capturam mais informações que a marcação da RelTempDCD. Esse aspecto faz com que ambas as anotações se completem e o número de TLINKs necessários a marcar seja diminuído, pois parte da informação temporal já foi representada pela RelTempDCD.

Definições e exemplos adicionais sobre a marcação de TLINKs, com questões referentes à cardiologia evidenciadas, são encontrados no Apêndice D, em um resumo do *guideline* para anotação de TLINKs.

4.5 ETAPA 5 – PROCESSO DE ANOTAÇÃO

O processo de anotação é a etapa central do projeto, pois anotações insatisfatórias podem dificultar a extração das RTs. Dentro das etapas descritas anteriormente, esta dos procedimentos envolvidos na anotação dos documentos está realçada na Figura 24.

Figura 24 – Etapas da pesquisa, realçando a etapa 5.



Fonte: O autor (2020).

Nesta tese, foram realizados três processos de anotação de forma separada. Uma estratégia similar foi utilizada na anotação do THYME *corpus*, tendo sido anotados de forma conjunta EVTs e ETs, adjudicando os documentos e só então anotando as RTs nos documentos adjudicados (STYLER *et al.*, 2014a). Na anotação do i2b2 2012, foi empregada uma abordagem de anotar as três camadas (EVTs, ETs e TLINKs) ao mesmo tempo, devido à maior eficiência na anotação, com a redução do tempo total (SUN; RUMSHISKY; UZUNER, 2013b).

Diferentemente da anotação do THYME *corpus* e do i2b2 2012, o objetivo central das anotações foi ter toda a atenção em uma camada por vez, para que, assim, as etapas de refinamento do *guideline* e treinamento do anotador fossem mais bem aproveitadas. Além disso, anotações de EVTs já adjudicadas tendem a ajudar na anotação de ETs, principalmente quando regras envolvendo EVTs são codificadas no *guideline*, sendo, assim, mais fácil anotar ETs que concordem com EVTs. Da mesma forma, ter lotes adjudicados de ETs e EVTs facilita a anotação de RTs, por ser necessário focar somente na identificação dos pares e do tipo de relação.

Para todas as camadas de anotação, foram selecionados anotadores com *expertise* do domínio da saúde. Sua escolha foi baseada nos resultados dos experimentos realizados por Roberts *et al.* (2007) e Sun, Rumshisky e Uzuner (2013a). Durante o estudo de Roberts *et al.* (2007), foi constatado que anotadores com *expertise* médica são mais propensos a achar relações entre entidades que dependem de um entendimento mais profundo do texto. Igualmente, no estudo de Sun, Rumshisky e Uzuner (2013a), foi verificado que anotadores com *expertise* médica tiveram mais sucesso em interpretar abreviações incomuns ou ambíguas, assim como achar TLINKs baseados em relações causais entre conceitos. Portanto, em textos ambulatoriais, em que existem questões implícitas, abreviaturas específicas, baixa qualidade de escrita e diversos fatores adicionais que trazem dificuldade na extração (já detalhados na seção 2.1.2), profissionais que trabalham com esses tipos de texto têm melhor compreensão das informações contidas neles.

Assim, foram selecionados alunos do curso de Medicina da Pontifícia Universidade Católica do Paraná. O projeto contou também com uma aluna de Programa Institucional de Bolsas de Iniciação Científica de Medicina, do sexto ano, que auxiliou com conhecimento específico durante a criação dos *guidelines* de EVTs e RTs, assim como participou como anotadora nessas duas etapas.

A ferramenta de anotação utilizada neste projeto foi a MAE2, apresentada anteriormente. A ferramenta foi usada durante a anotação do i2b2 2012, sendo ressaltada por Sun, Rumshisky e Uzuner (2013a) a característica de trazer todo o texto clínico na tela, o que é fundamental no caso da marcação de TLINKs que ocorrem em sentenças e seções distintas do texto. Outra característica destacada é a possibilidade da checagem rápida das relações envolvendo determinada entidade, bem como de todas as entidades associadas a determinada relação. No geral, MAE2 é uma ferramenta prática de usar, sendo somente necessário fornecer um arquivo *Document Type Definition* (DTD) para anotação de uma tarefa.

Um arquivo DTD define qual será a estrutura do documento XML, trazendo informações sobre quais rótulos serão utilizados, assim como quais atributos estão relacionados com cada rótulo (PUSTEJOVSKY; STUBBS, 2012). Na Figura 25, são mostrados os padrões DTD usados neste projeto. O primeiro trecho do arquivo (sinalizado pelo número 1) envolve a anotação de EVTs. Neste caso, foi delimitada a marcação do EVT; pelo elemento “#PCDATA”, foi indicada a entrada de dados textuais. Todos os atributos do EVT foram conectados à menção do EVT pelo trecho

“ATTLIST”. Para cada atributo, todas as possíveis marcações foram fornecidas entre parênteses e separadas por um delimitador “|”. Para anotação de ETs, o padrão de especificação foi similar ao de EVTs, residindo a diferença em atributos como Valor, Quant e Cps receberem valores textuais, sinalizados por “#PCDATA”. Na anotação de ETs, foram fornecidas ao anotador as anotações de EVTs adjudicadas. Sendo assim, os trechos 1 e 2 foram mantidos no DTD para anotação de ETs. No trecho 3, é mostrado o trecho do arquivo DTD responsável pela anotação de TLINKs, sendo mantidos os trechos 1 e 2 de DTD para fornecer ao anotador as anotações de ETs e EVTs adjudicadas; sem isso, não seria possível marcar relações entre EVTs e ETs. As marcações de TLINKs consistiram em criar uma relação vazia, adicionar dois elementos (EVTs ou ETs) e marcar sua relação entre um dos tipos predefinidos.

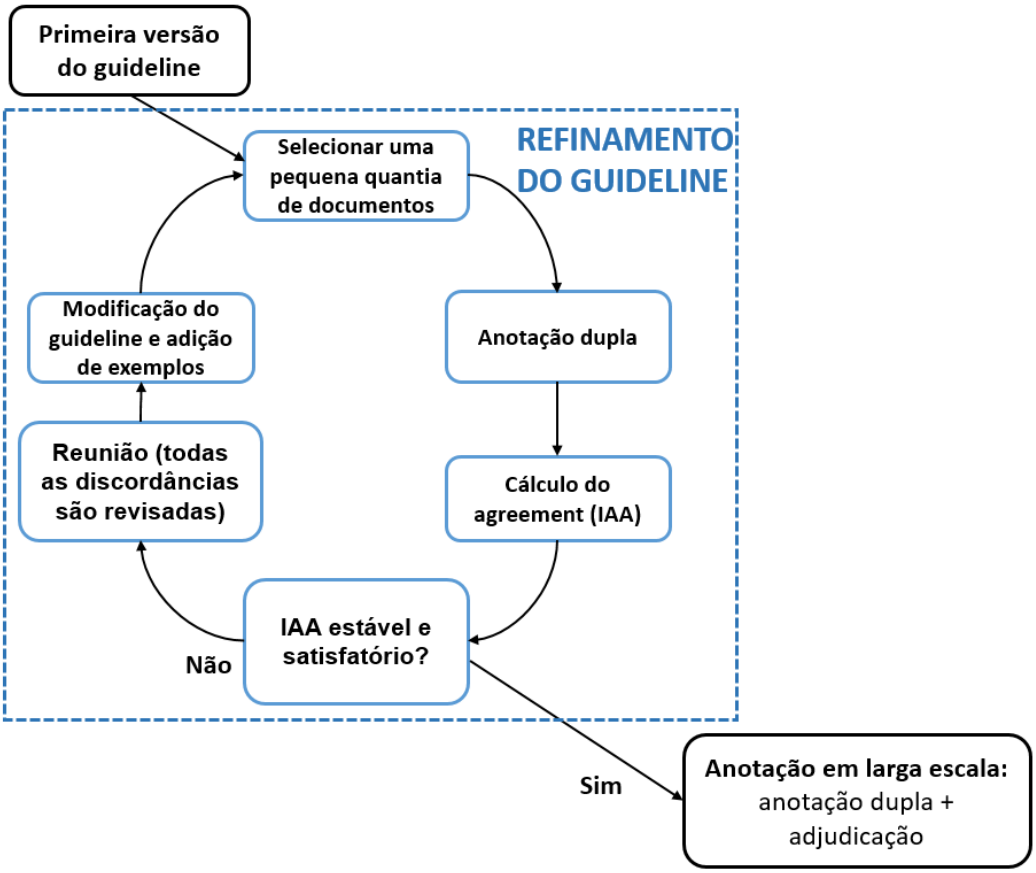
Figura 25 – Padrão DTD utilizado para anotação de RTs, sinalizando os componentes utilizados para anotação de EVTs (1), ETs (2) e RTs (3).

```
<!ENTITY name "Annotation_v1">

<!-- ~ Annotation DTD ~ EVENTOS + EXPRESSÕES TEMPORAIS + RELAÇÕES TEMPORAIS -->

<!ELEMENT EVENTO ( #PCDATA ) >
<!ATTLIST EVENTO Tipo (Teste | Problema | Tratamento | Departamento Clínico | Evidência | Ocorrência) #IMPLIED >
<!ATTLIST EVENTO Polaridade (Positiva | Negativa) # IMPLIED >
<!ATTLIST EVENTO Modalidade (Factual | Não-Factual) # IMPLIED >
<!ATTLIST EVENTO RelTempDCD (ANTES | DEPOIS | SOBREPOSTO | ANTES/SOBREPOSTO) # IMPLIED >
1

<!ELEMENT TIMEX3 ( #PCDATA ) >
<!ATTLIST TIMEX3 Tipo (Data | Tempo | Duração | Frequência) # IMPLIED >
<!ATTLIST TIMEX3 Valor CDATA # IMPLIED >
<!ATTLIST TIMEX3 Mod (MAIS | MENOS | APROX | COMEÇO | FINAL | METADE | NA ) # IMPLIED "NA">
<!ATTLIST TIMEX3 Quant CDATA # IMPLIED >
<!ATTLIST TIMEX3 Cps CDATA # IMPLIED >
2

<!ELEMENT RELATION EMPTY >
<!ATTLIST RELATION Relação_Temporal ( CONTÉM | SOBREPOSTO |  COMEÇA_EM | TERMINA_EM | ANTES ) #
REQUIRED >
3
```

Fonte: O autor (2020).

O ciclo de refinamento utilizado para este projeto foi uma adaptação do ciclo de anotação proposto pelo CLEF, detalhado anteriormente. O ciclo adaptado é mostrado na Figura 26.

Figura 26 – Processo de refinamento do *guideline* para todas as camadas de anotação.



Fonte: O autor (2020).

Todas as camadas de anotação seguiram o mesmo ciclo de refinamento do *guideline*. A partir da versão inicial deste, foram passados documentos para anotação, por meio de anotação dupla, sendo então calculados os valores de IAA. A diferença desta tese é que, além de verificar se são estáveis, foi observado se os valores são satisfatórios. A condição de valores satisfatórios envolve a obtenção de valores de IAA semelhantes ou superiores aos de IAA obtidos na anotação do i2b2 2012 e Clinical TempEval 2016 (o *corpus* da edição de 2016 é mais utilizado para pesquisa). Assim, uma anotação foi considerada estável e satisfatória se apresentou valores de IAA superiores aos obtidos nos *corpora* de ambas as *shared tasks* por dois lotes.

Em casos em que IAA não era estável e satisfatório, foram propostas reuniões para revisar todas as discordâncias daquele lote de treino. Para isso, foi implementada uma função nas ferramentas de avaliação da anotação para trazer, em formato de tabela, todos os acertos, acertos parciais (quando relevante) e discordâncias. Um exemplo da tabela gerada para anotação de EVTs é mostrado na Figura 27, evidenciando duas discordâncias na marcação do atributo RelTempDCD. Vale salientar que todos os atributos foram mostrados na tabela, porém, para fins de

demonstração, somente o trecho marcado e o atributo RelTempDCD são representados na figura. Após esta etapa de verificação das anotações, foram feitas as modificações no *guideline* e adicionados novos exemplos a partir das questões levantadas na reunião. Esse ciclo se repetiu até as condições do IAA impostas serem satisfeitas.

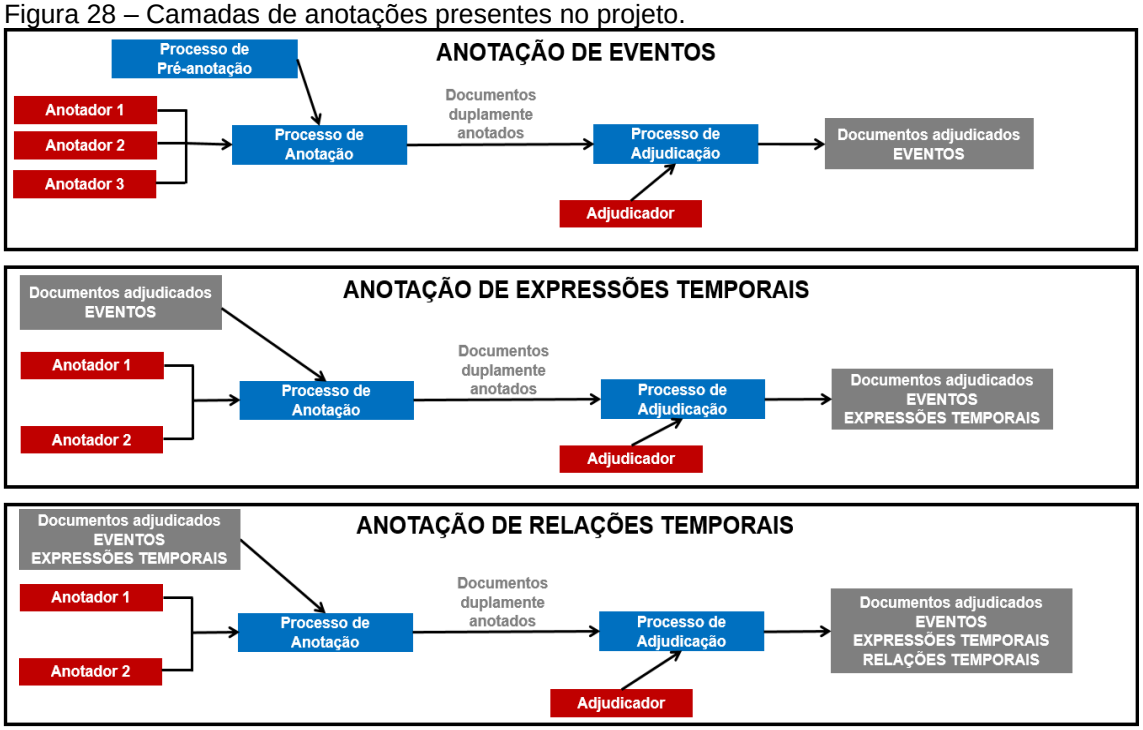
Figura 27 – Função da ferramenta de cálculo do IAA para trazer acertos, acertos parciais e discordâncias, evidenciando duas discordâncias de anotação no exemplo.

TEXTO	ACERTO DCD	ACERTO MOD	ACERTO POL	ACERTO TRECHO	ACERTO TIPO	ann1_Trecho	ann2_Trecho	ann1_DCD	ann2_DCD
9617	1	1	1	1	1	Comorbidades	Comorbidades	Antes/Sobreposta	Antes/Sobreposta
9617	1	1	1	1	1	relata	relata	Sobreposta	Sobreposta
9617	0	1	1	1	1	HAS	HAS	Sobreposta	Antes/Sobreposta
9617	1	1	1	1	1	Medicamentos	Medicamentos	Antes/Sobreposta	Antes/Sobreposta
9617	1	1	1	1	1	Renagel 800 mg	Renagel 800 mg	Antes/Sobreposta	Antes/Sobreposta
9617	1	1	1	1	1	ECO	ECO	Depois	Depois
9617	1	1	1	1	1	retorno	retorno	Depois	Depois
9617	1	1	1	1	1	valvuloplastia mitral e	valvuloplastia mitral e	Antes	Antes
9617	1	1	1	1	1	IRC	IRC	Antes/Sobreposta	Antes/Sobreposta
9617	1	1	1	1	1	diálise	diálise	Antes/Sobreposta	Antes/Sobreposta
9617	1	1	1	1	1	transplante renal	transplante renal	Depois	Depois
9617	1	1	1	1	1	avaliação cardiológica	avaliação cardiológica	Sobreposta	Sobreposta
9617	1	1	1	1	1	transplante renal	transplante renal	Depois	Depois
9617	1	1	1	1	1	Dopença renal policística	Dopença renal policística	Antes/Sobreposta	Antes/Sobreposta
9617	0	1	1	1	1	diabetes	diabetes	Sobreposta	Antes/Sobreposta
9617	1	1	1	1	1	Cabornato de cálcio 500mg	Cabornato de cálcio 500mg	Antes/Sobreposta	Antes/Sobreposta

Fonte: O autor (2020).

Após os processos de refinamento do *guideline* e treinamento dos anotadores serem finalizados, foram realizadas as anotações em larga escala. Os 126 textos foram divididos em 14 lotes, totalizando nove documentos por lote de anotação. O projeto de anotação, como um todo, é sumarizado na Figura 28.

Figura 28 – Camadas de anotações presentes no projeto.



Fonte: O autor (2020).

O processo de anotação de EVTs contou com três anotadores, os quais participaram do processo de treinamento e refinamento do *guideline*. Na anotação em larga escala, foi anotado um lote por semana, considerando a dupla anotação por três anotadores, com cada um anotando seis textos por semana.

Foi desenvolvida uma ferramenta de pré-anotação de EVTs para diminuir o trabalho, fornecendo determinadas marcações do EVT, com seus respectivos tipos. As pré-anotações foram baseadas em um dicionário construído ao longo de todas as rodadas de anotações de refinamento do *guideline* e das quatro rodadas de anotação em larga escala. Todas as anotações consideradas corretas foram adicionadas ao dicionário, assim como expressões regulares para incluir a dosagem nas pré-anotações de medicamentos. Nesta tese, não poderia existir um mesmo trecho de texto usado para marcações de diferentes EVTs, não havendo qualquer tipo de sobreposição de entidades. Sendo assim, foram desenvolvidas regras para assegurar que a marcação de determinado evento fosse sempre a mais abrangente. Por exemplo, na frase “relata dispneia ao esforço”, se os termos “dispneia” e “dispneia ao esforço” estivessem contidos no dicionário, a ferramenta deveria automaticamente marcar “dispneia ao esforço” como Problema.

As ferramentas de pré-anotação são valiosas em um cenário ambulatorial, porque há menções específicas (por exemplo, sintomas, medicamentos e exames

laboratoriais) recorrentes em textos de cardiologia. Isso pode ser exemplificado pelos exames de rotina definidos no protocolo para pacientes com hipertensão arterial da 7º Diretriz Brasileira de Hipertensão Arterial (MALACHIAS *et al.*, 2016). Os testes de rotina são: análise de urina, potássio plasmático, glicemia em jejum, ritmo de filtração glomerular estimado, creatinina plasmática, colesterol total, HDL-c, triglicérides plasmáticas, ácido úrico plasmático e eletrocardiograma. Assim, sempre que um paciente com hipertensão faz uma consulta de rotina, recomenda-se pelo menos solicitar esses exames.

Foram fornecidos aos anotadores os lotes já com os eventos pré-anotados. Os anotadores foram livres para fornecer suas próprias anotações e para aceitar, modificar ou excluir as pré-anotações. Após a anotação dos 14 lotes, foram gerados 126 textos duplamente anotados, que passaram por um processo de adjudicação. Foi seguido o mesmo procedimento da anotação do CLEF *corpus*, detalhado em Roberts *et al.* (2009), em que o adjudicador não pode contestar/anular anotações concordantes entre ambos os anotadores, podendo somente resolver discordâncias de anotação. O doutorando foi responsável pela adjudicação desses lotes, sendo gerado o padrão ouro da anotação de EVTs. Em seguida, foi realizada a anotação de ETs, como mostrado na Figura 28.

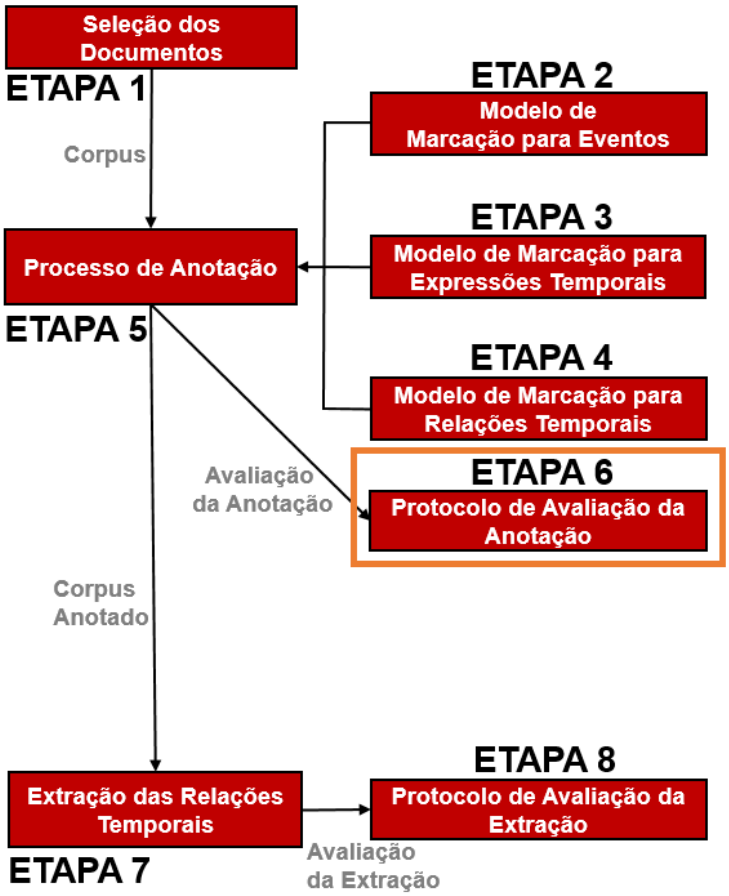
O padrão ouro gerado pela anotação de EVTs foi fornecido para dois anotadores, que não participaram da anotação da etapa anterior. Um deles era o próprio doutorando e o outro, uma estudante de Medicina envolvida em diversos processos de anotação em textos clínicos, tendo sido, inclusive, anotadora durante a anotação de ETs no projeto de Azevedo (2019). Por já ter participado de uma anotação similar, foram necessários poucos lotes para um valor de IAA satisfatório. Ao final da anotação, foram gerados 126 textos duplamente anotados, sendo o doutorando responsável pela adjudicação. Após esta etapa, foi gerado o padrão ouro de ETs. Após essas etapas, o padrão ouro conteve anotações de EVTs e ETs.

O último passo foi a anotação de TLINKs. Foi mantida uma anotadora participante da anotação de EVTs e selecionada uma nova. Após o procedimento de treinamento e refinamento do *guideline*, os 126 textos foram duplamente anotados e adjudicados pelo doutorando. Após essa etapa, foi gerado o padrão ouro contendo todas as camadas de anotação (EVTs, ETs e TLINKs).

4.6 ETAPA 6 – PROTOCOLO DE AVALIAÇÃO DA ANOTAÇÃO

A avaliação das anotações de EVTs, ETs e RTs envolveu o cálculo dos valores de IAA para cada camada de anotação manual. Dentro das etapas descritas anteriormente, esta de descrição da avaliação das anotações está realçada na Figura 29.

Figura 29 – Etapas da pesquisa, realçando a etapa 6.



Fonte: O autor (2020).

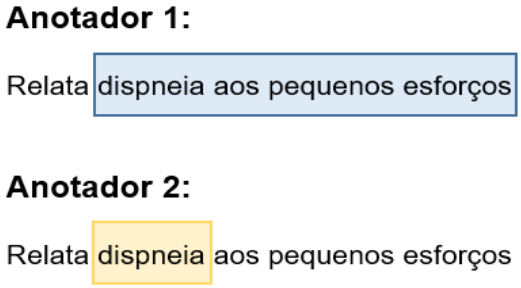
Para avaliação da anotação de EVTs e ETs, houve a avaliação da marcação no texto (*span*) e dos atributos. Para o *span*, foi utilizado o F1-*score* entre dois anotadores, sendo uma métrica comprovada pelo estudo de Hripcsak e Rothschild (2005). As equações de *Precision* (Equação 4.1A), *Recall* (Equação 4.1B) e F1-*score* (Equação 4.1C) envolveram Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (4.1)$$

(A) (B) (C)

O F1-*score* foi utilizado como métrica de avaliação das anotações do THYME e i2b2 2012, sendo, assim, aplicado neste projeto. Os valores de IAA foram reportados tanto no cenário de *exact matching* quanto de *partial matching*. No primeiro cenário, os anotadores concordaram em uma anotação apenas quando ambos marcaram o mesmo intervalo de texto para a menção; no segundo, concordaram em uma anotação quando ambos marcaram extensões de textos que se sobrepuseram de alguma forma. Na Figura 30, é mostrado um exemplo para marcação de EVTs; se um anotador marcasse “dispneia aos poucos esforços” como EVT e outro, “dispneia” somente, seria um parcial no *partial matching*, porém uma discordância no *exact matching*. No *partial matching*, o anotador recebeu “meio acerto” por cada acerto parcial.

Figura 30 – Exemplo de acordo parcial entre anotadores.



Fonte: O autor (2020).

Para cálculo dos valores de IAA para os atributos dos EVTs (Tipo, Polaridade, Modalidade e RelTempDCD) e ETs (Tipo, Valor, Mod, Freq e Cps), foi seguido o protocolo utilizado no i2b2 2012, empregado durante a anotação do TimeBank (PUSTEJOVSKY *et al.*, 2006), em que o IAA para cada atributo envolve o cálculo da *accuracy*, mostrada na Equação 4.2, considerando somente como VP menções que tiveram *match* na marcação do texto.

$$Accuracy = \frac{\mathrm{VN} + \mathrm{VP}}{\mathrm{VN} + \mathrm{VP} + \mathrm{FN} + \mathrm{FP}} \qquad (4.2)$$

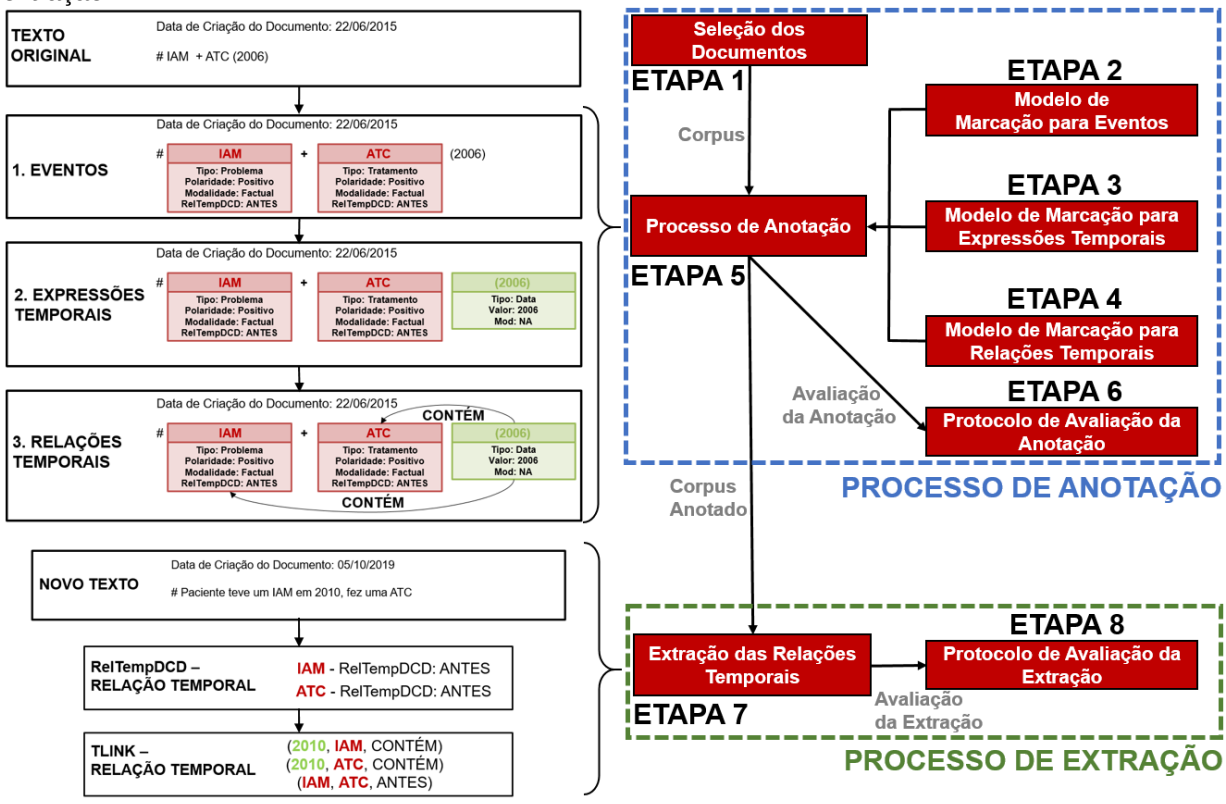
Nas anotações de ETs, foi adicionada uma regra na avaliação do IAA para menções do tipo Duração e Frequência, seguindo os mesmos passos da avaliação do i2b2 2012. A normalização de menções do tipo Duração foi indicada por expressões “P” (período) e “PT” (período em horas/minutos), além das unidades Y = ano, M (em marcações P) = mês, D = dia, H = hora, M (em marcações PT) = minuto. Sendo assim, poderiam existir casos de anotação em que o valor anotado por um dos anotadores seria “P1D” e “PT24H” pelo outro. Apesar de terem o mesmo conteúdo, a normalização é expressa de forma distinta, com unidades diferentes. Nesses casos de unidades distintas, o código de avaliação do i2b2 2012 converte todas as normalizações para a unidade hora. O mesmo aspecto foi adicionado para cálculo do IAA neste projeto.

Na anotação do i2b2 2012, o IAA para TLINKs não leva em conta os EVTs e ETs já adjudicados, uma vez que todas as camadas de anotação são feitas simultaneamente; sendo assim, todas as discordâncias e anotações parciais influenciam o IAA de TLINKs. Nesta tese, existe uma similaridade maior com a anotação do THYME, em que todos os lotes de EVTs e ETs são adjudicados, para então ser marcados os TLINKs. Por isso, para avaliação das anotações, foi utilizada a mesma estratégia do THYME *corpus*, calculando o F1-*score*. Só foi considerado um acerto da anotação quando ambas as menções e o tipo de relação marcado entre elas eram iguais.

5 ENCAMINHAMENTOS METODOLÓGICOS – PROCESSO DE EXTRAÇÃO

Das etapas relacionadas ao processo de extração de RTs, a sétima (seção 5.1) esteve relacionada com a treinamento e teste dos algoritmos para extração de RTs. Ao fim dela, foram gerados modelos que, a partir de textos com anotações prévias de EVTs e ET, extraíssem TLINKs e RelTempDCD. Todo esse processo é exemplificado na Figura 31. Nesta tese, não é realizada a extração de EVTs e ETs, sendo considerados os valores presentes nas anotações. Na etapa 8 (seção 5.2), foi realizada a avaliação da extração de RTs, verificando o desempenho para extração de TLINKs e RelTempDCD.

Figura 31 – Etapas do projeto de pesquisa, sinalizando as relacionadas ao processo de anotação e de extração.



Fonte: O autor (2020).

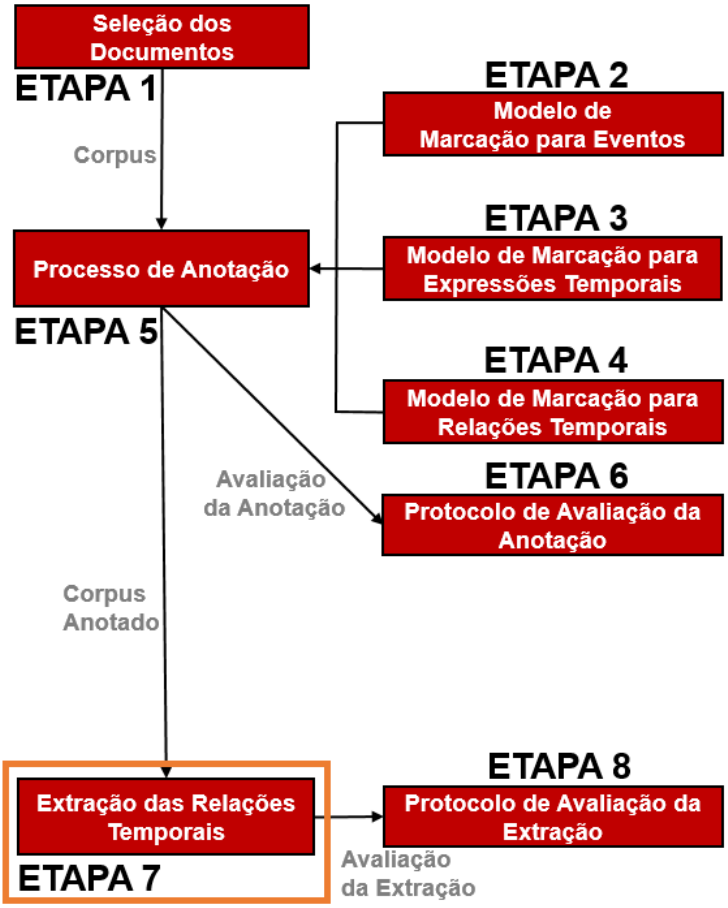
As etapas referentes ao processo de extração serão detalhadas a seguir.

5.1 ETAPA 7 – EXTRAÇÃO DE RELAÇÕES TEMPORAIS

A etapa de extração de RTs envolve diversas questões, como classificadores especializados, desenvolvimento de heurísticas para limitar a quantidade de pares

candidatos e engenharia de *features*. Das etapas descritas anteriormente, esta está realçada na Figura 32.

Figura 32 – Etapas da pesquisa, realçando a etapa 7.



Fonte: O autor (2020).

A extração de RTs foi dividida nas tarefas de extração de RelTempDCD e TLINK, foco desta seção.

A extração de RelTempDCD já é bem resolvida na literatura, com resultados de extração para o *corpus* Clinical TempEval 2016 já excedendo os valores de IAA. Na verdade, RelTempDCD são menos complexas de extrair do que TLINKs, por não sofrerem com o problema do severo desbalanceamento do conjunto de dados durante o treinamento, que se deve à necessidade da geração de pares candidatos. Tipicamente, um EVT é conectado a uma DCD; nos piores casos, como no i2b2 2012, pode ser necessário selecionar uma dentre duas DCDs (data de admissão ou alta). Na extração de TLINKs, um EVT pode ser conectado a qualquer outro EVT ou ET na mesma sentença ou ainda em sentenças distintas. Quando as menções estão em sentenças distintas, é usado um critério para delimitação de pares, tanto baseado em

uma janela de sentenças tanto em uma janela de *tokens*. Em ambos os casos, considerando TLINKs em mesma sentença e em sentenças distintas, existe um desbalanceamento severo no conjunto de dados durante o treinamento.

Na Figura 33, é exemplificada a questão das dificuldades encontradas durante a criação de pares candidatos. Considerando somente a sentença “IAM com ATC em 2013”, existem três pares positivos (marcados em verde), enquanto três pares negativos são gerados. No caso de considerar sentenças distintas, uma sentença com apenas um EVT gerou mais seis pares, todos negativos. De uma proporção de 1:1 (1 par positivo para cada negativo), houve uma mudança para 1:4 somente ao considerar pares em relação a um EVT adicional. Vale ressaltar que, neste caso, é uma sentença curta, porém em sentenças longas a quantidade de pares negativos aumenta proporcionalmente.

Figura 33 – Exemplo de pares gerados para RTs entre elementos em mesma sentença e em sentenças distintas.

Fonte: O autor (2020).

Devido à questão do desbalanceamento e às diferentes características das relações, são usualmente criados modelos especializados. Uma divisão utilizada por diversos autores foi separar os TLINKs entre relações entre EVTs e entre EVTs e ETs. Alguns dos melhores resultados para o i2b2 2012 *corpus* e os *corpora* relacionados ao Clinical TempEval (baseados no THYME *corpus*) utilizaram essa abordagem, a saber: Lin *et al.* (2016a, 2016b, 2017), Tang *et al.* (2013), D’Souza e Ng (2014), Lee *et al.* (2016, 2018), Tourille *et al.* (2017a) e Dligach *et al.* (2017). Além disso, TLINKs entre EVT e ET sofrem menos impacto do desbalanceamento pelo menor número de ETs nos textos, diminuindo a quantidade de possíveis pares. Dados esses motivos, neste projeto foi realizada a mesma divisão.

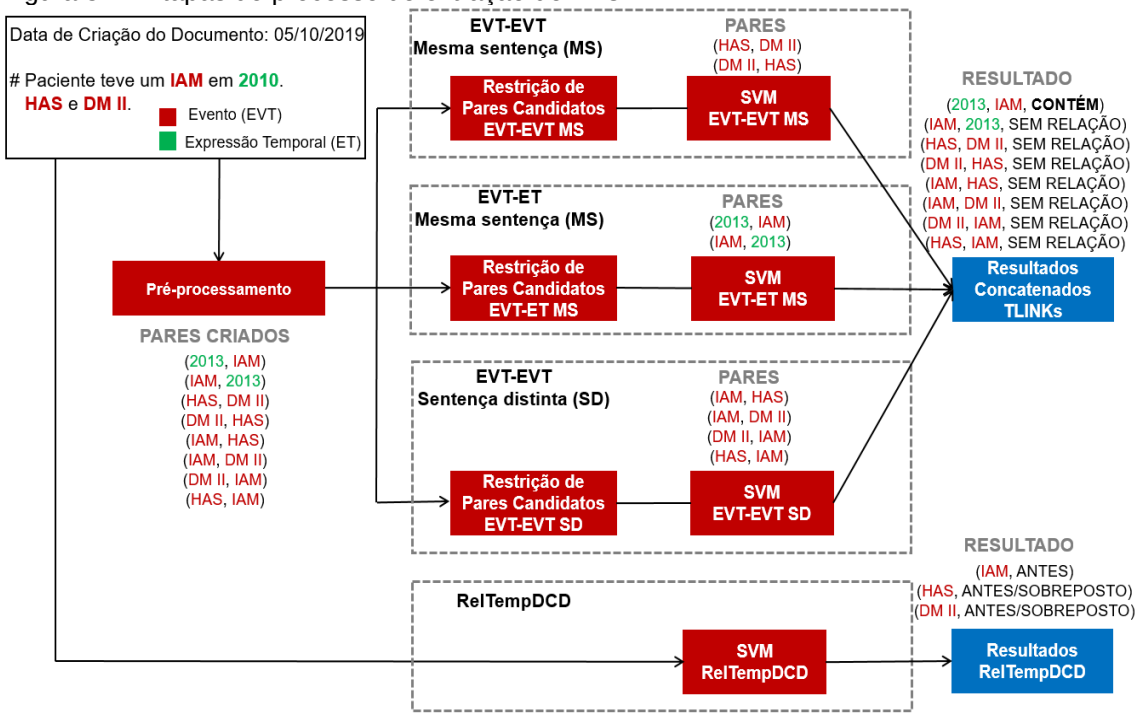
Também é comum autores fazerem diferenciações adicionais entre TLINKs envolvendo menções em mesma sentença e em sentenças distintas. TLINKs com menções em diferentes sentenças são complexos, pois, além de a quantidade de pares positivos diminuir com o aumento das sentenças consideradas, o número de possíveis pares cresce rapidamente quanto mais sentenças são consideradas.

Alguns autores desenvolveram abordagens considerando somente relações entre elementos na mesma sentença, descartando completamente sentenças distintas. Essa abordagem foi empregada em alguns dos melhores resultados envolvendo os *corpora* relacionados ao Clinical TempEval, incluindo: Dligach *et al.* (2017), Lin *et al.* (2016b, 2017, 2018), Galvan *et al.* (2018), MacAvaney, Cohan e Goharian (2017) e Tourille *et al.* (2017c). Contudo, ao considerar essa abordagem, certos TLINKs importantes em sentenças distintas são perdidos. Como observado por Tourille *et al.* (2017c), no *corpus* Clinical TempEval 2016, perto de 76% dos exemplos positivos de treinamento ocorrem na mesma sentença. Sendo assim, nesse cenário são perdidas 24% das relações, sem ao menos os sistemas desenvolvidos tentarem classificar corretamente. Por esse motivo, nesta tese o objetivo foi trabalhar com relações em sentenças distintas, buscando ampliar a cobertura.

No conjunto de treinamento do *corpus* anotado para este projeto, cerca de 95% dos pares positivos de TLINKs entre EVT e ET ocorreram na mesma sentença, existindo poucos casos de relações em sentenças distintas. Assim, somente relações entre EVTs foram consideradas em um contexto de sentenças distintas.

Todos os componentes para extração de RTs são sumarizados na Figura 34. De forma geral, a tarefa de extração de RTs foi dividida em quatro distintas categorias de tarefa. A primeira etapa consistiu na criação dos pares candidatos, ou seja, gerar os pares para treinamento e teste dos algoritmos. Na figura, são mostrados os pares candidatos gerados para duas sentenças. Primeiramente, os pares somente foram gerados, porém nos passos consequentes foram separados em suas respectivas categorias. Vale ressaltar que RTs do tipo RelTempDCD não necessitaram de criação de pares ou heurísticas, somente do treinamento do classificador.

Figura 34 – Etapas do processo de extração de RTs.



Fonte: O autor (2020).

Na Figura 34, verifica-se que cada categoria de TLINK tem suas próprias heurísticas para restrição dos pares candidatos e classificador especializado. Além deste, toda a engenharia de *features* era personalizada para cada categoria, sendo levantados autores para cada uma delas. Com o final da extração, o resultado dos classificadores individuais foi concatenado e obtida uma métrica geral para classificação de TLINKs. Na figura, percebe-se que todos os pares criados foram classificados com algum tipo de relação, quase todos sem relação.

A estratégia para extração de TLINKs e RelTempDCD foi baseada em SVM, focando na engenharia de *features* e otimização de parâmetros. Para extração de RelTempDCD, o estado da arte envolve SVM, porém, para extração de TLINKs, resultados superiores no *corpus* Clinical TempEval foram obtidos por abordagens baseadas em aprendizado profundo.

Modelos que consideram o contexto de maneira bidirecional atualmente mantêm os melhores resultados nos *corpora* relacionados ao Clinical TempEval. Ressaltam-se dois trabalhos: a abordagem baseada em *Bidirectional Long Short-Term Memory* (LSTM), utilizada por Lin *et al.* (2018), e o modelo baseado em *Bidirectional Encoder Representations From Transformers* (BERT) (DEVLIN *et al.*, 2019), empregado por Lee *et al.* (2019), que fizeram o *fine-tuning* do BioBERT, um modelo de BERT pré-treinado em *corpora* biomédicos.

Apesar de os melhores resultados no THYME *corpus*, nos recortes do Clinical TempEval 2016 e 2017, serem baseados em aprendizado profundo, optou-se por trabalhar com SVM (aprendizado de máquina tradicional), pelos seguintes motivos:
Os *corpora* relacionados ao THYME *corpus* são bem maiores que o *corpus* anotado neste projeto, sendo que abordagens baseadas em aprendizado profundo se beneficiam de maior quantidade de dados anotados. O *corpus* Clinical TempEval 2016 contém 23.243 TLINKs, enquanto o deste projeto possui 2.116 TLINKs, sendo cerca de 11 vezes menor. Considerando apenas TLINKs do tipo Contém (único tipo utilizado nos *corpora* do Clinical TempEval), o *corpus* anotado neste projeto é mais de 20 vezes menor.

a) A proposta desta tese é avaliar as diferentes características das três categorias de relações (EVTs em mesma sentença, EVTs em sentenças distintas e EVT e ET em mesma sentença).
b) A proposta desta tese é avaliar e desenvolver novas heurísticas para diminuição dos pares candidatos, focando principalmente em ter uma alta cobertura de TLINKs entre elementos em sentenças distintas.
c) A proposta desta tese é construir uma base sólida, fundamentada em aprendizado de máquina, trazendo features utilizadas pelos autores de melhores resultados e features propostas neste projeto, com o objetivo de avaliá-las em um cenário de textos ruidosos e de baixa qualidade de escrita (detalhado na seção 2.1.2).

Dos autores que tiveram resultados expressivos nos *corpora* associados ao i2b2 2012 e Clinical TempEval por meio de SVM, destacam-se Lin *et al.* (2016a), Lee *et al.* (2016, 2018) e Tang *et al.* (2013). Além deles, para domínio geral, ressalta-se o trabalho de Mirza e Tonelli (2016). Esses autores utilizaram a biblioteca LIBLINEAR (FAN *et al.*, 2008) para seus experimentos com SVM.

Para problemas multiclasses, o LIBLINEAR utiliza uma estratégia de *one-vs-the-rest*, sendo amplamente utilizado por ser eficiente para o treinamento de problemas em larga escala (FAN *et al.*, 2008). Apesar de o SVM ter a capacidade de mapear os dados para espaços dimensionais maiores por uso de *kernels*, se o conjunto de *features* é muito largo (mais *features* do que instâncias), esse mapeamento não linear não melhora os resultados (CHEN, C. *et al.*, 2015; HSU; CHANG; LIN, 2003). Justamente pelo fato de não utilizar *kernels*, o LIBLINEAR pode

treinar um conjunto maior de *features* com um classificador linear (CHEN, C. *et al.*, 2015).

Foi empregado o L2-regularized L2-loss dual SVM, implementado no LIBLINEAR, selecionando uma regularização baseada em L2 devido ao objetivo de todas as *features* terem um papel na classificação (LIN *et al.*, 2016a). Em *kernels* lineares, somente um parâmetro pode ser otimizado, o fator de penalidade indicado por C, representando a classificação incorreta (BEN-HUR; WESTON, 2010). Durante o treinamento, foram testados o padrão de C (C = 1) e a otimização do valor de C por meio de *grid-search*. Para valores altos de C, uma alta penalidade foi aplicada para erros/erros de margem (BEN-HUR; WESTON, 2010). A mudança dos valores de C controlou o erro de treinamento, erro de teste, número de vetores de suporte e margem do SVM (THARWAT, 2019).

Para os valores de C otimizados por *grid-search*, foi seguida a abordagem de Hsu, Chang e Lin (2003), aumentando o valor exponencialmente. A escolha de delimitação dos valores exponenciais foi baseada em Hsu, Chang e Lin (2003), Rossi e Carvalho (2008) e Mantovani *et al.* (2015), tendo sido buscados valores de 2^{-2} até 2^{13}.

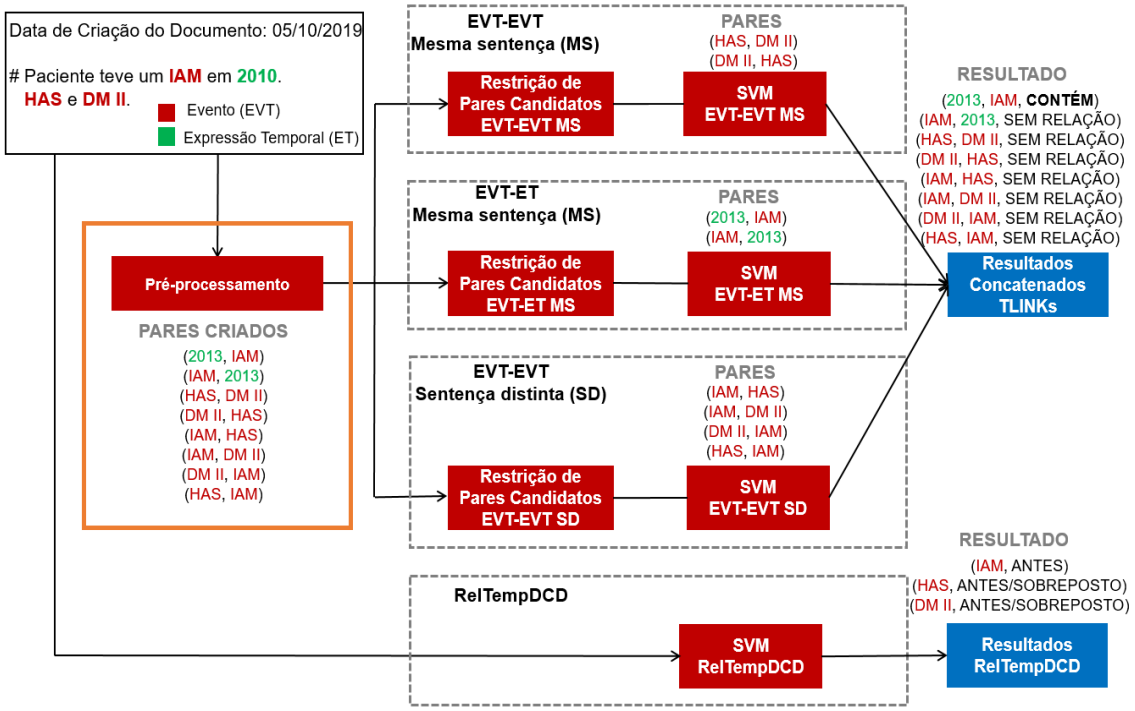
Além da otimização do valor de C, foi proposto trabalhar com *cost-sensitive learning* para mitigar o efeito do desbalanceamento. A abordagem foi utilizada por Lin *et al.* (2016a) e Lee *et al.* (2016, 2018). Para corrigir a falta de balanceamento nos dados, foi proposto adicionar diferentes custos para erros de classificação para cada classe, adicionando um peso inversamente proporcional à sua frequência (BEN-HUR; WESTON, 2010). Se fosse um problema de duas classes e existissem 20 pares positivos entre cem candidatos, o peso correto seria 5 (100/20) para a classe positiva e 1,25 (100/80) para a classe negativa (LEE *et al.*, 2016).

A seguir, serão detalhadas todas as características dos componentes específicos da extração de TLINKs e RelTempDCD. Na seção 5.1.1, será detalhado o pré-processamento; nas seções 5.1.2, 5.1.3, 5.1.4 e 5.1.5, os componentes para extração. Para TLINKs, serão detalhadas as heurísticas utilizadas para delimitação da quantidade de pares candidatos.

5.1.1 Pré-processamento

A etapa de pré-processamento (realçada na Figura 35) envolveu o pré-processamento do texto, os procedimentos para criação das *features* e a criação de pares candidatos (como são gerados os exemplos positivos e negativos). Todos serão detalhados a seguir.

Figura 35 – Etapas do processo de extração de RTs, com a etapa de pré-processamento destacada.



Fonte: O autor (2020).

Em relação ao pré-processamento do texto, houve a normalização, buscando reduzir o impacto da formatação flexível, e a tokenização, transformando o texto em sentenças e *tokens*.

Na normalização, primeiramente foram aplicadas as seguintes abordagens: (i) transformação de números com separadores decimais do tipo ponto em vírgula por expressões regulares; (ii) correções de casos envolvendo vírgulas, pontos e ponto e vírgula, como a correção de “hoje.Eu” para “hoje. Eu”; (iii) eliminação de dígitos, por meio de expressões regulares; (iv) transformação de múltiplos espaços em branco em apenas um, por meio de expressões regulares. Quanto à eliminação dos dígitos, existem alguns domínios em que números podem carregar significado, mas há aplicações em que a inclusão de números é menos informativa. Para esta tese, optou-

se por não trabalhar com números, devido à questão esparsa das representações por *Bag-of-Words* (BOW), bigrama e trigrama.

Na tokenização, o texto foi dividido em sentenças e destas foram obtidos os *tokens*. Em textos clínicos, é usual o delimitador de sentença ser uma tecla "*enter*", indicando uma quebra de linha, sem qualquer pontuação associada. Além disso, existem certos pontos de má formatação de pontuações e utilização de símbolos que devem ser levados em conta por meio de pré-processamento. Se esses aspectos não forem considerados durante a delimitação de uma sentença, vão ocorrer casos com sentenças extremamente longas, inclusive casos em que o documento inteiro é considerado uma única sentença. Para cada novo documento gerado, o texto foi dividido com base na quebra de linha; após essa divisão, os trechos resultantes foram pré-processados e obtida uma lista ordenada das sentenças no documento por meio do método *Sent_tokenize* da biblioteca NLTK, com o idioma ajustado para português. Para obter os *tokens* a partir das sentenças, foi usado o método *Word_tokenize* da biblioteca NLTK.

Após a etapa de tokenização, foram aplicadas as seguintes abordagens da normalização: (i) exclusão de pontuações por expressões regulares; (ii) transformação dos textos para letras minúsculas. Pontuações são importantes em algumas análises, mas consideradas não informativas em muitas aplicações (DENNY; SPIRLING, 2017). Em textos clínicos, existe uma falta de padronização em pontuações e símbolos, por isso foram descartadas pontuações para esses tipos de *feature*. Transformar o texto em minúscula reduz o vocabulário, generalizando a aplicação de PLN, sendo um passo recomendado neste cenário de pesquisa, em que textos podem ser totalmente escritos em letras maiúsculas (LANE; HOWARD; HAPKE, 2019).

Em relação à criação das *features*, foram utilizados os pré-processamentos já detalhados anteriormente. As *features* envolvendo representações por meio de BOW, bigrama e trigrama foram geradas pelo método *CountVectorizer* da biblioteca scikit-learn[2]. As *features* baseadas em *one-hot encoding* foram codificadas pelos métodos *Categorizer* e *Dummy Encoder* da biblioteca DaskML[3]. Todas as *features* numéricas foram normalizadas com base no método *StandardScaler*, disponível na biblioteca scikit-learn, buscando valor de média zero e variância unitária. Todas as *features* para

[2] Disponível em: https://scikit-learn.org/stable/.
[3] Disponível em: https://ml.dask.org/.

cada componente estão detalhadas no Apêndice E, assim como o pré-processamento utilizado (como BOW e *one-hot encoding*).

Para criação dos pares candidatos, foram aplicadas as etapas de normalização e tokenização já mencionadas. A criação dos pares candidatos é essencial para a extração de RTs. Em uma anotação de relações, são fornecidos somente os exemplos positivos, sendo os negativos, necessários para treinar modelos, determinados por meio de heurísticas. Como já mencionado, é usual limitar a quantidade de pares de alguma forma ou criar separações baseadas em sentenças ou números de *tokens*. Nesta tese, foi empregado um critério baseado em sentenças, sendo necessário primeiramente definir o que é considerado uma sentença. Foi proposta uma abordagem baseada em utilização de um contador e um caractere de identificação do começo do evento. Foi obtida uma lista ordenada de todas as menções no documento, tanto EVTs quanto ETs a partir da posição no documento. Para cada elemento da lista ordenada, foi criado um texto, em que foi copiado o texto original, com a adição de um caractere para indicar a posição inicial da menção no documento. Esse procedimento é exemplificado pela adição do caractere “$” no EVT “atrioseptoplastia”, na Figura 36.

Figura 36 – Texto ambulatorial com adição do caractere “$” para indicar início de uma menção no documento.

```
# 2 CX DE TROCA DE MITRAL + $ATRIOSEPTOPLASTIA (2006)
# FA PERMANENTE
# 23/09/14 PROTESE METALICA EM POSIÇÃO MITRAL NORMOFUNCIONANTE.
VE APRESENTA DIMENSÕES NORMAIS COM FUNÇÃO SISTOLICAPRESERVADA.
# MEDICAMENTOS EM USO:
WARFARIM 5MG
SELOKEN 100MG
# PACIENTE VEM PARA CONSULTA DE ROTINA.
NEGA QUAISQUER SINTOMAS.
NEGA INTERCORRÊNCIAS DA ÚLTIMA CONSULTA PRA CÁ.
# AO EXAME :  BEG, CORADA, HIDRATADA, LOTE, EUPNÉICA.
MV + SEM RA. RITMO CARDÍACO IRREGULAR, BULHAS NORMOFONÉTICAS, PRESENÇA DE ESTALIDO METÁLICO.
FC 60
PA 110X60 MMHG
CONDUTA : MANTEMOS SELOKEM.ENCAMINHAMOS A PACIENTE PARA AMBULATÓRIO DE VALVOPATIAS.
```

Fonte: O autor (2020).

Em seguida, a lista ordenada de sentenças foi varrida por um laço de repetição, até que a sentença com o caractere “$” fosse encontrado, ocasião em que foi obtido o número da sentença na lista de sentenças, ao qual foi adicionada a menção na lista ordenada de menções. O próximo passo foi criar todos os possíveis pares entre as

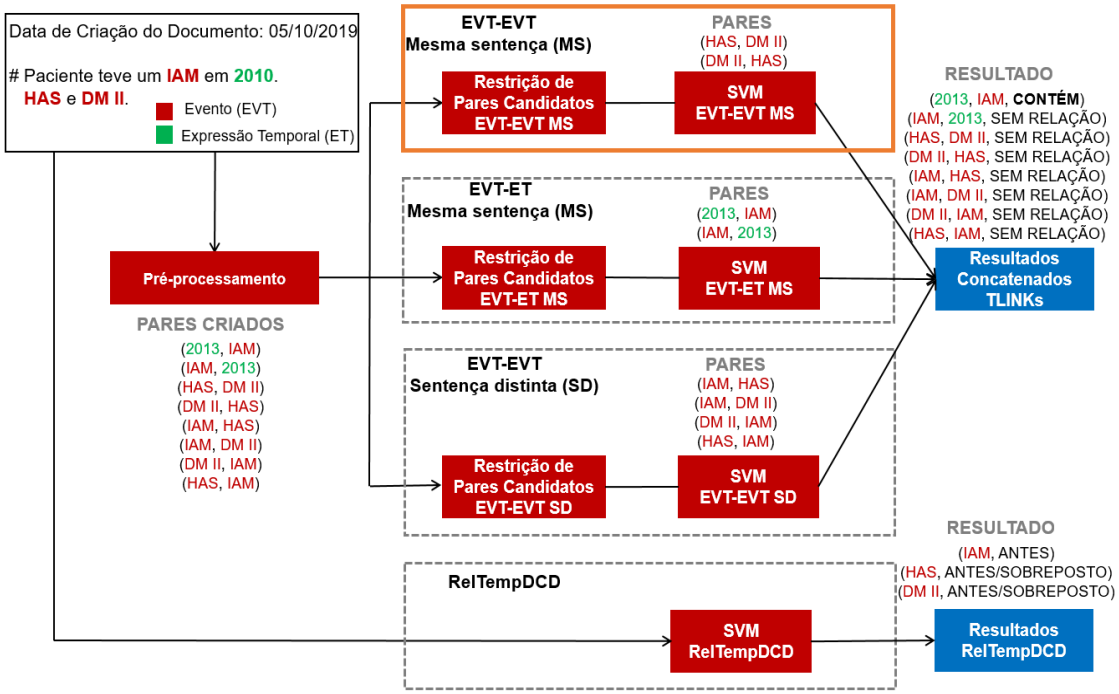
menções (excluindo pares entre o mesmo elemento) da lista ordenada do documento. Para cada par gerado, existindo uma relação anotada (exemplo positivo) entre as menções, foi adicionado ao par o rótulo do respectivo tipo de relação (por exemplo, Contém); caso contrário, foi adicionado o rótulo Sem Relação, para indicar que se tratava de um exemplo negativo para o classificador. Após esse procedimento, foi utilizado o número da sentença como parâmetro para determinar se os pares candidatos eram relativos a elementos na mesma sentença ou sentenças distintas.

Após todos esses passos, foram obtidos todos os possíveis pares de TLINKs entre EVTs em mesma sentença, entre EVTs em sentenças distintas e entre EVT e ET em mesma sentença. Esse mesmo procedimento foi realizado para todos os documentos e pares gerados, concatenados em uma lista por categoria de TLINK.

5.1.2 Modelo para extração de TLINKs entre eventos em mesma sentença

O desenvolvimento de modelos de extração de TLINKs entre EVTs em mesma sentença envolveu a definição de heurísticas para limitar a quantidade de pares negativos e a engenharia de *features*. Ambas as questões serão detalhadas nesta seção. Dentro das etapas do processo de extração de TLINKs, esta está realçada naFigura 37.

Figura 37 – Etapas do processo de extração de RTs, com a etapa de extração de RTs entre EVTs em mesma sentença destacada.



Fonte: O autor (2020).

Ao contrário de TLINKs entre EVT e ET em que quase toda ET na sentença é importante, em TLINKs entre EVTs, nem todos os EVTs em uma sentença o são no contexto do domínio clínico (LIN *et al.*, 2017). Sendo assim, TLINKs entre EVTs acabam tendo uma maior proporção de pares negativos, influenciando o desempenho dos classificadores, tanto que, no conjunto de treinamento deste *corpus* anotado, considerando somente relações em mesma sentença, existem cerca de 18 pares negativos para cada par positivo em TLINKs entre EVTs, enquanto para TLINKs entre EVT e ET há cerca de oito pares negativos. Para diminuir essa proporção, foram propostos experimentos com duas heurísticas.

A primeira delas foi a heurística apresentada por Tourille *et al.* (2016), ao trabalhar com SVM em experimentos no *corpus* Clinical TempEval 2016, tendo sido testada neste projeto para as três categorias de TLINKs distintas. Esta abordagem se tornou popular para experimentos com *corpora* relacionados ao Clinical TempEval, sendo utilizada por: Lin *et al.* (2018), Tourille *et al.* (2017a, 2017c), Liu *et al.* (2019) e Dligach *et al.* (2017). Inclusive, os melhores resultados para os *corpora* do Clinical TempEval 2016 e 2017 empregam esta heurística.

A heurística citada por Tourille *et al.* (2016) transforma o problema de classificação de dois tipos de relação (Contém e Sem Relação) em um problema de classificação de três tipos (Contém, Contido_Por e Sem Relação). A relação nomeada Contida_Por foi adicionada para indicar que uma menção está contida por outra. Assim, a heurística proposta consistiu em considerar todos os pares da esquerda para a direita e, quando necessário, trocar a relação do tipo Contém por Contida_Por. Como o processo usual considera todos os possíveis pares da esquerda para a direita e vice-versa, isso acaba gerando muitos pares adicionais. Com esta abordagem, o número de possíveis pares foi cortado pela metade ao custo de considerar uma relação adicional, porém todos os exemplos positivos foram mantidos. Com exceção dos TLINKs do tipo Sobreposto, os demais tinham sua categoria oposta/inversa. TLINKs do tipo Sobreposto são reflexivos, ou seja, não importando a ordem do elemento, a relação é mantida, sendo preciso somente trocar a ordem das menções quando necessário.

Além dessa heurística, foi proposta uma pelo próprio doutorando, baseada em uma observação do conjunto de treinamento sobre os EVTs do tipo Teste. Definida a heurística, só existem pares entre EVTs do tipo Teste (ambos sendo marcações de

Testes), quando: (i) algum dos EVTs do tipo Teste tem o trecho “avali” (com o objetivo de englobar termos como “avaliação” e “avaliar”) em sua marcação; (ii) algum dos EVTs do tipo Teste tem um dos trechos “lab” ou “exam” (tentando cobrir todas as diferentes formas de mencionar exames laboratoriais, como “laboratório” e “exames”) em sua marcação; (iii) algum dos EVTs do tipo Teste tem o trecho “fc” (menção de frequência cardíaca dentro do exame de eletrocardiograma) em sua marcação.

De 2.003 pares positivos entre EVTs do tipo Teste no conjunto de treinamento, apenas 187 eram exemplos positivos. Com a heurística proposta, foram criados somente 342 pares, dos quais 184 eram exemplos positivos. Sendo assim, apenas três pares positivos não foram cobertos pela heurística, alcançando 98,4% de todos os pares positivos no conjunto de treinamento.

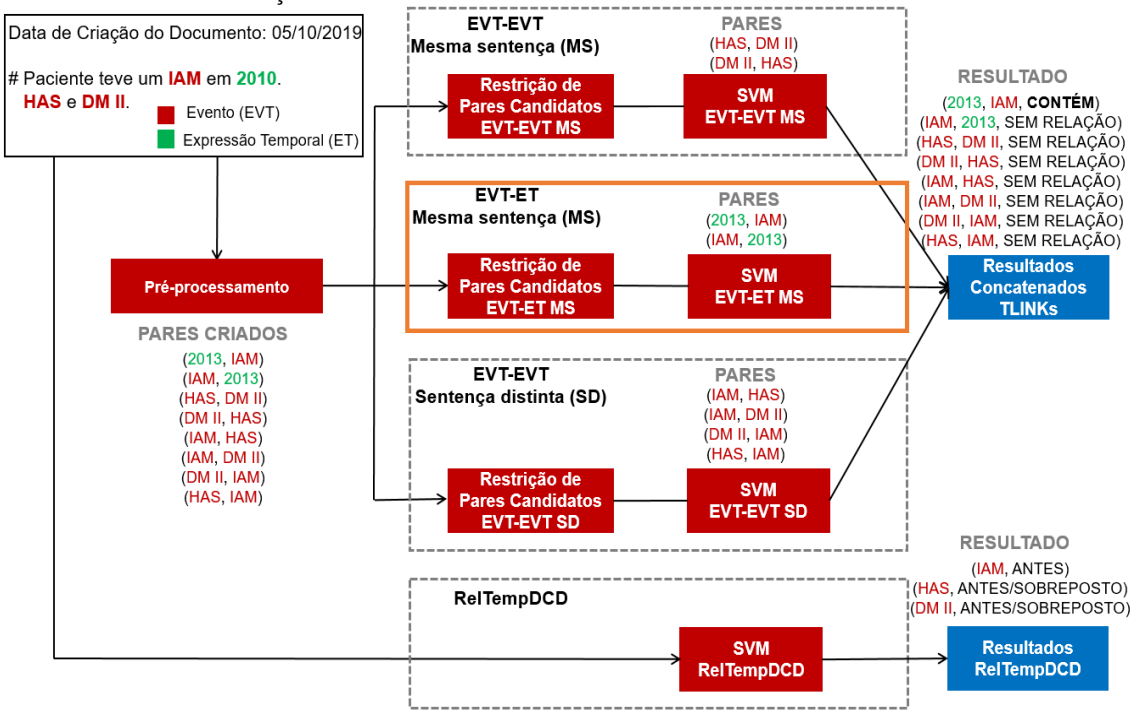
Com as heurísticas definidas, o próximo passo foi mencionar a engenharia de *features*. Foram utilizadas *features* propostas nos estudos de melhor desempenho para os *corpora* do Clinical TempEval 2015, 2016 e 2017 e i2b2 2012.

Features usadas para extração de TLINKs podem ser divididas em *features* propostas para descrever cada “ponto” da relação, ou seja, as entidades, e *features* propostas para descrever a relação entre a fonte e o alvo (MACAVANEY; COHAN; GOHARIAN, 2017). O conjunto de *features* empregado para esse tipo de relação é descrito no Apêndice E.

5.1.3 Modelo para extração de TLINKs entre eventos e expressões temporais em mesma sentença

O desenvolvimento de modelos de extração de TLINKs entre EVT e ET em mesma sentença envolveu a utilização de uma heurística para limitar a quantidade de pares negativos e a engenharia de *features*. Ambas as questões serão detalhadas nesta seção. Dentro das etapas do processo de extração de TLINKs, esta está realçada na Figura 38.

Figura 38 – Etapas do processo de extração de RTs, com a etapa de extração de RTs entre EVTs e ETs em mesma sentença destacada.



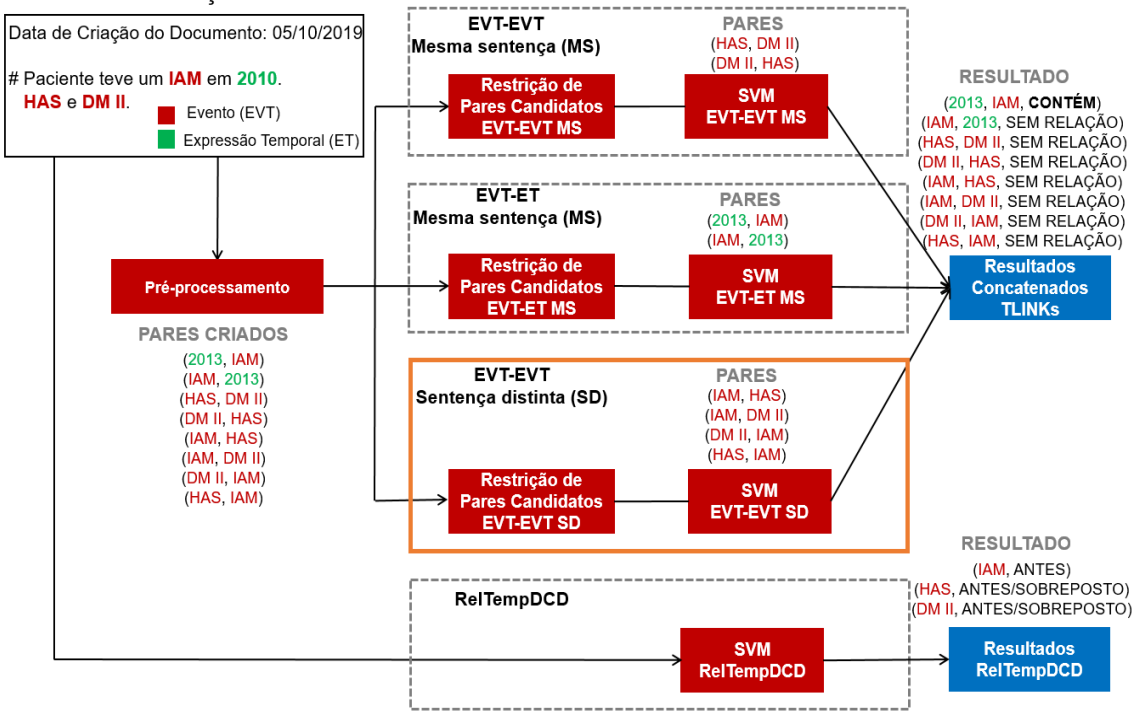
Fonte: O autor (2020).

A heurística testada foi detalhada anteriormente, sendo a proposta por Tourille *et al.* (2016), considerando somente pares da esquerda para a direita, trocando o tipo de relação quando necessário. O conjunto de *features* empregado para este tipo de relação é descrito no Apêndice E.

5.1.4 Modelo para extração de TLINKs entre eventos em sentenças distintas

O desenvolvimento de modelos de extração de TLINKs entre EVTs em sentenças distintas envolveu a definição de heurísticas para limitar a quantidade de pares negativos e a engenharia de *features*. Ambas as questões serão detalhadas nesta seção. Dentro das etapas do processo de extração de TLINKs, esta está realçada na Figura 39.

Figura 39 – Etapas do processo de extração de RTs, com a etapa de extração de RTs entre EVTs em diferentes sentenças destacada.



Fonte: O autor (2020).

Em trabalhos como de Tourille *et al.* (2016, 2017b), foi proposto considerar relações entre elementos em até três sentenças distintas, justificado pela cobertura de 89% das relações positivas no *corpus* Clinical TempEval 2016.

Estudos como de Tang *et al.* (2013), D'Souza e Ng (2014), Cherry *et al.* (2013), Lin *et al.* (2016a) e Lee *et al.* (2016) tiveram resultados promissores por meio de estratégia de filtragem de pares candidatos em sentenças distintas. Sendo assim, nesta tese, optou-se por trabalhar com TLINKs em sentenças distintas, com a criação de heurísticas para filtrar os pares a somente elementos que "provavelmente" tivessem uma relação.

As estratégias de Tang *et al.* (2013), D'Souza e Ng (2014) e Cherry *et al.* (2013) focavam casos de correferências em diferentes sentenças, com ênfase em aspecto como comparação de atributos dos EVTs e núcleo sintático. Lee *et al.* (2016) propuseram diversas heurísticas baseadas em seções de documentos e combinação de atributos para filtrar pares em mesma sentença e em sentenças distintas. Além disso, adicionaram algumas heurísticas específicas para proporcionar maior limitação à criação de pares (mais de duas sentenças), objetivando manter somente pares que eram muito prováveis de ter uma relação. O componente de melhor resultado de Lee

et al. (2016) foi justamente o classificador responsável por TLINKs entre elementos com distâncias de mais de duas sentenças.

Foram testadas duas heurísticas: a primeira delas foi a heurística proposta por Tourille *et al.* (2016), considerando somente pares da esquerda para a direita, trocando o tipo de relação quando necessário; a segunda envolveu observações no conjunto de treinamento.

Como é possível verificar na Tabela 1, a cobertura dos pares positivos aumenta conforme o número de sentenças adjacentes é incrementado. Em casos de uma sentença adjacente, somente foram gerados pares de um EVT com outros EVTs considerando o limite de uma sentença adjacente (tanto antes quanto depois dele). Vale salientar que não foram criados pares dos EVTs com outros da mesma sentença, uma vez que são representados em outra categoria de TLINK.

Tabela 1 – Quantidade de relações entre EVTs em sentenças distintas no conjunto de treinamento, de acordo com a quantidade de sentenças adjacentes consideradas, trazendo cobertura, total de pares gerados e proporção.

Sentenças adjacentes	Total de pares positivos	Pares positivos considerados	Cobertura	Total de pares gerados	Proporção (pares negativos para cada par positivo)
1	190	81	42,63%	12.044	147,69
2	190	120	63,16%	24.062	199,52
3	190	155	81,58%	33.268	213,63
4	190	171	90,00%	42.024	244,75
5	190	175	92,11%	49.282	280,61
6	190	177	93,16%	56.320	317,19
7	190	179	94,21%	62.016	345,46
8	190	180	94,74%	66.930	370,83
9	190	181	95,26%	71.582	394,48
10	190	182	95,79%	75.640	414,60
11	190	183	96,32%	79.144	431,48
12	190	184	96,84%	82.100	445,20
13	190	185	97,37%	84.916	458,01
14	190	186	97,89%	87.410	468,95
15	190	187	98,42%	89.376	476,95
16	190	188	98,95%	91.158	483,88
17	190	189	99,47%	92.634	489,13
18	190	189	99,47%	93.842	495,52
19	190	189	99,47%	94.840	500,80
20	190	190	100,00%	95.670	502,53

Fonte: O autor (2020).

Observando os valores da Tabela 1, percebe-se que a proporção de pares negativos para cada par positivo foi muito superior aos demais componentes que trabalham com TLINKs em mesma sentença. Para uma cobertura de apenas 42,63%, foram criados quase 148 pares negativos para cada par positivo, considerando que cada par positivo foi dividido em diversos tipos de TLINK (por exemplo, Contém ou Sobreposto).

Para uma cobertura de pelo menos 90%, existia uma proporção de quase 245 pares negativos para cada par positivo. Então, foi proposta uma heurística com o mesmo objetivo de Lee *et al.* (2016): filtrar os pares a somente elementos que "provavelmente" contenham uma relação.

Após análises do conjunto de treinamento e conversas com um especialista da área da saúde, foi verificado que, entre esses EVTs em sentenças distintas, seria importante trazer os do tipo Problema relacionados a EVTs do tipo Teste, evidenciando que aquele Problema foi encontrado durante aquele Teste. Devido à formatação do texto, ocorreram casos em que os EVTs do tipo Problema foram separados por pontos ou quebras de linha, colocando-os em sentenças distintas, o que acabou fazendo com que essas menções ficassem em sentenças longe da menção do teste (por exemplo, eletrocardiograma). Além disso, foi verificado que existiam diversos casos de EVTs do tipo Teste envolvendo exames de laboratório e menções de avaliação relacionadas a outros testes em sentenças distintas. Exames de laboratório foram separados por quebras de linha ou pontos, sendo considerados elementos em sentenças distintas.

Um último caso de menções contemplando testes foi identificado: EVTs do tipo Ocorrência relacionados a EVTs do tipo Teste em sentenças distintas. Este TLINK ocorreu quando certos exames foram pedidos pelo médico durante a seção Planejamento; assim, foram marcadas relações entre os exames e a menção do retorno.

Dos 190 TLINKs entre EVTs em sentenças distintas, 87 (45,79%) envolveram TLINKs entre EVTs do tipo Teste; 71 (37,37%), TLINKs entre um EVT do tipo Teste e outro do tipo Problema; e 19 (10%), TLINKs entre um EVT do tipo Teste e outro do tipo Ocorrência. Somente focando em criar regras para esses três tipos de TLINK, foram cobertos 177 (93,16%) dos 190 possíveis TLINKs.

Após um período de análise dessas relações de interesse, foram determinadas as seguintes regras para criação de pares. Nesse sentido, somente foram

considerados TLINKS entre EVTs em diferentes sentenças quando: (i) em EVTs do tipo Teste, se algum tivesse o trecho “avali”, “lab” ou “exam” e ambos, RelTempDCD de Antes ou Depois; (ii) em um EVT do tipo Teste e outro do tipo Problema, se o primeiro tivesse o trecho “eco” (englobando “eco”, “ecocardiograma”, “ecocardio”, “ecocardio com stress”, “ecott”, “ecodoppler” e “ecocardio tt”), “cintilo” (englobando “cintilo” e “cintilografia”) ou “ecg” (abreviatura de ecocardiograma) e ambos, RelTempDCD de Antes ou Depois; (iii) em um EVT do tipo Teste e outro do tipo Ocorrência, se este tivesse o termo “retorno” e ambos, RelTempDCD de Depois.

Sumarizando a cobertura das regras propostas no conjunto de treinamento, tem-se:

a) De 87 relações envolvendo EVTs do tipo Teste, a heurística conseguiu capturar 87 (100% de cobertura).
b) De 71 relações envolvendo um EVT do tipo Teste e outro do tipo Problema, a heurística conseguiu capturar 70 (99,06% de cobertura).
c) De 19 relações envolvendo um EVT do tipo Teste e outro do tipo Ocorrência, a heurística conseguiu capturar 18 (94,12% de cobertura), com a única exceção sendo uma menção de retorno com grafia incorreta.

Dos 177 TLINKs que possivelmente poderiam ser extraídos, as regras conseguiram capturar 175. Sendo assim, de todos os 190 TLINKs entre EVTs em sentenças distintas no conjunto de treinamento, com a heurística criaram-se pares em que 175 deles estavam presentes. A partir da Tabela 2, é possível observar que, pela heurística proposta, a proporção de números negativos para cada par positivo reduziu drasticamente, sendo possível considerar até 20 sentenças adjacentes, com uma proporção de 5,51 e uma taxa de cobertura de 92,11%.

Tabela 2 – Quantidade de relações entre EVTs em sentenças distintas no conjunto de treinamento com a heurística criada, de acordo com a quantidade de sentenças adjacentes consideradas, trazendo cobertura, total de pares gerados e proporção.

Sentenças adjacentes	**Total de pares positivos**	**Pares positivos considerados**	**Cobertura**	**Total de pares gerados**	**Proporção (pares negativos para cada par positivo)**
1	190	71	37,37%	233	2,28
2	190	108	56,84%	424	2,93
3	190	141	74,21%	560	2,97
4	190	156	82,11%	640	3,10
5	190	160	84,21%	779	3,87

6	190	162	85,26%	861	4,31
7	190	164	86,32%	946	4,77
8	190	165	86,84%	980	4,94
9	190	166	87,37%	1001	5,03
10	190	167	87,89%	1020	5,11
11	190	168	88,42%	1043	5,21
12	190	169	88,95%	1061	5,28
13	190	170	89,47%	1081	5,36
14	190	171	90,00%	1093	5,39
15	190	172	90,53%	1107	5,44
16	190	173	91,05%	1114	5,44
17	190	174	91,58%	1125	5,47
18	190	174	91,58%	1128	5,48
19	190	174	91,58%	1136	5,53
20	190	175	92,11%	1139	5,51

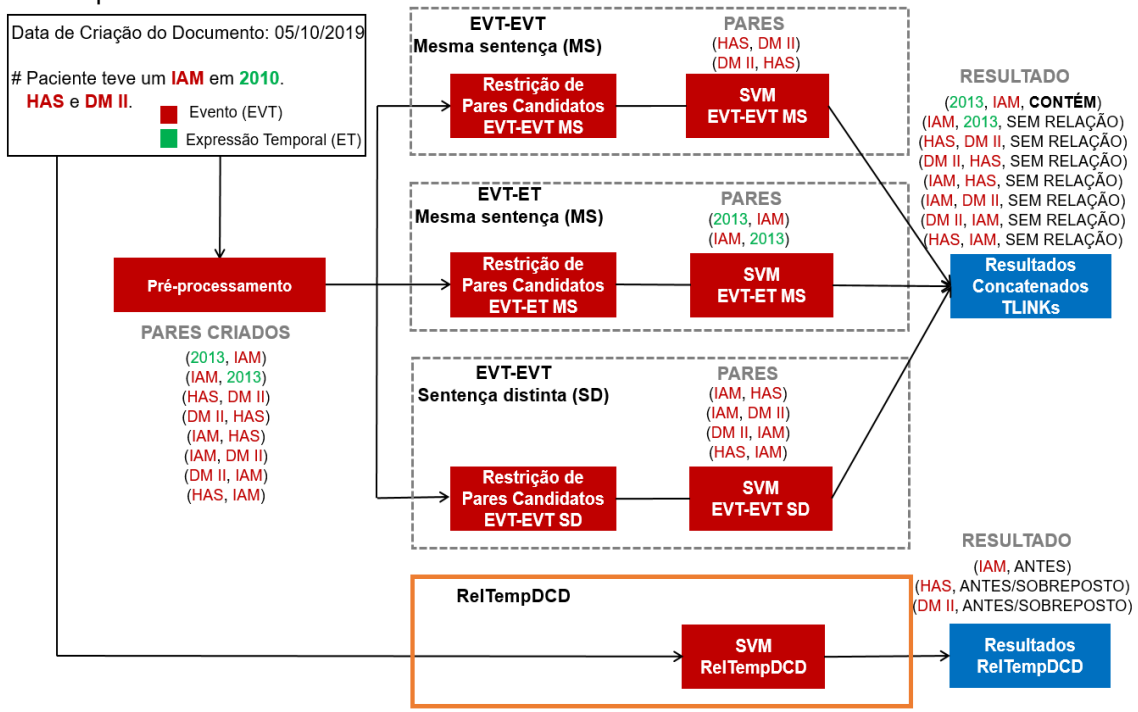
Fonte: O autor (2020).

O conjunto de *features* empregado para este tipo de relação é descrito no Apêndice E.

5.1.5 Modelo para extração de RelTempDCD

Para este tipo de RT, não existiu o problema da criação de pares candidatos, tendo todo EVT uma relação com a DCD, tornando um problema de classificação sem pares negativos. Para cada EVT, sempre houve uma relação com um dos seguintes tipos: Antes, Sobreposto, Antes/Sobreposto e Depois. Dentro das etapas do processo de extração de TLINKs, esta está realçada na Figura 40.

Figura 40 – Etapas do processo de extração de RTs, com a etapa de extração de RTs do tipo RelTempDCD destacada.



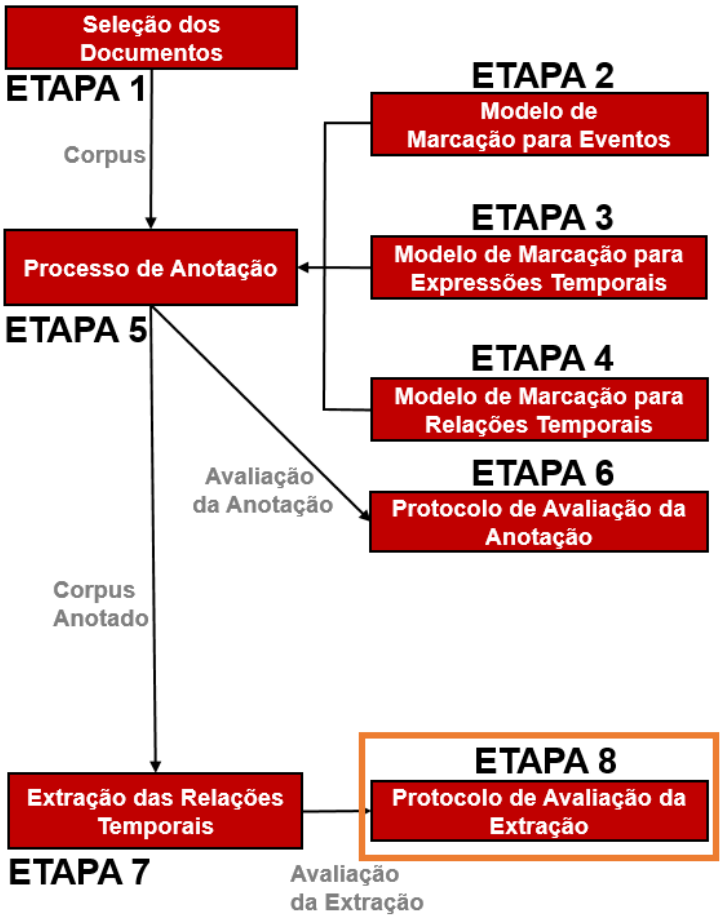
Fonte: O autor (2020).

Neste tipo de RT, as *features* utilizadas focaram em descrever o EVT e seu contexto. Sendo assim, houve uma simplificação em relação aos demais tipos, em que existiam diversas *features* para descrever a relação entre menções presentes no texto. O conjunto de *features* empregado para este tipo de relação é descrito no Apêndice E.

5.2 ETAPA 8 – PROTOCOLO DE AVALIAÇÃO DA EXTRAÇÃO

O protocolo de avaliação da extração define como será feita a avaliação do desempenho dos algoritmos para extração de RTs. Dentro das etapas descritas anteriormente, esta está realçada na Figura 41. Primeiramente, serão trazidas algumas definições comuns das avaliações do treinamento e teste, como divisão do conjunto e métricas de avaliação gerais. Em seguida, serão detalhados os aspectos específicos da avaliação do treinamento, na seção 5.2.1, e do teste, na seção 5.2.2.

Figura 41 – Etapas da pesquisa, realçando a etapa 8.



Fonte: O autor (2020).

Para os experimentos, o conjunto de dados foi dividido em dois, sendo um conjunto de treinamento (80%) e outro de teste (20%), seguindo a divisão proposta por Viani *et al.* (2019) e pela *shared task* Clinical TempEval 2017, sendo esta similar à divisão de 75% para treinamento e 25% para teste proposta na *shared task* Clinical TempEval 2016.

No conjunto de treinamento, foram desenvolvidas as heurísticas para limitação de pares candidatos (não levando em conta observações no conjunto de teste) e os experimentos relacionados aos classificadores. O conjunto de teste permaneceu isolado, sendo utilizado somente para testar os modelos finalizados.

Para avaliação dos modelos, tanto no treinamento quanto no teste, foi usado o F1-*score*, métrica de avaliação mais empregada em aprendizado em *datasets* desbalanceados (HE; MA, 2013). Como envolve a ponderação dos valores de *Precision* e *Recall*, ter um valor alto assegura que tanto a *Precision* quanto o *Recall* são relativamente altos (TAN; STEINBACH; KUMAR, 2016).

Na avaliação de TLINKs, ocorre um severo desbalanceamento em favor da classe relativa aos pares negativos com o sistema somente recebendo crédito pelos pares positivos que acerta. No critério de avaliação do Clinical TempEval 2016, em que só existe um tipo de relação (Contém), claramente é descrito que um sistema só recebe crédito por uma marcação se ambas as menções estão marcadas corretamente assim como um TLINK do tipo Contém entre elas (BETHARD *et al.*, 2016).

Outra questão é considerar que as classes envolvendo exemplos negativos "inflam" o F1-*score* do sistema, principalmente o micro F1-*score*, devido ao severo desbalanceamento. Foi, então, criada uma função de avaliação, em que a classe Sem Relação (responsável pelos exemplos negativos) foi considerada apenas no cálculo dos valores de FP e FN das demais classes, porém não foram gerados resultados específicos para essa classe. Nesse cenário, foi possível avaliar o desempenho geral dos sistemas pelo micro F1-*score*, considerando todos os termos entre todas as classes juntamente no cálculo (TANEV; MAGNINI, 2006).

5.2.1 Avaliação do treinamento

No treinamento, o objetivo da avaliação envolveu avaliar modelos desenvolvidos, buscando um modelo que generalizasse bem para o conjunto de teste.

Todos os experimentos no conjunto de treinamento envolveram *cross-validation*, a fim de treinar e testar os modelos em diferentes partições do conjunto de dados. Dentre as abordagens de *cross-validation*, foi selecionada a *stratified cross-validation*, criando os conjuntos para preservar a distribuição original dos dados em todos os subconjuntos, garantindo que treinamento e validação teriam aproximadamente a mesma distribuição de classes (KONONENKO; KUKAR, 2007). Usualmente, é utilizada 10-*fold cross-validation*, porém uma única 10-*fold cross-validation* pode não ser suficiente para estimar o erro quando os dados são limitados, sendo comum repetir o processo de *cross-validation* dez vezes, obtendo a média dos resultados (WITTEN; FRANK; HALL, 2011). Para o treinamento, todos os testes dos modelos envolveram 10x 10-*fold cross-validation* pelo método *RepeatedStratifiedKFold* da biblioteca scikit-learn, fixando o número de divisões em dez, o número de repetições em dez e um valor de *seed* para que os procedimentos fossem reprodutíveis.

Além do valor de F1-*score* obtido por 10x 10-*fold cross-validation,* foi realizada uma avaliação da significância estatística dos resultados. Para cada experimento, o resultado foi comparado a um modelo-base, definido a partir das *features* envolvidas nos melhores resultados da literatura. Esse seria, teoricamente, o "melhor modelo" a ser contraposto pelos modelos testados, os quais tinham variantes de configurações e/ou *features* que não eram tão utilizadas por autores, porém poderiam ser interessantes nesse contexto. O objetivo foi testar se a diferença no desempenho médio entre dois modelos era real ou não, ou seja, se a média de desempenho era igual (aceite de H0) ou diferente (rejeite de H0). Devido aos pareados, uma hipótese poderia ser utilizar o *paired t-test*, porém a condição de independência das amostras é violada, havendo uma subestimação da variância, porque as amostras não são independentes, ou seja, os diferentes conjuntos de treinamento e teste se sobrepõem de alguma forma (BOUCKAERT; FRANK, 2004; NADEAU; BENGIO, 2003).

A proposta de Nadeau e Bengio (2003) consistiu em uma correção da estimação da variância para levar em conta a dependência entre as amostras. O *correct resampled t-test* é mostrado na Equação 5.1, sendo o valor de t calculado com base em: k (número de experimentos), n_2 (percentual utilizado para teste de *cross-validation*), n_1 (percentual utilizado para treinamento na *cross-validation*), σ^2 (variância das diferenças) e $\overline{d}$ (diferença das médias).

$$\boldsymbol{t = \frac{\overline{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right) \times \sigma_d^2}}} \qquad (5.1)$$

A partir de experimentos envolvendo testes estatísticos adequados para avaliação de dois algoritmos, Bouckaert e Frank (2004) recomendam o uso do *correct resampled t-test* proposto por Nadeau e Bengio (2003), em um cenário de 10x *10-fold cross-validation*. A configuração do 10x 10-*fold cross-validation* envolveu o mesmo método e configurações mencionados anteriormente.

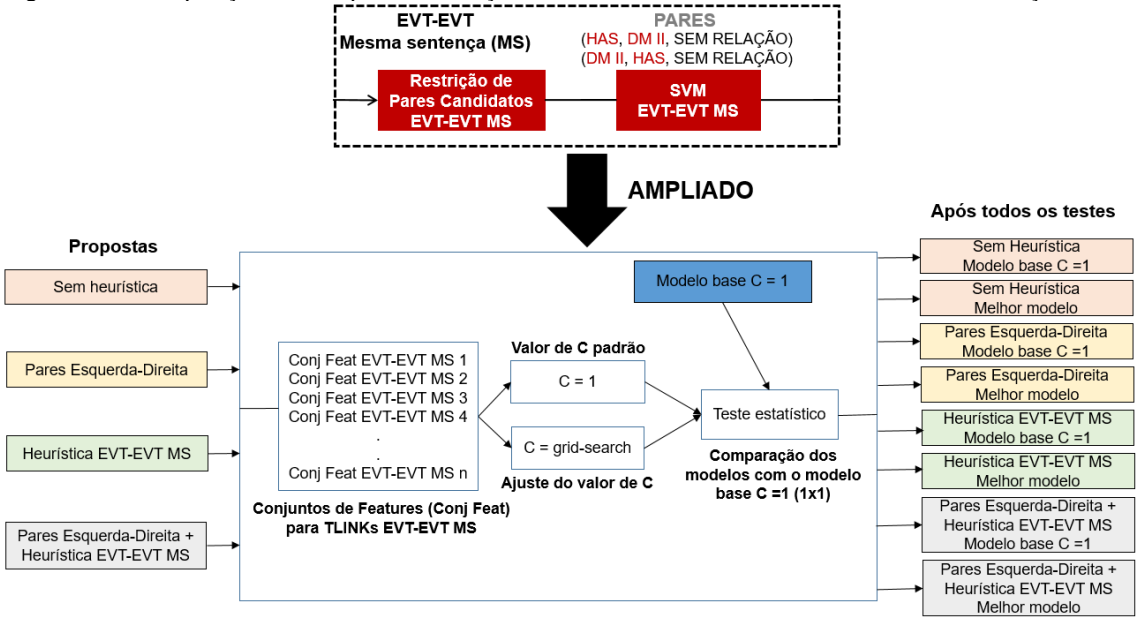
Para esta tese, a avaliação das *features* e da otimização dos valores de C deu-se pelo *correct resampled t-test*, realizado dez vezes em 10-*fold cross-validation*, considerando um nível de significância (*alpha*) de 0,05, n_1 de 0,9, n_2 de 0,1 e k de 100.

Como já mencionado, existem quatro classificadores especializados para extração de RTs: o classificador para TLINKs entre EVTs em mesma sentença (seção

5.1.2), TLINKs entre EVT e ET em mesma sentença (seção 5.1.3), TLINKs entre EVTs em sentenças distintas (seção 5.1.4) e RelTempDCD (seção 5.1.5). Todos eles, com exceção de RelTempDCD, foram testados com base nas heurísticas selecionadas, por sua vez testadas para todos os conjuntos de *features* selecionados. Para RelTempDCD, houve somente o teste de todos os conjuntos de *features*.

Nesta etapa de treinamento, os testes se restringiram aos classificadores em separado, verificando o melhor desempenho para cada heurística dentre os modelos e otimizações. Cada um dos quatro classificadores especializados será abordado adiante, começando pelo classificador para TLINKs entre EVTs em mesma sentença. Na Figura 42, é mostrado o *design* do experimento para sua extração.

Figura 42 – Ampliação da etapa de extração de TLINKs entre EVTs em mesma sentença.



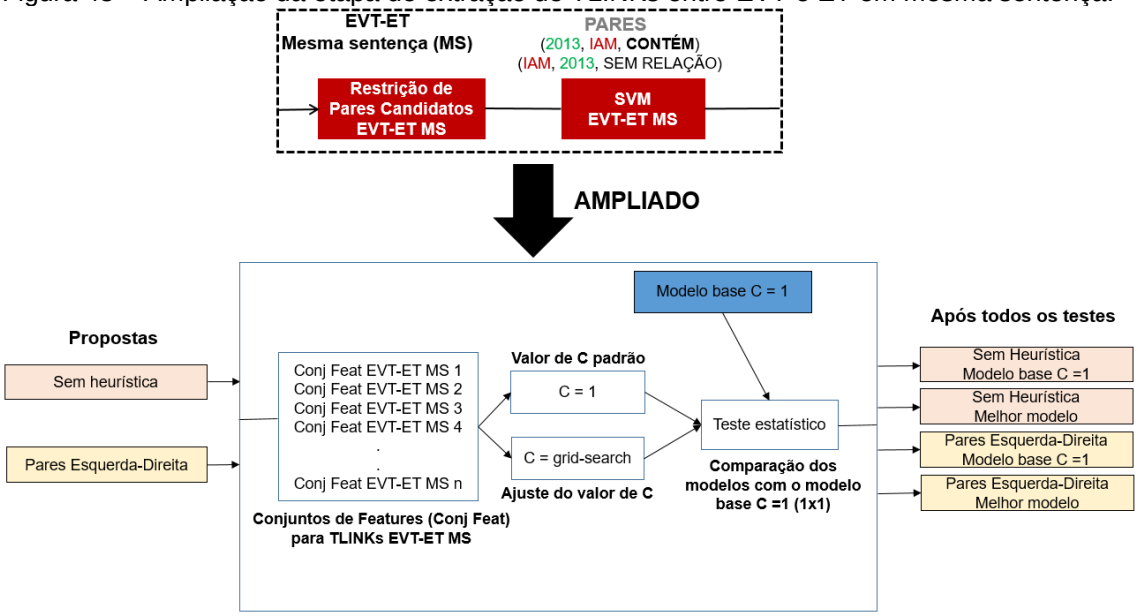
Fonte: O autor (2020).

Como é possível notar, foram criadas quatro propostas envolvendo o uso de heurísticas: (i) não usar nenhuma heurística (“Sem heurística”); (ii) utilizar a heurística proposta por Tourille *et al.* (2016), considerando somente pares da esquerda para a direita, trocando a relação quando necessário (“Pares Esquerda-Direita”); (iii) utilizar a heurística criada na seção 5.1.2, focada na diminuição de pares de EVTs do tipo Teste (“Heurística EVT-EVT MS”); (iv) combinação de “Pares Esquerda-Direita” e “Heurística EVT-EVT MS”. A marcação Começa_Em foi fundida/mesclada com Sobreposto devido ao baixo número de casos, principalmente em um cenário de 10-*fold cross validation*.

Foram testadas todas as propostas em todos os conjuntos de *features* (Apêndice F), considerando o parâmetro C igual a 1 (parâmetro *default* do SVM) como o valor obtido pelo *grid-search* em 10-*fold cross-validation*. Todos esses experimentos foram testados por 10x 10-*fold corrected resampled t-test* em comparação a um modelo-base (cada proposta tinha seu modelo-base), considerando como de "melhor desempenho". No final, cada proposta teve o "melhor resultado" selecionado, sendo esse o modelo de maior F1-*score* e média estatisticamente diferente em relação ao modelo-base. Ambos, modelo-base e melhor resultado, foram mantidos para cada proposta.

O próximo classificador mencionado refere-se a TLINK entre EVT e ET em mesma sentença. Na Figura 43, é mostrado o *design* do experimento para sua extração dos TLINKs. Diferentemente do TLINK anterior, este experimento envolveu somente duas propostas: (i) "Sem heurística"; (ii) "Heurística Pares Esquerda-Direita".

Figura 43 – Ampliação da etapa de extração de TLINKs entre EVT e ET em mesma sentença.



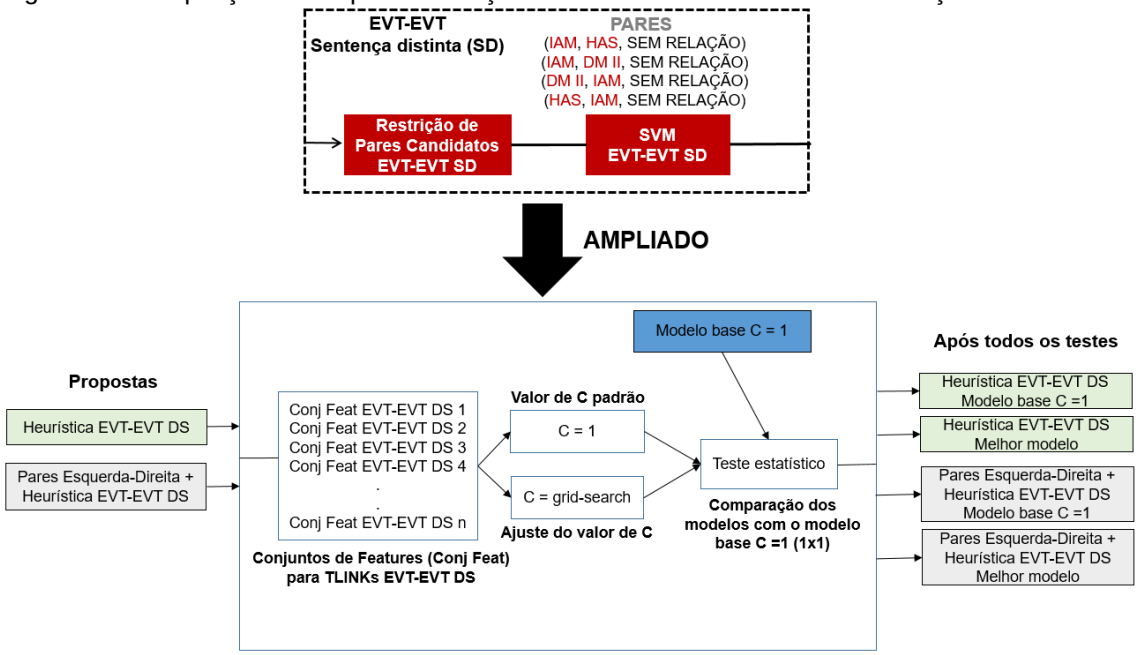
Fonte: O autor (2020).

Devido ao baixo número de exemplos positivos para Antes, Começa_Em e Termina_Em em experimentos envolvendo 10-*fold cross-validation*, os tipos Começa_Em e Termina_Em foram fundidos/adicionados à marcação Sobreposto e a marcação Antes foi descartada.

Foram testadas todas as propostas em todos os conjuntos de *features* (Apêndice F), sendo mantidos o modelo-base e o modelo de melhor desempenho para cada proposta.

O próximo experimento envolveu TLINKs entre EVTs em sentenças distintas. Na Figura 44, é mostrado o *design* do experimento para sua extração.

Figura 44 – Ampliação da etapa de extração de TLINKs entre EVTs em sentenças distintas.



Fonte: O autor (2020).

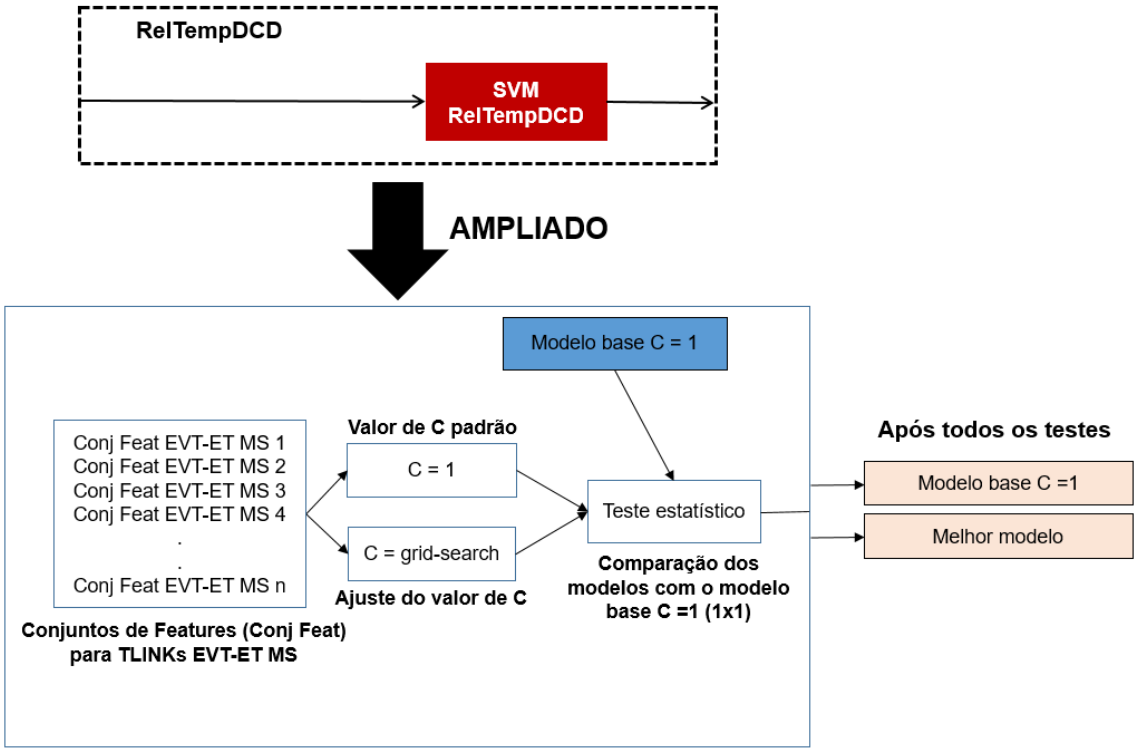
Assim como no caso de TLINKs entre EVTs em mesma sentença, foram criadas duas propostas envolvendo o uso de heurísticas: (i) utilizar a heurística criada na seção 5.1.4, focada na diminuição de pares de EVTs do tipo Teste (“Heurística EVT-ETV SD”); (ii) combinação de “Pares Esquerda-Direita” e “Heurística EVT-EVT SD”. Nesses experimentos, o objetivo foi verificar a eficiência da heurística proposta.

Foi fixada uma janela de sentença de 20, ou seja, cada EVT poderia ter pares com outros EVTs em até 20 sentenças (tanto antes quanto depois). Só foi possível trabalhar com 20 sentenças devido ao baixo número de pares gerados pelas heurísticas, positivos ou negativos.

Foram testadas todas as propostas em todos os conjuntos de *features* (Apêndice F), sendo mantidos o modelo-base e o modelo de melhor desempenho para cada proposta.

O próximo experimento envolveu RelTempDCD, não existindo a necessidade de propor heurísticas. Na Figura 45, é mostrado o *design* do experimento para sua extração.

Figura 45 – Ampliação da etapa de extração de RelTempDCD.



Fonte: O autor (2020).

Foram testadas todas as propostas em todos os conjuntos de *features* (Apêndice F), sendo mantidos o modelo-base e o modelo de melhor desempenho para cada proposta.

5.2.2 Avaliação do teste

Na avaliação do teste, o objetivo foi verificar o desempenho do modelo para dados ainda não vistos pelos modelos. Consistiu em três etapas distintas. Na primeira e segunda, foram feitas as avaliações individuais dos componentes, ou seja, foram verificados os modelos de melhor resultado e base no conjunto de teste para cada tipo de RT. Na avaliação da classificação, foi calculado o desempenho do modelo para sua entrada de dados; assim, se devido a heurísticas ou limitação de janela não fossem consideradas todas as possíveis relações daquele tipo de RT, isso não era levado em conta. Portanto, os modelos não foram penalizados por não classificar instâncias reduzidas do conjunto de treinamento pelas heurísticas, não sendo adicionados como FN. Por exemplo, se um modelo fosse testado com cem pares positivos, seria avaliado em relação a eles; se outro modelo para o mesmo tipo de RT fosse treinado com 300 pares positivos, seria avaliado em relação a eles. Esta

avaliação visou a julgar o desempenho da classificação, por isso foi chamada avaliação da classificação. Vale ressaltar que, devido à não restrição de pares, os modelos envolvendo RelTempDCD não precisavam desta segunda etapa de avaliação.

Apesar de a avaliação da classificação trazer informações sobre o desempenho dos modelos, não foi possível comparar as propostas diferentes. Por isso, nesta etapa, chamada avaliação local, os modelos foram penalizados, em suas avaliações, com casos que não se propuseram a classificar. Por exemplo, se existissem 750 TLINKs entre EVTs em mesma sentença, o número de casos positivos avaliados seria de 750, independentemente das heurísticas utilizadas.

Na última etapa, houve uma avaliação total do sistema. Os resultados de todos os componentes em nível local para TLINKs foram concatenados e gerados os resultados finais. RTs do tipo RelTempDCD foram mantidas em separado, por serem distintas. Os modelos gerados são mostrados no Quadro 14. Esta avaliação envolveu somente os melhores modelos para cada proposta. Cada modelo utiliza um conjunto de propostas; quando era mencionado o termo “mesma sentença” no modelo, isso indicava que ele não lidava com TLINKs entre EVTs em sentenças distintas e o termo “sentenças distintas”, o inverso. Todos os modelos foram penalizados pelos pares de TLINKS entre EVT e ET que não foram extraídos.

Quadro 14 – Modelos finais gerados para extração de RTs no conjunto de teste.

Modelos finais	**TLINK EVTs em mesma sentença**	**TLINK EVT e ET em mesma sentença**	**TLINK EVTs em sentenças distintas**	**TLINK EVT e ET em sentenças distintas**
Sem heurística	Sem heurística	Sem heurística	-	-
Pares Esquerda-Direita	Pares Esquerda-Direita	Pares Esquerda-Direita	-	-
Heurísticas propostas (mesma sentença)	Heurística EVT-EVT MS	Sem heurística	-	-
Heurísticas propostas (sentenças distintas)	Heurística EVT-EVT MS	Sem heurística	Heurística EVT-EVT SD	-
Heurísticas propostas + Pares Esquerda-Direita (mesma sentença)	Heurística EVT-EVT MS + Pares Esquerda-Direita	Pares Esquerda-Direita	-	-
Heurísticas propostas + Pares Esquerda-Direita (sentenças distintas)	Heurística EVT-EVT MS + Pares Esquerda-Direita	Pares Esquerda-Direita	Heurística EVT-EVT SD + Pares Esquerda-Direita	-
RelTempDCD	-	-	-	-

Fonte: O autor (2020).

6 RESULTADOS

Os resultados envolveram a definição dos modelos de anotação, as estatísticas relacionadas aos *corpora* anotados e, por fim, os resultados da avaliação, tanto da anotação quanto da extração de RTs.

6.1 MODELO PARA ANOTAÇÃO DE EVENTOS

O *guideline* para anotação de EVTs foi a base da anotação, alicerce para as demais camadas. No Apêndice B, é fornecido todo um detalhamento do *guideline*, trazendo diversos exemplos para cada categoria de anotação, sempre com explicações e fazendo correlação com o SOAP. Esse *guideline* foi extremamente detalhado, procurando fornecer a maior quantidade de exemplos possíveis aos anotadores, para que, em caso de dúvida, sempre fosse possível realizar uma busca no *guideline*. Sempre que havia dúvida de anotação, eram adicionados exemplos e observações ao *guideline,* a partir das discussões realizadas. Devido a alguns problemas na diferenciação de marcações de “queixa” entre os tipos Problema e Evidência, foram adicionados diversos exemplos e explicações. O *guideline* foi finalizado com 84 páginas, sendo que a maior parte está relacionada às definições do que deveria ser marcado para cada tipo de EVT, assim como detalhamento de RelTempDCD para cada um deles. Um exemplo é mostrado na Figura 46.

Figura 46 – Exemplo de marcação do *guideline* de anotação de EVTs.

- **Menções de queixas (peço atenção pois existe uma diferença tênue entre queixas como ‘PROBLEMA’ ou ‘EVIDÊNCIA’):**
  REGRA: Quando se está EVIDENCIANDO (tipo “evidência”) uma queixa é preciso estar se queixando de algum PROBLEMA.
  - Nega queixas **[PROBLEMA]**

  Observação: Como esta menção de queixas não está relacionada com um PROBLEMA esta menção se torna o próprio problema a ser marcado.
  - Nega outras queixas **[PROBLEMA]**.
  - Paciente queixa-se de dor de cabeça **[PROBLEMA]**

  Observação: Como esta menção de queixa-se está relacionada a outro problema “dor de cabeça” esta menção de “queixa-se” não é um PROBLEMA.

Fonte: O autor (2020).

6.2 MODELO PARA ANOTAÇÃO DE EXPRESSÕES TEMPORAIS

O *guideline* para anotação de ETs é detalhado com explicações e exemplos no Apêndice C. Nele, similarmente ao de EVTs, o objetivo envolveu ser o mais específico

possível, trazendo exemplos e explicações para cada anotação mais complexa. O *guideline* foi finalizado com 35 páginas, em que a maior parte envolveu definições de Data e Frequência, as mais frequentes nos textos. As ETs do tipo Data ocorriam durante todo o texto e eram utilizadas em contextos distintos; menções do tipo Frequência usualmente estavam ligadas a medicamentos. Na Figura 47, são mostrados dois exemplos de marcações específicas no *guideline*.

Figura 47 – Exemplo de marcação do *guideline* de anotação de ETs.

- Insulina 55 / 11 / 10 **[Tipo: Frequência | mod: ND | Valor: P1D | freq: 3x | cps: ND]**

Observação: nestes casos onde existe a dosagem por período marcamos tudo como expressão temporal.

- enalapril 10 / 10 **[Tipo: Frequência | mod: ND | Valor: P1D | freq: 2x | cps: NA]**

Observação: neste caso “10/10” indica 10mg 2x por dia.

Fonte: O autor (2020).

6.3 MODELO PARA ANOTAÇÃO DE RELAÇÕES TEMPORAIS

O *guideline* para anotação de TLINKs foi o último a ser desenvolvido e levou em conta as marcações e *guidelines* das camadas de anotação anteriores. No Apêndice D, é fornecido todo um detalhamento dele, com diversas explicações e exemplos para cada categoria de anotação. Como marcações de TLINKs dependem do contexto das menções e da compreensão geral do texto pelo anotador, foi importante fornecer diversos exemplos distintos para serem utilizados como base de marcação. Um exemplo dessa abordagem é fornecido na Figura 48, com marcações completas envolvendo exames laboratoriais.

Figura 48 – Exemplo de marcação do *guideline* de anotação de RTs.

- **Lab [EVENTO: TESTE] 26/01/2013 [TIMEX3: DATA]** – **Cr [EVENTO: TESTE]** 1,2 / **Glic [EVENTO: TESTE]** 89/ **Ur [EVENTO: TESTE]** 35 / **TGP [EVENTO: TESTE]** 41

Marcação: 26/01/2013 **CONTÉM** Lab

Marcação: Lab **CONTÉM** Cr

Marcação: Lab **CONTÉM** Glic

Marcação: Lab **CONTÉM** Ur

Marcação: Lab **CONTÉM** TGP

Observação: neste caso é aplicado o padrão da menção geral do exame de laboratório estar contida (relação do tipo CONTÉM) dentro da data e os exames em específico estarem contidos (relação do tipo CONTÉM) na menção geral do exame de laboratório.

Fonte: O autor (2020).

6.4 ESTATÍSTICAS DO *CORPUS* ANOTADO

Nesta seção, são trazidos resultados relacionados às anotações, assim como detalhes específicos. Além disso, são feitas comparações das anotações com trabalhos similares encontrados na literatura.

Na Tabela 3, são fornecidas informações relativas ao *corpus* anotado, trazendo comparações com dois *corpora* relacionados à extração de TLINKs no domínio clínico, o i2b2 2012 e o Clinical TempEval 2016. Além deles, foi considerado o estudo de Viani *et al.* (2019), por ser um *corpus* de tamanho similar no domínio da cardiologia (apesar de não ter anotações de ETs e TLINKs).

Tabela 3 – Quantidade de documentos, *tokens*, eventos, ETs e RTs para o *corpus* e demais *corpora* relacionados, com sua respectiva média por texto.

Informação	***Corpus* anotado**		**i2b2 2012**		**Clinical TempEval 2016**		**Viani *et al.* (2019)**	
	Valor absoluto	**Média/ texto**	**Valor absoluto**	**Média/ texto**	**Valor absoluto**	**Média/ texto**	**Valor absoluto**	**Média/ texto**
Textos	126	-	310	-	591		75	
Tokens	20.907	165,9	177.940	574	517.113	875	57.225	763
EVTs	4.015	31,9	26.846	86,6	78.851	133,4	4.365	58,2
ETs	870	6,9	3.844	12,4	7.863	13,3	-	
TLINKs	2.116	16,8	54.560	176	23.243	39,33	-	

Fonte: O autor (2020).

A partir da Tabela 3, observa-se uma menor quantidade de documentos anotados nesse *corpus*, assim como uma menor média de *tokens* por texto, sendo notada uma significativa diferença de tamanho, inclusive em relação ao *corpus* de

Viani *et al.* (2019), que tem proporções de anotações de EVTs similares. A questão da compilação de informações e do demasiado uso de abreviações para encurtar o tempo de escrita influenciou o tamanho do *corpus*.

Foram anotados 126 documentos, com um total de 4.015 anotações de EVTs, 870 anotações de ETs e 2.116 anotações de TLINKs, com as quantidades de anotações sendo menores em comparação aos demais trabalhos da Tabela 3. Para a anotação de EVTs, esse aspecto fica evidente em comparação com o *corpus* Clinical TempEval 2016. O *corpus* relacionado ao i2b2 2012 não foi utilizado para comparação de TLINKs devido a este incorporar a RelTempDCD como uma anotação de TLINK, "inflando" o número de TLINKs por texto. No *corpus* Clinical TempEval, existem 11 vezes mais anotações de TLINKs, evidenciando uma diferença de tamanho significativa entre ambos. Percebe-se, nos *corpora* da Tabela 3, um baixo número de marcações de ETs por texto, principalmente em comparação aos EVTs anotados.

Na Tabela 4, são fornecidas estatísticas sobre os atributos dos EVTs, trazendo suas respectivas categorias, assim como sua contagem de anotações. O atributo RelTempDCD é apresentado em uma tabela separada (Tabela 5) para melhor comparação com os demais esquemas de anotação.

Tabela 4 – Atributos dos EVTs com suas respectivas categorias de marcação, número total de marcações por categoria e percentual entre parênteses.

Atributo	**Categoria**	**Total**
Tipo	Problema	1.444 (35,97%)
	Teste	1.017 (25,33%)
	Tratamento	971 (24,18%)
	Ocorrência	318 (7,92%)
	Evidência	216 (5,38%)
	Departamento Clínico	49 (1,22%)
Polaridade	Positiva	3.571 (88,94%)
	Negativa	444 (11,06%)
Modalidade	Factual	3.973 (98,95%)
	Não Factual	42 (1,05%)

Fonte: O autor (2020).

Quanto ao atributo Tipo, houve uma maior quantidade de anotações nesse *corpus* de Problemas, Testes e Tratamentos, similar ao encontrado na anotação do i2b2 2012, em que 32,4% das anotações eram relacionadas a Problemas, 16,4%, a Testes e 24,4%, a Tratamentos. Esses percentuais de anotação para Problemas e Testes foram similares aos encontrados na anotação desta tese, com valores de

32,97% e 24,18%, respectivamente. Contudo, no *corpus* i2b2 2012, foram encontradas menos anotações de Testes; somente 16,4% das anotações envolviam Testes, valor inferior aos 25,33% anotados neste *corpus*. Isso ocorreu devido ao menor número de menções de exames laboratoriais no *corpus* i2b2 2012, aspecto observado durante uma análise dele. Na anotação de Viani *et al.* (2019), 26,9% de anotações envolveram Testes, sendo um valor mais próximo ao obtido neste *corpus*.

Para o atributo RelTempDCD, notou-se um maior percentual de anotações de Antes e Antes/Sobreposto, justificado pelo fato de a maior parte das informações presentes no SOAP residir no histórico do paciente na seção Subjetiva. Como é mostrado na Tabela 5, tanto na anotação do *corpus* Clinical TempEval 2016 quanto na anotação do *corpus* MERLOT, existiu maior incidência de marcações envolvendo Sobreposto. Ressalta-se que o *corpus* MERLOT foi utilizado para comparação devido a ter as mesmas categorias de anotação do *corpus* produzido neste projeto. Tanto no MERLOT quanto no Clinical TempEval 2016, existiu um percentual menor de anotações de Antes/Sobreposto, em comparação com o *corpus* anotado neste projeto.

Tabela 5 – Quantidade de anotações das categorias de RelTempDCD para este *corpus*, com seu valor absoluto e percentual e comparações em relação aos trabalhos relacionados.

Atributo	**Categoria**	**Neste *corpus***	**Clinical TempEval 2016**	**MERLOT**
RelTempDCD	Antes	1.658 (41,30%)	29.170 (36,97%)	1.936 (10,68%)
	Antes/Sobreposto	1.138 (28,34%)	4.240 (5,37%)	2.643 (14,58%)
	Sobreposto	725 (18,06%)	37.091 (47,01%)	12.211 (67,36%)
	Depois	494 (12,30%)	8.400 (7,38%)	1.3337 (7,38%)

Fonte: O autor (2020).

Na Tabela 6, são fornecidas informações sobre os percentuais de anotação do atributo Tipo das ETs, com uma comparação com a anotação do i2b2 2012 e o trabalho de Azevedo (2019).

Tabela 6 – Tipo de ET, trazendo o número total de marcações por tipo e o percentual neste *corpus*, com comparações em relação aos trabalhos relacionados.

Atributo	**Tipo**	**Neste *corpus***	**i2b2 2012**	**Azevedo (2019)**
Tipo	Frequência	443 (50,92%)	10,10%	39%
	Data	344 (39,54%)	70,5%	43%
	Duração	83 (9,54%)	16,70%	17%
	Tempo	0 (0%)	2,7%	1%

Fonte: O autor (2020).

A partir da Tabela 6, verifica-se que as ETs estavam mais fortemente ligadas a Datas na anotação do i2b2 2012, o que difere da anotação deste *corpus*, que tem percentuais de anotação similares aos de Azevedo (2019), com uma diferença entre os de Frequência e Duração. A anotação deste *corpus* teve mais menções de Frequência, possivelmente devido ao fato de ter mais menções de medicamentos em uso, sendo estes representados nos textos com sua respectiva periodicidade. Além disso, nota-se, em todos os *corpora*, uma baixa quantidade de anotações de Tempo, evidenciando um pequeno número de marcações de horas específicas do dia.

A quantidade de anotações de TLINKs por tipo é mostrada na Tabela 7. Devido à similaridade com a anotação do THYME *corpus*, foi utilizado como parâmetro de comparação um experimento de anotação no THYME *corpus* realizado por Styler *et al.* (2014a). Não foram feitas comparações deste *corpus* com os recortes do THYME *corpus* para as *shared tasks* do Clinical TempEval, por nestas terem sido considerados somente TLINKs do tipo Contém, não existindo informações sobre as anotações das demais categorias.

Tabela 7 – Marcações de TLINKs, com seu valor absoluto e percentual, além da comparação com um trabalho relacionado.

Anotação	**Tipo**	**Neste corpus**	**Styler *et al.* (2014a)**
TLINK	Contém	1.138 (53,78%)	5.112 (64,42%)
	Sobreposto	888 (41,97%)	1.205 (12,19%)
	Antes	69 (3,26%)	1.004 (12,65%)
	Começa_Em	19 (0,90%)	488 (6,15%)
	Termina_Em	1 (0,05%)	126 (1,59%)

Fonte: O autor (2020).

Em ambas as anotações, nesta e de Styler *et al.* (2014a), existiu um maior número de marcações do tipo Contém, questão característica da anotação de *narrative containers*. Na anotação deste *corpus*, quando comparada à anotação de Styler *et al.* (2014a), nota-se um elevado percentual de marcações do tipo Sobreposto, uma vez que medicamentos em uso são associados com suas frequências por marcações do tipo Sobreposto e marcações de ETs do tipo Duração são relacionados a EVTs com marcações do tipo Sobreposto. Estas são utilizadas em diversos contextos por sua definição mais abrangente. Além disso, neste *corpus*, a quantidade de anotações para os tipos Antes, Começa_Em e Termina_Em foi menos representativa, com cerca de 4,21% do total das anotações.

6.5 AVALIAÇÃO DA ANOTAÇÃO

Nesta seção, primeiramente são trazidas algumas questões referentes às anotações, como o tempo que levaram, assim como a quantidade de *rounds* de treinamento/refinamento necessários. Em seguida, são apresentados os valores de IAA para as três camadas de anotações deste *corpus*, com comparações em relação aos demais trabalhos presentes na literatura.

Para a anotação de EVTs, foram necessários seis *rounds* de treinamento, contando com sete reuniões para discussão e modificação do *guideline*, processo que durou cerca de dois meses. Para a anotação real, 14 lotes foram anotados durante três meses. Para ETs, foram necessários somente dois *rounds* de treinamento e uma reunião para discussão e modificação do *guideline*, totalizando duas semanas de treinamento. Para a anotação real, 14 lotes foram anotados durante um mês e meio. Para anotação de TLINKs, foram necessários quatro lotes de treinamento, com cinco reuniões para discussão e modificação do *guideline*, processo que durou cerca de um mês e meio. Para a anotação real, 14 lotes foram anotados durante cerca de três meses.

De forma geral, a anotação de EVTs se mostrou mais complexa, sendo necessário um maior número de reuniões para ter uma base sólida de anotação. A anotação de TLINKs, que poderia ser complexa pela questão da interpretação do contexto e pela diversidade das relações que podem ser marcadas, foi menos complexa que a anotação de EVTs.

Em relação à avaliação de anotação, na Tabela 8 são apresentados os valores de IAA para as anotações de EVTs, ETs e TLINKs. Os valores de IAA dos EVTs e ETs envolveram sua marcação/delimitação no texto (*span*), com valores de *exact* e *partial matching*. O IAA de TLINKs inclui o total de acerto do par e a relação entre as menções.

Tabela 8 – IAA para marcação de EVTs (*span*), ETs (*span)* e TLINKs.

Anotação	***F1-score* (*Exact*)**	***F1-score* (*Partial*)**
EVT	0,9066	0,9329
ET	0,9479	0,9602
TLINK	0,8835	-

Fonte: O autor (2020).

O processo de anotação de EVTs teve um valor de IAA de 0,9066 na configuração de *exact matching*, exibindo valores positivos em comparação aos encontrados em i2b2 2012 (0,83), Clinical TempEval 2016 (0,864), Viani *et al.* (2019) (0,855), TimeBank[4] (0,78) e Styler *et al.* (2014a) (0,8083). Na configuração de *partial matching*, houve um valor de IAA de 0,9329, positivo quando comparado aos valores obtidos por Viani *et al.* (2019) (0,922), i2b2 2012 (0,87) e TimeBank (0,81).

No processo de anotação de ETs, houve um valor de 0,9479 na configuração de *exact matching*, indicando um resultado positivo quando comparado aos valores encontrados em TimeBank (0,83), i2b2 2012 (0,73), Styler *et al.* (2014a) (0,8047) e Clinical TempEval 2016 (0,731). Na configuração de *partial matching*, foram obtidos valores similares à anotação do TimeBank (0,96), com valor de IAA de 0,9602.

Para anotação de TLINKs, obteve-se um IAA de 0,8835, sendo positivo quando comparado aos valores de 0,651 e 0,5630 para o Clinical TempEval 2016 e Styler *et al.* (2014a), respectivamente, ambos relacionados ao THYME *corpus*.

Tabela 9 – IAA para atributos dos EVTs e ETs, trazendo os valores de *exact* e *partial matching*.

Anotação	**Atributo**	***Accuracy* (*Exact*)**	***Accuracy* (*Partial*)**
EVT	Tipo	0,9917	0,9870
	Polaridade	0,9891	0,9887
	Modalidade	0,9967	0,9954
	RelTempDCD	0,9546	0,9528
ET	Tipo	0,9740	0,9723
	Valor	0,9468	0,9434
	Mod	0,9913	0,9892
	Quant	0,9858	0,9837
	Cps	0,9834	0,9837

Fonte: O autor (2020).

Em relação aos atributos dos EVTs e ETs, os valores de IAA obtidos para o processo de anotação estão sinalizados na Tabela 9. Foram encontrados valores de IAA de 0,9917 para Tipo, 0,9891 para Polaridade, 0,9967 para Modalidade e 0,9546 para RelTempDCD. Os resultados foram similares aos apresentados no *corpus* Clinical TempEval 2016 (0,966 para Tipo, 0,984 para Polaridade e 0,964 para Modalidade), com exceção da RelTempDCD, que teve valor de IAA de 0,721. Também foram similares aos apresentados na anotação do *corpus* i2b2 2012 (0,93 para Tipo, 0,97 para Polaridade e 0,96 para Modalidade). No estudo de Styler *et al.*

[4] Disponível em: http://www.timeml.org/timebank/documentation-1.2.html;

(2014a), um experimento de anotação no THYME *corpus*, foi obtido um valor de IAA para RelTempDCD de 0,7189.

Para as anotações de ETs, foram obtidos valores de IAA de 0,9740 para Tipo, 0,9468 para Valor, 0,9913 para Mod, 0,9858 para Quant e 0,9834 para Cps. Foram encontrados valores de IAA similares aos do TimeBank (1 para Tipo, 0,9 para Valor e 0,95 para Mod), com exceção do valor perfeito de concordância no TimeBank para Tipo. Para a anotação do *corpus* Clinical TempEval 2016, somente o Tipo era computado, tendo sido obtido um IAA de 0,941, similar ao desta. Na anotação do i2b2 2012, foram achados valores de IAA de 0,9 para Tipo, 0,75 para Valor e 0,83 para Mod. Além desses trabalhos, salienta-se a anotação de Azevedo (2019), em que, apesar de usar Cohen's *kappa* para avaliação, métrica diferente da utilizada nesta tese, foi obtido um valor de IAA de 0,8922 para o atributo Valor.

Tabela 10 – IAA individual para cada tipo de EVT, trazendo valores para a anotação no texto (*span*) e o valor de *Accuracy* para RelTempDCD.

Tipo	**F1-*score* (*Exact*)**	**F1-*score* (*Partial*)**	***Accuracy* RelTempDCD (*Exact*)**	***Accuracy* RelTempDCD (*Partial*)**
Problema	0,8845	0,9161	0,9125	0,9101
Tratamento	0,8951	0,9324	0,9802	0,9795
Teste	0,9478	0,9661	0,9815	0,9802
Evidência	0,9744	0,9744	0,9905	0,9905
Ocorrência	0,8668	0,8811	0,9455	0,9472
Departamento Clínico	0,8391	0,8506	0,9189	0,9211

Fonte: O autor (2020).

Na Tabela 10, são mostrados os valores de IAA para marcações no texto (*span*) para os tipos de EVT, evidenciando um menor IAA para marcação de Problemas, Ocorrências e Departamentos Clínicos. Nota-se um alto valor para marcações de Testes e Evidências. Para RelTempDCD, foram observados valores altos para Testes, Evidências e Tratamentos, além de valores mais baixos para Problemas, Departamentos Clínicos e Ocorrências, principalmente para os dois primeiros.

6.6 AVALIAÇÃO DA EXTRAÇÃO DE RELAÇÕES

Nas próximas seções, cada um dos quatro classificadores especializados tem seu desempenho avaliado nas diferentes propostas. Detalhes sobre os resultados relacionados ao treinamento estão disponíveis no Apêndice G.

Na seção 6.6.5, são realizadas concatenações dos resultados obtidos pelos classificadores especializados para TLINKs, obtendo um resultado final da extração.

6.6.1 Extração de TLINKs entre eventos em mesma sentença

Para este tipo de TLINK, foram testadas quatro propostas: (i) classificação sem qualquer heurística (“Sem heurística”); (ii) criação de pares somente da esquerda para a direita (“Pares Esquerda-Direita”); (iii) heurística proposta nesta tese (“Heurística EVT-EVT MS”); (iv) concatenação das abordagens “Pares Esquerda-Direita” e “Heurística EVT-EVT MS” (“Pares Esquerda-Direita + Heurística EVT-EVT MS”). A quantidade de exemplos para o conjunto de treinamento e de teste para este tipo de TLINK é mostrada na Tabela 11.

Tabela 11 – Quantidade de marcações por categoria para TLINKs entre EVTs em mesma sentença no conjunto de treinamento e de teste, com seus respectivos percentuais.

Marcação	Conjunto de treinamento	Conjunto de teste	Total
Contém	344 (25,05%)	118 (27,19%)	462 (25,57%)
Sobreposto	253 (18,43%)	97 (22,35%)	350 (19,37%)
Antes	52 (3,79%)	12 (2,76%)	64 (3,54%)
Começa_Em	11 (0,80%)	0 (0%)	11 (0,61%)
Termina_Em	0 (0%)	0 (0%)	0 (0%)
Total	**660 (100%)**	**227 (100%)**	**887 (100%)**

Fonte: O autor (2020).

Na Tabela 12, são apresentados os desempenhos dos modelos no conjunto de teste. A partir dos resultados, observou-se uma piora significativa para todas as propostas, com exceção da proposta “Pares Esquerda-Direita + Heurística EVT-EVT MS”, em que a diminuição não foi brusca (0,0261 em nível de classificador). Considerando um nível de comparação local, pelo valor do F1-*score* (local) de 0,7810, verifica-se que a abordagem “Pares Esquerda-Direita + Heurística EVT-EVT MS” obteve o melhor desempenho no conjunto de teste também.

Tabela 12 – Resultados para TLINKs entre EVTs em mesma sentença no conjunto de teste para todas as propostas, com o modelo-base e o melhor modelo.

Proposta	Modelo	F1-*score* (classificador)	Suporte	F1-*score* (local)	Suporte
Sem heurística	Modelo-base SH	0,5614	227	0,5614	227
	Melhor modelo SH	0,6130	227	0,6130	227
Pares Esquerda-Direita	Modelo-base PED	0,7045	227	0,7045	227

Heurística EVT-EVT MS	Melhor modelo PED	0,7045	227	0,7045	227
	Modelo-base HEMS	0,6283	226	0,6270	227
	Melhor modelo HEMS	0,6522	226	0,65076	227
Pares Esquerda-Direita + Heurística EVT-EVT MS	Modelo-base HEMS-PED	0,7703	226	0,7685	227
	Melhor modelo HEMS-PED	0,7828	226	**0,7810**	227

Fonte: O autor (2020).
Notas: SH = sem heurística; PED = pares esquerda-direita; HEMS = heurística EVT-EVT em mesma sentença.

O conjunto de *features* de melhor desempenho para a proposta “Pares Esquerda-Direita + Heurística EVT-EVT MS” envolveu as mesmas do modelo-base, com a adição de duas, as *features* Posição e Termos Antes/Depois. Na Tabela 13, é apresentado o desempenho dessa abordagem para todos os tipos de marcação. Vale salientar que as marcações Depois e Contido_Por foram adicionadas devido à abordagem “Pares Esquerda-Direita”, sendo marcações opostas a Antes e Contém, respectivamente. Nota-se que o desempenho para a classe majoritária foi o maior, impactando diretamente no valor de micro F1-*score*.

Tabela 13 – Resultados obtidos em nível local no conjunto de teste da melhor estratégia para extração de TLINKs entre EVTs em mesma sentença.

Marcação	***Precision***	***Recall***	**F1-*score***	**Suporte**
Depois	0,6667	0,6667	0,6667	3
Antes	0,5000	0,5556	0,5263	9
Contido_Por	0,0000	0,0000	0,0000	2
Contém	0,9115	0,8879	0,8996	116
Sobreposto	0,7000	0,6495	0,6738	97
Micro	0,8009	0,7621	**0,7810**	227

Fonte: O autor (2020).

6.6.2 Extração de TLINKs entre eventos e expressões temporais em mesma sentença

Foram testadas duas propostas: (i) classificação sem qualquer heurística (“Sem heurística”); (ii) criação de pares somente da esquerda para a direita (“Pares Esquerda-Direita”). A quantidade de exemplos para o conjunto de treinamento e de teste para este tipo de TLINK é mostrada na Tabela 14.

Tabela 14 – Quantidade de marcações por categoria para TLINKs entre EVTs em mesma sentença no conjunto de treinamento e teste, com seus respectivos percentuais.

Marcação	**Conjunto de treinamento**	**Conjunto de teste**	**Total**

Contém	310 (22,58%)	80 (18,43%)	390 (21,58%)
Sobreposto	395 (28,77%)	124 (28,57%)	519 (28,72%)
Antes	1 (0,07%)	1 (0,23%)	2 (0,11%)
Começa_Em	6 (0,44%)	2 (0,46%)	8 (0,44%)
Termina_Em	1 (0,07%)	0 (0%)	1 (0,06%)
Total	**713 (100%)**	**207 (100%)**	**920 (100%)**

Fonte: O autor (2020).

Na Tabela 15, são apresentados os desempenhos desses modelos no conjunto de teste, com o melhor sendo o próprio modelo-base. A partir dos resultados no conjunto de teste, observa-se uma melhora expressiva do melhor modelo para a proposta “Sem heurística” (“melhor modelo SH”). Para a proposta “Pares Esquerda-Direita”, também existiu uma melhora do desempenho no conjunto de teste. Em nível de comparação local, pelo valor do F1-*score* (local) de 0,9057, verifica-se que a abordagem “Pares Esquerda-Direita” também teve o melhor desempenho no conjunto de teste.

Tabela 15 – Resultados para TLINKs entre EVTs e ETs em mesma sentença no conjunto de teste para todas as propostas, com o modelo-base e o melhor modelo.

Proposta	**Modelo**	**F1-*score* (classificador)**	**Suporte**	**F1-*score* (local)**	**Suporte**
Sem heurística	Modelo-base SH	0,7396	206	0,7380	207
	Melhor modelo SH	0,8608	206	0,8586	207
Pares Esquerda-Direita	Modelo-base PED	0,9078	206	0,9057	207
	Melhor modelo PED	0,9078	206	0,9057	207

Fonte: O autor (2020).
Notas: SH = sem heurística; PED = pares esquerda-direita.

Na Tabela 16, é apresentado o desempenho dessa abordagem para todos os tipos de marcação. Vale salientar que a marcação Contido_Por foi adicionada devido à abordagem “Pares Esquerda-Direita”, sendo a marcação oposta/inversa a Contém. Nota-se que a maior parte das menções do tipo Contém foi transformada em Contido_Por; isso ocorreu devido ao fato de a menção da direita conter a menção da esquerda. Além disso, verifica-se um alto *Recall* para ambas as classes majoritárias.

Tabela 16 – Resultados obtidos em nível local no conjunto de teste da melhor estratégia para extração de TLINKs entre EVTs e ETs em mesma sentença.

Marcação	***Precision***	***Recall***	**F1-*score***	**Suporte**
Contido_Por	0,8442	0,9848	0,9091	66
Contém	1	0,5	0,6667	14
Sobreposto	0,9023	0,9524	0,9266	126
Antes	0	0	0	1

Micro	0,8848	0,9275	0,9057	207

Fonte: O autor (2020).

6.6.3 Extração de TLINKs entre eventos em diferentes sentenças

Foram testadas duas propostas: (i) heurística proposta nesta tese (“Heurística EVT-EVT SD”); (ii) “Heurística EVT-EVT SD” em conjunto com “Pares Esquerda-Direita” (“Pares Esquerda-Direita + Heurística EVT-EVT SD”). A quantidade de exemplos para o conjunto de treinamento e de teste é mostrada na Tabela 17.

Tabela 17 – Quantidade de marcações por categoria para TLINKs entre EVTs em sentenças distintas no conjunto de treinamento e de teste, com seus respectivos percentuais.

Marcação	Conjunto de treinamento	Conjunto de teste	Total
Contém	182 (82,35%)	78 (89,66%)	260 (84,42%)
Sobreposto	7 (3,17%)	2 (2,30%)	9 (2,92%)
Antes	1 (0,45%)	1 (1,15%)	2 (0,65%)
Começa_Em	0 (0%)	0 (0%)	0 (0%)
Termina_Em	0 (0%)	0 (0%)	0 (0%)
Total	**190 (100%)**	**81 (100%)**	**271 (100%)**

Fonte: O autor (2020).

Na Tabela 18, são apresentados os desempenhos desses modelos no conjunto de teste, com o melhor sendo o próprio modelo-base. A partir dos resultados, observa-se uma queda de *performance* do melhor modelo no treinamento para o conjunto de teste, de 0,1495 em nível de classificador para a proposta “Pares Esquerda-Direita + Heurística EVT-EVT SD”.

Tabela 18 – Resultados para TLINKs entre EVTs em sentenças distintas no conjunto de teste para todas as propostas, trazendo o modelo-base e o melhor modelo a partir dos experimentos.

Proposta	Modelo	F1-*score* (classificador)	Suporte	F1-*score* (local)	Suporte
Heurística EVT-EVT SD	Modelo-base HESD	0,7013	77	0,6545	81
	Melhor-modelo HESD	0,6463	77	0,6310	81
Pares Esquerda-Direita + Heurística EVT-EVT SD	Modelo-base PED-HESD	0,7595	77	0,7407	81
	Melhor modelo PED-HESD	0,7595	77	0,7407	81

Fonte: O autor (2020).
Notas: SH = sem heurística; PED = pares esquerda-direita.

Na Tabela 19, é verificado o desempenho desta abordagem para todos os tipos de marcação. Vale salientar que a marcação Contido_Por foi adicionada devido à

abordagem “Pares Esquerda-Direita”, sendo uma marcação oposta a Contém. Nota-se que o desempenho para esse tipo de TLINK é inferior ao dos demais classificadores especializados apresentados nas seções anteriores.

Tabela 19 – Resultados obtidos em nível local no conjunto de teste da melhor estratégia para extração de TLINKs entre EVTs em sentenças distintas.

Marcação	***Precision***	***Recall***	**F1-*score***	**Suporte**
Contido_Por	0,5714	0,5	0,5333	8
Contém	0,7568	0,8	0,7778	70
Sobreposto	0	0	0	1
Antes	0	0	0	2
Micro	0,7407	0,7407	0,7407	81

Fonte: O autor (2020).

6.6.4 Extração de RelTempDCD

A quantidade de exemplos para o conjunto de treinamento e teste para este tipo de TLINK é mostrada na Tabela 20.

Tabela 20 – Quantidade de marcações por categoria para RelTempDCD no conjunto de treinamento e de teste, com seus respectivos percentuais.

Marcação	**Conjunto de treinamento**	**Conjunto de teste**	**Total**
Antes	1269 (41,15%)	389 (41,78%)	1658 (41,30%)
Sobreposto	550 (17,83%)	175 (18,80%)	725 (18,06%)
Antes/Sobreposto	890 (28,86%)	248 (26,64%)	1138 (28,34%)
Depois	375 (12,16%)	119 (12,16%)	494 (12,30%)
Total	**3084 (100%)**	**931 (100%)**	**4015 (100%)**

Fonte: O autor (2020).

Na Tabela 21, são apresentados os desempenhos desses modelos no conjunto de teste. A partir dos resultados, observa-se uma queda de 0,0118 no desempenho do melhor modelo para o conjunto de teste.

Tabela 21 – Resultados para RelTempDCD, trazendo o modelo-base e o melhor modelo a partir do experimento no conjunto de teste.

Modelo	**F1-*score* (local)**	**Suporte**
Modelo-base RelTempDCD	0,9023	931
Melhor modelo RelTempDCD	0,9270	931

Fonte: O autor (2020).

Na Tabela 22, é verificado o desempenho desta abordagem para todos os tipos de marcação.

Tabela 22 – Resultados obtidos em nível local no conjunto de teste da melhor estratégia para extração de RelTempDCD.

Marcação	*Precision*	*Recall*	F1-*score*	Suporte
Antes	0,9150	0,9409	0,9278	389
Antes/Sobreposto	0,9344	0,9194	0,9268	248
Depois	0,8991	0,8235	0,8596	119
Sobreposto	0,9607	0,9771	0,9688	175

Fonte: O autor (2020).

6.6.5 Resultado da extração

Nesta seção, os classificadores especializados desenvolvidos são avaliados conjuntamente, ou seja, para o conjunto todo de teste como o desempenho real. Vale ressaltar que são considerados os 37 exemplos positivos de TLINKs entre EVTs e ETs em sentenças distintas, que foram totalmente descartados no desenvolvimento dos classificadores. Esses TLINKs tinham padrões distintos, não sendo possível criar heurísticas para limitação dos pares, além de serem poucas as ocorrências. Alguns dos casos só ocorriam devido à maneira bem específica em que o texto foi anotado, como, por exemplo, relações entre problemas diagnosticados em exames em sentenças distintas devido à quebra de linhas ou pontuações entre resultados, não existindo uma menção específica da data. Nesse cenário, o usual seria existir uma menção do exame, como eletrocardiograma ou menção geral de exame laboratorial, sendo, assim, um TLINK entre EVTs, porém houve essa particularidade na escrita do texto. Outro caso bem específico foi a questão de duração de sintomas: alguns profissionais separam o problema (sintoma) da duração por pontos, independentemente de serem aspectos relacionados ao mesmo sintoma, dentro do mesmo relato do paciente.

Na Tabela 23, são mostrados todos os TLINKs, tanto no contexto entre EVTs quanto entre EVT e ET, totalizando 521 exemplos positivos no conjunto de teste. Nesta seção, será verificado o resultado do sistema para extração deles.

Tabela 23 – Quantidade de TLINKs por categoria no conjunto de treinamento e teste, assim como o valor total.

Relação	Categoria	Conjunto de treinamento	Conjunto de teste	Total
TLINK	EVT e EVT	850 (53,32%)	308 (59,12%)	1.158 (54,75%)

EVT e ET	744 (46,68%)	213 (40,88%)	957 (45,25%)
Total	**1.594 (100%)**	**521 (100%)**	**2.115 (100%)**

Fonte: O autor (2020).

A partir de todos os modelos de melhor resultado em todas as propostas, foram feitos os testes presentes na Tabela 24. Ressalta-se que “Heurísticas propostas” (mesma sentença) envolve somente “Heurística EVT-EVT MS”, enquanto “Heurísticas propostas” (sentenças distintas), tanto “Heurística EVT-EVT MS” quanto “Heurística EVT-EVT SD”.

Tabela 24 – Resultados finais obtidos pelo sistema em diferentes configurações dos melhores modelos no conjunto de teste.

Modelo	***Precision***	***Recall***	**F1-*score***	**Suporte**
Sem heurística	0,7506	0,5893	0,6602	521
Pares Esquerda-Direita	0,8237	0,6545	0,7294	521
Heurísticas propostas (mesma sentença)	0,7565	0,6142	0,6780	521
Heurísticas propostas (sentenças distintas)	0,7314	0,7159	0,7236	521
Heurísticas propostas + Pares Esquerda-Direita (mesma sentença)	0,8430	0,7006	0,7652	521
Heurísticas propostas + Pares Esquerda-Direita (sentenças distintas)	0,8268	0,8157	**0,8213**	521

Fonte: O autor (2020).

A partir dos resultados presentes na Tabela 24, foi verificado que o melhor modelo geral foi obtido pela combinação das heurísticas propostas neste projeto com a heurística “Pares Esquerda-Direita”, proposta por Tourille *et al.* (2016). Na Tabela 25, é verificado o desempenho desta abordagem para todos os tipos de marcação. No geral, o melhor desempenho foi obtido para marcação Contido_Por, seguida de Contém e Sobreposto. As demais marcações, que tinham um baixo suporte, tiveram o pior desempenho na classificação.

Tabela 25 – Resultados finais obtidos para extração de TLINKs no conjunto de teste, trazendo as métricas de avaliação por marcação.

Marcação	***Precision***	***Recall***	**F1-*score***	**Suporte**
Depois	0,6667	0,6667	0,6667	3
Antes	0,5000	0,4167	0,4545	12
Contido_Por	0,8214	0,9079	0,8625	76
Contém	0,8557	0,8300	0,8426	200
Sobreposto	0,8206	0,7957	0,8079	230
Micro	0,8268	0,8157	0,8213	521

Fonte: O autor (2020).

Ao final, foi obtido um micro F1-*score* de 0,8213 para extração de TLINKs no conjunto de teste, assim como 0,9270 para extração de RelTempDCD.

7 DISCUSSÃO

Este capítulo é organizado em três seções. Na primeira, são discutidas questões da criação do *guideline*, anotação e seus respectivos resultados, além de fornecer detalhes sobre os procedimentos de anotação e suas limitações. Na segunda, o foco muda para a extração de RTs, trazendo aspectos relevantes sobre a criação dos modelos, conclusões sobre o processo de extração de RTs e limitações. Na terceira, ocorre o fechamento da tese, sendo trazidas as considerações finais.

7.1 CRIAÇÃO DOS *GUIDELINES* E ANOTAÇÃO

A extração de relações é um tema de pesquisa dependente de anotações prévias; sem conceitos a ser relacionados, não existem relações. Isso adiciona uma dificuldade para criação de *corpora* disponíveis para pesquisa. Para extração de TLINKs, existem camadas de anotação de EVTs e ETs, sendo estas fundamentais na definição daquelas. Por exemplo, considerando o esquema de anotação de TLINKs deste projeto, na sentença “lab 29/04/15: TSH 15, T4L 0,97” existiriam relações entre “lab” e “29/04/15”, com a menção geral do exame laboratorial contida na data, assim como os exames “TSH” e “T4L” contidos na menção geral. Se no conjunto adjudicado da anotação de EVTs não existisse a menção “lab”, devido a alguma discordância ou simples erro de anotação/adjudicação, o termo central dessa frase, menção de “lab”, seria perdido e, com isso, diversas das marcações corretas não poderiam ser anotadas durante a anotação de TLINKs, trazendo um viés na anotação desse texto.

Um dos objetivos específicos deste projeto foi a criação de *guidelines* para as anotações de EVTs, ETs e RTs; devido a essas questões, foi proposto desenvolver *guidelines* claros e extensos sobre o que deve ou não ser anotado em cada uma das anotações, assim como justificativas. Se o anotador entende o porquê de aquela anotação ser importante naquele contexto, se torna menos propício a esquecê-la durante a marcação. Buscou-se obter um *guideline* visando a dois aspectos: (i) replicabilidade; (ii) alta cobertura.

A questão de replicabilidade é essencial, principalmente em um cenário de expansão futura das anotações do *corpus* ou troca de anotadores. A busca por esses dois aspectos levou a *guidelines* extensos, detalhados e com uma grande quantidade de exemplos sobre cada tipo de anotação.

Essa estratégia de criação dos *guidelines* pode ser considerada bem-sucedida, tomando por base os valores de IAA positivos obtidos para todas as camadas de anotação durante a criação do *corpus*. O segundo objetivo específico envolveu a criação de *corpus* anotado com EVTs, ETs e RTs, completado ao término da adjudicação.

Nas seções seguintes, serão trazidos aspectos relacionados aos *guidelines* e anotações para cada camada de anotação, informando aspectos efetivos e questões que envolvem melhorias.

7.1.1 Criação do *guideline* e anotação de eventos

Para anotação de EVTs, utilizar o atributo Tipo para tornar a definição dos EVTs mais específica foi uma estratégia efetiva. A vantagem de utilizar tipos específicos de EVT envolve a maior compreensão, mediante tipos intuitivos e de fácil entendimento, além de proporcionar uma caracterização mais efetiva para cada marcação para cada tipo. A mesma abordagem foi utilizada por Viani *et al.* (2019), tendo sido obtidos valores de IAA similares ao deste projeto na configuração de *partial matching*. Tanto na anotação de Viani *et al.* (2019) quanto nesta, o atributo Tipo foi adaptado a partir do esquema utilizado na *shared task* i2b2 2012. A partir dos valores de IAA já relatados na seção 6.5, evidenciam-se valores positivos em comparação aos trabalhos relacionados, sendo esse aspecto mais evidente para marcação do texto (*span*) e principalmente para RelTempDCD.

Neste projeto, uma maior concordância entre os anotadores foi obtida na marcação textual das menções (*span*) de Testes (valor de IAA de 0,9478 em *strict matching*) e Evidências (valor de IAA de 0,9744 em *strict matching*). Isso ocorreu devido a marcações de Testes serem mais simples, sendo menções gerais de exame, exames de diagnósticos ou exames laboratoriais, sendo esse último mais facilmente anotado e de grande volume nos textos. Já Evidências têm forte relação com menções de um único *token*, como “nega”, “relata” ou “mostra”, sendo estas de mais fácil anotação. Um aspecto positivo desta anotação foi a obtenção de valores parecidos de IAA para Problemas, Tratamentos e Ocorrências, indicando que a estratégia baseada na documentação clínica por meio de OLD CARTS foi efetiva para delimitação de Problemas, principalmente para sintomas. Apesar disso, as divergências durante o processo de anotação dupla ainda estiveram relacionadas a essa questão de

delimitação dos Problemas, com marcações como “dispneia aos moderados esforços” sendo consideradas como “dispneia” por um anotador e “dispneia aos moderados esforços” por outro.

Outro aspecto que trouxe discordâncias foram marcações bem específicas presentes em alguns textos, como “pés incham”, que não estavam abordados no *guideline* e não eram frequentes. Dentre todos os tipos, menções de Departamentos Clínicos tiveram o pior valor de IAA, sendo verificada uma discordância na anotação de termos como “cardio”, “cardiologia”, “cardio geral”, “posto” e “UPA”, apesar de isso estar bem especificado no *guideline*. Como foram anotadas somente 49 marcações de Departamentos Clínicos, essas discordâncias impactaram no valor de IAA (0,8391 na configuração de *strict matching*).

Para o atributo RelTempDCD, o formato estruturado do SOAP beneficiou a anotação. Ressalta-se que, apesar de terem sido utilizadas as mesmas categorias de anotação do THYME *corpus*, foram obtidas proporções para cada marcação diferente. Na anotação deste projeto, houve um percentual bem maior de anotações de Antes/Sobreposto, provavelmente devido à questão de considerar a marcação Antes/Sobreposto em casos em que o EVT tenha uma caracterização de continuidade, mesmo que implícita, como no caso de doenças crônicas, comorbidades, medicamentos em uso e menções de Tratamentos, como “marca-passo”. Houve menos marcações de Sobreposto, uma vez que somente questões relacionadas a menções de relato do paciente e condições referentes ao exame físico e visual eram consideradas Sobreposto. Pelo valor de IAA obtido nesta anotação, conclui-se que a estratégia de se basear no SOAP e criar padrões de anotação baseados em como profissionais de saúde interpretam o documento foi efetiva, principalmente em um cenário em que profissionais da saúde anotam os textos.

A limitação da anotação de EVTs deste projeto está relacionada com a marcação do texto (*span*). Para Problemas, existem certas informações importantes que deveriam ser consideradas para a entidade fazer sentido, porém estão representadas de maneira disjunta. Por exemplo, no trecho “VE = hipertrofiado, cavidade, função sistólica e contração preservadas, alt de relaxamento”, seria importante ter uma marcação relacionando “alt de relaxamento” (alteração de relaxamento) com “VE” (ventrículo esquerdo) no exame, apesar de estarem separados no texto. Por exemplo, no *corpus* da *shared task* do ShARe/CLEF 2013 (SUOMINEN *et al.*, 2013), foram anotadas menções disjuntas para um tipo de menção

específico. No contexto de textos clínicos, o conceito de entidades disjuntas pode inclusive ser ampliado para questão de sintomas, uma vez que, em contexto como “dor torácica diariamente, tipo fisgada”, a ET “diariamente” torna impossível uma marcação conjunta de “dor torácica” e “tipo fisgada” (uma caracterização da dor).

7.1.2 Criação do *guideline* e anotação de expressões temporais

Para a anotação de ETs, houve um aspecto facilitador: utilizar como base o *guideline* construído por Azevedo (2019), assim como ter a possibilidade de fazer levantamentos das anotações no *corpus* anotado por Azevedo (2019) auxiliou enormemente no processo de criação do *guideline* e identificação das questões a ser modificadas.

A questão levantada por Azevedo (2019) e Viani *et al.* (2019) relacionada à dificuldade de anotadores diferenciarem ETs entre Datas e Durações foi resolvida pela consideração do valor do atributo RelTempDCD durante a marcação de ETs. Uma vez que, se baseando no atributo RelTempDCD, o anotador não precisava mais interpretar o EVT associado àquela ET, somente verificar o valor do atributo, isso causava menos conflito devido a diferenças de interpretação. Só foi possível utilizar esta estratégia devido à anotação de ETs ter ocorrido após a adjudicação dos lotes da anotação de EVTs. No geral, pelos valores de IAA obtidos durante todo o processo, conclui-se que o processo de anotação de ETs foi positivo.

Uma questão de melhoria envolveria a normalização de ETs. Em certos contextos, como “IAM há 10 anos” ou “última consulta há 3 meses”, existe uma questão de imprecisão presente. Usualmente, esse tipo de menção, apresentado durante a seção Subjetiva do SOAP, apresenta certo grau de incerteza na normalização. Quando se menciona “há 10 anos”, isso é um valor aproximado, podendo ser no começo ou no final do décimo ano ou até estar relacionado a outro período (por exemplo, no décimo segundo ano), havendo nesta anotação um arredondamento para dez anos. Sendo assim, normalizar para “exatamente 10 anos atrás” acabava sendo uma estimativa incorreta. Em um cenário de múltiplas consultas do mesmo paciente, poderiam existir casos de o mesmo EVT começando em momentos distintos a cada documento devido à normalização.

7.1.3 Criação do *guideline* e anotação de relações temporais

O *guideline* de TLINKs foi complexo de ser criado. Foram utilizadas as categorias do THYME *corpus*, porém os *guidelines* foram criados para melhor atender às características dos textos. Existem certas questões particulares de textos de cardiologia, como no trecho “IAM + ATC”, em que há a relação implícita de que “IAM” veio antes de “ATC”; estas precisam ser levadas em conta na criação dos *guidelines*.

De acordo com Styler *et al*. (2014a), a abordagem de *narrative containers*, proposta por Pustejovsky e Stubbs (2011), foi utilizada para anotação do THYME *corpus* por funcionar bem com a estrutura de narrativa contida em textos clínicos e de domínio geral, sendo útil para interpretar *timelines*. No entanto, no contexto de textos ambulatoriais, não existe uma estrutura de narrativa; muito da informação está contida na sentença, não tendo qualquer relação com a sentença ao lado. Outro aspecto importante é a presença de sentenças extremamente curtas, como “nega dispneia” ou “HAS”, diferentemente de textos de domínio geral, em que existem sentenças mais longas e mais bem estruturadas. Apesar de não existir uma questão de narrativa, usualmente há EVTs e ETs centrais na frase ou contexto, sendo esse aspecto notado, por exemplo, em menções de exames de laboratório ou exames de diagnóstico. Sendo assim, o objetivo foi anotar o *corpus* baseado nesta abordagem.

Pelos valores de IAA obtidos no *corpus*, foi verificado que a estratégia de anotação foi efetiva, com os *guidelines* sendo representativos e fornecendo exemplos suficientes e adequados. Foram obtidos valores de IAA positivos frente à literatura, porém é necessário levar em consideração que, em um contexto de sentenças mais longas e de estrutura narrativa, podem existir marcações mais complexas em nível de sentença ou texto, aumentando o grau de dificuldade da anotação.

Em relação às categorias de TLINKs anotadas, verificou-se uma baixa utilização das categorias Começa_Em, Termina_Em e Antes. A categoria Termina_Em só foi utilizada em uma marcação e Começa_Em, em 19 marcações. Assim, quando esses exemplos eram divididos entre os três classificadores de TLINKs, não existiam exemplos suficientes para o treinamento em 10-*fold cross-validation*, com esses dois tipos de marcação usualmente sendo fundidos/mesclados com a marcação Sobreposto, similar ao procedimento realizado para o *corpus* i2b2 2012 (SUN; RUMSHISKY; UZUNER, 2013a). Essas marcações poderiam ser totalmente descartadas ou já adicionadas à categoria Sobreposto em uma nova

versão do *guideline*. Marcações da categoria Antes eram mais frequentes, mas bem menos frequentes que as categorias majoritárias Sobreposto e Contém. Nesse contexto, não é proposto descartar esse tipo de relação, pois são extremamente importantes no contexto clínico, servindo para especificar questões como ordem de tratamentos ou relações entre problemas e tratamentos realizados.

Uma proposta de modificação no *guideline* seria adicionar uma categoria para melhor representação de marcações simultâneas. Por exemplo, na sentença “HAS em acompanhamento há 2 anos”, existem os TLINKs: (i) “HAS” Sobreposto “acompanhamento”; (ii) “HAS” Sobreposto “há 2 anos”; (iii) “acompanhamento” Sobreposto “há 2 anos”. Nesse contexto, existe o mesmo tipo de marcação entre os EVTs “acompanhamento” e “HAS” com a ET “há 2 anos”, apesar de “HAS” ocorrer há mais de dois anos e “acompanhamento”, há exatamente dois anos, existindo uma característica simultânea que não é representada na marcação.

7.2 MODELO PARA EXTRAÇÃO DE RELAÇÕES TEMPORAIS

Nesta seção, são sumarizados os resultados relacionados ao desempenho de cada classificador, assim como o desempenho geral.

7.2.1 Modelo para extração de TLINKs entre eventos em mesma sentença

Para este tipo de TLINK, o melhor modelo envolveu o uso das duas heurísticas, “Pares Esquerda-Direita” e “Heurística EVT-EVT MS”. A heurística “Pares Esquerda-Direita” se mostrou efetiva para reduzir a quantidade de pares candidatos, diminuindo pela metade; no entanto, ainda sobraram muitos pares negativos entre EVTs do tipo Teste, que só possuíam pares positivos em contextos específicos. Sendo assim, combinando ambas as abordagens, foi obtido o melhor desempenho no conjunto de treinamento e teste.

Apesar de um F1-*score* de 0,7810, o resultado ainda foi inferior ao modelo para a extração de TLINKs entre EVT e ET em mesma sentença, fato verificado também nos demais trabalhos de literatura. Esse mesmo aspecto foi observado nos trabalhos de Lin *et al.* (2016b, 2017) e Dligach *et al.* (2017) em experimentos no THYME *corpus*.

Neste *corpus* anotado, existe uma média de 6,90 ETs e 31,87 EVTs por texto, ou seja, mesmo considerando somente TLINKs em mesma sentença, muitos dos

pares criados para TLINKS entre EVTs acabam sendo pares negativos. Esse fator impacta na extração de TLINKs pelo classificador.

7.2.2 Modelo para extração de TLINKs entre eventos e expressões temporais em mesma sentença

Para este tipo de TLINK, a abordagem “Pares Esquerda-Direita”, única heurística testada, foi efetiva, tanto no conjunto de treinamento quanto de teste, inclusive, existindo uma diferença de somente 0,002 do F1-*score* do treinamento para o de teste, tendo sido obtido um F1-*score* de 0,9057 neste. Vale ressaltar que valores de *Recall* altos para as categorias Contido_Por (0,9848) e Sobreposto (0,9524) indicam uma alta cobertura dos casos positivos.

A abordagem “Pares Esquerda-Direita” é propícia para esse tipo de texto, pois os padrões encontrados são claros. Usualmente, menções de exames, tanto gerais de exames laboratórios quanto de diagnóstico, estão situadas à esquerda da ET. Isso faz com que seja trocado o tipo de TLINK para Contido_Por, bem específico para esses casos. Uma parte substancial dessas marcações de Sobreposto envolve a relação entre medicamentos e frequência de uso, ET do tipo Frequência, bem específico de marcação, um dos motivos pela alta cobertura de casos.

Esse foi o melhor classificador entre os especializados para TLINKs. A maior parte dos trabalhos relacionados recentes não apresenta valores para componentes específicos, sendo usualmente somente apresentado um valor final. Devido a esse fato, foi realizada uma comparação com o valor obtido por Dligach *et al.* (2017). Para tal, foi necessário verificar o desempenho geral do componente para extração de TLINK entre EVT e ET, levando em conta também os pares descartados entre EVT e ET em sentenças distintas. Após esse procedimento, foi obtido um valor de F1-*score* de 0,8930, positivo em relação ao valor de 0,70 obtido por Dligach *et al.* (2017) em experimentos no *corpus* Clinical TempEval 2016 com um modelo baseado em *convolutional neural network*.

7.2.3 Modelo para extração de TLINKs entre eventos em sentenças distintas

Este tipo de TLINK é usualmente o mais afetado com a questão dos pares candidatos; qualquer aumento no número de sentenças aumenta consideravelmente

a quantidade de pares, tanto que determinados autores ignoram todos os TLINKs em sentenças distintas.

O melhor modelo envolveu o uso de duas heurísticas, “Pares Esquerda-Direita” e “Heurística EVT-EVT SD”, em conjunto. Essa combinação teve melhor desempenho do que somente considerar “Heurística EVT-EVT SD”. A motivação por trás desta surgiu com base em conversas com profissionais da área da saúde, em que foi evidenciado que seria importante trazer TLINKs entre Problemas e Testes (Testes em que foram encontrados esses Problemas) em sentenças distintas, sendo esse tipo de TLINK importante para compreensão da *timeline* do paciente. Além disso, foram adicionados alguns filtros para captar outros tipos específicos de TLINK. Foi verificado que esse classificador, na sua melhor configuração obtida no conjunto de treinamento, não conseguiu ter desempenho similar no conjunto de teste, com uma queda de 0,15 em nível de classificador no F1-*score* do conjunto de treinamento para o conjunto de teste.

Devido às heurísticas serem bem específicas, considerando o atributo RelTempDCD, o atributo Tipo e *tokens* da menção, não houve uma generalização tão efetiva para o conjunto de teste.

7.2.4 Modelo para extração de RelTempDCD

Para a extração de RelTempDCD, não existiram questões complexas como criação de pares candidatos ou determinações de heurísticas, sendo um problema de classificação de um atributo de EVT, em que este pode ter somente uma de quatro possíveis marcações. Há algumas marcações, como Antes e Antes/Sobreposto, que podem ser confusas devido à dependência de interpretação de continuidade de determinado EVT, porém, no geral, as categorias são bem distanciadas entre si.

Foi observado por Lee *et al.* (2016), durante experimentos no *corpus* Clinical TempEval 2016, que seu melhor desempenho foi para a classe Sobreposto. Esse mesmo aspecto ocorreu com os experimentos deste *corpus*, por ser uma categoria de marcação relacionada a um grupo de EVTs muito específicos, que ocorreram durante a DCD, como EVTs referentes a exames físicos, e EVTs do tipo Evidências relacionados a relatos do paciente (por exemplo, “nega” e “relata”).

O melhor modelo obteve F1-*score* de 0,9388 no conjunto de treinamento e 0,927 no conjunto de teste. Esse valor é positivo quando comparado ao melhor

resultado reportado para o *corpus* Clinical TempEval 2016 (0,87) por Tourille *et al.* (2017a).

Durante experimentos com um *corpus* de textos cardiológicos para italiano, Viani *et al.* (2019) obtiveram valores de F1-*score* de 0,857 para Sobreposto, 0,834 para Antes e 0,793 para Depois. Já nesta tese, os valores por marcação foram de 0,9688 para Sobreposto, 0,9278 para Antes, 0,8596 para Depois e 0,9268 para Antes/Sobreposto, tendo sido obtidos valores positivos para todas as marcações.

7.2.5 Desempenho geral

Nesta seção, são sumarizados aspectos sobre o desempenho geral do sistema, feitas comparações com trabalhos similares, além de observações em nível geral sobre limitações e futuros trabalhos.

Como evidenciado anteriormente, o melhor sistema envolveu ambas as heurísticas propostas nesta tese (“Heurística EVT-EVT MS” e “Heurística EVT-EVT SD”), assim como “Pares Esquerda-Direita”, proposta por Tourille *et al.* (2016) e amplamente utilizada em trabalhos envolvendo o THYME *corpus*. Pelos resultados obtidos, a heurística “Pares Esquerda-Direita” (F1-*score* de 0,7294), sozinha, já teve resultados superiores no conjunto de testes em relação às propostas neste projeto concatenadas (F1-*score* de 0,7236). Esse aspecto é relevante, pois, nesse cenário, o modelo envolvendo a heurística “Pares Esquerda-Direita” ignora TLINKs entre EVT e ET em sentenças distintas, enquanto o modelo com ambas as heurísticas propostas nesta tese as considera; ainda assim, houve desempenho inferior.

Adicionando somente “Heurística EVT-EVT MS” ao modelo envolvendo “Pares Esquerda-Direita”, houve melhora do F1-*score* para 0,7652. Adicionando ambas as heurísticas propostas, o valor atingiu 0,8213, o melhor resultado alcançado para extração de TLINK nesta tese.

Vale ressaltar que, apesar de o componente envolvendo TLINKs entre EVTs em sentenças distintas ter tido pior desempenho no conjunto de teste, considerando os demais componentes, foi um classificador eficiente. Uma estratégia comum para extração de TLINK envolvia descartar totalmente aqueles em sentenças distintas. Então, qualquer modelo que tente a extração, mesmo que o resultado não seja dos melhores, é positivo, uma vez que qualquer acerto contribui para a melhora do resultado. Outro aspecto relevante é que não foi encontrado nenhum modelo proposto

na literatura para trabalhar com uma janela de 20 sentenças, como nesta tese, ou seja, tentando buscar a maior quantidade possível de informações de determinados tipos considerados relevantes.

Outro aspecto pertinente de ser mencionado é a questão de a heurística “Pares Esquerda-Direita” ser efetiva neste *corpus*. Nos trabalhos encontrados, era usualmente aplicada somente para o *corpus* Clinical TempEval 2016, considerando somente TLINKs de tipo Contém. Apesar de neste projeto tipos adicionais serem considerados, esta abordagem se mostrou efetiva em todos os cenários testados.

Apesar das distintas características deste *corpus* em comparação ao THYME (diferente especialidade, tamanho e escrita dos textos), comparando o melhor modelo obtido neste projeto com o melhor modelo obtido na literatura, encontrado durante a revisão sistemática, evidencia-se um resultado positivo em relação aos trabalhos de Lin *et al.* (2019), com F1-*score* de 0,684, e Lin *et al.* (2018), com F1-*score* de 0,630, ambos detalhados na revisão sistemática do Apêndice A.

As limitações do estudo envolveram baixa quantidade de textos anotados, com dificuldade de obtenção de textos adicionais para a especialidade cardiológica. Outra limitação relacionou-se ao método de extração, apesar de a proposta ser analisar tipos de TLINK distintos, verificando quais características de cada classificador especializado e quais fatores beneficiam a *performance* destes. Ainda assim, pode ser considerada uma limitação não ter modelos-base relacionados com aprendizagem profunda.

Por último, um trabalho futuro seria considerar a questão do *closure* da relação. Alguns trabalhos geram uma expansão do conjunto de treinamento por *closure*, criando TLINKs adicionais que poderiam ser inferidos. Trabalhos como de Galvan *et al.* (2018) e Tourille *et al.* (2016) não aplicaram *closure* para inferir TLINKs. Apesar de ser uma abordagem utilizada em trabalhos relacionados a *corpora* disponibilizados por *shared tasks*, como i2b2 2012 e Clinical TempEval 2016, uma análise mais apurada é necessária antes da sua aplicação. Como indicado por Derczynski (2016), apesar de o *closure* ser uma técnica útil para aumentar o volume de dados, os resultados de adicioná-los por ela podem ser imprevisíveis, não sendo sempre indicada. Nesta tese, não foi aplicado *closure* nas avaliações ou no treinamento, porém foi especificado para os anotadores sempre marcarem todos os pares que pudessem ser inferidos na sentença em casos de marcação de TLINKs entre EVTs

do tipo Sobreposto. No entanto, não foi realizado um estudo mais aprofundado sobre o tema.

7.3 CONSIDERAÇÕES FINAIS

Nesta tese, foram realizadas diversas etapas para atingir o objetivo final do projeto. A primeira envolveu uma revisão sistemática, que permitiu uma maior compreensão sobre o tema e seleção de autores-chave. A segunda contemplou a criação de três *guidelines* e a realização de três processos de anotação, todos contando com estudantes de Medicina como anotadores. Ressalta-se que o processo de anotação e os *guidelines* criados são replicáveis para outras especialidades, uma vez que começaram a ser criados sem definições relacionadas à cardiologia, podendo servir de ponto de partida para trabalhos futuros.

Os resultados da anotação foram positivos, obtendo-se valores de IAA positivos em comparação aos trabalhos relacionados. Para a extração de RTs, mesmo com um *corpus* pequeno, foi possível conseguir resultados satisfatórios em comparação com a literatura. Apesar das diferenças dos textos e das especialidades frente aos trabalhos da literatura, foram obtidos valores de F1-*score* positivos para todos os componentes, além de ter sido comprovada a eficácia das heurísticas para redução de pares, tanto propostas nesta tese quanto da literatura.

Para que o modelo gerado possa ser implementado para extração de RTs em um cenário real, em alguma aplicação específica, seria necessário ter métodos para extração de EVTs e ETs, pois, nesta tese, é usado o padrão ouro para EVTs e ETs.

Um trabalho futuro seria um estudo de ablação para verificar o efeito das *features* na extração, identificando o impacto de cada uma no processo de extração de RTs para cada componente. Esse seria um estudo interessante, pois traria detalhes adicionais sobre cada tipo de RT e que tipo de informação melhor beneficia a extração de RTs. Outro trabalho futuro seria aumentar o número de exemplos para os classificadores por meio de novos documentos anotados ou técnicas de *data augmentation*.

Foram identificadas diversas lacunas que poderiam ser preenchidas por esta tese na introdução. A lacuna de nenhum estudo para o idioma português foi preenchida ao fim de todas as etapas da pesquisa. A relacionada à falta de *corpora* com os três níveis de anotações propostos (EVTs, ETs e RTs) foi preenchida ao final

das etapas 1 a 5. A lacuna de nenhum estudo com os três níveis de anotação propostos para domínio cardiológico foi preenchida ao final das etapas 1 a 5. A lacuna da falta de estudos envolvendo a extração de RTs para textos reais (ruidosos) foi preenchida ao final de todas as etapas da pesquisa.

Os objetivos específicos de elaborar *guidelines* para domínio clínico e de elaborar um *corpus* anotado com EVTs, ETs e RTs a partir de textos ambulatoriais de cardiologia foram completados, fato verificado pelos valores de IAA obtidos e pelos resultados de extração de RTs. Os objetivos específicos de criar modelos especializados para extração de RTs e de avaliar os modelos desenvolvidos e heurísticas propostas foram alcançados, tendo sido verificado que as heurísticas e modelos foram efetivos, com resultados significativos frente à literatura.

A pergunta de pesquisa desta tese – como pode ser feita a extração automática de RTs de narrativas clínicas para um *corpus* da área clínica anotado em português para auxiliar no sequenciamento de eventos clínicos? – foi respondida, sendo evidenciado que a extração de RTs pode ser feita mediante *guidelines* bem definidos, um processo de anotação consistente (nos três níveis) e um sistema de extração com componentes especializados (heurísticas específicas).

O objetivo geral de construir um modelo para extração de RTs em narrativas clínicas em língua portuguesa foi completado, assim como todos os objetivos específicos propostos.

REFERÊNCIAS

ALEIXO, P.; PARDO, T. A. S. P. **CSTNews**: um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory). [S.l.: s.n.], 2008.

ALLEN, J. F. Maintaining knowledge about temporal intervals. **Readings in Qualitative Reasoning About Physical Systems**, [s.l.], v. 26, n. 11, p. 361-372, 1983.

ALMEIDA, A. M. G. *et al.* Applying multi-label techniques in emotion identification of short texts. **Neurocomputing**, [s.l.], v. 320, p. 35-46, 2018.

ALPAYDIN, E. **Introduction to machine learning**. London: The MIT Press, 2014.

ANDRADE, P. J. N. *et al.* Angioplastia coronariana versus cirurgia de revascularização: revisão de estudos randomizados. **Arquivos Brasileiros de Cardiologia**, Rio de Janeiro, v. 97, n. 3, p. e60-e69, 2011.

ARSTEIN, R. Inter-annotator agreement. *In*: IDE, N. (Ed.). **Handbook of linguistic annotation**. Dordrecht: Springer Netherlands, 2017.

AYAT, N. E.; CHERIET, M.; SUEN, C. Y. Automatic model selection for the optimization of SVM kernels. **Pattern Recognition**, [s.l.], v. 38, n. 10, p. 1733-1745, 2005.

AZEVEDO, R. F. **Temporal tagging of noisy clinical texts written in Brazilian Portuguese**. 2019. Dissertation (Master in Informatics) – Pontifícia Universidade Católica do Paraná, Curitiba, 2019.

BAPTISTA, J.; HAGÈGE, C.; MAMEDE, N. J. Identificação, classificação e normalização de expressões temporais do português: a experiência do Segundo HAREM e o futuro. *In*: MOTA, C.; SANTOS, D. (Ed.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas**: o Segundo HAREM. Porto: Linguateca, 2008. p. 33-54.

BARROS, M. *et al.* ULISBOA at SemEval-2016 task 12: extraction of temporal expressions, clinical events and relations using IBEnt. *In:* INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings […]**. [S.l.: s.n.], 2016. p. 1263-1267.

BEN-HUR, A.; WESTON, J. A user's guide to support vector machines. *In*: CARUGO, O.; EISENHABER, F. (Ed.). **Data mining techniques for the life sciences**. [S.l.]: Humana Press, 2010. p. 223-239.

BENNETT, J. A. *et al.* Validity and reliability of the NYHA classes for measuring research outcomes in patients with cardiac disease. **Heart and Lung: Journal of Acute and Critical Care**, [s.l.], v. 31, n. 4, p. 262-270, 2002.

BETHARD, S. ClearTK-TimeML: a minimalist approach to TempEval 2013. *In*: JOINT

CONFERENCE ON LEXICAL AND COMPUTATIONAL SEMANTICS, 2., 2013, Atlanta. **Proceedings [...]**. [S.l.: s.n.], 2013. v. 2. p. 10-14.

BETHARD, S. *et al.* SemEval-2015 task 6: Clinical TempEval. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 9., 2015, Denver. **Proceedings [...]**. [S.l.: s.n.], 2015. p. 806-814.

BETHARD, S. *et al.* SemEval-2016 task 12: Clinical TempEval. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1052-1062.

BETHARD, S. *et al.* SemEval-2017 task 12: Clinical TempEval. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 11., 2017, Vancouver. **Proceedings [...]**. [S.l.: s.n.], 2017. p. 565-572.

BICKLEY, L.; SZILAGYI, P. G. **Bates' guide to physical examination and history-taking**. Lippincott: Williams & Wilkins, 2012.

BODENREIDER, O. The unified medical language system (UMLS): integrating biomedical terminology. **Nucleic Acids Research**, [s.l.], v. 32, n. Suppl. 1, p. D267-D270, 2004.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. *In*: ANNUAL WORKSHOP ON COMPUTATIONAL LEARNING THEORY, 5., 1992, Pittsburgh. **Proceedings [...]**. [S.l.: s.n.], 1992. p. 144-152.

BOUCKAERT, R. R.; FRANK, E. Evaluating the replicability of significance tests for comparing learning algorithms. *In*: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 8., 2004, Sydney. **Proceedings [...]**. Berlin: Springer, 2004. p. 3-12.

BRAMSEN, P. *et al.* Finding temporal order in discharge summaries. *In*: AMIA ANNUAL SYMPOSIUM, 2006, Washington, DC. **Proceedings [...]**. [S.l.]: AMIA, 2006a.

BRAMSEN, P. *et al.* Inducing temporal graphs. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2006, Sydney. **Proceedings [...]**. [S.l.: s.n.], 2006b. p. 189-198.

CAMERON, S.; TURTLE-SONG, I. Learning to write case notes using the SOAP format. **Journal of Counseling and Development**, [s.l.], v. 80, n. 3, p. 286-292, 2002.

CAMPILLOS, L. *et al.* A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). **Language Resources and Evaluation**, [s.l.], v. 52, n. 2, p. 571-601, 2018.

CAPURRO, D. *et al.* Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multi-site

assessment. **eGEMs**, [s.l.], v. 2, n. 1, p. 7-11, 2014.

CASELLI, T.; MORANTE, R. Vuacltl at semeval 2016 task 12: a CRF pipeline to clinical tempeval. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1241-1247.

CHANG, Y. C. *et al.* TEMPTING system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. **Journal of Biomedical Informatics**, [s.l.], v. 46, n. Suppl., p. S54-S62, 2013.

CHEN, C. *et al.* A comprehensive analysis of detection of online paid posters. *In*: ULUSOY, Ö.; TANSEL, A. U.; ARKUN, E. (Ed.). **Recommendation and search in social networks**. Berlin: Springer, 2015. p. 101-118.

CHEN, Q. *et al.* An automatic system to identify heart disease risk factors in clinical texts over time. **Journal of Biomedical Informatics**, [s.l.], v. 58, p. S158-S163, 2015.

CHENG, Y. *et al.* Temporal relation discovery between events and temporal expressions identified in clinical narrative. **Journal of Biomedical Informatics**, [s.l.], v. 46, n. Suppl., p. S48-S53, 2013.

CHERRY, C. *et al.* À la recherche du temps perdu: extracting temporal relations from medical text in the 2012 i2b2 PLN challenge. **Journal of the American Medical Informatics Association**, [s.l.], v. 20, n. 5, p. 843-848, 2013.

CHIKKA, V. R. CDE-IIITH at SemEval-2016 task 12: extraction of temporal information from clinical documents using machine learning techniques. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1237-1240.

CHOLLET, F. **Deep learning with Python**. [S.l.]: Manning Publications, 2017.

COHAN, A.; MEURER, K.; GOHARIAN, N. Guir at semeval-2016 task 12: temporal information processing for clinical narratives. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1248-1255.

COLLOVINI, S.; GOULART, R.; VIEIRA, R. Identificação de expressões anafóricas e não anafóricas com base na estrutura do sintagma. *In*: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 2., 2004, Salvador. **Anais [...]**. [S.l.: s.n.], 2004.

CONSELHO FEDERAL DE ENFERMAGEM (COFEN). **Guia de recomendações para registro de enfermagem no prontuário do paciente e outros documentos de enfermagem**. Brasília, DF, 2016.

CORMACK, J. *et al.* Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge. **Journal of Biomedical Informatics**, [s.l.], v. 58, p. S120-S127, 2015.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, [s.l.], v. 297, n. 20, p. 273-297, 1995.

D'SOUZA, J.; NG, V. Classifying temporal relations in clinical data: a hybrid, knowledge-rich approach. **Journal of Biomedical Informatics**, [s.l.], v. 46, n. Suppl., p. S29-S39, 2013.

D'SOUZA, J.; NG, V. Knowledge-rich temporal relation identification and classification in clinical notes. **Database**, [s.l.], v. 2014, p. 1-20, 2014a.

D'SOUZA, J.; NG, V. Annotating inter-sentence temporal relations in clinical notes. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 9., 2014, Reykjavik. **Proceedings […]**. [S.l.: s.n.], 2014b. p. 2758-2765.

DALIANIS, H. Characteristics of patient records and clinical corpora. *In*: DALIANIS, H. **Clinical text mining**. Cham: Springer, 2018. p. 21-34.

DENG, L.; LIU, Y. (Ed.). **Deep learning in natural language processing**. Berlin: Springer, 2018.

DENNY, M.; SPIRLING, A. **Text preprocessing for unsupervised learning**: why it matters, when it misleads, and what to do about it. [S.l.: s.n.], 2017.

DERCZYNSKI, L. **Representation and learning of temporal relations.** *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 26., 2016, Osaka. **Proceedings […]**. [S.l.: s.n.], 2016. p. 1937-1948.

DEVLIN, J. *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2019, Minneapolis. **Proceedings […]**. [S.l.: s.n.], 2019. v. 1. p. 4171-4186.

DLIGACH, D. *et al.* Neural temporal relation extraction. **Eacl**, [s.l.], v. 2, p. 746-751, 2017.

DUDCHENKO, A.; GANZINGER, M.; KOPANITSA, G. Machine learning algorithms in cardiology domain: a systematic review. **The Open Bioinformatics Journal**, [s.l.], v. 13, n. 1, p. 25-40, 2020.

FAN, R. E. *et al.* LIBLINEAR: a library for large linear classification. **Journal of Machine Learning Research**, [s.l.], v. 9, n. 2008, p. 1871-1874, 2008.

FELDMAN, S. PLN meets the jabberwocky natural language processing in information retrieval. **Online (Wilton)**, [s.l.], v. 23, n. 3, p. 62-72, 1999.

FERRO, L. *et al.* **Tides temporal annotation guidelines**: version 1.0. 2. McLean: The MITRE Corporation, 2001.

FERRO, L. *et al.* Defining a state-of-the-art POS-tagging environment for Brazilian Portuguese clinical texts. **Research on Biomedical Engineering**, [s.l.], n. 3, p. 267-276, 2020.

FONSECA, E. *et al.* Corp: uma abordagem baseada em regras e conhecimento semântico para a resolução de correferências. **Linguamática**, Minho, v. 9, n. 1, p. 3-18, 2017.

FORT, K. **Collaborative annotation for reliable natural language processing**: technical and sociological aspects. [S.l.]: John Wiley & Sons, 2016.

FORT, K. *et al.* Towards a methodology for named entities annotation. *In:* LINGUISTIC ANNOTATION WORKSHOP, 3., 2009, Suntec. **Proceedings [...]**. [S.l.: s.n.], 2009. p. 142-145.

FRIES, J. A. Brundlefly at SemEval-2016 task 12: recurrent neural networks vs. joint inference for clinical temporal information extraction. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1274-1279.

GALVAN, D. *et al.* Investigating the challenges of temporal relation extraction from clinical text. *In*: INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS, 9., 2018, Brussels. **Proceedings [...]**. [S.l.: s.n.], 2018. p. 55-64.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

GOLDBERG, Y. Neural network methods for natural language processing. **Synthesis Lectures on Human Language Technologies**, [s.l.], v. 10, n. 1, p. 1-309, 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**: machine learning book. Massachusetts: MIT Press, 2016.

GOODWIN, T.; HARABAGIU, S. M. A probabilistic reasoning method for predicting the progression of clinical findings from electronic medical records. *In*: SUMMITS ON TRANSLATIONAL SCIENCE, 2015, San Francisco. **Proceedings [...]**. [S.l.]: AMIA, 2015.

GROUIN, C. *et al.* Eventual situations for timeline extraction from clinical reports. **Journal of the American Medical Informatics Association**, [s.l.], v. 20, n. 5, p. 820-827, 2013.

GROUIN, C.; MORICEAU, V. LIMSI at SemEval-2016 task 12: machine-learning and temporal information to identify clinical events and time expressions. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1225-1230.

GROUIN, C.; MORICEAU, V.; ZWEIGENBAUM, P. Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal

relations from clinical records. **Journal of Biomedical Informatics**, [s.l.], v. 58, p. S133-S142, 2015.

HE, H.; MA, Y. **Imbalanced learning**. [S.l.]: Wiley-IEEE, 2013.

HENRIKSSON, A. *et al.* Identifying adverse drug event information in clinical notes with distributional semantic representations of context. **Journal of Biomedical Informatics**, [s.l.], v. 57, p. 333-349, 2015.

HRIPCSAK, G.; ROTHSCHILD, A. S. Agreement, the F-measure, and reliability in information retrieval. **Journal of the American Medical Informatics Association**, [s.l.], v. 12, n. 3, p. 296-298, 2005.

HSU, C. C.; CHANG, C. C.; LIN, C. C. **A practical guide to support vector classification**. [S.l.: s.n.], 2003.

HUANG, P. Y. *et al.* NTU-1 at SemEval-2017 task 12: detection and classification of temporal events in clinical data with domain adaptation. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 11., 2017, Vancouver. **Proceedings [...]**. [S.l.: s.n.], 2017. p. 1010-1013.

IDE, N. Introduction. *In*: IDE, N. **Handbook of linguistic annotation**. Dordrecht: Springer Netherlands, 2017. p. 1-18.

JAGANNATHA, A. N.; YU, H. Bidirectional RNN for medical event detection in electronic health records. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016.

JAIN, A.; KULKARNI, G.; SHAH, V. Natural language processing. **International Journal of Computer Sciences and Engineering**, [s.l.], v. 6, n. 1, p. 161-167, 2018.

JENSEN, P. B.; JENSEN, L. J.; BRUNAK, S. Mining electronic health records: towards better research applications and clinical care. **Nature Reviews Genetics**, [s.l.], v. 13, n. 6, p. 395-405, 2012.

JOHNSON, K. W. *et al.* Enabling precision cardiology through multiscale biology and systems medicine. **JACC: Basic to Translational Science**, [s.l.], v. 2, n. 3, p. 311-327, 2017.

JONNAGADDALA, J. *et al.* Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. **BioMed Research International**, [s.l.], v. 2015, p. 1-10, 2015.

JORGE, M. L. C.; PARDO, T. Experiments with CST-based multidocument summarization. *In*: WORKSHOP ON GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 5., 2010, Uppsala. **Proceedings [...]**. [S.l.: s.n.], 2010. p. 74-82.

KAUARK, F. S.; MANHÃES, F. C.; MEDEIROS, C. H. **Metodologia da pesquisa**: um guia prático. Ibicaraí: Via Litterarum, 2010.

KHALIFA, A.; VELUPILLAI, S.; MEYSTRE, S. UtahBMI at SemEval-2016 task 12: extracting temporal information from clinical text. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1256-1262.

KONONENKO, I.; KUKAR, M. **Machine learning and data mining**. [S.l.]: Horwood, 2007.

KREIMEYER, K. *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. **Journal of Biomedical Informatics**, [s.l.], v. 73, p. 14-29, 2017.

KUBLER, S.; ZINSMEISTER, H. **Corpus linguistics and linguistically annotated corpora**. [S.l.]: Bloomsbury Academic, 2015.

KURDI, M. Z. **Natural language processing and computational linguistics**: speech, morphology and syntax. [S.l.]: John Wiley & Sons, 2016.

LANE, H.; HOWARD, C.; HAPKE, H. **Natural language processing in action**. [S.l.]: Manning, 2019.

LEAMAN, R.; KHARE, R.; LU, Z. Challenges in clinical natural language processing for automated disorder normalization. **Journal of Biomedical Informatics**, [s.l.], v. 57, p. 28-37, 2015.

LEE, H. J. *et al.* UTHealth at SemEval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1292-1297.

LEE, H. J. *et al.* Identifying direct temporal relations between time and events from clinical notes. **BMC Medical Informatics and Decision Making**, [s.l.], v. 18, n. Suppl 2, 2018.

LEE, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, [s.l.], n. September, p. 1-7, 2019.

LEEUWENBERG, A.; MOENS, M. F. Kuleuven-liir at semeval 2016 task 12: Detecting narrative containment in clinical records. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1280-1285.

LEEUWENBERG, A.; MOENS, M. F. Structured learning for temporal relation extraction from clinical records. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 15., 2017, Valencia. **Proceedings [...]**. [S.l.: s.n.], 2017a. v. 1. p. 1150-1158.

LEEUWENBERG, T.; MOENS, M. F. KULeuven-LIIR at SemEval-2017 Task 12: cross-domain temporal information extraction from clinical records. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 11., 2017, Vancouver. **Proceedings [...]**. [S.l.: s.n.], 2017b. p. 1030-1034.

LENERT, L. A. Toward medical documentation that enhances situational awareness learning. *In*: AMIA ANNUAL SYMPOSIUM, 2016, Chicago. **Proceedings [...]**. [S.l.]: AMIA, 2016.

LIDDY, E. D. Enhanced text retrieval using natural language processing. **Bulletin of the American Society for Information Science and Technology**, [s.l.], v. 24, n. 4, p. 14-16, 1998.

LIDDY, E. D. Natural language processing. **Encyclopedia of Library and Information Science**, [s.l.], v. 39, n. 1, p. 60-62, 2001.

LIN, C. *et al.* Descending-path convolution kernel for syntactic structures. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 52., 2014, Baltimore. **Proceedings [...]**. [S.l.: s.n.], 2014. v. 2. p. 81-86.

LIN, C. *et al.* Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. **Journal of the American Medical Informatics Association**, [s.l.], v. 22, n. e1, p. e151-e161, 2015.

LIN, C. *et al.* Multilayered temporal modeling for the clinical domain. **Journal of the American Medical Informatics Association**, [s.l.], v. 23, n. 2, p. 387-395, 2016a.

LIN, C. *et al.* Improving temporal relation extraction with training instance augmentation. *In:* WORKSHOP ON BIOMEDICAL NATURAL LANGUAGE PROCESSING, 15., 2016, Berlin. **Proceedings [...]**. [S.l.: s.n.], 2016b. p. 108-113.

LIN, C. *et al.* Representations of time expressions for temporal relation extraction with convolutional neural networks. *In:* WORKSHOP ON BIOMEDICAL NATURAL LANGUAGE PROCESSING, 16., 2017, Vancouver. **Proceedings [...]**. [S.l.: s.n.], 2017. p. 322-327.

LIN, C. *et al.* Self-training improves recurrent neural networks performance for temporal relation extraction. *In*: INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS, 9., 2018, Brussels. **Proceedings [...]**. [S.l.: s.n.], 2018. p. 165-176.

LIU, B. Sentiment analysis and opinion mining. **Synthesis Lectures on Human Language Technologies**, [s.l.], v. 5, n. 1, p. 1-167, 2012.

LIU, S. *et al.* Attention neural model for temporal relation extraction. *In*: CLINICAL NATURAL LANGUAGE PROCESSING WORKSHOP, 2., 2019, Minneapolis. **Proceedings [...]**. [S.l.: s.n.], 2019. p. 134-139.

LOTUFO, P. A. Trends in cardiovascular diseases and heart disease death rates

among adults aged 45-64: Brazil, 2000-2017. **Sao Paulo Medical Journal**, São Paulo, v. 137, n. 3, p. 213-215, 2019.

LU, X. **Computational methods for corpus annotation and analysis**. Berlin: Springer, 2014.

LUO, Z. *et al.* Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. *In*: AMIA ANNUAL SYMPOSIUM, 2011, Washington, DC. **Proceedings [...]**. [S.l.]: AMIA, 2011.

MACAVANEY, S.; COHAN, A.; GOHARIAN, N. GUIR at SemEval-2017 task 12: a framework for cross-domain clinical temporal information extraction. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 11., 2017, Vancouver. **Proceedings [...]**. [S.l.: s.n.], 2017. p. 1024-1029.

MALACHIAS, M. *et al.* 7th Brazilian guideline of arterial hypertension: presentation. **Arquivos Brasileiros de Cardiologia**, Rio de Janeiro, v. 107, n. 3, p. XV-XIX, 2016

MANTOVANI, R. G. *et al.* To tune or not to tune: recommending when to adjust SVM hyper-parameters via meta-learning. *In*: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2015, Killarney. **Proceedings [...]**. [S.l.]: IEEE, 2015. p. 1-8.

MARSLAND, S. **Machine learning**: an algorithmic perspective. [S.l.]: CRC Press, 2015.

MCDONALD, C. J.; TANG, P. C.; HRIPCSAK, G. Electronic health record systems. *In*: SHORTLIFFE, E. H.; CIMINO, J. J. (Ed.). **Biomedical informatics**. London: Springer, 2014. p. 391-423.

MENEZES FILHO, L. A.; PARDO, T. A. S. **Detecção de expressões temporais para sumarização multidocumento**. Porto Alegre: SBC, 2011.

MEYSTRE, S. M. *et al.* Extracting information from textual documents in the electronic health record: a review of recent research. **Yearbook of Medical Informatics**, [s.l.], v. 17, n. 1, p. 128-144, 2008.

MEYSTRE, S. M. *et al.* Clinical data reuse or secondary use: current status and potential future progress. **Yearbook of Medical Informatics**, [s.l.], v. 26, n. 1, p. 38-52, 2017.

MILLER, T. *et al.* Discovering temporal narrative containers in clinical text. *In*: WORKSHOP ON BIOMEDICAL NATURAL LANGUAGE PROCESSING, 2013, Sofia. **Proceedings [...]**. [S.l.: s.n.], 2013. p. 18-26.

MIRZA, P.; TONELLI, S. Catena: causal and temporal relation extraction from natural language texts. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 26., 2016, Osaka. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 64-75.

MOHARASAN, G.; HO, T.-B. Extraction of temporal information from clinical

narratives. **Journal of Healthcare Informatics Research**, [s.l.], v. 3, n. 2, p. 220-244, 2019.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. Massachusetts: The MIT Press, 2012.

MOTA, C.; SANTOS, D. (Ed.). **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas**: o Segundo HAREM. Porto: Linguateca, 2008.

MOWERY, D. L. *et al.* Distinguishing historical from current problems in clinical reports–which textual features help? *In:* WORKSHOP ON BIOMEDICAL NATURAL LANGUAGE PROCESSING, 2009, Boulder. **Proceedings […]**. [S.l.: s.n.], 2009. p. 10-18.

NADEAU, C.; BENGIO, Y. Inference for the generalization error. **Machine Learning**, [s.l.], v. 52, p. 239-281, 2003.

NEGI, K. *et al*. A novel method for drug-adverse event extraction using machine learning. **Informatics in Medicine Unlocked**, [s.l.], v. 17, p. 100190, 2019.

NIKFARJAM, A.; EMADZADEH, E.; GONZALEZ, G. Towards generating a patient's timeline: extracting temporal relationships from clinical notes. **Journal of Biomedical Informatics**, [s.l.], v. 46, n. Suppl., p. 1-21, 2013.

OLIVEIRA, L. E. S. *et al.* SemClinBr: a multi institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. **arXivLabs**, [s.l.], 2020.

PALSHIKAR, G. K. Techniques for named entity recognition: a survey. *In*: IRMA INTERNATIONAL (Ed.). **Bioinformatics**: concepts, methodologies, tools, and applications. [S.l.]: IGI Global, 2013. p. 400-426.

PARDO, T. A. S. *et al.* **Sumarização automática**: principais conceitos e sistemas para o português brasileiro. São Carlo: [s.n.], 2008.

PEARCE, P. F. *et al.* The essential SOAP note in an EHR age. **Nurse Practitioner**, [s.l.], v. 41, n. 2, p. 29-36, 2016.

POZZOBON, A. **Etimologia e abreviatura de termos médicos**: um guia para estudantes, professores, autores e editores em medicina e ciências relacionadas. Lajeado: Ed. UNIVATES, 2011.

PUSTEJOVSKY, J. *et al.* TimeML: robust specification of event and temporal expressions in text. **New Directions in Question Answering**, [s.l.], v. 3, p. 28-34, 2003.

PUSTEJOVSKY, J. *et al.* Temporal and event information in natural language text. **Language Resources and Evaluation**, [s.l.], v. 39, n. 2-3, p. 123-164, 2005.

PUSTEJOVSKY, J. *et al.* **Timebank 1.2 documentation**. [S.l.: s.n.], 2006.

PUSTEJOVSKY, J. *et al.* ISO-TimeML: an international standard for semantic annotation. *In*: LREC, 2010, Malta. **Proceedings [...]**. [S.l.: s.n.], 2010. p. 394-397.

PUSTEJOVSKY, J.; STUBBS, A. Increasing informativeness in temporal annotation. *In*: LINGUISTIC ANNOTATION WORKSHOP, 5., 2011, Portland. **Proceedings [...]**. [S.l.: s.n.], 2011. p. 152-160.

PUSTEJOVSKY, J.; STUBBS, A. **Natural language annotation for machine learning**: a guide to corpus-building for applications. [S.l.]: O'Reilly Media, 2012.

QI, P. *et al.* Stanza: a Python natural language processing toolkit for many human languages. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: SYSTEM DEMONSTRATIONS, 58., 2020. **Proceedings [...]**. [S.l.: s.n.], 2020. p. 101-108.

RAGHAVAN, P.; FOSLER-LUSSIER, E.; LAI, A. M. Learning to temporally order medical events in clinical text. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 50., 2012, Jeju Island. **Proceedings [...]**. [S.l.: s.n.], 2012a. v. 2. p. 70-74.

RAGHAVAN, P.; FOSLER-LUSSIER, E.; LAI, A. M. Temporal classification of medical events. *In*: WORKSHOP ON BIOMEDICAL NATURAL LANGUAGE PROCESSING, 2012, Montréal. **Proceedings [...]**. [S.l.: s.n.], 2012b. p. 29-37.

RAGHAVAN, P.; FOSLER-LUSSIER, E.; LAI, A. M. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. *In*: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2012, Montréal. **Proceedings [...]**. [S.l.: s.n.], 2012c. p. 731-741.

RAGHAVAN, P. *et al.* Cross-narrative temporal ordering of medical events. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 52., 2014, Baltimore. **Proceedings [...]**. [S.l.: s.n.], 2014. v. 1. p. 998-1008.

RIM, K. Mae2: portable annotation tool for general natural language use. *In*: JOINT ACL-ISO WORKSHOP ON INTEROPERABLE SEMANTIC ANNOTATION, 12., 2016, Portoroz. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 75-80.

ROBERTS, A. *et al.* Building a semantically annotated corpus of clinical texts. **Journal of Biomedical Informatics**, [s.l.], v. 42, n. 5, p. 950-966, 2009.

ROBERTS, A. *et al.* The CLEF corpus: semantic annotation of clinical text. *In*: AMIA ANNUAL SYMPOSIUM, 2007, Chicago. **Proceedings [...]**. [S.l.]: AMIA, 2007.

ROBERTS, K. *et al.* The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. **Journal of Biomedical Informatics**, [s.l.], v. 58, p. S111–S119, 2015.

ROBERTS, K.; RINK, B.; HARABAGIU, S. M. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. **Journal of the American Medical Informatics Association**, [s.l.], v. 20, n. 5, p. 867-875, 2013.

ROSENBLOOM, S. T. *et al.* Generating clinical notes for electronic health record systems. **Applied Clinical Informatics**, [s.l.], v. 1, n. 3, p. 232-243, 2010.

ROSSI, A. L. D.; CARVALHO, A. C. P. L. F. Bio-inspired optimization techniques for SVM parameter tuning. *In*: BRAZILIAN SYMPOSIUM ON NEURAL NETWORKS, 10., 2008, Salvador. **Proceedings […]**. [S.l.]: IEEE, 2008. p. 57-62.

SARATH, P. R.; MANIKANDAN, R.; NIWA, Y. Hitachi at SemEval-2017 task 12: system for temporal information extraction from clinical notes. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 11., 2017, Vancouver. **Proceedings […]**. [S.l.: s.n.], 2017. p. 1005-1009.

SAURÍ, R. *et al.* **TimeML annotation guidelines**: version 1.2. [S.l.: s.n.], 2006.

SHEIKHALISHAHI, S. *et al.* Natural language processing of clinical notes on chronic diseases: systematic review. **Journal of Medical Internet Research**, [s.l.], v. 21, n. 5, p. 1-18, 2019.

SHORTLIFFE, E. H.; BARNETT, G. O. Biomedical data: their acquisition, storage, and use. *In*: SHORTLIFFE, E. H.; CIMINO, J. J. (Ed.). **Biomedical informatics**. London: Springer, 2014. p. 39-66.

SILVA, C. R. O. E. **Metodologia e organização do projeto de pesquisa**: guia prático. Fortaleza: UFC, 2004.

SOHN, S. *et al.* Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. **Journal of the American Medical Informatics Association**, [s.l.], v. 20, n. 5, p. 836-842, 2013.

STRÖTGEN, J.; GERTZ, M. Domain-sensitive temporal tagging. **Synthesis Lectures on Human Language Technologies**, [s.l.], v. 9, n. 3, p. 1-151, 2016.

STUBBS, A. *et al.* Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2. **Journal of Biomedical Informatics**, [s.l.], v. 58, p. S67-S77, 2015.

STYLER, W. F. *et al.* Temporal annotation in the clinical domain. **Transactions of the Association for Computational Linguistics**, [s.l.], v. 2, p. 143-154, 2014a.

STYLER, W. F. *et al.* **THYME annotation guidelines**. [S.l.: s.n.], 2014b.

SUN, W.; RUMSHISKY, A.; UZUNER, O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J**ournal of the American Medical Informatics Association**, [s.l.], v. 20, n. 5, p. 806-813, 2013a.

SUN, W.; RUMSHISKY, A.; UZUNER, O. Annotating temporal information in clinical narratives. **Journal of Biomedical Informatics**, [s.l.], v. 46, n. Suppl., p. S5-S12, 2013b.

SUOMINEN, H. *et al.* Overview of the ShARe/CLEF eHealth evaluation lab 2013. *In*: INTERNATIONAL CONFERENCE OF THE CROSS-LANGUAGE EVALUATION FORUM FOR EUROPEAN LANGUAGES, 2013, Berlin. **Proceedings [...]**. Berlin: Springer, 2013. p. 212-231.

SWARTZ, M. H. **Textbook of physical diagnosis E-book**: history and examination. [S.l.]: Elsevier Health Sciences, 2020.

TALLEY, N. J.; O'CONNOR, S. **Clinical examination**: a systematic guide to physical diagnosis. [S.l.]: Elsevier Health Sciences, 2013.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. [S.l.]: Pearson Education India, 2016.

TANEV, H.; MAGNINI, B. Weakly supervised approaches for ontology population. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 11., 2006, Trento. **Proceedings [...]**. [S.l.: s.n.], 2006.

TANG, B. *et al.* A hybrid system for temporal information extraction from clinical text. **Journal of the American Medical Informatics Association**, [s.l.], v. 20, n. 5, p. 828-835, 2013.

TAO, C. *et al.* CNTRO: a semantic web ontology for temporal relation inferencing in clinical narratives. *In*: AMIA ANNUAL SYMPOSIUM, 2010, Washington, DC. **Proceedings [...]**. [S.l.]: AMIA, 2010.

THANAKI, J. **Python natural language processing**. [S.l.]: Packt, 2017.

THARWAT, A. Parameter investigation of support vector machine classifier with kernel functions. **Knowledge and Information Systems**, [s.l.], v. 61, n. 3, p. 1269-1302, 2019.

TORII, M. *et al.* Risk factor detection for heart disease by applying text analytics in electronic medical records. **Journal of Biomedical Informatics**, [s.l.], v. 58, p. S164-S170, 2015.

TOURILLE, J. *et al.* LIMSI-COT at SemEval-2016 task 12: temporal relation identification using a pipeline of classifiers. *In:* INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 10., 2016, San Diego. **Proceedings [...]**. [S.l.: s.n.], 2016. p. 1136-1142.

TOURILLE, J. *et al.* Temporal information extraction from clinical text. **Eacl**, [s.l.], v. 2, p. 739-745, 2017a.

TOURILLE, J. *et al.* LIMSI-COT at SemEval-2017 task 12: neural architecture for

temporal information extraction from clinical narratives. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 11., 2017, Vancouver. **Proceedings [...]**. [S.l.: s.n.], 2017b.

TOURILLE, J. *et al.* Neural architecture for temporal relation extraction: A bi-lstm approach for detecting narrative containers. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 55., 2017, Vancouver. **Proceedings [...]**. [S.l.: s.n.], 2017c. v. 2. p. 224-230.

UZZAMAN, N. *et al.* Semeval-2013 task 1: Tempeval-3: evaluating time expressions, events, and temporal relations. *In*: JOINT CONFERENCE ON LEXICAL AND COMPUTATIONAL SEMANTICS, 2., 2013, Atlanta. **Proceedings [...]**. [S.l.: s.n.], 2013. v. 2.

VELUPILLAI, S. *et al.* Recent advances in clinical natural language processing in support of semantic analysis. **Yearbook of Medical Informatics**, [s.l.], v. 10, n. 1, p. 183-193, 2015a.

VELUPILLAI, S. *et al.* Blulab: temporal information extraction for the 2015 clinical temporal challenge. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 9., 2015, Denver. **Proceedings [...]**. [S.l.: s.n.], 2015b. p. 815-819.

VERHAGEN, M. *et al.* SemEval-2007 task 15: TempEval temporal relation identification. **Computational Linguistics**, [s.l.], n. June, p. 75-80, 2007.

VERHAGEN, M. Temporal closure in an annotation environment. **Language Resources and Evaluation**, [s.l.], v. 39, n. 2-3, p. 211-241, 2005.

VERHAGEN, M. *et al.* SemEval-2010 task 13: TempEval-2. *In*: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 5., 2010, Uppsala. **Proceedings [...]**. [S.l.: s.n.], 2010. p. 57-62.

VIANI, N. *et al.* Supervised methods to extract clinical events from cardiology reports in Italian. **Journal of Biomedical Informatics**, [s.l.], v. 95, n. May, p. 103219, 2019.

WANG, Y. *et al.* Clinical information extraction applications: a literature review. **Journal of Biomedical Informatics**, [s.l.], v. 77, n. November 2017, p. 34-49, 2018.

WEED, L. L. Medical records, patient care, and medical education. **Irish Journal of Medical Science**, [s.l.], v. 39, n. 6, p. 271-282, 1964.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining**. [S.l.: s.n.], 2011.

WORLD HEALTH ORGANIZATION (WHO) *et al.* **Global action plan for the prevention and control of noncommunicable diseases 2013-2020**. Geneva: WHO, 2013.

WU, S. *et al.* Deep learning in clinical natural language processing: a methodical review. **Journal of the American Medical Informatics Association**, [s.l.], v. 27, n. 3, p. 457-470, 1 mar. 2020.

XU, Y. *et al.* An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. **Journal of the American Medical Informatics Association**, [s.l.], v. 20, n. 5, p. 849-858, 2013.

ZHU, L.; YANG, H.; YAN, Z. Mining medical related temporal information from patients' self-description. **International Journal of Crowd Science**, [s.l.], v. 1, n. 2, 2017.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis Lectures on Artificial Intelligence and Machine Learning**, [s.l.], v. 3, n. 1, p. 1-130, 2009.

APÊNDICE A – REVISÃO SISTEMÁTICA

Temporal Relation Extraction in Clinical Texts: A Systematic Review

Yohan Bonescki Gumiel †

Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil, yohan.bonescki@pucpr.edu.br

Lucas Emanuel Silva e Oliveira

Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil, lucas.oliveira@pucpr.br

Vincent Claveau

IRISA – CRNS, Université de Rennes 1, Rennes, Rennes, France, vincent.claveau@irisa.fr

Natalia Grabar

CRNS, Univ. Lille, Lille, 59000 Lille, France, natalia.grabar@univ-lille3.fr

Emerson Cabrera Paraiso

Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil, paraiso@ppgia.pucpr.br

Claudia Moro

Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil, c.moro@pucpr.br

Deborah Ribeiro Carvalho

Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil, ribeiro.carvalho@pucpr.br

† Corresponding Author.

ABSTRACT

The unstructured data contained in electronic health records, represented by clinical texts, are a massive source of information for healthcare because they describe the patient journey, including clinical findings, procedures, and information about the continuity of care. The publication of several studies that focused on the extraction of temporal relations from electronic health records during the last decade and the realization of multiple shared tasks is a tribute to the importance of this research theme. Therefore, we propose a review of temporal relation extraction in clinical texts. We analyzed 101 articles and concluded that temporal relations between events and document creation time were addressed well by the present researchers with results on par with those determined by human annotators in some studies. However, there is room for improvement in temporal relations regarding events and temporal expressions. In the early publications, rule-based and simpler approaches were used; later, machine learning methods with several heuristics for candidate pair selection, training set expansion, and specialized classifiers were used. In later publications, the best approaches depended heavily on feature engineering that adapted frameworks and/or methods from previous publications or that developed frameworks based on deep learning.

CCS CONCEPTS

• Computing methodologies ~ Artificial intelligence ~ Natural language processing ~ Information extraction • Applied computing ~ Life and medical sciences ~ Health informatics

KEYWORDS

Temporal relation extraction, Natural Language Processing, Clinical Data

INTRODUCTION

Electronic health records (EHRs) are repositories of patient information containing structured and unstructured data. Some examples of structured data are medications, laboratory exams, and radiology exams [51], while the clinical narratives (i.e., free text data) represent the non-structured data. In fact, much of the data that could be used for healthcare quality improvement (e.g., creation of a patient timeline) and clinical research (e.g., drug-drug interactions) are stored in free-text or other formats that limit reusability in some manner [13]. A study of six hospitals by Capurro *et al.* [14] showed that, on average, 75% of all data elements were not available as structured data or computable database fields.

The volume of the information available from EHRs makes the process of reviewing such data by humans a long and laborious task [35]. It is difficult to imagine that any person or team could work with these unstructured data and transform them by manual labor to a structured format [22]. Thus, computational tools are needed to automatically or semi-automatically extract information from EHRs data [113]. Natural language processing (PLN), when applied to clinical narratives, has the potential to benefit both healthcare research and biomedical research [61].

The use of PLN in temporal relation extraction can positively affect the healthcare field, as it can extract from clinical narratives rich and contextual information that is not available elsewhere, including rich temporal and background information about current status/conditions, and even information about a patient's past (e.g., a treatment that was done a long time ago) [96,121]. In addition to containing information about the past and present, the records also provide information about the future (e.g., foreseen interventions and treatments). Extracting temporal relations can help to identify interactions between drugs that lead to adverse event reactions, as in Iyer *et al.* [49], and can help to detect adverse events related to the usage of a specific drug, as in Liu *et al.* [83]. These are research activities in which temporality is applied because the effects are verified over time. Temporal relations can also provide information to identify a patient's current diseases, as Capurro *et al.* [13] identified patients with acute kidney injury, or to predict certain disease occurrence over the hospital stay, as Bejan *et al.* [5] predicted pneumonia status during intensive care unit (ICU) stay. Temporal relation extraction enables timeline creation, which can improve healthcare as it provides events (e.g., medical problems, treatments) and time references in an organized format, summarizing text information in a graphical format.

Yet, there are some difficulties in working with narrative texts, such as nonstandard expressions, abbreviations, assumptions, and the extensive use of domain knowledge [74]. The texts may present flexible formatting (e.g., missing punctuation), atypical grammar (e.g., omission of expected words such as verbs or objects), tachygraphy (e.g., abbreviations, acronyms, abbreviated phrases with local expressions or dialects), and orthographic errors [61,88]. Moreover, most of these issues are related to the short time a clinician has to write down the details of the appointment with the patient. For Wang *et al.* [141], working with low-quality texts (a characteristic commonly attributed to clinical narratives) leads to problematic extraction of syntactic, grammatical, and also semantic features, which affects the performance of a system that relies on these attributes.

Some aspects of time can be captured either by using structured data, which contains information with its related time mention, or by the EHRs timestamps, which can be directly extracted from the document header. Yet, a large amount of information about disease progression can only be found in the unstructured part of the EHRs (clinical texts), and temporal relation extraction is needed to correlate the temporal expressions with events [36]. In the clinical domain, events are situations that occur in patient records that are clinically relevant (e.g., treatments, problems, tests); temporal expressions are expressions that allude to temporal mentions (e.g., duration mentions or date mentions); and temporal relations are the relations between events and temporal expressions [96]. A temporal expression can be either a time mention in the free text or the document creation time (DCT). Events are typically annotated with attributes (e.g., polarity and type) and temporal expression typically involves the expression normalization to the ISO8601[5] (ISO for date and time format).

Clinical narratives are examples of unstructured data, and therefore cannot be accessed by software applications the same way as structured data, as computers cannot do temporal tasks (e.g., timeline creation or answer temporal questions) without a reasoning system based on manual annotations [5,26]. Thus, there is a need to transform this unstructured information to structured information, allowing the connection of patient conditions and events mentions to a clinical timeline [22,53]. Thus, temporal relation extraction over clinical narratives is a prime target for the development of automated PLN methods based on patient data [123]. Indeed, automatically

[5] https://www.iso.org/iso-8601-date-and-time-format.html

extracting temporal information and discovering its temporal relations can promote an understanding of several clinical aspects, such as disease progression, the causes of disease, the monitoring of patient operations, the effectiveness and side-effects of drugs, and the patient's timeline [74,145].

In addition to the free-text issues, the temporal relation extraction task itself presents some challenges. Allen [2], when working with temporal relations in English, noted that temporal mentions could be implicitly mentioned in the text and vague. In general, text annotation is already a complicated process, but the annotation of temporal relations is much more complicated. For instance, temporal relation extraction in the clinical area has lower inter-agreement than other clinical annotation tasks, such as event and temporal expression annotation tasks [96]. Furthermore, the domain has a high impact on temporal relation extraction. The task of extracting temporal information from narratives may be more complex than general domain texts (e.g., newswires) because of the lack of formalism and writing quality [126]. Lastly, if the process of determining temporal relations is difficult for humans with guidelines, it is even more challenging for an automated system.

For extracting temporal information, especially implicit information, domain-specific knowledge and PLN frameworks developed for the clinical domain are necessary [96]. The need for specific knowledge and tools may be a limiting factor, especially in the clinical domain, owing to the lack of resources and available data. Frameworks developed for the general domain need adaptations to fulfill the needs of the clinical area. Examples are the adaptation of HeidelTime [118], a framework for temporal expression extraction and normalization to the clinical domain created by Hamon and Grabar [44] for English and French and by De Azevedo *et al*. [3] for Brazilian Portuguese. Moreover, sophisticated PLN techniques are typically provided by frameworks that are language-dependent, hindering the use of languages other than English. The amount of available data is also a limiting factor: deep learning approaches rely on a large amount of data to address generality, and access to clinical domain data is difficult owing to data privacy.

Hopefully, several shared tasks have been organized to provide data that the research community can use to develop their temporal extraction techniques, and then compare their results with. There are several shared tasks that focus on temporal relations in the clinical domain, and several articles that aim to deal with temporality in different clinical areas have been published. In addition to all the work done in temporal information extraction for the general domain, the interest in temporal relation extraction from clinical narratives began to grow with the 2012 Informatics for Integrating Biology and Bedside (i2b2) [123], and then with Clinical TempEval in semEval2015 [7], semEval2016 [8], and semEval2017 [9]. With the intent to discuss the approaches used (for shared-task related, or general articles), highlight the main aspects, and point out the best methods in studies, we performed a systematic review that followed the PRISMA statement [94].

The objective of this article is to present a review of the state of the art in temporal relation extraction in clinical texts. The question this review aims to answer is: "What is the effectiveness of machine learning and rule-based techniques in identifying temporal relations in clinical texts?" Our secondary objective is to provide insights on the domain evolution over time, leveraging temporal relation extraction objectives and developed frameworks.

This review was structured as follows. Section 2 provides an overview of temporal relation extraction, including representations and examples for both clinical and general domains. In Section 3, the methodological steps are detailed, and we provide global quantitative results about the selected publications. Previous temporal relations shared tasks, in both general and clinical domains, that focused on temporal relation extraction divided the temporal relations into relations between mentions (either events or temporal expressions) over the text, which we will refer to as TLINKs and relations between events and the DCT, which we will refer to as DCT relations. We elaborate on the DCT relations-related articles in Section 4 and TLINKs-related articles in Section 5. In Section 6, we present shared-task datasets and approaches for temporal relation extraction in the general domain, describe the relevant publications, and compare the approaches used. In Section 7, we conclude this article. From this review, the reader can expect an overview of temporal relations and an investigation of the best techniques and frameworks for temporal relation extraction.

TEMPORAL RELATION EXTRACTION

Temporal relation extraction involves linking mentions (either events or temporal expressions) and defining a relation type that represents how they are temporally related to each other. A definition of event and temporal expression can be influenced by the text domain, as we have different needs for clinical and general domain texts. For the general domain, there are representations such as ISO-TimeML [103], which defines events as mentions involving verbs, nominalizations, adjectives, predicative clauses, and prepositional phrases, and temporal expressions are defined as mentions of dates, times (specific time during a day), durations, and sets [102,112].

For the clinical domain, there are representations such as THYME-TimeML [121], where events are any mention that is considered to be relevant when constructing a clinical timeline, and temporal expressions are similar to the TimeML representation but with specific domain cases (e.g., medication frequency). The differences between representations (considering both events and temporal expressions) in the general and clinical domains are explained in sub-section 2.2.
The temporal relation type is based on the markup language definition depending on the research objective. In sub-section 2.1, we discuss the most-used temporal relation representations for both general and clinical domains.

TEMPORAL RELATIONS REPRESENTATIONS

A landmark for temporal relation extraction was the interval-based algebra proposed by Allen in 1983. Several studies adopted Allen's representation [2], which quickly became a pattern for temporal modeling [126]. Allen's representation held that given two points in time or intervals of time, any relation between the two could be represented by seven relations: "BEFORE," "MEET," "OVERLAP," "DURING," "STARTS," "FINISHES," and "EQUAL" [2]. Considering the inverse relations ("EQUAL" does not have an inverse relation) there are 13 possible relations. Allen's relations are shown in Figure 1 (column Allen's Algebra).
Among the temporal information representations, TimeML is a temporal markup language that was developed exclusively to annotate events, temporal expressions, and relations over text [103]. In 2010, ISO-TimeML was released, which is a revised and interoperable version of TimeML, by Pustejovsky *et al.* [103]. TimeML was developed by researchers of the PLN community aiming to move temporal information in free-text format into a structured data format [22]. TimeML represents relations between events and temporal expressions using a format based on Allen's representation.
In terms of relations, the main differences between Allen's representation and TimeML is that TimeML does not address "OVERLAP" relations, and that the relation "EQUAL" in Allen's algebra is represented over four relations in TimeML: "IDENTITY," "SIMULTANEOUS," "HOLD," and "HELD BY" [136]. The TimeML relations are displayed in Figure 1 (column TimeML).

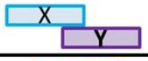
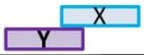
REPRESENTATION	Allen`s representation	TimeML	I2b2 TLINK	I2b2 DCT relation	Clinical TempEval TLINK	Clinical TempEval DCT relation
1. 	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y
2. 	X MEETS Y	X I_BEFORE Y	X BEF_OVERLAP Y	X BEF_OVERLAP Y		X BEF_OVERLAP Y
3. 	X OVERLAPS Y		X OVERLAPS Y	X OVERLAPS Y	X OVERLAPS Y	X OVERLAPS Y
4. 	X DURING Y	X IS_INCLUDED Y	X DURING Y	X DURING Y		
5. 	X STARTS Y	X BEGINS Y			X BEGINS_ON Y	
6. 	X FINISHES Y	X ENDS Y				
7. 	X FINISHED_BY Y	X ENDS_BY Y	X ENDS_BY Y	X ENDS_BY Y	X ENDS_ON Y	
8. 	X AFTER Y	X AFTER Y	X AFTER Y	X AFTER Y		X AFTER Y
9. 	X MEET_BY Y	X I_AFTER Y				
10. 	X OVERLAPPED_BY Y					
11. 	X CONTAINS Y	X INCLUDES Y			X CONTAINS Y	
12. 	X STARTED_BY Y	X BEGINS_BY Y	X BEGINS_BY Y	X BEGINS_BY Y		
13. 	X EQUALS Y	X IDENTIFY Y X SIMULT. Y X HOLD Y X HELD_BY Y	X SIMULT. Y	X SIMULT. Y		

Figure 1: Temporal relation types present in Allen's representation; TimeML annotation scheme, 2012 i2b2 dataset and Clinical TempEval datasets, relations in gray were annotated in the dataset but not used during the shared-tasks.

Owing to the clinical domain annotation particularities and to better represent the clinical information, some adaptations were made to the annotation process and to the TimeML. Here, we will detail some of these adaptations regarding the 2012 i2b2 and Clinical TempEval datasets. All Clinical TempEval shared-tasks are based on the Temporal History of Your Medical Events (THYME) corpus. These datasets are the primary source of research regarding temporal relation extraction over clinical texts.

For the 2012 i2b2 dataset, in addition to an adaptation of the TimeML annotation guidelines [100], preliminary guidelines from THYME were used [123]. In the 2012 i2b2 dataset, the annotated relations were: "BEFORE," "AFTER," "OVERLAPS/SIMULTANEOUS," "BEFORE_OVERLAP," "DURING," "BEGUN_BY," and "ENDED_BY." For the i2b2 2012 shared-task, only three types of relationships were used ("BEFORE," "OVERLAPS," and "AFTER"). There was a simplification process owing to both lower agreement and lower annotation number for certain relations, with the merging of certain types of relations (e.g., "SIMULTANEOUS," "OVERLAPS," and "DURING" merged as "OVERLAP") [123]. The relation types are detailed in Figure 1 (columns i2b2 TLINK and i2b2 DCT). Relations represented in black were used in the challenge (e.g., BEFORE) and relations represented in gray were merged (e.g., "DURING"). This difference in the annotation number for certain relation types is shown in Sun, Rumshisky and Uzuner [124], exemplified by 2.7% of the annotations (before temporal closure) being of the "ENDED_BY" type, while 66.6% (before temporal closure) were of "OVERLAPS/SIMULTANEOUS."

There are two columns in Figure 1 for the 2012 i2b2 temporal representation because we performed two different temporal relation extractions over the dataset. In 2012 i2b2, there were relations (1) between mentions (either events or temporal expressions) over the text, which we will refer to as TLINKs (Figure 1 – column i2b2 TINK) and (2) between events and section times (SECTIME), which we will refer to as DCT relations (Figure 1 – column i2b2 DCT). Section times are types of temporal expressions regarding either date of admission or date of discharge, depending on where the event is located within the text. Section times can be considered as a type

of DCT, and so are referred to as DCT relation. The 2012 i2b2 challenge corpus was composed of 310 discharge summary notes averaging 86.6 events, 12.4 temporal expressions, and 176 TLINKs per note.
The ISO-TimeML was expanded to clinical domain into THYME guidelines, named THYME-TimeML, to annotate the THYME corpus. In THYME corpus annotation, five different relations were used: "BEFORE," "OVERLAPS," "BEGINS_ON," "ENDS_ON," and "CONTAINS." Unlike the i2b2 corpus, the THYME corpus was annotated with the narrative container concept introduced by Pustejovsky and Stubbs [104]. They [104] emphasized the importance of an annotation schema for temporal relations that resulted in maximally annotated temporal relation information while not relying on models that were too difficult to apply.
The choice of using narrative containers, according to Styler *et al.* [121], comes from the difficulty in capturing every possible relation and the rise in disagreement that happens when annotators try to do so. By using this choice of annotation, whenever is possible, time expressions and events are connected to a narrative container (event or temporal expression anchor) that defines their temporal interval. Several events or temporal expressions can be connected to the same anchor, which contains them (represented in "CONTAINS" row in Figure 1). Events and temporal expressions that are in the same narrative container can be related, as a single element, with other containers [55]. The biggest advantage is the reduction in the amount of annotation needed [55]. By the usage of the narrative container approach, and also by eliminating confusing inverse relations (e.g., "DURING' and "AFTER"), relations annotation agreement improved over the i2b2 annotations [121].
In the Clinical TempEval 2015, 2016, and 2017 shared-tasks, which were based on the THYME corpus, there were two types of temporal relations. There are two columns in Figure 1 for the Clinical TempEval temporal representation because we performed two different temporal relation extractions over the dataset. In Clinical TempEval, the relations were: (1) between mentions (either events or temporal expressions) over the text, which we will refer to as TLINKs (Figure 1 – column Clinical TempEval TINK) and (2) between events and DCT (considered as an attribute of the event), which we will refer to as DCT relations (Figure 1 – column Clinical TempEval DCT). Every event has a DCT relation with one of the following types: "BEFORE," "AFTER," "OVERLAP," "BEFORE_OVERLAP," or "AFTER." The "BEFORE_OVERLAP" is included only in DCT relations, and represents that the event occurred in the past and still occurs in the DCT, which can be the case with a chronic disease, for example, that exists before the clinical document creation and continues to exist during its writing. The TLINKs in the Clinical TempEval shared-tasks (Figure 1 – column Clinical TempEval TLINK) were simplified to consider only the "CONTAINS" type. Actually, "CONTAINS" relations were the most frequently annotated relation (66% of the annotated TLINKs) over the THYME corpus, followed by "OVERLAPS" with 14.6% [121].
Regarding the Clinical TempEval challenge corpus, the 2015 Clinical TempEval corpus was composed of 440 documents averaging 136.05 events, 13.43 temporal expressions, 37.43 TLINKs, and 136.05 DCT relations (as each event has a DCT relation) per document. The 2016 edition had 591 documents averaging 133.42 events, 13.30 temporal expressions, 39.33 TLINKs, and 133.42 DCT relations per document. The 2017 edition aimed at cross-domain extraction with different domains for the training and testing data. In phase 1, unsupervised domain adaptation (UDA), the evaluation was performed on brain cancer notes, given colon cancer notes. In phase 2, supervised domain adaptation (SDA), a small portion of the brain cancer annotated notes were added as input [111]. According to Bethard *et al.* [9], this task is a much more challenging task than the previous ones (2015 and 2016 editions). The 2017 Clinical TempEval was composed of 769 documents averaging 120.83 events, 12.70 temporal expressions, 33.28 TLINKs, and 120.83 DCT relations per document.
In addition to the 2012 i2b2 shared-task and THYME related shared-tasks, the 2014 i2b2/UTHealth challenge was related to heart disease mentions, focusing on the discovery of potential risk factors [127]. There was no dual evaluation for DCT relation extraction, as temporality was not the focus. Thus, we will briefly present some information about the challenge, and further information can be obtained from Stubbs *et al.* [120]. Shortly, risk factors could be connected to the DCT with one or more relations ("BEFORE," "AFTER," and "DURING"), turning it into a multi-label classification task [20].

TEMMPORAL ANNOTATION EXAMPLES

To exemplify the differences regarding general domain and clinical domain annotation schemes, we present two examples to discuss the main differences between them. Figure 2 (A) represents a relation over the TimeML

annotation of a simple sentence “John left 2 days ago,” which is composed of an event, a temporal expression, and a temporal relation. This example was based on one of the annotation examples presented by Saurí *et al.* [112]. In TimeML, temporal expressions are represented by the TIMEX3 tag, which captures mentions of dates (e.g., December 10), times (e.g., at four o’clock), duration (e.g., 2 months), and sets (e.g., everyday) [102]. Every temporal expression is normalized to a “value” according to the ISO 8601 standard. The temporal expression “2 days ago” is related to the DCT(“t0”), which is shown by the “AnchorTimeID” value. The term “ago” makes sense only if we consider the date on which the text is written (DCT). As the DCT normalized value is “2002-07-10,” the temporal expression normalized value is “2002-07-08.”

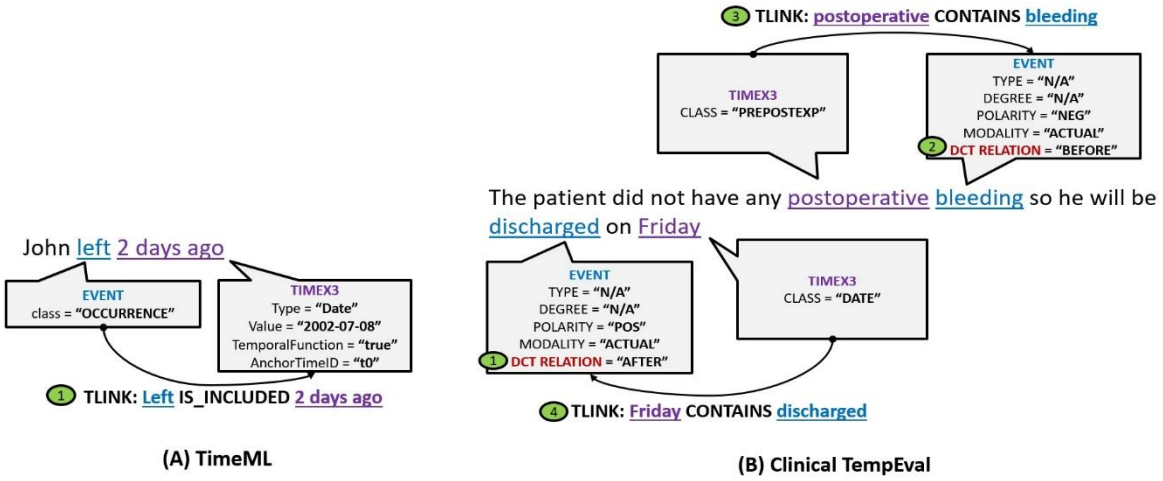


Figure 2: TimeML (A) and Clinical TempEval (B) temporal relation example.

Briefly, an event can be expressed by verbs, nominalizations, adjectives, predicative clauses, and prepositional phrases [112]. The verb “left” is an event of the class “OCCURRENCE,” which includes all events that do not fit any of the other categories. In a conventional TimeML markup, every event would have to be connected to at least one “MAKEINSTANCE” tag. This tag creates the “realization” of the event and is developed for addressing cases, whereas we can have multiple instances related to the same event [102]. The MAKEINSTANCE tag would be connected to the temporal expression (TIMEX3) in this case. To simplify our explanation, we skip the MAKEINSTANCE tag and directly connect the event with the TIMEX3.

As “2 days ago” refers to an entire day, and “left” is an event that occurred sometime during that period of time equivalent to a day, we can infer that the event “left” is totally contained by (‘IS_INCLUDED” relation) in “full day,” which happened “2 days ago.” In addition to these tags, we could still mark a SIGNAL tag that involves words (e.g., before, after, while, on) that explicitly denote relations [102]. In this sentence, we do not have any SIGNAL tags, but in a sentence such as “John left 2 days before the attack,” the text span of “before” would be marked as a SIGNAL. Further details about MAKEINSTANCE, SIGNAL tags and complete examples can be found in Saurí *et al.* [112] and Pustejovsky *et al.* [102].

With the previous simple example, we can see the tags and markups that could be made over a simple general-domain sentence with TimeML. To illustrate the differences between domains, we will adapt an example regarding Clinical TempEval shared-tasks based on THYME-TimeML, Figure 2 (B), which is shown in Bethard *et al.* [8].

THYME-TimeML does not involve tags such as “SIGNALS” and “MAKEINSTANCE,” as these tags were simplified. In the THYME corpus, an event is any mention that is considered to be relevant when constructing a clinical timeline [121]. This involves several types of mentions such as medical problems (e.g., headache), treatments (e.g., medications and procedures), and exams (e.g., diagnostic imaging or physical and visual examination). Verbs such as “denied,” “discharge,” “continued,” and “showed” are also marked, but the annotation is less focused on verbs than in TimeML. In this example, both “bleeding” and “discharged” are marked as events.

An event has the attributes “TYPE,” “DEGREE,” “POLARITY,” and DCT relation. For both events (“bleeding” and “discharged”), the “TYPE” is marked as “N/A” because it is the default value. According to THYME annotation guidelines [151], in practice, “DEGREE” is used to infer that something is substantially or slightly true (e.g., “slight pain”). For both events, the “DEGREE” attribute is marked as “N/A” (default value). The

"POLARITY" attribute can be either marked as "pos" (an event that happened or is going to happen) or "neg" (the opposite). In this example, "bleeding," the "POLARITY" attribute was marked as "neg" because the patient did not have any bleeding, while the event "discharged" was marked as "pos" for the "POLARITY" attribute because the patient would be discharged on Friday. For both events, the "MODALITY" attribute is marked as "actual," as they represent events that have already happened or that are going to happen in the future. If we have a "discharge" mention in a sentence, such as "possible discharge on Friday," its "MODALITY" attribute would be changed to "hypothetical." The DCT relation is marked as "before" for "bleeding" because it is inferred by the textual construction of the phrase that the bleeding (in this case absence of bleeding) occurred before the DCT. For the event "discharged," the "DCT relation" value is marked as "after" because of the usage of the simple future in passive voice phrase construction.

In Figure 2 (B), we have two temporal expressions (represented by TIMEX3 tags), which are "postoperative" and "Friday." It is notable that unlike in TimeML, there is no value normalization. The mention "Friday" is typical of TIMEX3 with its "TYPE"/"CLASS" being marked as "date." Two significant changes are the coverage of "set" class (as it includes the frequencies of medications), and the addition of a "prepostexp" class (covering temporal terms as preoperative, postoperative, and intraoperative). The "set" class involves expressions such as "b.i.d" (twice a day) and q2h (every two hours), which are associated with medication prescription, as in the sentence "Metformin 850 mg b.i.d." As explained above, the mention "postoperative" is considered a TIMEX3 with a "class" of "prepostexp."

Regarding the TLINKs, in this example, we have two relations marked as "CONTAINS." As our example is related to the Clinical TempEval shared-tasks, the "CONTAINS" relation is the only relation type between mentions (events or temporal expressions). As the postoperative period did not contain any bleeding, it means that bleeding (with negative "POLARITY") as totally contained within the postoperative period. Similarly, we can infer that as the patient will be "discharged" on "Friday," "discharged" is totally contained ('CONTAINS" relation) in "full day," which is "Friday."

The benefit of extracting such temporality is shown in Figure 3. Using a "simple" approach of merely connecting every event to its DCT, we cannot infer any order. Thus, in this scenario, both "bleeding" and "discharged" occurred at the same time, which is not true. By adding more information to the annotation of the DCT relations, we can differentiate the "bleeding," which is a past event that has already happened, from the "discharged," which is a future event that has not happened yet. The issue of using only the DCT relation is that you become tied to relation types; in Clinical TempEval there are only four buckets ("BEFORE," "AFTER," "BEFORE_OVERLAP," and "OVERLAP") into which events can be divided. Moreover, DCT relations are too generic for certain temporal relation extraction studies, as categories such as "BEFORE DCT" are too extensive because they do not refer to a certain point or closed period in time, but refer instead to a wide period of time.

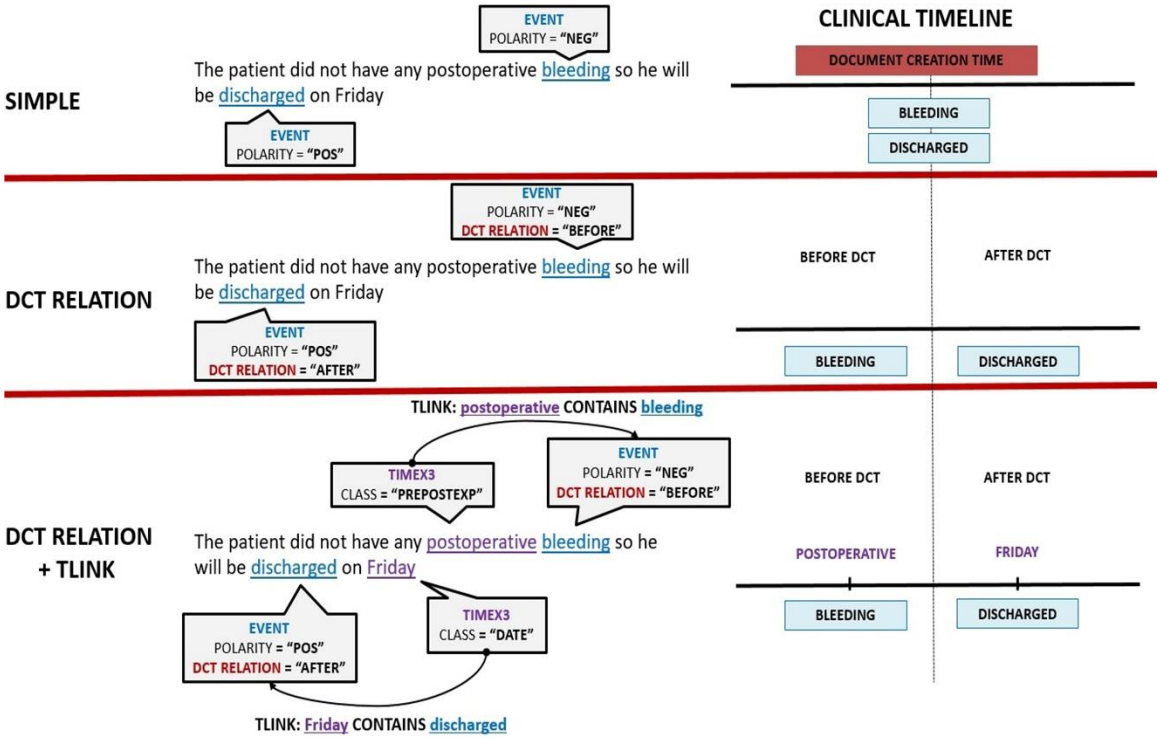


Figure 3: Clinical timeline creation for "simple," "DCT relation," and "DCT relation + TLINK" approaches.

Adding TLINKs, as shown in Figure 3 ("DCT RELATION + TLINK" row), to anchor events to specific periods of time represented by temporal expressions improves the timeline representation. For example, "discharged," in addition to belonging to the wide period of time "AFTER DCT," which could be either one day after or ten years after, also belongs to a period of time within a single day ("Friday"). As not every event or temporal expression has associated TLINKs, the usage of DCT relations enables some sort of event ordering, but TLINKs provide a more detailed representation.
In this section, we present some ways to express temporal relations, while providing examples of annotation. In the next section, we provide details of our systematic review, explaining the methodological process and criteria that resulted in the selected articles.

METHODOLOGY

The databases selected for this review were PubMed Central (MedLine), Science Direct, and ACL Anthology. The descriptors used while searching over these databases were: ("temporal relation" OR "temporal relations" OR "temporal extraction" OR "temporal information" OR "temporal relationship" OR "temporal relationships" OR "timeline") AND ("clinical text" OR "clinical texts" OR "clinical narratives" OR "clinical narrative" OR "clinical reports" OR "clinical report").

SELECTION OF PAPERS TO REVIEW

In Figure 4, we summarized our methodological steps. We identified 2,092 articles over the databases, and an additional 16 articles were identified by reading the selected articles and finding a direct relationship between these additional articles and the review. After analyzing the title and abstract criteria and the full-text criteria, we selected 101 articles. These 101 articles were evaluated according to their approaches to dealing with temporality and their quantitative results.

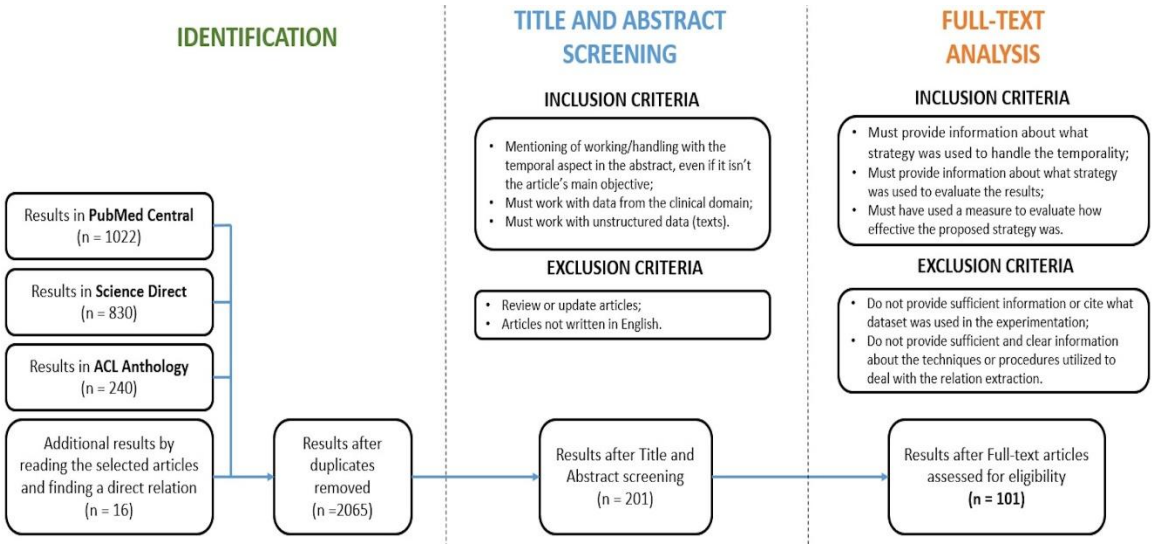


Figure 4: Methodological steps used for this systematic review.

GLOBAL QUANTITATIVE RESULTS

Sun and colleagues [125] published a survey showing some preliminary studies from 2006 to 2012. An increase in publications happened in 2013, after the 2012 i2b2 challenge (represented in Figure 5). Most of the selected publications were from datasets in the English language: 92 out of 101 reviewed articles. Publications in languages

other than English included: Xu *et al.* [144], Zhu, Yang, and Yan [150], Su *et al.* [122] and Liu *et al.* [84] in Chinese; Seol *et al.* [114] and Lee and Choi [66] in Korean; Afzal *et al.* [1] in Dutch; Tourille *et al.* [130] in French; Viani *et al.* [139] in Italian; and Henriksson *et al.* [46] in Swedish. One could conclude that there is room for research in languages other than English.

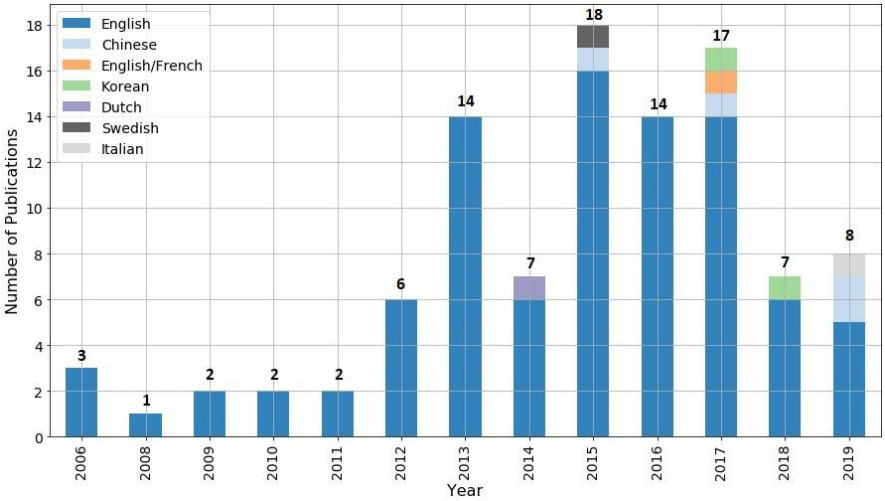


Figure 5: Number of publications by year separated by language in chronological.

PUBLICATION SUMMARIZATION

As mentioned in Introduction our analysis was divided into two temporal relation extraction tasks, DCT relation (Section 4) and TLINK (Section 5) extraction. For each task we sub-divided into rule-based approaches (sub-sections 4.1 and 5.1), machine learning-based and hybrid approaches (sub-sections 4.2 and 5.2) and deep learning-based approaches (sub-sections 4.3 and 5.3). These divisions are summarized over Figure 6. For both machine learning-based and hybrid approaches and deep learning-based approaches we also summarized the used approaches.

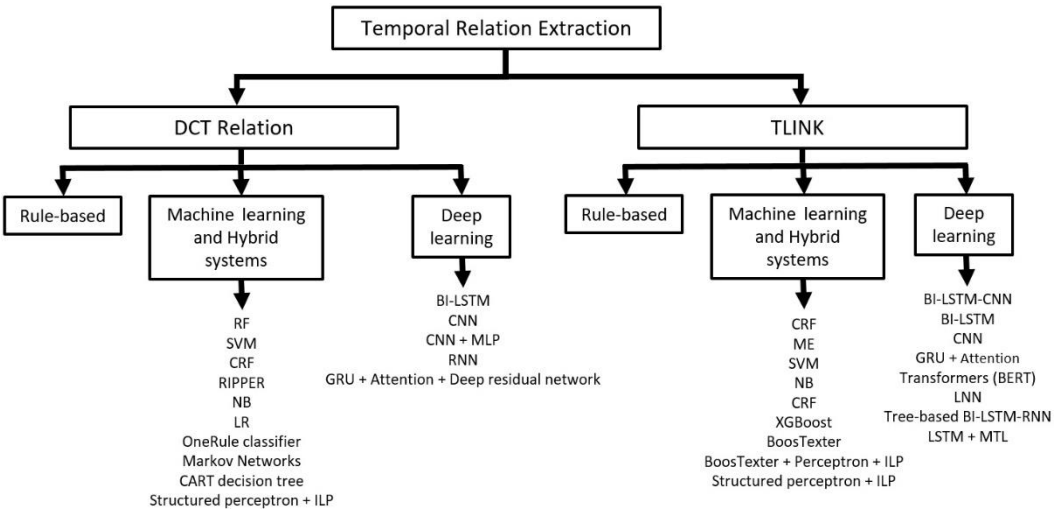


Figure 6: Publication summarization according the task and the used approach.

DCT RELATIONS

DCT relations are considered to be any relationship between an EVENT and a DCT. The results are sorted according to the strategy used to deal with the relationship with three categories: rule-based/pattern matching systems (see sub-section 4.1), and machine learning and hybrid systems (see sub-section 4.2) and deep learning (see sub-section 4.3). As we provide an in-depth overview and evaluation of all selected articles, we compile a summary (see sub-section 4.4) with highlights and conclusions.

DCT RELATIONS RULE-BASED SYSTEMS

The systems that extract DCT relations exclusively with a rule-based approach are listed in Table 1. The table contains information about the article, the main objective of the article, the strategy used to extract temporality, and the article result, which is related with the "Sep DCT" (separated DCT relation evaluation). If the article had a separate evaluation, the result regards the temporal extraction, if not, the results regard the system main objective.

Table 1: Articles related to DCT relation that used full rule-based systems.

Authors	Objective	Strategy	Results	Sep DCT
Iyer *et al.* [49]	Identify DDI signals	Linked events to DCT	AUROC 81.5%	No
Iyer *et al.* [48]	Identify DDI signals	Linked events to DCT	AUROC > 80% (two datasets)	No
Liu *et al.* [83]	Differentiate DAE from indications	Linked events to DCT	AUROC 85%	No
Lee *et al.* [62]	Clopidogrel-induced bleedings identification	Linked events to DCT	Fm 0.9248	No
Yetisgen-yildiz *et al.* [147]	Predict pneumonia status during ICU stay	Linked events to DCT	Fm 0.49	No
Bejan *et al.* [5]	Predict pneumonia status during ICU stay	Linked events to DCT	Fm 0.837	No
Wu *et al.* [143]	Asthma status identification with temporal aggregation	Linked events to DCT	Fm 0.8571, 0.7463	No
Harkema *et al.* [45]	ConText devolvement	Regex	Fm 0.73 Historical, 0.86 Hypothetical	Yes
Afzal *et al.* [1]	ConText adaptation to Dutch	Regex	Fm 0.94-0.97 Recent, 0.26- 0.54 Historical, 0.13-0.44 Hypothetical	Yes
Gaizauskas *et al.* [37]	Preliminary CLEF study	Rules	Fm 0.6225	Yes
Khalifa and Meystre [54]	UTHealth 2014 RFE	df	Fm 0.875	No
Yang and Garibaldi [146]	UTHealth 2014 RFE	df + specific rules	Fm 0.915	No
Urbain [132]	UTHealth 2014 RFE	df	Fm 0.890	No
Karystianis *et al.* [53]	UTHealth 2014 RFE	df + specific rules + ann refinement	Fm 0.8776	No
Shivade *et al.* [115]	UTHealth 2014 RFE	Rules	Fm 0.907	No
Chang *et al.* [18]	UTHealth 2014 RFE	Context-aware refinement approach	Fm 0.897	No
Chang *et al.* [19]	UTHealth 2014 RFE	Rules	Fm 0.5628	No
Wang *et al.* [141]	i2b2 2012 TRE	Rules	Mach ratio 0.69 (NTC)	Yes

Legend: AUROC: area under the receiver operator curve, DAE: drug-adverse events, DCT: document creation time, DDI: drug–drug interaction, ICU: intensive care unit, Fm: F-measure, RFE: risk factor extraction, TRE: temporal relation extraction, regex: regular expression, df: default value, ann: annotation, NTC: not comparable.

In some cases, the temporality involves merely connecting the events and DCTs. This typically happens whenever temporality is not the main focus, but merely a step in the information extraction methodology, or whenever the process of connecting the event to its DCT has sufficient temporal information for the application. Articles related to drug-drug interaction (DDI) and drug-adverse events (DAE) connect an event to its DCT to create timelines containing events and drugs mentions, as in Iyer *et al.* [48] to adjust the disproportionality ratios,

or as in Iyer *et al.* [49] to generate odds ratios to learn DDI signals. Another example is Liu's *et al.* study [83], which creates a timeline to generate drug-disease pairs and uses a support vector machine (SVM) classifier to distinguish between a drug-indication and a DAE. Temporality was used as a feature by Lee *et al.* [62] for an SVM classifier for the detection of clopidogrel-induced bleedings.

Regarding disease identification, some articles linked an event to its DCT, as in Yetisgen-Yidilz *et al.* [147] and Bejan *et al.* [5] for predicting pneumonia status during ICU stay; in Yetisgen-Yidilz *et al.* [147], which used a Maximum Entropy (ME) classifier; and in Bejan *et al.* [5], which used an SVM classifier. Wu *et al.* [143] tested several temporal aggregation methods for asthma status identification.

In some cases, temporality involved having more information than only the DCT of the event. It required more detailed information. This was done by defining time bins for which the temporality could be more specific. One example was creating time bins to differentiate between events that occurred before the DCT and events that happened after the DCT. The time bins typically depend on the final extraction objective.

Harkema *et al.* [45] developed ConText, a regular expression-based algorithm to extract negation, experiencer, and temporality from clinical texts. The time bins were defined as "historical" (events that started more than two weeks before the DCT), "recent" (events that started less than two weeks before the DCT), and "hypothetical" (neither recent nor historical events). Afzal *et al.* [1] adapted ConText to Dutch, adding new regular expressions to cover with "historical" events. Harkema *et al.* [45] achieved better results for the time bins.

Gaizauskas *et al.* [37] developed a preliminary study related to the Clinical e-Science Framework (CLEF) project by designing a full rule-based system to extract DCT relation and classify it into the following time bins: "BEFORE," "AFTER," and "IS_INCLUDED."

For The 2014 i2b2/UTHealth challenge dataset, we have reviewed 15 articles with seven of them using rules to deal with temporality.

Some authors used approaches based on a "default" time attribute (time bins) and looked for the most frequent values in the training set for each risk factor. For example, in more than 98% of the cases of coronary artery disease (CAD) mentions, the time attribute was of the "continuing" type ("before DCT" relation and also "during DCT" relation), and this was considered the "default value" for CAD mentions [146].

Several authors used the "default value" assignment. Khalifa and Meystre [54] adapted several frameworks to deal with the challenge, and used the "default value" strategy for the temporality. Urbain [132] also used the "default values" for DCT relation, and showed that using a distributional semantic model to capture event context improved the results. The semantic model increased the F-measure from 0.838 to 0.89.

In addition to the "default value" assignment, some authors added rules. Karystianis *et al.* [53] used a lexicalized rule-based approach, and Yang and Garibaldi [146] used rules for the task of finding and managing exceptions. They [146] also explored several strategies to refine the annotations, improving the F-measure from 0.861 to 0.915 with the annotation refinement alone.

Some authors developed a full rule-based system. Shivade *et al.* [115] elaborated a system that was fully based on rules, and when creating rules for DCT relation, considered ConText output, which determines the temporal aspect of concepts. According to the authors, it improved the results. Some authors tried approaches to enrich the context, refine the annotations, or even produce new annotations. Chang *et al.* [18] proposed a context-aware approach that refined recognized contexts until the time attribute was extracted, compared the results with a non-context-aware method (the "default value" assignment), and achieved a higher F-measure using the context-aware method owing to higher precision.

The 2014 i2b2/UTHealth had no specific method to measure the DCT relation module performance for each system and used only a score based on the whole system performance. The only exception was in a study by Grouin, Moriceau, and Zweigenbaum [43], who proposed an evaluation script to separate the evaluation into two factors—the risk factor detection and the temporal relation extraction. Regardless of this evaluation issue, we noticed that articles that relied only on "default value" assignment and did not add any specific rule to deal with special cases or machine learning approaches to its DCT relation system, achieved the poorest overall results (see sub-sections 4.2, 4.3 and 4.4 for more details).

For the i2b2 2012 challenge dataset Chang *et al.* [19] and Wang *et al.* [141] used a rule-based system for DCT relations. The study by Chang *et al.* [19] classified its own performance as "poor" for this type of temporal relation. Wang *et al.* [141] used a rule-based approach that relied only on syntactic and mention attributes and properties. For the i2b2 2012 dataset, machine learning-based or hybrid systems supported by rules achieved better results than the DCT relation extraction (see sub-section 4.2).

DCT RELATIONS MACHINE LEARNING AND HYBRID SYSTEMS

For classifying events into time bins, machine learning approaches and hybrid approaches (those that consider both machine learning and rules) were popular. The systems that extracted DCT relations with machine learning or hybrid systems are listed in Table 2. The table contains information about the article, the main objective of the article, the strategy used to extract temporality, and the article result, which is related with the "Sep DCT" (separated DCT relation evaluation). If the article had a separate evaluation, the result regards the temporal extraction, if not, the results regard the system main objective.

Table 2: Articles related to DCT relation that used machine learning or hybrid systems.

Authors	Objective	Best strategy	Results	Sep DCT
Henriksson *et al.* [46]	DAE IE	2 RF clfs	Fm 0.243 "PAST," 0.220 "FUTURE"	Yes
Mowery *et al.* [95]	Comparison between ConText and ML rule learners	RIPPER clf	Acc 0.971	Yes
Zhu, Yang, and Yan [150]	TIE from online health communities	3 SVM clfs	Fm 0.535	Yes
Raghavan, Fosler-Lussier, and Lai [108]	Investigating the task of working with time bins	CRF clf	Fm 0.84	Yes
Raghavan, Fosler-Lussier, and Lai [106]	co-reference resolution (time bins as features)	CRF clf	Fm 0.7574 AD, 0.9322 BA, 0.4610 OA, 0.5744 WBA and 0.9630 AA	Yes
Lin *et al.* [75]	Identify DAE (temporal features)	SVM clf	Fm of 0.829	No
Torii *et al.* [127]	UTHealth 2014 RFE	21 RIPPER clfs + voting	Fm 0.9185	No
Grouin, Moriceau, and Zweigenbaum [43]	UTHealth 2014 RFE	OneRule clfs	Fm 0.857	No
Chen *et al.* [20]	**UTHealth 2014 RFE**	**label-powerset strategy + SVM clfs**	**Fm 0.9268**	**No**
Goodwin and Harabagiu [39]	UTHealth 2014 RFE	Markov networks + rules	Fm 0.9098	No
Cormack *et al.* [29]	UTHealth 2014 RFE	CART DT + df	Fm 0.917	No
Jonnagaddala *et al.* [52]	UTHealth 2014 RFE	NB clf + rules	Fm 0.8302	No
Roberts *et al.* [110]	**UTHealth 2014 RFE**	**3 SVM clfs + df + rules + ann refinement**	**Fm 0.9277**	**No**
Cherry *et al.* [23]	**i2b2 2012 TRE**	**SVM clf**	**Fm 0.6954**	**No**
D'Souza and Ng [30]	**i2b2 2012 TRE**	**CRF clf**	**Fm 0.693**	**No**
Tang *et al.* [126]	**i2b2 2012 TRE**	**2 SVM clfs**	**Fm 0.6932**	**No**
Xu *et al.* [145]	i2b2 2012 TRE	2 SVM clfs	Fm 0.6849	No
Lin *et al.* [74]	**i2b2 2012 TRE and CTE TRE**	**2 SVM clfs (i2b2) and SVM clf (CTE)**	**Fm 0.695 (i2b2), 0.807 (CTE)**	**No**
Nikfarjam, Emadzadeh, and Gonzalez [96]	i2b2 2012 TRE	SVM clf + rules	Fm 0.63	No
Roberts, Rink, and Harabagiu [109]	i2b2 2012 TRE	SVM clf + rules	Fm 0.5594	No
Styler *et al.* [121]	THYME corpus TRE	SVM clf	Fm 0.474	Yes
Velupillai *et al.* [135]	**CTE 2015 TRE**	**CRF clf**	**Fm 0.791**	**Yes**
Fries [36]	CTE 2016 TRE	LR clf + rules	Fm 0.743	Yes
Khalifa, Velupillai, and Meystre [55]	**CTE 2016 TRE**	**CRF clf**	**Fm 0.844**	**Yes**
Chikka [24]	CTE 2016 TRE	CRF clf	Fm 0.714	Yes
Caselli and Morante [15]	CTE 2016 TRE	CRF clf	Fm 0.712	Yes
Grouin and Moriceau [42]	CTE 2016 TRE	CRF clf	Fm 0.687	Yes
Lee *et al.* [63]	CTE 2016 TRE	SVM clf	Fm 0.835	Yes
Tourille *et al.* [130]	**CTE 2016 + MERLOT TRE**	**SVM clf (CTE) and SVM clf (MERLOT)**	**Fm 0.87 (CTE), 0.83 (MERLOT)**	**Yes**
Tourille *et al.* [129]	CTE 2016 TRE	RF clf	Fm 0.807	Yes

Authors	Objective	Best strategy	Results	Sep DCT
Cohan, Meurer, and Goharian [27]	CTE 2016 TRE	LR clfs	Fm 0.815	Yes
MacAvaney, Cohan, and Goharian [87]	CTE 2017 TRE	CRF clf	Fm 0.40 UDA, 0.50 SDA	Yes
Sarath, Manikandan, and Niwa [111]	**CTE 2017 TRE**	**CRF clf**	**Fm 0.45 UDA, 0.52 SDA**	**Yes**
Huang *et al.* [47]	CTE 2017 TRE	SVM clf	Fm 0.49 SDA	Yes
Tourille *et al.* [131]	**CTE 2017 TRE**	**SVM clf**	**Fm 0.519 UDA, 0.591 SDA**	**Yes**
Leeuwenberg and Moens [70]	**CTE 2016 TRE**	**Structured perceptron + ILP**	**Fm 0.846**	**Yes**
Leeuwenberg and Moens [68]	**CTE 2017 TRE**	**Structured perceptron + ILP**	**Fm 0.49 UDA, 0.56 SDA**	**Yes**
Viani *et al.* [139]	Cardiology notes TRE	SVM clf	Fm 0.857 "OVERLAP," 0.834 "BEFORE," 0.793 "AFTER"	Yes

Legend: DAE: drug-adverse events, ML: machine learning, RF: random forest, CRF: conditional random fields, DCT: document creation time, SVM: support vector machine, LR: logistic regression, NB = naïve bayes, DT: decision trees, AD: after discharge, BA: before admission, OA: on admission, WBA: way before admission, AA: after admission, UDA: unsupervised domain adaptation, SDA: supervised domain adaptation, Fm: F-measure, RFE: risk factor extraction, acc: accuracy, TRE: temporal relation extraction, ann: annotation, , IE: information extraction, clf: single classifier, clfs: more than one classifier, df : default value, CTE: Clinical TempEval, TIE: temporal information extraction, ILP: integer linear programming.

Several articles that were not related to a shared-task dataset were based on traditional machine learning. Henriksson *et al.* [46] defined time bins as "past" (before the DCT) and "future" (after the DCT) using a random forest (RF) classifier for each time bin. Mowery *et al.* [95] compared ConText to the Decision Tree, Ripper, and Rule Learner (RL) algorithms. The time bins are the same as in the ConText articles in sub-section 4.1, and the best results were obtained using Ripper. Zhu, Wang, and Yan [150] developed an SVM classifier for each relation type ("AFTER," "BEFORE," and "OVERLAP") when working with Chinese consultation posts gathered from an online health community.

Both the studies by Raghavan, Fosler-Lussier, and Lai [108], [106] developed a framework based on a conditional random fields (CRF) classifier to classify events into one of the following time bins: "way before admission," "before admission," "on admission," "after admission," and "after discharge." These articles differentiate clearly between past events according to their time distance from admission, and they also differentiate future events after admission (events while the patient is admitted) and after discharge. In the first study by Raghavan, Fosler-Lussier, and Lai [108], the main focus was on how to work with time bins, but in the second [106], the time bins were a methodology step for managing co-reference resolution.

For The 2014 i2b2/UTHealth dataset, in addition to the rule-based systems mentioned in sub-section 4.1, machine learning and hybrid systems were used in seven publications.

In the machine learning systems, Torii *et al.* [127] proposed a framework that used common features, and also used hot-spot features with 21 RIPPER classifiers and majority voting. Grouin, Moriceau, and Zweigenbaum [43] tested the performance of the OneRule classifier versus SVM, training one OneRule classifier for each type of risk factor and achieving results superior or equal to those of the SVM for each risk factor. Chen *et al.* [20] used a label-powerset strategy to transform the classification into a single-label classification problem, and then used SVM. The features used included unigrams, bigrams, and the term frequency-inverse document frequency (TF-IDF) of words.

Goodwin and Harabagiu [39] proposed a hybrid system with a probabilistic framework based on Markov networks and leveraged rules. Cormack *et al.* [29] explored a rule-based method that did not require the use of domain knowledge, using the "default value" strategy, and also used a CART decision tree model for processing numeric test results. Jonnagaddala *et al.* [52] used a naïve Bayes classifier combined with rules to perform the temporal task. Roberts *et al.* [110] used three SVM classifiers, one for each type of relation ("AFTER DCT," "BEFORE DCT," and "DURING DCT"), and the output of the classifiers was passed to a set of constraints and exceptions. This set of constraints and exceptions was based on "default values" and crafted rules. Roberts *et al.* [110] also explored the impact of fine-grained annotation on the performance of the system, and tested two sets of annotations. One set was the original annotations, and the other set was fine-grained, mention-level annotations (NLM annotation) created by Roberts *et al.* [110]. The NLM annotations, which were made especially for the challenge, gave better results than the original annotation.

For the 2014 i2b2/UTHealth dataset, the ML classifiers (full classifiers or hybrid approaches) achieved the best overall results. Both Roberts *et al.* [110] and Chen *et al.* [20] relied on several SVM classifiers, Roberts *et al.* [110] also added rules to cover constraints and exceptions. Torii *et al.* [127] used RIPPER models and majority voting, Cormack *et al.* [29] used CART decision tree and rules, Yang and Garibaldi [146] and Shivade *et al.* [115] relied on rule-crafting to achieve good results, but Roberts *et al.* [110] and Chen *et al.* [20] obtained the best results. As shown by some authors, the impact of refining the annotations, producing a fine-grained annotation, and using context-aware approaches resulted in better overall results than when those techniques were not used.

As mentioned in sub-section 4.1, most articles related to the i2b2 2012 dataset used machine learning or hybrid approaches. These approaches are described below.

Some authors developed machine learning approaches (no rules added) using a single classifier for both relations between event and section-times (event-admission relations and event-discharge relations). Cherry *et al.* [23] used an SVM classifier to create seven categories containing all the combinations of section-times and relation types (including a "no-relation" category). D'Souza and Ng [30] used a single CRF classifier for both section-times, transforming the classification problem into a sequence labeling task.

Defining classifiers according to section-times was used by Tang *et al.* [126], who used an SVM classifier for event-admission relations, and another SVM classifier for event-discharge relations to create specialized classifiers for each Section-time. Xu *et al.* [145] also used a separate SVM classifier for each Section-time. Similarly, Lin *et al.* [74] created separate classifiers for each Section-time to develop an open-source clinical temporal relation discovery system that was part of Apache cTAKES temporal module, considering only features that were extracted with Apache cTAKES. This module was later used by Lin *et al.* [75] to add temporality to the identification of liver toxicity due to the usage of methotrexate in patients with rheumatoid arthritis.

Hybrid systems by Nikfarjam, Emadzadeh, and Gonzalez [96] and Roberts, Rink, and Harabagiu [109] used rules to link events to a Section-time using a classifier afterwards. Roberts and colleagues used an SVM classifier.

Regarding the i2b2 2012 dataset, we concluded that the best results were obtained in studies that used specialized classifiers for relations between events and each Section-time (one for event-admission and another for event-discharge), especially with SVM classifiers in the studies of Tang *et al.* [126] and Lin *et al.* [74]. Best results were also obtained in studies that used a single classifier. Cherry *et al.* [23] used SVM, and D'Souza and Ng [30] used CRF to create categories that involved all relations (time bins) for both section-times. Studies that created rules to link events to section-times, and then used a single classifier, as did Nikfarjam, Amadzadeh, and Gonzalez [96] and Roberts, Rink, and Harabagiu [109] achieved the worst overall results.

The remaining studies were related to the THYME corpus. Approximately 44% of the studies in this sub-section (machine learning/hybrid systems) sub-section 4.3 (deep learning systems) were related to the THYME corpus. The study by Styler *et al.* [121] was a preliminary work about the THYME corpus before the release of the Clinical TempEval 2015 challenge dataset, which annotated a portion of the data and applied ClearkTK-TimeML [6], an SVM-based system that was among the best performers in the TempEval 2013 [133] shared-task for the general domain. The domain (clinical vs. general) impacted the system performance.

Regardless of the Clinical TempEval dataset (2015, 2016, or 2017), the majority of the frameworks were based only on machine learning (no rule crafting or pattern matching was used), except for the study by Fries [36], which used logistic regression (LR) in combination with rules.

Studies such as that by Velupillai *et al.* [135], Khalifa, Velupillai, and Meystre [55], Chikka [24], Caselli and Morante [15], Grouin and Moriceau [42], MacAvaney, Cohan, and Goharian [87] and Sarath, Manikandan, and Niwa [111] treated DCT relation classification as a sequence labeling task, and used a single CRF classifier. Sarath, Manikandan, and Niwa [111] used the ClearTK-TimeML named-entity recognition (NER) chucking classifier, which is CRF-based.

A single SVM classifier was used for a multi-class classification task by Lin *et al.* [74], Lee *et al.* [63], Tourille *et al.* [130], Tourille *et al.* [131], and Huang *et al.* [47]. Tourille *et al.* [129] used an RF classifier. Cohan, Meurer, and Goharian [27] used a LR classifier for each type of relation.

The publications of Leeuwenberg and Moens [68, 70] focused in structured machine learning, jointly predicting DCT relations and TLINKs using a structured perceptron model and integer linear programming (ILP).

Similar to THYME relations, the publication by Viani *et al.* [139] worked with Italian cardiology texts, using the same DCT relation types as in the THYME corpus, and an SVM classifier to classify events into one of the time bins.

For the THYME corpus, we concluded that the best approaches were the ones that used a single classifier for the task, either SVM or CRF. The difference between those approaches relied on feature engineering and added features to leverage the surrounding context and the attributes and characteristics of the event.

DCT RELATIONS DEEP LEARNING SYSTEMS

For classifying events into time bins, deep learning systems were used. The systems that extracted DCT relations with deep learning are listed in Table 3. The table contains information about the article, the main objective of the article, the strategy used to extract temporality, and the article result, which is related with the "Sep DCT" (separated DCT relation evaluation). If the article had a separate evaluation, the result regards the temporal extraction, if not, the results regard the system main objective.

Table 3: Articles related to DCT relation that used deep learning systems.

Authors	Objective	Best strategy	Results	Sep DCT
Li, Jagannatha, and Yu [73]	Extracting presence and period assertions	GRU + deep residual networks + Att	Micro Acc 0.8410 period assertion	Yes
Su *et al.* [122]	RFE	CNN	Fm 0.9812	Yes
Chokwijitkul *et al.* [25]	UTHealth 2014 RFE	BI-LSTM	Fm 0.9081	No
Li and Huang [72]	CTE 2016 TRE	CNN + MLP	Fm 0.788	Yes
Long *et al.* [85]	CTE 2017 TRE	RNNs	Fm 0.32 SDA	Yes

Legend: DCT: document creation time, UDA: unsupervised domain adaptation, SDA: supervised domain adaptation, Fm: F-measure, Acc = accuracy, RFE: risk factor extraction, TRE: temporal relation extraction, GRU: gated recurrent unit, CNN: convolutional neural network, MLP: multilayer perceptron, LSTM: long short-term memory, RNN: recurrent neural network, att: attention mechanism, CTE: Clinical TempEval.

Several articles developed frameworks based on deep learning. Li, Jagannatha, and Yu [73] aimed at extracting a presence assertion (similar to the i2b2 2010 shared task) but also extracted a period assertion, classifying events into "history" (past event), "current" (actual event), "future," and "unknown" time bins. The developed framework was based on gated recurrent unit (GRU) networks, deep residual networks, and an attention mechanism, and it achieved superior results when compared with both SVM and LSTM. Su *et al.* [122] tested both SVM and convolutional neural networks (CNNs), CNN model adapted from Kim [57] for risk factor identification over Chinese clinical texts using objectives similar to that of the 2014 i2b2 UTHealth. The best results were achieved with CNNs for time assertion. The time bins were related to the duration of hospital stay (DHS): "before the DHS," "during the DHS," "after the DHS," and "continuing."

Regarding the 2014 i2b2/UTHealth dataset, Chokwijitkul *et al.* [25] evaluated a neural network-based model, with word embeddings trained using word2vec [89] on the challenge dataset, and tested CNNs, long short-term memory (LSTM), bidirectional long short-term memory (BI-LSTM), and GRU, achieving the best results with BI-LSTM. For these dataset publications in sub-section 4.2 (machine learning and hybrid systems), they have achieved the best results when compared to deep learning.

The remaining studies were related to the THYME corpus. For the Clinical TempEval 2016 dataset Li and Huang [72] used a CNN with a multilayer perceptron (MLP). For the Clinical TempEval 2017 dataset Long *et al.* [85] used a recurrent neural network (RNN) classifier for each relation type.

DCT RELATIONS CONCLUSIONS

Depending on the primary objective of the article, merely connecting an event to its DCT could provide sufficient temporal information for research; whereas, a more detailed representation leveraging not only the connection between events and DCT but also classifying their relation into time bins increase the extraction difficulty. There is a trade-off between adding time bins, adding more specific temporality, and task difficulty. For example, classifiers show superior performance classifying events into "past" and "future," as in Henriksson *et al.* [46], compared with having more detailed time bins as in the THYME corpus ("BEFORE," "AFTER," "OVERLAP," and "BEFORE/OVERLAP"). In contrast, a classifier needs to leverage information to distinguish events that

happened in the past ("BEFORE" time bin) from events that occurred in the past and continued until the DCT ("BEFORE/OVERLAP" time bin).
For i2b2 2014 UTHealth dataset, best results were obtained using several SVM classifiers by Roberts *et al.* [110] and Chen *et al.* [20], with Roberts *et al.* [110] also adding rules. For the i2b2 2012 dataset approaches based on SVM (two classifiers, one for each Section-time), as in Tang *et al.* [126] and Lin *et al.* [74], and CRF (single classifier creating categories for all possible combinations of relations and section-times), as in D'Souza and Ng [30] and Cherry *et al.* [23], achieved best results.
For the THYME corpus, the use of a single SVM classifier or a single CRF classifier achieved the best results. We highlight the CRF based approaches of Velupillai *et al.* [135], Khalifa, Velupillai, and Meystre [55] and Sarath, Manikandan and Niwa [111], and the SVM based approaches of Lin *et al.* [74], Tourille *et al.* [130] and Tourille *et al.* [131].
Overall the best results are related to the use of machine learning or hybrid systems based on either SVM or CRF classifiers, except for Leeuwenberg and Moens [68, 70] which achieved significant results with a structured perceptron model. The feature set for these SVM and CRF based approaches involved features as token information (e.g., shape, n-grams) and syntactic features (e.g., part-of-speech tagging) over events and nearby tokens, information about nearby verbs (e.g., tense and part-of-speech tagging), event attribute information (e.g., event type), and also information about document section and nearby temporal expressions. Some authors also relied on semantic features based on lexicons and the Unified Medical Language System (UMLS).

TLINKS

TLINKs are considered to be any type of relation between a pair of mentions (either events or temporal expressions) in which temporal expressions are not the DCT. The results were sorted according to the strategy used to define the relation with two categories: rule-based/pattern matching systems (see sub-section 5.1), machine learning and hybrid systems (see sub-section 5.2) and deep learning systems (see sub-section 5.3). As we provide an in-depth overview and evaluation of all the selected articles, we compile a summary (see sub-section 5.4) with highlights and conclusions.

TLINKS RULE-BASED SYSTEMS

The systems that extracted TLINKs with rule-based systems are detailed in Table 4. The table contains information about the article, the article main objective, the strategy used to extract temporality, and the article results, which are related to the "sep TLINK" (separated TLINK relation evaluation). If the article had a separate evaluation, the result is regarding the temporal extraction, if not the results are regarding the system main objective.

Table 4: Article-related TLINK extraction with full rule-based systems.

Authors	Objective	Strategy	Results	Sep TLINK
Li and Patrick [71]	Extracting temporal information from a noisy corpus	WS: extract TIs after EVTs and correlate them (rule)	Fm 0.743 (Identification of temporal EVTs)	No
Denny *et al.* [32]	Extracting TIs and status (EVTs) for patient screening	WS: correlated TI to its nearest EVT (token distance)	Fm 0.9296 in identifying TIs and correlating with EVTs	No
Liu *et al.* [82]	Correlating lab test results in clinical notes to structured data	WS: linked EVTs to TIs (rules)	Fm 0.966	No
Savova *et al.* [113]	Discovering drug treatment patterns	Pattern matching to link dates to EVTs	Pr 0.80-0.9738 for the categories	No
Viani *et al.* [138]	TRE from clinical texts in psychiatric domain	Jointly normalize TIs and correlate EVTs with TIs	Fm 0.58 (in identifying a relation)	Yes
Wang *et al.* [141]	Safety surveillance reports and i2b2 2012 TRE	Rules based on element information and basic syntactic	Match ratio 0.75 (VAERS + FAERS), 0.57 (i2b2 2012) for timestamp generation	Yes

Authors	Objective	Strategy	Results	Sep TLINK
Lee and Choi [66]	Extracting temporal segments	Rules based on observations to delimiter segments	Fm 0.86	No
Xu *et al.* [144]	Extracting tuples of EVTs, TIs and descriptions	Rules to correlate EVTs and TIs in same segment of text	Event-time correct linking in 98.5% of the cases	Yes
Seol *et al.* [114]	Identifying problem-action relations in clinical texts	Linked EVTs to TIs by rules into a CSU unit	Fm 0.788 (CSU classification)	No
Gaizauskas *et al.* [37]	Temporal information extraction	WS: rules to link EVTs to dates	Fm 0.7169	Yes
Stubbs and Harshfield [119]	Adapting TTK to clinical domain	Rules to link EVTs and TIs	Acc 0.84	No
Zhou, Parsons and Hripcsak [149]	Evaluating TimeText in discharge summaries	EE relations with TimeText	96.5% of relations were correct (specialist evaluation)	Yes
Capurro *et al.* [13]	Identifying patients with acute kidney injury	Patient pattern matching	Correctly classified 88% of the patients	No

Legend: WS: within-sentence relation, FAERS: Food and Drug Administration (FDA) adverse event reporting system, CSU: clinical semantic unit, TTK: TARSQI toolkit, VAERS: US vaccine adverse event reporting system, TI: time, EVT: event, Fm: F-measure, Pr: precision, TRE: temporal relation extraction, EE: event-event.

A method to deal with temporal relations is by creating rules to identify relations between events and expressions without inferring any relation type (merely identifying the relation).

This approach was used by Li and Patrick [71], who first identified temporal expression in a sentence, and then identified the events, correlating them afterward. Denny *et al.* [32] also used rules, assigning each temporal expression to its nearest (smaller token distance) event. Liu *et al.* [82] used rules to correlate glucose and HbA1c lab test results in clinical texts with structured data to link events and temporal expressions within the same sentence. Savova *et al.* [113] defined temporal dates (temporal expressions regarding dates) for each drug mention to identify drug treatment patterns (which included medication usage order). Viani *et al.* [138] developed a rule-based system to jointly normalize temporal expression and identify temporal relations between events and temporal expressions, and labeled them as "yes" (indicating a temporal relation) or "no" (indicating no temporal relation). This work was related to temporal identification relating to the beginning of psychosis symptoms. Wang *et al.* [141] used a rule-based approach that relied only on syntactic and mention attributes and properties to extract relations between events and timestamps (temporal expressions of "date" type) over the i2b2 2012 dataset.

Lee and Choi [66] worked with temporal segmentation, in which each segment contained topics with the same temporal or topical content. The segmentation was done by a rule-based system based on observations on Korean discharge summaries. Xu *et al.* [144] extracted description-related tuples from Chinese medical text, linking events to temporal expressions in the same statement (a segment of text until a period symbol) to identify tuples composed of temporal expressions, events, and descriptions. Similarly, Seol *et al.* [114] constructed a clinical semantic unit (CSU), similar to a tuple, composed of events and temporal expressions using rules to identify problem–action relations in Korean clinical texts.

In addition to identifying the existence of a relation, identifying the type of relation increases the amount of temporal information extracted over the text, and improves the temporal ordering. Gaizauskas *et al.* [37] developed a preliminary study related to the CLEF project by designing a rule-based system to extract within-sentence relations between events and temporal dates (temporal expression regarding dates) that considered the relation types "BEFORE," "AFTER," and "IS_INCLUDED." Stubbs and Harshfield [119] adapted the TARSQI Toolkit (TTK) [137] Blinker module, which was developed for the newswire domain, to the clinical domain. Blinker uses rules to create relations between events and temporal expressions, with the relation being based on the TimeML markup language. Zhou, Parsons, and Hripcsak [149] used the TimeText [148] temporal relation extraction module to extract relations between pairs of events to evaluate the TimeText performance on discharge summaries. Capurro *et al.* [13] searched for patients that matched certain criteria in clinical texts by performing a temporal query search using the framework ClinicalTime, which leverages temporality using Allen's relations.

TLINKS MACHINE LEARNING AND HYBRID SYSTEMS

For temporal relation extraction, especially that regarding challenge task datasets, machine learning and deep learning systems achieves better results than does rules. Some authors use hybrid approaches by adding rules to cover cases that classifiers have trouble handling.
The systems that extracted TLINKs using machine learning or hybrid systems (machine learning and rules) are detailed in Table 5. The table contains information about the article, the main objective of the article, the strategy used to extract temporality, and the article result, which is related to the "sep TLINK" (separated TLINK relation evaluation). If the article had a separate evaluation, the result is regarding the temporal extraction, if not, the result is regarding the system main objective.

Table 5: Articles related to TLINK extraction with machine learning or hybrid systems.

Authors	Objective	Strategy	Candidate pair selection	Result	Sep TLINK
Luo *et al.* [86]	Extracting constraints from eligibility criteria	CRF clf	-	Fm 0.7981	No
Bransem *et al.* [10]	TO through temporal segments	BoosTexter	-	Acc 0.783	Yes
Bransem *et al.* [11]	TO through temporal segments	BoosTexter + Perceptron + ILP	-	Acc 0.84	Yes
Raghavan *et al.* [105]	Event alignment	ME clf	-	Fm 0.673	Yes
Raghavan, Fosler-Lussier, and Lai [107]	Event ordering (pairwise clf, ranking)	SVM clf (pairwise), SVM clf (ranking)	-	Acc 0.8216 (ranking), 0.7133 (pairwise)	Yes
Chang *et al.* [19]	i2b2 2012 TRE	2 ME clfs + rules	Rules	Fm 0.5628	No
Grouin *et al.* [41]	i2b2 2012 TRE	9 clfs + rules	Cross-product	Fm 0.6231	No
Roberts, Rink, and Harabagiu [109]	i2b2 2012 TRE	SVM ranker + SVM clf	Rules	Fm 0.5594	No
Moharasan and Ho [93]	i2b2 2012 TRE	CS: NB clf, WS: 2 NB clfs	Rules	Fm 0.671	No
Cherry *et al.* [23]	**i2b2 2012 TRE**	**WS: 2 ME clfs, CS: 1 ME clf + rules**	**WS: APP, CS: rules**	**Fm 0.6954**	**No**
Lin *et al.* [74]	**i2b2 2012 and CTE 2015 TRE**	**WS: 2 SVM clfs, CS: 2 SVM clfs + rules (CS only in i2b2 2012). CSL; TS expansion**	**WS: APP, CS: rules**	**Fm 0.695 (i2b2) , 0.321 (CTE)**	**No**
Xu *et al.* [145]	**i2b2 2012 TRE**	**WS: 3 SVM clfs, CS: 3 SVM clfs**	**WS: APP, CS: rules**	**Fm 0.6849**	**No**
Sohn *et al.* [116]	i2b2 2012 TRE	SVM clf + rules	WS: APP, CS: rules	Fm 0.537	No
Cheng *et al.* [22]	i2b2 2012 TRE	WS: ME clf + conflict resolution, CS: rules	WS: rules	Fm 0.43	No
Nikfarjam, Emadzadeh, and Gonzalez [96]	i2b2 2012 TRE	WS: 2 SVM clfs + temporal graph, CS: rules	WS: APP, graph	Fm 0.63	No
Tang *et al.* [126]	**i2b2 2012 TRE**	**WS: 2 SVM clfs, CS: 2 SVM clfs**	**WS: rules, CS: rules**	**Fm 0.6932**	**No**
D'Souza and Ng [30]	**i2b2 2012 TRE**	**Based on Tang *et al.* [126] + rules**	**Based on Tang *et al.* [126]**	**Fm 0.693**	**No**
D'Souza and Ng [31]	**i2b2 2012 TRE**	**Based on Tang *et al.* [126] + rules**	**Based on Tang *et al.* [126]**	**Fm 0.702**	**No**
Lee *et al.* [64]	Re-annotated i2b2 2012 TRE to direct relations	SVM clfs + rules + CSL	WS: APP	Fm 0.6377 (NTC)	No
D'Souza and Ng [117]	Annotated missing cross-sentence relations i2b2 2012 TRE	D'Souza and Ng [30], D'Souza and Ng [31]	D'Souza and Ng [30], D'Souza and Ng [31]	Fm 0.341 (NTC)	No
Miller *et al.* [90]	THYME corpus TRE	SVM clf	WS: APP	Fm 0.737 (NTC)	Yes

Authors	Objective	Strategy	Candidate pair selection	Result	Sep TLINK
Lin *et al.* [80]	THYME corpus TRE	SVM clf	WS: APP	Fm 0.708 (NTC)	Yes
Styler *et al.* [121]	THYME corpus TRE	2 SVM clfs	WS: APP	Fm 0.204 (NTC)	Yes
Velupillai *et al.* [135]	CTE 2015 TRE	CRF clf + rules	Rules	Fm 0.181	Yes
Khalifa, Velupillai, and Meystre [55]	CTE 2016 TRE	WS: 2 SVM clfs, CS: 2 SVM clfs	WS: APP, CS: rules	Fm 0.511	Yes
Caselli and Morante [15]	CTE 2016 TRE	WS: 2 CRF clfs	WS: APP	Fm 0.453	Yes
Barros *et al.* [4]	CTE 2016 TRE	4 CRF clfs	WS: APP, CS: rules	Fm 0.264	Yes
Tourille *et al.* [130]	CTE 2016 + MERLOT TRE	WS: SVM clf. 3-class	WS: APP	Fm 0.53 (CTE), 0.65 (MERLOT)	Yes
Tourille *et al.* [129]	CTE 2016 TRE	WS: SVM clf, CS: SVM clf. Rules; 3-class	WS: APP, CS: rules	Fm 0.538	Yes
Lin *et al.* [77]	**CTE 2016 TRE**	**WS: 2 SVM clfs. TS expansion**	**WS: APP**	**Fm 0.594**	**Yes**
Lee *et al.* [63]	**CTE 2016 TRE**	**WS: 2 SVM clfs, CS: 4 SVM clfs. CSL**	**WS: APP + pair filtering, CS: rules**	**Fm 0.573**	**Yes**
Chikka [24]	CTE 2016 TRE	CRF clf	-	Fm 0.313	Yes
Leeuwenberg and Moens [67]	CTE 2016 TRE	2 SVM clfs	WS: APP + restrictions, CS: rules	Fm 0.551	Yes
Leeuwenberg and Moens [70]	CTE 2016 TRE	Structured perceptron + ILP	APP over TK + rules	Fm 0.608	Yes
Sarath, Manikandan, and Niwa (SARATH; MANIKANDAN; NIWA, 2017)	CTE 2017 TRE	WS: 2 clf ensembles, CS: 2 clf ensembles. CSL	WS: APP, CS: rules. Pair filtering; rules	Fm 0.23 UDA, 0.15 SDA	Yes
MacAveney, Cohan, and Goharian [87]	**CTE 2017 TRE**	**WS: XGBoost clf**	**WS: APP**	**Fm 0.34 UDA, 0.25 SDA**	**Yes**
Leeuwenberg and Moens [68]	**CTE 2017 TRE**	**Structured perceptron + ILP**	**APP over TK**	**Fm 0.32 UDA, 0.28 SDA**	**Yes**
Huang *et al.* [47]	CTE 2017 TRE	WS: 2 SVM clfs	WS: APP	Fm 0.26 (SDA)	Yes

Legend: CRF: conditional random fields, SVM: support vector machine, ME: maximum entropy, NB: naive Bayes, WS: within-sentence relation, CS: cross-sentence relation, APP: all possible pairs, TS: training set, SRL: semantic role labeling, 3-class: transforming to a 3-class classification task, CSL: cost-sensitive learning, CTE: Clinical TempEval, clf: single classifier, clfs: more than one classifier, Fm: F-measure, Acc: accuracy, ILP: integer linear programming, TO: temporal ordering, TRE: temporal relation extraction, NTC: not comparable, UDA: unsupervised domain adaptation, SDA: supervised domain adaptation, ILP: integer linear programming.

Luo *et al.* [86] developed a temporal ontology based on eligibility criteria to extract temporal constraints based on a CRF classifier. An event may contain no temporal constraints, or one or more temporal constraints, and each constraint contains temporal expressions and temporal relations.

Regarding temporal segmentation, Bransem *et al.* [10] and Bransem *et al.* [11] identified temporally related segment pairs by considering their relations ("BEFORE," AFTER," and "NO RELATION"). Bransem *et al.* [10] and Bransem *et al.* [11] used BoosTexter classifier, based on boosting. Bransem *et al.* [11] also used a perceptron classifier and ILP.

The temporal alignment of events was extracted by Raghavan *et al.* [105] using a weighted finite-state transducer (WFST) approach. The temporal relation extraction was one of the minor steps, and pairs of events were classified using a maximum entropy classifier into one of the categories ("BEFORE," AFTER," and "OVERLAP"). Also working with events, Raghavan, Fosler-Lussier and Lai [107] evaluated event ranking and pairwise classification for event ordering. The pairwise classification (considering pairs of events) into the relations ("BEGINS," "END," "SIMULTANEOUS," "INCLUDES," or "BEFORE") was done with an SVM classifier, and the event-ranking was done with an SVM ranker. The best results were achieved with event-raking over pairwise classification.

As i2b2 2012 was the first dataset involving temporal relation extraction from clinical texts, several articles (17) were selected in this review related to this dataset. From these publications, 15 were based on machine learning or are hybrid systems.

The results over the i2b2 2012 are summarized according to three aspects: the candidate pair selection technique, the approach to extract within-sentence relations (pairs of mentions in the same sentence), and cross-sentence relations (pairs of mentions in different sentences). These aspects have a major impact on the results.

As any events or temporal expression mention can generate a candidate pair when a classifier is trained, this would create a high number of negative samples in the training set, as most of the pairs would not have a temporal relation. Owing to the predominance of the classes with negative samples, there would be great imbalance in the class distribution, which negatively impacts the classification results [109]. Cross-sentence relations are even more complicated, as the number of positive pairs diminishes, while the number of possible pairs rises as more sentences are considered.

Regarding candidate pair generation, Chang *et al.* [19] defined heuristic rules, Grouin *et al.* [41] used a cross product-based approach, Roberts, Rink, and Harabagiu [109] considered all mentions in previous and current sentences regarding the selected mention (excluding mentions that occur after the select mention). Moharasan and Ho [93] used rules based on event attributes, and the dependency parsing method to generate pairs between mentions (events and temporal expressions). Several authors differentiated clearly between the approach used for generating pairs for within-sentence relations and those for cross-sentence relations, which improved the overall results.

For within-sentence relations, Cherry *et al.* [23], Lin *et al.* [74], Xu *et al.* [145], and Sohn *et al.* [116] considered all possible pairs regarding mentions (both events and times). Lin *et al.* [74] also used cost-sensitive learning to reduce the imbalance problem and added extra training data, utilizing a technique based on event expansion with the UMLS. Cheng *et al.* [22] defined rules for within-sentence relations, and Nikfarjam, Emadzadeh, and Gonzalez [96] considered all possible pairs that were inferred over a graph strategy. Tang *et al.* [126] used a candidate pair approach for within-sentence relations, considering all possible consecutive pairs of mentions in a sentence, as well as pairs that showed a dependency relation according to a dependency parsing tree. Tang's *et al.* [126] approach was used also by D'Souza and Ng [30] and D'Souza and Ng [31].

For cross-sentence relations, Tang *et al.* [126] and Cherry *et al.* [23] focused on event-event relations when generating candidate pairs. Tang *et al.* [126] considered all possible pairs between the first and last events in both sentences (consecutive sentences), and any pair between two events with the same semantic type and same head noun. Tang's *et al.* [126] approach was used also by D'Souza and Ng [30] and D'Souza and Ng [31]. Cherry *et al.* [23] created all possible pairs of events, using a five-sentence window limited to pairs in which the events had matching event attributes. Lin *et al.* [74] considered all possible pairs between the first and last events in both sentences (consecutive sentences). Tang *et al.* [126] considered all possible event-time pairs in consecutive sentences, and also considered coreference pairs defined by rules. Xu *et al.* [145] considered all possible pairs in adjacent sentences (neighbor sentences). Sohn *et al.* [116] defined rules to generate cross-sentence pairs.

According to the results, we concluded that the within-sentence candidate pairs methods that considered all possible pairs between mentions, as used by Cherry *et al.* [23], Lin *et al.* [74], and Xu *et al.* [145], and heuristics that considered consecutive pairs and dependency relation as in Tang *et al.* [126], D'Souza and Ng [30], and D'Souza and Ng [31] achieved the best results. For the cross-sentence relations, the best results were based on approaches by Tang *et al.* [126], Cherry *et al.* [23], D'Souza and Ng [30], and D'Souza and Ng [31] that focused on defining heuristics for selecting certain types of event-event relations, the approach by Xu *et al.* [145] that considered every pair of events and temporal expressions in adjacent sentences, and the approach by Lin *et al.* [74] that considered all possible pairs of events and temporal expressions in consecutive sentences, and defined heuristics for selecting event-event relations and co-referenced pairs.

The best results expanded the training set with transitive closure, generating additional relations that could be marked and were left unlabeled. Dataset imbalance was a problem because the number of gold-standard relations was small in relation to the total number of possible relations. One way to address this imbalance would be to add the transitive closure of the ground truth relations when training the classifier [23]. Adding the transitive would help to reduce the dominant class, which is "no relation," and to increase the number of positive samples. According to Lin *et al.* [74], a large number of implicit relations were left unlabeled because they could be inferred.

Regarding the approaches to extracting relations, some authors classified both within sentences and across sentences using the same approach. Roberts, Rink, and Harabagiu [109] used an SVM ranker to detect relations,

and an SVM classifier to infer the relation type. Grouin *et al.* [41] divided the extraction problem into 57 situations, selected 9 of the situations, defined the best classifiers for each situation, and then added rules and applied a combination of the winners to select the result. Chang *et al.* [19] used a hybrid approach, prioritizing rules over ML. The ML approach was based on two ME classifiers, one to identify relations and another to infer relation type, and had specific rules for within-sentence and cross-sentence relations. Sohn *et al.* [116] also used a hybrid approach, with an SVM classifier in combination with rules, and specific rules for within-sentence and cross-sentence relations.

Most authors distinguished between within-sentence relation extraction and cross-sentence relation extraction. For within-sentence relations, Cherry *et al.* [23], Tang *et al.* [126], Lin *et al.* [74], D'Souza and Ng [30], and D'Souza and Ng [31] used a classifier for event-event relations, and a classifier for event-time relations. D'Souza and Ng [30] and D'Souza and Ng [31] first used rules that had 80% accuracy, and if none applied, they then used an ML-based approach. Tang *et al.* [126], Lin *et al.* [74], D'Souza and Ng [30] and D'Souza and Ng [31] used SVM classifiers, and Cherry *et al.* [23] used ME classifiers. Xu *et al.* [145] also considered event-event and event-time relations with SVM classifiers for both relations, but also used an additional SVM classifier for time-time relations. Cheng *et al.* [22] used an ME classifier and a conflict resolution mechanism. Nikfarjam, Emadzadeh, and Gonzalez [96] used a temporal graph module, and if no relation was inferred by the module, they then used an SVM classifier for event-event relations, and another SVM classifier for event-time relations. Moharasan and Ho [93] used a naïve Bayes classifier for within-sentence relations.

For cross-sentence relations, Tang *et al.* [126], D'Souza and Ng [30], and D'Souza and Ng [31] used an SVM classifier for events in consecutive sentences, and used an SVM classifier for events that were co-referenced. Cherry *et al.* [23] trained a ME classifier for event-event relations of the "OVERLAP" relation type, and used rules to infer "BEFORE" and "AFTER relation types. According to Cherry *et al.* [23], "OVERLAP" relations are reflexive, and are more likely to pay off in terms of recall. Lin *et al.* [74] used an SVM classifier for event-event pairs in consecutive sentences, an SVM classifier for event-time in consecutive sentences, and rules to classify conferenced events of the "OVERLAP" relation type. Xu *et al.* [145] used an SVM classifier for event-event relations, an SVM classifier for event-time relations, and an SVM classifier for time-time relations. Nikfarjam, Emadzadeh, and Gonzalez [96] and Cheng *et al.* [22] used a rule-based system to classify cross-sentence relations. Moharasan and Ho [93] used a naïve Bayes classifier for cross-sentence relations, and another naïve Bayes classifier for cross-section relations.

The best results Tang *et al.* [126], Cherry *et al.* [23], Lin *et al.* [74], D'Souza and Ng [30], D'Souza and Ng [31], and Xu *et al.* [145] used one classifier for event-event relations, and another classifier for event-time relations within sentences. All of them used an SVM classifier for event-event relations and an SVM for event-time relations, with the exception of Cherry *et al.* [23], who used ME classifiers instead. For cross-sentence relations, specialized classifiers also achieved superior results, as in Tang *et al.* [126], where SVM classifiers were used for relations between events in consecutive sentences and events that were co-referenced. Among the best results, Lin *et al.* [74] and Cherry *et al.* [23] used rules for inferring relations in addition to the ML-based approaches. D'Souza and Ng [30] and D'Souza and Ng [31] used approaches that were based on Tang's *et al.* [126] approach, but added high accuracy rules to complement the ML approach, and focused on feature engineering. D'Souza and Ng [31] added features based on discourses relations and semantic relations is the state-of-the-art for the i2b2 dataset.

Some studies that used the i2b2 2012 dataset re-annotated specific relations or focused on specific extraction tasks. Lee *et al.* [64] focused on within-sentence relations between event-times, and re-annotated the i2b2 2012 dataset to consider only direct relations. Direct relations are relations in which the temporal expressions modify the event (or otherwise), or in which the temporal expression and event are both arguments or adjuncts of the same predicate. Lee *et al.* [64] developed an SVM system using cost-sensitive learning and rules. The results that Lee at al. [64] achieved on the re-annotated dataset with their approach were better than the results of Tang *et al.* [126] in this re-annotated dataset.

D'Souza and Ng [117] supplied missing cross-sentence annotations in the i2b2 2012 dataset by defining ten eligibility criteria for annotating pairs in adjacent sentences, and used the same approach as in D'Souza and Ng [31] and D'Souza and Ng [30] with pruning heuristics. D'Souza and Ng [117] achieved higher F-measure with the expanded i2b2 2012 corpus in comparison with the original i2b2 2012 corpus, showing the improvement in classifier performance using the expanded corpus.

Most of our articles were related to the THYME corpus. From the articles detailed in sub-sections 5.2 and 5.3, a total of 29 were related to the THYME corpus.

Some studies, such as Miller *et al.* [90], Lin *et al.* [80], and Styler *et al.* [121] that worked with a small portion of THYME dataset were early publications in which datasets smaller than the THYME dataset were used during the shared task. Miller *et al.* [90] worked with within-sentence relations, and considered all possible pairs using an SVM approach with tress kernels to extract relations. Miller's *et al.* [90] best results were with a combination of flat features (FF), bag trees (BT), path-enclosed trees (PET), and path trees (PT). Lin's *et al.* [80] approach was based on Miller *et al.* [90], adding a descending path kernel (DPK) to the system. Styler *et al.* [121] worked with within-sentence relations, considering all possible pairs using the ClearTK-TimeML, which was based on an SVM classifier for event-event relations, and an SVM classifier for event-time relations, with small adaptations to fit the THYME corpus.

Regarding the Clinical TempEval 2015 dataset, there was a low number of participants owing to the long authorization process. Only two articles, those by Velupillai *et al.* [135] and Lin *et al.* [74], were related to the Clinical TempEval 2015 dataset. Velupillai *et al.* [135] considered pairs between mentions, limiting the range to three mentions from left to right if adjacent (neighbor) mentions were in separate sentences, and then merged the sentences. Velupillai's *et al.* [135] approach was based on rules, which were generated by the Moonstone system, and a CRF classifier. Lin *et al.* [74] restricted the pairs to all possible within-sentence pairs using an SVM classifier for event-event relations and an SVM classifier for event-time relations. Lin *et al.* [74] used cost-sensitive learning to counter dataset imbalance, and also added extra training data by utilizing a technique based on event expansion with the UMLS. Lin's *et al.* [74] approach achieved superior results restricting within-sentence relations, reducing the possible candidate pairs, and specific machine learning training for event-event and event-time relations also improved results (as seen in i2b2 2012 evaluation). Cost-sensitive learning, which was used by Lin *et al.* [74], was also effective for SVM-based approaches over the i2b2 2012 dataset, and training set expansion relieved the imbalance by adding additional positive training samples.

For the Clinical TempEval 2016 dataset, we have reviewed 20 articles that are detailed below in this sub-section (5.2) and the next sub-section (5.3). These publications are summarized according to three aspects: the candidate pair selection technique, the approach to extract within-sentence relations, and cross-sentence relations.

Candidate pair selection was crucial, as all events/times pairs can be candidate pairs, which raises the number of candidates relative to the small number of positive examples. Generating many more negative instances than positive ones, especially for cross-sentence relations, is not the best scenario for training a classifier [63].

For candidate pair selection of candidates in the same sentence (within-sentence), the most common strategy was to considerer all possible pairs between mentions, as was done by Khalifa, Velupillai, and Meystre [55], Caselli and Morante [15], Barros *et al.* [4], Tourille *et al.* [130], Tourille *et al.* [129], Lin *et al.* [77]. Lee *et al.* [63] also considered all possible pairs, but filtered unlikely candidates based on rules created by observing the THYME corpus annotations. The rules removed pairs if mentions were not in the same section, or if an event had a modality attribute of "factual" while the other was "hypothetical," or if an event had a DCT relation of "BEFORE" while the other was "AFTER." These cases will surely not have any relation because they could not be related in any way. Leeuwenberg and Moens [67] also considered all possible pairs, but added some restrictions based on new lines.

For candidate pairs in cross-sentence relations, the most common approach was to restrict the approach to within-sentence relations, discarding all possible cross-sentence pairs. As observed by Tourille *et al.* [129] close to 76% of positive sample pairs occurred within the same sentences in the training set, and there was a trade-off between covering more pairs to consider cross-sentence relations and drastically raising the negative sample proportion, as more sentences were considered when creating candidate pairs. The authors that focused only within-sentence relations were Caselli and Morante [15], Tourille *et al.* [130] and Lin *et al.* [77].

Several authors considered cross-sentence relations by creating heuristics to generate candidate pairs. Some authors restricted the number of sentences within the mentions, Barros *et al.* [4] restricted pairs to adjacent (neighbor) sentences, Tourille *et al.* [129] considered all possible pairs in a three-sentence window (containing approximately 89% of the positive examples). In addition to restricting the sentence window, Lee *et al.* [63] added rules, and considered the same approach used for within-sentence relations (detailed above) for pairs that were two sentences apart, but with additional rules to omit pairs that were more than two sentences apart.

Khalifa, Velupillai, and Meystre [55] created different heuristics for event-event and event-time pairs. For event-event pairs across consecutive (neighbor) sentences, the first and last events from the current sentence were paired with the first and last from both neighbor sentences (left and right). For event-time across consecutive sentences, each temporal expression was paired to the first and last events of both neighbor sentences (left and right). Leeuwenberg and Moens [67] added candidate pairs that were in the same line, adding heuristics based on comma and colon symbols to consider more pairs.

Leeuwenberg and Moens [70] developed a sentence agnostic approach for candidate pair selection, restricting candidate pairs to pairs of mentions that occurred in a window of 30 tokens and the mentions should occur in the same paragraph.

An approach that was popular to use for dataset imbalance was to transform the two-class classification task ("CONTAINS" and "NO_RELATION") into a three-class classification task ("CONTAINS," "NO_RELATION," and "IS_CONTAINED) by adding an "IS_CONTAINED" class to indicate that a mention is contained in another. All possible pairs were then generated from left to right, and whenever was needed, "CONTAINS" was changed to "IS_CONTAINED." This approach was used by Tourille *et al.* [129] and Tourille *et al.* [130]. The advantage of this method is that it cuts in half the number of candidate pairs. This method was widely used in deep learning based models (sub-section 5.3).

Some studies attempted to classify both within-sentence and cross-sentence relations by using specialized approaches for these relations. For within-sentence relations, Khalifa, Velupillai, and Meystre [55] and Lee *et al.* [63] used an SVM classifier for event-event relations, and an SVM classifier for event-time relations. Khalifa, Velupillai, and Meystre [55] also used an SVM classifier for event-event relations, and an SVM classifier for event-time relations across sentences. Lee *et al.* [63] added more specific classifiers for cross-sentence relations, with an SVM classifier for event-event relations and an SVM classifier for event-time relations that considered two adjacent sentences, an SVM classifier for event-time relations and an SVM classifier for event-event relations that considered more than two adjacent sentences. Lee *et al.* [63] also used cost-sensitive learning in addition to the SVM to reduce the effect of dataset imbalance. The use of specialized classifiers in relations across more than two adjacent sentences resulted in the best performance in Lee's *et al.* [63] framework, mainly because of the candidate pair selection that kept only pairs that were likely to contain a relation [63]. Tourille *et al.* [129] used an SVM classifier for within-sentence relations, and an SVM classifier for cross-sentence relations, but also added rules to extract certain relations.

Regarding the approaches that dealt with within-sentence relations only (and did not focus on cross-sentence relation), Caselli and Morante [15] used a CRF classifier for event-event relations, and a CRF classifier for event-time relations. Tourille *et al.* [130] used SVM classifiers to perform within-sentence relation extraction they trained an SVM classifier to extract relations from the THYME corpus (TempEval 2016 dataset) and an SVM classifier to extract relations from MERLOT (French annotated corpus), and compared relation extraction methods in different languages scenarios. Lin *et al.* [77] used an SVM classifier for event-event relations and an SVM classifier for event-time relations. They focused on training set expansion with a framework based on UMLS, using different expansion criteria for each relation (event-event, event-time).

Several approaches used the same framework to classify both within-sentence and cross-sentence relations jointly. Chikka [24] used a CRF-based approach, Barros *et al.* [4] also used a CRF-based approach with a CRF classifier for event-event relations, a CRF classifier for event-time relations, a CRF classifier for time-time relations, and a CRF classifier for time-event relations. Barros *et al.* [4] also used rules. The approach used by Leeuwenberg and Moens [67] was based on cTAKES temporal system [74] with additional features, and used an SVM classifier for event-event relations, and an SVM classifier for event-time relations.

Leeuwenberg and Moens [70] classified used both within-sentence and cross-sentence relations jointly, with structured perceptron model that jointly predicted TLINKs and DCT relations, and used sub-sampled negative examples during training to achieve faster training and lower the impact of imbalance.

Among the machine learning-based approaches Leeuwenberg and Moens [70] achieved the best results with the structured machine learning approach based on structured perceptron and ILP. Followed by Lee *et al.* [63] which was based on several specialized classifiers, heuristics to diminish the possible candidate pairs, and cost-sensitive learning. These approaches used by Lee *et al.* [63] have been successful for machine learning approaches, as summarized over our i2b2 2012 dataset analysis.

For the Clinical TempEval 2017 dataset, we have reviewed seven articles that are detailed over this sub-section (5.2) and the next sub-section (5.3). These publications are summarized according to three aspects: the candidate pair selection technique, the approach to extract within-sentence relations, and cross-sentence relations.

For candidate pair selection from within-sentence relations, the most common approach is to consider all possible pairs, as used by Sarath, Manikandan, and Niwa [111], MacAveney, Cohan, and Goharian [87] and Huang *et al.* [47].

For candidate pair selection considering cross-sentence relations, some authors restricted the candidate pairs to within-sentence pairs and discarded all cross-sentence possible pairs. The authors that used this approach were MacAveney, Cohan, and Goharian [87] and Huang *et al.* [47]. Sarath, Manikandan, and Niwa [111] considered all possible pairs over a two-sentence window.

The approach used by Leeuwenberg and Moens [68] was based on Leeuwenberg and Moens [70], in which candidate pairs involved mentions that occurred within a window of 30 tokens, and occurred in the same paragraph.
Some additional approaches that reduce the dataset imbalance impact over the classifiers were used. Sarath, Manikandan, and Niwa [111] filtered unlikely candidate pairs as Lee *et al.* [63] by using heuristic rules based on Khalifa, Velupillai, and Meystre [55], and also using cost-sensitive learning to reduce the imbalance impact over the classification.
Regarding the approaches used to extract the within-sentence relations, some authors used machine learning-based frameworks. Sarath, Manikandan, and Niwa [111] used an ensemble of classifiers for event-event relations, and an ensemble of classifiers for event-time relations. The ensemble of classifiers involved gradient boosted trees, XGBoost, extra trees, and RF classifiers. MacAveney, Cohan, and Goharian [87] used an XGboost classifier. Huang *et al.* [47] used an SVM classifier for event-event relations, and an SVM classifier for event-time relations.
Regarding cross-sentence relations, MacAveney, Cohan, and Goharian [87] and Huang *et al.* [47] focused only on within-sentence relations. Sarath, Manikandan, and Niwa [111] used an ensemble of classifiers for event-event relations, and an ensemble of classifiers for event-time relations.
Leeuwenberg and Moens [68] classified both within-sentence and cross-sentence relations jointly, with structured perceptron model based on the previous publication of Leeuwenberg and Moens [70]. To deal with the cross-domain adaptation Leeuwenberg and Moens [68] used constraints on the output labeling formulated by Leeuwenberg and Moens [70] and randomly replaced tokens of events with an unknown token word (UNK). For phase 2, Leeuwenberg and Moens [68] added weighting to the target domain data (brain cancer).
The best results over the Clinical TempEval 2017 dataset were based on deep learning. MacAveney, Cohan, and Goharian [87] and Leeuwenberg and Moens [68] achieved the best results over machine learning-based approaches.

TLINKS DEEP LEARNING

The systems that extracted TLINKs using deep learning-based systems are detailed in Table 6. The table contains information about the article, its primary objective, the strategy used to extract temporality, and the article result, which is related to the "sep TLINK" (separated TLINK relation evaluation). If the article had a separate evaluation, the result is regarding the temporal extraction; if not, the result is regarding the system primary objective.

Table 6: Articles related to TLINK extraction with Deep learning.

Authors	Objective	Strategy	Candidate pair selection	Result	Sep TLINK
Liu *et al.* [84]	Temporal indexing	BI-LSTM-CNN	Rules	Fm 0.7597	Yes
Wang *et al.* [142]	i2b2 2012 TRE (EE "OVERLAP" relations)	BI-LSTM + TS expansion	-	Fm 0.6217 (NTC)	No
Cohan, Meurer, and Goharian [27]	CTE 2016 TRE	WS: SRL+ dependency parse tree	WS: APP	Fm 0.506	Yes
Tourille *et al.* [128]	CTE 2016 TRE	WS: BI-LSTM, CS: BI-LSTM; 3-class	WS: APP, CS: rules	Fm 0.613	Yes
Dligach *et al.* [34]	CTE 2016 TRE	WS: 2 CNNs; 3-class; XML markup	WS: APP	Fm 0.515 EE, 0.700 ET (NTC)	Yes
Lin *et al.* [78]	CTE 2016 TRE	WS: SVM clf + CNN; XML markup	WS: APP	Fm 0.621	Yes
Liu *et al.* [81]	CTE 2016 TRE	WS: GRU + att; 3-class; XML markup	WS: APP	Fm 0.690 (NTC)	Yes
Galvan *et al.* [38]	**CTE 2016 TRE**	**WS: tree-based BI-LSTM-RNN**	**WS: APP**	**Fm 0.629**	**Yes**
Lin *et al.* [76]	**CTE 2016 + CTE 2017 TRE**	**WS: BI-LSTM; TS expansion; 3-class. XML markup**	**WS: APP**	**Fm 0.630 (CTE 2016), 0.547 (CTE 2017 UDA)**	Yes

Authors	Objective	Strategy	Candidate pair selection	Result	Sep TLINK
Leeuwenberg and Moens [69]	**CTE 2016 TRE**	**MTL + LSTM**	**APP over TK**	**Fm 0.628**	**Yes**
Lin *et al.* [79]	**CTE 2016 + CTE 2017 TRE**	**BERT; TS expansion; 3-class**	**APP over TK**	**Fm 0.684 (CTE 2016), 0.565 (CTE 2017 UDA)**	**Yes**
Jeblee and Hirst [50]	CTE 2016 TRE (Ranking)	List-Net (Cao *et al.* [12])	-	MSE 0.072 (NTC)	Yes
Tourille *et al.* [131]	CTE 2017 TRE	WS: BI-LSTM. CS: BI-LTM; 3-class	WS: APP, CS: rules	Fm 0.328 UDA, 0.316 SDA	Yes

Legend: GRU: gated recurrent unit, CNN: convolutional neural network, RNN: recurrent neural network, LSTM: long short-term memory, WS: within-sentence relation, CS: cross-sentence relation, APP: all possible pairs, TS: training set, SRL: semantic role labeling, 3-class: transforming to a 3-class classification task, CTE: Clinical TempEval, clf: classifier, Fm: F-measure, TRE: temporal relation extraction, NTC: not comparable, UDA: unsupervised domain adaptation, SDA: supervised domain adaptation, EE: event-event, ET: event-time, att: attention mechanism, LNN: linear neural network, TK: token window, MTL: multi-task learning, MSE: mean squared error.

Among the deep learning-based approaches, Liu *et al.* [84] worked with Chinese clinical notes to extract temporal indexing by extracting temporal relations and selecting only one pair for each event mention. Candidate pairs were an event, and temporal expression pairs, used rules to restrict the temporal expressions that were linked to each event. The relations ("AFTER," "BEFORE," "SIMULTANEOUS," and "NONE") were predicted using a model based on both LSTM (BI-LSTM) and CNNs. Compared with the SVM, CNN, and LSTM models, the hybrid model (RNN-CNN) yielded better results.

Regarding the i2b2 2012 dataset, only one of the 17 publications used deep learning. Recent publications about the i2b2 2012 dataset focused on refining the annotations and then evaluate the refinement with machine learning approaches. The only deep learning-based approach was proposed by Wang *et al.* [142] to deal with event-event relations of the type "OVERLAP" considering a binary classification problem: "OVERLAP" or "NONE." Wang [142] generated additional training data using a transformer model and tested it on the i2b2 2012 dataset. Wang *et al.* [142] used a BI-LSTM model, inspired by Tourille *et al.* [128] model, and achieved better results with the augmented/generated data than with the original dataset (no augmentation) or the duplicated original dataset (all samples were duplicated).

Regarding the Clinical TempEval 2016 dataset, 10 publications were deep learning-based. These publications are summarized below according to three aspects: the candidate pair selection technique, the approach to extract within-sentence relations, and cross-sentence relations.

For candidate pair selection of candidates in the same sentence (within-sentence), the most common strategy was to considerer all possible pairs between mentions as was done by Tourille *et al.* [128], Dligach *et al.* [34], Lin *et al.* [78], Liu *et al.* [81], Galvan *et al.* [38], Lin *et al.* [76] and Cohan, Meurer, and Goharian [27].

As in the previous sub-section, some publications focused only on within-sentence relations, like Dligach *et al.* [34], Lin *et al.* [78], Liu *et al.* [81], Galvan *et al.* [38], Lin *et al.* [76] and Cohan, Meurer, and Goharian [27]. Tourille *et al.* [128] considered cross-sentence relations but considered all possible pairs in a three-sentence window (containing approximately 89% of the positive examples).

Some authors considered sentence agonistic approaches for candidate pair selection. Leeuwenberg and Leeuwenberg and Moens [69] and Lin *et al.* [79] restricted candidate pairs to pairs of mentions that occurred in certain token window, Leeuwenberg and Moens [69] considered a 30 token window and Lin *et al.* [79] considered a 60 token window. Leeuwenberg and Moens [69] strategy of considering a 30 token window gave a positive-to-negative sample ratio of 1:36.

From the strategies used for candidate pair selection, we concluded that some of the best results over the dataset, represented by Lin *et al.* [76], Galvan *et al.* [38], Liu *et al.* [81], Lin *et al.* [78], and Dligach *et al.* [34], restricted relations to within-sentence relations, considered all possible pairs within a sentence, and discarded cross-sentence relations. Among the top four results over the dataset (from studies that were comparable because they used the same evaluation metrics), Lin *et al.* [76] and Galvan *et al.* [38] used this approach. Tourille *et al.* [128] achieved competitive results by considering all within-sentence relations and restricting cross-sentence according to the sentence window. Sentence agnostic approaches based on token windows also achieved some of the best results. This approach was used by Lin *et al.* [79] and Leeuwenberg and Moens [69]. Among the top four results over the dataset, Leeuwenberg and Moens [69] and Lin *et al.* [79] used this approach. Lin *et al.* [79] have the best results over this dataset with this approach.

It was noticeable that several studies that achieved comparative results over the dataset transformed the two-class classification task into a three-class classification task (approach detailed in sub-section 5.2), cut in half the

number of candidate pairs. This transformation was used by Lin *et al.* [79], Liu *et al.* [81], Lin *et al.* [76], Dligach *et al.* [34], and Tourille *et al.* [128]. Among the top four results over the dataset Lin *et al.* [79] and Lin *et al.* [76] used this approach. Lin *et al.* [79] have the best results over this dataset with this approach.

Some studies attempted to classify both within-sentence and cross-sentence relations by using specialized approaches for these relations. Tourille *et al.* [128] used a BI-LSTM model for within-sentence relations and a BI-LSTM model for cross-sentence relations, with character embeddings based on Lample's *et al.* [58] BI-LSTM model and word embeddings. They used one embedding per mention attribute and one embedding per cTAKES attribute. According to Tourille *et al.* [128], both mention and cTAKES features improved the results.

Regarding the approaches that dealt with within-sentence relations only (and did not focus on cross-sentence relation), Cohan, Meurer, and Goharian [27] used an approach with semantic role labeling (SRL) based on Collobert *et al.* [28] and a dependency parsing tree. Dligach *et al.* [34] used a CNN model for event-event relations, and a CNN model for event-time relations, and augmented the token sequence with XML markup for relation arguments. Lin *et al.* [78] used an SVM for event-event relations based on the THYME system [74, 77], and a CNN model for event-time relations, using the same XML markup for the CNN as Dligach *et al.* [34], but representing temporal expressions with a single token of the class attribute. Liu *et al.* [81] used a GRU model with an attention mechanism on top, and the XML markup based on that of Dligach *et al.* [34] for events, and XML markup based on that of Lin *et al.* [78] for temporal expressions. Galvan *et al.* [38] adapted the tree-based BI-LSTM-RNN model by Miwa and Bansal [92], making new sentence-level annotations to adapt the input. Lin *et al.* [76] used a BI-LSTM model similar to that used by Tourille *et al.* [128] but expanded the training set with self-training, creating "silver instances" using a high-precision BI-LSTM model. Lin *et al.* [76] also used XML markup based on that used by Dligach *et al.* [34] for events and used XML markup based on that used by Lin *et al.* [78] for temporal expressions.

Several approaches used the same framework to classify both within-sentence and cross-sentence relations jointly. The approach by Leeuwenberg and Moens [69] was based on multi-task learning (MTL), and trained word representations jointly on relation extraction and context prediction by sharing weights and using an LSTM model to classify the relation. For context prediction using a continuous skip-gram (SG) model trained over the MIMIC III and THYME corpora, the model was trained on a combination of the loss functions for both tasks. Lin *et al.* [79] used bidirectional encoder representations from transformers (BERT) [33] for relation extraction of both within-sentence relations and cross-sentence relations, expanding the training set with the addition of "silver annotations" with high confidence in the dataset. The training set expansion was based on Lin *et al.* [76], but used BERT to generate the "silver annotations." Lin *et al.* [79] used BioBERT [65], a pre-trained BERT model trained over biomedical corpora.

All developed systems with good results over the Clinical TempEval 2016 were based on deep learning. The only exception was the system by Lin *et al.* [78], which used an SVM-based THYME system for event-event relations. LSTM-based models were used by Leeuwenberg and Moens [69] in multi-task learning (MTL), and LSTM was used to classify relations by Tourille *et al.* [128], who used a BI-LSTM-based model, and by Lin *et al.* [76] who used a BI-LSTM model that relied on self-training. CNN-based models were used by Lin *et al.* [78] and Dligach *et al.* [34] with XML relation argument augmentation. The strategy of using XML relation argument augmentation was also used by Liu *et al.* [81] and Lin *et al.* [76]. Galvan *et al.* [38] tree-based BI-LSTM-RNN model based on Miwa and Bansal [92] also achieved good results over the dataset. Liu *et al.* [81] used a GRU model with an attention mechanism, and Lin *et al.* [79] also used an attention mechanism along with BERT. Lin *et al.* [79] used an approach based on BERT by considering BERT with no data augmentation, and achieved an F-measure of 0.669 in comparison with existing state-of-the-art approaches used in the studies of Lin *et al.* [76], Galvan *et al.* [38], and Leeuwenberg and Moens [69], whose F-measure range from 0.628 to 0.630. Lin *et al.* [79] achieved an F-measure of 0.684 (more than 5% better than existing state-of-the-art approaches) with a transformer base model along with data augmentation, data augmentation strategy similar to that of Lin *et al.* [76]. These results show the power of the transformer base models.

Regarding the Clinical TempEval 2016 dataset, the Jeblee and Hirst [50] study was not directly comparable because of its dataset adaption. Jeblee and Hirst [50] adapted the THYME corpus to consider only "BEFORE" and "OVERLAP" relations by merging the other types. They attempted to do temporal ranking because temporal pairwise classification exhibits a high number of possible pairs with "NO RELATION," which worsens as the pairwise token window increases [50]. The temporal ranking model was the List-Net from Cao *et al.* [12], which is based on a linear neural network.

Regarding the Clinical TempEval 2017 dataset, three publications are deep learning-based. As the publications of Lin *et al.* [76] and Lin *et al.* [79] used both Clinical TempEval 2016 and Clinical TempEval 2017 they are

already detailed above over the Clinical TempEval 2016 dataset, so we only highlight the strategies used to deal with cross-domain. Similarly, the framework proposed by Tourille *et al.* [131] is based on Tourille *et al.* [128] framework developed over Clinical TempEval 2016 dataset and thus below we only highlight the strategies used to deal with cross-domain.

Regarding approaches to classify cross-domain relations, Tourille *et al.* [131] tested both blocking the training of pre-trained word embedding during network training, and randomly replacing tokens of events with an UNK. They achieved slightly better results (0.01 higher F-measure) by blocking further training of word embeddings than by using the UNK strategy. In within-sentence relations, they also added a token to identify sentence breaks. Lin *et al.* [76] also used the UNK strategy.

The best results for the Clinical TempEval 2017 dataset were based on deep learning, except for MacAveney, Cohan, and Goharian [87], who used an XGboost classifier.

Based on the best results, we conclude that adding "silver instances" using unlabeled data, as done by Lin *et al.* [76] and Lin *et al.* [79], improved the results, as it provided more data for training while reducing the imbalance. Models based on BI-LSTM, such as those used by Tourille *et al.* [131] and Lin *et al.* [76], captured both previous and future context by processing data in both directions with separate RNNs [40], which improved the results. The BERT model is a multi-layer bidirectional transformer encoder, which uses bidirectional pre-training of language representations by pre-training representations over the unbalanced text, jointly leveraging "directional context" (both left and right context) in all layers [33]. The best results for this dataset were achieved by Lin *et al.* [79], followed by Lin *et al.* [76].

TLINK CONCLUSIONS

Identifying TLINKs consists of two steps. First, a relation between a pair of mentions (events and temporal expression mentions) is identified, followed by the second step of extracting the relation type (e.g., "BEFORE," stating that a mention occurred before the other).

Several studies dealt only with the TLINK identification, correlating events to temporal expressions by rules (detailed in sub-section 5.1). However, for both steps (pair identification and relation extraction) machine learning systems, hybrid systems (machine learning and rule-based) and deep learning systems were used.

Most of the articles related to TLINK extraction were based on shared-task datasets. This was primarily owing to the difficulty of developing a temporal relation-annotated corpus. For TLINK extraction events, the temporal expressions need to be previously annotated, which adds two more annotations processes composed of rounds of guideline refinement, annotator training, text annotation, and adjudication. Furthermore, temporal relation annotation is a complex annotation task, as is shown by the lower inter-annotator agreement in comparison with the event and temporal expression inter-annotator agreements over the i2b2 2012 and Clinical TempEval shared-tasks.

For the i2b2 2012 dataset, the best results over the dataset are still machine learning-based, mostly due to the lack of recent publications aiming to achieve state of the art over the dataset. Best approaches are represented by Tang *et al.* [126], D'Souza and Ng [30], D'Souza and Ng [31], Cherry *et al.* [23], Xu *et al.* [145], and Lin *et al.* [74]. Besides machine learning, most of them also relied on rules. Achieving best results over the dataset involved creating several specialized classifiers for within-sentence relations (e.g., event-event and event-time pairs), and cross-sentence relations. For cross-sentence relations, superior results were achieved by using heuristics to reduce the number of possible pairs to achieve the best overall result. The best result over i2b2 2012 belonged to D'Souza and Ng [31], based on Tang *et al.* [126], with additional high-accuracy rules to complement the ML approach and also a focus on feature engineering.

In the past, the best results over the THYME corpus were obtained using specific traditional machine learning classifiers for both within-sentence and cross-sentence relations and differentiated between event-event pairs and event-time pairs, as in Lin *et al.* [74], Lee *et al.* [63], and Lin *et al.* [77]. Lin *et al.* [74] and Lee *et al.* [63] also used cost-sensitive learning to reduce the dataset imbalance.

Over time, neural network-based models came into use and began to achieve better results than traditional machine learning-based approaches. Approaches based on CNNs with Lin *et al.* [78] and Dligach *et al.* [34]. BI-LSTM with Tourille *et al.* [128], Lin *et al.* [76], and Galvan *et al.* [38], who used a BI-LSTM-RNN model. Leeuwenberg and Moens [69] with LSTM together with MTL. Attention-based models with Liu *et al.* [81], GRU model with attention mechanism, and with Lin *et al.* [79], who used attention with BERT.

Lin's *et al.* [79] approach based on BERT with no data augmentation achieved a higher F-measure in comparison with existing state-of-the-art approaches used in the studies of Lin *et al.* [76], Galvan *et al.* [38], and Leeuwenberg and Moens [69] over the Clinical TempEval 2016 dataset. This showed the power of transformer base models. Lin *et al.* [79] achieved an F-measure of more than 5% higher when compared with the existing state-of-the-art approaches with self-training based data augmentation.
Regarding the Clinical TempEval 2017 dataset, the best results over the dataset were from Lin *et al.* [76] and Lin *et al.* [79]. Models such as BI-LSTM and BERT, which learn context bidirectionally, obtained the best results for cross-domain relations. Moreover, adding "silver instances" with self-training improved the results. As neural network methods rely on a massive amount of labeled data, adding "silver instances," even if some "silver instances" were inaccurate, tended to improve performance. In both Clinical TempEval 2016 and Clinical TempEval 2017 datasets, the best results were obtained by Lin *et al.* [79] with the BERT-based model.
The BERT model was pre-trained over two unsupervised tasks, and the self-attention mechanism in the transformer allowed it to be fine-tuned to several PLN tasks [33]. BioBERT is a BERT-based model for the clinical domain that was based on pre-training over biomedical corpora developed by Lee *et al.* [65]. Fine-tuning BioBERT to relation extraction by considering tokens in a certain window, transforming the two-class task into a three-class task, and augmenting instances of self-training were factors that contributed to Lin *et al.* [79] achieving the state-of-the-art over Clinical TempEval 2016 and 2017 datasets.
In the next section, we present shared-task datasets and approaches for temporal relation extraction in the general domain, describe the relevant publications, and make comparisons between the approaches used.

TEMPORAL RELATION EXTRACTION IN GENERAL DOMAIN

The 2013 TempEval shared-task was related to the TempEval-3 corpus (TE-3), which was based on the AQUAINT and TimeBank [101] corpora, with two temporal relation extraction tasks. One of them was Task C, for which the system should identify pairs and classify the relations given gold standard events and temporal expressions. The other was Task C (relation only), for which the system should only classify relations among identified pairs. For both tasks, the relations were based on the TimeML relations described in Section 2. According to Cassidy *et al.* [16], in TimeBank the annotators only labeled relations that were key to understanding the document, which resulted in sparse annotations. Cassidy *et al.* [16] annotated the TimeBank-Dense, which increased the number of annotations and considered additional relation categories (e.g., relation between temporal expressions and DCT), and simplified the relation types. According to Cassidy *et al.* [16], making decisions regarding relations like "BEFORE" and "I_BEFORE" (immediately before) can complicate an already-difficult annotation. TimeBank-Dense also added a "VAGUE" relation for pairs where no clear temporal relation could be inferred.
The ratio of relations between events and temporal expressions is 0.7 for TimeBank, 0.8 for TE-3, and 6.3 for TimeBank-Dense, showing the increase in the number of labeled relations per document. Both TE-3 and TimeBank-Dense are used for temporal relations over general domain texts.
Regarding the TE-3, we focused on ClearTK-TimeML, which had the best results over the TempEval 2013 shared-task and is used as a baseline for comparison with newly developed systems. ClearTK-TimeML is based on several SVM classifiers, and was applied to the clinical domain by Styler *et al.* [121] using the THYME corpus (see sub-section 5.2). For the TE-3, we also focused on Laokulrat *et al.* [60] with UTTime, which is a hybrid approach based on rules and RL classifiers that benefit from inverse relations during training. For TimeBank-Dense, we focused on CAEVO [17], a sieve-based approach that proposed smaller specialized classifiers while leveraging rules. CAEVO is typically used as a baseline for TimeBank-Dense.
Vo and Bagheri [140] used an Open Information Extraction (OpenIE) strategy to generate an event network graph, and then used rules (for direct relations) and event flow (for indirect relations) to infer the relation. Laokulrat, Miwa, and Tsuruoka [59] also used graphs, using stacked learning with LR classifiers and time graphs.
Ning, Feng, and Roth [97] used a structured perceptron model and ILP to extract relations, which was similar to the structured learning approach by Leeuwenberg and Moens [70], which is detailed in sub-sections 5.2. Cheng and Miyao [21] used a BI-LSTM and dependency paths, and processing data in both directions with RNNs also improved the results, similar to the results for clinical texts (mentioned in subsection 5.4).
Some of the best results over the TE-3 and TimeBank-Dense corpora were achieved by frameworks that predicted temporal relations in conjunction with another classification problem, Mirza and Tonelli [91] and Ning *et al.* [98] extracted causal relations, and Vashishtha, Van Durme, and White [134] extracted event duration. As

mentioned by Ning *et al.* [98] temporal relations and causal relations interact, and decisions are based on evidence from both. Mirza and Tonelli [91] used a sieve-based architecture based on CAEVO to predict causal and temporal relations using SVM classifiers and rules. Ning *et al.* [98] jointly predicted causal and temporal relations with ILP and constrained conditional models. Vashishtha, Van Durme, and White [134] used ELMo [99] contextual embeddings, a tuner (to reduce the dimension of ELMo), and attention mechanisms to jointly predict event duration and temporal relations with an MLP on a generated dataset. Afterward, Vashishtha, Van Durme, and White [134] used transfer learning to extract temporal relations from TE-3 and TimeBank-Dense datasets. According to Vashishtha, Van Durme, and White [134], this approach placed more importance on event duration during prediction, while improving the understanding of the temporality regarding complex events. Both ELMo and BERT (best results over THYME dataset) are state-of-the-art language representation models based on bidirectionality. ELMo is a BI-LSTM language model (ML) with character convolutions [99], and BERT is a multi-layered bidirectional transformer encoder [33]. A recent review of word representations for the clinical domain can be found in Khattak *et al.* [56].

CONCLUSION

Clinical texts suffer from writing and formatting issues, as pointed out in the introduction, which makes them noisy texts that directly impact information extraction performance. Noisy texts impact rule-based systems performance, and it also affects ML and deep learning approaches performance, as patterns become harder to extract by the systems.

As ML approaches, especially deep learning approaches, depend on the amount of available data, data scarcity can directly impact performance in studies where there is insufficient annotated data or where there is unbalanced data. Depending on the language scope, data availability can vary.

Regarding the approaches to DCT relations and TLINKs, the extraction of DCT relations have been solved, and the achieved results exceeded the inter-annotator agreement for the Clinical TempEval 2016 dataset. DCT relations are easier to extract than TLINKs because they do not suffer from dataset imbalance during training owing to candidate pair generation. Typically, an event is connected to a DCT, or in the worst cases, as in i2 2012, an event can be connected to two DCTs (admission or discharge). In TLINKs, an event can be connected to every other event or temporal expression in a sentence (within-sentence relation), or in different sentences (cross-sentence relation), according to a certain window, either based on a sentence window or a token window, which creates a severe imbalance over the dataset during training.

For DCT relations, using either an SVM classifier, as in Tang *et al.* [126], Cherry *et al.* [23], Tourille *et al.* [130], Lin *et al.* [74], Tourille *et al.* [131], and Viani *et al.* [139], or a CRF classifier, as in D'Souza and Ng [30], Velupillai *et al.* [135], Khalifa, Velupillai, and Meystre [55], and Sarath, Manikandan, and Niwa [111] are sufficient for most problems. The major impacting factor is the feature engineering.

For TLINKs, the best approaches are based on deep learning, with frameworks based on: BI-LSTM, as in Tourille *et al.* [128] and Lin *et al.* [76], BI-LSTM-RNN as in Galvan [38], multi-task learning in conjunction with LSTM for prediction, as in Leeuwenberg and Moens [69], and self-attention, as in Liu *et al.* [81] and Lin *et al.* [79]. Training set expansion, either by UMLS, as in Lin *et al.* [74] or by self-training, as in Lin *et al.* [76] and Lin *et al.* [79] improved the results. Transforming the problem into a three-class problem, as in Tourille *et al.* [128], Lin *et al.* [76], Lin *et al.* [79], and Liu *et al.* [81] diminished the dataset imbalance during training.

The best results belonged to Lin *et al.* [79], who applied BERT to relation extraction, utilizing the BioBERT pre-trained model. The problem with BERT-based models is the need for a large corpus, which can be difficult for languages other than English, and the amount of time and resources required for pre-training. BioBERT was pre-trained for 23 days with eight NVIDIA V100 GPUs [65], which impacts its applicability to languages other than English.

This review addressed temporal relation extraction in clinical texts by describing the approaches and highlighting common characteristics among the best results. We verified that the best results for DCT relations relied on traditional ML approaches, with the use of SVM and CRF classifiers combined with feature engineering. However, the TLINKs best results relied on deep learning approaches with heuristics to diminish candidate pairs. Training set expansion also improved the results.

Although we achieved our review purposes, we did not discuss which areas within the clinical domain are directly affected by temporal relation extraction, and how improving the temporal relation extraction framework

results could benefit these areas in the future. This review also lacks perspective on possible applications in the clinical domain. These topics could be addressed in future studies regarding temporal relation extraction.

ACKNOWLEDGMENTS

This work was supported by the Brazilian Government Agency Coordination for the Improvement of Higher Education Personnel (CAPES).

REFERENCES

[1] Zubair Afzal, Ewoud Pons, Ning Kang, Miriam C.J.M. Sturkenboom, Martijn J. Schuemie, and Jan A. Kors. 2014. ContextD: An algorithm to identify contextual properties of medical terms in a dutch clinical corpus. *BMC Bioinformatics* 15, 1 (2014), 1–12. DOI: https://doi.org/10.1186/s12859-014-0373-3

[2] James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26, 11 (1983), 361–372. DOI: https://doi.org/10.1145/182.358434

[3] Rafael F. de Azevedo, Joao P. Santos Rodrigues, Mayara R. da Silva Reis, Claudia M. C. Moro, and Emerson Cabrera Paraiso. 2018. Temporal tagging of noisy clinical texts in Brazilian Portuguese. In *Proc. of the 13th International Conference on Computational Processing of the Portuguese Language*, 231–241. https://doi.org/10.1007/978-3-319-99722-3

[4] Marcia Barros, Andre Lamurias, Gonçalo Figueiro, Marta Antunes, Joana Teixeira, Alexandre Pinheiro, and Francisco M. Couto. 2016. ULISBOA at SemEval-2016 Task 12: Extraction of temporal expressions, clinical events and relations using IBEnt. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1263–1267. https://doi.org/10.18653/v1/S16-1196

[5] Cosmin A. Bejan, Lucy Vanderwende, Heather L. Evans, Mark M. Wurfel, and Meliha Yetisgen-Yildiz. 2013. On-time clinical phenotype prediction based on narrative reports. In *AMIA Annu. Symp. Proc. 2013*. 103–110.

[6] Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *2nd Joint Conf. on Lexical and Computational Semantics (* SEM), Volume 2: Proc. of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*. 10–14.

[7] Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: clinical TempEval. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*. 806–814. https://doi.org/10.18653/v1/s15-2136

[8] Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: clinical TempEval. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1052–1062. https://doi.org/10.18653/v1/S16-1165

[9] Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: clinical TempEval. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 565–572. https://doi.org/10.18653/v1/S17-2093

[10] Philip Bramsen, Pawan Deshpande, Yoong K. Lee, and Regina Barzilay. 2006. Finding temporal order in discharge summaries. In *AMIA Annu. Symp. Proc. 2006*. 81–85.

[11] Philip Bramsen, Pawan Deshpande, Yoong K. Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*. 189-198. https://doi.org/10.3115/1610075.1610105

[12] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proc. of the 24th International Conference on Machine Learning*. 129–136. https://doi.org/10.1145/1273496.1273513

[13] Daniel Capurro, Mario Barbe, Claudio Daza, Josefa Santa María, Javier Trincado, and Ignacio Gomez. 2015. ClinicalTime: Identification of patients with acute kidney injury using temporal abstractions and temporal pattern matching. In *AMIA Summits on Translational Science Proc. 2015*. 46–50.

[14] Daniel Capurro, Meliha Yetisgen, Erik van Eaton, Robert Black, and Peter Tarczy-Hornoch. 2014. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: A multi-site assessment. *EGEMs* 2, 1 (2014), 7–11. DOI: https://doi.org/10.13063/2327-9214.1079

[15] Tommaso Caselli and Roser Morante. 2016. VUACLTL at SemEval 2016 task 12: A CRF pipeline to clinical TempEval. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1241–1247. https://doi.org/10.18653/v1/S16-1193

[16] Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 501–506. https://doi.org/10.3115/v1/p14-2082

[17] Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics* 2, (2014), 273–284. DOI: https://doi.org/10.1162/tacl_a_00182

[18] Nai Wen Chang, Hong Jie Dai, Jitendra Jonnagaddala, Chih-Wei Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2015. A context-aware approach for progression tracking of medical concepts in electronic medical records. *Journal of Biomedical Informatics* 58 (2015). S150-2157. DOI: https://doi.org/10.1016/j.jbi.2015.09.013

[19] Yung-Chun Chang, Hong-Jie Dai, Johnny Chi-Yang Wu, Jian-Ming Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2013. TEMPTING system: A hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of Biomedical Informatics* 46, (2013), S54–S62. DOI: https://doi.org/10.1016/j.jbi.2013.09.007

[20] Qingcai Chen, Haodi Li, Buzhou Tang, Xiaolong Wang, Xin Liu, Zengjian Liu, Shu Liu, Weida Wang, Qiwen Deng, Suisong Zhu, Yangxin Chen, and Jingfeng Wang. 2015. An automatic system to identify heart disease risk factors in clinical text over time. *Journal of Biomedical Informatics* 58, (2015), S 158-63. DOI: https://doi.org/10.1016/j.jbi.2015.09.002

[21] Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*. 1–6. https://doi.org/10.18653/v1/P17-2001

[22] Yao Cheng, Peter Anick, Pengyu Hong, and Nianwen Xue. 2013. Temporal relation discovery between events and temporal expressions identified in clinical narrative. *Journal of Biomedical Informatics* 46, (2013), S48–S53. DOI: https://doi.org/10.1016/j.jbi.2013.09.010

[23] Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry de Bruijn. 2013. À la recherche du temps perdu: Extracting temporal relations from medical text in the 2012 i2b2 PLN challenge. *Journal of the American Medical Informatics Association* 20, 5 (2013), 843–848. DOI: https://doi.org/10.1136/amiajnl-2013-001624

[24] Raghavendra V. Chikka. 2016. CDE-IIITH at SemEval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1237–1240. https://doi.org/10.18653/v1/S16-1192

[25] Thanat Chokwijitkul, Anthony Nguyen, Hamed Hassanzadeh, and Siegfried Perez. 2019. Identifying risk factors for heart disease in electronic medical records: A deep learning approach. In *Proc. of the BioPLN 2018 workshop*. 18–27. https://doi.org/10.18653/v1/w18-2303

[26] Kim Clark, Deepak Sharma, Rui Qin, Christopher G. Chute, and Cui Tao. 2014. A use case study on late stent thrombosis for ontology-based temporal reasoning and analysis. *Journal of Biomedical Semantics* 5, 1 (2014), 1–9. DOI: https://doi.org/10.1186/2041-1480-5-49

[27] Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. GUIR at SemEval-2016 task 12: Temporal information processing for clinical narratives. *In Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1248–1255. https://doi.org/10.18653/v1/S16-1194

[28] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, (2011), 2493–2537. DOI: https://doi.org/10.1.1.231.4614

[29] James Cormack, Chinmoy Nath, David Milward, Kalpana Raja, and Siddhartha R. Jonnalagadda. 2015. Agile text mining for the 2014 i2b2/UTHealth cardiac risk factors challenge. *Journal of Biomedical Informatics* 58, (2015), S120–S127. DOI: https://doi.org/10.1016/j.jbi.2015.06.030

[30] Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations in clinical data: A hybrid, knowledge-rich approach. *Journal of Biomedical Informatics* 46, (2013), S29–S39. DOI: https://doi.org/10.1016/j.jbi.2013.08.003

[31] Jennifer D'Souza and Vincent Ng. 2014. Knowledge-rich temporal relation identification and classification in clinical notes. *Database* 2014, (2014), 1–20. DOI: https://doi.org/10.1093/database/bau109

[32] Joshua C. Denny, Josh F. Peterson, Neesha N. Choma, Hua Xu, Randolph A. Miller, Lisa Bastarache, and Neeraja B. Peterson. 2010. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *Journal of the American Medical Informatics Association* 17, 4 (2010), 383–388. DOI: https://doi.org/10.1136/jamia.2010.004804

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT 2019*, 4171–4186.

[34] Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 746–751.

[35] Elizabeth Ford, John A. Carroll, Helen E. Smith, Donia Scott, and Jackie A. Cassell. 2016. Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association* 23, 5 (2016), 1007–1015. DOI: https://doi.org/10.1093/jamia/ocv180

[36] Jason A. Fries. 2016. Brundlefly at SemEval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1274–1279. https://doi.org/ 10.18653/v1/S16-1198

[37] Rob Gaizauskas, Henk Harkema, Mark Hepple, and Andrea Setzer. 2006. Task-oriented extraction of temporal information: The case of clinical narratives. In *Proc. of the 13th International Symposium On Temporal Representation and Reasoning (time'06)*. 188–195. https://doi.org/10.1109/TIME.2006.27

[38] Diana Galvan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. 2018. Investigating the challenges of temporal relation extraction from clinical text. In *Proc. of the 9th International Workshop on Health Text Mining and Information Analysis*. 55–64. https://doi.org/10.18653/v1/w18-5607

[39] Travis Goodwin and Sanda M. Harabagiu. 2015. A probabilistic reasoning method for predicting the progression of clinical findings from electronic medical records. In *AMIA Summits on Translational Science Proceedings 2015*. 61–65.

[40] Alex Graves, Abdel Rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. *In 2013 IEEE international conference on acoustics, speech and signal processing*. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

[41] Cyril Grouin, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum. 2013. Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association* 20, 5 (2013), 820–827. DOI: https://doi.org/10.1136/amiajnl-2013-001627

[42] Cyril Grouin and Véronique Moriceau. 2016. LIMSI at SemEval-2016 task 12: machine-learning and temporal information to identify clinical events and time expressions. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1225–1230. https://doi.org/10.18653/v1/s16-1190

[43] Cyril Grouin, Véronique Moriceau, and Pierre Zweigenbaum. 2015. Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records. *Journal of Biomedical Informatics* 58, (2015), S133–S142. DOI: https://doi.org/10.1016/j.jbi.2015.06.014

[44] Thierry Hamon and Natalia Grabar. 2014. Tuning HeidelTime for identifying time expressions in clinical texts in English and French. In *Proc. of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*. 101–105. https://doi.org/ 10.3115/v1/W14-1116

[45] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics* 42, 5 (2009), 839–851. DOI: https://doi.org/10.1016/j.jbi.2009.05.002

[46] Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics* 57, (2015), 333–349. DOI: https://doi.org/10.1016/j.jbi.2015.08.013

[47] Po-Yu Huang, Hen-Hsen Huang, Yu-Wun Wang, Ching Huang, and Hsin-Hsi Chen. 2017. NTU-1 at SemEval-2017 task 12: Detection and classification of temporal events in clinical data with domain adaptation. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 1007–1010. https://doi.org/10.18653/v1/S17-2177

[48] Srinivasan V. Iyer, Rave Harpaz, Paea LePendu, Anna Bauer-Mehren, and Nigam H. Shah. 2014. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association* 21, 2 (2014), 353–362. DOI: https://doi.org/10.1136/amiajnl-2013-001612

[49] Srinivasan V. Iyer, Paea Lependu, Rave Harpaz, Anna Bauer-Mehren, and Nigam H. Shah. 2013. Learning signals of adverse drug-drug interactions from the unstructured text of electronic health records. In *AMIA Summits on Translational Science Proc. 2013*. 83–87.

[50] Serena Jeblee and Graeme Hirst. 2019. Listwise temporal ordering of events in clinical notes. In *Proc. of the 9th International Workshop on Health Text Mining and Information Analysis*. 177–182. https://doi.org/10.18653/v1/w18-5620

[51] Peter B. Jensen, Lars J. Jensen, and Soøren Brunak. 2012. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics* 13, 6 (2012), 395–405. DOI: https://doi.org/10.1038/nrg3208

[52] Jitendra Jonnagaddala, Siaw-Teng Liaw, Pradeep Ray, Manish Kumar, Hong-Jie Dai, and Chien-Yeh Hsu. 2015. Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. *BioMed Research International* 2015, (2015), 1–10. DOI: https://doi.org/10.1155/2015/636371

[53] George Karystianis, Azad Dehghan, Aleksandar Kovacevic, John A. Keane, and Goran Nenadic. 2015. Using local lexicalized rules to identify heart disease risk factors in clinical notes. *Journal of Biomedical Informatics* 58, (2015), S183–S188. DOI: https://doi.org/10.1016/j.jbi.2015.06.013

[54] Abdulrahman Khalifa and Stéphane Meystre. 2015. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics* 58, (2015), S128–S132. DOI: https://doi.org/10.1016/j.jbi.2015.08.002

[55] Abdulrahman Khalifa, Sumithra Velupillai, and Stephane Meystre. 2016. UtahBMI at SemEval-2016 task 12: Extracting temporal information from clinical text. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1256–1262. https://doi.org/10.18653/v1/S16-1195

[56] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics* 4, (2019), 100057. DOI: https://doi.org/10.1016/j.yjbinx.2019.100057

[57] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMPLN)*. 1746–1751. https://doi.org/10.3115/v1/d14-1181

[58] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 260–270. https://doi.org/10.18653/v1/N16-1030

[59] Natsuda Laokulrat, Makoto Miwa, and Yoshimasa Tsuruoka. 2015. Stacking approach to temporal relation Classification with temporal inference. *Journal of Natural Language Processing* 22, 3 (2015), 171–196. DOI: https://doi.org/10.5715/jPLN.22.171

[60] Natsuda Laokulrat, Yoshimasa Tsuruoka, Makoto Miwa, and Takashi Chikayama. 2013. UTTime: Temporal relation classification using deep syntactic features. In *2nd Joint Conf. on Lexical and Computational Semantics (* SEM), Volume 2: Proc. of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*. 88–92.

[61] Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics* 57, (2015), 28–37. DOI: https://doi.org/10.1016/j.jbi.2015.07.010

[62] Hee-Jin Lee, Min Jiang, Yonghui Wu, Christian M. Shaffer, John H. Cleator, Eitan A. Friedman, Joshua P. Lewis, Dan M. Roden, Josh Denny, and Hua Xu. 2017. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. In *AMIA Summits on Translational Science Proc. 2017*. 185–192.

[63] Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval-2016 task 12: An end-to-end system for temporal information extraction from clinical notes. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1292–1297. https://doi.org/10.18653/v1/S16-1201

[64] Hee Jin Lee, Yaoyun Zhang, Min Jiang, Jun Xu, Cui Tao, and Hua Xu. 2018. Identifying direct temporal relations between time and events from clinical notes. *BMC medical informatics and decision making* 18, 2 (2018). DOI: https://doi.org/10.1186/s12911-018-0627-5

[65] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan H. So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2019, 1–7. DOI: https://doi.org/10.1093/bioinformatics/btz682

[66] Wangjin Lee and Jinwook Choi. 2018. Temporal segmentation for capturing snapshots of patient histories in Korean clinical narrative. *Healthcare informatics research* 24, 3 (2018), 179–186. DOI: https://doi.org/10.4258/hir.2018.24.3.179

[67] Artuur Leeuwenberg and Marie-Francine Moens. 2016. KULeuven-LIIR at SemEval 2016 task 12: Detecting narrative containment in clinical records. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1280–1285. https://doi.org/10.18653/v1/S16-1199

[68] Artuur Leeuwenberg and Marie-Francine Moens. 2017. KULeuven-LIIR at SemEval-2017 task 12: Cross-domain temporal information extraction from clinical records. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 1029–1033. https://doi.org/10.18653/v1/S17-2181

[69] Artuur Leeuwenberg and Marie-Francine Moens. 2018. Word-level loss extensions for neural temporal relation classification. In *Proc. of the 27th International Conference on Computational Linguistics*. 3436–3447.

[70] Artuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 1150–1158. https://doi.org/10.18653/v1/e17-1108

[71] Min Li and Jon Patrick. 2012. Extracting temporal information from electronic patient records. In *AMIA Annual Symposium Proc. 2012*. 542–51.

[72] Peng Li and Heng Huang. 2016. UTA DLPLN at SemEval-2016 task 12: Deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1268–1273. https://doi.org/10.18653/v1/s16-1197

[73] Rumeng Li, Abhyuday N. Jagannatha, and Hong Yu. 2017. A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. In *AMIA Annual Symposium Proc. 2017*. 1149–1158.

[74] Chen Lin, Dmitriy Dligach, Timothy A. Miller, Steven Bethard, and Guergana K. Savova. 2016. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association* 23, 2 (2016), 387–395. DOI: https://doi.org/10.1093/jamia/ocv113

[75] Chen Lin, Elizabeth W. Karlson, Dmitriy Dligach, Monica P. Ramirez, Timothy A. Miller, Huan Mo, Natalie S. Braggs, Andrew Cagan, Vivian Gainer, Joshua C. Denny, and Guergana K. Savova. 2015. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association* 22, e1 (2015), e151–e161. DOI: https://doi.org/10.1136/amiajnl-2014-002642

[76] Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. In *Proc. of the 9th International Workshop on Health Text Mining and Information Analysis*. 165–176. https://doi.org/10.18653/v1/w18-5619

[77] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2016. Improving temporal relation extraction with training instance augmentation. In *Proc. of the 15th Workshop on Biomedical Natural Language Processing*. 108–113. https://doi.org/10.18653/v1/W16-2914

[78] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2017. Representations of time expressions for temporal relation extraction with convolutional neural networks. In *Proc. of BioPLN* 2017.322–327. https://doi.org/10.18653/v1/W17-2341

[79] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proc. of the 2nd Clinical Natural Language Processing Workshop*. 65–71. https://doi.org/ 10.18653/v1/W19-1908

[80] Chen Lin, Timothy Miller, Alvin Kho, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, and Guergana Savova. 2014. Descending-path convolution kernel for syntactic structures. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*. 81–86. https://doi.org/ 10.3115/v1/P14-2014

[81] Sijia Liu, Liwei Wang, Vipin Chaudhary, and Hongfang Liu. 2019. Attention neural model for temporal relation extraction. In *Proc. of the 2nd Clinical Natural Language Processing Workshop*. 134–139. https://doi.org/10.18653/v1/w19-1917

[82] Sijia Liu, Liwei Wang, Donna Ihrke, Vipin Chaudhary, Cui Tao, Chunhua Weng, and Hongfang Liu. 2017. Correlating lab test results in clinical notes with structured lab data: A case study in hbA1c and glucose. In *AMIA Summits on Translational Science Proc.* 2017. 221–228.

[83] Yi Liu, Paea Lependu, Srinivasan Iyer, and Nigam H. Shah. 2012. Using temporal patterns in medical records to discern adverse drug events from indications. In *AMIA Summits on Translational Science Proc. 2012*. 47–56.

[84] Zengjian Liu, Xiaolong Wang, Qingcai Chen, Buzhou Tang, and Hua Xu. 2019. Temporal indexing of medical entity in Chinese clinical notes. *BMC medical informatics and decision making* 19, Suppl 1 (2019). DOI: https://doi.org/10.1186/s12911-019-0735-x

[85] Yu Long, Zhijing Li, Xuan Wang, and Chen Li. 2017. XJPLN at SemEval-2017 task 12: Clinical temporal information ex-traction with a hybrid model. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 1011–1015. https://doi.org/10.18653/v1/S17-2178

[86] Zhihui Luo, Stephen B. Johnson, Albert M. Lai, and Chunhua Weng. 2011. Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. In *AMIA annual symposium Proc. 2011*. 843–52.

[87] Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 1021–1026. https://doi.org/10.18653/v1/S17-2180

[88] Stéphane M. Meystre, Guergana Savova, Karin Kipper-Schuler, and John F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of medical informatics* 17, 1 (2008), 128–144. DOI: 10.1055/s-0038-1638592

[89] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of the 26th International Conference on Neural Information Processing Systems*. 3111--3119.

[90] Timothy Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana Savova. 2013. Discovering temporal narrative containers in clinical text. In *Proc. of the 2013 Workshop on Biomedical Natural Language Processing*. 18–26.

[91] Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and temporal relation extraction from natural language texts. In *Proc. of COLING 2016, the 26th International Conference on Computational Linguistics*. 64–75.

[92] Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*. 1105–1116. https://doi.org/10.18653/v1/p16-1105

[93] Gandhimathi Moharasan and Tu-Bao Ho. 2019. Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research* 3, 2 (2019), 220–244. DOI: https://doi.org/10.1007/s41666-019-00049-0

[94] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Systematic reviews and meta-analyses: The PRISMA statement. *Annals of internal medicine* 151, 4 (2009), 264–269. DOI: https://doi.org/10.7326/0003-4819-151-4-200908180-00135

[95] Danielle L Mowery Ms, Henk Harkema, John N Dowling Ms, Jonathan L Lustgarten, and Wendy W Chapman. 2009. Distinguishing historical from current problems in clinical reports — which textual features help?. In *Proc. of the BioPLN 2009 Workshop*. 10–18.

[96] Azadeh Nikfarjam, Ehsan Emadzadeh, and Graciela Gonzalez. 2013. Towards generating a patient's timeline: Extracting temporal relationships from clinical notes. *Journal of biomedical informatics* 46, SUPPL. (2013), 1–21. DOI: https://doi.org/10.1016/j.jbi.2013.11.001

[97] Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1027–1037. https://doi.org/10.18653/v1/D17-1108

[98] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proc. of the 56th Annual Meeting of the Association for Computational*. 2278–2288. https://doi.org/10.18653/v1/p18-1212

[99] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2227–2237. https://doi.org/10.18653/v1/N18-1202

[100] James Pustejovsky, Jose Castano, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering* 3, (2003), 28–34.

[101] James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Beth Sundheim, David Day, Lisa Ferro, and Dragomir. 2003. The TIMEBANK corpus. In *Proc. of Corpus Linguistics 2003*. 647-656.

[102] James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language resources and evaluation* 39, 2–3 (2005), 123–164. DOI: https://doi.org/10.1007/s10579-005-7882-7

[103] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proc. of the 7th International Conference on Language Resources and Evaluation*. 394–397.

[104] James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proc. of the 5th Linguistic Annotation Workshop*, 152–160.

[105] Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad, and Albert M. Lai. 2014. Cross-narrative temporal ordering of medical events. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*. 998–1008. https://doi.org/10.3115/v1/p14-1094

[106] Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai. 2012. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. In *Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 731–741.

[107] Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai. 2012. Learning to temporally order medical events in clinical text. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*. 70–74.

[108] Preethi Raghavan, Eric Fosler-lussier, and Albert M Lai. 2012. Temporal classification of medical events. In *Proc. of the 2012 Workshop on Biomedical Natural Language Processing*. 29–37.

[109] Kirk Roberts, Bryan Rink, and Sanda M. Harabagiu. 2013. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association* 20, 5 (2013), 867–875. DOI: https://doi.org/10.1136/amiajnl-2013-001619

[110] Kirk Roberts, Sonya E. Shooshan, Laritza Rodriguez, Swapna Abhyankar, Halil Kilicoglu, and Dina Demner-Fushman. 2015. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. *Journal of Biomedical Informatics* 58, (2015), S111–S119. DOI: https://doi.org/10.1016/j.jbi.2015.06.010

[111] P. R. Sarath, Ravikiran Manikandan, and Yoshiki Niwa. 2017. Hitachi at SemEval-2017 task 12: System for temporal information extraction from clinical notes. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 1002–1006. https://doi.org/10.18653/v1/S17-2176

[112] Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, and Andrea Setzer. 2006. TimeML annotation guidelines, version 1.2.1. (March 2005). Retrieved from http://www.timeml.org/timeMLdocs/AnnGuide121.pdf

[113] Guergana K. Savova, Janet E. Olson, Sean P. Murphy, Victoria L. Cafourek, Fergus J. Couch, Matthew P. Goetz, James N. Ingle, Vera J. Suman, Christopher G. Chute, and Richard M. Weinshilboum. 2012. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. *Journal of the American Medical Informatics Association* 19, E1 (2012), 361–370. DOI: https://doi.org/10.1136/amiajnl-2011-000295

[114] Jae Wook Seol, Wangjin Yi, Jinwook Choi, and Kyung Soon Lee. 2017. Causality patterns and machine learning for the extraction of problem-action relations in discharge summaries. *International journal of medical informatics* 98, (2017), 1-12. DOI: https://doi.org/10.1016/j.ijmedinf.2016.10.021

[115] Chaitanya Shivade, Pranav Malewadkar, Eric Fosler-Lussier, and Albert M. Lai. 2015. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. *Journal of Biomedical Informatics* 58, (2015), S103–S110. DOI: https://doi.org/10.1016/j.jbi.2015.08.025

[116] Sunghwan Sohn, Kavishwar B. Wagholikar, Dingcheng Li, Siddhartha R. Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2013. Comprehensive temporal information detection from clinical text: Medical events, time, and TLINK identification. *Journal of the American Medical Informatics Association* 20, 5 (2013), 836–842. DOI: https://doi.org/10.1136/amiajnl-2013-001622

[117] Jennifer D'Souza and Vincent Ng. 2014. Annotating inter-sentence temporal relations in clinical notes. In *Proc. of the 9th International Conference on Language Resources and Evaluation*. 2758–2765.

[118] Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation* 47, 2 (2013), 269–298. DOI: https://doi.org/10.1007/s10579-012-9179-y

[119] Amber Stubbs and Benjamin Harshfield. 2010. Applying the TARSQI toolkit to augment text mining of EHRs. In *Proc. of the 2010 Workshop on Biomedical Natural Language Processing*. 141–143.

[120] Amber Stubbs, Christopher Kotfila, Hua Xu, and Özlem Uzuner. 2015. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of Biomedical Informatics* 58, (2015), S67–S77. DOI: https://doi.org/10.1016/j.jbi.2015.07.001

[121] William F. Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* 2, (2014), 143–154. DOI: https://doi.org/10.1162/tacl_a_00172

[122] Jia Su, Jinpeng Hu, Jingchi Jiang, Jing Xie, Yang Yang, Bin He, Jinfeng Yang, and Yi Guan. 2019. Extraction of risk factors for cardiovascular diseases from Chinese electronic medical records. *Computer methods and programs in biomedicine* 172, (2019), 1–10. DOI: https://doi.org/10.1016/j.cmpb.2019.01.007

[123] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* 20, 5 (2013), 806–813. DOI: https://doi.org/10.1136/amiajnl-2013-001628

[124] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics* 46, (2013), S5–S12. DOI: https://doi.org/10.1016/j.jbi.2013.07.004

[125] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Temporal reasoning over clinical text: The state of the art. *Journal of the American Medical Informatics Association* 20, 5 (2013), 814–819. DOI: https://doi.org/10.1136/amiajnl-2013-001760

[126] Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C. Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association* 20, 5 (2013), 828–835. DOI: https://doi.org/10.1136/amiajnl-2013-001635

[127] Manabu Torii, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T Wiley, and Daniel S Zisook. 2015. Risk factor detection for heart disease by applying text analytics in electronic medical records. *Journal of Biomedical Informatics* 58, (2015), S164–S170. DOI: https://doi.org/10.1016/j.jbi.2015.08.011.Risk

[128] Julien Tourille, Olivier Ferret, Aurelie Névéol, and Xavier Tannier. 2017. Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers. *In Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*. 224–230. https://doi.org/10.18653/v1/P17-2035

[129] Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016. LIMSI-COT at SemEval-2016 task 12: Temporal relation identification using a pipeline of classifiers. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 1136–1142. https://doi.org/10.18653/v1/S16-1175

[130] Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. Temporal information extraction from clinical text. In *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 739–745. https://doi.org/10.18653/v1/e17-2117

[131] Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. LIMSI-COT at SemEval-2017 task 12: Neural architecture for temporal information extraction from Clinical Narratives. In *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 595–600. https://doi.org/10.18653/v1/S17-2098

[132] Jay Urbain. 2015. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. *Journal of Biomedical Informatics* 58, (2015), S143–S149. DOI: https://doi.org/10.1016/j.jbi.2015.08.009

[133] Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 2013. Semeval-2013 task 1: {T}empeval-3: Evaluating time expressions, events, and temporal relations. In *2nd Joint Conf. on Lexical and Computational Semantics (* SEM), Volume 2: Proc. of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*. 1–9.

[134] Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. 2906–2919. https://doi.org/10.18653/v1/P19-1280

[135] Sumithra Velupillai, Danielle L. Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W. Chapman. 2015. BluLab: Temporal information extraction for the 2015 clinical TempEval challenge. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*. 815–819.

[136] Marc Verhagen. 2005. Temporal closure in an annotation environment. *Language Resources and Evaluation* 39, 2–3 (2005), 211–241. DOI: https://doi.org/10.1007/s10579-005-7884-5

[137] Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the TARSQI toolkit. In *COLING 2008: Companion volume: Demonstrations*. 189–192.

[138] Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Robert Stewart, Rashmi Patel, and Sumithra Velupillai. 2019. Annotating temporal relations to determine the onset of psychosis symptoms. *Studies in health technology and informatics* 264, (2019), 418–422. DOI: https://doi.org/10.3233/SHTI190255

[139] Natalia Viani, Timothy A. Miller, Carlo Napolitano, Silvia G. Priori, Guergana K. Savova, Riccardo Bellazzi, and Lucia Sacchi. 2019. Supervised methods to extract clinical events from cardiology reports in Italian. *Journal of Biomedical Informatics* 95, (2019), 103219. DOI: https://doi.org/10.1016/j.jbi.2019.103219

[140] Duc-Thuan Vo and Ebrahim Bagheri. 2019. Extracting temporal event relations based on event networks. In *European Conference on Information Retrieval* 2019. 844–851. https://doi.org/10.1007/978-3-030-15712-8_61

[141] Wei Wang, Kory Kreimeyer, Emily J. Woo, Robert Ball, Matthew Foster, Abhishek Pandey, John Scott, and Taxiarchis Botsis. 2016. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. *Journal of Biomedical Informatics* 62, (2016), 78–89. DOI: https://doi.org/10.1016/j.jbi.2016.06.006

[142] Zixu Wang, Julia Ive, Sumithra Velupillai, and Lucia Specia. 2019. Is artificial data useful for biomedical natural language processing algorithms?. In *Proc. of the 18th BioPLN Workshop and Shared Task*. 240–249. https://doi.org/10.18653/v1/w19-5026

[143] Stephen T. Wu, Young J. Juhn, Sunghwan Sohn, and Hongfang Liu. 2014. Patient-level temporal aggregation for text-based asthma status ascertainment. *Journal of the American Medical Informatics Association* 21, 5 (2014), 876–884. DOI: https://doi.org/10.1136/amiajnl-2013-002463

[144] Dong Xu, Meizhuo Zhang, Tianwan Zhao, Chen Ge, Weiguo Gao, Jia Wei, and Kenny Q. Zhu. 2015. Data-driven information extraction from Chinese electronic medical records. *PLoS One* 10, 8 (2015). DOI: https://doi.org/10.1371/journal.pone.0136270

[145] Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I. Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* 20, 5 (2013), 849–858. DOI: https://doi.org/10.1136/amiajnl-2012-001607

[146] Hui Yang and Jonathan M. Garibaldi. 2015. A hybrid model for automatic identification of risk factors for heart disease. *Journal of Biomedical Informatics* 58, (2015), S171–S182. DOI: https://doi.org/10.1016/j.jbi.2015.09.006

[147] Meliha Yetisgen-yildiz, Fei Xia, Lucy Vanderwende, and Mark M. Wurfel. 2011. Identifying patients with pneumonia from free-text intensive care unit reports identifying patients with pneumonia from free-text ICU reports. In *Proc. of Learning from Unstructured Clinical Text Workshop of the International Conference on Machine Learning 2011*.

[148] Li Zhou, Carol Friedman, Simon Parsons, and George Hripcsak. 2005. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. In *AMIA Annual Symposium Proc. 2005*. 869–873.

[149] Li Zhou, Simon Parsons, and George Hripcsak. 2008. The Evaluation of a temporal reasoning system in processing clinical discharge summaries. *Journal of the American Medical Informatics Association* 15, 1 (2008), 99–106. DOI: https://doi.org/10.1197/jamia.M2467

[150] Lichao Zhu, Hangzhou Yang, and Zhijun Yan. 2017. Mining medical related temporal information from patient's self-description. *International Journal of Crowd Science* 1, 2 (2017), 110–120. DOI: https://doi.org/10.1108/IJCS-08-2017-0018

[151] Will Styler, Guergana Savova, Martha Palmer, James Pustejovsky, Tim O'Gorman, and Piet C. de Groen. 2017. THYME annotation guidelines. (February 2014). Retrieved January 20, 2020 from http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

APÊNDICE B – RESUMO DE *GUIDELINE* PARA ANOTAÇÃO DE EVENTOS

Nesta anotação de EVTs, foram considerados quaisquer acontecimentos ou menções importantes na criação de uma *timeline* do paciente. Para cada EVT, atributos foram marcados com o objetivo de melhor defini-lo: Tipo, Polaridade, Modalidade e RelTempDCD.

TIPO

A definição de cada tipo foi adaptada para melhor cobrir os textos ambulatoriais de cardiologia, sempre fazendo um paralelo com o SOAP.

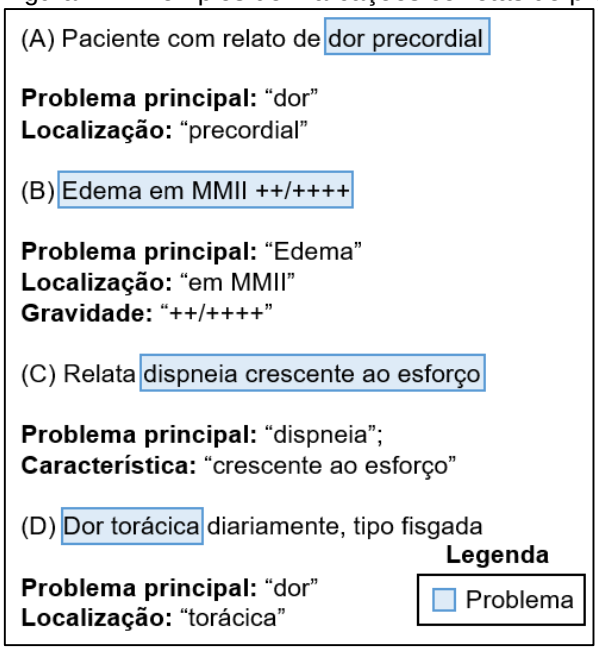
- **Problemas**

O conceito de Problema foi previamente definido, porém é necessário ressaltar quais elementos não foram marcados como Problema de forma alguma durante a anotação:

- Condições normais como “BEG” (bom estado geral), “corado” e “LOTE” (lúcido, orientado no tempo e espaço), questões usualmente registradas pelos profissionais de saúde durante o exame físico.
- Resultados de exames que apontavam normalidade, como “ventrículo esquerdo com diâmetros normais”.
- Resultados de exames laboratoriais que apontavam valores anormais, como “creatinina 2 mg/dl”; neste caso, a “creatinina” era marcada como um teste.
- Resultados de medições fora do normal, como “pressão arterial 145/95”; neste caso, a “pressão arterial” era marcada como teste. Somente quando houvesse qualquer tipo de inferência sobre o valor previamente descrito, como no trecho “pressão arterial elevada”, este passaria a ser uma menção de problema.
- Verbos que tinham relação direta com problemas em contexto; em “falta de ar piorou”, “falta de ar” era o problema central, porém “piorou” era uma condição dele, sendo marcado como uma menção de ocorrência.

Esses fatores são de suma importância devido a características da anotação. Principalmente por existir somente um rótulo de tipo para cada EVT, incorretamente adicionar uma condição como “piorou” ao problema (como mostrado no exemplo) eliminaria a possibilidade de adicionar esse termo corretamente como uma ocorrência.

Figura 1 – Exemplos de marcações corretas de problemas, com sinalização dos atributos.



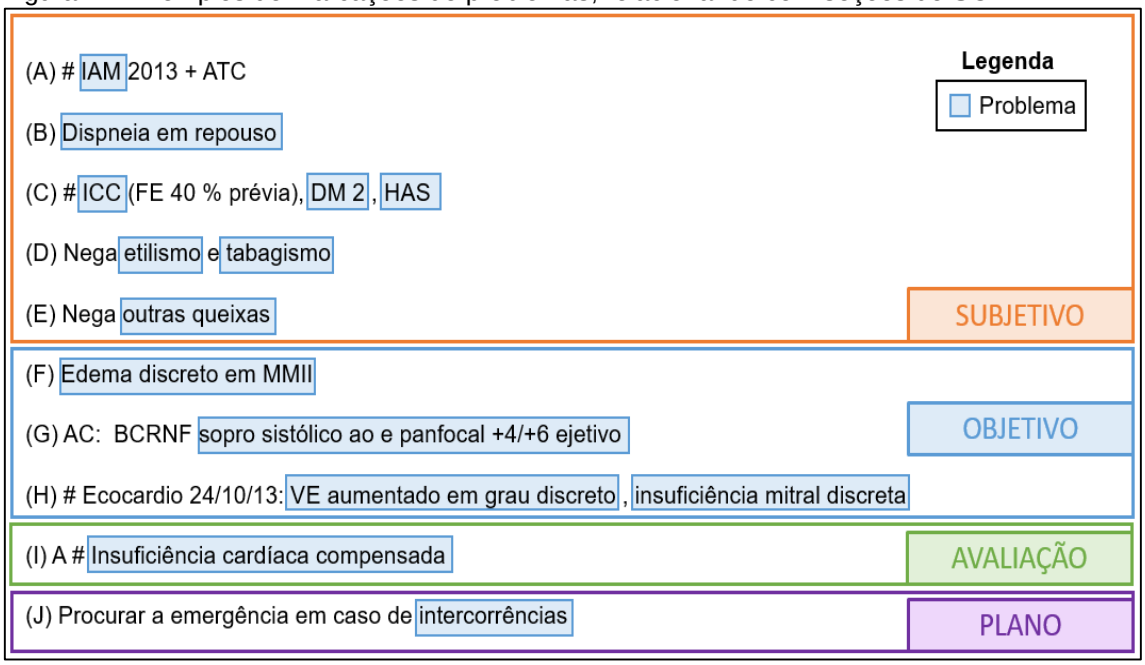
Fonte: O autor (2020).

Alguns exemplos de marcações são mostrados na Figura 1. No exemplo A, é importante trazer a localização, uma vez que é uma dor específica na região do peito em frente do coração. Já no exemplo B, tem-se um edema encontrado durante o exame físico, sendo fundamental evidenciar que é nos membros inferiores (MMII) e que sua gravidade é 2 em uma escala de 4. No exemplo C, tem-se um caso muito comum em textos cardiológicos, que é a caracterização da dispneia, sendo crucial trazer suas características. Questões como “aos grandes esforços”, “crescente ao esforço” e “em repouso” melhor definem esse problema. No exemplo D, existe uma questão importante: durante a anotação, não terem sido consideradas menções disjuntas. Dessa forma, apesar de “tipo fisgada” ser uma característica importante, por não estar interligada ao trecho contendo a menção “dor torácica”, é desconsiderado. Neste caso, “diariamente” entraria como uma ET.

Na seção Subjetiva, podem ser encontradas menções de problemas relativos a sintomas associados, assim como problemas médicos relacionados à história passada do paciente, além de doenças crônicas e histórico social (incluindo tabagismo

e etilismo). Alguns exemplos desses problemas encontrados na seção Subjetiva são mostrados na Figura 2, itens A até E. No exemplo A, tem-se um caso de “IAM” (infarto agudo do miocárdio), um problema pertencente à história do paciente, tendo sido feita uma angioplastia. Em B, tem-se um problema envolvendo um sintoma relatado, sendo uma “dispneia em repouso”. No exemplo C, há três menções de doenças crônicas, “ICC” (insuficiente cardíaca crônica), “DM 2” (*diabetes mellitus* tipo II) e “HAS” (hipertensão arterial). Tanto etilismo quanto tabagismo têm códigos na CID-10, por isso foram considerados problemas (isso inclui ex-tabagista). No exemplo E, tem-se um caso de “outras queixas”, sendo comum durante a entrevista do paciente menções envolvendo “queixas” ou “outras queixas” com o intuito de negar problemas/sintomas adicionais.

Figura 2 – Exemplos de marcações de problemas, relacionando com seções do SOAP.



Fonte: O autor (2020).

Na seção Objetiva, existem problemas achados durante os exames físico e visual, além de problemas contidos em exames de diagnóstico. Dados provenientes de exame físico envolvem qualquer informação obtida da observação geral, ausculta, palpação e percussão, assim como as reações do paciente (PEARCE *et al.*, 2016). Alguns exemplos de problemas são mostrados na Figura 2, nos itens F, G e H. Os exemplos F e G envolvem problemas achados durante o exame físico, sendo o sopro sistólico achado durante a ausculta cardíaca. No item H, há dois problemas, “VE

aumentado em grau discreto” e “insuficiente mitral discreta”, que foram encontrados durante um exame de ecocardiograma.

Na seção Avaliação, é feito o diagnóstico pelo profissional de saúde, levando em contato os elementos das seções Subjetiva e Objetiva. Exemplificando esse aspecto, na Figura 2I foi constatada uma “insuficiência cardíaca compensada” durante o diagnóstico efetuado pelo profissional de saúde.

Menções de problemas na seção Planejamento (Plano) envolvem EVTs condicionais futuros, contemplando frases em estrutura como: “se” determinado problema ocorrer, “então” faça determinada ação. Na Figura 2J, existe um exemplo desse tipo de menção, havendo a observação feita pelo profissional de saúde, indicando ao paciente buscar a emergência caso alguma intercorrência.

- **Tratamentos**

O conceito de Tratamento foi previamente definido, porém é necessário ressaltar quais elementos não foram marcados como tratamentos durante a anotação. Não poderiam ser anotadas como Tratamento menções de verbos relacionadas à realização do tratamento e locais ou departamentos em que foram realizados os tratamentos.

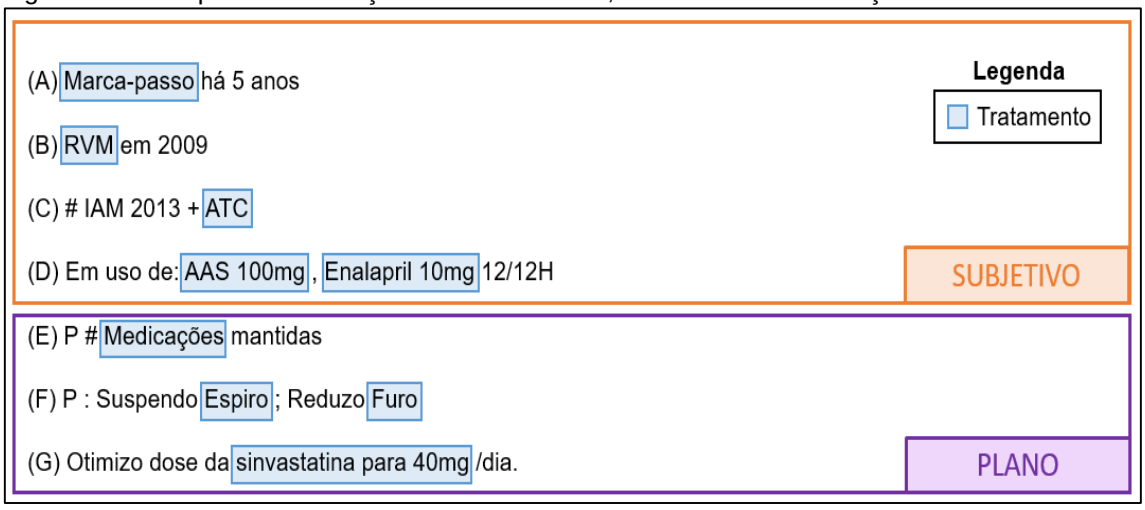
Assim como na anotação de Problema, o objetivo era trazer marcações mais específicas para tratamentos. Para marcações de procedimentos e intervenções, a localização era essencial (quando mencionada), sendo correta a anotação de “*stent* em coronária direita” e não somente “*stent*” quando a informação sobre a localização é fornecida. No caso de medicamentos, o ideal era sempre trazer a dosagem junto ao medicamento (quando mencionado), como o caso de “Sinvastatina 20 mg”.

Com as particularidades da anotação de tratamentos descritas, a seguir são detalhados os elementos considerados tratamentos, fazendo uma correlação direta com as seções da estrutura SOAP.

Na seção Subjetiva, os tratamentos envolviam as medicações em uso e procedimentos e intervenções realizados durante o passado do paciente (histórico médico passado). Na Figura 3, são mostrados alguns exemplos nos itens A até D. No exemplo A, tem-se uma menção de “marca-passo”, comum em textos de cardiologia. Nos exemplos B e C, há menções de “RVM” (revascularização do miocárdio) e “ATC” (angioplastia), ambas sendo alternativas bem aceitas para tratamento de insuficiência

coronariana (ANDRADE *et al.*, 2011). No exemplo D, há medicações em uso; nessas menções, é incluída a dosagem.

Figura 3 – Exemplos de marcações de tratamentos, relacionando com seções do SOAP.



Fonte: O autor (2020).

Na seção Planejamento (Plano), as menções de tratamentos também envolvem medicações, procedimentos e intervenções, porém, nesse contexto, estão relacionadas com as medidas propostas pelo profissional de saúde durante a consulta a partir do diagnóstico (prescrição). Na Figura 3, são mostrados alguns exemplos nos itens E, F e G. No exemplo E, existe uma menção geral de “medicações” (indicando as medicações em uso), temporalmente relacionada a “mantidas” durante as anotações das RTs. No exemplo F, tem-se a suspensão de “Espiro” (Espironolactona) e redução de “Furo” (Furosemida); nas anotações das RTs, a suspensão e redução serão relacionadas com o medicamento. No exemplo G, tem-se a otimização da dose, que seria o ajuste dela.

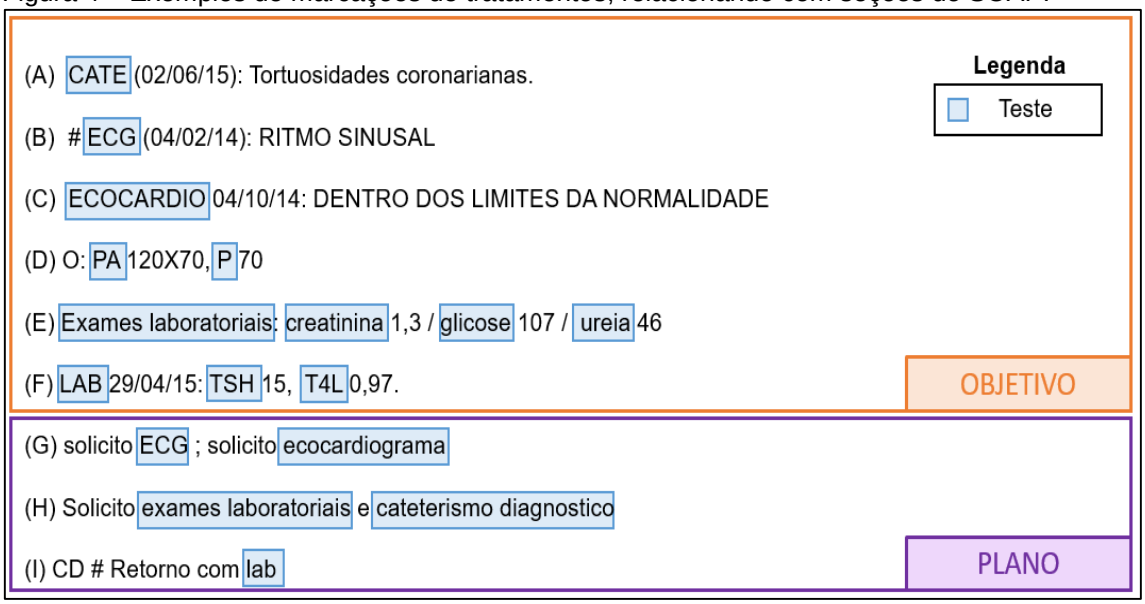
- **Testes**

Não poderiam ser anotadas como Testes: menções de verbos relacionadas à realização ou não realização de um teste e resultados junto aos testes.

Para marcações de testes, é essencial trazer o termo mais completo possível, como, por exemplo, corretamente anotando “ecocardiograma transesofágico” e não restringir a anotação a “ecocardiograma”, quando toda a informação é fornecida, ou “exame laboratorial”, não restringindo a “exame” ou “laboratorial”.

Na seção Objetiva, as menções de teste envolvem exames físico, visual e laboratoriais. Na Figura 4, são mostrados alguns exemplos nos itens A até F. Nos exemplos A, B e C, há menções de “CATE” (cateterismo cardíaco), “ECG” (eletrocardiograma) e “ECOCARDIO” (ecocardiograma), testes de diagnóstico usuais em texto de cardiologia. No exemplo D, existem menções de “PA” (pressão arterial) e “P” (pressão), medidas obtidas pelo exame físico. Nos exemplos E e F, tem-se menções de exames laboratoriais, em E todos por extenso, o que é raro em textos, e em F expressos por meio de abreviaturas. Vale salientar que menções gerais de exames laboratoriais são consideradas, pois, na etapa de anotação de RTs, estas vão “conter” as menções de exame.

Figura 4 – Exemplos de marcações de tratamentos, relacionando com seções do SOAP.



Fonte: O autor (2020).

Na seção Planejamento (Plano), existem menções de exames laboratoriais e testes/procedimentos de diagnóstico pedidos pelo profissional da saúde para a próxima consulta, representadas na Figura 4, exemplos G, H e I.

- **Evidências**

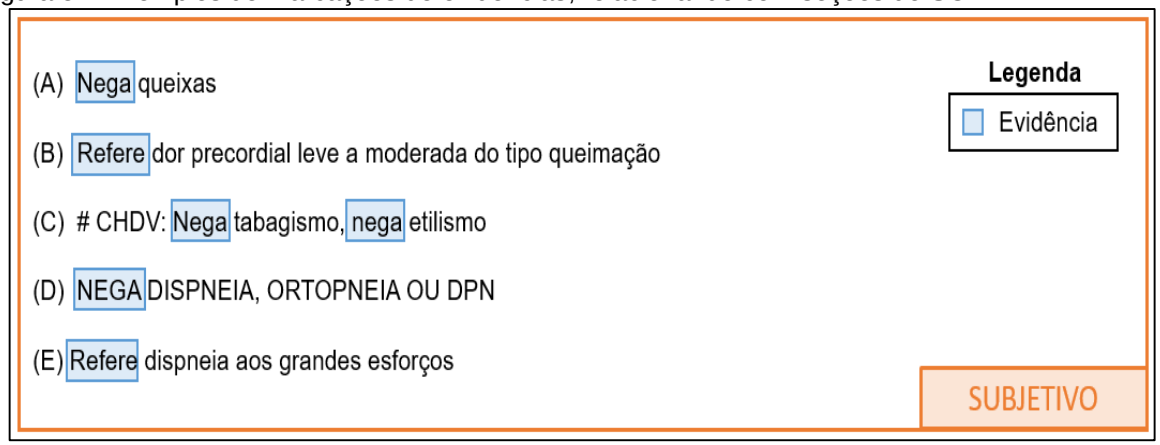
Evidências servem como conexão entre problemas e a sua fonte de informação. Em problemas obtidos por testes de diagnósticos, elas são menções como “mostrou” e “relevou”, que relacionam o teste com seu respectivo resultado. É usual ter somente o teste, uma sinalização como “:” e o resultado, porém existem

momentos em que elementos de ligação são usados pelo profissional de saúde na escrita do documento.

Os casos mais comuns de menções de evidência envolvem a entrevista do paciente, com determinados sintomas sendo “negados”, “relatados” e “queixados”. Nesse contexto, o problema é o sintoma, a fonte de informação é o paciente e a evidência é a ligação entre ambos.

Menções de evidências estão presentes na seção Subjetiva do SOAP, sendo alguns exemplos mostrados na Figura 5. No exemplo A, tem-se queixas, de maneira geral, sendo negadas. Nos exemplos C e D, há casos de “nega” em conjunto com questões importantes, como “tabagismo”, “etilismo”, “dispneia”, “ortoponeia” e “DPN” (dispneia paroxística noturna). Nos exemplos B e C, tem-se casos de sintomas referidos.

Figura 5 – Exemplos de marcações de evidências, relacionando com seções do SOAP.



Fonte: O autor (2020).

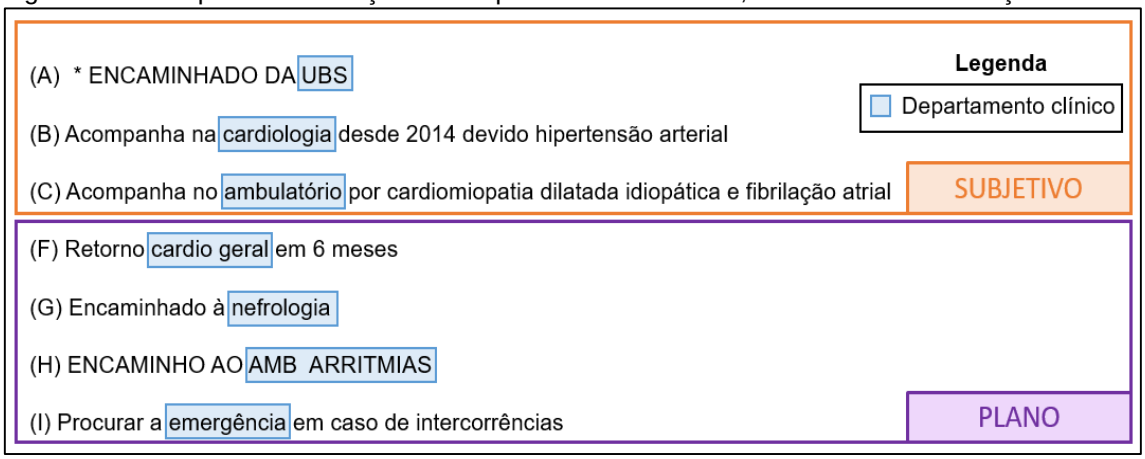
- **Departamentos Clínicos**

Qualquer local, departamento ou serviço que o paciente visitou, está visitando ou visitará com o objetivo de tratar, realizar exames ou buscar diagnóstico é considerado um Departamento Clínico. Verbos que indicam a ida ou retorno a departamentos clínicos não são considerados parte do EVT.

Na seção Subjetiva, departamentos clínicos aparecem no histórico médico do paciente. Existem menções de departamentos relativas ao passado do paciente, como, por exemplo, na Figura 6A, em que o paciente foi encaminhado da “UBS” (unidade básica de saúde). Há casos de acompanhamento nos exemplos B e C,

indicando uma continuidade de consultas em determinados departamentos clínicos, normalmente devido a alguma doença/condição crônica.

Figura 6 – Exemplos de marcações de departamentos clínicos, relacionando com seções do SOAP.



Fonte: O autor (2020).

Na seção Planejamento (Plano), menções de departamentos clínicos envolviam profissionais e locais que o paciente deve visitar no futuro, como nos exemplos F, G e H, com retornos e encaminhamentos a departamentos clínicos, sendo estes “cardio geral” (cardiologista geral), “nefrologia” e “amb arritmias” (ambulatório de arritmias). Existe também o caso de hipoteticamente o paciente ter que visitar um departamento quando de algum problema em específico, exemplificado na Figura 6I.

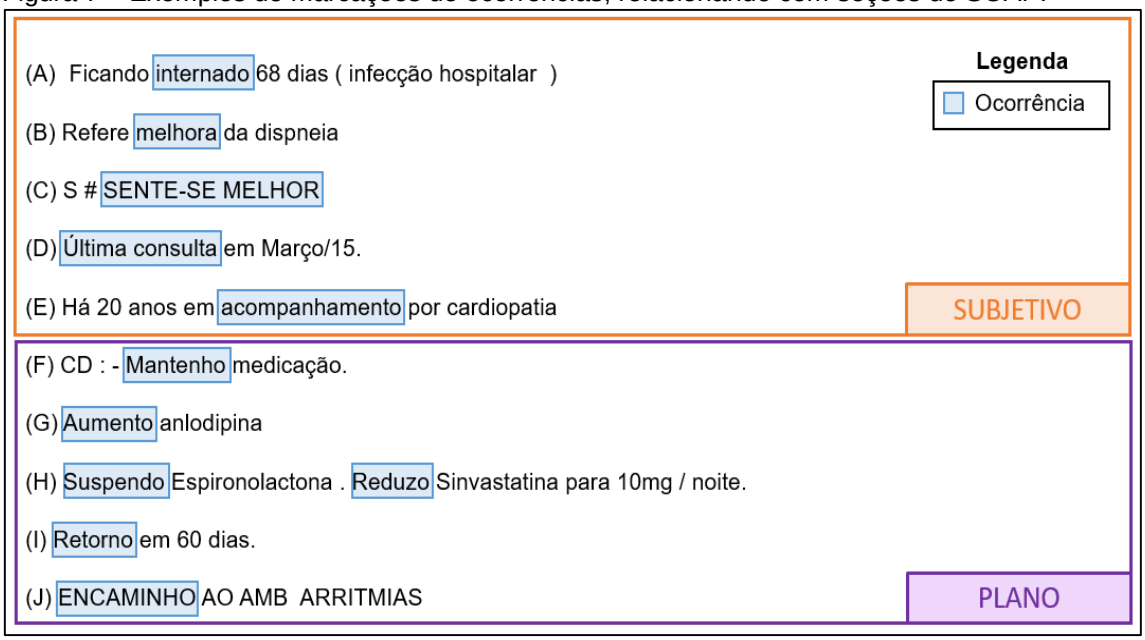
- **Ocorrências**

Ocorrências fazem parte da uma categoria genérica, porém são marcações essenciais, que muitas vezes modificam ou alteram o entendimento sobre os demais EVTs, sendo um fator complementar, como, por exemplo, a questão de “aumento”, “diminuição”, “melhora” e “piora” ligada a determinado problema, caso de menções de “suspensão” e “otimização” ligadas à prescrição de medicamentos, além de envolver menções relacionadas a “acompanhamento”, “retorno”, “consulta” e “internamento”, que são fatos importantes tanto no histórico quanto no plano de cuidado do paciente.

Menções de ocorrências estavam presentes na seção Subjetiva, envolvendo o histórico médico do paciente. Na Figura 7, são trazidos alguns exemplos, dos itens A até E. No exemplo A, tem-se uma menção de internamento, referente ao histórico do

paciente. No exemplo D, há outra menção envolvendo o histórico do paciente, neste caso, uma consulta anterior. Nos exemplos B e C, tem-se relatos do paciente durante a entrevista com o profissional de saúde; em B, o paciente menciona uma “melhora” relativa à sua “dispneia” e, em C, estar se sentindo melhor. O exemplo E indica o “acompanhamento” por um período de 20 anos devido a uma cardiopatia.

Figura 7 – Exemplos de marcações de ocorrências, relacionando com seções do SOAP.



Fonte: O autor (2020).

Ocorrências na seção Planejamento (Plano) são extremamente importantes devido a ser menções ligadas a tratamentos, especialmente mudanças na medicação. Há alguns exemplos na Figura 7, dos itens F até H. Menções como “aumento”, “suspendo”, “otimizo”, “reduzo” e “mantenho” são essenciais para sinalizar o cuidado prestado. Além disso, ocorrências como indicação de retorno (I) e encaminhamentos (J) são igualmente importantes durante o planejamento.

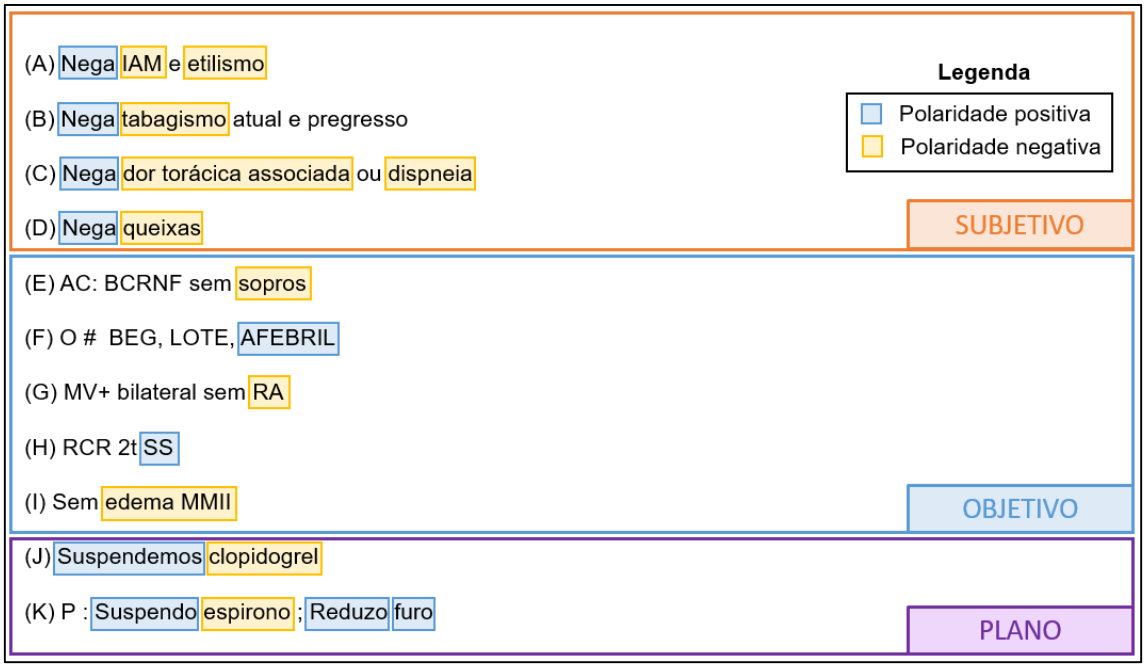
POLARIDADE

O atributo Polaridade tem relação com a ocorrência de um EVT, podendo ser positiva ou negativa. EVTs com polaridade positiva são EVTs “positivos” que aconteceram em determinado momento, estão acontecendo (durante a consulta) ou acontecerão (esperado). EVTs com polaridade negativa são o oposto, ou seja, EVTs que não aconteceram no passado, que não estão acontecendo ou que não

acontecerão. O valor padrão para este atributo era positivo, uma vez que a maior parte dos EVTs era positiva, sendo necessária alguma dica textual para inferir marcações negativas.

Na seção Subjetiva, pode haver exemplos de polaridade negativa envolvendo o termo “nega”, indicando que determinado EVT não existiu no histórico no paciente. Vale salientar que “nega” tem polaridade positiva, sendo somente um modificador. Alguns exemplos são mostrados na Figura 8, itens A até D, em que “nega” indica a não ocorrência de determinados sintomas (“dor torácica associada”, “dispneia” e “queixas”) e hábitos sociais (“tabagismo” e “elitismo”) no passado do paciente. No exemplo A, ainda existe a indicação de não ocorrência de um “IAM” (infarto agudo do miocárdio) no passado do paciente.

Figura 8 – Exemplos de marcações envolvendo atributo Polaridade, relacionando com seções do SOAP.



Fonte: O autor (2020).

Na seção Objetivo, tem-se menções relacionadas ao exame físico. Nos itens E, G e I, há problemas como “sopros”, “RA” (ruídos adventícios) e “edema MMII” (edema de membros inferiores), não constatados durante o exame físico. Há casos como dos exemplos F e H, nos quais, devido aos próprios prefixos dos EVTs os negarem, “a” em “afebril” e “s” em SS (sem sopro), se decidiu marcá-los como problemas (apesar de serem condições normais), porém com polaridade positiva.

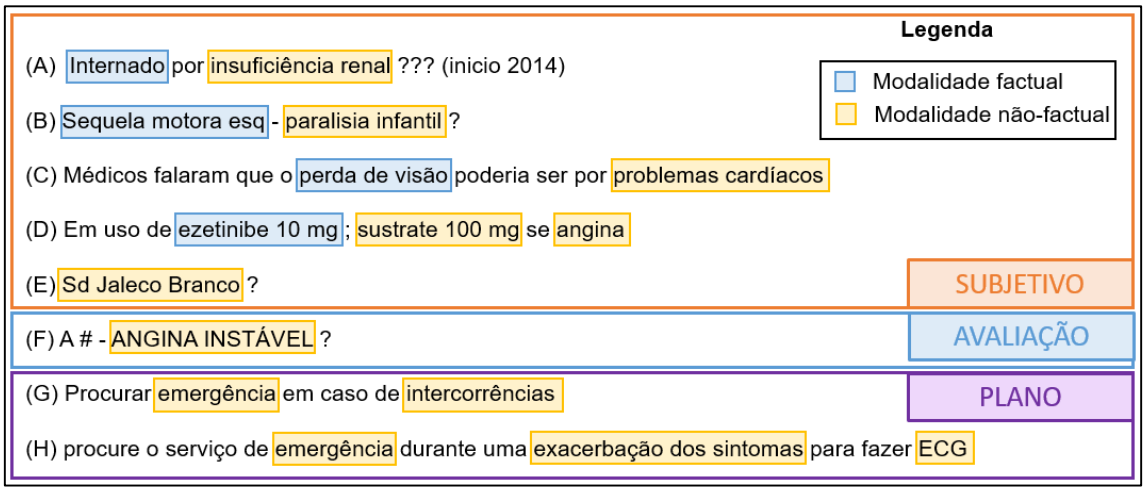
Na seção Planejamento (Plano), a polaridade se torna extremamente útil para indicar medicamentos suspensos pelo profissional de saúde, que têm polaridade negativa e são marcados no “futuro”, a fim de indicar que não existirão no futuro do paciente. Este caso é representado nos exemplos J e K. Vale salientar que, da mesma forma que “nega”, “suspendo” tem polaridade positiva, sendo somente um modificador.

MODALIDADE

A atributo Modalidade indica a “certeza” ou “incerteza” ligada a um EVT. A Polaridade tem relação com a ocorrência ou não de um EVT, já a Modalidade indica a certeza ou incerteza ligada àquele EVT, tendo uma relação direta com hipóteses. Suas marcações têm valor padrão de factual, sendo marcadas como não factual somente quando há “dicas textuais” para tal.

Na seção Subjetiva, existem marcações não factuais quando existe alguma incerteza do próprio paciente durante seu relato (como dúvida no motivo de uma internação) ou alguma incerteza do profissional de saúde sobre algum aspecto do relato do paciente. Na Figura 9, itens A até E, são mostrados alguns exemplos de marcação. No exemplo A, tem-se o caso de um internamento (factual), porém existe incerteza sobre o motivo, com uma hipótese de o motivo ser “insuficiência real”. No exemplo B, tem-se um problema factual, “sequela motora esq” (sequela motora esquerda), existindo uma hipótese de se dever à “paralisia infantil”. Similarmente, no exemplo C, há um caso de “perda de visão”, em que profissionais de saúde levantaram a hipótese de o motivo ser “problemas cardíacos”. No exemplo D, é mostrada a situação do uso de um medicamento caso algum problema específico. No exemplo E, existe uma hipótese de “Sd Jaleco Branco” (síndrome do jaleco branco), em que questões como controle pressórico são referidas como normais pelo paciente, porém apresentam alteração durante a consulta.

Figura 9 – Exemplos de marcações envolvendo atributo Modalidade, relacionando com seções do SOAP.



Fonte: O autor (2020).

Na seção Avaliação, existem marcações não factuais no momento em que há dúvidas em relação ao diagnóstico, não sendo algo totalmente certo. Na Figura 9F, existe um diagnóstico de possível “angina instável”.

Na seção Planejamento (Plano), as marcações não factuais envolvem EVTs condicionais, ou seja, “se determinado problema ocorrer, visite determinado local”. São instruções para casos de “possíveis” problemas que possam ser encontrados pelo paciente após a consulta. Nos exemplos G e H, existem menções condicionais: em caso de “intercorrências” ou “exacerbações dos sintomas”, buscar a “emergência”. No exemplo H, o profissional de saúde indica a busca da “emergência” para realizar um “ECG” (eletrocardiograma). No caso específico de um medicamento prescrito, sua marcação é factual, pois é “esperado” que o paciente faça uso dele.

RELTEMPDCD

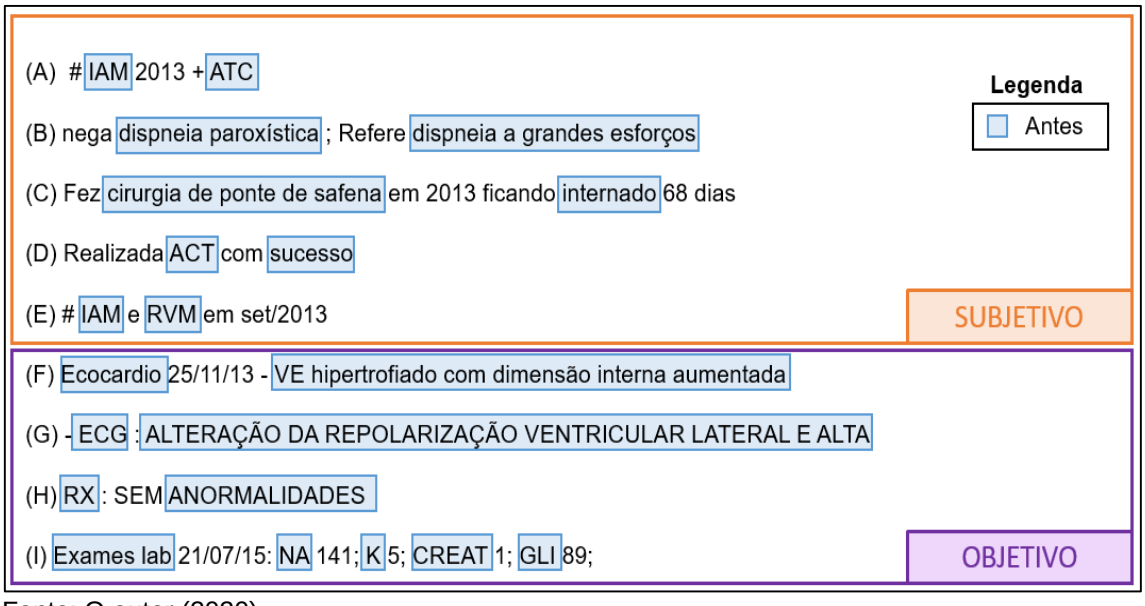
- **Antes**

Conhecer EVTs passados é essencial para o cuidado, trazendo informações importantes para o diagnóstico. Marcações com RelTempDCD de Antes estão usualmente presentes no histórico médico do paciente e em menções envolvendo testes realizados no passado dele (como ecocardiograma e cateterismo cardíaco). Quando não existe certeza de que um EVT se estende até a DCD, ou seja, sem

alguma regra específica para aquele caso ou alguma indicação textual no texto, a marcação correta é Antes.

Na seção Subjetiva, as marcações estão ligadas a menções do histórico do paciente, uma vez que, durante a entrevista, o profissional de saúde necessita obter informações importantes do que ocorreu no passado do paciente. Na Figura 10, exemplos A até E, são mostradas algumas marcações Antes. Nos exemplos A e E, tem-se casos de “IAM” (infarto agudo do miocárdio) mencionados no histórico do paciente e de tratamentos efetuados devido a esses problemas, “ATC” (angioplastia) e “RVM” (revascularização do miocárdio). No exemplo B, há diversos problemas negados ou referidos, salientando-se que, nesse contexto, o ato de “referir” ou “negar” ocorre no presente (marcação Sobreposto), porém o problema negado ou referido está relacionado ao passado do paciente. No exemplo C, existe menção de um tratamento passado, no caso, “cirurgia de ponte de safena”, indicando que o paciente ficou 68 dias “internado”, menção de Ocorrência. De forma similar, no exemplo D, foi realizada uma “ACT” (angioplastia), menção de Tratamento, que teve “sucesso”, menção de Ocorrência.

Figura 10 – Exemplos do atributo RelTempDCD com marcação Antes, relacionando com seções do SOAP.



Fonte: O autor (2020).

Na seção Objetiva, marcações Antes envolvem testes, assim como os problemas encontrados neles. Nos exemplos F até H, existem menções de testes, com seus respectivos resultados. No exemplo F, foi feito um “ecocardio”

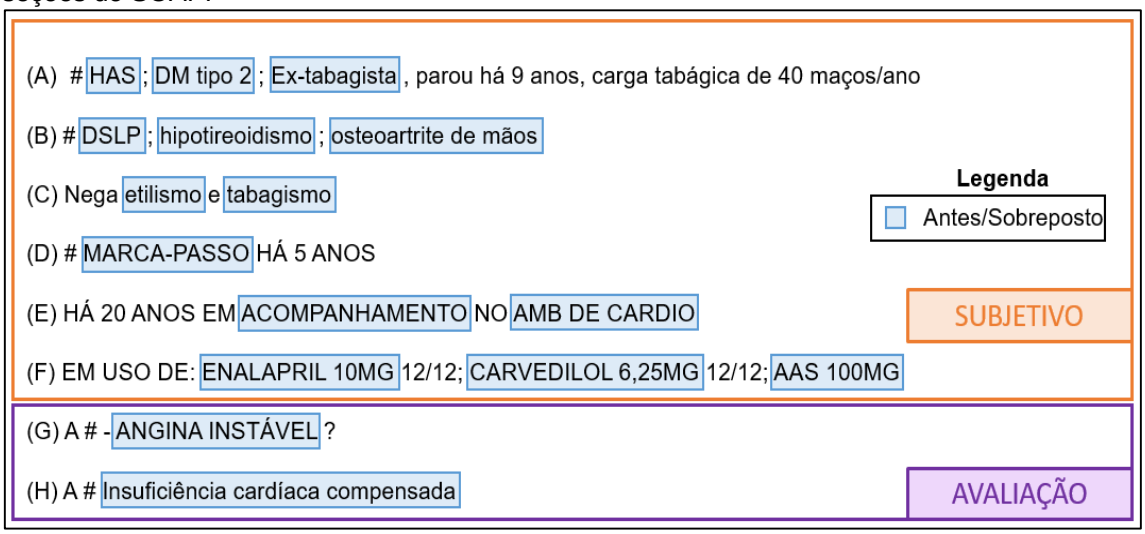
(ecocardiograma) e constatado “VE hipertrofiado com dimensão interna aumentada” (“VE” indicando ventrículo esquerdo). De forma similar, no exemplo G, houve uma alteração, menção de Problema, encontrada durante um exame de “ECG” (eletrocardiograma). No exemplo H, não foram encontradas “anormalidades” (marcado como um Problema com Polaridade negativa) durante o exame de imagem. Se houvesse alguma menção de Evidência, como “mostrou”, relacionando os testes aos problemas, essa menção também seria marcada como Antes. No exemplo I, tem-se um caso muito comum em textos de ambulatório, que são menções de exames de laboratório já realizados, assim como uma menção geral de “exames lab”, para englobar todos esses exames.

- **Antes/Sobreposto**

Marcações com RelTempDCD de Antes/Sobreposto estão usualmente presentes no histórico médico do paciente e no diagnóstico. Menções de doenças crônicas, medicamentos em uso, histórico social (tabagismo e etilismo) e comorbidades são sempre marcadas como Antes/Sobreposto. No entanto, os demais EVTs precisam ter um contexto de “continuidade” para poderem ser efetuadas marcações Antes/Sobreposto.

Na seção Subjetiva, tem-se casos de marcações de RelTempDCD com relações do tipo Antes/Sobreposto envolvendo o histórico médico do paciente, a partir das menções de medicamento em uso, doenças crônicas e comorbidades representadas, além de menções relacionadas ao histórico social, como etilismo e tabagismo. Essas questões são mostradas na Figura 11, itens A e B, com exemplos de “HAS”, “DM tipo 2”, “ex-tabagista” (também considerado um Problema), “DSLP”, “hipotireoidismo” e “osteoartrite de mãos”. No exemplo C, há casos de “etilismo” e “tabagismo” negados pelo paciente; assim, recebem Polaridade negativa. Um caso especial para textos cardiológicos é a marcação de menções de “marca-passo” (exemplo D) como Antes/Sobreposto devido à continuidade do tratamento. No exemplo E, “acompanhamento” e “amb de cardio” (ambulatório de cardiologia) são marcados como Antes/Sobreposto devido ao seu contexto. Um caso comum na seção Subjetiva é a menção de medicamentos em uso (exemplo F), sendo uma regra marcá-los como Antes/Sobreposto.

Figura 11 – Exemplos do atributo RelTempDCD com marcação Antes/Sobreposto, relacionando com seções do SOAP.



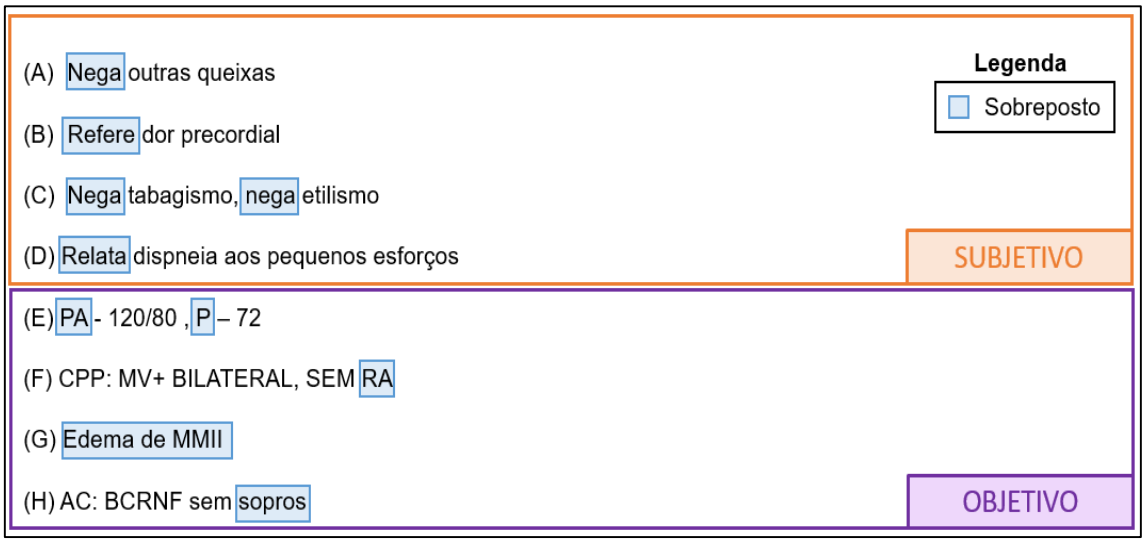
Fonte: O autor (2020).

Na seção Avaliação, geralmente só existem marcações de Problemas; devido a acontecerem do passado até o momento da consulta, quando são diagnosticados, são marcados como Antes/Sobreposto. Essas questões são mostradas na Figura 11, itens G e H, com menções de problemas diagnosticados, um deles (exemplo G) com atributo Modalidade com marcação não factual.

- **Sobreposto**

Marcações com RelTempDCD de Sobreposto estão associadas com EVTs relacionados à consulta, sendo uma marcação bem específica que reflete EVTs que ocorreram somente durante o encontro com o profissional de saúde. Na seção Subjetiva, essas marcações envolviam menções de “referir”, “negar” ou “relatar” vindas do paciente durante a entrevista do profissional de saúde. Alguns exemplos são fornecidos na Figura 12, dos itens A até D.

Figura 12 – Exemplos do atributo RelTempDCD com marcação Sobreposto, relacionando com seções do SOAP.



Fonte: O autor (2020).

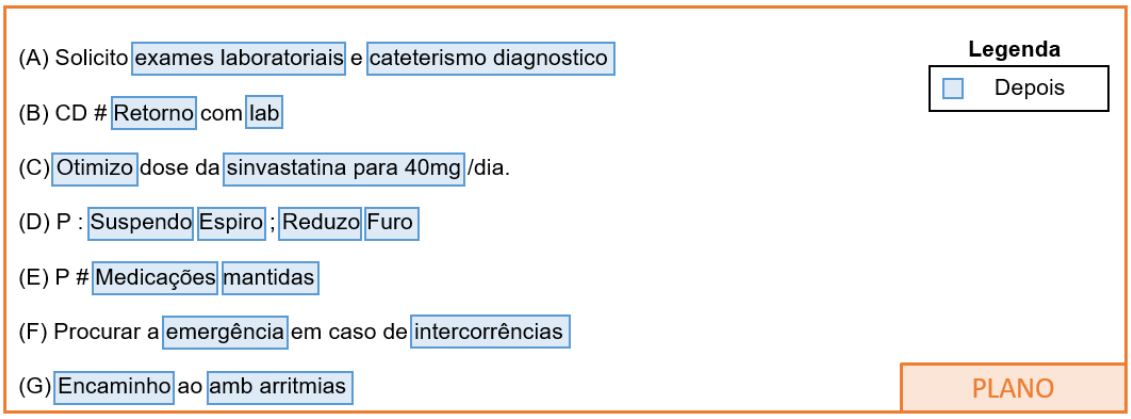
Na seção Objetiva, as marcações Sobreposto têm relação com testes e problemas encontrados durante o exame físico ou visual. Na Figura 12, há alguns exemplos nos itens E até H. No exemplo E, existem duas menções de testes realizados durante o exame físico, sendo “PA” (pressão) e “P” (pulso). Nos exemplos F e H, existem problemas como “RA” (ruídos adventícios) e “sopros” não sendo encontrados durante o exame físico, tendo marcações do atributo Polaridade como negativas. No exemplo G, foi encontrado um Problema durante o exame físico, no caso, um “edema de MMII” (edema de membros inferiores).

- **Depois**

Marcações com RelTempDCD de Depois estão relacionadas ao plano de ação proposto pelo profissional de saúde e se restringem a menções de Tratamentos, Testes, Ocorrências e Departamentos Clínicos no plano de ação. Este tipo de marcação se limita à seção Planejamento (Plano) do SOAP. Na Figura 13, tem-se exemplos de marcações de RelTempDCD como Depois. No exemplo A, há Testes, “exames laboratoriais” e “cateterismo diagnóstico”, que serão feitos e apresentados ao profissional de saúde na próxima consulta. No exemplo B, tem-se o caso de uma menção de “retorno” futuro com “lab” (exames laboratoriais). Nos exemplos C e D, existem menções de otimização, suspensão e redução marcadas como Depois, assim como os medicamentos envolvidos. No exemplo E, as “medicações” (medicações em uso citadas na seção Subjetiva) são “mantidas”, ou seja, estarão no futuro do paciente.

O exemplo F traz menções com marcações de modalidade não factual, existindo uma hipotética futura visita à “emergência” no caso de “intercorrências”. No exemplo G, tem-se um caso comum de encaminhamento do paciente a outro Departamento Clínico, neste caso, “amb arritmias” (ambulatório de arritmias).

Figura 13 – Exemplos do atributo RelTempDCD com marcação Depois, relacionando com seções do SOAP.



Fonte: O autor (2020).

APÊNDICE C – RESUMO DE *GUIDELINE* PARA ANOTAÇÃO DE EXPRESSÕES TEMPORAIS

TIPOS

- **Data**

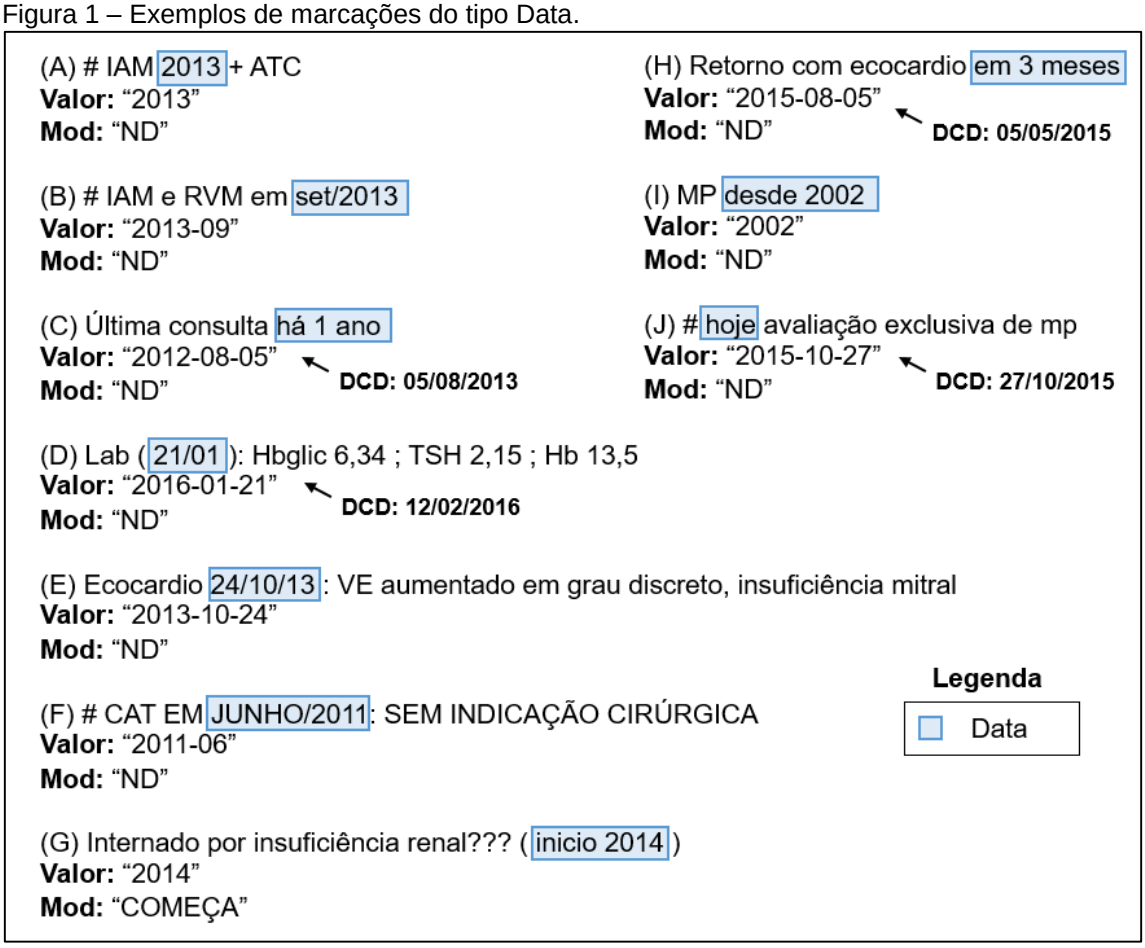
Datas são menções presentes em textos ambulatoriais. Como já mencionado, foram normalizadas, de acordo com a ISO 8601, no formato [YY-MM-DD], em que Y = ano, M = mês e D = dia. A normalização era representada pelo atributo nomeado Valor.

Existem menções de ETs do tipo Data na seção do histórico médico do paciente. Vale ressaltar que essas ETs usualmente têm a característica de serem vagas/incompletas, normalmente existindo somente a menção do ano ou do mês e ano, até porque, quando se está entrevistando o paciente sobre EVTs passados, a data relatada normalmente é aproximada. Essa questão é visualizada na Figura 1, exemplos A, B e G. No exemplo A, tem-se o uso de um “IAM” (infarto agudo do miocárdio) em 2013, tendo sido feita uma “ATC” (angioplastia) no mesmo ano. Devido a ser fornecido somente o ano, a normalização envolveu somente a especificação do ano. De forma similar, no exemplo B, há um caso de “IAM”, tendo sido feita “RVM” (revascularização do miocárdio); ambos os EVTs ocorreram em setembro de 2013. Dessa forma, a normalização envolveu somente mês e ano. Vale salientar que essa estrutura de “IAM” seguido da menção de algum tratamento, “ATC” ou “RVM”, é comum em textos cardiológicos. No exemplo G, tem-se um modificador (Mod) com valor igual a Começa, devido à palavra “início” trazer a informação de que o internamento ocorreu não apenas em 2014, mas, sim, no começo de 2014.

Menções de datas também estão presentes quando exames laboratoriais e outros testes/procedimentos diagnósticos são mencionados, existindo uma tendência de trazer ETs completas (dia, mês e ano) ou parcialmente completas (mês e dia). Isso acontece devido ao fato de que esses exames e testes/procedimentos vêm com a data preenchida no laudo, transcrita total ou parcialmente para o texto. Alguns casos são mostrados na Figura 1, exemplos D até F. No exemplo D, tem-se um caso de exame laboratorial, emq ue somente mês e dia estão presentes; neste caso, é usado o ano da DCD na normalização. Já no exemplo (F), há o caso de um exame, “CAT”

(cateterismo cardíaco), sendo fornecida informação sobre mês e ano, de modo que a normalização envolveu a especificação do mês e ano somente. Uma data completa é mostrada no exemplo E, com a normalização envolvendo especificação do dia, mês e ano.

Figura 1 – Exemplos de marcações do tipo Data.



Fonte: O autor (2020).

Existem também menções de ETs do tipo Data relativas, sendo necessário informação sobre a DCD para normalização. Esses casos são representados na Figura 1, exemplos C, H e J. No exemplo C, tem-se um caso comum, que seria a utilização do termo "há X anos" para tentar criar uma aproximação da data real. Vale salientar que se trata de um termo aproximado, vindo da informação obtida do paciente durante o relato. Neste caso, é calculada uma data um ano antes da DCD. No exemplo H, tem-se outro caso comum nos textos: especificações de retorno do paciente informadas na seção Planejamento (Plano) do SOAP pelo profissional de saúde. É usual nos textos de ambulatório existir uma indicação para o retorno, informada no padrão da quantidade de dias, meses ou anos a partir da atual consulta

(DCD). No exemplo H, é necessário calcular um período de três meses após a DCD para a normalização. Vale salientar que palavras como “há” e “em” sempre são mantidas nas ETs relativas, pois trazem informação sobre a relação entre as ETs. No exemplo J, tem-se um caso da menção “hoje”, sendo então normalizada para DCD.

Na Figura 1J, é mostrada a única exceção para a regra de Duração e Data estabelecida, nos casos em que “MP” (marca-passo) tem RelTempDCD de Antes/Sobreposto, porém, como ET envolve o termo “desde”, foi considerado Data e a RT tinha a função de especificar o aspecto da continuidade, seguindo, assim, o padrão para anotação de RTs de marcações envolvendo “desde” sugerido no *guideline* do THYME.

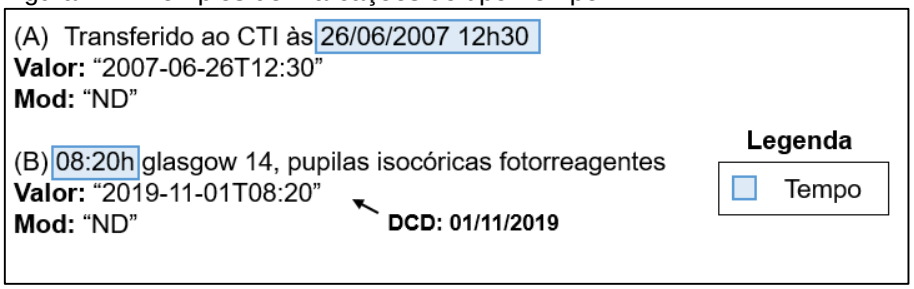
- **Tempo**

Menções de horários específicos no dia são comuns em textos de evolução de enfermagem, com diversas observações feitas sobre o estado do paciente e tratamentos aplicados. No contexto desta tese, com textos de ambulatório de cardiologia, este tipo de menção é extremamente raro. Durante toda a avaliação de textos de cardiologia, não foi encontradi nenhum texto com menções de horários, porém são detalhados alguns detalhes principais.

A normalização deste tipo de menção também é baseada na ISO 8601, porém a normalização é no formato [YYYY-MM-DD]T[HH:MM], com Y = ano, M = mês, D = dia, H = hora, M = minuto.

Na Figura 2, são mostrados dois exemplos de marcação. No exemplo A, tem-se um caso completo, em que tanto a data quanto a hora são fornecidas. No exemplo B, tem-se somente a menção do horário, sendo utilizada a DCD para normalizar (adicionada como informação referente ao dia).

Figura 2 – Exemplos de marcações do tipo Tempo.



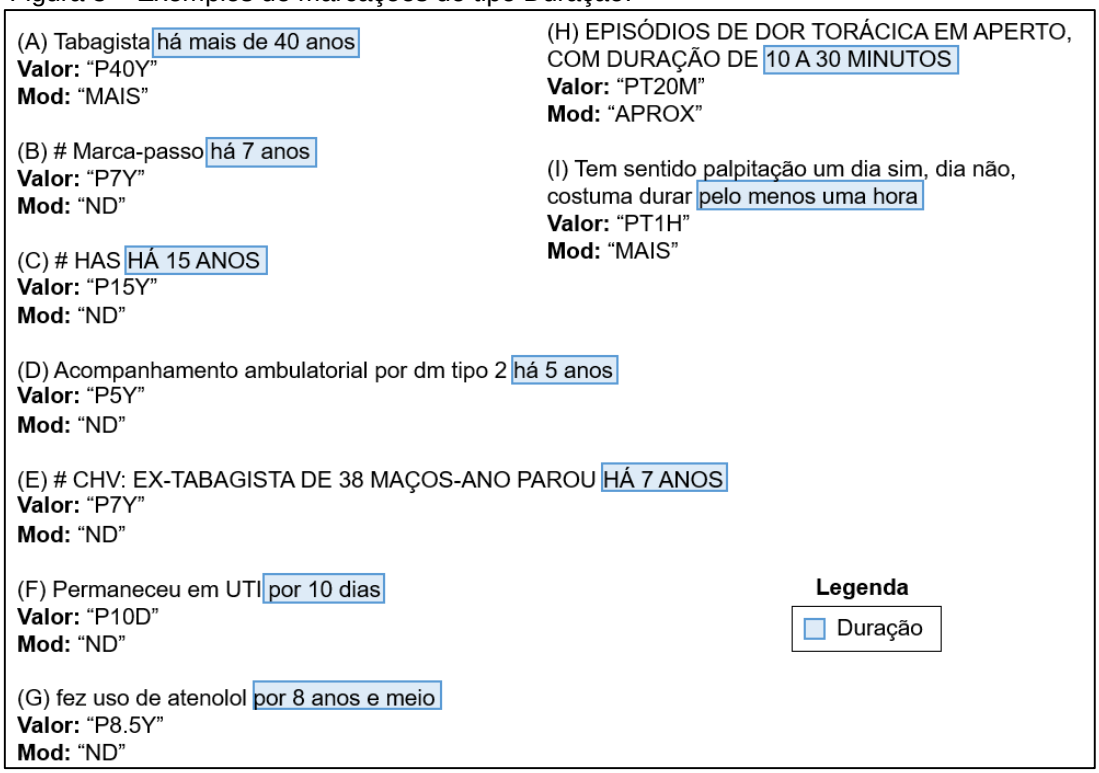
Fonte: O autor (2020).

- **Duração**

Menções de Duração são importantes para extração de RTs. Quando EVTs e ETs do tipo Duração são relacionados por meio de RTs, as ETs complementam os EVTs com informações acerca das suas extensões. O caso mais frequente de utilização de ET do tipo Duração é para descrição da duração de doenças crônicas, comorbidades, acompanhamentos contínuos, tratamentos contínuos (como marca-passo) e menções envolvendo tabagismo e etilismo, existindo uma tendência de utilizar o termo “há” na menção, pois é algo que se mantém do passado até a DCD.

Na Figura 3, há alguns casos representados nos exemplos A até E. Em todos, os EVTs relacionados têm marcação do atributo RelTempDCD como Antes/Sobreposto, indicando sua continuidade. Os EVTs envolvidos em cada exemplo são: (A) “tabagista”, (B) “marca-passo”, (C) “HAS” (hipertensão arterial sistêmica), (D) “acompanhamento ambulatorial” e (E) “ex-tabagista”. Por exemplo, a normalização para “há 15 anos” obedeceu a um padrão de “P” (indicando período), seguido do valor (7) e da unidade (“Y” para anos). No exemplo A, o modificador (Mod) teve marcação Mais devido à menção indicar uma duração superior a 40 anos.

Figura 3 – Exemplos de marcações do tipo Duração.



Fonte: O autor (2020).

Marcações do tipo Duração também envolvem durações relacionadas a EVTs passados ou futuros, como, por exemplo, com a utilização do termo “por”. Na Figura 16, os exemplos F e G representam essa questão. No exemplo G, tem-se um caso de menção do uso de um medicamento por determinado tempo, no caso, “atenolol” por um período de oito anos e meio. Em cenário similar, pode-se ter um caso de uso futuro de um medicamento por certo período, sendo uma menção pertencente à prescrição. No exemplo F, existe uma menção de “UTI” (unidade de terapia intensiva), indicando que o paciente permaneceu por um período de dez dias.

Como já mencionado, no OLD CARTS, tem-se um atributo do sintoma relacionado à sua duração. Para este projeto, a duração de um sintoma é marcada como um ET do tipo Duração; em seguida, por meio de RTs, essa ET é relacionada com o sintoma (Problema). Na Figura 3, exemplos H e I, essa situação é representada, com ambos os sintomas descritos com uma duração. Vale salientar que a frequência de ocorrência “um dia sim, dia não” do exemplo I seria uma ET do tipo Frequência.

- **Frequência**

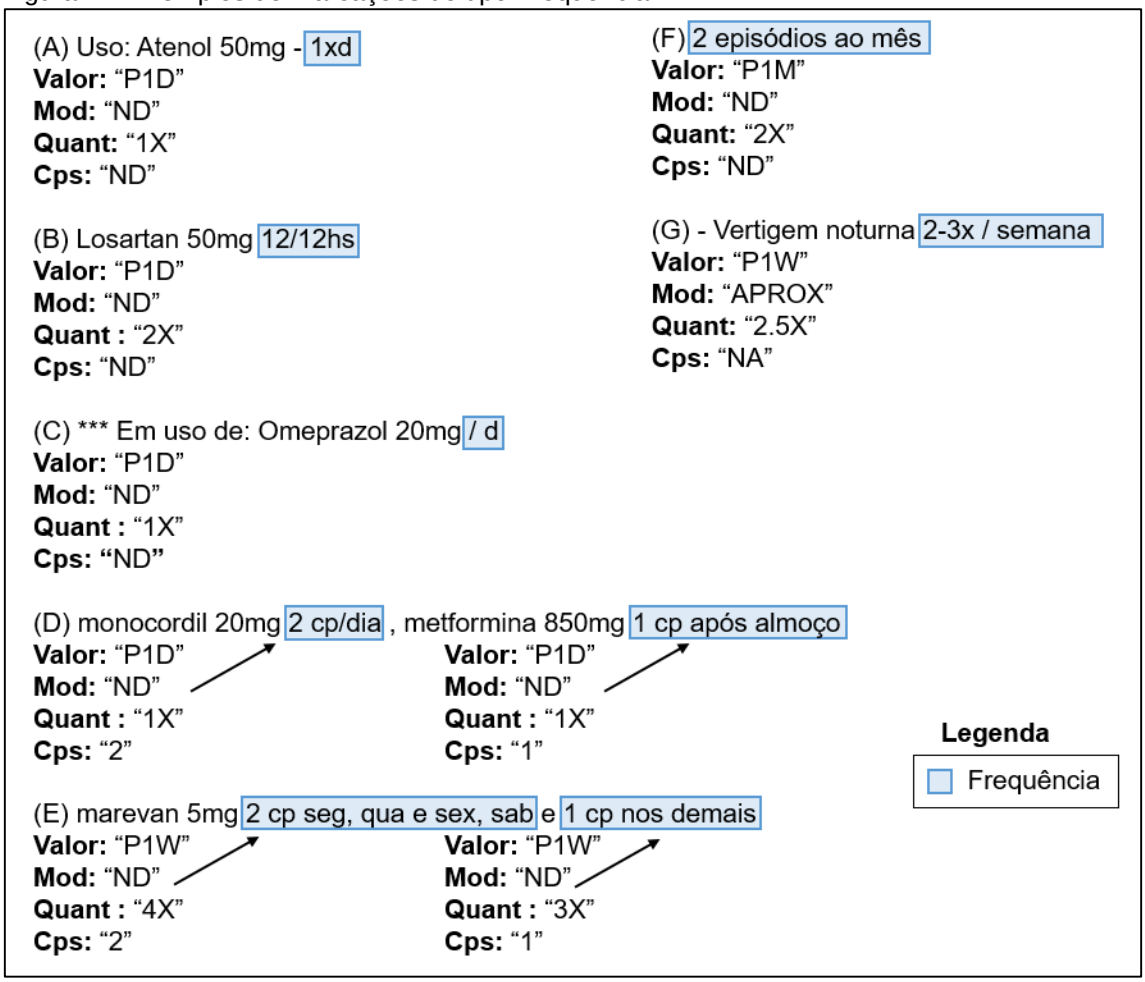
Marcações de ET do tipo Frequência têm relação com a frequência de determinado EVT, como, por exemplo, quantas vezes por dia determinado medicamento é usado pelo paciente. É um tipo de ET que, no contexto de textos ambulatoriais, usualmente está associado com a frequência de uso de medicamentos. Pode também estar associado com periodicidade de sintomas ou de outros aspectos relevantes no texto.

Os trechos relativos a medicamentos em uso normalmente contêm diversas menções deles (como “furosemida 40 mg”), cada qual com sua respectiva frequência de uso associada (como “8/8horas”). Como mencionado por Styler *et al.* (2014a), especificações temporais adicionais não são incomuns (como “3 vezes por dia durante as refeições”). Essas especificações adicionais acabam envolvendo detalhes específicos sobre refeições, dias da semana, mudança de dosagem e/ou quantidade de comprimidos.

Exemplos de marcações do tipo Frequência para medicamentos são mostrados na Figura 4, itens A até E. No exemplo A, tem-se o caso de um medicamento usado uma vez por dia, por isso o atributo Valor é marcado como “P1D” (período de um dia)

e o atributo Quant, como “1X” (uma vez). No exemplo C, há a menção “/ d”, que é semelhante a “1xd”, tendo as mesmas marcações de atributos. No exemplo B, tem-se um caso de uso de medicação a cada 12 horas em um período de um dia, totalizando duas vezes ao dia (Quant de “2X”). No exemplo D, é mostrado o atributo Cps, recebendo valores diferentes do padrão (ND), pois existem menções claras de comprimidos. No exemplo E, tem-se um caso complexo de marcação, existindo ETs envolvendo comprimidos e dias da semana. No caso de menções em que os comprimidos variam de acordo com o dia da semana, trabalha-se com o período de uma semana e leva-se em conta o total de dias. Assim, há duas ETs do tipo Frequência com quantidades diferentes de comprimidos relativas ao mesmo medicamento.

Figura 4 – Exemplos de marcações do tipo Frequência.



Fonte: O autor (2020).

Casos de ETs do tipo Frequência ligadas a sintomas são mostrados na Figura 4, exemplos F e G. Além de exibir ETs do tipo Duração, certos sintomas podem trazer informações sobre Frequência.

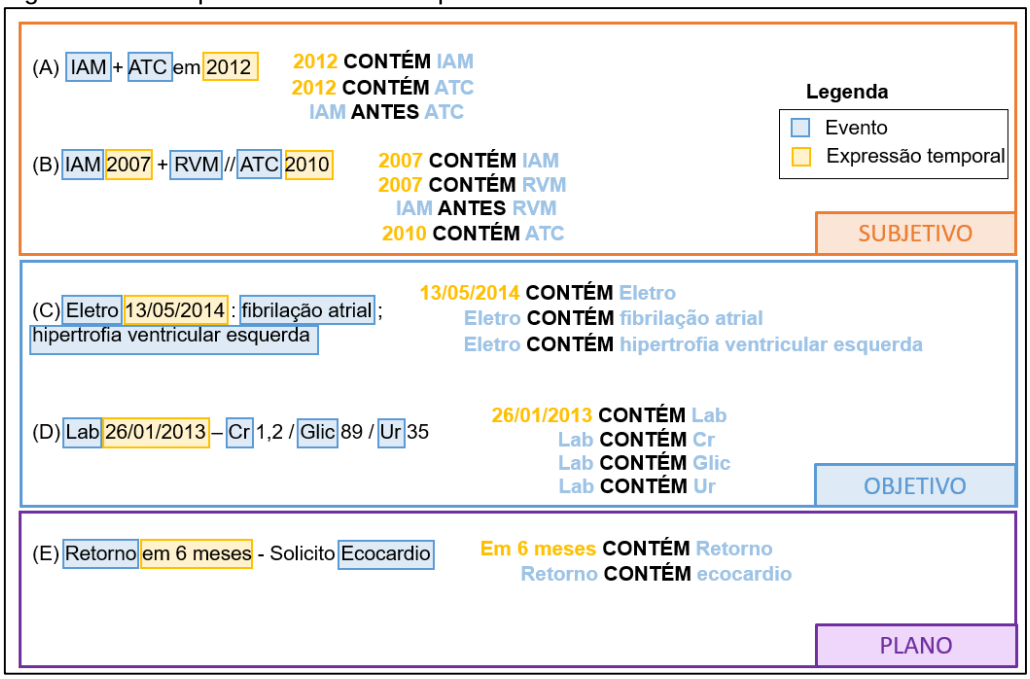
APÊNDICE D – RESUMO DE *GUIDELINE* PARA ANOTAÇÃO DE RELAÇÕES TEMPORAIS

- **Contém**

Relações do tipo Contém são essenciais no contexto clínico, existindo uma tendência de escrita de usar EVTs ou ETs dos tipos Data e Tempo como referências centrais. Inclusive, ETs dos tipos Data e Tempo são excelentes referências para "conter" outros EVTs, pois se referem a pontos específicos no tempo.

Há alguns elementos presentes nos textos ambulatoriais que são oportunos para utilizar relações do tipo Contém. Na seção Subjetiva do SOAP, existe o histórico médico do paciente, com uma tendência de localizar os EVTs temporalmente com base em menções de ET do tipo Data, apesar de, em certos casos, serem menções incompletas ou relativas. Na Figura 1, há dois casos, apresentados nos exemplos A e B. No exemplo A, tem-se o caso de uma ET do tipo Data ("2012"), contendo tanto o problema "IAM" (infarto agudo do miocárdio) quanto "ATC" (angioplastia), tratamento utilizado. Vale salientar que, como "IAM" veio antes da "ATC", é marcada uma relação do tipo Antes. Similarmente, no exemplo (B), o tratamento utilizado é "RVM" (revascularização do miocárdio). Adicionalmente, foi realizada uma "ATC" em "2010", evidenciada pela relação do tipo Contém.

Figura 1 – Exemplos de TLINKs do tipo Contém.



Fonte: O autor (2020).

Na seção Objetiva, existem menções de exames de laboratório e de diagnóstico. Todas as menções de exames específicos de laboratório realizados (como creatinina) têm uma tendência de ser localizadas temporalmente com base em menções gerais de “exames de laboratório” e ETs do tipo Data, normalmente completas. Essa questão pode ser vista na Figura 1D, com a ET do tipo Data contendo a menção de exame geral “Lab”, que, por sua vez, contém os exames de laboratório.

De forma similar, problemas encontrados em exames de diagnóstico têm uma tendência de ser localizados temporalmente com base nas menções dos exames de diagnóstico e ETs do tipo Data, normalmente completas. Na Figura 1C, tem-se um caso de marcação, com a ET do tipo Data contendo “eletro” (eletrocardiograma) e esta menção contendo os problemas encontrados.

Na seção Planejamento, há EVTs, como exames, relacionados com menções de retornos futuros e ET do tipo Data relativas (como “em 30 dias”). Na Figura 1E, tem-se um caso deste tipo, com a ET do tipo Data “em 6 meses” contendo “retorno”, que, por sua vez, contém “ecocardio” (ecocardiograma).

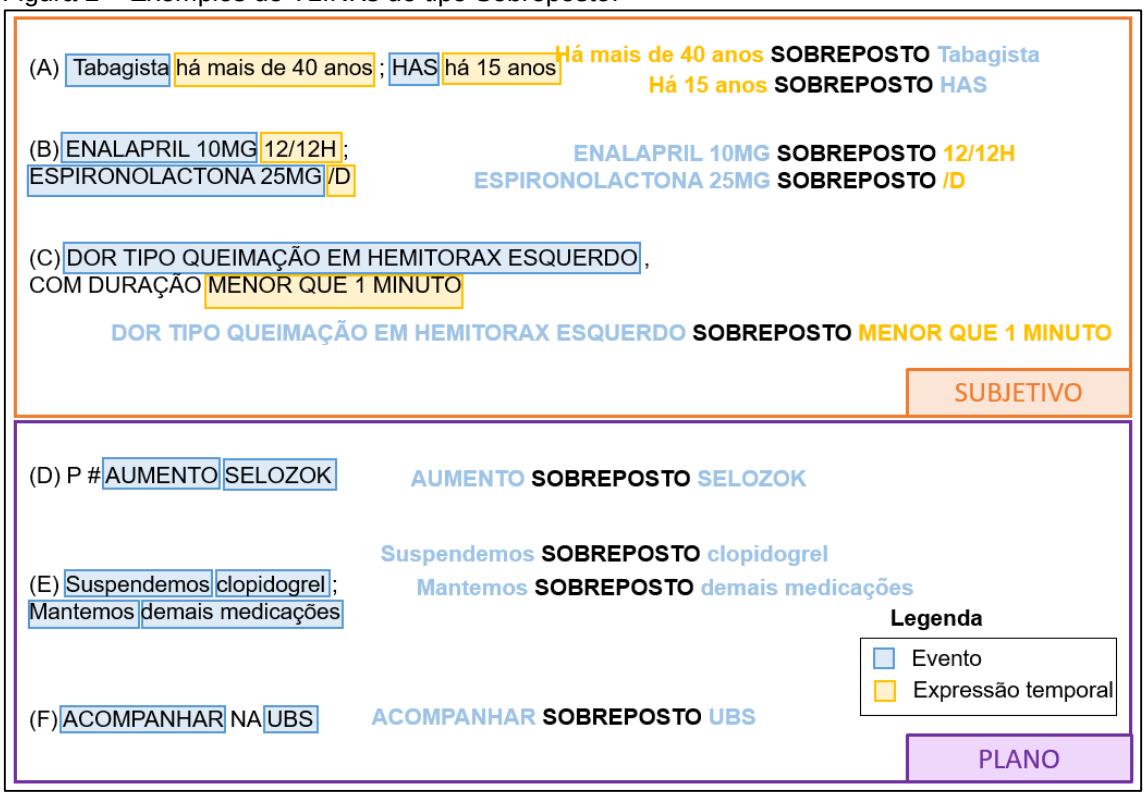
- **Sobreposto**

Com as relações do tipo Contém, são as mais presentes nos textos clínicos. As relações do tipo Sobreposto são usadas em diversos contextos dentro do texto clínico e, apesar de serem genéricas, fornecem informações importantes. São usadas na seção Subjetiva do SOAP na questão do histórico médico do paciente, relacionando questões contínuas, como doenças crônicas, comorbidades e acompanhamentos, com suas respectivas durações, como mostrado na Figura 2A, com as relações das menções “tabagista” e “HAS” (hipertensão arterial sistêmica) com suas respectivas durações em anos, representadas por ETs do tipo Duração.

Outra questão importante dentro da seção Subjetiva são os medicamentos em uso, até porque pacientes costumam utilizar diversos medicamentos. Neste caso, são marcadas relações do tipo Sobreposto entre os medicamentos e suas respectivas frequências de uso. Este tipo de marcação é mostrado na Figura 2B, com ambos os medicamentos, “enalapril 10 mg” e “espironolactona 25 mg”, relacionadaos com suas frequências de uso (ETs do tipo Frequência). Durante o relato do paciente, marcações

do tipo Sobreposto podem ser usadas para relacionar o problema relatado com demais questões do relato, como a duração do sintoma, como mostrado na Figura 2C.

Figura 2 – Exemplos de TLINKs do tipo Sobreposto.



Fonte: O autor (2020).

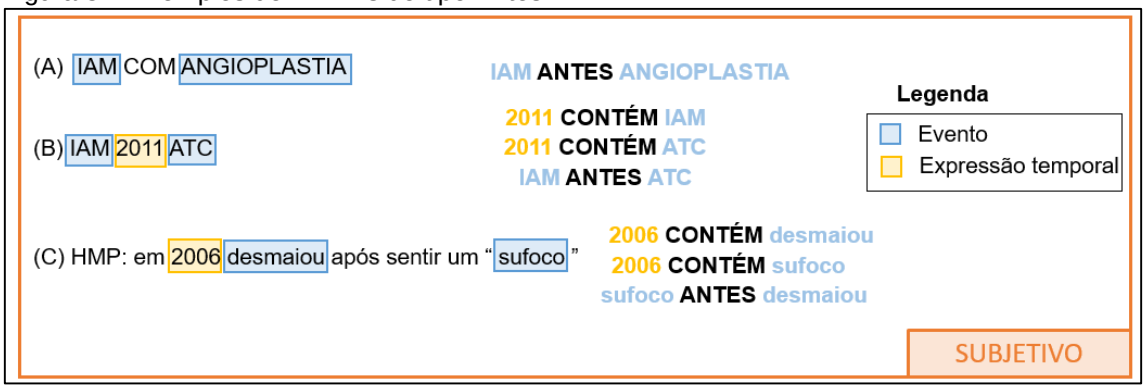
As marcações Sobreposto são muito utilizadas na seção Planejamento, relacionando menções de ocorrências como “aumento”, “suspendemos” e “mantemos” com suas respectivas medicações, como mostrado na Figura 2, exemplos D e E. Vale salientar que, nesse contexto, “clopidogrel” é um EVT com Polaridade negativa, pois está sendo suspenso. Tem-se um exemplo de acompanhamento que será realizado na “ubs” (unidade básica de saúde), mostrado no exemplo F.

- **Antes**

Marcações do tipo Antes têm a função de indicar ordem entre menções. Palavras como “pós”, “após” e “seguido” ajudam a indicar casos dessas menções. Na Figura 3, tem-se alguns exemplos de relações do tipo Antes. Nos exemplos A e B, há casos comuns em textos de cardiologia: a menção de problema, “IAM” (infarto agudo do miocárdio), seguido do tratamento utilizado, angioplastia, em ambos os casos. Vale destacar que, nos casos apresentados, é necessário conhecimento do profissional de

saúde para determinar que “IAM” veio antes do tratamento. No exemplo C, tem-se um caso de desmaio após um sufoco, sendo a palavra “pós” uma dica para marcar Antes.

Figura 3 – Exemplos de TLINKs do tipo Antes.



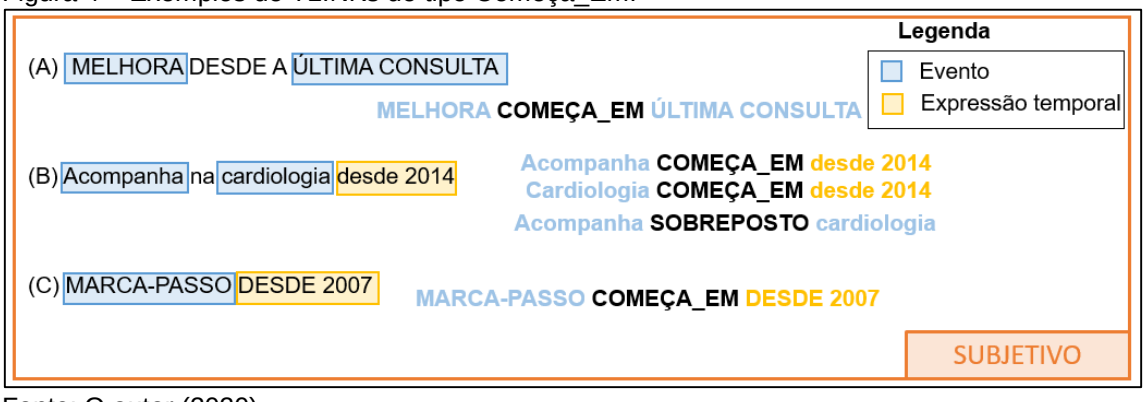
Fonte: O autor (2020).

- **Começa_Em**

Marcações do tipo Começa_Em indicam que o EVT começa ou no EVT ou na ET a que está relacionado. Isso normalmente está relacionado com marcações de ETs menos detalhadas (com informações de mês e/ou dia faltantes). Qualquer tipo de relação com marcação pontual (ET bem definida) pode ser do tipo Antes.

Alguns exemplos de marcações são mostrados na Figura 4. As do tipo Começa_Em normalmente estavam associadas com o uso da expressão “desde”, usualmente envolvendo menções de “marca-passo” e “acompanhamento”.

Figura 4 – Exemplos de TLINKs do tipo Começa_Em.



Fonte: O autor (2020).

- **Termina_Em**

Marcações do tipo Termina_Em indicam que o EVT termina ou no EVT ou na ET a que está relacionado. Isso normalmente está relacionado com marcações de

ETs menos detalhadas (com informações de mês e/ou dia faltantes). Qualquer tipo de relação com uma marcação pontual (ET bem definida) pode ser do tipo Antes.

Usualmente, marcações do tipo Termina_Em eram transformadas em marcações do tipo Começa_Em.

APÊNDICE E – *FEATURES* UTILIZADAS PELOS CLASSIFICADORES

Neste apêndice, são definidas e explicadas todas as *features* relacionadas aos experimentos de extração, sendo que cada componente de extração tem suas próprias *features* e autores-base.

- ***Features* para TLINKs entre eventos em mesma sentença**

Os conjuntos de *features* utilizados para este tipo de TLINK estão representados no Quadro 1. Dentre elas, “Contexto Rep 1” e “Contexto Rep 2” envolvem a obtenção do contexto próximo de cada menção por uso de n-gramas. Na *feature* “Contexto Rep 1”, ocorre a representação por unigramas e BOW e, na *feature* “Contexto Rep 2”, por unigramas, bigramas e trigramas. Com uma abordagem baseada em unigramas, se perde muito do significado da ordem das palavras, sendo possível somente identificar sua presença no contexto (LANE; HOWARD; HAPKE, 2019). Adicionando bigramas e trigramas, certo aspecto do significado é mantido, porém a maioria dos bigramas é rara, ainda mais trigramas, criando matrizes ainda mais esparsas (LANE; HOWARD; HAPKE, 2019). Essas *features* visam a verificar o efeito de diferentes n-gramas de contexto para este tipo de texto clínico. Adicionalmente, na Representação 2 (Rep 2), foi adicionado POS pelo *POS*-tagger treinado para textos clínicos no trabalho de Ferro *et al.* (2020).

Quadro 1 – Lista de *features* para TLINKs entre EVTs em mesma sentença.

Feature	**Descrição**	**Autores-base**
Contexto Rep 1	*Tokens* ao redor dos EVTs em uma janela de três *tokens* (três antes e três depois). Representação por BOW. Cada EVT tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2018), Xu *et al.* (2013)
Contexto Rep 2	*Tokens* e POS ao redor dos EVTs em uma janela de dois *tokens* (dois antes e dois depois). Representação por unigramas, bigramas e trigramas. Cada EVT tem sua própria representação.	Tang *et al.* (2013)
Menção Rep 1	*Tokens* e POS dos EVTs. Representação por BOW. Cada EVT tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2016, 2018)
Menção Rep 2	*Tokens* dos EVTs. Representação por BOW. Cada EVT tem sua própria representação.	MacAveney, Cohan e Goharian (2017)
Tenso	Tenso de todos os verbos finitos da sentença. Representação por BOW.	Lee *et al.* (2016, 2018), Lin *et al.* (2016a), Tang *et al.* (2013)

Tokens dentre	*Tokens* dentre (no meio de) as menções de EVTs. Representação por BOW.	Lin *et al.* (2016a), Cherry *et al.* (2013), Xu *et al.* (2013), Lee *et al.* (2018)
Atributos	Atributos Tipo, Polaridade, Modalidade e RelTempDCD de cada EVT. Representados por meio de *one-hot encoding*. Cada atributo de cada EVT tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2016, 2018), MacAveney, Cohan e Goharian (2017), Mirza e Tonelli (2016)
Concatenação de atributos	Concatenação dos atributos dos EVTs: Tipo com Tipo, Polaridade com Polaridade, Modalidade com Modalidade e RelTempDCD com RelTempDCD. A ordem do elemento na relação influencia a concatenação. Representada por meio de *one-hot encoding*.	Xu *et al.* (2013), Cherry *et al.* (2013), MacAveney, Cohan e Goharian (2017), Mirza e Tonelli (2016)
Número de *tokens*	Contagem do número de *tokens* presentes entre as menções de EVT. Representado por valores numéricos.	Lin *et al.* (2016a), Tang *et al.* (2013), Lee *et al.* (2018)
Número de menções	Contagem do número de menções (EVTs e ETs) entre as menções de EVTs. Representado por valores numéricos.	Lee *et al.* (2016), Tang *et al.* (2013), Cherry *et al.* (2013), Mirza e Tonelli (2016)
UMLS	Tipos semânticos de UMLS relacionados a cada EVT. Representação por BOW. Cada EVT tem sua própria representação.	Lin *et al.* (2016), Cherry *et al.* (2013), Xu *et al.* (2013)
Posição	Posição dos EVTs no texto, calculada a partir da divisão da posição inicial pela quantidade de caracteres do texto. Representada por valores numéricos. Cada EVT tem sua própria representação.	Tang *et al.* (2013)
Termos Antes/Depois	Busca por termos na sentença que possam indicar que um EVT veio antes de outro. Termos predefinidos em um dicionário criado. Representados por um valor binário, indicando que algum dos termos foi encontrado ou não.	-

Fonte: O autor (2020).

O conjunto de *features* “Menção Rep 1” e “Menção Rep 2” envolve a obtenção dos *tokens* dos EVTs por uso de BOW. Como EVTs podem ter múltiplos *tokens*, como, por exemplo, “dispneia em repouso”, é utilizada uma representação por meio do BOW. Se outro EVT na sentença fosse uma menção de “dispneia”, por exemplo, a informação contida na representação poderia auxiliar a identificar uma correferência, que deveria ser marcada como um TLINK do tipo Sobreposto. Além da representação do *token*, é proposto na *feature* “Menção Rep 2” trazer uma representação adicional do POS dos *tokens* dos EVTs; por exemplo, trazer a informação de que uma menção de EVT do tipo Ocorrência se refere a um verbo, como “manter”, pode auxiliar na questão de definir um TLINK entre essa ocorrência e um medicamento.

As *features* “Tenso”, “*Tokens* dentre”, “Número de *tokens*” e “Número de menções” buscam obter informações sobre o contexto entre as menções, sendo as últimas duas relacionadas a informações quantificadas. A *feature* “Tenso” obtém o

tenso dos verbos finitos na sentença, determinado por meio do Stanza (QI *et al.*, 2020).

Das demais *features*, ressalta-se “Concatenação de atributos”. A *feature* “Atributos” busca trazer informações sobre as menções, porém visa a descrever o contexto da relação. Ela se torna ainda mais interessante em um cenário em que a ordem da menção na relação é mantida. Por exemplo, se um EVT do tipo Teste apresenta um TLINK do tipo Contém com um EVT do tipo Problema, trazer uma concatenação dos atributos como “teste_problema” pode ajudar a indicar uma possível relação.

A *feature* “UMLS” é uma representação por meio de BOW dos tipos semânticos da UMLS, tendo cada tipo semântico relacionado ao EVT uma representação. É interessante uma representação por BOW, devido ao fato de um EVT poder ter diversos tipos semânticos associados. Foi feita uma busca por um dicionário criado a partir da UMLS, tendo ocorrido algumas alterações devido à ambiguidade de termos. Partiu-se do termo completo, em casos de termo composto, reduzindo sua expansão até alguma correspondência. Por exemplo, o termo “dispneia em repouso” seria buscado como “dispneia em repouso”, “dispneia em” e “dispneia”. Se ainda não houvesse correspondência, passaria para o próximo termo do EVT, neste caso, “em repouso” e “em”. Se ainda não houvesse correspondência, passaria para o próximo termo do EVT, no caso, “repouso”. Se, após todos esses passos, não fosse encontrada alguma correspondência, a coluna Tipo Semântico receberia o valor “SemUMLS”.

A *feature* “Posição”, proposta por Tang *et al.* (2013), se baseou nas seções do documento, porém para esses textos não existe tal informação. Por isso, foi calculada apenas a posição do EVT no documento, sendo um valor percentual entre 0 e 1.

Por último, a *feature* “Termos Antes/Depois” foi proposta pelo doutorando para melhor definir casos de TLINKs do tipo Antes. Certas palavras e símbolos específicos indicam um TLINK desse tipo. Nesse cenário, foi testado trazer uma *feature* binária indicando a presença ou não desses termos na frase. Os termos foram levantados conforme uma busca no conjunto de treinamento, sendo eles: “para”, “depois”, “por”, “após”, “apos”, '+', “com”, “foi”, “=>“, “->“, “antes”, “pós”, “pos”, “==>“, “-->“, “===>“ e “-->”.

- ***Features* para TLINKs entre eventos e expressões temporais em mesma sentença**

Os conjuntos de *features* utilizados para este tipo de TLINK estão representados no Quadro 2, com suas respectivas descrições e autores-base. Foram utilizadas *features* propostas nos estudos de melhor desempenho para os *corpora* Clinical TempEval 2015, 2016 e 2017 e i2b2 2012. A maior parte foi detalhada anteriomente, por isso as explicações serão direcionadas às particularidades encontradas neste conjunto.

Quadro 2 – Lista de *features* para TLINKs entre EVT e ET em mesma sentença.

Feature	**Descrição**	**Autores-base**
Contexto Rep 1	*Tokens* ao redor das menções (EVT e ET) em uma janela de três *tokens* (três antes e três depois). Representação por BOW. Cada menção tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2018), Xu *et al.* (2013)
Contexto Rep 2	*Tokens* e POS ao redor das menções (EVT e ET) em uma janela de dois *tokens* (dois antes e dois depois). Representação por unigramas, bigramas e trigramas. Cada menção tem sua própria representação.	Tang *et al.* (2013)
Menção Rep 1	*Tokens* do EVT e da ET. Representação por BOW. Cada menção tem sua própria representação.	Lin *et al.* (2016a), MacAveney, Cohan e Goharian (2017)
Menção Rep 2	*Tokens* e POS do EVT e da ET. Representação por BOW. Cada menção tem sua própria representação.	Lee *et al.* (2016, 2018)
Tenso	Tenso de todos os verbos finitos da sentença. Representação por BOW.	Lee *et al.* (2016, 2018), Lin *et al.* (2016a), Tang *et al.* (2013)
Tokens dentre	*Tokens* dentre (no meio de) as menções. Representação por BOW.	Lin *et al.* (2016a), Cherry *et al.* (2013), Xu *et al.* (2013), Lee *et al.* (2018)
Atributos	Atributos relacionados ao EVT: Tipo, Polaridade, Modalidade e RelTempDCD. Atributo relacionado à ET: Tipo. Representados por meio de *one-hot encoding*. Cada atributo de cada menção tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2018), Mirza e Tonelli (2016)
Concatenação do tipo	Concatenação dos atributos dos EVTs: Tipo com Tipo, Polaridade com Polaridade, Modalidade com Modalidade e RelTempDCD com RelTempDCD. A ordem do elemento na relação influencia a concatenação. Representada por meio de *one-hot encoding*.	Xu *et al.* (2013), Cherry *et al.* (2013), MacAveney, Cohan e Goharian (2017)
Número de *tokens*	Contagem do número de *tokens* presentes entre as menções de EVTs. Representado por valores numéricos.	Lin *et al.* (2016a), Tang *et al.* (2013), Lee *et al.* (2018)
Número de menções	Contagem do número de menções (EVTs e ETs) entre as menções de EVTs. Representado por valores numéricos.	Lee *et al.* (2016), Tang *et al.* (2013), Cherry *et al.* (2013), Mirza e Tonelli (2016)

ET começo	Indicação se a ET está no começo da sentença. Representada por um valor binário.	Xu *et al.* (203)
Par mais próximo	Indicação do par mais próximo entre todos os pares da sentença. Representado por um valor binário.	Lin *et al.* (2016a)
Posição	Posições das menções no texto, calculadas a partir da divisão da posição inicial pela quantidade de caracteres do texto. Representada por valores numéricos. Cada menção tem sua própria representação.	Tang *et al.* (2013)
Conjunções	Verifica se há conjunções entre as menções. As conjunções foram predefinidas em um dicionário. Representadas por um valor binário.	Lin *et al.* (2016a), Tang *et al.* (20133), Cherry *et al.* (2013)

Fonte: O autor (2020).

Para este tipo de TLINK, foi criada uma *feature* concatenando os tipos de EVTs e ETs, mantendo a ordem da menção na relação. Alguns tipos de ET estão usualmente relacionados a determinados tipos de EVT. Por exemplo, um medicamento (EVT do tipo Tratamento) normalmente está associado a uma menção de frequência de uso (ET do tipo Frequência), sendo, assim, uma *feature* oportuna.

Foram adicionadas duas *features* envolvendo a distância. A primeira, “ET começo”, contempla a distância da ET até o começo da sentença, verificando se ocorre dentro dos quatro primeiros *tokens* desta, sendo uma *feature* importante em menções de exames laboratoriais e de diagnóstico. A segunda, “Par mais próximo”, verifica se a distância entre o par é a menor na sentença; usualmente, ETs e EVTs próximos tendem a formar TLINKs.

- ***Features* para TLINKs entre eventos em sentenças distintas**

Os conjuntos de *features* utilizados para este tipo de TLINK estão representados no Quadro 3, com suas respectivas descrições e autores-base. As utilizadas para extração de TLINKs entre EVTs em sentenças distintas foram similares às usadas para EVTs em mesma sentença, tendo havido certa dificuldade de obter *features* específicas para este tipo de TLINK, uma vez que, em boa parte dos trabalhos envolvendo o THYME *corpus*, ele é, inclusive, ignorado.

As *features* utilizadas são mostradas no Quadro 3, com suas respectivas descrições e autores-base. A maior parte já foi discutida, por isso as explicações serão direcionadas às particularidades encontradas neste conjunto.

Quadro 3 – Lista de *features* para TLINKs entre EVTs em sentenças distintas.

Feature	**Descrição**	**Autores-base**
Contexto Rep 1	*Tokens* ao redor dos EVTs em uma janela de três *tokens* (três antes e três depois). Representação por BOW. Cada EVT tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2018), Xu *et al.* (2013)
Contexto Rep 2	*Tokens* e POS ao redor dos EVTs em uma janela de dois *tokens* (dois antes e dois depois). Representação por unigramas, bigramas e trigramas. Cada EVT tem sua própria representação.	Tang *et al.* (2013)
Menção Rep 1	*Tokens* dos EVTs. Representação por BOW. Cada EVT tem sua própria representação.	MacAveney, Cohan e Goharian (2017)
Menção Rep 2	*Tokens* e POS dos EVTs. Representação por BOW. Cada EVT tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2016, 2018)
Tenso	Tenso de todos os verbos finitos da sentença. Representação por BOW. Uma representação para cada sentença.	Lee *et al.* (2016, 2018), Lin *et al.* (2016a), Tang *et al.* (2013)
Atributos	Atributos Tipo, Polaridade, Modalidade e TempRelDCD de cada EVT. Representados por meio de *one-hot encoding*. Cada atributo de cada EVT tem sua própria representação.	Lin *et al.* (2016a), Lee *et al.* (2016, 2018), MacAveney, Cohan e Goharian (2017), Mirza e Tonelli (2016)
Concatenação de atributos	Concatenação dos atributos dos EVTs: Tipo com Tipo, Polaridade com Polaridade, Modalidade com Modalidade e RelTempDCD com RelTempDCD. A ordem do elemento na relação influencia a concatenação. Representada por meio de *one-hot encoding*.	Xu *et al.* (2013), Cherry *et al.* (2013), MacAveney, Cohan e Goharian (2017), Mirza e Tonelli (2016)
Número de *tokens*	Contagem do número de *tokens* presentes entre as menções de EVT. Representado por valores numéricos.	Lin *et al.* (2016a), Tang *et al.* (2013), Lee *et al.* (2018)
Número de menções	Contagem do número de menções (EVTs e ETs) entre as menções de EVT. Representado por valores numéricos.	Lee *et al.* (2016), Tang *et al.* (2013), Cherry *et al.* (2013), Mirza e Tonelli (2016)
UMLS	Tipos semânticos de UMLS relacionados a cada EVT. Representação por BOW. Cada EVT tem sua própria representação.	Lin *et al.* (2016a), Cherry *et al.* (2013), Xu *et al.* (2013)
Posição	Posições dos EVTs no texto, calculadas a partir da divisão da posição inicial pela quantidade de caracteres do texto. Representada por valores numéricos. Cada EVT tem sua própria representação.	Tang *et al.* (2013)
Data Sentença	Indicador da existência de uma ET do tipo Data na sentença do EVT. Representação binária. Cada EVT tem sua própria representação.	Tang *et al.* (2013)

Fonte: O autor (2020).

Para este tipo de TLINK, foi adicionada uma *feature*, “Data Sentença”, para identificar se existia uma ET do tipo Data na sentença dos EVTs. Menções gerais de exames de laboratórios usualmente estão na mesma sentença de uma ET do tipo Data, com os exames de laboratório presentes em sentenças sem menções de data. De forma similar, exames de diagnóstico usualmente estão na mesma sentença de

uma ET do tipo Data, enquanto os problemas médicos encontrados constam em sentenças sem menções de data.

- ***Features* para RelTempDCD**

Os conjuntos de *features* utilizados para este tipo estão representados no Quadro 4, com suas respectivas descrições e autores-base. A maior parte já foi detalhada, por isso as explicações serão direcionadas às particularidades encontradas neste conjunto.

Quadro 4 – Lista de *features* para RelTempDCD.

Feature	**Descrição**	**Autores-base**
Contexto Rep 1	*Tokens* ao redor do EVT em uma janela de três *tokens* (três antes e três depois). Representação por BOW.	Lin *et al.* (2016a), Tourille *et al.* (2017b), Viani *et al.* (2019)
Contexto Rep 2	*Tokens* e POS ao redor do EVT em uma janela de cinco *tokens* (cinco antes e cinco depois). Representação por unigramas, bigramas e trigramas.	Tang *et al.* (2013), Lee *et al.* (2016), Tourille *et al.* (2017a)
Menção Rep 1	*Tokens* do EVT. Representação por BOW.	Lin *et al.* (2016a), Tourille *et al.* (2017b)
Menção Rep 2	*Tokens* e POS do EVT. Representação por BOW.	-
Tenso Rep 1	Tenso de todos os verbos finitos da sentença. Representação por BOW.	Tang *et al.* (2013), Lee *et al.* (2016), Lin *et al.* (2016a), Tourille *et al.* (2017a, 2017b), Viani *et al.* (2019)
Tenso Rep 2	Tenso de todos os verbos finitos e forma dos verbos não finitos da sentença. Representação de BOW.	-
Atributos	Atributos Tipo, Polaridade, Modalidade e RelTempDCD do EVT. Representados por meio de *one-hot encoding*.	Tang *et al.* (2013), Lin *et al.* (2016a), Lee *et al.* (2016), Mirza e Tonelli (2016), Tourille *et al.* (2017a, 2017b)
UMLS	Tipos semânticos de UMLS relacionados ao EVT. Representação por BOW.	Lin *et al.* (2016a)
Posição	Posições do EVT no texto, calculadas a partir da divisão da posição inicial pela quantidade de caracteres do texto. Representada por valores numéricos.	Tang *et al.* (2013), Lin *et al.* (2016a), Tourille *et al.* (2017a, 2017b), Viani *et al.* (2019)
Contexto menções	Tipos de outros EVTs ou ETs que estão na janela de uma menção (uma antes e uma depois). Representado por meio de *one-hot encoding*.	Tourille *et al.* (2017b)

Fonte: O autor (2020).

Na *feature* “Contexto Rep 2”, foi testado aumentar a janela para cinco, tentando capturar um contexto mais profundo por meio dos n-gramas. Também foi introduzida uma *feature* para este tipo de relação, “Contexto menções”, cuja função foi verificar o

contexto das menções em uma janela de um, ou seja, quais tipos de EVT e ET estavam presentes. Por exemplo, se o EVT estivesse em um contexto de duas ETs do tipo Frequência e fosse um medicamento, provavelmente RelTempDCD seria Antes/Sobreposto (medicação em uso) ou Depois (medicação que está sendo prescrita).

A *feature* “Tenso Rep 1” é similar às descritas anteriormente, residindo a diferença na *feature* “Tenso Rep 2”, por trazer os tensos dos verbos finitos, além da forma dos verbos não finitos. Esta *feature* foi proposta pelo doutorando, com o intuito de verificar se questões como gerúndio, infinitivo e particípio influenciam o aprendizado do algoritmo.

APÊNDICE F – EXPERIMENTOS NO CONJUNTO DE TREINAMENTO

Para cada um dos classificadores, foram feitos experimentos com determinados conjuntos de *features*, detalhados a seguir.

- **Experimentos de TLINKs entre eventos em mesma sentença**

As propostas para este tipo de TLINK foram testadas com todos os conjuntos de *features* detalhados no Quadro 1, já definidos no Apêndice E. As *features* “Tenso”, “*Tokens* dentre”, “Atributos”, “Concatenação de atributos”, “Número de *tokens*” e “Número de menções” eram fixas, tendo sido adicionadas em todos os conjuntos. As *features* “Contexto Rep 1” e “Contexto Rep 2” alternaram sua ocorrência durante todo o experimento, tendo o objetivo de verificar se considerar contexto adicional por meio de bigramas e trigramas beneficiaria a classificação. As *features* “Menção Rep 1” e “Menção Rep 2” alternaram sua ocorrência durante todo o experimento, a fim de verificar o efeito do POS dos EVTs. As *features* “UMLS”, “Posição” e “Termos Antes/Depois” eram adicionais, não sendo amplamente utilizadas na literatura, porém foram testadas.

Quadro 1 – Experimentos de TLINKs entre EVTs em mesma sentença.

Conjunto	**Contexto Rep 1**	**Contexto Rep 2**	**Menção Rep 1**	**Menção Rep 2**	**Tenso**	***Tokens* dentre**	**Atributos**	**Concatenação de atributos**	**Número de *tokens***	**Número de menções**	**UMLS**	**Posição**	**Termos Antes/Depois**
Modelo-base	x		x		x	x	x	x	x	x			
Conj Feat EVT-EVT MS 1	x			x	x	x	x	x	x	x			
Conj Feat EVT-EVT MS 2		x	x		x	x	x	x	x	x			
Conj Feat EVT-EVT MS 3		x		x	x	x	x	x	x	x			
Conj Feat EVT-EVT MS 4	x		x		x	x	x	x	x	x	x	x	x
Conj Feat EVT-EVT MS 5	x			x	x	x	x	x	x	x	x	x	x
Conj Feat EVT-EVT MS 6		x	x		x	x	x	x	x	x	x	x	x
Conj Feat EVT-EVT MS 7		x		x	x	x	x	x	x	x	x	x	x
Conj Feat EVT-EVT MS 8	x		x		x	x	x	x	x	x	x		
Conj Feat EVT-EVT MS 9	x			x	x	x	x	x	x	x	x		
Conj Feat EVT-EVT MS 10		x	x		x	x	x	x	x	x	x		
Conj Feat EVT-EVT MS 11		x		x	x	x	x	x	x	x	x		
Conj Feat EVT-EVT MS 12	x		x		x	x	x	x	x	x		x	
Conj Feat EVT-EVT MS 13	x			x	x	x	x	x	x	x		x	
Conj Feat EVT-EVT MS 14		x	x		x	x	x	x	x	x		x	

Conj Feat EVT-EVT MS 15		x		x	x	x	x	x	x	x		x	
Conj Feat EVT-EVT MS 16	x		x		x	x	x	x	x	x			x
Conj Feat EVT-EVT MS 17	x			x	x	x	x	x	x	x			x
Conj Feat EVT-EVT MS 18		x	x		x	x	x	x	x	x			x
Conj Feat EVT-EVT MS 19		x		x	x	x	x	x	x	x			x
Conj Feat EVT-EVT MS 20	x		x		x	x	x	x	x	x	x	x	
Conj Feat EVT-EVT MS 21	x			x	x	x	x	x	x	x	x	x	
Conj Feat EVT-EVT MS 22		x	x		x	x	x	x	x	x	x	x	
Conj Feat EVT-EVT MS 23		x		x	x	x	x	x	x	x	x	x	
Conj Feat EVT-EVT MS 24	x		x		x	x	x	x	x	x	x		x
Conj Feat EVT-EVT MS 25	x			x	x	x	x	x	x	x	x		x
Conj Feat EVT-EVT MS 26		x	x		x	x	x	x	x	x	x		x
Conj Feat EVT-EVT MS 27		x		x	x	x	x	x	x	x	x		x
Conj Feat EVT-EVT MS 28	x		x		x	x	x	x	x	x		x	x
Conj Feat EVT-EVT MS 29	x			x	x	x	x	x	x	x		x	x
Conj Feat EVT-EVT MS 30		x	x		x	x	x	x	x	x		x	x
Conj Feat EVT-EVT MS 31		x		x	x	x	x	x	x	x		x	x

Fonte: O autor (2020).

- **Experimentos de TLINKs entre eventos e expressões temporais em mesma sentença**

As propostas para este tipo de TLINK foram testadas com todos os conjuntos de *features* detalhados no Quadro 2. As *features* “Tenso”, “*Tokens* dentre”, “Atributos”, “Concatenação do tipo” e “Número de *tokens*” eram fixas, sendo adicionadas em todos os conjuntos. Da mesma forma que nos experimentos de TLINKs entre EVTs em mesma sentença, “Contexto Rep 1”, “Contexto Rep 2”, “Menção Rep 1” e “Menção Rep 2” alternaram sua ocorrência durante os experimentos. Neste caso, “Par mais próximo” e “Número de menções” também alternaram sua ocorrência durante os experimentos, tendo o objetivo de verificar se era mais informativo para o classificador a informação de que aquele par era o mais próximo na sentença ou o número de menções entre o par. As *features* “ET começo”, “Posição” e “Conjunção” eram adicionais, que não são amplamente utilizadas na literatura, porém foram testadas.

Quadro 2 – Experimentos de TLINKs entre EVT e ET em mesma sentença.

Conjunto	**Contexto Rep 1**	**Contexto Rep 2**	**Menção Rep 1**	**Menção Rep 2**	**Par mais próximo**	**Número de menções**	**Tenso**	***Tokens* dentre**	**Atributos**	**Concatenação do tipo**	**Número de *tokens***	**ET começo**	**Posição**	**Conjunção**
Modelo-base	x		x			x	x	x	x	x	x			

Conj Feat EVT-ET MS 1	x		x		x		x	x	x	x	x			
Conj Feat EVT-ET MS 2	x			x		x	x	x	x	x	x			
Conj Feat EVT-ET MS 3	x			x	x		x	x	x	x	x			
Conj Feat EVT-ET MS 4		x	x			x	x	x	x	x	x			
Conj Feat EVT-ET MS 5		x	x		x		x	x	x	x	x			
Conj Feat EVT-ET MS 6		x		x		x	x	x	x	x	x			
Conj Feat EVT-ET MS 7		x		x	x		x	x	x	x	x			
Conj Feat EVT-ET MS 8	x		x			x	x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 9	x		x		x		x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 10	x			x		x	x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 11	x			x	x		x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 12		x	x			x	x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 13		x	x		x		x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 14		x		x		x	x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 15		x		x	x		x	x	x	x	x	x	x	x
Conj Feat EVT-ET MS 16	x		x			x	x	x	x	x	x			x
Conj Feat EVT-ET MS 17	x		x		x		x	x	x	x	x			x
Conj Feat EVT-ET MS 18	x			x		x	x	x	x	x	x			x
Conj Feat EVT-ET MS 19	x			x	x		x	x	x	x	x			x
Conj Feat EVT-ET MS 20		x	x			x	x	x	x	x	x			x
Conj Feat EVT-ET MS 21		x	x		x		x	x	x	x	x			x
Conj Feat EVT-ET MS 22		x		x		x	x	x	x	x	x			x
Conj Feat EVT-ET MS 23		x		x	x		x	x	x	x	x			x
Conj Feat EVT-ET MS 24	x		x			x	x	x	x	x	x		x	
Conj Feat EVT-ET MS 25	x		x		x		x	x	x	x	x		x	
Conj Feat EVT-ET MS 26	x			x		x	x	x	x	x	x		x	
Conj Feat EVT-ET MS 27	x			x	x		x	x	x	x	x		x	
Conj Feat EVT-ET MS 28		x	x			x	x	x	x	x	x		x	
Conj Feat EVT-ET MS 29		x	x		x		x	x	x	x	x		x	
Conj Feat EVT-ET MS 30		x		x		x	x	x	x	x	x		x	
Conj Feat EVT-ET MS 31		x		x	x		x	x	x	x	x		x	
Conj Feat EVT-ET MS 32		x	x			x	x	x	x	x	x	x		
Conj Feat EVT-ET MS 33		x	x		x		x	x	x	x	x	x		
Conj Feat EVT-ET MS 34		x		x		x	x	x	x	x	x	x		
Conj Feat EVT-ET MS 35		x		x	x		x	x	x	x	x	x		
Conj Feat EVT-ET MS 36	x		x			x	x	x	x	x	x	x		
Conj Feat EVT-ET MS 37	x		x		x		x	x	x	x	x	x		
Conj Feat EVT-ET MS 38	x			x		x	x	x	x	x	x	x		
Conj Feat EVT-ET MS 39	x			x	x		x	x	x	x	x	x		
Conj Feat EVT-ET MS 40		x	x			x	x	x	x	x	x	x	x	
Conj Feat EVT-ET MS 41		x	x		x		x	x	x	x	x	x	x	
Conj Feat EVT-ET MS 42		x		x		x	x	x	x	x	x	x	x	
Conj Feat EVT-ET MS 43		x		x	x		x	x	x	x	x	x	x	
Conj Feat EVT-ET MS 44	x		x			x	x	x	x	x	x	x	x	
Conj Feat EVT-ET MS 45	x		x		x		x	x	x	x	x	x	x	
Conj Feat EVT-ET MS 46	x			x		x	x	x	x	x	x	x	x	
Conj Feat EVT-ET MS 47	x			x	x		x	x	x	x	x	x	x	

Conj Feat EVT-ET MS 48		x	x			x	x	x	x	x	x	x		x
Conj Feat EVT-ET MS 49		x	x		x		x	x	x	x	x	x		x
Conj Feat EVT-ET MS 50		x		x		x	x	x	x	x	x	x		x
Conj Feat EVT-ET MS 51		x		x	x		x	x	x	x	x	x		x
Conj Feat EVT-ET MS 52	x		x			x	x	x	x	x	x	x		x
Conj Feat EVT-ET MS 53	x		x		x		x	x	x	x	x	x		x
Conj Feat EVT-ET MS 54	x			x		x	x	x	x	x	x	x		x
Conj Feat EVT-ET MS 55	x			x	x		x	x	x	x	x	x		x
Conj Feat EVT-ET MS 56		x	x			x	x	x	x	x	x		x	x
Conj Feat EVT-ET MS 57		x	x		x		x	x	x	x	x		x	x
Conj Feat EVT-ET MS 58		x		x		x	x	x	x	x	x		x	x
Conj Feat EVT-ET MS 59		x		x	x		x	x	x	x	x		x	x
Conj Feat EVT-ET MS 60	x		x			x	x	x	x	x	x		x	x
Conj Feat EVT-ET MS 61	x		x		x		x	x	x	x	x		x	x
Conj Feat EVT-ET MS 62	x			x		x	x	x	x	x	x		x	x
Conj Feat EVT-ET MS 63	x			x	x		x	x	x	x	x		x	x

Fonte: O autor (2020).

- **Experimentos para TLINKs entre eventos em sentenças distintas**

As propostas para este tipo de TLINK foram testadas com todos os conjuntos de *features* detalhados no Quadro 3. As *features* “Tenso”, “*Tokens* dentre”, “Atributos”, “Concatenação do tipo” e “Número de *tokens*” eram fixas, sendo adicionadas em todos os conjuntos. Da mesma forma que nos experimentos de TLINKs entre EVTs em mesma sentença, “Contexto Rep 1”, “Contexto Rep 2”, “Menção Rep 1” e “Menção Rep 2” alternaram sua ocorrência durante os experimentos. As *features* “UMLS”, “Posição” e “Data sentença” eram adicionais, que não são amplamente utilizadas na literatura, porém foram testadas.

Quadro 3 – Experimentos de TLINKs entre EVTs em sentenças distintas.

Conjunto	**Contexto Rep 1**	**Contexto Rep 2**	**Menção Rep 1**	**Menção Rep 2**	**Tenso**	**Atributos**	**Concatenação de atributos**	**Número de *tokens***	**Número de menções**	**UMLS**	**Posição**	**Data sentença**
Modelo-base	x		x		x	x	x	x	x			
Conj Feat EVT-ET DS 1	x			x	x	x	x	x	x			
Conj Feat EVT-ET DS 2		x	x		x	x	x	x	x			
Conj Feat EVT-ET DS 3		x		x	x	x	x	x	x			
Conj Feat EVT-ET DS 4	x		x		x	x	x	x	x	x	x	x
Conj Feat EVT-ET DS 5	x			x	x	x	x	x	x	x	x	x
Conj Feat EVT-ET DS 6		x	x		x	x	x	x	x	x	x	x
Conj Feat EVT-ET DS 7		x		x	x	x	x	x	x	x	x	x

Conj Feat EVT-ET DS 8	x		x		x	x	x	x	x	x		
Conj Feat EVT-ET DS 9	x			x	x	x	x	x	x	x		
Conj Feat EVT-ET DS 10		x	x		x	x	x	x	x	x		
Conj Feat EVT-ET DS 11		x		x	x	x	x	x	x	x		
Conj Feat EVT-ET DS 12	x		x		x	x	x	x	x		x	
Conj Feat EVT-ET DS 13	x			x	x	x	x	x	x		x	
Conj Feat EVT-ET DS 14		x	x		x	x	x	x	x		x	
Conj Feat EVT-ET DS 15		x		x	x	x	x	x	x		x	
Conj Feat EVT-ET DS 16	x		x		x	x	x	x	x			x
Conj Feat EVT-ET DS 17	x			x	x	x	x	x	x			x
Conj Feat EVT-ET DS 18		x	x		x	x	x	x	x			x
Conj Feat EVT-ET DS 19		x		x	x	x	x	x	x			x
Conj Feat EVT-ET DS 20	x		x		x	x	x	x	x	x	x	
Conj Feat EVT-ET DS 21	x			x	x	x	x	x	x	x	x	
Conj Feat EVT-ET DS 22		x	x		x	x	x	x	x	x	x	
Conj Feat EVT-ET DS 23		x		x	x	x	x	x	x	x	x	
Conj Feat EVT-ET DS 24	x		x		x	x	x	x	x	x		x
Conj Feat EVT-ET DS 25	x			x	x	x	x	x	x	x		x
Conj Feat EVT-ET DS 26		x	x		x	x	x	x	x	x		x
Conj Feat EVT-ET DS 27		x		x	x	x	x	x	x	x		x
Conj Feat EVT-ET DS 28	x		x		x	x	x	x	x		x	x
Conj Feat EVT-ET DS 29	x			x	x	x	x	x	x		x	x
Conj Feat EVT-ET DS 30		x	x		x	x	x	x	x		x	x
Conj Feat EVT-ET DS 31		x		x	x	x	x	x	x		x	x

Fonte: O autor (2020).

- **Experimentos de RelTempDCD**

Os experimentos para RelTempDCD envolveram os conjuntos de *features* mostrados no Quadro 4. A *feature* “Atributos” era fixa, sendo adicionada em todos os conjuntos. Da mesma forma que nos experimentos de TLINKs entre EVTs em mesma sentença, “Contexto Rep 1”, “Contexto Rep 2”, “Menção Rep 1” e “Menção Rep 2” alternaram sua ocorrência durante os experimentos. Neste caso, “Contexto Rep 2” envolveu uma janela de cinco *tokens* de contexto. Além disso, “Tenso 1” e “Tenso 2” alternaram sua ocorrência durante os experimentos, com o objetivo de verificar se, além do tenso, a forma dos verbos não finitos poderia contribuir com a classificação. As *features* “UMLS”, “Posição” e “Contexto menções” eram adicionais, não obrigatórias, que não são amplamente utilizadas na literatura, porém foram testadas.

Quadro 4 – Experimentos de RelTempDCD.

Conjunto	Contexto Rep 1	Contexto Rep 2	Menção Rep 1	Menção Rep 2	Tenso 1	Tenso 2	Atributos	Posição	Contexto menções	UMLS
Modelo-base	x		x		x		x			
Conj Feat RelTempDCD 1	x		x			x	x			
Conj Feat RelTempDCD 2	x			x	x		x			
Conj Feat RelTempDCD 3	x			x		x	x			
Conj Feat RelTempDCD 4		x	x		x		x			
Conj Feat RelTempDCD 5		x	x			x	x			
Conj Feat RelTempDCD 6		x		x	x		x			
Conj Feat RelTempDCD 7		x		x		x	x			
Conj Feat RelTempDCD 8	x		x		x		x	x	x	x
Conj Feat RelTempDCD 9	x		x			x	x	x	x	x
Conj Feat RelTempDCD 10	x			x	x		x	x	x	x
Conj Feat RelTempDCD 11	x			x		x	x	x	x	x
Conj Feat RelTempDCD 12		x	x		x		x	x	x	x
Conj Feat RelTempDCD 13		x	x			x	x	x	x	x
Conj Feat RelTempDCD 14		x		x	x		x	x	x	x
Conj Feat RelTempDCD 15		x		x		x	x	x	x	x
Conj Feat RelTempDCD 16	x		x		x		x			x
Conj Feat RelTempDCD 17	x		x			x	x			x
Conj Feat RelTempDCD 18	x			x	x		x			x
Conj Feat RelTempDCD 19	x			x		x	x			x
Conj Feat RelTempDCD 20		x	x		x		x			x
Conj Feat RelTempDCD 21		x	x			x	x			x
Conj Feat RelTempDCD 22		x		x	x		x			x
Conj Feat RelTempDCD 23		x		x		x	x			x
Conj Feat RelTempDCD 24	x		x		x		x	x		
Conj Feat RelTempDCD 25	x		x			x	x	x		
Conj Feat RelTempDCD 26	x			x	x		x	x		
Conj Feat RelTempDCD 27	x			x		x	x	x		
Conj Feat RelTempDCD 28		x	x		x		x	x		
Conj Feat RelTempDCD 29		x	x			x	x	x		
Conj Feat RelTempDCD 30		x		x	x		x	x		
Conj Feat RelTempDCD 31		x		x		x	x	x		
Conj Feat RelTempDCD 32	x		x		x		x		x	
Conj Feat RelTempDCD 33	x		x			x	x		x	
Conj Feat RelTempDCD 34	x			x	x		x		x	
Conj Feat RelTempDCD 35	x			x		x	x		x	
Conj Feat RelTempDCD 36		x	x		x		x		x	
Conj Feat RelTempDCD 37		x	x			x	x		x	
Conj Feat RelTempDCD 38		x		x	x		x		x	
Conj Feat RelTempDCD 39		x		x		x	x		x	
Conj Feat RelTempDCD 40	x		x		x		x	x	x	

Conj Feat RelTempDCD 41	x		x			x	x	x	x	
Conj Feat RelTempDCD 42	x			x	x		x	x	x	
Conj Feat RelTempDCD 43	x			x		x	x	x	x	
Conj Feat RelTempDCD 44		x	x		x		x	x	x	
Conj Feat RelTempDCD 45		x	x			x	x	x	x	
Conj Feat RelTempDCD 46		x		x	x		x	x	x	
Conj Feat RelTempDCD 47		x		x		x	x	x	x	
Conj Feat RelTempDCD 48	x		x		x		x		x	x
Conj Feat RelTempDCD 49	x		x			x	x		x	x
Conj Feat RelTempDCD 50	x			x	x		x		x	x
Conj Feat RelTempDCD 51	x			x		x	x		x	x
Conj Feat RelTempDCD 52		x	x		x		x		x	x
Conj Feat RelTempDCD 53		x	x			x	x		x	x
Conj Feat RelTempDCD 54		x		x	x		x		x	x
Conj Feat RelTempDCD 55		x		x		x	x		x	x
Conj Feat RelTempDCD 56	x		x		x		x	x		x
Conj Feat RelTempDCD 57	x		x			x	x	x		x
Conj Feat RelTempDCD 58	x			x	x		x	x		x
Conj Feat RelTempDCD 59	x			x		x	x	x		x
Conj Feat RelTempDCD 60		x	x		x		x	x		x
Conj Feat RelTempDCD 61		x	x			x	x	x		x
Conj Feat RelTempDCD 62		x		x	x		x	x		x
Conj Feat RelTempDCD 63		x		x		x	x	x		x

Fonte: O autor (2020).

APÊNDICE G – RESULTADOS PARA O CONJUNTO DE TREINAMENTO

Nesta seção, são sumarizados os resultados para o conjunto de treinamento para cada uma das RTs.

- **Resultados do treinamento para TLINKs entre eventos em mesma sentença**

Os resultados no conjunto de treinamento para o modelo-base e o melhor modelo para cada proposta para este tipo de TLINK são mostrados na Tabela 1. Os melhores modelos envolveram os seguintes conjuntos: (i) para a proposta “Sem heurística”, o conjunto “Conj Feat EVT-EVT MS 27”; (ii) para “Pares Esquerda-Direita”, o conjunto “Modelo-base”; (iii) para “Heurística EVT-EVT MS”, o conjunto “Conj Feat EVT-EVT MS 31”; (iv) para “Pares Esquerda-Direita + Heurística EVT-EVT MS”, o conjunto “Conj Feat EVT-EVT MS 28”.

Tabela 1 – Resultados para TLINKs entre EVTs em mesma sentença no conjunto de treinamento para todas as propostas, com modelo-base e o melhor modelo.

Proposta	Modelo	F1-*score*	Desvio padrão	C	t (t crit = 1,66; alfa = 0,05)
Sem heurística	Modelo-base	0,7078	0,0422	1	-
	Melhor modelo	**0,7339**	0,0445	0,25	-3,0257
Pares Esquerda-Direita	Modelo-base	**0,7907**	0,0368	1	-
	Melhor modelo	**0,7907**	0,0368	1	-
Heurística EVT-EVT MS	Modelo-base	0,7048	0,0411	1	-
	Melhor modelo	**0,7364**	0,0443	0,25	-3,6652
Pares Esquerda-Direita + Heurística EVT-EVT MS	Modelo-base	0,8010	0,0390	1	-
	Melhor modelo	**0,8089**	0,0394	0,25	-1,9384

Fonte: O autor (2020).

A partir dos experimentos, é evidenciado que, em nível de classificador (não no nível local), tanto a proposta “Pares Esquerda Direita” quanto “Pares Esquerda-Direita + Heurística EVT-EVT MS” obtêm os melhores resultados durante o treinamento. Para “Pares Esquerda-Direita”, o melhor modelo envolveu o próprio modelo-base.

- **Resultados do treinamento para TLINKs entre eventos e expressões temporais em mesma sentença**

Os resultados no conjunto de treinamento para o modelo-base e o melhor modelo para cada proposta para este tipo de TLINK são mostrados na Tabela 2. Os melhores modelos envolveram os seguintes conjuntos de *features*: (i) para proposta “Sem heurística”, o conjunto “Conj Feat EVT-ET MS 19”; (ii) para “Pares Esquerda-Direita”, o conjunto “Modelo-base”. Nota-se que, em nível de classificador, existiu uma melhora significativa da proposta “Sem heurística” em seu melhor modelo, se beneficiando das mudanças de não considerar o POS das menções, pela utilização de “Menção Rep 1”, e da troca da *feature* “Número de menções” pela *feature* “Par mais próximo”, um valor binário indicando se esse par era mais próximo entre as menções. Já o melhor desempenho para a proposta “Pares Esquerda-Direita” ocorreu com a utilização do seu próprio modelo-base.

Tabela 2 – Resultados para TLINKs entre EVT e ET em mesma sentença no conjunto de treinamento para todas as propostas, com o modelo-base e o melhor modelo.

Proposta	**Modelo**	**F1-*score***	**Desvio padrão**	**C**	**t (t crit = 1,66; alfa = 0,05)**
Sem heurística	Modelo-base SH	0,6904	0,0336	1	-
	Melhor modelo SH	**0,8196**	0,0307	0,25	-9,8275
Pares Esquerda-Direita	Modelo-base PED	**0,9017**	0,0250	1	-
	Melhor modelo PED	**0,9017**	0,0250	1	-

Fonte: O autor (2020).
Notas: SH = Sem heurística; PED = Pares Esquerda-Direita.

- **Resultados do treinamento para TLINKs entre eventos em sentenças distintas**

Os resultados no conjunto de treinamento para o modelo-base e o melhor modelo para cada proposta para este tipo de TLINK são mostrados na Tabela 3. Os melhores modelos envolveram os seguintes conjuntos de *features*: (i) para proposta “Heurística EVT-EVT SD”, o conjunto “Conj Feat EVT-ET DS 29”; (ii) para “Pares Esquerda-Direita + Heurística EVT-EVT SD”, o conjunto “Modelo-base”.

Tabela 3 – Resultados para TLINKs entre EVTS em sentenças distintas no conjunto de treinamento para todas as propostas, com o modelo-base e o melhor modelo a partir dos experimentos.

Proposta	Modelo	F1-*score*	Desvio padrão	C	t (t crit = 1,66; alfa = 0,05)
Heurística EVT-EVT SD	Modelo-base HESD	0,7640	0,0695	1	-
	Melhor modelo HESD	**0,8578**	0,0535	0,25	-4,3828
Pares Esquerda-Direita + Heurística EVT-EVT SD	Modelo-base PED-HESD	**0,9090**	0,0555	1	-
	Melhor modelo PED-HESD	**0,9090**	0,0555	1	-

Fonte: O autor (2020).
Noras: HESD = Heurística EVT-EVT SD; PED= Pares Esquerda-Direita.

Nota-se que, em nível de classificador, existiu uma melhora significativa da proposta “Sem heurística” no seu melhor modelo. O melhor desempenho para a proposta “Pares Esquerda-Direita + Heurística EVT-EVT SD” foi obtido com a utilização do seu próprio modelo-base.

- **Resultados do treinamento para RelTempDCD**

Os resultados no conjunto de treinamento para o modelo-base e o melhor modelo para cada proposta para RelTempDCD são mostrados na Tabela 4. O melhor modelo envolveu o conjunto “Conj Feat RelTempDCD 14”.

Tabela 4 – Resultados para RelTempDCD, com o modelo-base e o melhor modelo a partir dos experimentos no conjunto de treinamento.

Modelo	F1-*score*	Desvio padrão	C	t (t crit = 1,66; alfa = 0,05)
Modelo-base RelTempDCD	0,9139	0,0144	1	-
Melhor modelo RelTempDCD	**0,9388**	0,0123	0,25	-4,3940

Fonte: O autor (2020).

Nota-se que existiu uma melhora nos resultados do modelo-base para o melhor modelo, devido à mudança das *features*. A *feature* relacionada ao contexto “Contexto Rep 1” do modelo-base foi trocada por “Context Rep 2”, alterando a configuração do contexto de uma janela de três (-3, +3) para cinco (-5, +5), considerando bigramas e trigramas tanto dos *tokens* quanto do POS destes. Além disso, foram adicionadas as *features* “Posição” (indicando a posição no documento) e “Contexto Menções” (trazendo os tipos de menção que estão do lado direito e esquerdo).

ANEXO A – APROVAÇÃO DO COMITÊ DE ÉTICA

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 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 a iniciar 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

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)