UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Metodologia de Pré-processamento Textual para Extração de Informação sobre Efeitos de Doenças em Artigos Científicos do Domínio Biomédico

Pablo Freire Matos

São Carlos
Setembro/2010

Pablo Freire Matos

Metodologia de Pré-processamento Textual para Extração de Informação sobre Efeitos de Doenças em Artigos Científicos do Domínio Biomédico

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Ricardo Rodrigues Ciferri
Coorientador: Thiago Alexandre Salgueiro Pardo

São Carlos
Setembro/2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M433mp

Matos, Pablo Freire.
Metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico / Pablo Freire Matos. -- São Carlos : UFSCar, 2010.
159 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Banco de dados. 2. Mineração de textos. 3. Artigos científicos. 4. Domínio biomédico. I. Título.

CDD: 005.74 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Metodologia de Pré-processamento Textual para Extração de Informação sobre Efeitos de Doenças em Artigos Científicos do Domínio Biomédico"

PABLO FREIRE MATOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Prof. Dr. Ricardo Rodrigues Ciferri
(Orientador - DC/UFSCar)

Profa. Dra. Solange Oliveira Rezende
(ICM/USP)

Prof. Dr. Sérgio Roberto Pereira da Silva
(DIN/UEM)

São Carlos
Setembro/2010

Ao Senhor meu Deus, responsável por eu estar aqui e pelas bênçãos realizadas em minha vida.

À minha família por estar, mesmo que distante, presente em minha vida; Em especial, a minha querida mãe que me ensinou a sempre dar prioridade aos meus estudos.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por me conceder saúde, forças, paciência e principalmente, perseverança para concluir este trabalho.

Aos meus pais, Lélia e Carlos, e meus irmãos, Karolly e Diego, por estarem juntos em todos os momentos, incentivando-me com palavras de conforto e motivação.

À Aline pelas orações direcionadas a conclusão deste projeto.

Ao Grupo de Banco de Dados da UFSCar e do ICMC/USP, e ao grupo do Projeto da Anemia Falciforme que me propiciaram momentos de aprendizado intelectual inestimáveis.

Aos colegas André Oliveira e Rodolfo Barbeiro, agradeço pela convivência desde o momento que nos conhecemos na entrevista de seleção do mestrado e depois quando viemos a morar juntos.

A todos os meus amigos e colegas do mestrado por compartilharem os momentos de tristeza e alegria, em especial, Mauricio Cerri, Jesus Portocarrero, David Buzatto, Paulo Ávila, Juliana Duque, Renata Tsuruda, Samara Martins, Caroline Perlin, Daniel Cugler, Thiago Siqueira, Rafael Miani, Marcus Teixeira, Rodrigo Bela, Edimilson Batista, Leonardo Lombardi, João Paulo Siqueira, Maykon Santana, Matheus Viana e Mayra Rodriguez.

À Arnaldo Candido sou muito grato pela genialidade em solucionar algumas dúvidas sobre a minha pesquisa ou simplesmente por arrumar um tempinho para conversarmos.

Ao professor e orientador Dr. Ricardo Rodrigues Ciferri, obrigado pela oportunidade de estudar na UNIVERSIDADE FEDERAL DE SÃO CARLOS. Com certeza, uma experiência que levarei por toda minha vida. Muito obrigado pela confiança que depositou em mim, pelo incentivo, pelas dicas e pelas correções que foram imprescindíveis para a concretização deste projeto.

Ao docente Dr. Thiago Alexandre Salgueiro Pardo que me coorientou com maestria. Agradeço-lhe pela oportunidade de estudar PLN no ICMC/USP, o que me proporcionou o

conhecimento de uma nova área até então desconhecida e pelas imprescindíveis contribuições acrescentadas neste projeto.

Às docentes Drª. Cristina Dutra de Aguiar Ciferri, Drª. Marilde Terezinha Prado Santos e Drª. Marina Teresa Pires Vieira por fazerem críticas construtivas em minhas apresentações, incentivando-me a procurar as respostas para os questionamentos realizados. Agradeço a esta última por incentivar-me e por apoiar-me na escrita de artigo.

À docente Drª. Marcela Xavier Ribeiro pela oportunidade de atuar como tutor virtual na disciplina de Projeto de Banco de Dados, o que além de auxiliar-me financeiramente após o término da bolsa de mestrado, me propiciou um crescimento profissional.

Ao então presidente do CNPq, o médico Dr. Marco Antonio Zago, idealizador do projeto Anemia Falciforme e em especial à médica Drª. Ana Cristina Silva Pinto, especialista do domínio, que concedeu horas do seu importante tempo para solucionar as dúvidas que ajudaram no melhor entendimento do problema a ser superado.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro.

À Camila Cassiavilani Passos, bibliotecária da UFSCar, pelas várias correções realizadas nas referências desta dissertação.

Por fim, agradeço a todos que contribuíram direta ou indiretamente por mais esta conquista e a UFSCar por acolher-me e propiciar-me anos de experiência e conhecimento que vão servir para toda a minha vida.

"*Uma busca começa sempre com a sorte do PRINCIPIANTE. E termina sempre com a prova do CONQUISTADOR.*"

(Paulo Coelho)

"*NENHUM detalhe é tão pequeno que não justifique o suor. E NENHUM feito é tão grande que não desperte SONHOS.*"

(JACK DEFINITIVO, 2001, p. 432)

"*Não se tem CERTEZA de nada. Essa é a única CERTEZA que tenho.*"

(John Forbes Nash Jr.)

RESUMO

Existe um grande volume de informação não estruturada (i.e., em formato textual) sendo publicada cada vez mais em meios eletrônicos, particularmente em bibliotecas digitais. Assim, com o passar do tempo, o ser humano fica cada vez mais restringido a uma limitada quantidade de texto que é capaz de processar e assimilar. No sentido de identificar as informações relevantes de um texto e com o objetivo de estruturar e armazenar essas informações em um banco de dados, a fim de propiciar uma futura descoberta de relacionamentos interessantes entre as informações extraídas, nesta dissertação é proposta uma metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico. A metodologia é composta por quatro etapas: **Entrada de Dados** (Etapa 1), **Classificação de Sentenças** (Etapa 2), **Identificação de Termos Relevantes** (Etapa 3) e **Gerenciamento de Termos** (Etapa 4). Esta metodologia utiliza três abordagens de extração de informação encontradas na literatura: abordagem baseada em aprendizado de máquina, abordagem baseada em dicionário e abordagem baseada em regras. A primeira abordagem é desenvolvida na Etapa 2, na qual um algoritmo de aprendizado de máquina supervisionado é responsável em classificar as sentenças. A segunda e a terceira abordagens são desenvolvidas na Etapa 3, na qual um dicionário de termos validados pelo especialista e regras desenvolvidas por meio de expressões regulares foram utilizados para identificar termos relevantes nas sentenças. A validação da metodologia foi realizada por meio de sua instanciação para uma área do domínio biomédico, mais especificamente usando artigos sobre a doença Anemia Falciforme. Nesse sentido, dois estudos de caso foram realizados tanto na Etapa 2 quanto na Etapa 3. O valor da acurácia obtida na classificação de sentenças foi acima de 60% e o valor da medida-F para a classe efeito negativo foi acima de 70%. Estes valores correspondem aos resultados alcançados com o algoritmo de aprendizado de máquina *Support Vector Machine* juntamente com a aplicação do filtro Remoção de Ruído. A medida-F obtida com a identificação de termos relevantes foi acima de 85% para a extração fictícia (i.e., classificação manual realizada pelo especialista) e acima de 80% para a extração real (i.e., classificação automática realizada pelo classificador). O valor de medida-F acima de 70% do classificador e o valor de medida-F acima de 80% da extração real mostra a relevância da classificação de sentenças na metodologia proposta. É importante ressaltar que sem a classificação de sentença, muitos falsos positivos seriam identificados nos artigos completos.

Palavras-chave: Pré-processamento Textual, Extração de Informação, Artigos Completos, Domínio Biomédico.

ABSTRACT

There is a large volume of unstructured information (i.e., in text format) being published in electronic media, in digital libraries particularly. Thus, the human being becomes restricted to an amount of text that is able to process and to assimilate over time. In this dissertation is proposed a methodology for textual preprocessing to extract information about disease effects in the biomedical domain papers, in order to identify relevant information from a text, to structure and to store this information in a database to provide a future discovery of interesting relationships between the extracted information. The methodology consists of four steps: **Data Entrance** (Step 1), **Sentence Classification** (Step 2), **Identification of Relevant Terms** (Step 3) and **Terms Management** (Step 4). This methodology uses three information extraction approaches from the literature: machine learning approach, dictionary-based approach and rule-based approach. The first one is developed in Step 2, in which a supervised machine learning algorithm is responsible for classify the sentences. The second and third ones are developed in Step 3, in which a dictionary of terms validated by an expert and rules developed through regular expressions were used to identify relevant terms in sentences. The methodology validation was carried out through its instantiation to an area of the biomedical domain, more specifically using papers on Sickle Cell Anemia. Accordingly, two case studies were conducted in both Step 2 and in Step 3. The obtained accuracy in the sentence classification was above of 60% and F-measure for the negative effect class was above of 70%. These values correspond to the results achieved with the *Support Vector Machine* algorithm along with the use of the Noise Removal filter. The obtained F-measure with the identification of relevant terms was above of 85% for the fictitious extraction (i.e., manual classification performed by the expert) and above of 80% for the actual extraction (i.e., automatic classification performed by the classifier). The F-measure of the classifier above of 70% and F-measure of the actual extraction above 80% show the relevance of the sentence classification in the proposed methodology. Importantly to say that many false positives would be identified in full text papers without the sentence classification step.

Keywords: Textual preprocessing, Information Extraction, Full Papers, Biomedical Domain.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
API	*Application Programming Interface* (Interface de Programação de Aplicativos)
ATR	*Automatic Term Recognition* (Reconhecimento Automático de Termo)
EI	Extração de Informação
HMM	*Hidden Markov Model* (Modelo Oculto de Markov)
HTML	*Hypertext Markup Language* (Linguagem de Marcação de Hipertexto)
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Databases* (Descoberta de Conhecimento em Banco de Dados)
KDT	*Knowledge Discovery in Texts* (Descoberta de Conhecimento em Textos)
LEP	Lista de Exclusão de Palavra
LET	Lista de Exclusão de Termo
MD	Mineração de Dados
MT	Mineração de Textos
NCBI	*National Center for Biotechnology Information*
NER	*Named Entity Recognition* (Reconhecimento de Entidade Nomeada)
PDF	*Portable Document Format* (Formato de Documento Portável)
PLN	Processamento de Língua Natural
POS	*Part-Of-Speech* (Etiquetador Gramatical)
QA	*Question & Answering* (Perguntas e Respostas)
RI	Recuperação de Informação
SCA	*Sickle Cell Anemia* (Doença Anemia Falciforme)
SVM	*Support Vector Machine* (Máquina de Vetor de Suporte)
URL	*Uniform Resource Locator* (Localizador de Recurso Uniforme)
XML	*Extensible Markup Language* (Linguagem de Marcação Extensível)

SUMÁRIO

1 INTRODUÇÃO

As informações relevantes estão mais em formato textual do que em imagens, gráficos, arquivos de música e vídeo ou até mesmo em equações. Segundo Tan (1999) e Chen (2001), 80% das informações das empresas e de conteúdo *on-line* do mundo estão em documentos textuais. Estudos têm revelado que entre 80% e 98% de todos os dados disponíveis nos computadores consistem em documentos não estruturados ou semiestruturados, como e-mails, páginas HTML, arquivos PDF e muitos outros documentos textuais (CHEUNG; LEE; WANG, 2005). Além disso, a quantidade de informação disponível eletronicamente está aumentando consideravelmente nos últimos anos (GANTZ et al., 2007).

No domínio biomédico, também existe uma grande quantidade de informação publicada por meio de artigos científicos que impossibilita e inviabiliza a leitura de todos os artigos por um ser humano. A título de exemplo, o PubMed, repositório *on-line* gerenciado pelo *National Center for Biotechnology Information* (NCBI) e pelo *National Library of Medicine*, contém mais de 18 milhões de publicações médicas, incluindo registros do MEDLINE e outros jornais científicos (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2010). O MEDLINE (http://gateway.nlm.nih.gov), sistema *on-line* de busca e análise de literatura médica, é um banco de dados que contém mais de 16 milhões de referências a artigos de revistas da área de ciências com uma maior concentração em biomedicina (NATIONAL LIBRARY OF MEDICINE, 2008). Já o Entrez é um sistema de recuperação de banco de dados integrado do NCBI que fornece acesso a um conjunto de 35 bases que juntas contêm mais de 350 milhões de registros (SAYERS et al., 2009).

A grande quantidade de informação armazenada nesses bancos de dados e disponíveis eletronicamente demonstra que existe um crescimento acelerado da produção e armazenamento das informações tanto em forma de resumos quanto de artigos científicos completos. Diante da imensa quantidade de informação disponível em formato textual, os seres humanos não são capazes de processar (i.e., ler e assimilar) toda essa informação.

Nesse contexto, abordagens de extração de informação vêm sendo utilizadas como solução para estruturar as principais informações do texto, a fim de propiciar uma futura descoberta de relacionamentos interessantes entre as informações extraídas, sem a qual seria humanamente inviável devido ao grande volume de informação. No sentido de identificar as informações relevantes do texto e com o objetivo de estruturar e de armazenar essas informações em um banco de dados, nesta dissertação é proposta uma metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico.

1.1 Motivação

Diversos trabalhos existentes na literatura extraem informação de proteína (MIKA; ROST, 2004b), interações de proteína (ONO et al., 2001) ou gene e proteína (LEONARD; COLOMBE; LEVY, 2002). A maioria dessas informações é selecionada de resumos (i.e., *abstracts*) do MEDLINE. Por mais bem resumido que seja o *abstract* não é possível escrever em poucas palavras todos os resultados obtidos com os experimentos. Muitas informações importantes deixam de ser extraídas.

Corney et al. (2004) destacam o benefício em extrair informação em artigos completos – mais da metade da informação extraída é do corpo do artigo – apesar de algumas dificuldades em converter artigos no formato PDF para formato de texto e do tempo extra exigido para processamento do texto (a análise realizada em resumo leva de 3 a 5 segundos e em artigo completo de 6 a 10 minutos, segundo Corney et al. (2004)). Na Figura 1 é mostrada a densidade de informação sobre nome de gene e proteína em um artigo completo, sendo a localização 0% e 100% o início e o fim do artigo, respectivamente. O pico de informação da parte esquerda corresponde ao resumo e o pico no meio corresponde às seções de discussão e resultado do artigo.

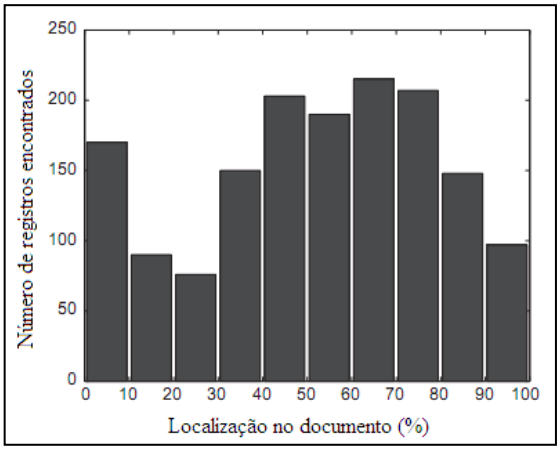

Figura 1 – Densidade de informação em um artigo completo.
Fonte: Adaptado de Corney et al. (2004).

Ainda segundo Corney et al. (2004), a Figura 1 sugere que não apenas algumas seções devam ser analisadas, mas sim o artigo completo. Schuemie et al. (2004) corroboram com a importância da extração em artigos completos: há seções que possuem mais informações do que outras e quanto mais seções pesquisadas mais informação será extraída. Contudo, Schuemie et al. (2004) ressalvam que leva-se mais tempo para processar e os artigos são mais difíceis de adquirir devido às proteções de *copyright*. Apesar dessas dificuldades, há um senso comum no que diz respeito à relevância dessas informações.

Nesse contexto, este trabalho tem como desafio extrair as informações em artigos completos sobre efeitos de doenças relacionados ao domínio biomédico.

1.2 Hipóteses e Objetivos

Os seres humanos têm a habilidade de distinguir padrões linguísticos no texto e podem superar obstáculos que os computadores não têm facilidade, como gírias, variação de ortografia e informação contextual. Entretanto, embora o ser humano tenha a capacidade de compreensão do idioma, os computadores têm a vantagem de possuir recursos como processamento de texto em alta velocidade e em grande volume (GUPTA; LEHAL, 2009). Assim, devido à dificuldade e à limitação do ser humano de processar textos à medida que as informações crescem a um grande volume, esta dissertação se baseia na seguinte hipótese:

Hipótese 1: É possível usar abordagens de extração de informação para identificar automaticamente termos relevantes do domínio biomédico. A aplicação destas abordagens em um conjunto de textos poderá ser realizada em um tempo aceitável, ou seja, em poucos segundos por artigo analisado. A aplicação destas abordagens também gera uma aceitável taxa de precisão e revocação que varia em torno de 70% a 90%. Por fim, com a aplicação destas abordagens ocorrerá uma redução da participação do especialista na identificação manual dos termos.

A maioria dos trabalhos de extração de informação extrai informação somente da seção *abstract* dos artigos, devido à praticidade de obter os documentos textuais desta fonte. Contudo, é sabido que em artigos completos existem uma maior quantidade de informação, tendo algumas seções que possuem mais informações do que em outras e quanto mais seções pesquisadas mais informação relevante poderá ser identificada e extraída (CORNEY et al., 2004; SCHUEMIE et al., 2004). Tem-se como segunda hipótese deste trabalho:

Hipótese 2: Extrair termos relacionados a efeitos de doenças no domínio biomédico de outras seções do artigo, além do seu resumo, permite obter uma maior quantidade de informação relevante. Extrair estes termos de todas as seções do artigo, no entanto, aumenta

consideravelmente a quantidade de falsos positivos e implica na análise desnecessária de seções que não discorrem sobre efeitos. Nesta dissertação, portanto, será analisada as seguintes seções: *abstract, results* e *discussion*.

Como terceira hipótese tem-se:

Hipótese 3: O uso de duas etapas separadas e consecutivas para primeiro classificar as sentenças em classes de interesse e depois para identificar e extrair termos apenas nas sentenças classificadas nestas classes de interesse possibilita um bom resultado no processo de extração de informação de termos relacionados a efeitos de doenças no domínio biomédico. O uso de dicionário, para identificar precisamente os termos conhecidos e curados que ocorrem nas sentenças e o de regras para identificar novos termos relevantes com o uso de padrões gera um bom desempenho em termos de precisão, de revocação e de medida-F para a classe efeito negativo, respectivamente, acima de 70%, acima de 85% e acima de 80%.

Portanto, o objetivo desta dissertação é propor uma metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico. A metodologia proposta é composta por quatro etapas: Entrada de Dados (Etapa 1), Classificação de Sentenças (Etapa 2), Identificação de Termos Relevantes (Etapa 3) e Gerenciamento de Termos (Etapa 4). A partir dos documentos textuais fornecidos na Etapa 1, as sentenças são classificadas em suas respectivas classes (Etapa 2). A Etapa 2 objetiva distinguir as informações de interesse das informações irrelevantes. Em seguida, na Etapa 3, os termos relevantes são identificados e extraídos das sentenças de interesse. Por fim, na Etapa 4, os termos são armazenados em um banco de dados, após a validação dos termos pelo especialista. A metodologia proposta utiliza três abordagens de extração de informação encontradas na literatura, a saber: abordagem baseada em aprendizado de máquina, abordagem baseada em dicionário e abordagem baseada em regra. A primeira abordagem é desenvolvida na Etapa 2, na qual um algoritmo de aprendizado de máquina supervisionado é responsável em classificar as sentenças. Estas sentenças são classificadas em uma das três classes: efeito positivo, efeito negativo e outros. A segunda e a terceira abordagens são desenvolvidas na Etapa 3, na qual um dicionário de termos validados pelo especialista e regras desenvolvidas por meio de expressões regulares foram utilizados para identificar termos relevantes nas sentenças. Os termos extraídos na Etapa 3 são termos sobre a classe efeito negativo.

1.3 Organização do Trabalho

Esta dissertação está organizada em oito capítulos distribuídos na seguinte ordem. **Capítulo 1**: é abordado o contexto onde este trabalho se encontra, a motivação para a definição do tema, as hipóteses e os objetivos deste trabalho; **Capítulo 2**: são apresentadas as áreas de conhecimento da mineração de textos (MT), quais os tipos mais comumente utilizados para descoberta de conhecimento textual e por fim, as etapas do processo de MT; **Capítulo 3**: é contextualizado o reconhecimento automático de termo na área da terminologia e quais as abordagens comumente utilizadas para extração de informação; **Capítulo 4**: são discutidos trabalhos que extraem informações de textos não estruturados utilizando as abordagens de dicionário, de regra e de aprendizado de máquina explicadas no Capítulo 3; **Capítulo 5**: é descrito a metodologia proposta nesta dissertação para extrair informação no domínio biomédico. **Capítulo 6**: é instanciada a metodologia proposta no Capítulo 5 com exemplos de uma área do domínio biomédico; **Capítulo 7**: é realizado um estudo de caso nas duas etapas da metodologia (i.e., classificação de sentenças e identificação de termos relevantes); e **Capítulo 8**: é concluído o trabalho, apresentando as contribuições alcançadas, as sugestões de trabalhos futuros, além das produções científicas e técnicas desenvolvidas durante o mestrado.

2 MINERAÇÃO DE TEXTOS

Mineração de Textos (MT) (TAN, 1999), também conhecida como Descoberta de Conhecimento Textual (FELDMAN; DAGAN, 1995) ou Mineração de Dados Textuais (HEARST, 1999), refere-se ao processo de extrair informações úteis em documentos no formato textual não estruturado por meio da identificação de conhecimento e exploração de padrões. A MT é vista como uma extensão da Mineração de Dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A Mineração de Dados procura desvendar conhecimento de bases de dados estruturadas e extrair padrões e tendências de grandes volumes de dados, normalmente, de um domínio específico. Diferentemente da Mineração de Textos que utiliza técnicas de pré-processamento para identificar e extrair características representativas de documentos em formato não estruturado, a MD aplica algoritmos de descoberta de padrões em dados estruturados em um banco de dados.

Na comunidade biomédica a MT é considerada como um processo de destacar a informação relevante (seja recuperando ou extraindo) de uma grande coleção de dados textuais; ajuda os especialistas do domínio a entender a grande quantidade de texto por meio da extração de informação (Seção 2.1.3), recuperação de informação (Seção 2.1.4) e descoberta de relacionamentos implícitos (Seção 2.1.5) (SPASIC et al., 2005).

A MT surge como uma tecnologia para processar esses textos seja para sumarizar (tornar um texto sucinto), categorizar (classificar em classes definidas), agrupar, obter respostas relacionadas a perguntas (perguntas e respostas) ou simplesmente extrair informação que pode ser combinada com categorização, por exemplo.

Este capítulo está dividido da seguinte forma: na Seção 2.1, são destacadas as áreas de conhecimento que estão envolvidas com a Mineração de Textos; em seguida, na Seção 2.2 são discutidos alguns tipos de descoberta de conhecimento em textos que podem ser utilizadas nas áreas de conhecimento citadas na Seção 2.1; na Seção 2.3 são apresentadas as etapas do processo de mineração; e por fim, na Seção 2.4, são apresentadas as considerações finais.

2.1 Áreas de Conhecimento relacionadas à Mineração de Textos

A Mineração de Textos é um campo de pesquisa interdisciplinar e segundo Hotho, Nürnberger e Paass (2005), está relacionada a áreas como: extração de informação, recuperação de informação, aprendizado de máquina, estatística, processamento de língua natural e mineração de dados. Esta última é utilizada como área "fim" com o objetivo de extrair padrões relevantes a partir do cruzamento de informações. As outras são utilizadas como áreas "meio", auxiliando no processo de seleção (recuperação de informação) e extração de informação (aprendizado de máquina e processamento de língua natural) (HOTHO; NÜRNBERGER; PAASS, 2005).

A seguir é apresentado sucintamente cada uma dessas áreas que contribui com a mineração textual. Essas áreas possuem técnicas que auxiliam em uma ou mais etapas do processo de Mineração de Textos explicado na Seção 2.3.

2.1.1 Processamento de Língua Natural

O termo Processamento de Língua Natural (PLN) ou Linguística Computacional é, normalmente, usado para descrever a função de um sistema de computador que analisa e sintetiza a língua falada ou escrita (JACKSON; MOULINIER, 2002). PLN tem como propósito ser apoiado por um sistema computadorizado que compreende a língua natural assim como um ser humano. A palavra "Natural" indica a distinção da escrita e fala humana das línguas não formais, mas artificiais como matemática ou lógica.

Há um grande interesse em PLN em encontrar informações relevantes e "escondidas" nas diversas fontes que se encontram em formato textual não estruturado na Internet e nas Intranets corporativas. PLN utiliza de técnicas estatísticas para interpretar e determinar o significado de partes do texto (JACKSON; MOULINIER, 2002).

PLN tem o intuito de analisar e representar naturalmente a ocorrência de textos em um ou mais níveis de análise linguística. Os níveis de análise linguística são divididos em seis categorias (JURAFSKY; MARTIN, 2000): **Fonética e Fonológica**: estudo dos sons linguísticos; **Morfológica**: estudo das características, formação e construção das palavras; **Sintática**: estudo das relações entre as partes/palavras das sentenças. Análise das palavras em uma sentença, a fim de descobrir a estrutura gramatical da mesma; **Semântica**: estudo do significado. Determinar os possíveis significados de uma sentença, incluindo desambiguação das palavras no contexto; **Pragmática**: estudo da compreensão do uso da língua em situações que requerem conhecimentos do mundo; **Discursiva**: estudo das unidades linguísticas maiores do que uma única expressão (discurso), interpretando a estrutura e o significado do texto.

Segundo Spasic et al. (2005), o primeiro passo para o processamento de texto automático é a tokenização, a qual identifica as unidades básicas do texto conhecidas como *tokens*, utilizando delimitadores explícitos como espaço em branco ou pontuação. Após a tokenização pode ser realizado o processamento léxico ou sintático, a saber:

- Léxico, inclui: lematização, processo que substitui a palavra flexionada pela forma básica sem número e gênero (cantaremos→cantar); *stemming*, processo que reduz a palavra ao seu radical (cantaremos→cant); etiquetador gramatical (*Part-Of-Speech* - POS) identifica a categoria gramatical de cada palavra do texto utilizando a marcação de etiquetas. Geralmente é morfológico (identifica substantivo, adjetivo, artigo) ou morfossintático (identifica as funções sintáticas como sujeito, predicado, aposto).
- Sintático envolve a análise da estrutura sintática de uma sentença, inclui: *shallow parser* e *deep parser*. A primeira identifica os grupos nominais e verbais, como sintagmas; enquanto que a segunda gera representação completa da estrutura gramatical de uma sentença.

2.1.2 Aprendizado de Máquina

Aprendizado de Máquina (AM) é uma área da Inteligência Artificial que lida com problemas de aprendizado computacional a fim de adquirir conhecimento de forma automática. Um sistema de aprendizado tem a função de analisar informações e generalizá-las, para a extração de novos conhecimentos. Para isso usa-se um programa de computador para automatizar o aprendizado (MONARD; BARANAUSKAS, 2003).

O aprendizado utiliza do princípio da indução (inferência lógica) com o intuito de obter conclusões genéricas a partir de um conjunto de exemplos. Um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados. As hipóteses geradas por meio dessa inferência podem ou não preservar a verdade.

Para a indução derivar conhecimento novo representativo, os exemplos das classes têm que estar bem definidos e ter uma quantidade suficiente de exemplos, obtendo assim hipóteses úteis para um determinado tipo de problema. Quanto mais exemplos relevantes selecionados para treinamento no indutor, mais bem classificado será o novo conjunto de dados. O objetivo do algoritmo de indução é construir um classificador que possa determinar a classe a qual um exemplo não rotulado pertence. É possível rotular um novo exemplo devido à generalização.

O aprendizado indutivo pode ser dividido em supervisionado (AS) e não supervisionado (ANS), como pode ser visto na Figura 2. O AS é utilizado para classificação

dos exemplos em classes predefinidas: resolve problemas preditivos. O ANS é utilizado para agrupamento, agrupando exemplos semelhantes: resolve problemas descritivos. Classificação e agrupamento são, respectivamente, exemplos desses dois tipos de aprendizado.

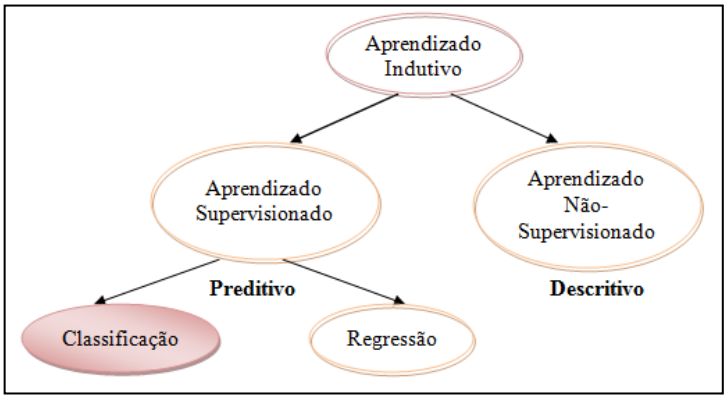

Figura 2 – Hierarquia do aprendizado.
Fonte: Adaptado de Monard e Baranauskas (2003).

Como as classes já estão definidas, o interesse deste trabalho encontra-se no aprendizado supervisionado, mais especificamente na classificação que trabalha com rótulos de classes discretos (e.g., paciente normal, paciente com doença A), diferentemente da regressão que lida com valores contínuos (e.g., pacientes maiores de 18 anos com altura de 1,8 metros).

Monard e Baranauskas (2003) classificam AM em alguns paradigmas, a saber: **Simbólico**, representações simbólicas de um problema por meio da análise de exemplos e contraexemplos como expressão lógica, árvore de decisão, regras ou rede semântica. Exemplo: Algoritmos de árvore de decisão como ID3, C4.5; **Estatístico**, utiliza modelos estatísticos para encontrar uma aproximação do conceito induzido. Exemplo: *Support Vector Machine* (SVM) e aprendizado Bayesiano; **Baseado em Exemplos**, classifica um novo exemplo com base em uma classificação similar conhecida. Exemplo: Raciocínio baseado em caso e método do k-vizinhos mais próximos (*k-nearest neighbor*, *kNN*); **Conexionista**, inspirada no modelo biológico do sistema nervoso. Exemplo: Redes Neurais; e **Evolutivo**, modelo biológico de aprendizado. Exemplo: Analogia com a teoria de Darwin.

Medidas Utilizadas pelo Classificador

A taxa de erro (também conhecida como taxa de classificação incorreta) é uma medida comumente utilizada para avaliar um classificador. Sendo n o número de exemplos, o $erro(h)$, calculado pela Equação (1), compara a classe verdadeira de cada exemplo y_i com o rótulo atribuído pelo classificador induzido $h(x_i)$. A expressão $\|y_i \neq h(x_i)\|$ retorna 1 se a condição for verdadeira e zero caso contrário.

$$erro(h) = \frac{1}{n}\sum_{i=1}^{n} \|y_i \neq h(x_i)\| \quad (1)$$

O complemento da taxa de erro é a precisão do classificador, denotada por $precisão(h)$, Equação (2).

$$precisão(h) = 1 - erro(h) \quad (2)$$

Há um limiar (erro máximo) que é estabelecido para um classificador. O erro chamado de erro majoritário é calculado em um conjunto de exemplos T a partir da distribuição das classes, Equação (3).

$$erro\text{-}maj(T) = 1 - \max_{i=1,\cdots,k} dist(C_i) \quad (3)$$

Matos et al. (2009a) apresentam mais detalhes das medidas de desempenho utilizadas para avaliar um classificador, além de destacar os métodos de particionamento dentre os quais destaca-se o k-*Fold Cross-Validation* (os exemplos são treinados em k-1 *folds* e testados com o *fold* restante. Este processo é repetido k vezes, cada vez considerando um *fold* diferente para teste) e algumas das técnicas de seleção de características utilizadas com o objetivo de reduzir a dimensionalidade dos dados.

2.1.3 Extração de Informação

Extração de Informação (EI) preocupa-se em localizar partes específicas em documentos em língua natural, extraindo informações estruturadas a partir de textos não estruturados. EI pode ser descrita como: Extração de fatos essenciais de tipos predefinidos de um documento em língua natural; Representação de cada fato como um modelo (*template*) cujo *slots* são preenchidos com base no que é encontrado no texto (ANANIADOU; MCNAUGHT, 2006).

Em geral, sistemas de EI são úteis: se a informação a ser extraída é encontrada de forma explícita e nenhuma conclusão é necessária; quando um pequeno número de modelos é suficiente para resumir as partes relevantes do documento; e a informação desejada é expressa localmente no texto (FELDMAN; SANGER, 2007).

Ainda segundo Feldman e Sanger (2007), técnicas de EI podem ser parte da tarefa de Mineração de Textos para facilitar a extração do conhecimento. Contudo, o domínio deve ter um padrão que se encaixa num modelo. Os resultados da EI, informações estruturadas, geralmente são armazenados em um banco de dados para, posteriormente, serem utilizados em algoritmos de Mineração de Dados para identificar padrões interessantes.

Existem cinco tarefas de EI como mostrados na Tabela 1. Todos os tipos são fracamente dependentes de domínio, como exemplificado por Cunningham (2006): mudar o domínio a ser processado de notícias financeiras para outro tipo de notícias implica algumas alterações no sistema; mudar o assunto de notícias para artigos científicos envolve grandes mudanças.

Tabela 1 – Cinco tarefas de extração de informação.
Fonte: McNaught e Black (2006).

Tarefa	Descrição
Entidade Nomeada	Extrai nome, lugares, etc.
Correferência	Identifica relações entre entidades
Template Element	Extrai atributos descritivos de entidade nomeada
Template Relation	Extrai relacionamento específico de entidade nomeada (simples fatos)
Scenario Template	Extrai eventos. Um ou mais *slots* são preenchidos com *template element* ou *template relation* para cada tipo de evento extraído

No domínio biomédico a tarefa mais utilizada é o Reconhecimento de Entidade Nomeada (PARK; KIM, 2006). Reconhecimento de Entidade Nomeada (NER) identifica referências para tipos de objetos particulares, como nome de pessoas, empresas e localizações. NER é uma das áreas de extração de informação mais estudadas, não apenas em biomedicina, mas também em outras áreas (ANANIADOU; MCNAUGHT, 2006).

2.1.4 Recuperação de Informação

Segundo Lancaster (1968 apud VAN RIJSBERGEN, 1979), um sistema de Recuperação de Informação (RI) não informa ao usuário sobre o assunto que o mesmo deseja encontrar, mas limita-se a localizar os documentos relativos à consulta do usuário e informar sobre a existência ou não da informação desejada.

RI é diferente de EI. Esta última extrai informações relevantes não estruturadas dos documentos. Uma aplicação de EI analisa os textos não estruturados e apresenta as informações específicas no formato estruturado predefinido como explicado na Seção 2.1.3. Um sistema de RI recupera documentos relevantes baseado em uma consulta do usuário e baseia-se em busca por palavras-chave ou busca por similaridade.

O exemplo mais conhecido de um sistema de RI é o buscador Google que seleciona os documentos disponíveis na Web de acordo com a consulta definida pelo usuário. Segundo Jensen, Saric e Bork (2006), o PubMed é o sistema de RI mais conhecido no domínio biomédico. Este último é um sistema *ad hoc* que usa duas metodologias de RI: modelo booleano e de vetor.

Para processar grandes coleções de documento, a Mineração de Textos exige um grande poder computacional e tempo para analisar os textos. A RI pode contribuir restringindo os documentos com base na informação desejada do usuário, reduzindo o número de documentos que serão analisados e consequentemente diminuindo o tempo de espera. Por exemplo, Hu et al. (2005) extraem informação sobre fosforilação de resumos do MEDLINE. Para isso, numa primeira fase utiliza-se do sistema de RI para selecionar somente artigos relacionados à fosforilação, delimitando o domínio de extração.

No contexto deste trabalho, a Recuperação de Informação não será utilizada, pois os artigos no formato em PDF serão pré-selecionados pelo especialista do domínio.

2.1.5 Mineração de Dados

Mineração de Dados (MD) (HAN; KAMBER, 2006) faz parte do processo de Descoberta de Conhecimento em Banco de Dados (KDD) e às vezes os dois termos são usados de maneira indistinta (LUO, 2008). O processo de KDD pode consistir dos seguintes passos: seleção dos dados, pré-processamento (limpeza dos dados), transformação dos dados, busca por padrão (mineração de dados) e interpretação e avaliação dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O objetivo do processo de KDD é a aplicação de métodos específicos de mineração de dados para extração e descoberta de padrão. Contudo, o KDD utiliza-se de técnicas de aprendizado de máquina e estatística para avaliar e interpretar os padrões minerados e determinar quais padrões podem ser considerados como conhecimento novo.

Uma área que pode contribuir com o processo de KDD é o *data warehousing* que, segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), ajuda com o armazenamento de grandes volumes de dados, com a limpeza e com o acesso aos dados. O principal objetivo do *data warehousing* é fornecer suporte à decisão aos gerentes de negócios.

2.2 Descoberta de Conhecimento em Textos

Descoberta de Conhecimento em Textos (ou simplesmente KDT) visa explorar e descobrir padrões em textos. O termo KDT foi cunhado por Feldman e Dagan (1995) como sinônimo para Mineração de Textos.

Zweigenbaum et al. (2007) descrevem três tarefas de KDT (perguntas e respostas, sumarização e geração de hipótese) que podem ser utilizadas como uma tarefa para extrair informação. Fan et al. (2006) descrevem mais cinco tarefas de KDT além das três citadas anteriormente: extração de informação, categorização, agrupamento, visualização de informação e rastreamento de informação.

Nota-se que Zweigenbaum et al. (2007) consideram as três tarefas como parte da extração de informação; Fan et al. (2006) incluem a extração de informação como uma tarefa de KDT. Neste trabalho considera-se extração de informação (explicada na Seção 2.1.3) como uma área de conhecimento abrangente no contexto da Mineração de Textos e por isso, não será explicada nesta seção.

A seguir são apresentadas algumas dessas tarefas que podem ser utilizadas para descoberta de padrão em textos: perguntas e respostas, sumarização, categorização e agrupamento.

2.2.1 Perguntas e Respostas

Mollá e Vicedo (2007) definem Perguntas e Respostas (tradução do inglês *Question Answering* - QA) como uma tarefa onde é possível um computador responder perguntas arbitrárias formuladas em língua natural. Pesquisa em QA está sendo desenvolvida a partir de duas diferentes perspectivas científicas: Inteligência Artificial e Recuperação de Informação.

Diferentemente de retornar uma lista de documentos a partir de grandes coleções de texto (objetivo da Recuperação de Informação discutida na Seção 2.1.4), QA tenta fornecer respostas curtas e específicas para perguntas, disponibilizando informações de apoio relacionadas à fonte original do documento, caso o usuário queira verificar a origem da informação (ZWEIGENBAUM et al., 2007).

Sistemas de QA são especialmente úteis em situações em que o usuário precisa conhecer uma informação específica, por exemplo, "Qual a quantidade anual de crianças brasileiras que nascem com a doença X?" Geralmente deseja-se obter estas informações em tempo hábil sem perder muito tempo pesquisando.

Inicialmente sistemas de QA foram desenvolvidos para aplicações genéricas e mais recentemente em domínios restritos (ZWEIGENBAUM et al., 2007). Mollá e Vicedo (2007) descrevem informações sobre QA em domínio restrito e apresentam uma lista de sistemas de Perguntas e Respostas como JAVELIN, QUETAL, AQUA e START.

2.2.2 Sumarização

O objetivo da sumarização de texto automático é identificar as informações importantes de um documento e apresentá-las de forma sucinta e coerente. Diferentemente de sistemas de EI que geralmente tem como entrada (para processamento) textos, em sistemas de sumarização a entrada é uma coleção de documentos. O resultado desses sistemas também difere: no primeiro são geradas informações estruturadas, enquanto no segundo a informação representa uma síntese do documento original (ZWEIGENBAUM et al., 2007).

Sumarização pode ajudar usuários a encontrar rapidamente os pontos principais de um documento. Radev, Hovy e McKeown (2002) definem um sumário como um texto que é produzido de um ou mais textos que expressa informação essencial dos textos originais e não é maior do que a metade desses e geralmente é menos representativo do que os mesmos; o objetivo principal é apresentar um resumo das principais ideias de um documento.

Os sistemas de sumarização variam dependendo da tarefa a ser realizada. Afantenos, Karkaletsis e Stamatopoulos (2005) discutem alguns dos fatores que se devem pensar quando se intenciona gerar sumários: tipos de entradas de documentos (um ou mais documentos, monolíngue ou multilíngue, texto ou multimídia como imagem e vídeo); propósito do sumário (informativo, indicativo ou crítico; domínio específico ou genérico); e possíveis maneiras de apresentação do sumário (extrato ou resumo).

2.2.3 Categorização

Como explicado na Seção 2.1.2, a categorização é uma tarefa de aprendizado supervisionado. A categorização de documento, também conhecida como classificação, é uma tarefa importante na Mineração de Textos. Segundo Ikonomakis, Kotsiantis e Tampakas (2005), classificação de texto desempenha papel importante na extração de informação, sumarização, busca de texto e perguntas e respostas. Na comunidade científica a abordagem dominante para categorização é baseada em técnicas de aprendizado de máquina (SEBASTIANI, 2002).

Classificação de texto é a tarefa de classificar um documento em categorias predefinidas. O processo de classificação é apresentado de forma geral na Figura 3. Um conjunto de documentos pré-classificados em categorias é considerado para treinamento (**a**). Este é analisado a fim de derivar um modelo de classificação (**b**). Esse modelo muitas vezes precisa ser refinado em um processo de teste (não mostrado na figura) para validar o aprendizado. Assim, o esquema de classificação validado pode ser utilizado para a classificação de outros documentos (**c**), classificando o documento (**d**) nas categorias definidas anteriormente (**a**).

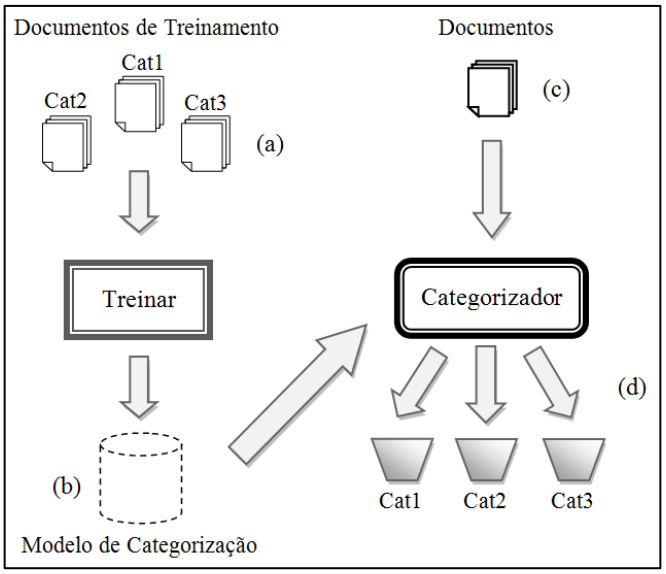

Figura 3 – Categorização de documentos.
Fonte: Adaptado de Dörre, Gerstl e Seiffert (1999).

2.2.4 Agrupamento

Como explicado na Seção 2.1.2, agrupamento é uma tarefa de aprendizado não supervisionado. O processo de agrupamento é apresentado de forma geral na Figura 4. Diferentemente da categorização, o agrupamento irá agrupar os documentos (**a**) sem o conhecimento de nenhuma categoria pré-classificada, separando os grupos com base na similaridade dos dados (**b**).

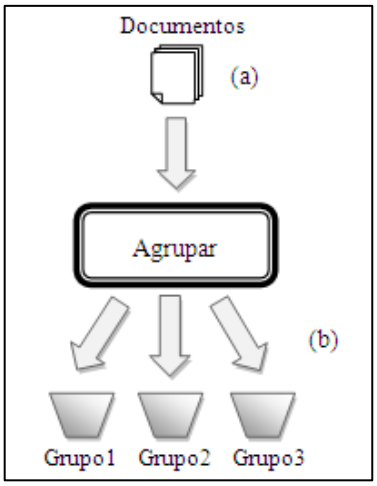

Figura 4 – Agrupamento de documentos.
Fonte: Adaptado de Dörre, Gerstl e Seiffert (1999).

O agrupamento lida com problema que consiste em grande parte em analisar os dados de entrada e sugerir um grupo, de acordo com similaridades observadas nos dados. É extremamente útil quando não se tem o conhecimento prévio do domínio.

2.3 Etapas do Processo de Mineração de Textos

Geralmente o processo de Mineração de Textos é dividido em quatro etapas (IMAMURA, 2001; MARTINS, 2003): coleta de documentos, pré-processamento, extração de padrões, e análise e avaliação dos resultados. Na coleta de documentos automática utiliza-se de ferramentas para recuperar informação e auxiliar o usuário a encontrar a informação que deseja mais rapidamente; após recuperar os documentos textuais, é realizado um pré-processamento para estruturar os mesmos e em seguida extrair informações relevantes; tendo os dados armazenados, por exemplo, em um banco de dados, padrões podem ser extraídos a fim de encontrar informações úteis; por último, deseja-se avaliar o resultado gerado a partir dos passos anteriores.

Apesar dessas quatro etapas, há algumas variações do processo de MT na literatura como em Rezende (2003), Mathiak e Eckstein (2004), Fan et al. (2006), Stavrianou, Andritsos e Nicoloyannis (2007), Feldman e Sanger (2007) e Aranha (2007). A seguir é resumida cada uma das quatro etapas conforme mostrada na Figura 5.

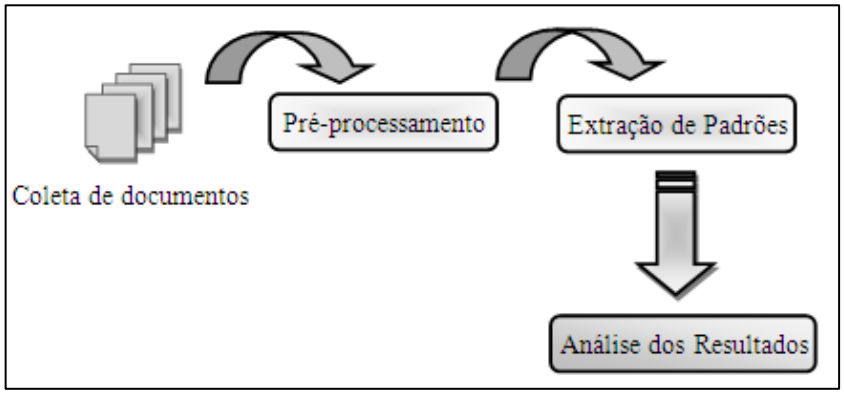

Figura 5 – Processo de Mineração de Textos em quatro etapas.

Neste trabalho de pesquisa em nível de mestrado é proposta uma metodologia para extrair informação que atua na fase de pré-processamento.

2.3.1 Coleta de Documentos

A primeira fase do processo de MT é a localização dos documentos que serão utilizados nas fases posteriores. Um dos problemas para coletar esses documentos é descobrir onde os dados estão armazenados. Existem vários locais onde essas informações possam ser encontradas como biblioteca em documentos impressos ou mídias digitais, computador em arquivos armazenados no disco rígido, e de forma geral e abrangente, na Internet. Esta última é um repositório de uma infinidade de documentos espalhados pela rede.

Para auxiliar na coleta de documentos, existem vários motores de busca (*search engines*). Estes motores são sistemas computacionais criados para localizar informação a

partir de palavras-chave e tem como objetivo auxiliar a encontrar uma informação. O mais conhecido desses motores de busca é o Google. No domínio biomédico encontram-se vários desses motores que auxiliam os pesquisadores a encontrar um artigo de forma rápida e precisa como o Entrez (NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION, 2009), cujo repositório armazena 350 milhões de registros correspondentes a 35 diferentes bases, incluindo o PubMed e PubMed Central.

2.3.2 Pré-processamento

O principal objetivo da etapa de pré-processamento de textos, segundo Aranha (2007), é estruturar os dados para serem submetidos a algum algoritmo de indexação ou Mineração de Dados. Ainda segundo Aranha (2007), pré-processamento normalmente significa dividir o texto em palavras (tokenizar), aplicar técnicas de *stemming*, remover as *stopwords* e classificar as palavras segundo a classe gramatical (técnicas de PLN discutidas na Seção 2.1.1). Stavrianou, Andritsos e Nicoloyannis (2007) aconselham analisar o texto antes de, por exemplo, remover as *stopwords* ou aplicar técnicas de lematização no texto, pois cada problema tem necessidades diferentes. Portanto, uma técnica que serve para uma aplicação pode não servir para outra.

Dois trabalhos encontrados na literatura atuam na fase de pré-processamento: Imamura (2001) e Aranha (2007). O primeiro projeta e constrói um módulo de pré-processamento de texto em português. O segundo apresenta um novo modelo de pré-processamento para minerar textos em português, utilizando técnicas de inteligência computacional. Segundo Carrilho Junior (2007), pré-processar textos é a fase mais oneroso do processo de MT, uma vez que não existe somente uma técnica que possa ser aplicada para, por exemplo, extrair informação de proteína no domínio biomédico.

2.3.3 Extração de Padrões

Após os documentos serem estruturados adequadamente, técnicas de extração de conhecimento podem ser utilizadas para identificar padrões e tendências nos dados. Algoritmos de Mineração de Dados são desenvolvidos para encontrar esses padrões. Segundo Aranha (2007), esses algoritmos são provenientes de diversas áreas de conhecimento como: aprendizado de máquina, estatística, redes neurais e banco de dados.

2.3.4 Análise e Avaliação dos Resultados

No final do processo utiliza-se de métricas para avaliar se o resultado gerado a partir dos passos anteriores está adequado. Nota-se que essas medidas também servem para validar

cada um dos passos anteriores individualmente: na coleta de documentos pode-se avaliar a qualidade da recuperação da informação; no pré-processamento, avaliar a qualidade da extração de informação; e na extração de padrões, avaliar o quão confiável são os padrões identificados.

A seguir são destacadas as principais métricas utilizadas em sistemas de extração de informação como precisão, revocação e medida-F e em sistemas de aprendizado de máquina como acurácia.

Precisão e revocação são medidas amplamente utilizadas para avaliar a qualidade dos resultados em diversas áreas do conhecimento. Precisão é uma medida de fidelidade, enquanto a revocação (conhecida também como cobertura ou sensibilidade) é uma medida de completude.

As medidas de precisão e revocação são medidas padrão da Recuperação de Informação (RI), Cleverdon (1966 apud SILVA, 2006). As mesmas são utilizadas para contribuir com a avaliação de sistemas de RI que tem o objetivo de recuperar documentos relevantes a partir da consulta de um usuário, porém diversas outras áreas, como Extração de Informação e Inteligência Artificial incluindo Aprendizado de Máquina e Processamento de Língua Natural, utilizam dessas medidas para avaliação.

A seguir, tomando como base as informações contidas na Tabela 2, as seguintes medidas são definidas.

Tabela 2 – Matriz de confusão de duas classes (Positivo/Negativo).

Condição Atual / Teste	***P***	***N***
p	Verdadeiro Positivo (VP)	Falso Positivo (FP)
n	Falso Negativo (FN)	Verdadeiro Negativo (VN)

- **Precisão:** Taxa com que todos os exemplos classificados como positivos são realmente positivos. Nenhum exemplo negativo é incluído.

$$Precisão = \frac{VP}{VP + FP} \tag{4}$$

- **Revocação:** Taxa com que se classifica como positivos todos os exemplos que são positivos. Nenhum exemplo positivo é deixado de fora. Apresenta uma indicação do quanto do total de informação relevante foi recuperada.

$$Revocação = \frac{VP}{VP + FN} \tag{5}$$

- ***Medida-F* (*F-Measure*)**: Média harmônica ponderada da precisão e revocação. Considera-se $P = Precisão$ e $R = Revocação$.

$$Medida - F = \frac{2 \times P \times R}{P + R} \quad (6)$$

- **Acurácia**: Mais frequentemente utilizada para avaliação de problemas de classificação de aprendizado de máquina.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (7)$$

Mais detalhes com exemplos destas medidas podem ser encontradas no relatório técnico sobre métricas de avaliação em Matos et al. (2009b).

2.4 Considerações Finais

Neste capítulo foram discutidas algumas áreas de conhecimento (Seção 2.1) que contribuem para minerar textos: Processamento de Língua Natural fornece algumas técnicas como POS, *shallow parser* e *deep parser* que podem ser usadas para processar textos; Aprendizado de Máquina Supervisionado classifica novos exemplos a partir do treinamento de exemplos expressivos; Extração de Informação (EI) extrai informações relevantes em dados não estruturados; Recuperação de Informação (RI) contribui parcialmente com a extração, restringindo a quantidade de documentos a serem processados pela EI; e Mineração de Dados (MD) identifica padrões a partir de dados estruturados armazenados em um banco de dados.

A MD é uma área de descoberta de conhecimento que fornece algoritmos para extrair padrões interessantes que são difíceis de serem examinados manualmente. Pode atuar na descoberta de padrões identificando relacionamentos e tendências em dados estruturados. Também foi destacada a diferença entre EI e RI, e entre MD e Mineração de Textos (MT).

Foram apresentadas várias tarefas de descoberta de conhecimento em textos dentre as quais: perguntas e respostas, encontra resposta para uma pergunta realizada pelo usuário; sumarização, identifica as informações importantes de um texto e apresenta de forma sucinta e coerente; categorização, classifica um documento em categorias predefinidas; e agrupamento, agrupa um documento em grupos que são definidos a partir da análise dos dados, diferentemente da categorização onde as categorias são conhecidas

Por fim, foi apresentado um processo de MT em quatro etapas (coleta de documentos, pré-processamento, extração de padrões, e análise e avaliação dos resultados), o qual foi sintetizado com base em algumas propostas encontradas na literatura. Foram apresentadas

algumas técnicas de pré-processamento, área de estudo deste trabalho e também foram discutidas métricas que podem ser utilizadas para avaliar o resultado das etapas do processo de MT.

No próximo capítulo será discutido sobre Extração Automática, mais especificamente Reconhecimento Automático de Termo, que reconhece e extrai unidades léxicas de documentos, as quais correspondem a conceitos de domínio.

3 EXTRAÇÃO AUTOMÁTICA

Reconhecimento de Entidade Nomeada (NER) refere-se à tarefa de reconhecimento de entidades como nome de pessoas e nome de empresas. No domínio biomédico, as entidades são genes, proteínas e doenças. Segundo Park e Kim (2006), NER é diferente de Reconhecimento Automático de Termo (ATR). Enquanto NER classifica tipos conhecidos de entidades do mundo real, ATR associa um dado termo com um conceito em um *framework* semântico bem definido.

Estas duas áreas de pesquisa, NER e ATR, no entanto também se misturam. Segundo Sekine (2004), há um relacionamento entre a pesquisa de terminologia e a entidade nomeada; na área biomédica, por exemplo, os nomes de proteínas e genes são certamente termos. Para extrair esses termos são utilizadas algumas técnicas herdadas de NER; Segundo Park e Kim (2006), um sistema de reconhecimento de termo pode utilizar um módulo de NER para reconhecer entidades nomeadas no texto. Para ratificar a mistura desses conceitos, tanto NER (PARK; KIM, 2006) quanto ATR (ANANIADOU; NENADIC, 2006) utilizam das mesmas abordagens (i.e., abordagens baseadas em dicionário, regras e aprendizado de máquina) para a extração de informação.

Neste trabalho é utilizado o Reconhecimento Automático de Termo (Seção 3.1) para a extração de informação (Seção 3.2).

3.1 Reconhecimento Automático de Termo

A terminologia representa na literatura biomédica um dos principais desafios para a Mineração de Textos (ANANIADOU; MCNAUGHT, 2006). Uma vez que existe uma grande quantidade de neologismos na terminologia biomédica, é necessário fornecer ferramentas que extraiam automaticamente novos termos para associá-los a bancos de dados biomédicos, vocabulários controlados e ontologias. O processamento terminológico abrange aspectos como a extração, tratamento da variação de termos, classificação e mapeamento de termos.

Segundo Ananiadou, Friedman e Tsujii (2004), os termos textuais encontrados na literatura biomédica, como nomes de genes, proteínas, organismos, drogas e produtos químicos, representam conceitos de domínio utilizados pela comunidade científica e seria

impossível compreender ou extrair informação de um artigo sem a identificação e a associação precisa desses termos.

Para ajudar na identificação e extração dos termos utiliza-se o Reconhecimento Automático de Termo (ATR) que pode ser dividido em três passos (Figura 6): o **Reconhecimento de Termo** diferencia os termos dos não termos; a **Classificação de Termo** classifica os termos reconhecidos em classes do domínio; e o **Mapeamento de Termo** associa automaticamente termos com novos conceitos representados por uma ontologia.

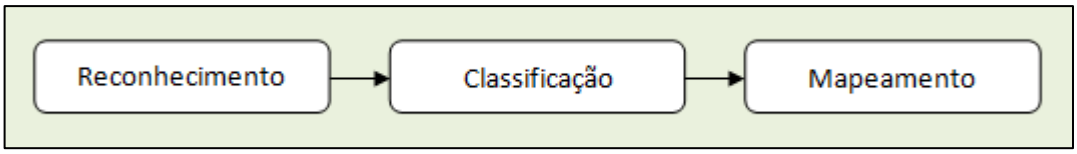

Figura 6 – Passos para a identificação de termos no texto.
Fonte: Krauthammer e Nenadic (2004).

Alguns dos passos podem ser realizados conjuntamente, por exemplo, reconhecimento e classificação de termos que, respectivamente, podem identificar os termos e associá-los as classes predefinidas do domínio biomédico (e.g., genes, proteínas ou doenças). Segundo Krauthammer e Nenadic (2004), a separação ou não das fases com o objetivo de melhorar a identificação de termos ainda é uma questão aberta de pesquisa. Obviamente, caso estes passos estejam separados, diferentes soluções para identificar termos podem ser utilizadas em problemas específicos.

Variações léxicas, sinônimos (i.e., conceito representado com vários termos) e termos homônimos (i.e., termos com vários significados) são obstáculos que impedem que alguns termos sejam identificados precisamente no texto. Identificar precisamente os termos é difícil devido às mudanças constantes seja por um termo que aparece por um pequeno período ou outros termos que aparecem frequentemente, mas depois deixam de aparecer no texto. O problema também é a falta de padronização dos nomes. Existem alguns guias para criação de novos tipos de entidades biomédicas, todavia as orientações nem sempre são seguidas. Portanto, esses nomes sem padrão são um obstáculo para sistemas de identificação automática de termos (KRAUTHAMMER; NENADIC, 2004).

A importância da terminologia desencadeou pesquisa significativa na área biomédica que resultou em várias abordagens utilizadas para selecionar, classificar e identificar ocorrências de termos em textos biomédicos (ANANIADOU; MCNAUGHT, 2006). Neste trabalho de pesquisa em nível de mestrado, concentra-se em reconhecer termos (i.e., identificar e extrair termos) do domínio biomédico, utilizando para isso a combinação de três abordagens para extração de informação (i.e., aprendizado de máquina, dicionário e regras). O

reconhecimento de termos permitirá o preenchimento de instâncias de tipos-relacionamento em um banco de dados, assim como a carga de novos termos ainda não presentes neste banco de dados. A metodologia proposta será explicada no Capítulo 5.

3.2 Abordagens para Extração de Informação

Cohen e Hunter (2008) apresentam duas abordagens para extração de informação: abordagem baseada em regras e baseada em aprendizado de máquina. A primeira faz o uso de algum tipo de conhecimento; a segunda utiliza-se de classificadores para classificar sentenças ou documentos. Krauthammer e Nenadic (2004) e Ananiadou e Nenadic (2006) apresentam uma terceira abordagem, além dessas duas anteriores: abordagem baseada em dicionário que utiliza informações de um dicionário para auxiliar na identificação dos termos ou das entidades no texto. Essas abordagens são as três predominantes para extração de informação no domínio biomédico.

Cada uma dessas abordagens tem vantagens e desvantagens. Frequentemente gasta-se tempo significativo para desenvolver um sistema baseado em regras, as quais são dependentes do domínio. Enquanto que um sistema baseado em aprendizado de máquina tipicamente exige uma grande quantidade de dados para treinamento. Costuma-se utilizar uma combinação das duas abordagens: geralmente classifica-se os documentos e em seguida, utilizam-se regras para extrair os termos (COHEN, K; HUNTER, 2008).

Em seguida serão apresentadas essas três abordagens na seguinte ordem: abordagem baseada em dicionário (Seção 3.2.1), abordagem baseada em regras (Seção 3.2.2) e abordagem baseada em aprendizado de máquina (Seção 3.2.3).

3.2.1 Abordagem Baseada em Dicionário

A abordagem baseada em dicionário utiliza uma lista de termos para identificar ocorrências de termos no texto. Casamento de padrão geralmente é utilizado entre as entradas contidas no dicionário e as palavras encontradas nas sentenças. Nadeau e Sekine (2007) apresentam algumas técnicas que pode ser utilizadas para reconhecimento e classificação de entidade nomeada como $stemming$ e lematização (apresentadas na Seção 2.1.1), distância de edição (TSURUOKA; TSUJII, 2003 apud NADEAU; SEKINE, 2007) e algoritmo *Soundex* (RAGHAVAN; ALLAN, 2004 apud NADEAU; SEKINE, 2007).

Bancos de dados biológicos armazenam informações de conceitos da biologia como genes, estrutura de proteínas, informações sobre reações químicas, doenças e organismos (REBHOLZ-SCHUHMANN; KIRSCH; COUTO, 2005). Ainda segundo Rebholz-

Schuhmann, Kirsch e Couto (2005), alguns recursos terminológicos podem ajudar a relacionar essas informações biológicas, que são citadas em publicações científicas, com informações armazenadas em um banco de dados. Exemplos desses recursos são: *Gene Ontology* (GO) e *Unified Medical Language System* (UMLS).

Ao contrário dos nomes de pessoas e locais no domínio geral, nomes de proteínas e genes têm sido gerenciados por meio de banco de dados por grandes organizações como o NCBI (http://www.ncbi.nlm.nih.gov/) e o *European Bioinformatics Institute* (http://www.ebi.ac.uk/) (PARK; KIM, 2006). Exemplos desses bancos de dados são: LocusLink, informações de gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene); SWISS-PROT, proteínas (http://www.expasy.org/sprot/); FlyBase, informações de gene, especificamente da Drosófila, espécie de pequenas moscas (http://flybase.org/); e GenBank, sequências de nucleotídeos e aminoácidos (http://www.ncbi.nlm.nih.gov/Genbank/).

Em seguida são apresentados alguns trabalhos que utilizam a abordagem baseada em dicionário que, por sua vez, é povoado com informações de algum banco de dados, como os citados anteriormente. São descritos trabalhos que extraem informações sobre gene, proteína e interações de proteína-proteína. Algumas técnicas são utilizadas para aumentar a probabilidade das palavras do dicionário serem identificadas no texto, como: *stemming* (ONO et al., 2001) e lematização (SCHUEMIE et al., 2007) para padronizar, respectivamente, palavras pelo radical e pela forma básica sem número e gênero; *stopwords* (KOU; COHEN; MURPHY, 2005) para diminuir os falsos positivos; e aproximação de string (KRAUTHAMMER et al., 2000; TSURUOKA; TSUJII, 2004) que calcula a similaridade entre palavras.

Trabalhos que utilizam Dicionário para Extração de Informação

As principais informações dos trabalhos que serão apresentados a seguir são resumidas na Tabela 3. Destacam-se as técnicas utilizadas para extrair informação e se foi utilizado um etiquetador *Part-Of-Speech* (POS), qual o dicionário utilizado, o domínio de atuação da extração, e por fim, quais os valores de precisão e revocação que foram obtidos na extração de informação.

Krauthammer et al. (2000) combinam nomes de proteína e gene contidos em um dicionário com o BLAST (i.e., ferramenta de comparação de sequências de nucleotídeos e aminoácidos). Os nomes são convertidos em uma sequência de nucleotídeos que é o formato de entrada do BLAST, substituindo cada caractere do nome com uma combinação única de nucleotídeo (exemplo na Tabela 4). Esses nomes são extraídos do banco de dados GenBank.

Dos nomes que não foram incluídos no banco de dados 4,4% foram identificados. Precisão e revocação obtidas foram, respectivamente, de 71,7% e 78,8%.

Tabela 3 – Resumo dos trabalhos com dicionário.

Autor	Técnicas	Etiquetador POS	Dicionário	Domínio	Precisão	Revocação
Krauthammer et al. (2000)	Aproximação (BLAST)	Não	GenBank[1]	Proteína e Gene	71,7%	78,8%
Ono et al. (2001)	Etiquetador de Brill e *stemming*	Sim	Construção manual	Interações proteína-proteína	94%	84,6%
Tsuruoka et al. (2004)	Aproximação (distância de edição) e expansão com UMLS	Não	UMLS[2]	Proteína	71,7%	62,3%
Egorov et al. (2004)	Tokenização e abreviação	Não	LocusLink e outras bases[1]	Proteína (mamífero)	98%	88%
Kou et al. (2005)	HMM, Etiquetador de Brill e *stopwords*	Sim	PIR-NREF[1]	Proteína	50,1%	68,8%
Schuemie et al. (2007)	Várias regras	Não	Vários bancos de dados	Proteína e Gene	-------	-------

[1] Banco de dados utilizado para construir o dicionário.
[2] Recurso terminológico utilizado para obter termos biomédicos.

Tabela 4 – Exemplo de tradução para o formato do BLAST.
Fonte: Adaptado de Krauthammer et al. (2000).

Tabela de conversão	A	AAAC	G	AAGC	P	ACCC	Z	AGAT	1	AGCG
Exemplo	zgap1				AGATAAGCAAACACCCAGCG					

Ono et al. (2001) propõem um método para extrair informação de interações de proteína-proteína de resumos do MEDLINE utilizando um dicionário que contém nomes de proteínas (i.e., somente proteínas de levedura e *Escherichia coli*), padrões de palavra e simples regras de POS (etiquetador de Brill). O método de extração automática identifica nomes de proteínas na sentença, usando um dicionário construído manualmente. Em seguida, a sentença é processada com regras de POS e então, são extraídas interações de proteína-proteína utilizando casamento de padrão. O dicionário construído contém 6.084 moléculas e 16.722 sinônimos (para as proteínas de levedura) e 4.405 termos (para a proteína *E.coli*). A média da precisão e da revocação alcançada para as proteínas é, respectivamente, de 94% e 84,6%.

Outro método baseado em dicionário é proposto por Tsuruoka e Tsujii (2004) que tem como objetivo reconhecer nomes de proteínas. Este trabalho pode ser dividido em duas fases: na primeira fase foram identificados os textos candidatos usando um dicionário; na segunda fase, os textos candidatos foram filtrados por meio do algoritmo de aprendizado de máquina Naïve Bayes, obtendo uma melhora da medida-F de 10,8% e aumentando a precisão com uma pequena perda de revocação. Para atenuar o problema de baixa revocação causado pelas variações da ortografia foram utilizadas duas técnicas: a primeira usa um algoritmo de procura de string por aproximação (i.e., distância de edição) ao invés de procura de string por casamento exato; a segunda expande o dicionário utilizando o UMLS (http://www.nlm.nih.gov/research/umls/) com a geração de variações de palavras. Com o uso dessas técnicas obteve-se uma melhora de 1,6%. A precisão e revocação obtidas foram, respectivamente, 71,7% e 62,3%.

ProtScan é um sistema desenvolvido por Egorov, Yuryev e Daraselia (2004) que utiliza uma abordagem baseada em dicionário para identificação de nomes de proteínas da classe mamífero em resumos do MEDLINE. São construídos dois dicionários: um deles serve para identificar os nomes nas sentenças; o outro serve para eliminar os falsos positivos e evitar desambiguação. Os dicionários foram gerados a partir do banco de dados LocusLink e de outras bases. Técnicas como algoritmo de tokenização e de abreviação são utilizadas. Obteve-se uma precisão de 98% e revocação de 88%.

Kou, Cohen e Murphy (2005) propõem um novo método de aprendizado denominado Dict-HMMs em que um dicionário é convertido para um modelo oculto de Markov (HMM) que reconhece frases do dicionário, assim como as variações destas frases. Dict-HMMs extrai somente nomes de proteínas que tem uma alta similaridade com os nomes armazenados no dicionário. O método proposto foi testado com três bancos de dados: a média da precisão e revocação com esses bancos foi, respectivamente, 50,1% e 68,8%. O Dict-HMMs obteve melhor revocação em comparação com alguns sistemas anteriores, os quais obtiveram uma melhor precisão. Assim, o diferencial deste sistema é a revocação. A vantagem é que o modelo pode ser treinado com uma pequena quantidade de dados. O etiquetador POS utilizado foi o de Brill. Essa pequena quantidade de dados é selecionada dos nomes de proteínas mais relevantes contidos no dicionário. O dicionário utilizado foi o PIR-NREF (http://pir.georgetown.edu/) que contém aproximadamente 500.000 nomes de proteínas.

Schuemie et al. (2007) avaliaram algumas técnicas para aumentar a revocação na identificação de nomes de genes e proteínas, utilizando a combinação de um dicionário construído a partir de informações armazenadas em vários bancos de dados com regras para

gerar variações de ortografia. Uma lista de regras é utilizada de trabalhos anteriores. Schuemie et al. (2007), além de utilizarem várias regras geradas por trabalhos anteriores, acrescentam algumas, por exemplo: se os termos contêm letras e números, então desconsidera-se a diferença de maiúscula e minúscula. Testes foram realizados com todas as regras (aumentando consideravelmente a revocação e diminuindo a precisão) e com algumas regras (diminuiu um pouco a revocação, mas aumentou a precisão). Apesar de várias regras serem utilizadas, muitas não influenciaram a revocação e algumas influenciaram negativamente a precisão. Em suma, a combinação dos bancos de dados aumentou significantemente a revocação em comparação com o uso de um único banco de dados. A precisão e a revocação foram avaliadas em três diferentes conjuntos de dados e com quatro tipos de organismos. Como o resultado em cada conjunto é discrepante, não foi possível obter uma média realista dessas medidas.

3.2.2 Abordagem Baseada em Regras

Regras permitem descrever, precisamente, os elementos de um conjunto, seja este finito ou infinito, sem a necessidade de enumerá-los explicitamente. A seguir alguns exemplos de padrões extraídos a partir de regras:

- O padrão a seguir identifica sentenças ou grupos de sentenças que contenham variações da palavra "*interact*", tendo no meio da sentença o nome de um gene "genexx" e que contenha variações da palavra "*bind*" (GHANEM et al., 2002).

$$interact\backslash s([a-z]*(\backslash s)+)*genexx[a-z]+\backslash s([a-z]*(\backslash s)+)*bind\backslash s$$

- Os dois padrões a seguir encontram relacionamento entre gene e doença. Para identificação desses relacionamentos pode-se utilizar análise linguística e semântica (COHEN, K.; HUNTER, 2008).

$$\langle gene\rangle\ plays\ a\ role\ in\ \langle disease\rangle$$
$$\langle disease\rangle\ is\ associated\ with\ \langle gene\rangle$$

Em seguida serão apresentados alguns trabalhos que utilizam regras para extrair informação no domínio biomédico.

Trabalhos que utilizam Regras para Extração de Informação

Serão apresentados três trabalhos de diferentes autores e anos que utilizam essencialmente regras para extrair informação de resumos do MEDLINE, com ou sem o auxílio de um dicionário. Na extração de informação a técnica de Processamento de Língua

Natural *Part-Of-Speech* (POS) pode ser utilizada. Os níveis de extração podem variar de nomes simples ou compostos, ou frase. A quantidade de resumos do MEDLINE utilizado para avaliar a extração de informação também é mostrada. As principais informações são resumidas na Tabela 5.

Tabela 5 – Trabalhos com regras.

Autor	Etiquetador POS	Níveis de Extração	Dicionário	Resumos do MEDLINE	Sistema	Domínio	Medida-F
Fukuda et al. (1998)	Não	Nome simples e composto	Não	30	KeX	Proteína	96,7%
Franzén et al. (2002)	Não	Nome simples e composto	Sim	200	YaPex	Proteína	KeX = 49,5% Yapes = 77,1%
Hu et al. (2005)	Sim	Frase	Não	300	RLIMS-P	Fosforilação de Proteína	92,7%

PROPER (**PRO**tein **P**roper-noun phrase **E**xtracting **R**ules), introduzido por Fukuda et al. (1998), é um método baseado em regras e um dos primeiros sistemas que extrai nomes de proteínas em publicações biológicas. As regras são geradas manualmente para extrair termos simples e compostos sem utilizar um dicionário. Obteve-se uma medida-F de 96,7% em 30 resumos do MEDLINE sobre a proteína SH3. As regras codificadas manualmente com base em observação do conjunto de dados contribuíram para o alto desempenho do método. A partir do método PROPER, foi desenvolvido o sistema de extração de informação KeX (http://www.hgc.jp/service/tooldoc/KeX/intro.html).

Franzén et al. (2002) desenvolveram o sistema YaPex para identificação automática de nomes de proteínas em 200 resumos do MEDLINE que utiliza regras desenvolvidas manualmente. O YaPex (http://www.sics.se/humle/projects/prothalt/) consiste de duas análises: léxica e sintática. Na primeira são selecionadas, por exemplo, as palavras com sufixos (e.g., *-ase* e *-in*) e que contenham letras maiúsculas ou números (e.g., HsMad2, U3-55k). Na segunda utiliza-se o analisador gramatical ENFDB para identificar nomes simples ou compostos. Os nomes de proteínas identificados são armazenados em um dicionário para ajudar na seleção de novos termos que não foram identificados pelo ENFDG. Para isso, utilizam-se as variações desses nomes para encontrar palavras similares no texto. O banco de dados SWISS-PROT auxilia na identificação dos termos principais. Algumas heurísticas de Fukuda et al. (1998) são utilizadas na análise léxica. Expressões regulares são aplicadas para reduzir a baixa precisão, por exemplo: padrões de sufixos de palavras (nomes de substâncias

químicas) ou palavras e expressões de fórmulas químicas, expressões aritméticas e sequências de aminoácidos.

Comparando-se os sistemas YaPex e KeX, o YaPex identificou mais nomes de proteínas do que o KeX. Franzén et al. (2002) avaliaram que o analisador sintático ENFDG contribuiu nessa identificação, selecionando adequadamente nomes simples e compostos. Seis diferentes análises foram realizadas para avaliar os sistemas. A maior diferença de medida-F entre os dois sistemas foi no limite *right* (i.e., nome que encontra-se do lado direito de uma sentença), respectivamente, de 77,1% e 49,5%.

Outro sistema baseado em regras é o RLIMS-P (***R**ule-based **LI**terature **M**ining **S**ystem for Protein **P**hosphorylation*), cujo objetivo é extrair informação de fosforilação de proteína de resumos do MEDLINE (HU et al., 2005). Foi desenvolvido com base no algoritmo de Ravikumar (2004 apud HU et al., 2005). Padrões foram criados depois de examinar diferentes formas usadas para descrever interações de fosforilação em 300 resumos do MEDLINE e 10 artigos.

Dois tipos de tarefas foram implementadas no sistema RLIMS-P: *citation mapping* e *evidence tagging*. A primeira tem a função de recuperar informação de artigos do MEDLINE relacionados à fosforilação, para a qual obteve precisão e revocação, respectivamente, de 91,4% e 96,4%. A segunda tem o objetivo de extrair informação sobre fosforilação dos artigos anotados, para a qual obteve precisão e revocação, respectivamente, de 97,9% e 88,0%. medida-F é de 92,7%.

O sistema RLIMS-P utiliza *shallow parsing* e extrai informação do texto utilizando casamento de padrões desenvolvidos manualmente. No pré-processamento, o texto é dividido em sentenças e tokenizado por palavras e pontuação. Cada palavra é associada às etiquetas POS, como advérbio, verbos, adjetivos, etc. Utiliza reconhecimento de entidade nomeada para detectar acrônimo e termo (NARAYANASWAMY; RAVIKUMAR; VIJAY-SHANKER, 2003 apud HU et al., 2005).

As sentenças podem ser casadas com um simples padrão "⟨AGENT⟩ *phosphorylate* ⟨THEME⟩ *at* ⟨SITE⟩", onde ⟨AGENT⟩ representa uma enzima (e.g., quinase catalisadora de fosforilação), ⟨THEME⟩ significa um substrato (i.e., proteína sendo fosforilada) e ⟨SITE⟩ indica um P-Site (i.e., resíduo de aminoácido sendo fosforilado). Este passo é para detectar sentenças com estrutura sintática de acordo com o padrão estabelecido, por exemplo: "*Active **p90Rsk2** was found to be able to phosphorylate histone **H3** at **Ser10***". São usados alguns padrões de etiquetas POS para identificar grupos de verbos e frases com substantivos.

Classificação semântica é utilizada no sistema RLIMS-P para melhorar a precisão da extração de frases com substantivo. A classificação utiliza sufixos, frases e palavras informativas, por exemplo, "*mitogen activated protein kinase*" é classificada como uma proteína por causa da palavra-chave "*kinase*". Outras regras e heurísticas são desenvolvidas com base na detecção de apositivo, conjunção e pares (i.e., sentença e acrônimo). Um exemplo de um par sentença/acrônimo é "*mitogen activated protein kinase*" e "MAPK". Os detalhes são encontrados em Narayanaswamy, Ravikumar e Vijay-Shanker (2003 apud HU et al., 2005).

Por fim, padrões baseado em regras são identificados na forma verbal (i.e., padrões com diferentes formas, como: "*phosphorylate/phosphorylated/phosphorylating/ phosphorylates*") e nominal (i.e., seleciona a palavra mais frequentemente encontrada: "*phosphorylation*").

3.2.3 Abordagem Baseada em Aprendizado de Máquina

Sistemas de aprendizado de máquina (AM) são normalmente projetados para um conjunto de classes específicas. São utilizados dados de treinamento para aprender as características úteis e relevantes para o reconhecimento e a classificação de termos. No que diz respeito à extração de termo utilizando a abordagem baseada em AM, uma sequência de palavras é considerada como um termo de uma determinada classe, se a mesma preenche os critérios de acordo com as características aprendidas a partir de um conjunto de termos predefinidos. Segundo Ananiadou e McNaught (2006), o principal desafio do AM é selecionar um conjunto de características representativas que podem ser utilizadas para o reconhecimento e a classificação precisa de novos termos. Outro desafio é detectar qual o limite de um termo composto por várias palavras.

Em seguida são apresentados alguns trabalhos que utilizam da abordagem de aprendizado de máquina para extrair informações biomédicas. Na Seção 2.1.2 encontra-se uma introdução sobre aprendizado de máquina. Vale destacar que a metodologia proposta de extração de informação nesta dissertação usa aprendizado de máquina para classificar as sentenças de interesse sobre as quais posteriormente serão utilizadas as abordagens de dicionário e de regra para o reconhecimento e a extração propriamente dita dos termos. Por outro lado, os trabalhos descritos a seguir usam o AM com o objetivo de extrair diretamente os termos de um documento.

Trabalhos que utilizam Aprendizado de Máquina para Extração de Informação

Na literatura biomédica encontra-se vários trabalhos que extraem informação utilizando a abordagem de aprendizado de máquina. Não tem-se aqui a intenção de aprofundar em cada um desses trabalhos. No próximo parágrafo são apresentados três trabalhos que identificam nomes de proteínas utilizando aprendizado de máquina com os valores de medida-F obtidos. Em seguida é resumido alguns trabalhos atuais que utilizam diferentes classificadores.

Nobata, Collier e Tsujii (1999) utilizaram os algoritmos de árvore decisão e de classificação Bayesiana para identificar frases que contêm nomes de proteínas, com base na composição de palavras (medida-F obtida foi de 70% a 80%). Collier, Nobata e Tsujii (2000) utilizaram o algoritmo HMM para treinar e detectar nomes de proteína no texto (73% de medida-F). Kazama et al. (2002) utilizaram o classificador SVM para identificar nomes biomédicos (e.g., proteínas, DNA e lipídeos) e obteve uma variação de medida-F de 54,4 a 73,6%.

Vários algoritmos de aprendizado de máquina são utilizados para reconhecimento de genes ou proteínas no domínio biomédico: Naïve Bayes (TSURUOKA; TSUJII, 2004); *Conditional Random Field* (MCDONALD; PEREIRA, 2005) e combinação de três classificadores (SVM e duas variações do HMM) (ZHOU, G. et al., 2005).

3.3 Considerações Finais

É crucial a identificação de termos para o processamento automático na literatura biomédica. O desempenho de métodos de Reconhecimento Automático de Termo (ATR) em domínio biomédico varia em torno de 70% a 90% de precisão e aproximadamente 70% de revocação (KRAUTHAMMER; NENADIC, 2004; ANANIADOU; MCNAUGHT, 2006). Para identificar esses termos são utilizadas três abordagens: abordagem baseada em dicionário, baseada em regras e baseada em aprendizado de máquina.

A **abordagem baseada em dicionário** tem a vantagem de armazenar informações relacionadas a um determinado domínio e possibilitar a identificação de termos como genes, proteínas e doenças no domínio biomédico. Um problema é a limitação de nomes que estão presentes no dicionário. Segundo Kou, William Cohen e Murphy (2005), extratores baseado em dicionário ao extrair nomes de proteínas geralmente têm uma baixa revocação, exceto se lidar com as variações de nome. Uma das maneiras de lidar com essas variações é utilizar técnicas como aproximação de string (e.g., distância de edição). Tsuruoka e Tsujii (2004)

ratificam o problema da variação, alertando para outro: nomes curtos armazenados no dicionário geram falsos positivos, diminuindo a precisão.

A abordagem baseada em regras tem algumas desvantagens: prolonga significativamente a construção de sistemas, reduz a capacidade de adaptação de regras em outro sistema e exclui termos que não correspondem aos padrões predefinidos. Tem em geral um desempenho melhor do que outras abordagens, no entanto há o problema de adaptação para novos domínios e classes (ANANIADOU; MCNAUGHT, 2006). A abordagem baseada em regras é mais adequada quando necessita-se de um sistema com alta precisão, a qual não é alcançada com a abordagem baseada em aprendizado de máquina devido à insuficiência de dados de treinamento (AGATONOVIC et al., 2008).

As vantagens de se utilizar a **abordagem baseada em aprendizado de máquina** são a independência de domínio e alta qualidade na predição. Os principais problemas relacionados aos algoritmos de AM são a necessidade de grandes quantidades de dados de treinamento e os dados precisam ser periodicamente retreinados após o advento de novos dados. Em geral, a classificação é prejudicada quando o conjunto de dados de uma classe é pequeno (classe minoritária) em relação a outras classes (ANANIADOU; MCNAUGHT, 2006). A qualidade na predição de sistemas de aprendizado de máquina depende da existência suficiente de dados de treinamento (TANABE; WILBUR, 2002a). Segundo Manning, Raghavan e Schütze (2008), não existe um algoritmo ótimo que resolva todos os problemas. Se o problema de classificação consistir de um número de categorias bem separadas, muitos algoritmos de classificação provavelmente trabalharão bem. Porém, os autores fazem uma ressalva: a maioria dos problemas contém uma grande quantidade de categorias muito similares.

Cada abordagem tem suas vantagens e desvantagens. Para desfrutar das características positivas dessas abordagens surge a necessidade de utilizar o que cada uma delas fornece de melhor. No próximo capítulo serão apresentados os trabalhos correlatos que extraem informação de artigos científicos e que utilizam a combinação das abordagens discutidas neste capítulo.

4 TRABALHOS CORRELATOS

No Capítulo 3 foi apresentada a discussão de vários trabalhos que extraem informação de resumos no domínio biomédico e que utilizam especificamente uma destas três abordagens, a saber: baseada em dicionário, em regras ou em aprendizado de máquina. Este capítulo tem o intuito de apresentar os trabalhos que também extraem informação nesse domínio e que utilizam a combinação destas três abordagens em resumos ou artigos completos.

Na Tabela 6 e na Tabela 7 são apresentados alguns trabalhos (ordenados por ano) encontrados na literatura que extraem informação de resumos e de artigos completos e que utilizam a combinação das três abordagens anteriormente mencionadas. Além de destacar qual a abordagem que está sendo utilizada, também são apresentadas algumas informações como o domínio da extração (e.g., gene e proteína), o sistema desenvolvido a partir das técnicas utilizadas para a extração de informação (se houver) e a utilização ou não de algum etiquetador. A seguinte nomenclatura foi utilizada em ambas as tabelas: D significa Dicionário; R significa Regras; AM significa Aprendizado de Máquina; e POS significa etiquetador *Part-Of-Speech*.

Tabela 6 – Trabalhos correlatos que extraem informação de resumos.

Autores	**Abordagem**			**Informação**			
	D	R	AM	Domínio	Sistema	Resumos do MEDLINE	Etiquetador POS
Leonard et al. (2002)	X	X	X	Gene e Proteína	-----	Sim	Não
Seki e Mostafa (2003)	X	X		Proteína	-----	Sim	Não
Mika e Rost (2004a, b)	X	X	X	Proteína	NLProt	Sim	Sim
GuoDong Zhou et al. (2004)		X	X	Proteína	PowerBioNE	Sim	Sim
Seki e Mostafa (2005)	X	X	X	Proteína	Protex	Sim	Não
Hanisch et al. (2005)	X	X		Gene e Proteína	ProMiner	Não	Não
Chun et al. (2006)	X		X	Gene e Doença	-----	Sim	Sim

Todos os trabalhos da Tabela 6 extraem informação de resumos do MEDLINE, com exceção de Hanisch et al. (2005) que utilizam o *benchmark* BioCreAtIvE. Alguns optaram em utilizar um etiquetador POS (MIKA; ROST, 2004a, b); já outros optaram em não utilizar etiquetador devido ao custo computacional (LEONARD; COLOMBE; LEVY, 2002; SEKI; MOSTAFA, 2003). As medidas de precisão e revocação, utilizadas para avaliar os resultados obtidos com a extração de informação, não foram apresentadas nesta tabela, pois na maioria dos trabalhos os valores obtidos dependem de alguns parâmetros inerentes de cada trabalho.

A seguir são apresentadas algumas informações dos trabalhos da Tabela 6 agrupados por abordagens utilizadas:

- Combinação de dicionário, aprendizado de máquina e regras: Leonard, Colombe e Levy (2002) utilizaram abordagem baseada em dicionário e regras para extrair nomes de genes e proteínas dos resumos do MEDLINE e em seguida, um classificador Bayesiano baseado na frequência das palavras é utilizado para pontuar os nomes relevantes; Mika e Rost (2004a, b) desenvolveram um sistema denominado NLProt (http://cubic.bioc.columbia.edu/services/NLProt/) que combina o algoritmo de aprendizado de máquina SVM com filtros baseados em regras e dicionário, a fim de identificar nomes e sequências de proteínas em resumos do PubMed; Seki e Mostafa (2005) utilizaram a combinação das três abordagens e não utiliza análise sintática nem etiquetador POS;
- Combinação de aprendizado de máquina e regras: GuoDong Zhou et al. (2004) desenvolveram o sistema PowerBioNE utilizando a abordagem baseada em aprendizado de máquina com os algoritmos HMM e k-vizinhos mais próximos e como pós-processamento utilizou-se o padrão para extrair regras automaticamente dos dados de treinamento;
- Combinação de regras e dicionário: Seki e Mostafa (2003) extraíram nomes de proteína usando regras e dicionário; Hanisch et al. (2005) desenvolveram o sistema ProMiner que utiliza as palavras geradas a partir da consulta de dicionário e extrai regras para reconhecer nomes compostos de gene e proteína;
- Combinação de dicionário e aprendizado de máquina: Chun et al. (2006) extraíram relações de gene e doença utilizando um dicionário construído a partir de seis bancos de dados. O algoritmo de aprendizado de máquina, Entropia Máxima, é utilizado para filtrar os falsos positivos gerados pelo dicionário.

Na Tabela 7 são resumidas as características de alguns trabalhos encontrados na literatura que extraem informação no domínio biomédico de artigos completos, os quais serão explicados nas seções subsequentes.

Tabela 7 – Trabalhos correlatos que extraem informação de artigos completos.

Autor	**Abordagem**			**Informação**				
	D	R	AM	Domínio	Sistema	Objetivo	POS	Avaliação[2]
Tanabe e Wilbur (2002a, b)	x	x	x	Gene e Proteína	ABGene	Extrair informação	Sim	Resumos Prec. 85,7% Rev. 66,7% Artigos Prec. 72,5% Rev. 50,7%
Corney et al. (2004)	x	x		Gene e Proteína	BioRAT	Povoar um banco de dados	Sim	Resumos Prec. 55,1% Rev. 20,3% Artigos Prec. 51,2% Rev. 43,6%
Bremer et al. (2004)	x	x		Gene e Proteína	--------	Povoar um banco de dados	Não	Prec. 63,5% Rev. 37,3%
Garten e Altman (2009)	x[1]	x[1]		Genes (G), Drogas (D) e Polimorfismos (P)	Pharmspresso	Destacar as sentenças de acordo com a consulta do usuário	Não	Revocação 78,1% (G) 74,4% (D) 60,8% (P) 50,3% (G e D)

[1] Ontologia e expressões regulares, respectivamente, do sistema Textpresso.
[2] Prec. significa Precisão e Rev. significa Revocação.

4.1 ABGene

O ABGene é um sistema treinado em resumos do MEDLINE e testado em um conjunto de artigos completos do domínio biomédico selecionados aleatoriamente para identificar nome de gene e proteína. Um etiquetador POS baseado em transformação é treinado em sentenças de resumos com ocorrência de gene destacada manualmente para induzir regras. Em seguida, regras e dicionário foram aplicados como pós-processamento.

Tanabe e Wilbur (2002b) realizaram duas adaptações no sistema ABGene (TANABE; WILBUR, 2002a) para extrair informação de artigos completos. Na primeira adaptação utilizou-se um classificador para atuar na classificação em nível de sentença de artigos completos. Definiu-se que sentenças abaixo de um limiar não contêm nomes de gene/proteína. Na segunda é realizado um pós-processamento a fim de extrair supostos grupos de nomes de

gene/proteína. Em 2,16 milhões de resumos do MEDLINE foram encontrados 2,42 milhões de nomes de gene e proteína. Separou-se em três grupos com limiar igual a: 10 (134.809 nomes), 100 (13.865 nomes) e 1.000 (1.136 nomes).

O treinamento foi feito com um conjunto de 1.000 artigos selecionados aleatoriamente do PubMed Central, totalizando 7.000 sentenças que foram selecionadas manualmente nos artigos. O teste foi realizado com um conjunto de 2.600 sentenças, a fim de avaliar como a heterogeneidade de artigos completos afeta o desempenho do ABGene. A média da precisão e revocação obtidas foram, respectivamente, 72,5% e 50,7% aquém da obtida em resumos (Tabela 7).

Tanabe e Wilbur (2002b) relataram alguns problemas na extração em artigos completos: falsos positivos como nomes de reagentes químicos são mais raros em resumos; vários falsos negativos encontram-se em tabelas e figuras. As principais técnicas utilizadas por Tanabe e Wilbur (2002a) são resumidas na Tabela 8.

Tabela 8 – Abordagem híbrida proposta por Tanabe e Wilbur (2002a).

PLN	Regras	Aprendizado de Máquina	Dicionário
Etiquetador POS de Brill (1994)	Expressão Regular	Aprendizado Bayesiano	Lista e banco de dados

O etiquetador POS utilizado gera automaticamente regras com palavras simples de nomes de gene e proteína. Em seguida, regras são desenvolvidas para extrair nomes compostos que são prevalentes na literatura. Algumas técnicas são utilizadas para filtrar os falsos positivos e falsos negativos, a saber:

Falsos positivos: dicionário e regras são utilizados para remover os falsos positivos. O dicionário contém 1.505 termos biológicos (ácidos, antígeno, etc.), 39 nomes de aminoácido, 233 enzimas, 593 células, 63.698 nomes de organismo do banco de dados do NCBI ou 4.357 termos não biológicos. Expressões regulares foram elaboradas para excluir drogas com sufixos comuns (e.g., *-ole*, *-ane*, *-ate*, etc.) e número seguido de medida (e.g., $25mg/ml$).

Falsos negativos: dicionário, aprendizado de máquina e regras são utilizados para recuperar os falsos negativos. O dicionário de 34.555 nomes simples e 7.611 nomes compostos é construído a partir do banco de dados LocusLink e do *Gene Ontology*. Os nomes com uma baixa frequência de trigramas ou uma palavra do contexto antes ou depois do nome também são selecionados. A palavra de contexto é gerada automaticamente por um algoritmo de probabilidade (peso Bayesiano ou *log odds score*) que indica a probabilidade de nomes de genes adjacentes aparecerem no texto. Expressões regulares adicionais são criadas para

permitir casamento de padrão de palavras com números e letras, e prefixos e sufixos comuns (e.g., *-gene, -like, -ase, homeo-*).

Também se utiliza o aprendizado Bayesiano para encontrar a probabilidade de um documento conter nome de gene/proteína, podendo, assim, não extrair informação de documentos que não contêm nomes relacionados. Para isso, documentos que contêm nomes de gene/proteína são treinados. Na classificação de novos documentos, documentos com valores de similaridade abaixo de um limiar são descartados.

A extração de informação em resumos obteve uma precisão de 85,7% e uma revocação de 66,7% usando a combinação da estratégia baseada em conhecimento (dicionário, regra e PLN) e estatística (aprendizado de máquina). Segundo Cohen e Hersh (2005), o ABGene é uma das abordagens baseada em regras mais bem-sucedida para reconhecimento de gene e proteína em textos biomédicos.

4.2 BioRAT

BioRAT (***Bio****logical* ***R****esearch* ***A****ssistant for* ***T****ext mining*, http://bioinf.cs.ucl.ac.uk/biorat/) é um sistema capaz de recuperar e analisar informação de resumos e artigos completos do domínio biomédico (CORNEY et al., 2004). Pesquisa por artigos (resumo e artigo completo) disponível no banco de dados PubMed a partir da consulta de entrada do usuário. Os artigos identificados na página no formato PDF são baixados e convertidos para o formato textual (não é informado como é realizado a conversão de PDF para o formato textual). Após recuperar os documentos relevantes, o sistema extrai fatos interessantes. Esses fatos podem ser utilizados para povoar o banco de dados automaticamente.

A extração de informação é baseada no conjunto de ferramentas desenvolvida pela Universidade de Sheffield denominada GATE (***G****eneral* ***A****rchitecture for* ***T****ext* ***E****ngineering*). GATE é utilizado para rotular as palavras (POS) para em seguida serem aplicados filtros para excluir verbos que não são proteínas. Dois componentes do GATE são utilizados: *gazetteers* e *templates*. O primeiro é utilizado para identificar palavras ou frases relacionadas a genes e proteínas. O segundo permite extrair informação automaticamente a partir de padrões textuais. Um exemplo de um simples *template* do sistema BioRAT é:

$$interaction\ of\ (PROTEIN_1)\ AND\ (PROTEIN_2),$$

onde "$PROTEIN_1$" e "$PROTEIN_2$" são *slots* para serem preenchidos com nomes de proteína, definido por um *gazetteer*. Exemplo de uma sentença que é identificada pelo *template* é:

"*Genetic evidence for the **interaction of Pex7p and Pex13p** is provided...*". Cada *template* é criado manualmente com o auxílio da interface gráfica do BioRAT.

As principais técnicas utilizadas por Corney et al. (2004) são resumidas na Tabela 9.

Tabela 9 – Abordagem de extração de informação proposta por Corney et al. (2004).

Recuperação de Informação	Extração de Informação	Dicionário	Regras
Resumos e artigos completos do PubMed	Fatos (gene e proteína)	*Gazetteers*	*Template*

BioRAT é comparado com o sistema de extração de informação SUISEKI (BLASCHKE; VALENCIA 2002 apud CORNEY et al., 2004). O sistema SUISEKI utiliza conhecimento estatístico como a frequência de palavras que ocorrem em uma frase. Os *frames* de SUISEKI, similares aos *templates* do BioRAT, contêm padrões relacionados a substantivos e verbos, mas não reconhecem conjunção, adjetivos ou outra classe de palavra.

Para avaliar o BioRAT foi utilizado o DIP (*Database of Interacting Proteins*) com 389 registros que contém 229 resumos do PubMed. O DIP é um banco de dados que contém interações entre proteínas, as quais serviram como base para comparar os resultados obtidos do SUISEKI com o BioRAT.

O sistema BioRAT utilizou um total de 19 *templates* derivados dos *frames* de SUISEKI e 127 *gazetteers* derivados do MeSH e outras fontes. A revocação alcançada por ambos sistemas em resumos é aproximadamente a mesma (BioRAT = 20,31% e SUISEKI = 22,33%). A taxa de revocação do BioRAT em artigo completo foi de 43,6%, sendo 25,6% do corpo do artigo e 18% do resumo. No entanto, a precisão do resumo foi maior do que no artigo completo, respectivamente, 55,07% e 51,25% (Tabela 7). Isto aconteceu devido as imperfeições no conjunto de *templates* usado pelo BioRAT. É destacado por Corney at al. (2004) que a solução para diminuir o erro de precisão é utilizar um esforço manual para aumentar as restrições dos *templates*.

4.3 Bremer et al. (2004)

Bremer et al. (2004) desenvolveram um sistema integrado que combina dicionários (i.e., de sinônimos, gene e proteína) com regras para extrair e organizar as relações genéticas de artigos completos. As relações extraídas são armazenadas em um banco de dados que inclui o código único do artigo (código do PubMed) e de quatros seções (resumo, introdução, materiais e método, resultados e discussão) para identificar o artigo selecionado e a seção das quais as informações foram extraídos.

Dois dicionários são criados com informação de nomes de gene e proteína (282.882), e sinônimos (274.845 sinônimos e 124 verbos de relação) para identificar sentenças que contém nomes de gene/proteína. O dicionário de gene e proteína foi construído a partir de vários banco de dados existentes como o LocusLink, o SWISS-PROT, dentre outros (alguns desses bancos foram apresentados na Seção 3.2.1). O dicionário de sinônimo contém variações de sinônimos (e.g., *inhibit* → *inhibits, inhibition, inhibited*), informações contextuais como prefixos e sufixos (e.g., *kinase, phosphate, receptor*) e verbos de interação que foram criados a partir da análise de 1.000 artigos por um processo semiautomático.

Os nomes armazenados no dicionário ajudaram a identificar sentenças que contêm um ou mais nomes de gene/proteína. A partir das sentenças identificadas, um conjunto de padrão de regras foi elaborado para extrair genes. As regras foram baseadas na combinação de nomes de gene/proteína, preposições e palavras-chave que indicam o tipo de relacionamento entre genes. Também foram criados padrões usando substantivos e verbos na forma passiva e ativa.

A extração de informação é dividida em quatro passos:

1. Tokenizar o texto em sentenças;
2. Analisar sentenças para identificar frases com substantivo e verbo;
3. Selecionar sentenças que contêm genes usando dicionários de nome de gene e proteína, e sinônimo;
4. Extrair gene utilizando regras de casamento de padrão.

A ferramenta de processamento textual LexiQuestMine da empresa SPSS (http://www.spss.com) foi utilizada para construir os dicionários de nomes de gene e proteína, sinônimos e padrões associados com genes.

Scripts foram desenvolvidos com o auxílio do software GetItRight (disponível comercialmente em http://www.cthtech.com/) para conectar e baixar artigos completos automaticamente no formato HTML. Um pré-processamento é realizado para converter o arquivo HTML para o formato XML. Para isso, por exemplo, removeu as etiquetas HTML, substituiu símbolos gregos (e.g., α → alfa) e eliminou as referências do artigo. No documento XML (Figura 7) foram incluídas etiquetas para cada seção, além de informações sobre o título e código do artigo. As figuras do artigo não foram incluídas no banco de dados, a fim de economizar espaço de armazenamento.

<?xml version='1.0'?><Doc>
<MedlineID>12514136</MedlineID>
<Title>
Determinants in mammalian telomerase RNA that mediate enzyme processivity and cross-species incompatibility
</Title>
<Abstract>
Abstract of document here ….
</Abstract>
<Introduction>
Introduction of document here ….
</Introduction>
<Methods>
Materials and methods section of document here …
</Methods>
<Results>
Results and discussion section of document here …..
</Results></Doc>

Figura 7 – Exemplo de um documento XML com etiquetas de quatro seções.
Fonte: Bremer et al. (2004).

Foram selecionados artigos no domínio da biologia molecular e da biomedicina, mais especificamente sobre tumores celebrais, de 20 revistas entre 1999 e 2003. Para avaliar o sistema, selecionou-se aleatoriamente 100 artigos, sendo cinco de cada revista e um de cada ano. Dez neurobiólogos analisaram manualmente esses 100 artigos e identificaram 141 nomes de gene. A precisão e revocação alcançadas foram, respectivamente, 63,5% e 37,3% (Tabela 7). A baixa precisão foi devido aos erros de padrão na identificação de nomes de gene/proteína em algumas sentenças e na falta de padrões com palavras compostas para explorar sentenças complexas. A baixa revocação foi devido à diversidade de 20 artigos diferentes.

Continuação do Trabalho de Bremer et al. (2004)

A partir da extração de informação de artigos científicos desenvolvida por Bremer et al. (2004), Natarajan e Berrar et al. (2006) implementaram um processo de mineração de textos como mostrado na Figura 8: artigos são baixados no formato HTML sem imagem e convertidos para o formato XML, utilizando a ferramenta GetItFull (NATARAJAN; HAINES et al., 2006); termos são extraídos do LexiQuestMine utilizando padrões; em seguida, no módulo Curador, os termos são padronizados utilizando um dicionário de sinônimos para serem, enfim, armazenados em um *data warehouse*. Os dados são utilizados posteriormente em uma rede de interação para visualizar as interações de gene e proteína.

A partir do armazenamento dos dados, Natarajan e Berrar et al. (2006) identificaram um relacionamento interessante entre o 1-fosfato de esfingosina e a invasividade de um tumor e notaram que a rede de interação desenvolvida tem potencial para melhorar o entendimento do papel desempenhado por tumores invasivos. Natarajan e Berrar et al. (2006) concluíram

que a extração automática de informações a partir de literatura biológica promete desempenhar um papel cada vez mais importante na descoberta de conhecimento biológico.

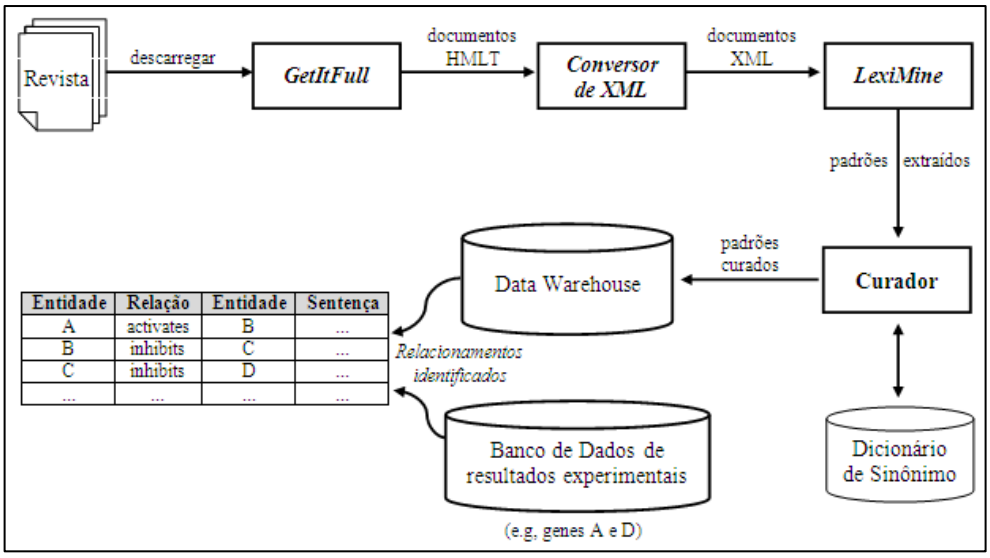

Figura 8 – Processo de extração de padrão e *data warehouse*.
Fonte: Adaptado de Natarajan e Berrar et al. (2006).

4.4 Pharmspresso

O sistema Pharmspresso (http://pharmspresso.stanford.edu) extrai informação sobre genes, drogas e polimorfismos de artigos completos da literatura pertinente à área da farmacogenômica, a partir da consulta determinada pelo usuário. Portanto, é um sistema de recuperação de informação que utiliza da extração de informação para recuperar as informações de acordo com a necessidade do usuário (GARTEN; ALTMAN, 2009).

Pharmspresso tem o objetivo de processar artigos completos no formato PDF, utilizando expressões regulares e indexar o conteúdo com base em uma ontologia de conceitos. O Pharmspresso é baseado no sistema Textpresso desenvolvido por Müller, Kenny e Sternberg (2004 apud GARTEN; ALTMAN, 2009).

Na Figura 9 é mostrado o processo de recuperação e extração de informação realizado pelo sistema. Artigos PDF são baixados, convertidos em formato textual e tokenizado em palavras e sentenças individuais. Em seguida, o texto é analisado para identificar palavras ou frases que são membros de categorias específicas de uma ontologia. Essas palavras ou frases identificadas são marcadas e indexadas para serem utilizadas em pesquisas futuras realizadas por palavras-chave definidas pelo usuário.

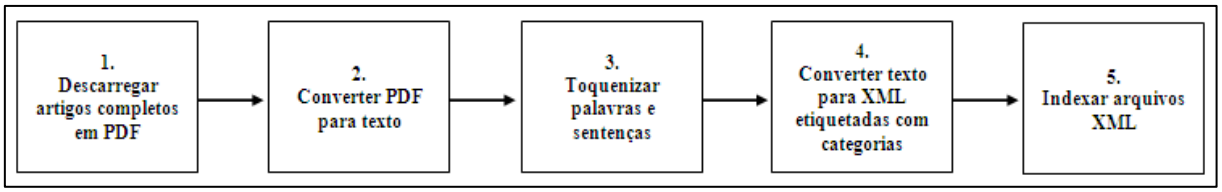

Figura 9 – Processo para recuperar e extrair informação do Pharmspresso.
Fonte: Adaptado de Garten e Altman (2009).

A ferramenta *open source* xpdf (http://www.foolabs.com/xpdf) foi utilizada para converter arquivos PDF para texto. *Scripts* em Perl foram adaptados do sistema Textpresso para tokenizar as sentenças e as palavras. A linguagem de programação Perl também foi utilizada para colocar as etiquetas no formato XML.

A avaliação do sistema Pharmspresso foi realizada em 45 artigos por 11 avaliadores (cientistas familiarizados com a literatura farmacogenética), os quais encontraram 178 genes, 191 drogas e 204 polimorfismos. O sistema encontrou, respectivamente, 78,1% (139), 74,4% (142) e 60,8% (124). Caso a consulta seja encontrar a relação de gene e droga, a percentagem é somente de 50,3% (Tabela 7). Os valores dessas medidas correspondem a revocação obtida.

Problemas com variações de nomes de gene foram encontrados, causando falsos positivos. Algumas das limitações do sistema são: o limite de 1.025 artigos completos predefinidos de 343 revistas diferentes; não há um mecanismo para pontuar as associações mais frequentemente mencionadas; e a impossibilidade de extrair informação de uma tabela convertida no formato de imagem.

4.5 Considerações Finais

A maioria dos sistemas apresentados neste capítulo extrai informação sobre gene ou proteína, utilizando a combinação das abordagens utilizadas no domínio biomédico (Tabela 7): dicionário, regras e aprendizado de máquina. Também existem trabalhos que optam por utilizar um etiquetador Part-Of-Speech. Com relação ao resultado gerado a partir da extração de informação, cada trabalho tem objetivos diferentes que podem ser sintetizados em: recuperar informação destacando as sentenças de acordo com a consulta definida pelo usuário e extrair informação para dar suporte à análise dos dados. As informações extraídas geralmente são armazenadas em um banco de dados para posterior identificação de padrões e relacionamentos interessantes.

No próximo capítulo será apresentada a metodologia proposta neste trabalho que utilizará das abordagens comumente desenvolvidas na literatura biomédica (i.e., aprendizado de máquina, regras e dicionário) para extrair informação.

5 METODOLOGIA PROPOSTA PARA EXTRAÇÃO DE INFORMAÇÃO NO DOMÍNIO BIOMÉDICO

Este capítulo objetiva apresentar a metodologia proposta de pré-processamento de informações não estruturadas, visando extrair informações relevantes sobre efeitos de doenças em artigos científicos do domínio biomédico. Nesta dissertação, metodologia é definida e usada como sendo um conjunto de etapas que são aplicadas para se atingir um determinador comum que é alcançar um resultado final desejável. A sequência das etapas é importante e deve ser respeitada. Em cada uma das etapas são utilizadas técnicas que permitem alcançar o resultado parcial da metodologia. A metodologia é composta por quatro etapas (Figura 10): Entrada de Dados (Etapa 1); Classificação de Sentenças (Etapa 2); Identificação de Termos Relevantes (Etapa 3); e Gerenciamento de Termos (Etapa 4). A partir dos documentos textuais fornecidos na Etapa 1, as sentenças são classificadas em suas respectivas classes (Etapa 2). Em seguida, os termos relevantes são identificados e extraídos das sentenças de interesse (Etapa 3) e após a validação dos termos pelo especialista, os termos são armazenados em um banco de dados (Etapa 4).

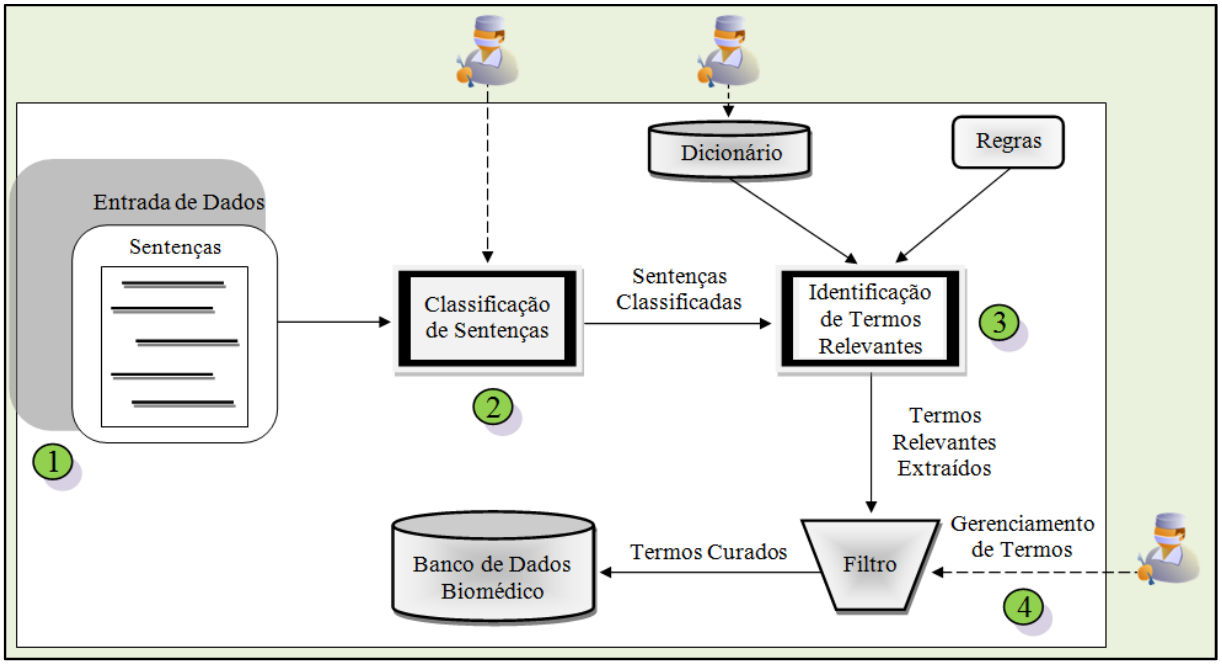

Figura 10 – Metodologia de pré-processamento para extração de informação.

Na etapa de **Entrada de Dados**, o conjunto de documentos é fornecido por especialistas do domínio. Estes documentos são formados por artigos científicos completos que inicialmente estão no formato PDF. Para facilitar a extração de informação destes artigos, é necessário converter os artigos do formato PDF para o formato de texto TXT ou para o formato XML, equivalente ao formato TXT mas acrescido de marcadores de localização.

Na etapa de **Classificação de Sentenças**, as sentenças dos artigos são classificadas utilizando técnicas de aprendizado de máquina supervisionado. O especialista pode auxiliar na construção do modelo de classificação, distinguindo as sentenças de interesse das sentenças que são irrelevantes.

Em seguida, na etapa de **Identificação de Termos Relevantes,** cada sentença classificada em uma classe de interesse é analisada utilizando expressões regulares (i.e., regras) com o intuito de identificar os termos relevantes e separá-los dentro de uma sentença. Todos os termos relevantes identificados são posteriormente extraídos e inseridos em um dicionário do domínio biomédico como um termo não curado. Este dicionário de terminologia contém termos que foram validados pelo especialista (i.e., termo curado), além de novos termos extraídos por regras (i.e., termo não curado). O dicionário possui duas funcionalidades: a primeira é ser o local no qual os termos extraídos dos artigos são armazenados como resultado do processo de extração de informação; a segunda é identificar sentenças que possuem termos baseado nos termos curados já presentes no dicionário.

Na etapa de **Gerenciamento de Termos,** o especialista auxilia na validação dos termos extraídos por meio de um filtro, isto é, ele exclui os termos irrelevantes ou transforma um termo não curado em um termo curado. Em seguida, os termos curados são inseridos no banco de dados biomédico. O dicionário e o banco de dados biomédico podem ser o mesmo repositório de dados. Na Figura 10, eles são representados separadamente para ilustrar suas funcionalidades de armazenamento dos termos curados, para auxiliar no processo de extração (i.e., dicionário) e como local no qual os termos curados identificados nas sentenças e novos temos não curados extraídos são armazenados (i.e., banco de dados biomédico). No decorrer do texto, ambos serão tratados como um mesmo repositório de dados, entretanto com estas funcionalidades distintas. O fato de ser utilizado o mesmo repositório faz com que termos extraídos (i.e., não curados) sejam facilmente transformados em termos curados, por meio da alteração do valor de um atributo do banco de dados, sem que seja necessário um processo de cópia de dados entre diferentes repositórios de dados (i.e., repositório de termos não curados para repositório de termos curados). Os termos validados podem ser utilizados em um novo ciclo de extração de informação, a fim de identificar termos em novas sentenças. Este recurso

é importante, pois uma regra pode identificar um termo em somente uma sentença X, mas após a validação deste termo pelo especialista, o dicionário pode identificar este mesmo termo em outras N sentenças.

A seguir é descrita cada uma das etapas da metodologia proposta.

5.1 Etapa 1 - Entrada de Dados

A entrada dos dados consiste de documentos textuais que estão inicialmente no formato PDF. Contudo, a metodologia proposta permite o uso de outros três formatos que são equivalentes ao formato inicial: HTML, XML e TXT. O documento no formato HTML é útil para visualizar as informações relevantes extraídas. Neste formato também é possível realçar as informações relevantes, a fim de apresentar em uma forma amigável para o especialista qual o termo importante naquele documento. Ademais, os documentos nos formatos HTML e PDF servem para o especialista ler o artigo inteiro. Os dois últimos formatos, XML e TXT, podem ser utilizados nas etapas de Classificação de Sentenças e Identificação de Termos Relevantes, as quais são detalhadas nas Seções 5.2 e 5.3.

Um exemplo de um documento XML é mostrado na Figura 11 (a). Cada documento XML mantém o mesmo conteúdo textual do documento PDF original, permitindo inclusive a identificação de uma sentença de um parágrafo de uma página. Esta identificação é feita por meio de marcadores específicos que identificam as principais informações do artigo como: nome da revista, título, ano e autor. Além disso, o documento XML contém algumas etiquetas que estão organizadas em nível hierárquico: seção » página » parágrafo » sentença. Assim, é possível processar somente determinadas seções do artigo.

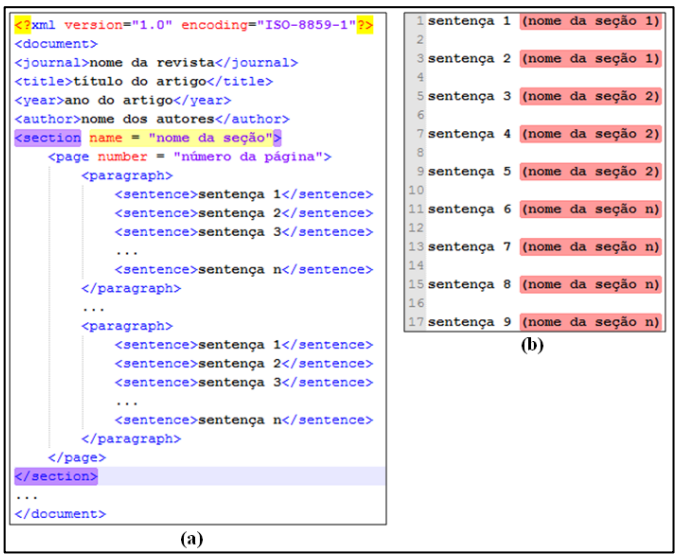

Figura 11 – Exemplo de um documento XML (a) e de um documento TXT (b).

Também é possível ter como entrada de dados documentos no formato TXT. Este formato é útil quando o artigo originalmente no formato PDF estiver protegido, ou quando não for possível converter um artigo do formato PDF para o formato XML, ou simplesmente utilizá-lo por conveniência. A criação do arquivo TXT é manual. A restrição para o processamento deste formato é que cada linha deve representar uma sentença e no final da sentença deve ser informado o nome da seção entre parênteses a qual a sentença faz parte. Na Figura 11 (b) é apresentado um exemplo.

5.2 Etapa 2 - Classificação de Sentenças

A partir dos documentos textuais é possível extrair informação. O primeiro passo da extração de informação é a Classificação de Sentenças, cujo objetivo é construir um modelo de classificação adequado que melhor represente as características das sentenças do domínio biomédico e com isso predizer qual a categoria de uma nova sentença. A classificação de sentenças supervisionada é composta por três fases: treinamento (Fase 1), teste (Fase 2) e uso do modelo (Fase 3). A classificação é supervisionada, pois os rótulos das classes são previamente conhecidos.

Na Fase 1, um classificador (ou modelo de classificação) é construído, a fim de descrever um determinado conjunto de dados. O modelo é criado a partir da análise do conjunto de treinamento. Este conjunto é rotulado em classes predefinidas, nas quais os rótulos possuem valores discretos e não ordenados. Na Figura 12 é mostrado um exemplo de sentenças rotuladas em suas respectivas classes.

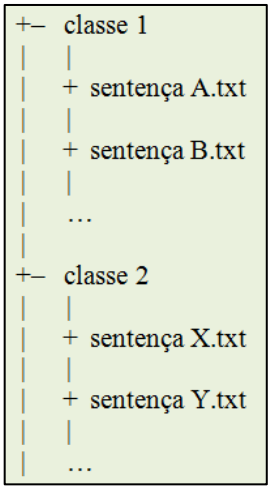

Figura 12 – Exemplo de sentenças rotuladas em suas respectivas classes.

A partir do modelo criado, é necessário avaliar se o modelo gerado é adequado para ser usado em sentenças cujo rótulo é desconhecido. Para isso, na Fase 2, sentenças que não foram utilizadas no treinamento são avaliadas com uma medida de desempenho. Tipicamente

são usadas as medidas acurácia, precisão, revocação e medida-F. Após a avaliação das sentenças, o modelo criado é utilizado na Fase 3 para classificar as novas sentenças dos artigos a serem processados, com o intuito de extrair informação do domínio biomédico.

A seguir é apresentado o processo de classificação de sentenças supervisionado que é composto por três etapas (Figura 13): Coleta dos Dados (Etapa 1), Pré-processamento (Etapa 2) e Categorização (Etapa 3).

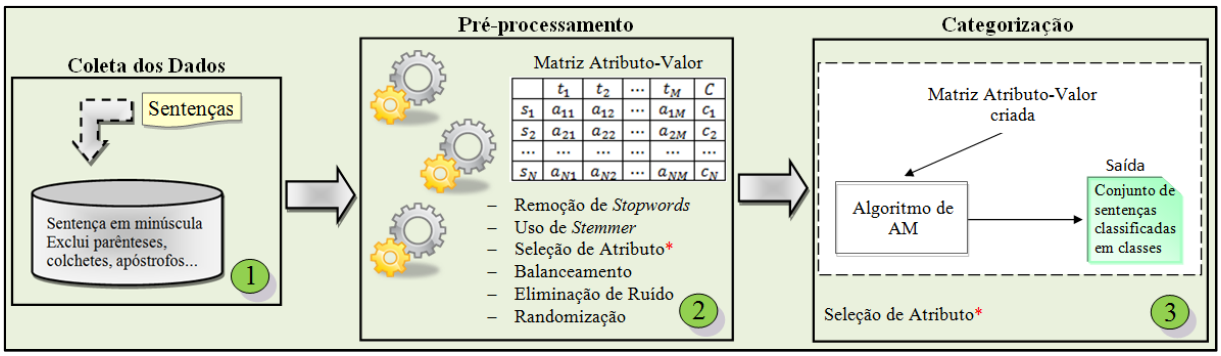

Figura 13 – Processo de classificação de sentenças supervisionado.

A Etapa 1 consiste na obtenção das sentenças a serem utilizadas para o treinamento e o teste do classificador. São aplicados alguns procedimentos no conjunto de sentenças, como: os caracteres das sentenças são colocados em letras minúsculas; são excluídos vírgula, ponto e vírgula, dois pontos, parênteses, colchetes, apóstrofos, sinal de mais ou menos (±) e excesso de espaço em branco. Nesta etapa, as categorias e o conjunto de sentenças correspondentes são definidos.

Na Etapa 2, as sentenças são estruturadas utilizando o modelo *bag-of-words*. Isto é necessário para que as sentenças possam ser manipuladas por algoritmos de aprendizado de máquina. Neste modelo as sentenças são organizadas em uma matriz atributo-valor. Cada linha i representa uma sentença s_i. Cada coluna l representa os termos ou uma sequência de termos (i.e., *n-grama*) $t_1, t_2, \dots, t_M$ presentes nas sentenças. Cada célula da matriz representa uma medida que relaciona a sentença e o termo. A medida binária pode ser utilizada com o intuito de representar a presença ou a ausência do termo na sentença. Ademais, cada sentença s_i está associada a uma classe c_i . Nesta Etapa pode-se aplicar algumas técnicas para a redução dos termos comuns e irrelevantes como a remoção de *stopwords* ou o ranqueamento dos termos mais importantes segundo algum critério, como a utilização de seleção de atributo. Ambas as técnicas visam reduzir a dimensionalidade do espaço de busca dos atributos, a fim de melhorar a precisão do algoritmo de indução. Também é possível aplicar filtros como: balanceamento das sentenças, eliminação de ruído ou randomização das sentenças. A técnica *stemmer* também pode ser utilizada.

Na Etapa 3 é realizada a classificação das sentenças propriamente dita. Um algoritmo de aprendizado de máquina é utilizado para classificar as sentenças em suas respectivas categorias. A técnica de seleção de atributo pode ser utilizada nesta etapa. O modelo criado é avaliado com alguma medida de desempenho, como acurácia, precisão e revocação. Este modelo é utilizado para classificar novas sentenças.

5.3 Etapa 3 - Identificação de Termos Relevantes

A partir das sentenças classificadas por um classificador, conforme descrito na Seção 5.2, é possível realizar a identificação de termos relevantes. Nesta etapa, é necessário identificar e extrair os termos relevantes em cada uma das sentenças de interesse classificadas na etapa anterior, ou seja, as sentenças que foram classificadas em uma classe de interesse são analisadas, sendo que as que foram classificadas em uma classe que não é de interesse são descartadas. Para isso, duas abordagens são utilizadas: dicionário e regra. A função do dicionário de terminologia é identificar em cada sentença do artigo qual é o termo curado que está presente na sentença. O objetivo é preencher o tipo-relacionamento artigo/termo, ou seja, inserir um dado que mostre que o termo ocorre em uma sentença. O intuito da regra é extrair novos termos automaticamente e inserir no dicionário. Os termos são inseridos no dicionário como termos não curados.

A seguir as duas abordagens de identificação de termos relevantes são explicadas mais detalhadamente.

5.3.1 Abordagem de Extração de Informação baseada em Dicionário

A abordagem de extração de informação baseada em dicionário tem o objetivo de identificar em quais sentenças os termos curados ocorrem e, consequentemente, preencher o tipo-relacionamento termo/artigo. Os termos curados são aqueles que foram validados manualmente por um especialista. É importante ressaltar que o dicionário não tem a função de identificar novos termos.

O dicionário terminológico é composto pelas tabelas presentes no esquema lógico derivado do esquema conceitual parcialmente representado na Figura 14 e pelas tabelas auxiliares Lista de Exclusão de Palavra (LEP) e Lista de Exclusão de Termo (LET). As tabelas LEP e LET são úteis para auxiliar na extração de termos que será explicada na seção a seguir (Seção 5.3.2).

A tabela LEP auxiliará na exclusão de palavras irrelevantes. Ela é composta por palavras comuns e gerais irrelevantes que não são do domínio biomédico e palavras irrelevantes do domínio biomédico que estão associadas a algum termo. As palavras comuns

contêm as 1000 palavras mais frequentes (http://www.bckelk.ukfsn.org/words/uk1000.html) que foram selecionadas dentre 4,6 milhões de palavras de um conjunto de 29 romances clássicos escritos por 18 autores do Reino Unido. Além dessas palavras, a tabela LEP contém palavras que são identificadas automaticamente nas sentenças processadas, ou seja, uma vez identificada uma nova palavra (i.e., um substantivo) acompanhada de uma determinada preposição, o substantivo é inserido automaticamente na tabela LEP. Essa identificação é feita por meio de uma regra específica, detalhada na Seção 5.3.2.

Já a tabela LET auxiliará na exclusão de um termo identificado erroneamente (i.e., um falso positivo). Ela é composta por termos substantivos simples, substantivos compostos e siglas do domínio biomédico que são irrelevantes e que sinalizam segmentos textuais que podem ser desconsiderados no processamento. Os termos inseridos na tabela LET são palavras que foram inseridas manualmente por meio de uma análise inicial em um conjunto de termos candidatos.

Na Figura 14 é mostrado um exemplo de um esquema conceitual (alguns atributos foram omitidos por questão de simplificação do esquema). Existem três tipos entidade (*Artigo*, *Termo* e *Variação*), sendo a última, um tipo entidade fraca.

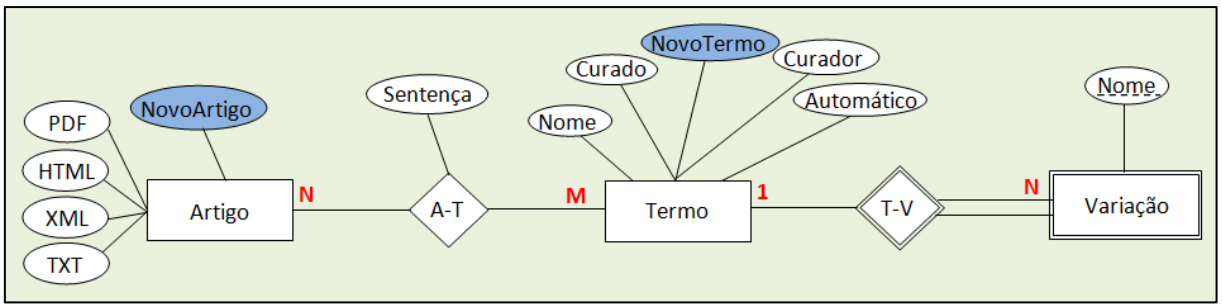

Figura 14 – Esquema conceitual biomédico.

O tipo entidade *Artigo* contém as informações do artigo, por exemplo, nome da revista, título, autor e artigo nos formatos PDF, HTML, XML e TXT. O tipo entidade *Termo* armazena informação sobre os termos relacionados ao domínio biomédico, por exemplo, nome do termo, se o termo foi curado, qual o nome do curador e se o termo foi inserido no dicionário por um processo automático ou manual. Um termo pode ser escrito de várias formas, isto é, pode ter variações. Assim, o tipo entidade fraca *Variação* armazena essas variações dos nomes de cada termo como sendo de um mesmo termo.

O repositório de termos contém termos que são curados e não curados. Somente os termos curados e suas variações são utilizados para identificar se o termo está presente na sentença (i.e., na função do dicionário). Na Figura 15 é apresentado um exemplo de termos curados e as variações de alguns termos. O “termo A“ possui as seguintes variações “termo

AA" e "termo AAs". A condição para existir termos e variações é que os nomes dos termos devem ser sempre nomes mais genéricos do que os nomes das variações e também incluir variações de número (i.e., singular e plural) e forma de representação (e.g. xxx of yyy e yyy xxx).

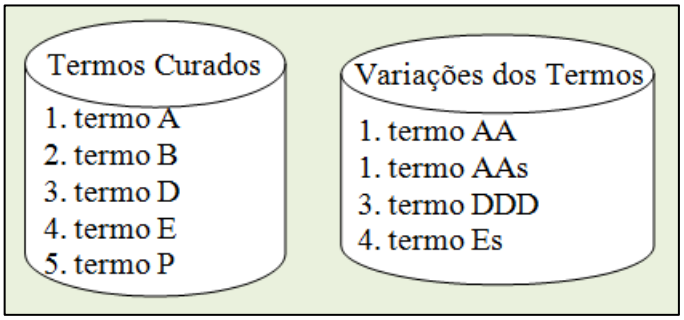

Figura 15 – Exemplo de termos curados e suas variações.

Para evitar uma nova identificação de um termo que já foi previamente identificado em um artigo, é necessário o uso de dois atributos para controlar o processamento: $NovoTermo$ e $NovoArtigo$ (atributos destacados na cor azul na Figura 14).

O Algoritmo 1 descreve o pseudocódigo da identificação de termos em novos artigos. Inicialmente, nas linhas 1 e 2 são selecionados do dicionário os artigos novos ($NovoArtigo = true$) e os termos que são curados e não são novos termos ($NovoTermo = false$). Em seguida, as sentenças classificadas do artigo novo em questão são selecionadas (linha 4). Os termos são identificados em cada sentença de interesse (linhas 5 a 9). Após o processamento das sentenças de cada artigo novo, o artigo é definido como processado (linha 10).

Algoritmo 1 – Identifica termos em novos artigos.

```
1  Artigos[] ← getNovoArtigo();
2  TermosCurado[] ← getNovoTermo(false);
3  for (i ← 1; i ≤ quantidade de Artigos; i ← i + 1)
4  |  Sentenças[] ← getSentençaClassificada(Artigos[i], true);
5  |  for (j ← 1; j ≤ quantidade das Sentenças; j ← j + 1)
6  |  |  for (k ← 1; k ≤ quantidade de TermosCurado; k ← k + 1)
7  |  |  |  identificaTermo(Sentenças[j], TermosCurado[k]);
8  |  |  end
9  |  end
10 |  setNovoArtigo(false);
11 end
```

O Algoritmo 2 mostra o pseudocódigo da identificação de termos em todos os artigos. Nas linhas 1 e 2 são selecionados do dicionário todos os artigos e todos os termos que são curados e novos ($NovoTermo = true$). Em seguida, as sentenças classificadas do artigo são

selecionadas (linha 4). Os termos são identificados em cada sentença de interesse (linhas 5 a 9). Após o processamento das sentenças de cada artigo, o artigo é definido como processado (linha 10). Por fim, todos os novos termos são definidos como termos processados (linhas 12 a 14).

Algoritmo 2 – Identifica termos em todos os artigos.

```
 1  Artigos[] ← getTodosArtigos();
 2  TermosCurado[] ← getNovoTermo(true);
 3  for (i ← 1; i ≤ quantidade de Artigos; i ← i + 1)
 4  |  Sentenças[] ← getSentençaClassificada(Artigos[i], true);
 5  |  for (j ← 1; j ≤ quantidade das Sentenças; j ← j + 1)
 6  |  |  for (k ← 1; k ≤ quantidade de TermosCurado; k ← k + 1)
 7  |  |  |  identificaTermo(Sentenças[j], TermosCurado[k]);
 8  |  |  end
 9  |  end
10  |  setNovoArtigo(false);
11  end
12  for (i ← 1; i ≤ TermoCurado; i ← i + 1)
13  |  setNovoTermo(false);
14  end
```

Resumindo os dois algoritmos explicados anteriormente: se o termo curado é novo, então o processa em todos os artigos; caso o termo curado não seja novo, então o processa somente em artigos que são novos, pois esse mesmo termo já foi processado nos artigos "antigos". Caso as duas condições não sejam verdadeiras (i.e., se não existe termos novos e se não existe artigos novos), então não há a necessidade de utilizar o dicionário para identificar em quais sentenças ocorrem os termos curados. Portanto, essas restrições evitarão um processamento desnecessário.

A carga de dados no dicionário pode ser realizada de forma manual ou automática. Na carga manual, os termos são inseridos com a participação do especialista do domínio por meio de um recurso de gerenciamento de termos (Seção 5.4). O dicionário deve ser construído por meio de uma carga inicial de dados. Já na carga de dados automática, os termos são identificados nos artigos científicos por meio da abordagem de regra discutida na Seção 5.3.2.

5.3.2 Abordagem de Extração de Informação baseada em Regras

Regras são construídas utilizando expressões regulares que proveem um mecanismo para identificar padrões textuais. Para auxiliar as expressões regulares na identificação dos termos relevantes contidos em uma sentença, a técnica de Processamento de Língua Natural *Part-Of-Speech* (POS) é utilizada.

O etiquetador POS consiste em rotular as palavras segundo a sua classe gramatical. Substantivo, adjetivo, advérbio, verbo e preposição são alguns exemplos de classes gramaticais. A etiquetação é baseada na própria definição da palavra, assim como no contexto ao qual a palavra está inserida. Exemplo de um contexto é o relacionamento de palavras associadas em uma sentença ou em um parágrafo.

Na metodologia proposta, duas estratégias utilizando padrões POS são aplicadas para extrair informação das sentenças: Verbo e Expressão com POS e somente POS. A primeira estratégia utiliza-se de verbos representativos e expressões compostas representativas para identificar se uma sentença contém ou não um termo. Caso a sentença contenha o verbo ou a expressão, então padrões POS são utilizados para extrair os termos relevantes na sentença. A segunda estratégia utiliza padrões POS mais específicos visando alcançar dois objetivos: o primeiro é identificar termos que a primeira estratégia não consegue identificar; o segundo é extrair termos com uma baixa ocorrência de falsos positivos.

Os padrões POS são uma sequencia de etiquetas POS que estão associadas a um conjunto de palavras. Por exemplo, o padrão JJ_NN é uma sequência de etiquetas POS, sendo a primeira etiqueta um adjetivo e a segunda etiqueta um substantivo, as quais podem ser associadas, respectivamente, as palavras *harmful* e *sickness* em *harmful*_JJ *sickness*_NN.

A seguir, a primeira e a segunda estratégias são explicadas detalhadamente.

Estratégia 1: Uso de verbo e expressão com POS para extração de termos relevantes

Na Figura 16 é apresentado um exemplo de extração de termos utilizando a Estratégia 1. Esta estratégia considera que um conjunto de três letras significa uma palavra etiquetada em sua classe gramatical, sendo que os termos relevantes são representados pelos caracteres RRR e os termos irrelevantes pelos caracteres III.

Para explicar o funcionamento da Estratégia 1, considere as três sentenças iniciais mostradas no passo 1 da Figura 16. As duas primeiras sentenças contêm um termo relevante sendo indicado por um verbo e uma expressão composta destacados na cor amarela. A terceira sentença, apesar de conter termos relevantes, representados por RRR, não contém nenhum verbo ou expressão composta representativos indicando um termo candidato e por isso, não é identificada pela Estratégia 1.

No passo 2 da Figura 16, duas sentenças dentre as três sentenças iniciais são selecionadas, uma vez que elas contêm um verbo ou uma expressão composta. O verbo ou a expressão delimitam em qual “Parte Específica” da sentença (i.e., antes ou depois do verbo ou da expressão composta) pode haver um termo relevante. Esta parte específica é destacada na

cor cinza como é mostrado no passo 2 da Figura 16. Primeiramente, é proposto que se aplique dois padrões POS na parte específica selecionada, a fim de eliminar falsos positivos: $[A-Za-z]\{1,3\}\ /\ [A-Za-z]\{1,3\}\_NN[PS]?$ e $(JJ)?\_NN\_(of\_IN)$. O primeiro padrão significa que o falso positivo a ser removido é uma palavra substantiva de uma até três letras, seguida de uma barra com uma até três letras. O objetivo deste padrão é excluir medidas como g/dL e cm/sec das sentenças, evitando a identificação de falso positivo. O segundo padrão significa que o falso positivo a ser removido é uma palavra substantiva seguida da preposição "*of*" (e.g., *analysis of*) ou é a composição de um adjetivo mais um substantivo também seguido da preposição "*of*" (e.g., *previous history of*). Os substantivos identificados por esse último padrão são inseridos automaticamente na Lista de Exclusão de Palavra, conforme previamente explicado na Seção 5.3.1. Na sentença 1 existe um falso positivo indicado pela cor vermelha. Este falso positivo é removido por um dos padrões anteriores.

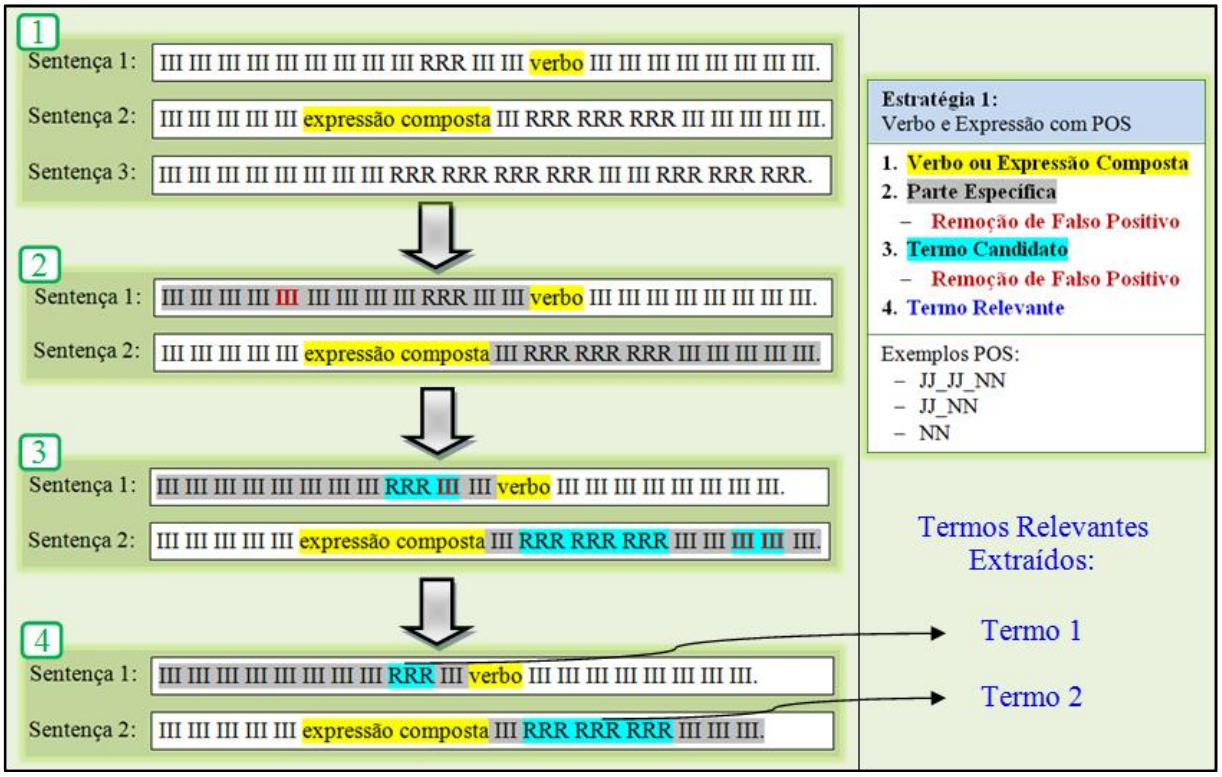

Figura 16 – Exemplo de termos extraídos pela Estratégia 1. Termos relevantes são representados pelos caracteres RRR e os termos irrelevantes pelos caracteres III.

Em seguida, no passo 3 da mesma figura, o possível termo destacado na cor turquesa, denominado de "Termo Candidato", é identificado por meio de padrões POS. Por exemplo, o termo candidato da sentença 1 poderia ter sido identificado pelo padrão POS JJ_NN. E os termos candidatos da sentença 2 poderiam ter sidos identificados, respectivamente, pelos padrões JJ_JJ_NN e JJ_NN. O termo candidato selecionado pode conter uma palavra que não faz parte do termo (e.g., termo candidato RRR III da sentença 1) ou o termo candidato pode

ser um falso positivo (e.g., termo candidato III III da sentença 2). No primeiro caso, a Lista de Exclusão de Palavra é consultada para remover a palavra que não faz parte do termo. No segundo caso, a Lista de Exclusão de Termo é consultada para remover o falso positivo.

Por fim, após a remoção de falsos positivos nos termos candidatos, no passo 4 da Figura 16, pode-se notar a extração de termos relevantes pela Estratégia 1. Os passos da Estratégia 1 estão resumidos na Figura 16 do lado direito.

O Algoritmo 3 descreve o pseudocódigo da extração de termos utilizando a Estratégia 1. Inicialmente, são selecionadas as sentenças classificadas, os verbos representativos e as expressões compostas representativas e os padrões POS da Estratégia 1 (linhas 1 a 3). Cada sentença é processada a fim de serem extraídos termos relevantes (linhas 4 a 23). Se existe um verbo ou uma expressão representativos na sentença analisada (linha 6), então a "Parte Específica" não está vazia (linha 7), indicando que existe um termo relevante nesta parte da sentença selecionada. Portanto, é necessário processar esta parte específica para remover falsos positivos com os padrões POS $[A-Za-z]\{1,3\}\,/\,[A-Za-z]\{1,3\}\_NN[PS]?$ e $(JJ)?\_NN\_(of\_IN)$ (linha 8). O substantivo identificado por este último padrão é inserido na Lista de Exclusão de Palavra.

Algoritmo 3 – Extrai termo utilizando a Estratégia 1.

```
 1 Sentenças[] ← getSentençaClassificada(true);
 2 VerboExpressão[] ← getVerboExpressão();
 3 PadrãoPOS[] ← getPadrãoPOS();
 4 for (i ← 1; i ≤ quantidade das Sentenças; i ← i + 1)
 5     for (j ← 1; j ≤ quantidade do VerboExpressão; j ← j + 1)
 6         ParteEspecífica ← getParteEspecífica(Sentenças[i], VerboExpressão[j]);
 7         if (ParteEspecífica não está vazia) then
 8             ParteEspecífica ← RemoverFalsoPositivo(ParteEspecífica);
 9             for (k ← 1; k ≤ quantidade do PadrãoPOS; k ← k + 1)
10                 TermoCandidato ← getTermoCandidato(ParteEspecífica, PadrãoPOS[k]);
11                 if (TermoCandidato não está vazio)then
12                     Termo ← UsarListaExclusãoTermo(TermoCandidato);
13                     if (Termo não está vazio)then
14                         print("É termo");
15                         Termo ← UsarListaExclusãoPalavra(Termo);
16                     else
17                         print("Não é termo");
18                     end
19                 end
20             end
21         end
22     end
23 end
```

Em seguida, cada padrão POS da Estratégia 1 é aplicado na parte específica, a fim de identificar um termo candidato (linha 10). Se for identificado o padrão POS para esta parte específica, então existe um termo candidato para o padrão POS em questão (linha 11). Este termo candidato pode ser um falso positivo. Assim, é necessário consultar a Lista de Exclusão de Termo (LET) para identificar se este termo candidato é realmente um falso positivo (linha 12). Se o termo candidato não contém uma palavra da LET, então o mesmo é um termo (linha 14). Caso contrário, não é um termo (linha 17). Por fim, sendo o termo candidato um termo, então é necessário remover as possíveis palavras irrelevantes presentes na Lista de Exclusão de Palavra (linha 15).

Estratégia 2: Uso de POS para a extração de termos relevantes

Na explicação da Estratégia 1 ilustrada na Figura 16, a extração de informação realizada pela Estratégia 1 evita a aplicação de padrões em toda a sentença, o que, por conseguinte, evita a extração de falsos positivos. Contudo, a Estratégia 1 não é capaz de extrair todos os termos presentes em uma sentença. Para extrair termos que a Estratégia 1 não consegue identificar e ao mesmo tempo não extrair muitos falsos positivos, propõe-se uma segunda estratégia que utiliza padrões POS mais específicos que são aplicados em toda a sentença.

Na Figura 17 é apresentado um exemplo de extração de termos utilizando a Estratégia 2. Esta estratégia também considera que o conjunto de três letras significa uma palavra etiquetada em sua classe gramatical, sendo que os termos relevantes são representados pelos caracteres RRR e os termos irrelevantes pelos caracteres III.

Para explicar o funcionamento da Estratégia 2, considere as três sentenças iniciais mostradas no passo 1 da Figura 17. Todas as três sentenças contêm um termo relevante. Nota-se que a primeira sentença contém um verbo representativo. No passo 2 da Figura 17, os termos candidatos, destacados na cor turquesa, são identificados por meio de padrões POS em duas das três sentenças. Apesar da terceira sentença possuir termos relevantes, neste caso supõe-se que estes termos não foram identificados pela Estratégia 2. Nas duas primeiras sentenças, foram identificados termos candidatos. Estes termos poderiam ter sido identificados pelos seguintes padrões POS: JJ_NN_NN_NN na sentença 1; e JJ_JJ_NN e JJ_JJ_NN_NN_NN na sentença 2. O termo candidato selecionado pode conter uma palavra que não faz parte do termo (e.g., termo candidato III RRR RRR RRR da sentença 1) ou o termo candidato pode ser um falso positivo (e.g., termo candidato III III III da sentença 2). No primeiro caso, a Lista de Exclusão de Palavra é consultada para remover a palavra que não faz

parte do termo. No segundo caso, a Lista de Exclusão de Termo é consultada para remover o falso positivo. Por fim, após a remoção de falsos positivos nos termos candidatos, no passo 3 da Figura 17, pode-se notar a extração de termos relevantes pela Estratégia 2. Os passos da Estratégia 2 estão resumidos na Figura 17 do lado direito.

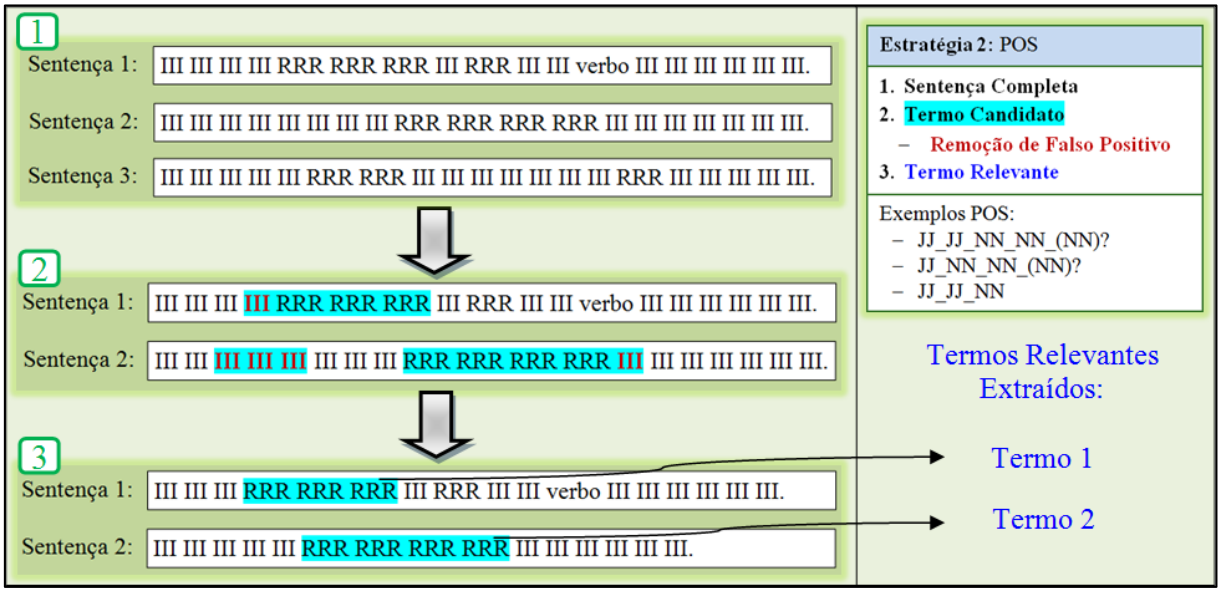

Figura 17 – Exemplo de termos extraídos pela Estratégia 2. Termos relevantes são representados pelos caracteres RRR e os termos irrelevantes pelos caracteres III.

O Algoritmo 4 descreve o pseudocódigo da extração de termo utilizando a Estratégia 2. Inicialmente, são selecionadas as sentenças classificadas e os padrões POS da Estratégia 2 (linhas 1 a 2). Cada sentença é processada, a fim de serem extraídos termos relevantes (linhas 3 a 16). Para cada sentença é aplicado um padrão POS da Estratégia 2, a fim de identificar um termo candidato (linha 5). Se for identificado o padrão POS para esta sentença, então existe um termo candidato para o padrão POS em questão (linha 6). Se o termo candidato não contém uma palavra da Lista de Exclusão de Termo, então o mesmo é um termo (linha 9). Caso contrário, não é um termo (linha 12). Sendo um termo, então é necessário remover as possíveis palavras irrelevantes presentes na Lista de Exclusão de Palavra (linha 10).

Algoritmo 4 – Extrai termo utilizando a Estratégia 2.

```
1  Sentenças[] ← getSentençaClassificada(true);
2  PadrãoPOS[] ← getPadrãoPOS();
3  for (i ← 1; i ≤ quantidade das Sentenças; i ← i + 1)
4  |  for (j ← 1; j ≤ quantidade do PadrãoPOS; j ← j + 1)
5  |  |  TermoCandidato ← getTermoCandidato(Sentenças[i], PadrãoPOS[j]);
6  |  |  if (TermoCandidato não está vazio)then
7  |  |  |  Termo ← UsarListaExclusãoTermo(TermoCandidato);
8  |  |  |  if (Termo não está vazio)then
9  |  |  |  |  print("É termo");
10 |  |  |  |  Termo ← UsarListaExclusãoPalavra(Termo);
11 |  |  |  else
12 |  |  |  |  print("Não é termo");
13 |  |  |  end
14 |  |  end
15 |  end
16 end
```

5.4 Etapa 4 - Gerenciamento de Termos

Como foi destacado anteriormente, o especialista tem um papel importante na validação dos termos extraídos automaticamente e também na inicialização do dicionário e na inserção de novos termos manualmente. Nesta etapa, o especialista pode realizar quatro operações:

1. Inserir novos termos: os novos termos inseridos são definidos automaticamente como termos curados. Assim, é possível os termos serem utilizados pela abordagem de dicionário para localização de termo em uma sentença e, portanto, identificar o tipo-relacionamento termo/sentença;
2. Hierarquizar termos: o especialista pode inserir ou atualizar um termo como uma variação de outro termo. A restrição para existir um termo com variações é que o nome do termo deve ser mais genérico do que os nomes das suas variações;
3. Validar termos extraídos: os termos identificados pela abordagem de regra são inseridos no dicionário como termos não curados. Assim, o especialista pode remover o termo caso seja um falso positivo ou definir o termo como termo curado, isto é, o termo é realmente um termo do domínio e, portanto, é um termo relevante;
4. Mover termos extraídos: o especialista também pode mover um termo de um tipo de categoria para outra. Por exemplo, considere que o termo X foi inserido na categoria Y. Contudo, o termo X está relacionado à categoria Z. Portanto, o termo X deve ser movido da categoria Y para a categoria Z. Este recurso é necessário quando um termo extraído não for armazenado no tipo de categoria correta.

5.5 Considerações Finais

Neste capítulo foi descrita a metodologia proposta de pré-processamento textual com o intuito de extrair informação em artigos científicos do domínio biomédico. Esta metodologia de pré-processamento textual é composta por quatro etapas: Entrada de Dados (Etapa 1), Classificação de Sentenças (Etapa 2), Identificação de Termos Relevantes (Etapa 3) e Gerenciamento de Termos (Etapa 4).

Na Etapa 1, os documentos textuais são selecionados. A Etapa 2 é uma etapa imprescindível na metodologia, pois esta etapa é responsável por distinguir, dentre todas as sentenças processadas, quais são as sentenças de interesse, isto é, quais as sentenças que possivelmente conterão algum termo relevante. Assim, na Etapa 3, a abordagem baseada em regra evita o processamento em sentenças que não são de interesse, o que evita também a extração de falsos positivos, já que podem existir padrões POS nas sentenças que não são de interesse e esses padrões podem, consequentemente, extrair um termo que não seja relevante. Portanto, os verbos representativos e as expressões compostas representativas e os padrões POS das estratégias de extração, somente são aplicados nas sentenças que forem classificadas em alguma classe de interesse. Os termos armazenados no dicionário também são identificados somente em sentenças de interesse, o que diminui o custo de processamento na Etapa 3, pois nem toda sentença é processada a fim de identificar se o termo curado está presente. Por fim, na Etapa 4, o especialista pode inserir novos termos, hierarquizar termos existentes ou novos termos, validar e mover os termos extraídos pela abordagem de regra.

Os termos são extraídos nos artigos e seções definidas pelo usuário, utilizando a abordagem de regra. Todos os termos extraídos automaticamente são inseridos no dicionário como termo não curado. O termo somente será utilizado pela abordagem de dicionário quando o mesmo for curado pelo especialista na Etapa 4 da metodologia. A vantagem de se utilizar somente termos curados para identificar a presença do termo na sentença é que os termos que serão usados pelo dicionário passaram obrigatoriamente pelo crivo do especialista, o que aumenta a confiança na identificação dos termos. Outra vantagem é que um termo, extraído em um artigo X, pode ser identificado pelo dicionário nos artigos X, Y e Z, já que pode existir um determinando termo em outros artigos. Por outro lado, a utilização dos termos extraídos depende da validação do especialista no processo de identificação de um novo termo.

No próximo capítulo esta metodologia é instanciada com exemplos de uma área do domínio biomédico.

6 INSTANCIAÇÃO DA METODOLOGIA PROPOSTA

No capítulo anterior foi descrita a metodologia proposta para a extração de informação no domínio biomédico. Neste capítulo, essa metodologia é instanciada com informações textuais sobre a doença Anemia Falciforme (PINTO et al., 2009), doravante chamada apenas de AF. O objetivo deste capítulo é mostrar como de fato é aplicada a metodologia proposta em um domínio específico e com isto ajudar no entendimento das etapas da metodologia.

As informações a serem extraídas são encontradas em artigos científicos escritos em inglês e são especificamente sobre *efeitos* da doença AF. Os efeitos podem ser efeito negativo da doença, efeito negativo do tratamento e efeito positivo, a saber:

- Efeito negativo da doença (ou complicação): qualquer efeito negativo inerente da doença, ou seja, decorrentes das hemácias falciformes, independente do uso de um determinado tratamento. Síndrome torácica aguda (*acute chest syndrome*), sequestro esplênico (*splenic sequestration*) e falha renal crônica (*chronic renal failure*) são alguns exemplos de complicações da AF. Sintomas da doença também são considerados complicações, como febre (*fever*), hemorragia (*hemorrhage*) e inflamação dos dedos do pé e da mão (*dactylitis*);
- Efeito negativo do tratamento (ou efeito colateral): problemas ocasionados por estímulos do tratamento, ou seja, são os efeitos negativos de um tratamento. O uso de certas drogas ou terapias pode causar nos pacientes com AF leucemia (*leukemia*), contágio por vírus devido à transfusão de sangue (HIV) e depressão (*depression*), dentre outros efeitos colaterais;
- Efeito positivo: melhorias ou benefícios ocasionados por estímulos do tratamento, ou seja, são os efeitos positivos de um tratamento, como remissão da doença (*disease remission*), melhora clínica (*clinical improvement*) e redução no tempo de internação (*reduction in hospitalization time*).

Na Figura 18 é apresentada a instanciação da metodologia proposta com informações sobre efeitos da doença AF. A Entrada de Dados (Etapa 1) consiste de artigos científicos que

estão no formato TXT ou XML. A Classificação de Sentenças (Etapa 2) objetiva distinguir as informações de interesse das informações irrelevantes. Nesta etapa, as classes de interesse são efeito negativo (complicação e efeito colateral), efeito positivo do tratamento e outros (i.e., não é efeito negativo ou efeito positivo). As sentenças de interesse são as sentenças que são classificadas nas classes efeito negativo ou efeito positivo, e por outro lado, as sentenças irrelevantes serão as sentenças classificadas na classe outros. Em seguida, na etapa de Identificação de Termos Relevantes (Etapa 3), os termos relevantes extraídos são referentes às sentenças da classe efeito negativo que é a classe de objeto de estudo deste trabalho. Os termos extraídos são sobre a classe efeito negativo devido à existência nos artigos científicos de uma maior quantidade de efeitos negativos do que de efeitos positivos, o que possibilitará uma maior quantidade de informações armazenadas no banco de dados a fim de possibilitar a aplicação de algoritmos de mineração de dados para descoberta de conhecimento. O fato da classe efeito positivo possuir uma menor quantidade de informação comparada com a classe efeito negativo e não ter sido utilizada neste trabalho, não significa que a classe efeito positivo não contenha informação importante a ser extraída. A identificação de termos relevantes da classe efeito positivo é uma questão que não foi investigada neste trabalho, contudo necessita ser instanciada pela Etapa 3 da metodologia proposta. No Gerenciamento de Termos (Etapa 4), os termos que realmente são efeitos negativos são armazenados no banco de dados com o auxílio da validação do especialista.

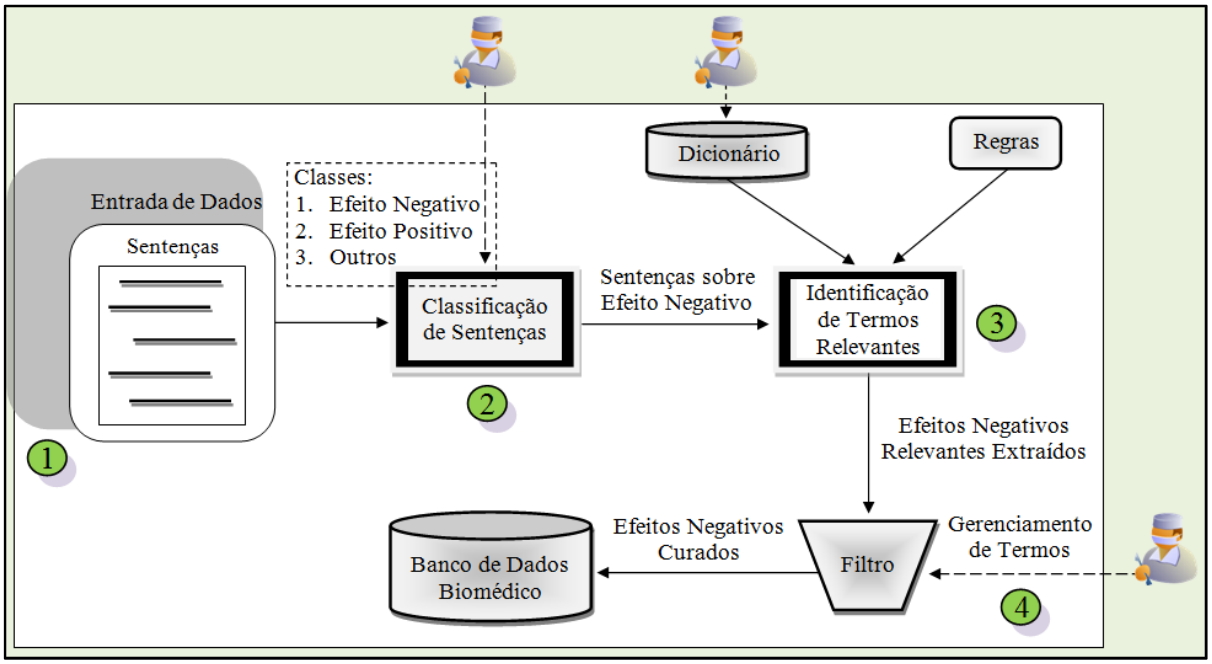

Figura 18 – Instanciação da metodologia de pré-processamento para extração de informação no domínio da AF.

Além da instanciação, também são apresentadas as principais características das ferramentas computacionais desenvolvidas durante o mestrado (Seção 6.5) para dar suporte às três últimas etapas da metodologia.

6.1 Entrada de Dados

A entrada de dados consiste de artigos científicos que estão originalmente no formato não estruturado PDF. Na Figura 19 é apresentado um exemplo de um artigo científico da doença AF originalmente no formato PDF. É necessário converter esse formato para um formato que permita o processamento textual. Dois formatos semiestruturados são aceitos para dar suporte ao processo de extração de informação: TXT e XML.

A conversão do formato PDF para o formato XML é feita por meio da ferramenta SCA-Translator desenvolvida por Carosia e Ciferri (2010). Na Figura 20 é apresentado o mesmo artigo da Figura 19 convertido pela ferramenta SCA-Translator para o documento XML. O documento XML mantém o mesmo conteúdo textual do documento PDF original, possibilitando identificar qual a página, o parágrafo e a seção de uma determinada sentença.

Por outro lado, a conversão do formato PDF para o formato TXT é feita de forma manual. Na Figura 21 é apresentado o mesmo artigo da Figura 19 no formato TXT. Figura 21 é possível identificar qual a seção que a sentença faz parte analisando o final de cada sentença (i.e., o nome da seção entre parênteses).

CLINICAL OBSERVATIONS, INTERVENTIONS, AND THERAPEUTIC TRIALS

Hydroxyurea for sickle cell disease in children and for prevention of cerebrovascular events: the Belgian experience

Béatrice Gulbis, David Haberman, Dominique Dufour, Catherine Christophe, Christiane Vermylen, Faustin Kagambega, Francis Corazza, Christine Devalck, Marie-Françoise Dresse, Kathleen Hunninck, Axel Klein, Phu Quoc Le, Michèle Loop, Philip Maes, Pierre Philippet, Eric Sariban, Chris Van Geet, and Alina Ferster

Hydroxyurea (HU) is considered to be the most successful drug therapy for severe sickle cell disease (SCD). Nevertheless, questions remain regarding its benefits in very young children and its role in the prevention of cerebrovascular events. There were 127 SCD patients treated with no attempt to reach maximal tolerated doses who entered the Belgian Registry: 109 for standard criteria and 18 who were at risk of stroke only. During 426 patient-years of follow-up for patients with standard criteria, 3.3 acute chest syndromes, 1.3 cerebrovascular events, and 1.1 osteonecrosis per 100 patient-years were observed. A subgroup of 32 patients followed for 6 years experienced significant benefit over this period. In each subgroup of children (younger than 2 years, 2-5, 6-9, and 10-19 years) followed for 2 years, clinical and biologic changes were similar, except for children younger than 2 years who had no total hemoglobin increase and remained at risk of severe anemia. In 72 patients evaluated by transcranial Doppler studies (TCD), 34 patients were at risk of primary stroke and only 1 had a cerebrovascular event after a follow-up of 96 patient-years. These results confirm the benefit of HU, even in very young children, and its possible role in primary stroke prevention. (Blood. 2005; 105:2685-2690)

Introduction

Since the first results of the Multicenter Study of Hydroxyurea (MSH) in adult sickle cell anemia patients published by Charache et al,[1] hydroxyurea (HU) has been successfully used in patients with sickle cell disease (SCD) who have recurrent painful crisis and/or acute chest syndrome (ACS)[2,3] as well as in patients with beta-thalassemia major.[4] In addition to its myelosuppressive effect and ability to induce fetal hemoglobin (HbF) synthesis, HU has been shown to reduce the adhesion of sickle erythrocytes to endothelial cells and to modulate endothelial cell activation and nitric oxide generation.[5]

Patients, materials, and methods

Patients enrolled

The Belgian Registry including SCD patients treated by HU was opened in 1993. Approval was obtained from the Pediatric Department of the Children University Hospital Reine Fabiola review board for this registry. Informed consent was provided according to the Declaration of Helsinki. Entry criteria in this registry, clinical monitoring, and laboratory measurements have been described in detail previously.[2] Briefly, to be eligible, patients had to have reported the year before entry at least 2 or more painful vaso-occlusive crises (VOCs) per year requiring hospitalization; one

Figura 19 – Exemplo de uma página de um artigo científico da AF no formato PDF.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<document>
<journal>CLINICAL OBSERVATIONS, INTERVENTIONS, AND THERAPEUTIC TRIALS</journal>
<title>Hydroxyurea for sickle cell disease in children and for prevention
of cerebrovascular events: the Belgian experience</title>
<year>2005</year>
<author>Beatrice Gulbis, David Haberman,... Alina Ferster</author>
<section name = "abstract">
    <page number = "1">
        <paragraph>
            <sentence>Hydroxyurea (HU) is considered to be the
            most successful drug therapy for severe
            sickle cell disease (SCD).</sentence>
            ...
            <sentence>These results
            confirm the benefit of HU, even in very
            young children, and its possible role in
            primary stroke prevention.</sentence>
        </paragraph>
    </page>
</section>
<section name = "introduction">
    <page number = "1">
        <paragraph>
            <sentence>Since the first results of the Multicenter
            Study of Hydroxyurea (MSH) in adult sickle cell anemia
            patients...</sentence>
            ...
        </paragraph>
        ...
    </page>
</section>
...
</document>
```

Figura 20 – Exemplo de um documento XML gerado pela ferramenta SCA-Translator.

```
1  A subgroup of 32 patients followed for 6 years experienced significant benefit over this period. (Abstract)
2
3  Hydroxyurea (HU) is considered to be the most successful drug therapy for severe sickle cell disease (SCD). (Abstract)
4
5  Acute promyelocytic leukemia was diagnosed in a 21-year-old female after 8 years of HU therapy. (Results)
6
7  Recurrent stroke was observed once in an 8-year-old girl, 6 years after the initial event. (Results)
8
9  One patient who died was described previously. (Results)
10
11 Repeated vaso-occlusive crises are sometimes present from infancy and early childhood. (Discussion)
12
13 One patient died of a fatal episode of splenic sequestration before completing 2 years of treatment. (Discussion)
14
15 In another cohort of 5 patients, HU was effective at preventing recurrent stroke. (Discussion)
16
17 More interesting are the data on patients at risk of secondary stroke. (Discussion)
```

Figura 21 – Exemplo de um arquivo TXT.

A principal diferença em extrair informação de artigos nos formatos XML e TXT é que no formato XML as informações textuais estão organizadas em nível hierárquico, assim é possível saber qual a sentença, qual o parágrafo, qual a página e qual a seção que o termo relevante extraído se encontra. Isto possibilita a identificação precisa da informação extraída. Portanto, é possível destacar o termo relevante no artigo no seu formato correspondente em HTML, o que possibilita ao especialista saber qual o contexto da informação extraída. Esta precisão não é possível se o formato usado é o TXT.

Por outro lado, a ferramenta SCA-Translator implementou um conjunto de tradutores que permite a conversão correta de artigos PDF de um conjunto de periódicos sobre a doença AF para o formato XML. Porém, para muitos periódicos esta ferramenta ainda não possui

tradutores e, portanto uma forma de processar os artigos PDF destes periódicos é por meio do uso de um arquivo TXT que é o segundo formato proposto de entrada de dados.

6.2 Classificação de Sentenças

A partir dos artigos científicos sobre a doença Anemia Falciforme sugeridos pelos especialistas, é possível extrair informação. O primeiro passo da extração de informação é a Classificação de Sentenças, cujo objetivo é construir um modelo de classificação adequado que melhor represente as características das sentenças de treinamento e com isso, predizer qual a categoria de uma nova sentença.

A classificação de sentenças supervisionada é composta por três fases: treinamento (Fase 1), teste (Fase 2) e uso do modelo (Fase 3). Na Fase 1, o classificador é construído, a fim de descrever o conjunto de sentenças. Este conjunto é rotulado em classes predefinidas. Na Figura 22 são mostradas as classes predefinidas relacionadas à doença AF com suas respectivas sentenças.

```
+– classe 1 → efeito negativo
|   |
|   + Recurrent stroke was observed once in an 8-year-old girl, 6 years after the initial event.
|   |
|   + In 6 cases, transient hematologic toxicity was reported.
|   |
|   …
|
+– classe 2 → efeito positivo
|   |
|   + The use of HU at MTD may bring additional benefit.
|   |
|   + Used carefully, with frequent monitoring, hydroxyurea therapy was safe.
|   |
|   …
|
+– classe 3 → outros
|   |
|   + The x-globin genotype was unknown for most of the patients.
|   |
|   + In some adolescents, poor compliance was evident.
|   |
|   …
```

Figura 22 – Exemplo da estrutura dos arquivos de treinamento.

A partir do modelo criado, é necessário avaliar se o modelo gerado é adequado para ser usado em sentenças cujo rótulo é desconhecido. Para isso, na Fase 2, sentenças que não foram utilizadas no treinamento foram avaliadas com a medida de desempenho acurácia. Para calcular a acurácia, o rótulo da sentença testada é comparado com o rótulo da sentença classificada. O método de particionamento *10-Fold Cross-Validation* foi utilizado para

estimar a acurácia do classificador. Após a avaliação das sentenças, o modelo criado foi utilizado na Fase 3.

O processo de classificação de sentenças supervisionado apresentado na Figura 13 da Seção 5.2 foi utilizado. O processo é composto por três etapas: Coleta dos Dados (Etapa 1), Pré-processamento (Etapa 2) e Categorização (Etapa 3).

Na Etapa 1, o conjunto de sentenças de treinamento foi definido manualmente com a ajuda do especialista do domínio da AF. Algumas sentenças deste conjunto e as respectivas classes destas sentenças podem ser vistas na Figura 22. Na Etapa 2, as sentenças são estruturadas utilizando o modelo *bag-of-words*. A matriz atributo-valor é construída utilizando a frequência mínima igual a dois para selecionar os atributos que ocorreram no mínimo duas vezes nas sentenças, ou seja, os atributos que ocorreram somente uma vez não foram considerados. Os atributos são formados de 1 a 3 gramas. A medida binária, que considera que o valor 1 representa a ocorrência do *n-grama* na sentença e o valor 0 caso contrário, foi utilizada. As técnicas de balanceamento das sentenças e remoção de ruído também foram utilizadas para respectivamente, balancear a distribuição das sentenças entre as classes e remover sentenças que estejam dificultando e atrapalhando o aprendizado.

Na Etapa 3 é realizada a classificação das sentenças propriamente dita. Seis algoritmos clássicos de aprendizado de máquina foram escolhidos para serem avaliados na classificação das sentenças. Os algoritmos escolhidos foram de diferentes paradigmas: *Support Vector Machine* (SVM) e Naïve Bayes (NB) são estatísticos; ID3, J48, Prism e OneR são algoritmos de aprendizado simbólico, os dois primeiros são algoritmos de árvore de decisão e os dois últimos são algoritmos de regras utilizados na representação simbólica. Os modelos criados para cada algoritmo foram avaliados com a medida de desempenho acurácia. Este modelo foi utilizado para classificar novas sentenças na Fase 3.

Na Figura 23 é mostrado um exemplo de sentenças que foram classificadas nas respectivas classes: efeito negativo, efeito positivo e outros. Na Seção 6.3, o objetivo é extrair as informações que estão presentes nas sentenças da classe “efeito negativo”. As sentenças que foram classificadas em “efeito positivo” e em “outros” são descartadas. No exemplo da Figura 23, o termo relevante a ser identificado na sentença de “efeito negativo” é “*sepsis*”.

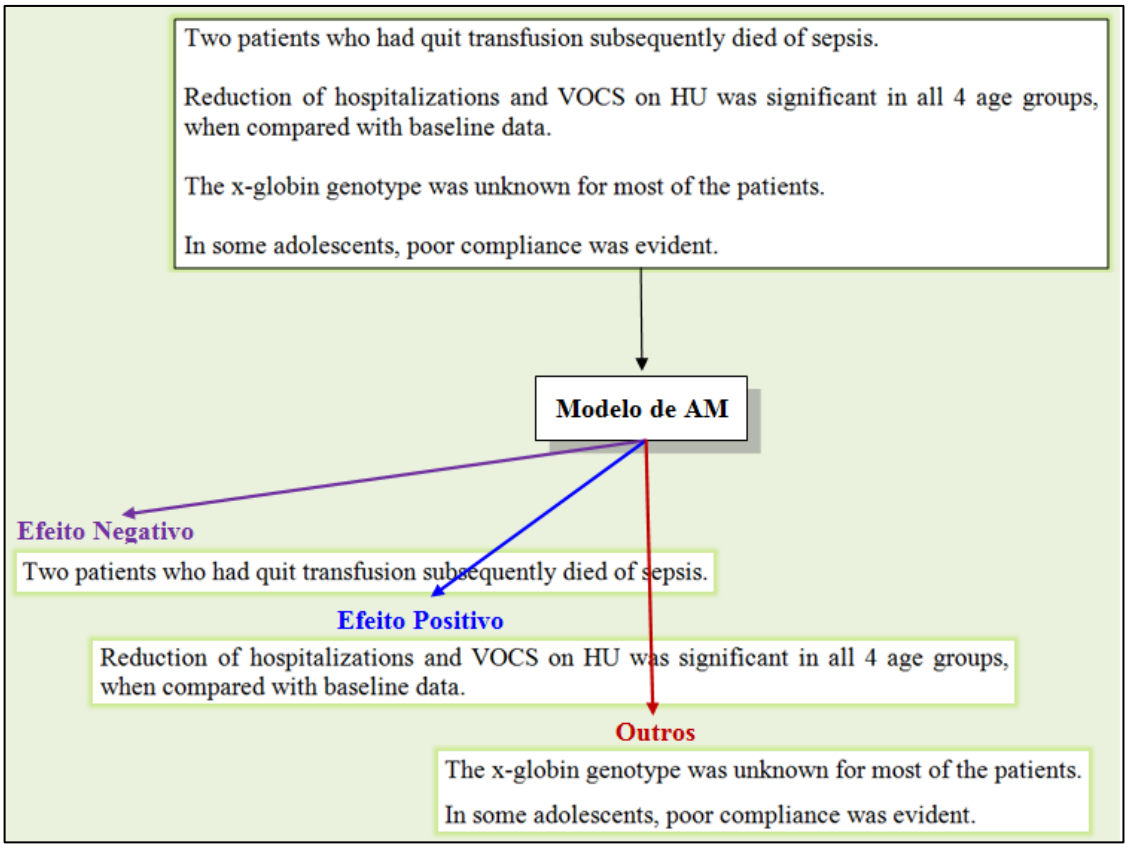

Figura 23 – Exemplo de sentenças da doença Anemia Falciforme e as suas respectivas classificações.

6.3 Identificação de Termos Relevantes

Nesta etapa é necessário identificar os termos relevantes em cada uma das sentenças de interesse (i.e., especificamente na classe efeito negativo). Para isso duas abordagens são utilizadas: dicionário e regras. O dicionário tem a função de identificar os termos curados armazenados no dicionário nas sentenças de interesse, a fim de preencher o tipo-relacionamento artigo/termo. O objetivo da regra é extrair automaticamente novos termos das sentenças de interesse e armazená-los no dicionário. Os termos já existentes no dicionário não são armazenados novamente. A inserção de termos no dicionário, somente é realizada com termos inexistentes no dicionário. É importante ressaltar que o dicionário não tem a funcionalidade de extrair novos termos.

A seguir são apresentados exemplos destas duas abordagens no domínio da AF.

6.3.1 Abordagem de Extração de Informação baseada em Dicionário

Como foi dito anteriormente, o dicionário tem a função de identificar os termos curados nas sentenças sobre efeitos negativos e, por conseguinte, preencher o tipo-relacionamento artigo/termo. O dicionário terminológico é composto pelas tabelas presentes no esquema lógico derivado do esquema conceitual parcialmente representado na Figura 24 e

pelas tabelas auxiliares Lista de Exclusão de Palavra (LEP) e Lista de Exclusão de Termo (LET).

Na Figura 24 é ilustrado parte do esquema conceitual do banco de dados desenvolvido neste mestrado, sendo que alguns atributos foram omitidos por questão de simplificação do esquema. Existem cinco tipos entidade (*Paper*, *Complication from Disease*, *Side Effect from Treatment*, *Complication Variation* e *Side Effect Variation*), sendo as duas últimas tipos entidade fraca. O esquema conceitual Entidade-Relacionamento Estendido (EER) completo do banco de dados pode ser visto no APÊNDICE A – ESQUEMA CONCEITUAL EER e o esquema lógico relacional mapeado a partir deste esquema conceitual pode ser encontrado no APÊNDICE B – ESQUEMA LÓGICO RELACIONAL.

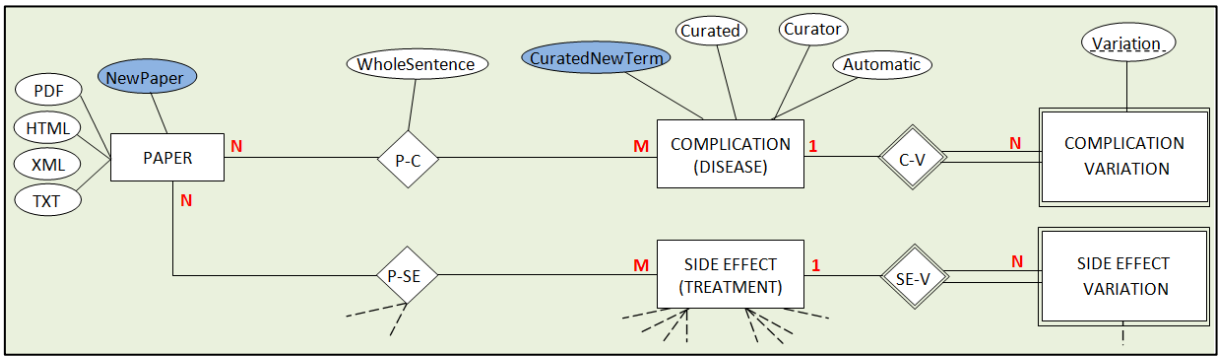

Figura 24 – Esquema conceitual da Anemia Falciforme.

O tipo entidade *Paper* contém as informações do artigo, por exemplo, nome da revista, título, autor e artigos nos formatos PDF, HTML, XML e TXT. Os tipos entidade *Complication* e *Side Effect* armazenam informação sobre os termos, respectivamente, relacionados a efeito negativo da doença e efeito negativo do tratamento. Alguns exemplos de atributos desses tipos entidade são: nome do termo, acrônimo, se o termo foi curado, qual o nome do curador e se o termo foi inserido no dicionário por um processo automático ou manual. Os tipos entidade fraca *Complication Variation* e *Side Effect Variation* armazenam as variações dos nomes de cada termo. Um termo pode ser escrito de várias formas, isto é, pode ter variações. Assim, os tipos entidade fraca *Variation* armazenam essas variações dos nomes de cada termo.

O banco de dados contém termos que são curados e não curados. Somente os termos curados e suas variações são utilizados para identificar se o termo está presente na sentença (i.e., funcionalidade do dicionário). Na Tabela 10 é apresentado um exemplo de termos curados e as suas variações. O nome do termo deve ser o nome mais genérico (e.g., *hemorrhage*) e os nomes das variações são os nomes mais específicos (e.g., *central nervous system hemorrhage* e *intracranial hemorrhage*).

Tabela 10 – Exemplos de termos e suas variações.

Termos	Variações
acute chest syndrome	*acute respiratory distress syndrome* *acute lung injury* *ali* *ards* *chest pain* *pulmonary insufficiency* *tachypnea*
cerebral vascular accident	*cerebrovascular event* *neurologic complication* *neurologic problem* *seizure* *stroke*
hemorrhage	*central nervous system hemorrhage* *intracranial hemorrhage*
pain	*painful episode* *pain crises* *pain crisis*

Na Tabela 38 e na Tabela 39 do APÊNDICE C – EFEITOS NEGATIVOS CURADOS ilustram, respectivamente, os efeitos negativos da doença (complicação) e os efeitos negativos do tratamento (efeito colateral) que foram previamente cadastrados no dicionário pelo especialista da doença AF.

Para evitar a identificação de um termo que já foi identificado previamente em um artigo, é necessário o uso de dois atributos para controlar o processamento: $CuratedNewTerm$ e $NewPaper$ (atributos destacados na cor azul na Figura 24). O Algoritmo 1 controla a identificação de termos em novos artigos. O Algoritmo 2 gerencia a identificação de novos termos em todos os artigos. Ambos os algoritmos foram explicados na Seção 5.3.1.

A LEP contém palavras comuns e gerais irrelevantes que não são relacionadas à doença AF (e.g., *other*, *different*, *underlying*) e palavras irrelevantes relacionadas à AF que estão associadas a algum termo relevante (e.g., em negrito: *painful* ***episodes***, ***recurrent*** *splenic sequestration*, ***primary*** *stroke*, ***multiple*** *vasoocclusive*). A LET contém termos substantivos simples (e.g., *dose*, *period*, *cohort*, *criteria*), substantivos compostos (e.g., *sickle cell disease*, *sickle cell anemia*) e siglas (e.g., *hb*, *scd*) relacionados à doença AF que são considerados como termos irrelevantes. A LEP auxiliará na exclusão de palavras irrelevantes que fazem parte do termo. A LET auxiliará a excluir um termo identificado erroneamente (i.e., falso positivo).

Na Tabela 11 e na Tabela 12 são mostrados exemplos de termos identificados e o respectivo termo relevante extraído após o uso das listas LEP e LET. Note que o termo relevante extraído somente será armazenado no dicionário se este mesmo termo não existir no dicionário. As tabelas LEP e LET são úteis para auxiliar na extração de termos que será explicada na seção a seguir (Seção 6.3.2).

Tabela 11 – Exemplo de remoção de palavra da tabela LEP.

Termo Identificado	LEP	Termo Relevante Extraído
secondary stroke	*secondary*	*stroke*
multiple vasoocclusive crises	*multiple*	*vasoocclusive crises*
recurrent splenic sequestration episodes	*recurrent, episodes*	*splenic sequestration*

Tabela 12 – Exemplo de remoção de termo que contém uma palavra da tabela LET.

Termo Identificado	LET	Termo Relevante Extraído
hydroxyurea administration	*hydroxyurea*	---
dose titration	*dose*	---
blood counts	*blood*	---

Na Figura 25 são apresentados exemplos de sentenças com os termos relevantes sublinhados. O dicionário auxiliará na identificação desses termos, a fim de preencher o tipo-relacionamento entre o termo e o artigo. Um exemplo deste tipo-relacionamento é a relação P-C (*Paper-Complication*) mostrado na Figura 24.

Anoxic brain injury occurred in three patients, central nervous system hemorrhage in three, and infarction in three.

Hypersplenism, which was diagnosed in six patients, is a classical complication of SCD, but recurrent splenic sequestration episodes are unusual in children aged 6 and 7 years, Hb SC children excepted, and suggest hydroxyurea-induced hypersplenism.

During 426 patient-years of follow-up for patients with standard criteria, 3.3 acute chest syndromes, 1.3 cerebrovascular events, and 1.1 osteonecrosis per 100 patient-years were observed.

Among the specific causes were pulmonary fat embolism and 27 different infectious pathogens.

Figura 25 – Exemplo de sentenças com termos relevantes sublinhados.

6.3.2 Abordagem de Extração de Informação baseada em Regras

A abordagem baseada em regras é utilizada para extrair automaticamente termos relevantes, por meio de padrões encontrados nas sentenças de interesse (i.e., sentenças sobre efeito negativo da doença AF). O etiquetador *Part-Of-Speech* (POS) é utilizado para

classificar as palavras em suas respectivas classes gramaticais (i.e., classificar nas classes substantivo, adjetivo, verbo, dentre outras).

Na Tabela 13 é apresentado um exemplo de uma sentença etiquetada. O padrão das etiquetas utilizado foi o padrão Penn Treebank (MARCUS; MARCINKIEWICZ; SANTORINI, 1993).

Tabela 13 – Exemplo de sentença etiquetada.

Sentença	Six patients with persistently abnormal TCD results developed stroke.
Sentença Etiquetada	Six_CD patients_NNS with_IN persistently_RB abnormal_JJ TCD_NNP results_NNS developed_VBD stroke_NN ._.

Inicialmente, foram analisadas manualmente algumas sentenças de artigos científicos sobre efeitos negativos da doença AF, com o intuito de formar o conjunto de regras a ser usado no processo de extração de informação. Com isto, foi possível identificar padrões para serem usados na formação das regras. Esses padrões foram utilizados em duas estratégias complementares para extrair informação das sentenças: Verbo e Expressão com POS (Estratégia 1) e somente POS (Estratégia 2). Estas estratégias foram explicadas na Seção 5.3.2. A seguir serão apresentados os padrões POS criados para cada uma dessas estratégias na classe de interesse efeitos negativos.

Estratégia 1: Uso de verbo e expressão com POS para extração de termos relevantes

Os padrões POS criados e utilizados na Estratégia 1 podem ser vistos na Tabela 14. As etiquetas com o símbolo til (~) significa negação e com o símbolo de interrogação (?) significa optativo (i.e., a etiqueta pode estar presente ou não). Considere o padrão 1.3 a título de exemplo: a expressão regular casará com este padrão se o termo for um adjetivo (JJ) seguido de um substantivo (NN) e que não comece com um adjetivo (~JJ) e nem termine com um substantivo (~NN).

Tabela 14 – Padrão POS da Estratégia 1.

Número	Padrão
1.0[1]	**(JJ_JJ_NN_NN_(NN)?)**
1.1[1]	(~JJ)_**(JJ_NN_NN_(NN)?)**
1.2[1]	**(JJ_JJ_NN)**_(~NN)
1.3	(~JJ)_**(JJ_NN)**_(~NN)
1.4	((~NN)&(~JJ))_**(NN_NN)**_((~NN)&(~JJ))
1.5	((~NN)&(~JJ))_**(NN)**_(~NN)
[1] Padrão também utilizado na Estratégia 2.	

Além das etiquetas JJ (adjetivo) e NN (substantivo), as etiquetas JJR (adjetivo comparativo), JJS (adjetivo superlativo), NNP (nome próprio) e NNS (substantivo no plural) foram utilizadas para formar as regras. Estas etiquetas não foram acrescentadas na Tabela 14 por questão de simplificação do padrão.

Esta estratégia consiste em aplicar os padrões POS da Tabela 14 em todas as sentenças que tiverem um verbo representativo ou uma expressão composta representativa. Entende-se por representativo, uma informação que pode caracterizar um termo importante na sentença. O verbo representativo encontra-se no passado, na voz passiva ou na voz ativa. O termo pode estar antes ou depois do verbo. Na Tabela 15 são apresentados os verbos representativos.

Tabela 15 – Verbos representativos.

Verbos no Passado	Voz (ativa ou passiva)	Termo (antes ou depois do verbo)
to document, to diagnose, to observe	passiva	antes
to observe of	passiva	depois
to develop, to occur	ativa	antes
to develop, to have	ativa	depois

Na Figura 26 é apresentado um exemplo de extração de termos de uma sentença com um verbo representativo "*to observe*" na voz passiva. No passo 1 da Figura 26, pode-se observar que os termos relevantes estão sublinhados e todos os termos encontram-se antes do verbo representativo que está destacado na cor amarela.

No passo 2 da Figura 26, a "Parte Específica" da sentença, destacada na cor cinza, é selecionada por meio da indicação do verbo. Primeiramente, é aplicado dois padrões POS na parte específica selecionada para eliminar falsos positivos: padrão 1 $[A-Za-z]\{1,3\}/[A-Za-z]\{1,3\}\_NN[PS]?$ e padrão 2 $(JJ)?\_NN\_(of\_IN)$. Nesta sentença, o padrão 2 é utilizado para remover o substantivo "*patient-years*" seguido da preposição "*of*". O falso positivo indicado pela cor vermelha na Figura 26 é removido e o substantivo "*patient-years*", caso não esteja presente na Lista de Exclusão de Palavra, é inserido automaticamente nesta lista com o intuito de remover futuros falsos positivos.

Em seguida, no passo 3 da Figura 26, os padrões POS da Tabela 14 são aplicados na parte específica a fim de identificar termos candidatos. Os termos candidatos podem ser vistos na figura na cor turquesa. Para remover os falsos positivos destacados na cor vermelho, a Lista de Exclusão de Palavra (LEP) e a Lista de Exclusão de Termo (LET) são consultadas. Com o auxílio da tabela LEP, os substantivos "*patients*" e "*patient-year*s" são eliminados. Perceba que este último substantivo foi inserido na tabela LEP automaticamente pelo padrão

POS aplicado no passo 2. Com o auxílio da tabela LET, o termo falso positivo "*standard criteria*" também é removido. Esta remoção somente foi possível porque a palavra "*criteria*", pertencente à tabela LET, indica que o termo candidato não é um termo relevante. Na Tabela 16 pode ser visto todos os termos candidatos identificados nesse exemplo e qual o padrão POS que identificou o termo candidato.

Por fim, após a remoção dos falsos positivos nos passos 2 e 3, os seguintes termos relevantes foram extraídos como mostrado no passo 4 da Figura 26: "*acute chest syndromes*", "*cerebrovascular events*" e "*osteonecrosis*".

Tabela 16 – Termos candidatos identificados na sentença da Figura 26.

Termos Candidatos	Padrão POS da Tabela 14
patients	1.5
standard criteria	1.3
acute chest syndromes	1.1
cerebrovascular events	1.3
osteonecrosis	1.5
patient-years	1.5

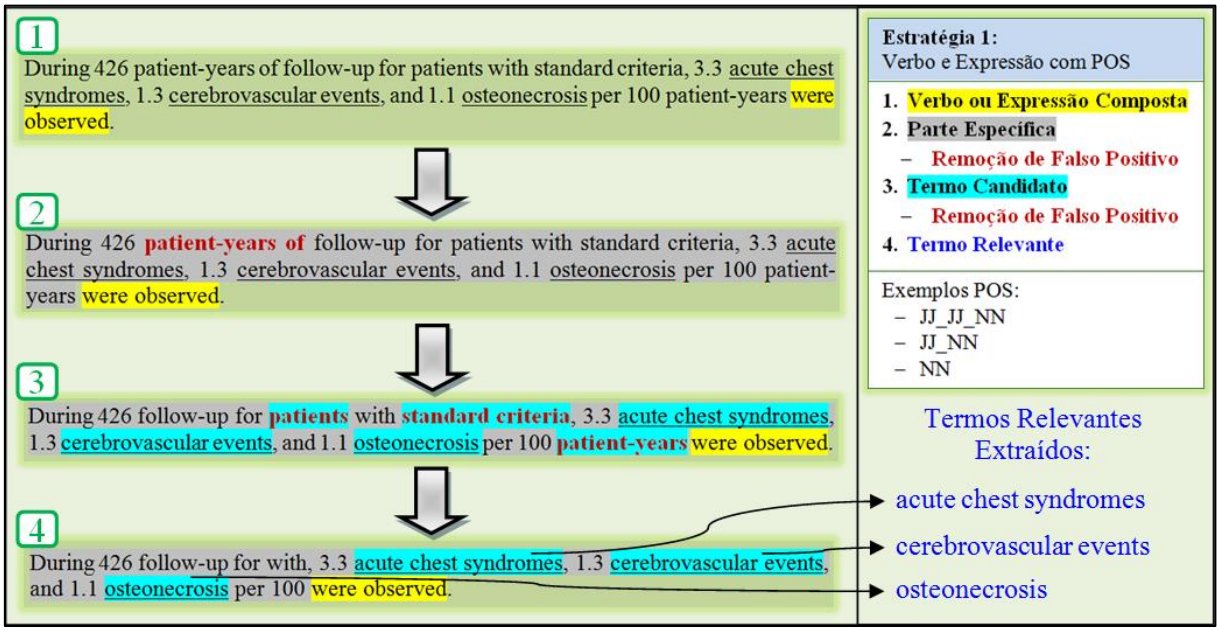

Figura 26 – Exemplo de termos extraídos de uma sentença por meio do verbo "*to observe"* na voz passiva.

Além dos verbos representativos, também é possível identificar um termo por meio de expressões compostas representativas. Na Tabela 17 são apresentadas essas expressões compostas. O termo pode estar antes ou depois da expressão. Na Figura 27 é apresentado um exemplo de extração de termos de uma sentença com a expressão composta "*caused by*".

Tabela 17 – Expressões compostas representativas.

Expressões Compostas	Termo (antes ou depois da expressão)
prevalence of, diagnosis of, at risk of, died of, died from, was associated with	depois
caused by	antes e depois

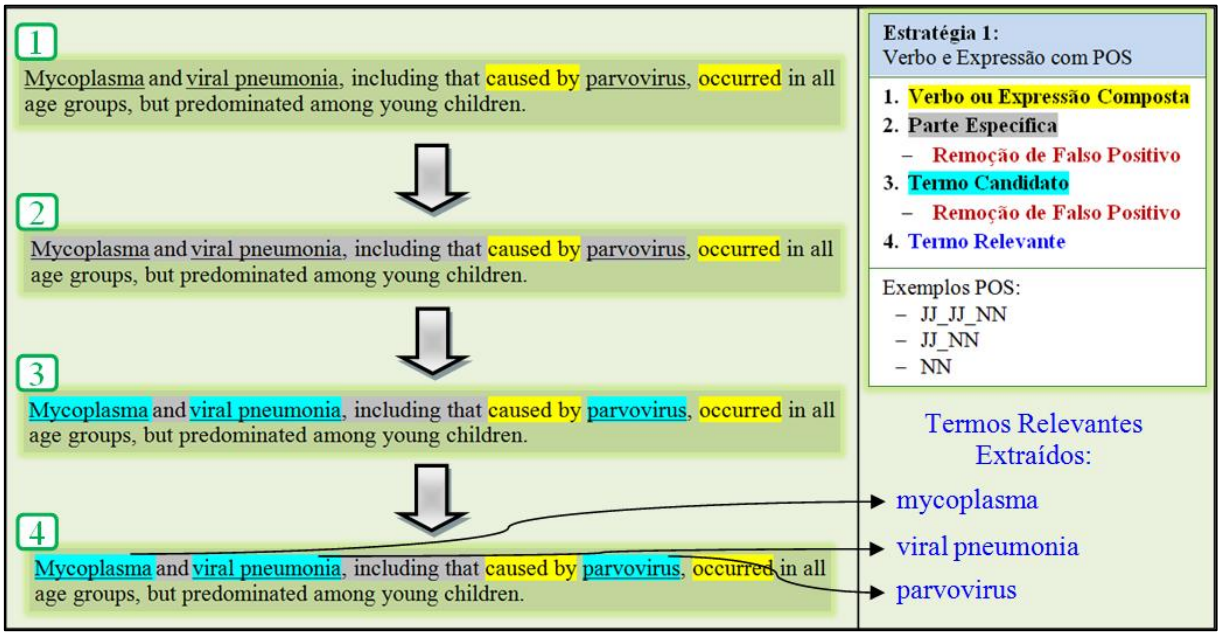

Figura 27 – Exemplo de termos extraídos de uma sentença por meio da expressão composta "*caused by*".

No passo 1 da Figura 27, pode-se observar que os termos relevantes estão sublinhados e dois deles encontram-se antes da expressão composta e um termo encontra-se depois da expressão composta. Nota-se que nesta sentença existe tanto uma expressão composta representativa quanto um verbo representativo. Ambos estão destacados na cor amarela. A identificação de termos relevantes por um verbo representativo foi explicada no exemplo anterior e não será explicada aqui novamente. Contudo, é importante ressaltar que os três termos identificados pela expressão composta "*caused by*" também são identificados pelo verbo "*occur*". No passo 2 da Figura 27, a "Parte Específica" da sentença, destacada na cor cinza, é selecionada por meio da indicação da expressão composta. Nenhum falso positivo foi identificado nesta sentença. Em seguida, no passo 3 da Figura 27 os padrões POS da Tabela 14 são aplicados na parte específica a fim de identificar termos candidatos. Os termos candidatos podem ser vistos na figura na cor turquesa. Nenhum falso positivo também foi identificado. Por fim, três termos relevantes foram extraídos por meio da expressão composta "*caused by*" e dos padrões POS, conforme mostrado no passo 4 da Figura 27. Os termos

"*mycoplasma*" e "*parvovirus*" foram identificados pelo padrão 1.5 da Tabela 14; e o termo "*viral pneumonia*" foi identificado pelo padrão 1.3.

Além da possibilidade de extrair termos utilizando separadamente verbo e expressão composta, também é possível identificar termos utilizando a combinação de verbo e expressão. Na Tabela 18 são apresentados quais são esses verbos representativos e qual a expressão composta representativa utilizada conjuntamente. A expressão regular somente casará com a sentença se, e somente se, existir um destes quatros verbos e se existir a expressão composta "*because of*" na sentença. O termo está depois da expressão composta.

Tabela 18 – Verbos com expressão composta.

Verbos	Voz (ativa ou passiva)	Expressão Composta	Termo (antes ou depois da expressão)
to stop, to interrupt, to initiate	passiva	*because of*	depois
to undergo	ativa	*because of*	depois

Na Figura 28 é apresentado um exemplo de extração de termos de uma sentença com o verbo representativo "*to stop*" na voz passiva, juntamente com a expressão composta "*because of*".

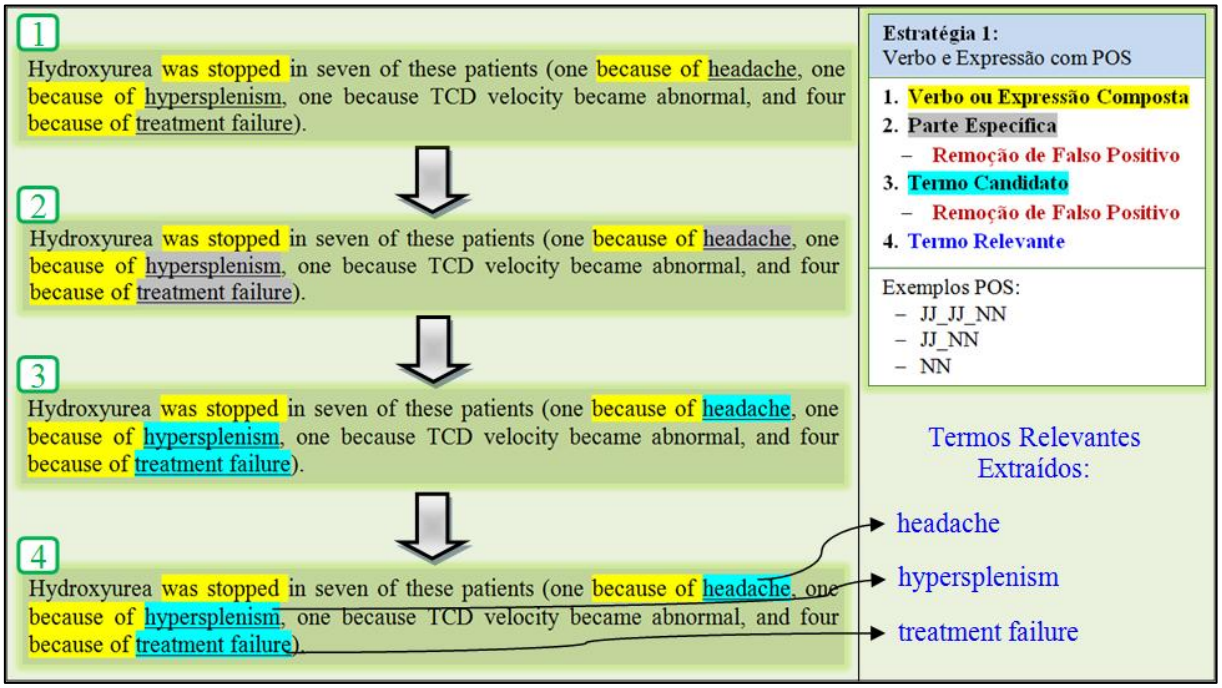

Figura 28 – Exemplo de termos extraídos de uma sentença por meio da ocorrência conjunta do verbo "*to stop*" e da expressão composta "*because of*".

No passo 1 da Figura 28, pode-se observar que os termos relevantes estão sublinhados e todos os termos encontram-se depois da expressão composta "*because of*" que está

destacada na cor amarela. No passo 2 da Figura 28, a "Parte Específica" da sentença, destacada na cor cinza, é selecionada por meio da indicação da expressão composta. Nenhum falso positivo foi identificado nesta sentença. Em seguida, no passo 3 da Figura 28 os padrões POS da Tabela 14 são aplicados na parte específica a fim de identificar termos candidatos. Os termos candidatos podem ser vistos na figura na cor turquesa. Nenhum falso positivo também foi identificado. Por fim, no passo 4 da Figura 28, três termos relevantes foram extraídos. Os termos "*headache*" e "*hypersplenism*" foram identificados pelo padrão 1.5 da Tabela 14; e o termo "*treatment failure*" foi identificado pelo padrão 1.4.

Todas as expressões regulares desenvolvidas para a Estratégia 1 podem ser encontradas em linguagem de alto nível e na linguagem de programação Java no APÊNDICE D – EXPRESSÕES REGULARES DA ESTRATÉGIA 1.

Estratégia 2: Uso de POS para a extração de termos relevantes

A Estratégia 1 utiliza verbos e expressões compostas representativos para identificar se uma sentença pode conter um termo relevante. A partir desta indicação, padrões POS são utilizados para extrair o termo relevante somente em uma parte específica da sentença. Esta parte específica pode ser antes ou depois do verbo ou da expressão. A vantagem de extrair informação em somente uma parte específica da sentença é que com isso diminui-se a possibilidade de se extrair falsos positivos. Contudo, alguns termos relevantes deixam de ser extraídos, seja porque parte da sentença não foi selecionada devido à restrição de seleção de uma parte específica da sentença ou porque a sentença não contém um verbo ou expressão representativos.

Desse modo, há a necessidade de utilizar padrões POS mais específicos aplicados em toda a sentença, a fim de identificar termos relevantes que a Estratégia 1 não conseguiu extrair. Esta segunda estratégia somente utiliza padrões POS para extrair um termo, diferentemente da primeira que além do POS também utiliza de verbos e expressões representativos. A partir da análise manual de algumas sentenças de artigos científicos sobre efeitos negativos da doença AF, os padrões POS apresentados na Tabela 19 foram criados e utilizados nesta estratégia.

Além das etiquetas JJ (adjetivo) e NN (substantivo), as etiquetas JJR (adjetivo comparativo), JJS (adjetivo superlativo) e NNS (substantivo no plural) foram utilizadas. Estas etiquetas não foram acrescentadas na Tabela 19 por questão de simplificação do padrão.

Tabela 19 – Padrão POS da Estratégia 2.

Número	Padrão
1.0[1]	**(JJ_JJ_NN_NN_(NN)?)**
1.1[1]	(~JJ)_**(JJ_NN_NN_(NN)?)**
1.2[1]	**(JJ_JJ_NN)**_(~NN)
2.0	(~JJ)_**(JJ_NN_IN_JJ_NN)**_(~NN)
2.1	**(**(~JJ)_**NN_IN)_(JJ_NN)**_(~NN)
3.0	(~JJ)_**(JJ_NN)_(IN_NN_NN_NN)**
3.1	(~JJ)_**(JJ_NN_IN_NN_NN)**_(~NN)
3.2	**(**(~JJ)_**JJ_NN_IN)_(NN)**_(~NN)
[1] Padrão também utilizado na Estratégia 1.	

Os três primeiros padrões também foram utilizados na Estratégia 1. Os outros padrões da Estratégia 1 não foram utilizados pela Estratégia 2, porque na Estratégia 2 os padrões POS são aplicados em todas as sentenças, o que pode extrair muitos falsos positivos. Na Estratégia 1 não acontece este problema, porque somente uma parte específica da sentença delimitada por um verbo ou expressão é processada.

Na Figura 29 é possível perceber que todos os termos relevantes (i.e., sublinhados) e que estão destacados na cor turquesa, não são extraídos pela Estratégia 1. Para a identificação desses termos, é necessário aplicar os passos explicados na Figura 17 da Seção 5.3.2.

Na Tabela 20 são mostrados os termos candidatos identificados, qual o padrão POS que identificou o termo e qual o termo após ter sido removido as palavras presentes na tabela LEP (palavras em vermelho na Figura 29, como *recurrent* e *episodes*).

When the acute chest syndrome was diagnosed, patients had hypoxia, decreasing hemoglobin values, and progressive multilobar pneumonia.

Anoxic brain injury occurred in three patients, central nervous system hemorrhage in three, and infarction in three.

Hypersplenism, which was diagnosed in six patients, is a classical complication of SCD, but recurrent splenic sequestration episodes are unusual in children aged 6 and 7 years, Hb SC children excepted, and suggest hydroxyurea-induced hypersplenism.

Among the specific causes were pulmonary fat embolism and 27 different infectious pathogens.

All these observations converge to suggest that hydroxyurea can prevent or delay functional asplenia, increasing the risk and lengthening the at-risk period for acute splenic sequestration episodes.

Persistent TCD elevation signals ongoing stroke risk.

Figura 29 – Exemplo de sentenças cujos termos destacados na cor turquesa são selecionados pela Estratégia 2.

Tabela 20 – Exemplos de termos candidatos identificados nas sentenças da Figura 29.

Termos Candidatos	Padrão POS da Tabela 19	Termo após o uso da LEP
progressive multilobar pneumonia	1.2	*progressive multilobar pneumonia*
central nervous system hemorrhage	1.0	*central nervous system hemorrhage*
recurrent splenic sequestration episodes	1.0	*splenic sequestration*
pulmonary fat embolism	1.2	*pulmonary fat embolism*
different infectious pathogens	1.2	*infectious pathogens*
acute splenic sequestration episodes	1.0	*acute splenic sequestration*
ongoing stroke risk	1.1	*stroke risk*

As expressões regulares desenvolvidas para a Estratégia 2 (POS) podem ser encontradas na linguagem de programação Java no APÊNDICE E – EXPRESSÕES REGULARES DA ESTRATÉGIA 2.

6.4 Gerenciamento de Termos

Nesta etapa, o especialista pode realizar quatro operações: inserir novos termos (Operação 1); hierarquizar termos (Operação 2), isto é, definir um termo como uma variação de um outro termo; validar termos extraídos automaticamente pela abordagem de regra (Operação 3), isto é, remover falso positivo ou definir o termo como termo relevante (i.e., termo curado); e mover um termo extraído de um tipo de categoria para outra (Operação 4).

Na Operação 1, o termo somente é inserido no dicionário se o mesmo for realmente um novo termo. Na Operação 2, o objetivo é definir um termo e suas variações para evitar identificação de um termo mais de uma vez pela abordagem de dicionário. Por exemplo, suponha que existam no dicionário os termos curados "*parvovirus*" e "*parvovirus b19 infection*". Considere que o dicionário seleciona estes dois termos para identificar se os mesmos estão presentes na sentença da Figura 30. A identificação realizada pelo dicionário estará equivocada, pois os dois termos serão identificados nessa sentença. O correto é somente a identificação do termo mais específico, que neste caso é "*parvovirus b19 infection*" (termo sublinhado na Figura 30). Portanto, é necessário que os termos sejam hierarquizados de um nível geral (termo "*parvovirus*") para o mais específico (variação "*parvovirus b19 infection*"). Na Tabela 21 são apresentados exemplos corretos de termos e suas variações, a fim de evitar esse problema.

In six patients (one in the hydroxyurea group) parvovirus b19 infection developed during treatment.

Figura 30 – Exemplo de sentença que mostra a identificação erradamente de dois termos.

Tabela 21 – Exemplos corretos de termos e suas respectivas variações.

Termos	Variações
hemorrhage	*central nervous system hemorrhage* *intracranial hemorrhage*
splenic sequestration	*acute splenic sequestration* *hypersplenism* *loss of spleen* *splenomegaly*
vasoocclusive	*vasoocclusive pain episode* *vasoocclusive crises* *vasoocclusive crisis* *vascular occlusion*

Na Operação 3, o objetivo é identificar se o termo não curado é realmente um termo ou é um falso positivo. Se for um falso positivo, o termo não curado deve ser excluído. Caso contrário, é definido como termo curado. Na Operação 4, o objetivo é mover o termo e suas variações que foram inseridos na categoria de efeito negativo do tratamento (efeito colateral) para a categoria correta. Por exemplo, considere que o termo "*priapism*" foi inserido na categoria efeito colateral. Contudo, o termo é uma complicação da doença AF e não um efeito colateral originado do tratamento. Portanto, o termo "*priapism*" deve ser movido da categoria efeito colateral para a categoria complicação.

O especialista tem um papel importante no gerenciamento dos termos. Nesta instanciação, a Dra. Ana Cristina Silva Pinto da Universidade de São Paulo do Campus de Ribeirão Preto foi a especialista responsável pela realização das operações aqui descritas para os termos presentes no dicionário sobre a doença AF.

6.5 Ferramentas Desenvolvidas

A seguir são apresentadas as três ferramentas que dão suporte as três últimas etapas da metodologia proposta para a extração de informação no domínio biomédico (Figura 31), a saber: ferramenta SCA-Classifier desenvolvida para auxiliar na Classificação de Sentenças (Etapa 2), ferramenta SCA-Extractor para auxiliar a etapa da Identificação de Termos Relevantes (Etapa 3) e SCA-TermManager para dar suporte ao especialista no Gerenciamento de Termos (Etapa 4). A sigla SCA, colocada como acrônimo inicial nos nomes das ferramentas, significa doença Anemia Falciforme em inglês (i.e., *Sickle Cell Anemia*).

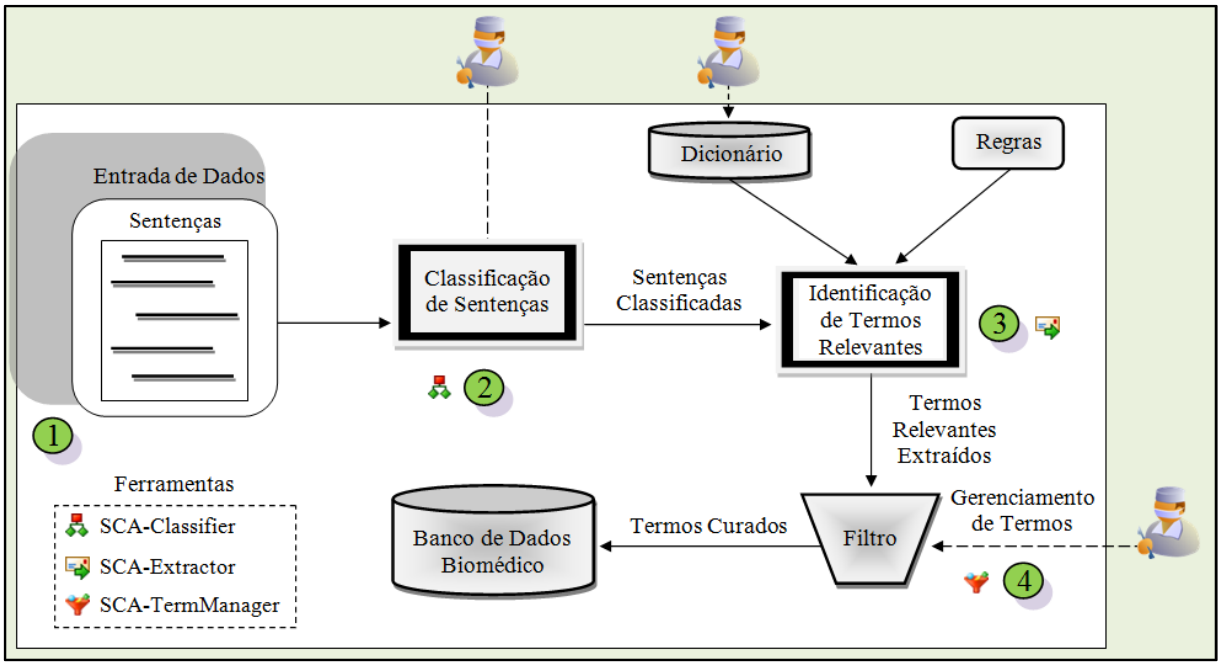

Figura 31 – Ferramentas desenvolvidas para as três últimas etapas da metodologia.

Todas as ferramentas foram desenvolvidas para o ambiente desktop, utilizando a linguagem de programação Java (SUN MICROSYSTEMS). Portanto, as mesmas podem ser executadas em diversos sistemas operacionais como Windows, Linux ou Mac OS. A versão mais atual das ferramentas pode ser baixada a partir da seguinte URL http://gbd.dc.ufscar.br/~pablofmatos/.

6.5.1 SCA-Classifier

A ferramenta SCA-Classifier tem o objetivo de ser um ambiente de aprendizado de máquina que possibilite ao usuário criar o melhor modelo de classificação que represente os dados de treinamento, de acordo com os resultados de acurácia de alguns classificadores em combinação com alguns filtros. Os algoritmos e os filtros utilizados pela ferramenta foram implementados utilizando da API (*Application Programming Interface*) do Weka (HALL et al., 2009). A ideia da criação deste ambiente é possibilitar ao usuário, que tenha um conhecimento inicial de conceitos de aprendizado de máquina, a realização de experimentos de uma forma prática e rápida.

Na Figura 32 são apresentadas as fases de treinamento e teste nesta ferramenta. Como pode ser observado, é possível selecionar quais filtros serão usados (passo 1), quais algoritmos de aprendizado de máquina serão usados (passo 2), além de outras características, como seleção de atributo, remoção de *stopword* ou uso de *stemmer* (passo 3).

Antes de iniciar o processo de classificação, é necessário selecionar se o conjunto de sentenças será treinado e testado ou se será criado o modelo de classificação (Figura 32, área

1). A diferença é que na primeira seleção (*Training & Test Classification*), o objetivo é permitir visualizar o resultado (e.g., precisão, revocação e acurácia) dos algoritmos selecionados no passo 2, sem ter a necessidade de se criar o modelo de classificação. O método de particionamento *10-Fold Cross-Validation* é utilizado para estimar a acurácia do classificador.

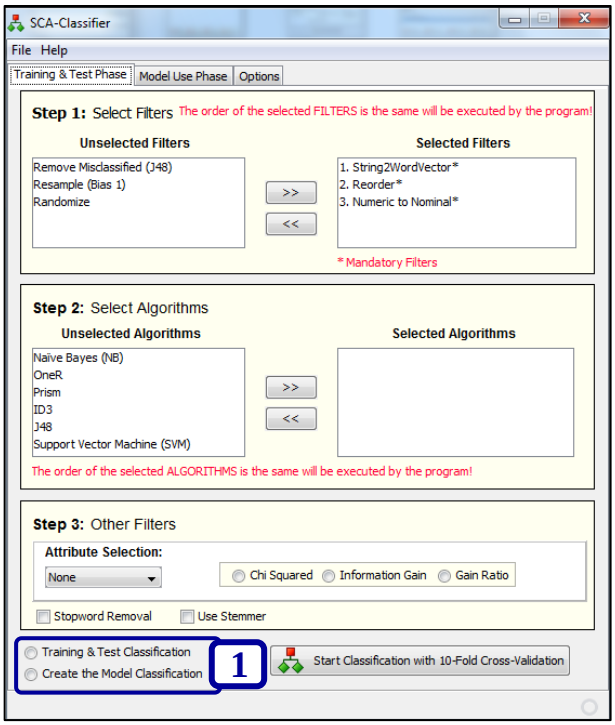

Figura 32 – Ferramenta SCA-Classifier: fases de treinamento e teste.

Após as fases de treinamento e de teste, o modelo criado pode ser utilizado para classificar novas sentenças. Esta fase é a fase do uso do modelo que pode ser observada na Figura 33. Nesta fase, o objetivo é classificar sentenças novas (i.e., sentenças que não foram treinadas) de acordo com o modelo de classificação criado. Antes de classificar as sentenças, é possível selecionar o formato de entrada dos dados (TXT ou XML) e se a classificação será realizada em todas as seções ou em seções específicas (i.e., *abstract*, *results* e *discussion*) (Figura 33, área 1).

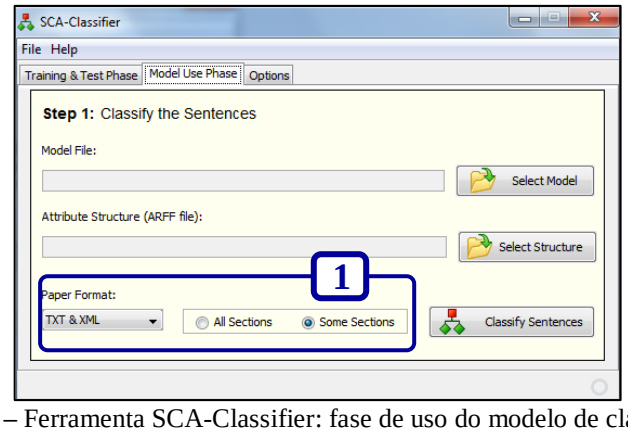

Figura 33 – Ferramenta SCA-Classifier: fase de uso do modelo de classificação.

6.5.2 SCA-Extractor

A ferramenta SCA-Extractor tem o objetivo de auxiliar o usuário médico na extração das principais informações dos artigos científicos sobre a doença Anemia Falciforme. Na Figura 34 é apresentado o módulo de extração de informação da ferramenta que consiste em dois passos: no primeiro passo, é necessário selecionar quais os artigos que vão ser processados e qual o formato do artigo (área 1 da figura). Os números dos artigos significam o identificador do artigo armazenado no banco de dados. O formato escolhido pode ser TXT ou XML. Uma das três amostras também pode ser escolhida. Cada amostra está relacionada com uma quantidade de sentenças. O número associado a cada amostra significa a quantidade de sentença daquela amostra. Por exemplo, Amostra600 possui 600 sentenças dos artigos 1, 2, 5, 6, 8 e 11. Estas amostras predefinidas são as amostras usadas em testes preliminares e depois em testes subsequentes para a validação do processo de extração.

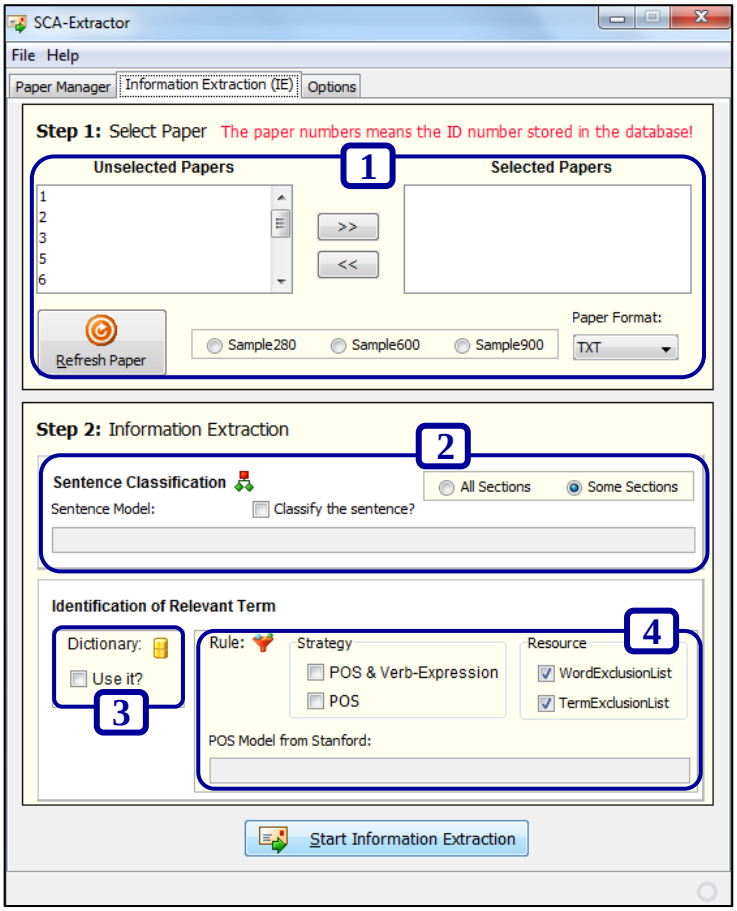

Figura 34 – Ferramenta SCA-Extractor: módulo de extração de informação.

No segundo passo, é necessário selecionar as abordagens de extração de informação que serão utilizadas: Aprendizado de Máquina – Classificação de Sentenças (área 2), Dicionário (área 3) ou Regra (área 4).

Na classificação de sentenças (área 2), precisa-se selecionar o modelo de classificação utilizado. O modelo é criado por meio da ferramenta SCA-Classifier, explicado na Seção 6.5.1. Pode-se escolher se a classificação será realizada em todas as seções ou em seções específicas (i.e., *abstract*, *results* e *discussion*). A permissão de selecionar somente estas três seções foi decidida a partir de uma análise manual nos artigos científicos, juntamente com o aval do especialista. Nesta análise percebeu que os efeitos negativo e positivo ocorrem majoritariamente nestas seções. O processamento em sentenças de algumas seções permite que menos sentenças sejam selecionadas e, por conseguinte, permite a extração de uma menor quantidade de falsos positivos. Por outro lado, informações sobre tratamento (não tratado nesta dissertação) foram identificadas em todo o artigo.

A função do dicionário (área 3) é identificar nas sentenças a presença dos termos curados e preencher o tipo-relacionamento artigo/termo. Na abordagem de dicionário somente são selecionados os termos que foram curados. Na abordagem de regra (área 4), é possível selecionar se a Lista de Exclusão de Palavra e a Lista de Exclusão de Termo serão utilizados para auxiliar na exclusão de falsos positivos. Também é possível escolher qual das estratégias de extração utilizar (Verbo e Expressão com POS, POS ou ambas).

Para utilizar uma das estratégias de extração, primeiramente é necessário selecionar qual o etiquetador *Part-Of-Speech* (POS) será utilizado. Existem vários etiquetadores POS publicados na literatura para o idioma inglês e qualquer um deles poderia ser utilizado. Na ferramenta SCA-Extractor, o etiquetador POS utilizado foi o da universidade de Stanford (THE STANFORD NATURAL LANGUAGE PROCESSING GROUP, 2010). Este etiquetador utiliza o algoritmo de aprendizado de máquina Entropia Máxima para calcular estatisticamente a probabilidade de uma palavra pertencer a uma determinada classe gramatical. Detalhes deste etiquetador podem ser encontrados em Toutanova e Manning (2000) e Toutanova et al. (2003). Para utilizar esse etiquetador é necessário escolher um dos três modelos de aprendizado disponíveis para o idioma inglês. A escolha dos modelos varia de acordo com a necessidade de desempenho. Quanto mais preciso for o modelo, mais tempo o etiquetador levará para etiquetar as palavras das sentenças. Na Tabela 22 são apresentadas as principais características desses modelos e o desempenho em relação à acurácia e ao tempo.

Os artigos, que são utilizados na extração de informação (Figura 34) devem ser previamente inseridos no banco de dados. O módulo de gerenciamento de artigo da Figura 35 foi desenvolvido para auxiliar nessa inserção. Para inserir novos artigos é necessário que os mesmos estejam enumerados para cada formato (e.g., 1.pdf, 1.html, 1.txt, 1.xml). O número

de cada artigo é único. Antes de inserir os artigos no banco de dados, é possível verificar quais são os artigos existentes.

Tabela 22 – Três modelos de etiquetagem POS da universidade de Stanford.

Modelo POS	Característica	Desempenho (Acurácia)	Desempenho (Tempo)
bidirectional-distsim-wsj-0-18.tagger	Treinado nas seções do *Wall Street Journal* (WSJ) 0-18 usando a arquitetura bidirecional e incluindo as características da palavra e de similaridade de distribuição.	– 97,28% de acertos nas seções 19-21 do WSJ – 90,46% de acertos em palavras desconhecidas	Baixo
left3words-distsim-wsj-0-18.tagger	Treinado nas seções WSJ 0-18 usando a arquitetura left3words e incluindo características da palavra e de similaridade de distribuição.	– 97,01% de acertos nas seções 19-21 do WSJ – 89,81% de acertos em palavras desconhecidas	Médio
left3words-wsj-0-18.tagger	Treinado nas seções WSJ 0-18 usando a arquitetura left3words e incluindo características da palavra.	– 96,97% de acertos nas seções 19-21 do WSJ – 88,85% de acertos em palavras desconhecidas	Alto

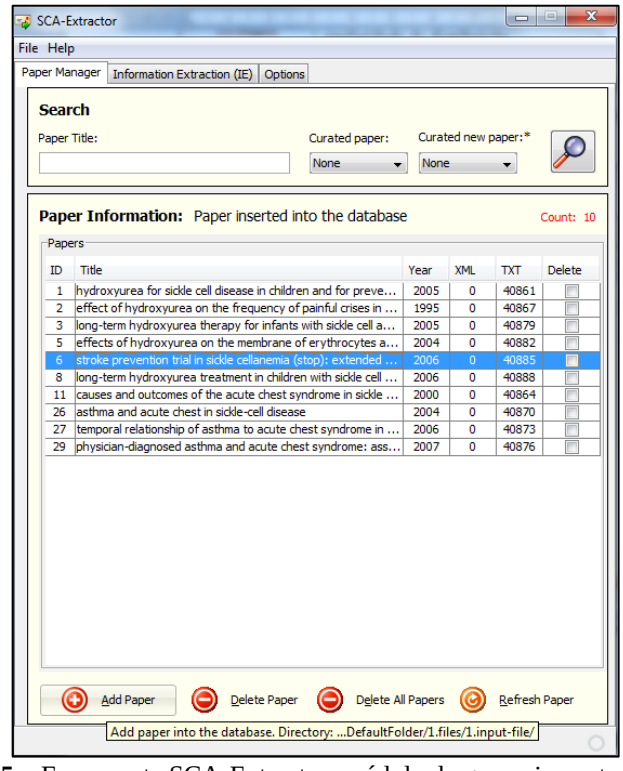

Figura 35 – Ferramenta SCA-Extractor: módulo de gerenciamento de artigo.

As principais classes da ferramenta SCA-Extractor foram modeladas na ferramenta de UML (*Unified Modeling Language*) Jude *Community* 5.4.1 (JUDE, 2010) e encontram-se no APÊNDICE F – DIAGRAMA DE CLASSES.

6.5.3 SCA-TermManager

A ferramenta SCA-TermManager foi desenvolvida para auxiliar o especialista no gerenciamento dos termos (Figura 36). O especialista pode realizar quatro operações: inserir novos termos (Operação 1); hierarquizar termos (Operação 2); validar termos extraídos automaticamente pela abordagem de regra (Operação 3); e mover um termo extraído de uma categoria para outra categoria (Operação 4).

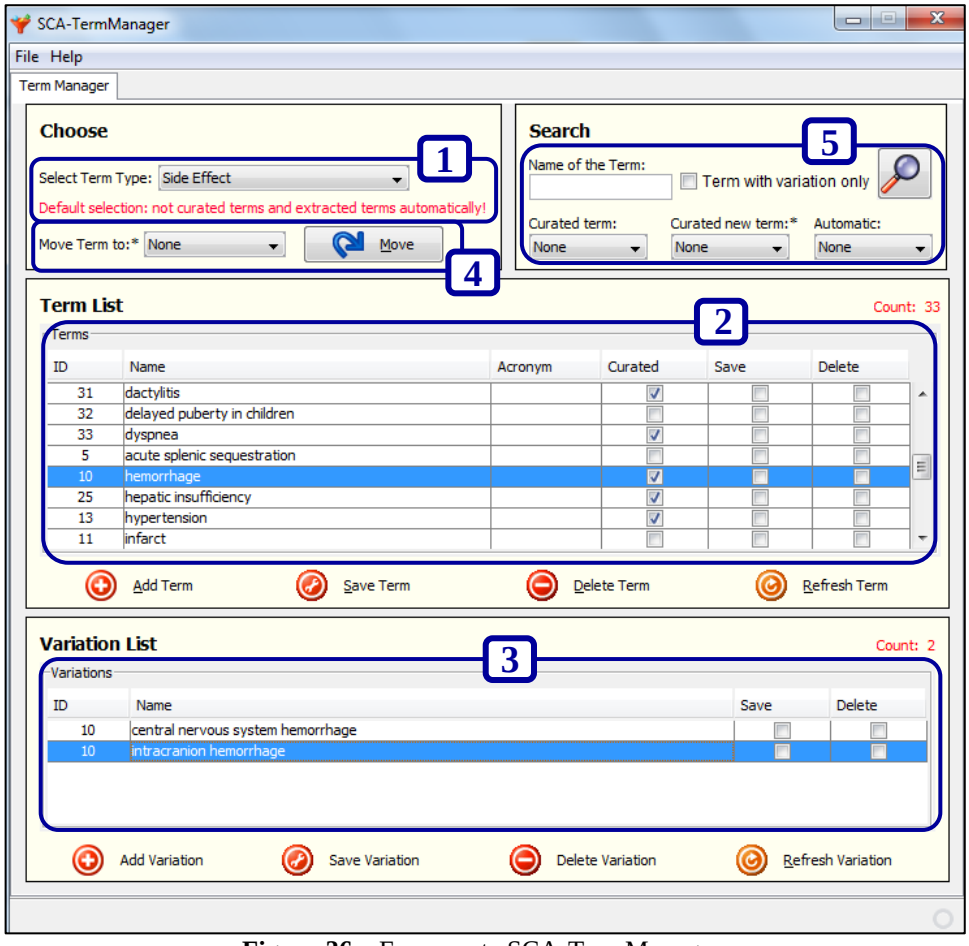

Figura 36 – Ferramenta SCA-TermManager.

A ferramenta possibilita a inserção de novos termos pelo especialista (Operação 1). Todos esses novos termos são inseridos no dicionário como termos curados. Para inserir novos termos, o especialista primeiramente deve selecionar qual o tipo/categoria do termo a ser curado (área 1 da Figura 36). Após a seleção da categoria do termo, o especialista pode inserir um novo termo (área 2). Na área 2 são mostrados os termos sobre efeito colateral (*side*

effect). Por padrão, são selecionados do banco de dados todos os termos que foram extraídos automaticamente pela abordagem de regra e que ainda não foram curados.

O especialista pode hierarquizar novos termos ou termos existentes (Operação 2). Para isto, é necessário adicionar os nomes das variações para um determinando termo (área 3). Ele também pode validar os termos extraídos (Operação 3). A validação pode ser feita de duas formas (área 2 da Figura 36): 1) removendo os falsos positivos, isto é, termos que foram extraídos, mas que não são termos; 2) ou definindo o termo extraído como termo curado.

Além dos termos poderem ser editados ou excluídos, os termos também podem ser movidos de uma categoria para outra (Operação 4). Por exemplo: suponha que o termo "*acute splenic sequestration*" não seja um termo de efeito colateral e sim, uma complicação. O especialista pode decidir mover o termo para a categoria complicação (área 4).

Além das quatros operações apresentadas anteriormente e fornecidas pela ferramenta SCA-TermManager, o especialista também pode pesquisar por alguns termos (área 5). Nesta área é possível visualizar os termos que foram curados ($curated\ term = yes$), os termos que foram curados, mas que já foram processados por algum artigo ($curated\ new\ term = no$), os termos que foram inseridos manualmente ($automatic = no$) ou ainda somente os termos com alguma variação.

6.6 Considerações Finais

Neste capítulo, a metodologia proposta para extração de informação descrita no Capítulo 5 foi instanciada para a doença Anemia Falciforme. A instanciação foi realizada em todas as etapas da metodologia: Entrada de Dados (Etapa 1), Classificação de Sentenças (Etapa 2), Identificação de Termos Relevantes (Etapa 3) e Gerenciamento de Termos (Etapa 4).

Na Etapa 1, os artigos científicos nos formatos TXT ou XML foram selecionados. A Etapa 2, por sua vez, é uma etapa muito importante, pois ela é responsável por distinguir as sentenças relevantes das irrelevantes. A Etapa 2 fornece uma confiança em relação ao uso do verbo "*to have*". O verbo "*to have*", por si só, não necessariamente indica que em uma determinada sentença tem um termo relevante. Por exemplo, considere a seguinte sentença: "*Children younger than 2 years had clinical response similar to older patients, except for days in the hospital."*. O uso do verbo "*to have*" nesta sentença informa que crianças menores de 2 anos tiveram resposta semelhante a pacientes mais velho. Esta sentença não contém nenhum termo relevante. Assim, a classificação de sentenças tem um papel fundamental no sentido de classificar esta sentença como sentença irrelevante.

A Etapa 3 tem o objetivo de extrair informação das sentenças que foram classificadas na classe de interesse. A partir de uma análise inicial em algumas sentenças sobre efeitos negativos da doença AF, foram desenvolvidas os padrões apresentados na Seção 6.3.2. Alguns problemas foram identificados a partir dessa análise: termos não foram extraídos por falta de padrão na sentença e por erro de etiquetagem em algumas palavras.

Na Figura 37 são apresentados dois exemplos de erro de etiquetagem. Os termos "*splenomegaly*" e "*parvovirus b19 infection"* não foram extraídos como um termo candidato.

Before hydroxyurea treatment, three of them had splenomegaly and two a previous history of splenic sequestration.

In six patients (one in the hydroxyurea group) parvovirus b19 infection developed during treatment.

Figura 37 – Sentenças etiquetadas erroneamente pelo etiquetador.

Na Tabela 23 pode ser observado que a etiqueta correta para o termo candidato "*splenomegaly*" é substantivo (NN) e não advérbio (RB). O padrão POS correto para identificar este termo é o padrão 1.5 da Tabela 14 (Estratégia 1). A etiqueta correta para a palavra "*parvovirus"* é adjetivo (JJ) e não verbo na terceira pessoa do singular (VBZ). O padrão POS correto para identificar este termo é o padrão 1.2 da Tabela 14 (Estratégia 1) e não o padrão 1.3. Portanto, erros causados pelo etiquetador podem atrapalhar a extração de termos pelas regras.

Tabela 23 – Termos não identificados corretamente.

Etiqueta Errada	Termo Candidato Errado	Etiqueta Correta	Termo Candidato Correto
RB	*Nenhum*	NN	*splenomegaly*
VBZ_JJ_NN	*b19 infection*	JJ_JJ_NN	*parvovirus b19 infection*

Ainda na Etapa 3, outros termos deixaram de ser identificados devido à falta de padrão e principalmente, por causa da necessidade de interpretação semântica e da dificuldade de encontrar um termo explícito de apoio nessas sentenças. Esta falta de padrão foi identificada mais especificamente em um artigo. Na Figura 38 são apresentadas algumas sentenças desse artigo.

5 On the other hand, hydroxyurea treatment reduced the percentage of reticulocytes (8.44% vs 4.46%) and of erythrocytes positive for CD36, CD71, CD49d and annexin V (Figures 1 and 2). **(Results)**

5 Hydroxyurea treatment reduced the percentage of annexin V+ erythrocytes in all patients, but one (Figure 1A). **(Results)**

5 The increase in MCV was positively correlated with the increase in F cells in the circulation and in hemoglobin concentration and negatively correlated with the percentage of reticulocytes and of cells expressing the adhesive molecules CD36, CD49d and annexin V (Figure 4). **(Discussion)**

5 This population also showed a concentration of erythrocytes expressing CD71, CD36 and CD49d and of reticulocytes (data not shown). **(Discussion)**

5 This increased adhesiveness has been attributed to the adhesion molecules CD36 and CD49d and, more recently, to PS exposure on the cell membrane. **(Discussion)**

5 PS exposure on the surface of sickle cells, in addition to affecting erythrocyte adhesion to the vascular endothelium, exacerbates anemia by enhancing phagocyte recognition and removal of these cells17 and favors the development of a thrombophilic state. **(Discussion)**

Figura 38 – Exemplos de sentenças que não foram identificadas nenhum padrão.

Por fim, na Etapa 4, o especialista pode inserir novos termos, hierarquizar termos existentes ou novos termos, validar e mover os termos extraídos pela abordagem de regra.

No próximo capítulo é apresentado o estudo de caso realizado a partir da instanciação realizada neste capítulo visando-se analisar a eficiência da metodologia proposta. Será realizada no Capítulo 7 a análise separada da etapa de classificação de sentenças usando a medida acurácia e a medida-F e da etapa de identificação de termos relevantes usando as medidas precisão, revocação e medida-F.

7 ESTUDOS DE CASO

Neste capítulo são realizados dois estudos de caso correspondentes às duas etapas da metodologia que foram discutidas no Capítulo 5, a saber: etapa de Classificação de Sentenças (Etapa 2) e etapa de Identificação de Termos Relevantes (Etapa 3). Estas etapas foram instanciadas para o domínio da doença Anemia Falciforme conforme descrito no Capítulo 6. Os experimentos realizados neste capítulo são sobre informações desta doença e tem o objetivo de avaliar a Etapa 2 e a Etapa 3 separadamente. Estas etapas contribuem para a extração de informação relevante que é a finalidade da metodologia proposta nesta dissertação.

O primeiro estudo de caso é realizado na Etapa 2 da metodologia proposta, cujo objetivo é utilizar um algoritmo de aprendizado de máquina supervisionado para classificar as sentenças nas respectivas classes. Nesta etapa, dois experimentos foram realizados: o primeiro experimento aplica-se as fases de treinamento e de teste, no qual o modelo de classificação é criado e testado com sentenças que não foram treinadas; o segundo experimento aplica-se na fase de uso do modelo que tem o objetivo de avaliar como o classificador criado no primeiro experimento se comporta na classificação de novas sentenças. Nos dois experimentos, a medida acurácia é utilizada para avaliar a classificação em relação a todas as classes e a medida-F é utilizada para avaliar a classificação de sentenças da classe de interesse. O segundo estudo de caso é na Etapa 3 da metodologia proposta, cujo objetivo é extrair os termos relevantes das sentenças classificadas na etapa anterior. As medidas precisão, revocação e medida-F são utilizadas para avaliar o percentual de termos extraídos.

Na Etapa 2 é realizada a classificação de sentenças a fim de distinguir as sentenças em três classes, a saber: efeito positivo, efeito negativo e outros (i.e., sentenças que potencialmente não possuem qualquer efeito). Em seguida, na Etapa 3, as abordagens baseadas em dicionário e em regra são utilizadas para identificar termos relevantes nas sentenças relacionadas à classe efeito negativo. O estudo de caso correspondente a Etapa 2 e a Etapa 3 são apresentados, respectivamente, na Seção 7.1 e na Seção 7.2.

7.1 Classificação de Sentenças

Este estudo de caso tem o objetivo de avaliar quais dos algoritmos de aprendizado de máquina que melhor representa as características das sentenças. Os algoritmos escolhidos foram *Support Vector Machine* (SVM) e Naïve Bayes (NB), ambos são algoritmos estatísticos; ID3, J48, Prism e OneR são algoritmos de aprendizado simbólico, os dois primeiros são algoritmos de árvore de decisão e os dois últimos são algoritmos de regras utilizados na representação simbólica.

Inicialmente, a partir da análise dos artigos científicos, juntamente com o aval do especialista do domínio, foi constatado que a maioria das sentenças de interesse estava presente em algumas seções específicas do artigo. Assim, decidiu-se selecionar somente estas sentenças, evitando a identificação de uma maior quantidade de falsos positivos. As seções escolhidas para análise são: *abstract*, *results* e *discussion*. Portanto, sentenças de outras seções como *introduction* e *methods* foram descartadas.

Em seguida foram selecionados 10 artigos científicos nos quais foram identificados 901 sentenças nas três seções citadas anteriormente. Detalhes dos 10 artigos selecionados podem ser vistos no APÊNDICE G – ENTRADA DE DADOS. Na Figura 39 (a) pode ser observada a distribuição das 901 sentenças em relação a cada seção de interesse. Dentre as 901 sentenças, 12% pertencem à seção resumo (i.e., *abstract*), 39% pertencem à seção resultados (i.e., *results*) e 49% pertencem à seção discussão (i.e., *discussion*).

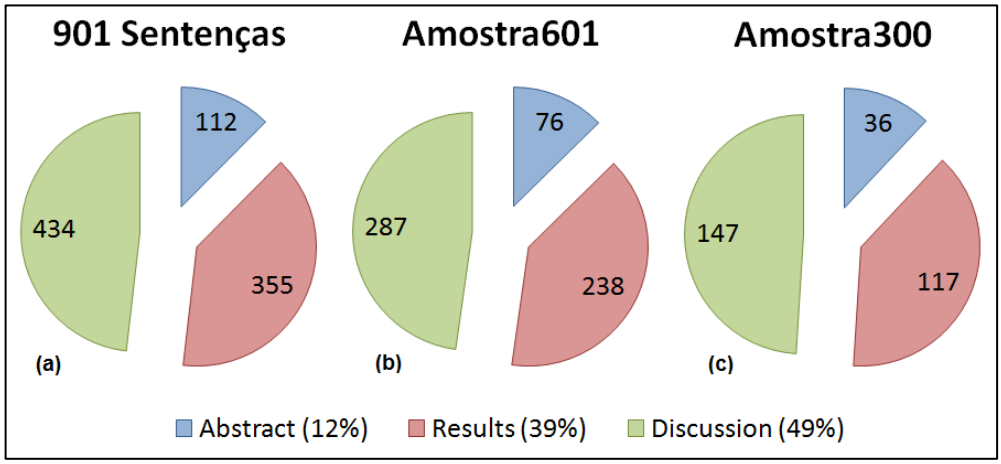

Figura 39 – Distribuição das 901 sentenças por seção de interesse (a). A distribuição original das sentenças por seção (a) foi mantida para a Amostra601 (b) e para a Amostra300 (c).

Com o objetivo de ter uma amostra para ser utilizada nas fases de treinamento e de teste e outra amostra para ser utilizada na fase de uso do modelo, foram selecionadas duas amostras. As sentenças de cada amostra foram selecionadas aleatoriamente a partir das 901 sentenças. A primeira e a segunda amostra contêm, respectivamente, 2/3 (i.e., 601 sentenças)

e 1/3 (i.e., 300 sentenças) das 901 sentenças. A proporção das sentenças por seção foi mantida para ambas as amostras, conforme ilustrado na Figura 39 (b) e (c). A divisão das amostras está de acordo com o método de particionamento *Holdout*, cujo valor de *p* utilizado é igual a 2/3 (MATOS et al., 2009a).

Na Figura 40 pode ser observada a distribuição das sentenças para as classes Efeito Positivo, Efeito Negativo e Outros em relação a cada uma das amostras. A classe Efeito Positivo contém 98 e 58 sentenças, respectivamente, 16% da Amostra601 e 19% da Amostra300. A classe Efeito Negativo contém 269 e 131 sentenças, respectivamente, 45% da Amostra601 e 44% da Amostra300. E a classe Outros, por sua vez, contém o restante das sentenças, isto é, 234 e 111 sentenças, respectivamente 39% da Amostra601 e 37% da Amostra300. Nota-se que mesmo as sentenças sendo selecionadas aleatoriamente, o percentual de sentenças por classe permaneceram equivalentes (i.e., 45% ≈ 44%, 39% ≈ 37% e 16% ≈ 19%).

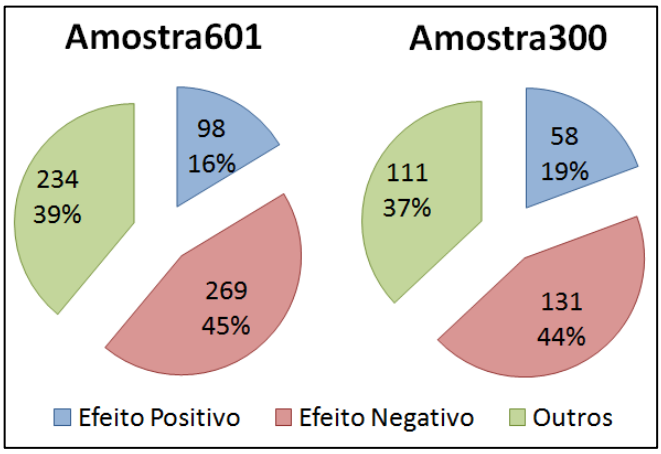

Figura 40 – Distribuição das sentenças por classe para cada uma das amostras. Estas duas amostras foram selecionadas aleatoriamente a partir das 901 sentenças.

A classificação das sentenças manualmente contou com o apoio do especialista do domínio que distinguiu quais as sentenças que continham efeito, das sentenças que não continham nenhum efeito. A seguir o primeiro e o segundo experimentos serão explicados. O primeiro utiliza a Amostra601 para treinar e testar o classificador com o método de particionamento 10-*Fold Cross-Validation* e o segundo utiliza o modelo criado a partir da Amostra601 para classificar as sentenças da Amostra300. O primeiro experimento representa a fase de treinamento e de teste da classificação e o segundo experimento representa a fase de uso do modelo.

7.1.1 Experimento 1: Fases de Treinamento e de Teste

Para a realização dos experimentos nas fases de treinamento e de teste, foi utilizada a ferramenta SCA-Classifier desenvolvida durante este mestrado e que foi explicada na Seção 6.5.1, a qual utilizou a API do Weka para implementar os algoritmos de aprendizado de máquina. Todos os parâmetros dos algoritmos utilizados foram os parâmetros padrão do Weka.

As sentenças foram estruturadas utilizando o modelo *bag-of-words*, sendo utilizados os atributos de 1 a 3 gramas para construir a matriz atributo-valor (MAV) e a medida binária para representar a presença ou a ausência do atributo na sentença. Os atributos com frequência abaixo de dois foram desconsiderados. Três diferentes testes são realizados na construção do modelo de classificação. O primeiro teste constrói a MAV utilizando a distribuição original das sentenças da Amostra601. O segundo teste constrói a matriz com a remoção de ruído nas sentenças e o terceiro teste constrói a matriz com o balanceamento das sentenças. Os testes são chamados, respectivamente, de Sem Filtro, Remoção de Ruído e Balanceamento. Estes três testes constroem as três MAV diferentes que são utilizadas por cada um dos seis algoritmos na construção do modelo de classificação, o que gera uma combinação de 18 resultados. Estes resultados foram avaliados utilizando a medida acurácia que é calculada por meio do método de particionamento *10-Fold Cross-Validation* (10-F CV).

Na Figura 41 é apresentada a distribuição das sentenças para cada classe de acordo com o tipo de teste. Nota-se que a distribuição original (i.e., sem uso de qualquer filtro) das sentenças está desbalanceada. A classe Efeito Negativo contém a maior quantidade de sentenças (45%), seguido da classe Outros (39%) e da classe Efeito Positivo (16%). O filtro Balanceamento é utilizado com o intuito de equilibrar este percentual. O método de balanceamento utilizado é o *over-sampling* (BATISTA; PRATI; MONARD, 2004), cuja função é igualar a quantidade de sentenças das classes minoritárias (i.e., as classes com menor quantidade de sentenças) com a quantidade de sentenças da classe majoritária sem aumentar a quantidade total de sentença da amostra. Na Figura 41 pode ser observado a quantidade de sentenças e o percentual para cada classe. Após o balanceamento, as classes Efeito Negativo e Outros contêm 194 sentenças (32%) e a classe Outros 213 sentenças (36%). O filtro de Remoção de Ruído elimina sentenças que estejam dificultando o aprendizado de um determinando algoritmo, ou seja, as sentenças que foram classificadas incorretamente por algum algoritmo são removidas. O algoritmo J48 foi utilizado para remover o ruído. Após a remoção de ruído, o percentual para as classes Efeito Positivo, Efeito Negativo e Outros é de, respectivamente, 13%, 46% e 41%. O percentual de redução para essas mesmas classes em

relação à distribuição original é de, respectivamente, 26% (98 – 72 = 26 sentenças), de somente 6% (269 – 252 = 17 sentenças) e de somente 5% (234 – 222 = 12 sentenças). Ao todo, houve uma redução de 9% das sentenças em relação à distribuição original (601 – 546 = 55 sentenças). Percebe-se que na distribuição original e no filtro Remoção de Ruído, a classe Efeito Positivo é pouco representativa.

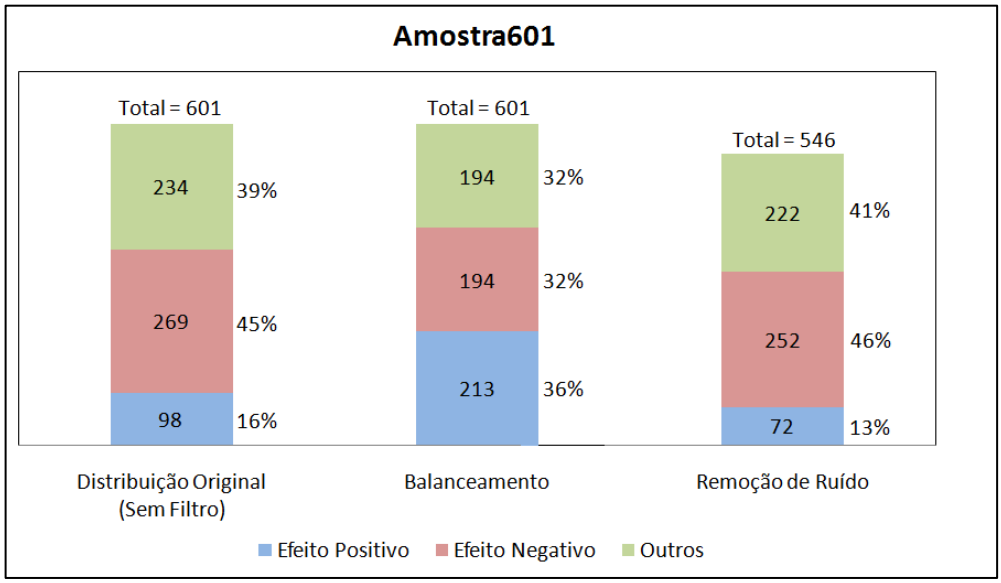

Figura 41 – Distribuição das 601 sentenças após aplicação de filtro.

Na Figura 42 é ilustrada a acurácia dos algoritmos de aprendizado de máquina Sem Filtro e com o filtro Remoção de Ruído calculada com o método de particionamento 10-F CV. Nota-se que com a utilização do filtro, houve um aumento no aprendizado em todos os algoritmos. É importante ressaltar o aumento expressivo de 19% do algoritmo J48. Este aumento é originado pelo uso do algoritmo J48 na remoção de ruído e lembrando que este aumento considerável no treinamento pode não ser percebido na classificação de novas sentenças (i.e., na fase de uso do modelo de classificação).

Na Figura 43 é ilustrada a acurácia, também calculada com o método de particionamento 10-F CV, dos algoritmos que utilizaram os filtros Remoção de Ruído e Balanceamento. Como pode ser observado, houve um aumento da acurácia com a utilização do filtro Balanceamento na maioria dos algoritmos, exceto o algoritmo OneR que teve uma redução de 4,69% e o algoritmo J48 que sofreu uma pequena redução de 0,61%.

Dos seis algoritmos testados, o OneR e o Prism foram os que obtiveram os piores resultados, tanto Sem Filtro quanto com o filtro Remoção de Ruído (resultado da Figura 42). Com a aplicação do filtro Balanceamento, o OneR obteve a pior acurácia (39,27%, Figura 43). Ainda na Figura 43 nota-se que o algoritmo Prism teve um aumento considerável de 14,72% em comparação com o resultado obtido do filtro Balanceamento (acurácia de 73,88%)

e com o filtro Remoção de Ruído (acurácia de 59,16%). Apesar deste aumento, o Prism não foi capaz de classificar todas as sentenças. O percentual de sentenças não classificadas para o filtro Balanceamento, para o filtro Remoção de Ruído e para Sem Filtro, foi respectivamente de, 17,63%, 28,57% e 29,28%. Este percentual significa, respectivamente, que 106 de 601 sentenças, que 156 de 546 sentenças e 176 de 601 sentenças não foram classificadas em nenhuma das três classes.

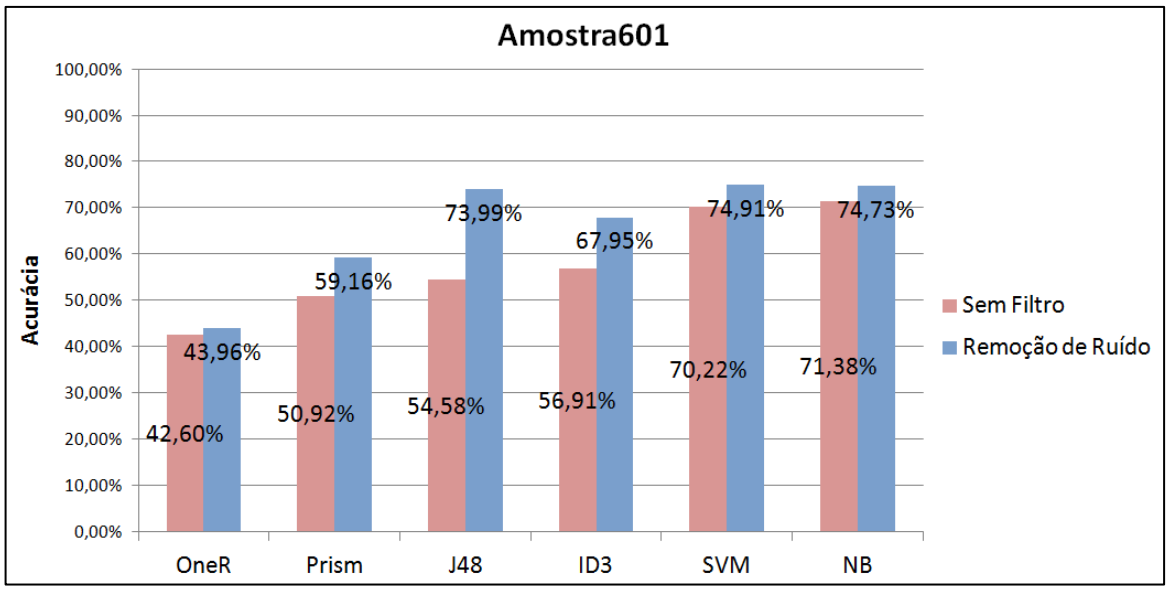

Figura 42 – Acurácia com 10-F CV dos algoritmos de aprendizado na Amostra601: Sem Filtro versus Remoção de Ruído.

A partir deste cenário, os algoritmos OneR e Prism não serão considerados para serem utilizados nos experimentos subsequentes. A acurácia dos outros quatros algoritmos (i.e., ID3, J48, NB e SVM) varia de acordo com a utilização do filtro ou não. Com o uso do filtro Remoção de Ruído, a diferença da maior acurácia (i.e., SVM) para a menor acurácia (i.e., ID3) é de apenas 6,96% (Figura 43). Com o uso do filtro Balanceamento, a diferença da maior acurácia (i.e., SVM) para a menor (i.e., J48) é de 13,81% (Figura 43), e Sem Filtro, a diferença da maior acurácia (i.e., NB) para a menor acurácia (i.e., J48), por sua vez, é de 16,8% (Figura 42).

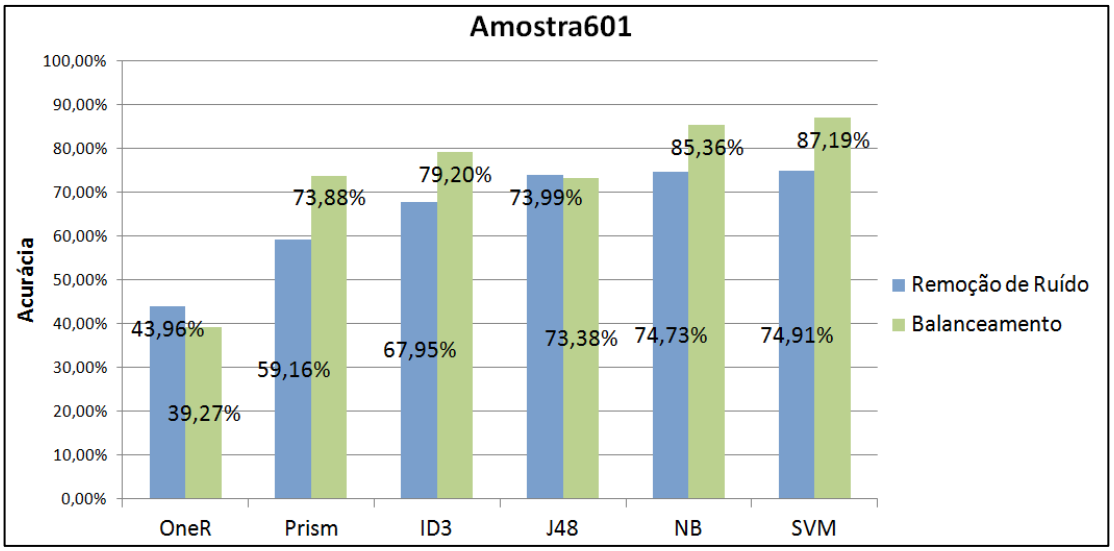

Figura 43 – Acurácia com 10-F CV dos algoritmos de aprendizado na Amostra601: Remoção de Ruído versus Balanceamento.

Para facilitar a visualização dos dados apresentados na Figura 42 e na Figura 43, a acurácia obtida pelos algoritmos J48, ID3, NB e SVM em relação à aplicação ou não de filtro é ilustrada na Figura 44. A acurácia de todos os algoritmos aumenta com a aplicação de um dos filtros. Portanto, convém utilizar um dos dois tipos de filtro para construir a matriz atributo-valor.

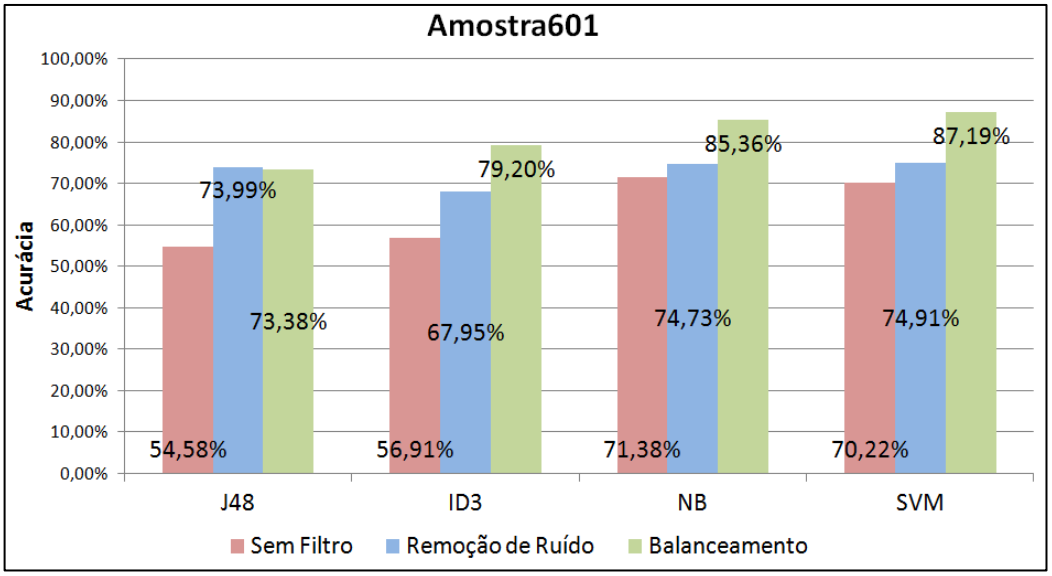

Figura 44 – Acurácia com 10-F CV na Amostra601 em relação à aplicação ou não de filtro.

Este trabalho está interessado em extrair termos relevantes sobre efeitos negativos. Nesse sentido, é interessante avaliar qual o percentual da medida-F dos quatros algoritmos em relação ao uso do filtro Remoção de Ruído e do filtro Balanceamento na classe efeito negativo. Na Figura 45 são ilustrados os percentuais obtidos. A maior medida-F dos algoritmos SVM (83,16%) e NB (82,42%) foi obtida utilizando o filtro Balanceamento. A

maior medida-F dos algoritmos ID3 (74,17%) e J48 (77,39%) foi obtida utilizando o filtro Remoção de Ruído. Nota-se que a medida-F do algoritmo J48 é maior tanto no filtro Remoção de Ruído quanto no filtro Balanceamento em relação ao algoritmo ID3. A diferença de percentual entre os algoritmos J48 e ID3 é de 3,22% no primeiro filtro e de 3,6% no segundo. Assim, apesar do algoritmo ID3 ter uma acurácia maior do que o J48 no uso Sem Filtro (56,91%, Figura 44) e com o filtro Balanceamento (79,20%, Figura 44), o ID3 possui uma medida-F para a classe efeito negativo menor do que o algoritmo J48.

Portanto, após esta análise inicial nos resultados das medidas acurácia e medida-F com a amostra de 601 sentenças, sendo treinadas e testadas com o método de particionamento 10-F CV, conclui-se que a aplicação de filtros melhora a distribuição das sentenças e a "pureza" das sentenças quando o ruído é eliminado. Ademais, os algoritmos J48, SVM e NB foram os que obtiveram os melhores resultados de medida-F para a classe efeito negativo que é a classe de objeto de estudo deste trabalho.

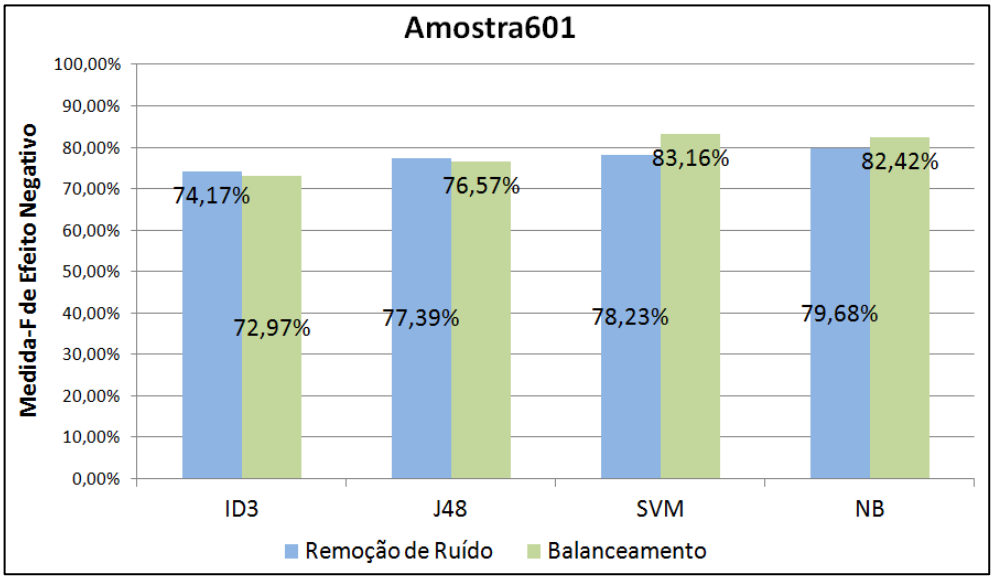

Figura 45 – Medida-F da classe efeito negativo na Amostra601 em relação aos filtros Remoção de Ruído e Balanceamento.

Em seguida, os três melhores algoritmos serão utilizados na fase de uso do modelo com a combinação do uso dos filtros Remoção de Ruído e Balanceamento. O intuito é identificar a combinação de algoritmo e filtro que melhor classifica as novas sentenças na classe efeito negativo.

7.1.2 Experimento 2: Fase de Uso do Modelo de Classificação

Este experimento tem o objetivo de avaliar qual a maior acurácia obtida dentre os algoritmos J48, SVM e NB na classificação das novas 300 sentenças e qual o percentual de medida-F do algoritmo de maior acurácia em relação aos filtros Remoção de Ruído e

Balanceamento. Estas novas sentenças são as sentenças da Amostra300 que serão classificadas a partir do modelo de classificação criado no experimento anterior com a Amostra601.

A distribuição das sentenças da Amostra300 por cada uma das classes pode ser vista na Figura 46. Na Figura 47 é ilustrada a acurácia obtida pelos três algoritmos de aprendizado de máquina em relação aos filtros Remoção de Ruído e Balanceamento. O algoritmo SVM obteve para ambos os filtros o maior percentual de acurácia (63,33%) em comparação aos algoritmos NB e J48.

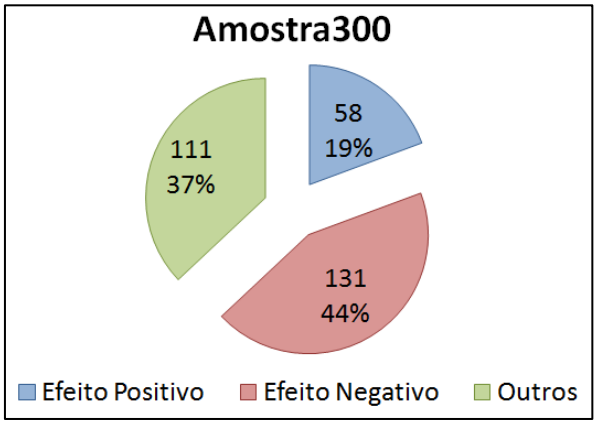

Figura 46 – Distribuição da Amostra300 por cada classe.

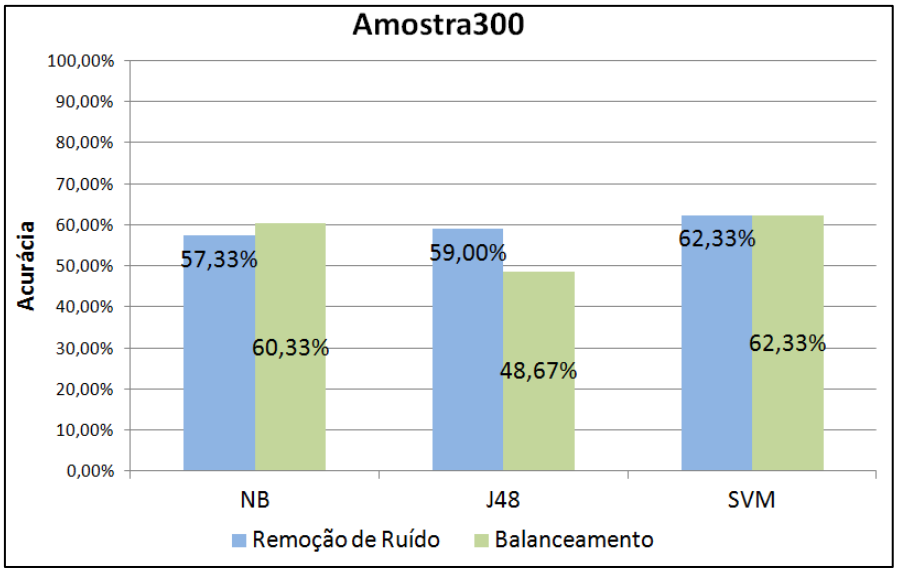

Figura 47 – Acurácia na Amostra300: Remoção de Ruído versus Balanceamento.

Na Figura 48 é ilustrada a medida-F da classe efeito negativo para os algoritmos NB, J48 e SVM em relação à aplicação dos filtros Remoção de Ruído e Balanceamento. O algoritmo SVM obteve o maior percentual tanto com o filtro Remoção de Ruído (71,81%) quanto com o filtro Balanceamento (71,43%). A diferença de percentual entre o SVM e o NB é de 5,14% para o filtro Remoção de Ruído e de 5,61% para o filtro Balanceamento. Já a

diferença entre o SVM e o J48 é de apenas 1,85% para o primeiro filtro e de 4,47% para o segundo filtro.

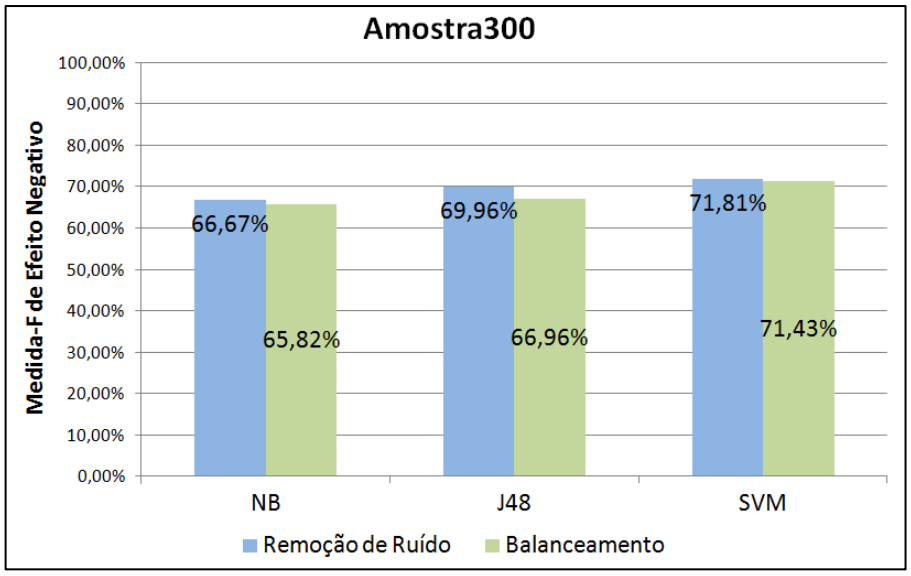

Figura 48 – Medida-F da classe efeito negativo na Amostra300: Remoção de Ruído versus Balanceamento.

Na Tabela 24 e Tabela 25 é apresentada a matriz de confusão do algoritmo SVM, respectivamente, com os filtros Remoção de Ruído e Balanceamento. A precisão e a revocação com o filtro Remoção de Ruído é, respectivamente de, 72,66% (93 de 128) e 70,99% (93 de 131). Já com o filtro Balanceamento, a precisão e a revocação é de 70,37% (95 de 135) e 72,52% (95 de 131). A precisão é maior com o filtro Remoção de Ruído do que com o filtro Balanceamento (diferença de 2,29%) e a revocação, por sua vez, é maior com o filtro Balanceamento do que com o filtro Remoção de Ruído (diferença de 1,53%). Ao calcular os valores de medida-F para os filtros Remoção de Ruído e Balanceamento, respectivamente, de 71,81% e 71,43%, o uso do filtro Remoção de Ruído é melhor 0,38% do que o filtro Balanceamento.

Tabela 24 – Matriz de confusão para o algoritmo SVM com Remoção de Ruído.

Classificado em / Avaliação Manual	**Efeito Negativo**	**Efeito Positivo**	**Outros**	Total
Efeito Negativo	93	2	36	131
Efeito Positivo	13	8	37	58
Outros	22	3	86	111
Total	128	13	159	300

Tabela 25 – Matriz de confusão para o algoritmo SVM com Balanceamento.

Classificado em / Avaliação Manual	**Efeito Negativo**	**Efeito Positivo**	**Outros**	Total
Efeito Negativo	95	13	23	131
Efeito Positivo	11	21	26	58
Outros	29	11	71	111
Total	135	45	120	300

Assim sendo, o algoritmo de aprendizado de máquina SVM obteve os melhores resultados de acurácia e medida-F e, portanto, será utilizado para classificar as 300 novas sentenças juntamente com a aplicação do filtro Remoção de Ruído. Este filtro será útil para construir a matriz atributo-valor utilizada pelo algoritmo SVM.

7.2 Identificação de Termos Relevantes

Nas fases de treinamento e de teste do estudo de caso anterior, foi criado o modelo de classificação para a Amostra601, utilizando o algoritmo de aprendizado de máquina SVM com a aplicação do filtro Remoção de Ruído. Também no estudo de caso anterior, o modelo de classificação criado foi utilizado para classificar as sentenças da Amostra300. O resultado da classificação na Amostra300 em relação à classe de objeto de estudo deste trabalho (i.e., classe efeito negativo) pode ser vista na Tabela 26.

Tabela 26 – Matriz de confusão para as sentenças que foram classificadas como sendo da classe efeito negativo.

Classificado em / Avaliação Manual	**Efeito Negativo**	**Efeito Positivo**	**Outros**	Total
Efeito Negativo	93	2	36	131
Efeito Positivo	13	----	----	----
Outros	22	----	----	----
Total	128	----	----	300

O objetivo deste estudo de caso é avaliar a identificação de termos relevantes nas sentenças que foram classificadas como sendo de efeito negativo. Das 300 sentenças apresentadas na Tabela 26, 93 sentenças são efeito negativo (i.e., verdadeiro positivo) e 35 sentenças (13 de efeito positivo e 22 de outros) não são efeitos negativos, mas foram classificadas como sendo efeito negativo (i.e., falso positivo) e 38 sentenças (2 de efeito

positivo e 36 de outros) são efeitos negativos, mas não foram classificadas como sendo efeito negativo (i.e., falso negativo).

Este estudo de caso tem o objetivo de avaliar a identificação dos termos relevantes nas sentenças e avaliar a influência do classificador na identificação desses termos. Nesse sentido, dois experimentos foram realizados, a saber: o primeiro experimento extrai informação das sentenças que são realmente efeito negativo (i.e., 131 sentenças, Tabela 26), avaliando a qualidade da extração sem influência da classificação automática; o segundo experimento extrai informação das sentenças que foram classificadas como sendo de efeito negativo (i.e., 128 sentenças, Tabela 26), avaliando assim a influência da classificação automática. No total existem 229 termos verdadeiros positivos nas 131 sentenças e 170 termos verdadeiros positivos nas 128 sentenças.

Para auxiliar na identificação desses termos, é utilizado o dicionário terminológico tanto no primeiro quanto no segundo experimento. Este dicionário contém os termos validados (i.e., curados) previamente pelo especialista do domínio. O dicionário é composto por 38 efeitos negativos da doença (i.e., complicações) e 19 efeitos negativos do tratamento (i.e., efeitos colaterais). A quantidade de variações dos termos sobre complicação e efeito colateral é de, respectivamente, 81 e 11. Os termos e as variações dos termos utilizados pela abordagem baseada em dicionário podem ser encontrados no APÊNDICE C – EFEITOS NEGATIVOS CURADOS.

É considerada a seguinte nomenclatura em relação à identificação de termos nos dois experimentos realizados a seguir. Os termos extraídos podem ser: termo completo, termo parcial, termo adicional, falso positivo e falso negativo. O termo extraído é considerado termo completo quando o termo identificado é exatamente o termo extraído (e.g., termo real *respiratory failure*, termo extraído *respiratory failure*); o termo extraído é considerado termo parcial quando o termo identificado não corresponde exatamente ao termo real (e.g., termo real *acute hepatic sequestration*, termo extraído *hepatic sequestration*); termo extraído é considerado termo adicional quando extrai mais palavras do que o termo real (e.g., termo real *chronic lung disease*, termo extraído *resultant chronic lung disease*); falso positivo é um termo que foi extraído, mas que não deveria ser extraído (e.g., *hydroxyurea therapy*); e falso negativo é um termo que deveria ser extraído, mas que não foi extraído (e.g., *thrombocytopenia*). Para avaliar os resultados da identificação de termos relevantes, foram utilizadas as medidas de precisão, revocação e medida-F. Os termos extraídos como sendo termo completo, termo parcial ou termo adicional são considerados verdadeiros positivos.

A seguir o primeiro experimento (i.e., avaliação da classificação manual versus extração) e em seguida o segundo experimento (i.e., avaliação da classificação automática versus extração) são explicados detalhadamente. O etiquetador POS utilizado nestes experimentos foi o modelo bidirecional desenvolvido pela universidade de Stanford, cujo percentual de acurácia foi o maior dentre os modelos. Os percentuais obtidos no treinamento e no uso do modelo de classificação foram, respectivamente, de 97,28% e de 90,46% conforme apresentados na Tabela 22 da Seção 6.5.2.

7.2.1 Experimento 1: Classificação Manual versus Extração

Este experimento tem o objetivo de avaliar a extração dos termos relevantes das 131 sentenças que contêm efeitos negativos. Ao todo existem 229 termos relevantes nessas sentenças. O resultado da extração não tem nenhuma influência do classificador e por isso, a extração realizada neste experimento é chamada de Extração Fictícia.

Na Figura 49 é apresentado o resultado da extração fictícia nas 131 sentenças, utilizando as abordagens de extração de informação baseada em regra, em dicionário, e em regra e dicionário juntos. A precisão do dicionário foi de 100%, pois todas as 131 sentenças continham termos verdadeiros positivos. Entretanto, a identificação dos termos pelo dicionário é limitada pela presença dos termos armazenados no dicionário e por isso, a revocação foi bem menor do que a precisão (68,12%). A precisão da regra, por sua vez, foi de 76,38% e a revocação de apenas 42,36%. A medida-F da regra e do dicionário foi, respectivamente, 54,59% e 81,04%. Ao utilizar as duas abordagens em conjunto, a precisão em relação à abordagem de dicionário teve uma redução de 13,51%. Esta redução é justificada pela identificação de falsos positivos pela abordagem de regra. Por outro lado, a revocação e a medida-F em relação à abordagem de dicionário teve um aumento de, respectivamente, 15,72% e 4,10%. Este aumento é devido à identificação de novos termos que a abordagem de regra identificou e que não foi identificado pela abordagem de dicionário.

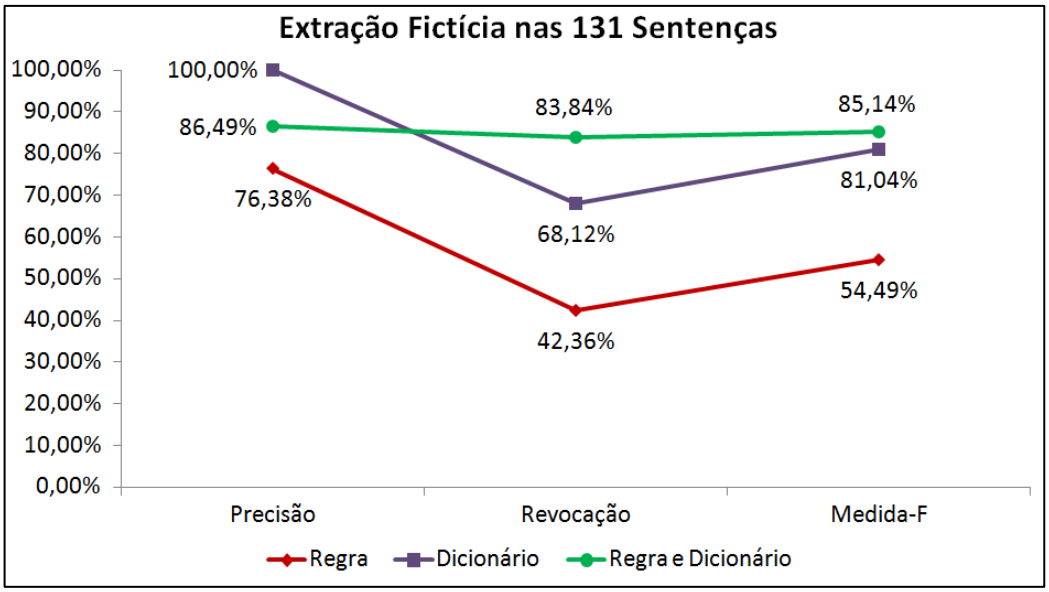

Figura 49 – Extração com regra e dicionário nas 131 sentenças classificadas manualmente pelo especialista.

Na Tabela 27 é possível observar a quantidade de termos verdadeiros positivos identificados pela regra, pelo dicionário e pelos dois em conjunto. O percentual total da abordagem baseada em regra e em dicionário de termo completo, de termo parcial e de termo adicional foi, respectivamente, 86,98%, 5,21% e 7,81%. No total, a abordagem baseada em regra, em dicionário e as duas usadas conjuntamente, identificaram, respectivamente, 97, 156 e 192 termos verdadeiros positivos.

Na Tabela 28 é apresentada a quantidade de verdadeiro positivo, de falso positivo e de falso negativo para cada uma das abordagens. Com a utilização da abordagem de regra e dicionário conjuntamente, a quantidade de verdadeiro positivo aumentou para 192 termos e a quantidade de falso negativo diminuiu, consideravelmente, para 37 termos. A quantidade de falso positivo permaneceu a mesma quantidade da abordagem de regra (i.e., 30 termos). É importante ressaltar, que a precisão do dicionário foi 100%, pois não foi identificado nenhum falso positivo. No total, foram identificados pela regra, pelo dicionário, e pela regra juntamente com dicionário, respectivamente, 259, 229 e 259 termos. Relembrando que a quantidade total de verdadeiro positivo nas 131 sentenças são 229 termos.

Tabela 27 – Verdadeiro positivo identificado pela regra e dicionário na extração fictícia.

Termos	Regra	Dicionário	Regra e Dicionário	Percentual
Termo Completo	80	141	167	86,98%
Termo Parcial	2	15	10	5,21%
Termo Adicional	15	0	15	7,81%
Total Verdadeiro Positivo	97	156	192	100,00%

Tabela 28 – Verdadeiro positivo, falso positivo e falso negativo em relação à regra e ao dicionário na extração fictícia.

Termos	Regra	Dicionário	Regra e Dicionário
Verdadeiro Positivo (VP)	97	156	192
Falso Positivo (FP)	30	0	30
Falso Negativo (FN)	132	73	37
Total (VP + FP + FN)	259	229	259

Na Tabela 29 é apresentado um exemplo de duas sentenças que a abordagem baseada em regra identificou termos verdadeiros positivos e que a abordagem baseada em dicionário não identificou. Na sentença 1, o termo "*laryngospasm*" foi identificado pela indicação do verbo "*occur*" e também pela indicação do verbo "*develop*". O termo "*pneumothorax"* foi identificado pelo verbo "*develop*" e o termo "*acute respiratory distress syndrome*" foi identificado pelo verbo "*have*" definido pela Estratégia 1 (Seção 6.3.2) e por um padrão *Part-Of-Speech (*POS) definido pela Estratégia 2 (Seção 6.3.2). Na sentença 2, os termos "*pulmonary disease*" e "*resultant chronic lung disease*" foram identificados por padrões POS definido pela Estratégia 2. O termo "*resultant chronic lung disease*" foi contabilizado como um termo adicional, pois o adjetivo "*resultant*" não faz parte do termo.

Tabela 29 – Sentenças que foram identificadas termos por meio de regra.

Sentença 1	***Laryngospasm*** *occurred in 10 patients, and* ***pneumothorax*** *developed in 2 patients, both of whom had had the* ***acute respiratory distress syndrome*** *before the bronchoscopy*
Sentença 2	*The longer patients with sickle cell disease live, the higher the frequency of recurrent* ***pulmonary disease*** *and* ***resultant chronic lung disease***

O *Gold Standard* foi obtido a partir das 131 sentenças, na qual foram identificados manualmente 229 termos. Para calcular o *Baseline*, foram extraídos das 131 sentenças todos os termos com os três padrões POS a seguir: adjetivo seguido de substantivo (padrão JJ_NN), substantivo seguido de substantivo (padrão NN_NN) e substantivos (padrão NN).

Na Tabela 30 são mostrados os verdadeiros positivos identificados pelo *Baseline*. Dentre os verdadeiros positivos, 48,86% dos termos foram substantivos. Os termos adjetivos seguido de substantivo identificados foram 43,38% e os substantivos seguido de substantivo foram apenas 7,76%. É importante ressaltar que os 45 termos parciais identificados pelo padrão JJ_NN são termos que precisam de padrões mais específicos para ser identificados completamente. Por exemplo, o termo "*transient hematologic toxicity*" poderia ser extraído completamente pelo padrão POS JJ_JJ_NN e o termo "*central nervous system abnormalities"* poderia ser extraído pelo padrão JJ_JJ_NN_NN. Percebe-se que com os padrões POS

apresentados na Tabela 30 não foram identificados nenhum termo verdadeiro positivo do tipo termo adicional. No total, foram identificados 219 verdadeiros positivos.

Tabela 30 – Verdadeiros positivos do *Baseline* nas 131 sentenças.

Padrão POS Utilizado	Termo Completo	Termo Parcial	Total	Percentual
JJ_NN	50	45	95	43,38%
NN_NN	17	0	17	7,76%
NN	107	0	107	48,86%
Total	174	45	219	100,00%

Na Tabela 31 são mostrados os valores para os falsos positivos identificados pelo *Baseline*. A grande maioria, 69,86% das palavras é substantivo. O restante do percentual é de adjetivo seguido de substantivo (18,47%) e de substantivo seguido de substantivo (11,67%). Ainda na Tabela 31, nota-se que os valores totais dos falsos positivos representados na quarta coluna são subtraídos dos verdadeiros positivos da terceira coluna. Isto é necessário, pois os valores dos falsos positivos (i.e., segunda coluna) correspondem a todas as palavras da amostra, incluindo os verdadeiros positivos. Os valores dos verdadeiros positivos (i.e., terceira coluna) são correspondentes aos valores da quarta coluna da Tabela 30. No total, foram identificados 720 falsos positivos.

Também foram identificados falsos negativos. Os falsos negativos foram termos que não foram etiquetados como substantivos nem como adjetivos (e.g., o verbo "*died*") e termos que são substantivos, mas foram classificados erroneamente pelo etiquetador POS como adjetivos (e.g., "*carcinogenic*") ou verbo (e.g., "*wheezing*"). Também houve termos classificados como palavras estrangeiras pelo classificador (e.g., "*cor pulmonale"*). Ao todo foram identificados 10 falsos negativos.

Tabela 31 – Falsos positivos do *Baseline* nas 131 sentenças.

Padrão POS Utilizado	Falso Positivo	Verdadeiro Positivo	Total	Percentual
JJ_NN	228	95	133	18,47%
NN_NN	101	17	84	11,67%
NN	610	107	503	69,86%
Total	939	219	720	100,00%

Na Figura 50 é apresentado o melhor resultado de medida-F da extração fictícia (i.e., regra e dicionário) em comparação com o *Baseline* e o *Gold Standard* nas 131 sentenças. A

diferença da medida-F da extração fictícia para o *Gold Standard* é 14,86% e para o *Baseline* é 47,64%.

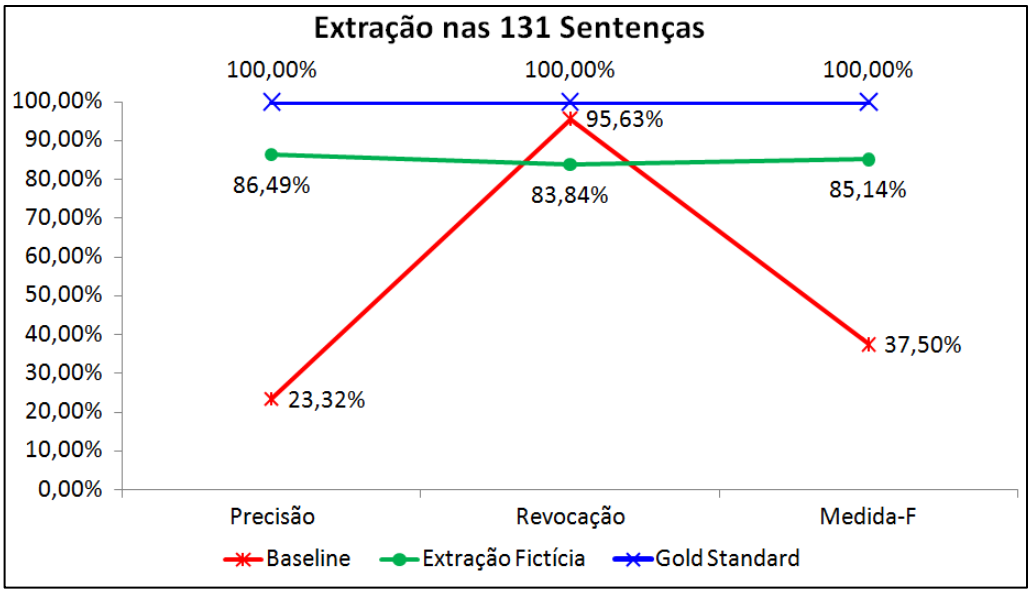

Figura 50 – Extração nas 131 sentenças classificadas manualmente pelo especialista comparado com o *Baseline* e o *Gold Standard*.

7.2.2 Experimento 2: Classificação Automática versus Extração

Este experimento tem o objetivo de avaliar a extração dos termos relevantes das 128 sentenças da Amostra300 que foram classificadas na classe efeito negativo (Tabela 26 no início da Seção 7.2). Ao todo existem 170 termos relevantes nessas sentenças. O resultado da extração pode ser influenciado pela classificação dessas sentenças. A extração realizada neste experimento é chamada de Extração Real, pois a classificação de sentenças é realizada automaticamente, diferentemente do primeiro experimento que considerou todas as sentenças classificadas corretamente pelo especialista.

Na Figura 51 é ilustrado o resultado da extração real nas 128 sentenças da Amostra300, utilizando as abordagens de extração de informação baseada em regra, em dicionário, e em regra e dicionário juntos. A precisão do dicionário foi de 87,31%, pois dentre as 128 sentenças existiam sentenças que eram falsos positivos e a revocação foi de 68,82%. A precisão da regra, por sua vez, foi de 62,96% e a revocação de apenas 50%. A medida-F da regra e do dicionário foi, respectivamente, 55,74% e 76,97%. Ao utilizar as duas abordagens em conjunto, a precisão em relação à abordagem de dicionário teve uma redução de 12,56%. Esta redução é justificada pela identificação de falsos positivos pela abordagem de regra. Por outro lado, a revocação e a medida-F em relação à abordagem de dicionário teve um aumento de, respectivamente, 18,24% e 3,46%. Este aumento é devido à identificação de novos termos que a abordagem de regra identificou e que não foi identificado pela abordagem de dicionário.

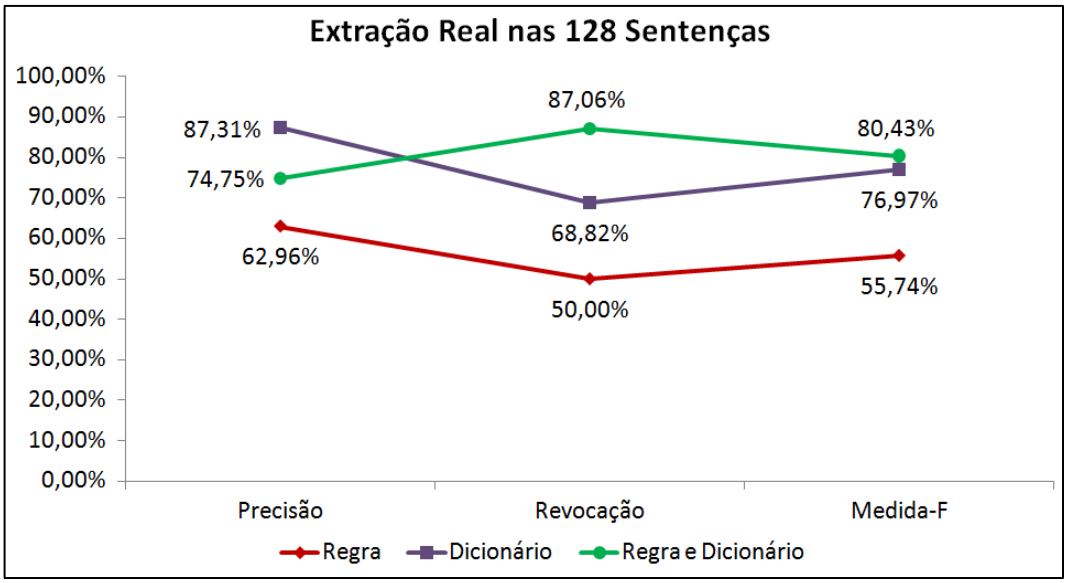

Figura 51 – Extração com regra e dicionário nas 300 sentenças classificadas automaticamente pelo algoritmo de classificação de sentenças.

Na Tabela 32 pode-se observar a quantidade de termos verdadeiros positivos identificados pela regra, pelo dicionário e pelos dois em conjunto. O percentual total das abordagens baseada em regra e em dicionário em relação ao termo completo, termo parcial e termo adicional foi, respectivamente, 86,49%, 4,73% e 8,78%. Nota-se que o dicionário nunca identifica termo adicional, exceto se o especialista validar erroneamente um "termo adicional" no dicionário. No total, a abordagem baseada em regra, em dicionário e as duas juntamente, identificaram, respectivamente, 85, 117 e 148 termos verdadeiros positivos.

Tabela 32 – Verdadeiro positivo identificado pela regra e dicionário na extração real.

Termos	Regra	Dicionário	Regra e Dicionário	Percentual
Termo Completo	71	107	128	86,49%
Termo Parcial	1	10	7	4,73%
Termo Adicional	13	0	13	8,78%
Total Verdadeiro Positivo	85	117	148	100,00%

Na Tabela 33 é apresentada a quantidade de verdadeiro positivo, de falso positivo e de falso negativo para cada uma das abordagens. Com a utilização das abordagens de regra e de dicionário em conjunto, a quantidade de verdadeiro positivo aumentou para 148 termos e a quantidade de falso negativo diminuiu, expressivamente, para 22 termos. A quantidade de falso positivo permaneceu a mesma quantidade da abordagem de regra (i.e., 50 termos). Certamente os 17 termos falsos positivos identificados pelo dicionário já estão incluídos entre os 50 termos falsos positivos identificados pela regra e por isso, a quantidade de falsos positivos não aumentou com o uso das duas abordagens. No total, foram identificados pela

regra, pelo dicionário e pela regra juntamente com dicionário, respectivamente, 220, 187 e 220 termos. Relembrando que a quantidade total de verdadeiro positivo nas 128 sentenças são 170 termos.

Tabela 33 – Verdadeiro positivo, falso positivo e falso negativo identificados pela regra e dicionário na extração real.

Termos	Regra	Dicionário	Regra e Dicionário
Verdadeiro Positivo (VP)	85	117	148
Falso Positivo (FP)	50	17	50
Falso Negativo (FN)	85	53	22
Total (VP + FP + FN)	220	187	220

Na Tabela 34 é apresentado um exemplo de duas sentenças que foram classificadas como sendo da classe efeito negativo. A primeira sentença é uma sentença da classe efeito positivo. Contudo, a presença de termos sobre efeito negativo como "*stroke*" e "*iron overload*" induz o classificador a classificar como sendo efeito negativo. Na segunda sentença, as palavras "*leukemia*" e "*diseases*" são palavras que aparecem na classe efeito negativo. Analisando o sentido das duas sentenças, nota-se que a sentença 1 é realmente um efeito positivo. Já a sentença 2 pode ter sido classificada erroneamente pelo especialista, pois a sentença expressa uma tendência de doenças mieloproliferativas que evoluem para leucemia aguda.

Tabela 34 – Sentenças classificadas como efeito negativo e que são falsos positivos. Os termos em negrito foram identificados pelo dicionário.

Sentença 1	*The use of HU instead of transfusion for **stroke** prevention will avoid the risk of **iron overload** and the need of iron chelation*
Sentença 2	*However, the relevance of these reports to sickle cell anemia is unclear because of the inherent tendency of myeloproliferative diseases to evolve into acute **leukemia***

Na Tabela 35 é apresentado um exemplo de sentenças que foram classificadas pelo especialista como sendo sentenças da classe Outros ou da classe Efeito Positivo e que foram classificadas pelo algoritmo de aprendizado de máquina como sendo da classe Efeito Negativo, ou seja, estas quatro sentenças são consideradas falso positivo. As duas primeiras são da classe Outros. Todavia ao analisar estas duas sentenças, os termos destacados em negrito ("*obstructive airway abnormalities" e "leukemia"*) são de fato efeitos negativos. Este tipo de erro pode ter acontecido durante a classificação das sentenças realizada pelo especialista e que é aceitável, pois todo o ser humano está suscetível a falhas. O crédito da classificação é do algoritmo de aprendizado de máquina que "acertou" a classe. Contudo, ao

analisar a terceira sentença, o termo em negrito ("*physician-diagnosed asthma"*) é um efeito negativo, mas a sentença foi considerada pelo especialista como um efeito positivo devido à redução da mortalidade associada à síndrome torácica aguda (i.e., ACS). A quarta sentença é de fato uma sentença sobre efeito negativo e que mais uma vez o algoritmo de aprendizado "acertou".

Estas sentenças apresentadas na Tabela 35 são quatro exemplos de sentenças que foram classificadas pelo algoritmo de aprendizado como sendo da classe Efeito Negativo e que foram classificadas pelo especialista como sendo da classe Efeito Positivo ou da classe Outros. Destas sentenças classificadas "erroneamente" como sendo da classe Efeito Negativo, 24 termos foram identificados pelas regras. Estes 24 termos, juntamente com outros 26 termos falsos positivos que foram identificados nas sentenças nas quais as classes foram corretamente classificadas como sendo de efeito negativo, foram os responsáveis pelos baixos percentuais de precisão (62,96%) e revocação (50%) obtidos com a abordagem de regra. O total de falsos positivos identificados pela abordagem baseada em regra (24 + 26 = 50 termos) pode ser visto na Tabela 33 e os baixos percentuais de precisão e de revocação podem ser observados na Figura 51.

Tabela 35 – Sentenças classificadas como efeito negativo e que são falsos positivos. Os termos em negrito foram identificados pela regra.

Sentença 1	*In adolescents, one third of asthmatics, that is those with asthma-like symptoms and one or more **obstructive airway abnormalities** may not have been diagnosed as asthmatic by a physician*
Sentença 2	*However, the relevance of these reports to sickle cell anemia is unclear because of the inherent tendency of myeloproliferative diseases to evolve into acute **leukemia***
Sentença 3	*We speculate that aggressive treatment with moderate to high doses of ICS and a leukotriene modifier (5-liopoxygenase inhibitor or LTRA) in SCD patients with **physician-diagnosed asthma** and SCD may reduce the mortality and morbidity associated with ACS*
Sentença 4	*Patients with **physician-diagnosed asthma** were 4.0 times (95% CI, 1.7, 9.5) more likely to develop ACS during the admission than patients without asthma*

O *Gold Standard* foi obtido a partir das 128 sentenças, na qual foram identificados manualmente 170 termos. Para calcular o *Baseline*, foram extraídos das 128 sentenças todos os termos com os três padrões POS: adjetivo seguido de substantivo, substantivo seguido de substantivo e substantivos. Na Tabela 36 são apresentados os verdadeiros positivos identificados pelo *Baseline*. Dentre os verdadeiros positivos, 46,95% dos termos foram adjetivos seguido de substantivo (padrão JJ_NN). Os termos substantivos (padrão NN) identificados foram 46,34% e os substantivos seguido de substantivo (padrão NN_NN) foram

apenas 6,71%. Para os padrões NN_NN e NN não foram identificados nenhum termo parcial. Percebe-se que com estes padrões POS não foram identificados nenhum termo verdadeiro positivo do tipo termo adicional. No total, foram identificados 164 verdadeiros positivos.

Tabela 36 – Verdadeiros positivos do *Baseline* nas 128 sentenças.

Padrão POS Utilizado	Termo Completo	Termo Parcial	Total	Percentual
JJ_NN	39	38	77	46,95%
NN_NN	11	0	11	6,71%
NN	76	0	76	46,34%
Total	126	38	164	100,00%

Na Tabela 37 são mostrados os valores para os falsos positivos identificados pelo *Baseline*. A grande maioria das palavras é substantivo (65,81%). O restante do percentual é de adjetivo seguido de substantivo (20,30%) e de substantivo seguido de substantivo (13,89%). Ainda na Tabela 37, nota-se que os valores totais dos falsos positivos representados na quarta coluna são subtraídos dos verdadeiros positivos da terceira coluna. Isto é necessário, pois os valores dos falsos positivos (i.e., segunda coluna) correspondem a todas as palavras da amostra, incluindo os verdadeiros positivos. Os valores dos verdadeiros positivos (i.e., terceira coluna) são correspondentes aos valores da quarta coluna da Tabela 36. No total, foram identificados 857 falsos positivos.

Também foram identificados falsos negativos. Os falsos negativos foram termos que não foram etiquetados como substantivos nem como adjetivos (e.g., o verbo "*died*") e termos classificados como palavras estrangeiras pelo classificador (e.g., "*cor pulmonale"*). Ao todo foram identificados 7 falsos negativos.

Tabela 37 – Falsos positivos do *Baseline* nas 128 sentenças.

Padrão POS Utilizado	Falso Positivo	Verdadeiro Positivo	Total	Percentual
JJ_NN	251	77	174	20,30%
NN_NN	130	11	119	13,89%
NN	640	76	564	65,81%
Total	1021	164	857	100,00%

Na Figura 52 é apresentado o melhor resultado de medida-F da extração real (i.e., regra e dicionário) em comparação com o *Baseline* e o *Gold Standard* nas 128 sentenças da Amostra300. A diferença da medida-F da extração real para o *Gold Standard* é 19,57% e para o *Baseline* é 52,91%.

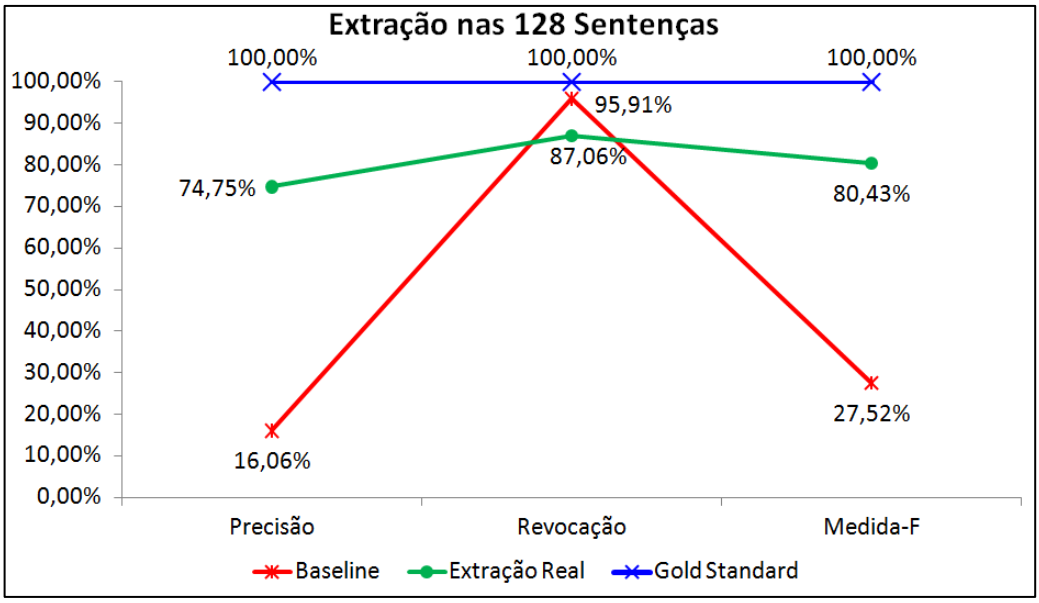

Figura 52 – Extração nas 128 sentenças classificadas automaticamente pelo algoritmo de classificação de sentenças comparado com o *Baseline* e o *Gold Standard*.

7.3 Considerações Finais

Neste capítulo foi descrita a avaliação realizada na metodologia proposta neste trabalho. A avaliação foi realizada em artigos científicos relacionados à doença Anemia Falciforme. As informações relacionadas à doença foram sobre efeito negativo e efeito positivo. Os efeitos foram identificados, com o aval do especialista, em três seções dos artigos: *abstract*, *results* e *discussion*. Esta restrição da seleção das sentenças nestas três seções restringiu a possibilidade de extrair muitos falsos positivos.

Duas etapas da metodologia foram avaliadas: a Classificação de Sentenças (Etapa 2) e a Identificação de Termos Relevantes (Etapa 3). Na Etapa 2 dois experimentos foram realizados: o primeiro experimento foi nas fases de treinamento e de teste da classificação de sentenças, cujo objetivo foi criar um modelo de classificação que melhor representasse as sentenças treinadas. Experimentos foram realizados com uma amostra de 601 sentenças (2/3 de 901 sentenças) com seis algoritmos clássicos de aprendizado de máquina (i.e., SVM, NB, ID, J48, Prism e OneR) em combinação com três diferentes testes (i.e., Sem Filtro, Remoção de Ruído e Balanceamento) para construir o modelo de classificação. Os melhores percentuais de medida-F na classe efeito negativo são apresentados na Figura 53; o segundo experimento foi realizado com uma amostra de 300 sentenças (1/3 de 901 sentenças) na fase de uso do modelo, cujo objetivo foi classificar as novas sentenças com o modelo de classificação criado no primeiro experimento. O melhor algoritmo foi o SVM em combinação com o filtro Remoção de Ruído, que obteve uma medida-F de 71,81% na classe efeito negativo (Figura 54).

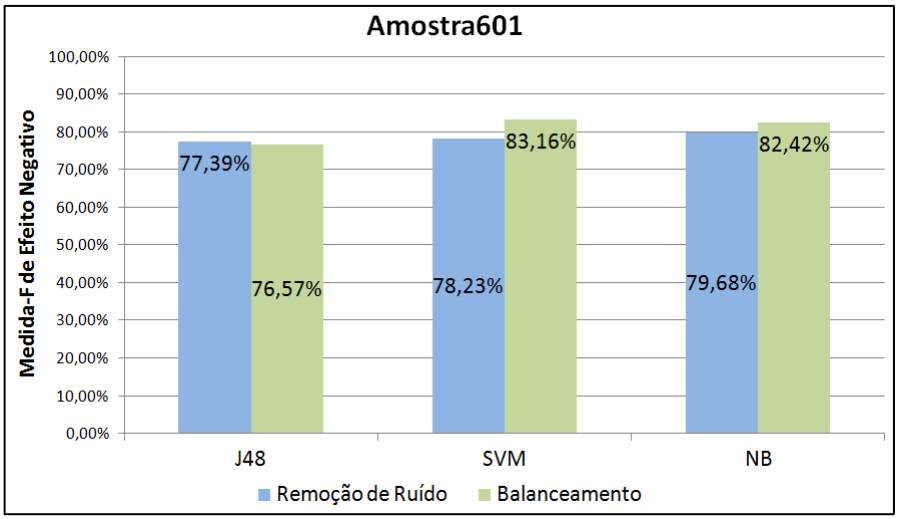

Figura 53 – Medida-F dos algoritmos J48, SVM e NB na classe efeito negativo na Amostra601: Remoção de Ruído versus Balanceamento.

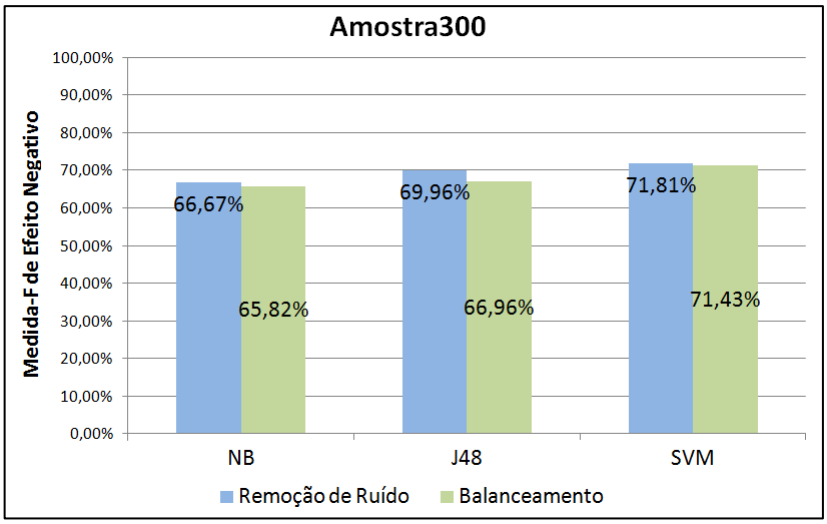

Figura 54 – Medida-F dos algoritmos NB, J48 e SVM na classe efeito negativo na Amostra300: Remoção de Ruído versus Balanceamento.

Em seguida na Etapa 3, também foram realizados dois experimentos: o primeiro experimento teve o intuito de avaliar a identificação dos termos relevantes nas sentenças sem a influência do classificador. Para isso, foram selecionadas as 131 sentenças da amostra de 300 sentenças que continham efeito negativo; o segundo experimento teve o intuito de extrair somente informação das sentenças que foram classificadas como sendo de efeito negativo (i.e., 128 sentenças da Amostra300) e assim avaliar a influência da classificação automática. No total existiam 229 termos verdadeiros positivos nas 131 sentenças e 170 termos verdadeiros positivos nas 128 sentenças. Na Figura 55 são apresentados os resultados dos dois experimentos da extração: extração fictícia e extração real. A extração é fictícia, pois é realizada nas 131 sentenças que foram classificadas manualmente pelo especialista (i.e., classificação manual). A extração é real, pois é realizada nas 128 sentenças que foram classificadas automaticamente pelo classificador (i.e., classificação automática). A precisão da

extração fictícia é 11,74% maior do que a extração real. Já a revocação da extração real é 3,22% maior do que a extração fictícia. A diferença da medida-F da extração fictícia para a extração real é 4,71%.

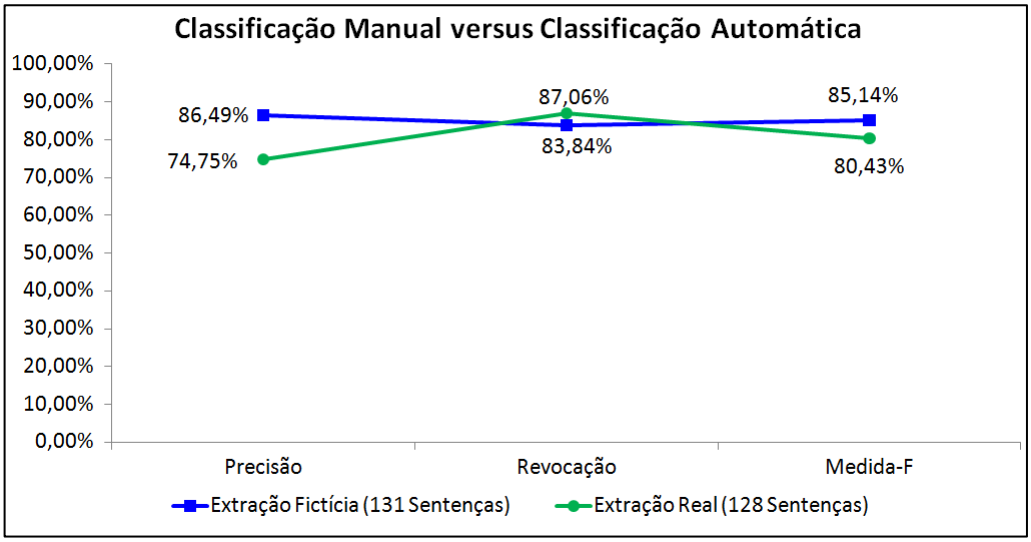

Figura 55 – Classificação Manual versus Classificação Automática.

A partir da análise dos dados da Etapa 2, percebe-se que a classificação de sentenças é uma etapa importante na metodologia, pois ela tem o objetivo de distinguir as sentenças relevantes das irrelevantes. Um classificador ideal com acerto de 100% obteria uma medida-F na extração de informação de 85,14%. O percentual obtido pela extração real foi próximo deste valor (80,43%), o que justifica o uso do classificador de sentenças na metodologia proposta.

8 CONCLUSÃO

Neste trabalho foi proposta uma metodologia de pré-processamento textual para extrair informação de efeitos de doenças em artigos científicos do domínio biomédico. A metodologia é composta por quatro etapas: **Entrada de Dados** (Etapa 1), **Classificação de Sentenças** (Etapa 2), **Identificação de Termos Relevantes** (Etapa 3) e **Gerenciamento de Termos** (Etapa 4). A partir dos documentos textuais fornecidos na Etapa 1, as sentenças são classificadas em suas respectivas classes (Etapa 2). A Etapa 2 objetiva distinguir as informações de interesse das informações irrelevantes. Em seguida, na Etapa 3, os termos relevantes são identificados e extraídos das sentenças de interesse. Por fim, na Etapa 4, os termos são armazenados em um banco de dados, após a validação dos termos pelo especialista. Os termos validados podem ser utilizados em um novo ciclo de extração de informação, a fim de identificar termos em novas sentenças. Este recurso é importante, pois a regra pode identificar um termo em somente uma sentença X, mas após a validação deste termo pelo especialista, o dicionário pode identificar este mesmo termo em outras N sentenças.

A Etapa 2 e a Etapa 3 foram validadas por meio de estudos de caso sobre informações relacionadas à doença Anemia Falciforme. Em cada uma das duas etapas foram realizados dois experimentos. Na Etapa 2, o primeiro experimento teve como objetivo criar e testar o modelo de classificação. Para isso, foi utilizada uma amostra de 601 sentenças classificadas manualmente pelo especialista em três classes: efeito positivo, efeito negativo e outros (i.e., não é efeito positivo ou efeito negativo). Este modelo foi avaliado com a medida acurácia a fim de avaliar a classificação em relação a todas as três classes e com a medida-F para avaliar a classificação da classe efeito negativo. Os percentuais de acurácia e de medida-F obtidos foram, respectivamente, acima de 85% e de 80%. Estes percentuais foram obtidos por meio do método de particionamento 10-*Fold Cross-Validation*. Em seguida, foi realizado o segundo experimento que teve o objetivo de avaliar como o modelo de classificação criado no primeiro experimento se comportava na classificação de novas sentenças. Para isso, foi utilizada uma amostra de 300 sentenças que também foi classificada manualmente pelo

especialista nas três classes citadas anteriormente. Os percentuais de acurácia obtidos foram acima de 60% e os percentuais de medida-F para a classe efeito negativo foram acima de 70%. As sentenças da Amostra601 e da Amostra300 foram selecionadas aleatoriamente a partir de um conjunto de 901 sentenças. A primeira e a segunda amostra correspondem, respectivamente, 2/3 e 1/3 das 901 sentenças.

A partir dos experimentos realizados na Etapa 2, foi realizada a avaliação dos dois experimentos realizados na Etapa 3. O primeiro experimento teve o objetivo de avaliar a extração de informação com a influência da classificação manual. Para isso, as 131 sentenças da Amostra300 que continham efeito negativo foram utilizadas, a fim de serem identificados os termos relevantes. Utilizando as abordagens de regra e de dicionário em conjunto, os percentuais de precisão e de medida-F para a classe efeito negativo foram acima de 85% e o percentual de revocação foi acima de 80%. O percentual de revocação foi próximo aos percentuais obtidos pelo *Gold Standard* (100%) e pelo *Baseline* (aproximadamente 96%). Os percentuais de precisão e de medida-F também foram próximos do percentual do *Gold Standard* (100%), entretanto os percentuais foram extremamente melhores do que o *Baseline* que foi aproximadamente de apenas 23% de precisão e somente 37% de medida-F. O segundo experimento teve como objetivo avaliar a extração de informação com influência da classificação automática. Assim, ao invés de utilizar as 131 sentenças que continham somente efeito negativo, neste experimento foram utilizadas as 128 sentenças classificadas pelo classificador, a partir da Amostra300, como sendo de efeito negativo. Os percentuais do classificador para a Amostra300 foram de 72,66% de precisão, 70,99% de revocação e 71,81% de medida-F. Estes percentuais foram obtidos no segundo experimento realizado na Etapa 2. Os percentuais de precisão, de revocação e de medida-F para a classe efeito negativo utilizando as abordagens de regra e de dicionário em conjunto foram, respectivamente, acima de 70%, acima de 85% e acima de 80%. O percentual de revocação foi próximo aos percentuais obtidos pelo *Gold Standard* (100%) e pelo *Baseline* (aproximadamente 96%). Os percentuais de precisão e de medida-F também foram próximos do percentual do *Gold Standard* (100%), entretanto os percentuais foram excepcionalmente melhores do que o *Baseline* que obteve aproximadamente apenas 16% de precisão e de somente 27% de medida-F. Comparando a extração com a classificação manual versus a classificação automática, há uma pequena diferença de 4,71% de medida-F. Em suma, a extração com a classificação manual é o ideal, contudo o especialista não dispõe de tempo para realizar constantemente a classificação manual das sentenças. Assim, a extração com a classificação automática é uma alternativa possível e viável, pois proporciona somente uma pequena diminuição da medida-F.

A seguir são apresentadas as contribuições deste trabalho (Seção 8.1), os possíveis trabalhos futuros (Seção 8.3) e por fim, as produções científicas e técnicas desenvolvidas durante o mestrado (Seção 8.4).

8.1 Contribuições

A principal contribuição deste trabalho é a proposta de uma metodologia de pré-processamento textual para extrair informação sobre efeitos de doenças em artigos completos do domínio biomédico. A metodologia proposta possui a grande vantagem de possibilitar a extração de informação em artigos científicos completos, o que é realizado por poucos trabalhos identificados na literatura. Na metodologia proposta neste trabalho, somente é possível extrair informação nos artigos completos sem extrair muitos falsos positivos, devido à utilização de um algoritmo de aprendizado de máquina na Etapa 2 (i.e., Classificação de Sentenças) cujo objetivo é distinguir as sentenças que são de interesse (i.e., sentenças que têm algum termo relevante) das sentenças irrelevantes. Apesar da medida-F obtida com a classificação de sentenças para a classe efeito negativo não ser 100%, o percentual de medida-F de 71,81% obtida com a Amostra300 é um percentual que contribui na identificação de termos relevantes, possibilitando extrair menos falsos positivos.

Além desta contribuição teórica, são destacadas as contribuições de cunho prático:

1. Criação e disponibilização de recursos como coleção de documentos do domínio, de termos do dicionário terminológico e de bases de regras desenvolvidas para extrair automaticamente os termos relevantes sobre a doença Anemia Falciforme;
2. Criação e disponibilização de ferramentas para a classificação de sentenças, para a extração de informação e para o gerenciamento de termos. A ferramenta de classificação faz o uso da API do Weka amplamente utilizada no contexto de aprendizado de máquina e de mineração de dados. As três ferramentas foram implementadas em Java.

8.2 Adaptabilidade da Metodologia Proposta

A metodologia proposta possui quatro etapas conforme apresentado neste trabalho. A maioria das etapas da metodologia (i.e., Etapa 1, Etapa 2 e Etapa 4) é independente do domínio e pode ser aplicada a qualquer domínio sem modificações. Já a Etapa 3 é dependente do domínio.

A Etapa 1 tem a restrição dos artigos científicos estarem nos formatos XML ou TXT. A Etapa 2 é totalmente independente do domínio e do idioma. Isto significa que nesta etapa pode-se classificar sentenças de qualquer problema que envolva a distinção de classes, por

exemplo, sentenças de artigos científicos que contêm informação sobre paciente, tratamento e fator de risco.

A Etapa 3 possui a dependência nas duas abordagens para extração de informação: dicionário e regra. Como a extração de informação realizada pelo dicionário é o casamento exato dos termos armazenados no dicionário e, por conseguinte, identificados nas sentenças, o custo de adaptação depende somente da existência de um novo dicionário do domínio. Já a extração de informação realizada pela regra necessita de uma adaptação maior, pois as regras são criadas a partir de uma análise no conjunto de sentenças do domínio, no qual são identificados termos relevantes que contêm um conjunto de padrões *Part-Of-Speech* que são rigorosamente dependentes dos termos a serem extraídos. Contudo, as duas estratégias apresentadas neste trabalho para identificar os termos relevantes em uma sentença utilizando a abordagem de regra podem ser utilizadas e adaptadas para a extração de informação de termos que não sejam efeitos de doenças.

A Etapa 4 também é totalmente independente do domínio de atuação. Isto significa que os termos podem ser gerenciados pelo especialista sem existir nenhuma restrição às informações armazenadas no banco de dados.

8.3 Trabalhos Futuros

A seguir são enumeradas as sugestões de trabalhos futuros:

– Hierarquização automaticamente dos termos extraídos. Por exemplo, suponha que os termos "*parvovirus infection*" e "*infection*" tenham sido extraídos automaticamente e armazenados na entidade "Termo" da Figura 56. Quando esses dois termos fossem utilizados para extrair informação na sentença "*Aplastic crisis associated with a parvovirus infection was documented in one patient*", ambos os termos seriam identificados, o que causaria erro de duplicidade de informação. Assim, é necessário armazenar o termo mais genérico na entidade "Termo" (i.e., *infection*) e os termos mais específicos na entidade "Variação" (i.e., *parvovirus infection*);

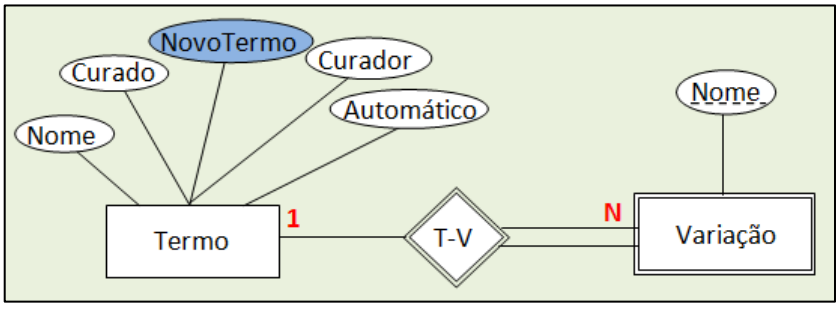

Figura 56 – Exemplo de esquema conceitual com as entidades termo e variação.

- Classificação automaticamente dos efeitos negativos extraídos em efeitos negativos da doença (complicação) e efeitos negativos do tratamento (efeito colateral);
- Criação de uma coleção de documentos anotada com informações relacionadas ao domínio da doença Anemia Falciforme, contando com a participação do especialista do domínio na validação das informações anotadas. Esta coleção poderá ser útil para auxiliar o processo de extração de informação automático, contribuindo com uma validação de forma rápida, prática e justa. Nesse sentido, um documento XML poderia ser criado para identificar os termos de interesse em cada sentença, assim como a classe a qual a sentença faz parte. Um exemplo deste documento XML pode ser visto na Figura 57. Pelo que se conhece, não existe uma coleção de documentos com informações anotadas sobre essa doença;

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<document>
<journal>journal name</journal>
<title>paper title</title>
<year>paper year</year>
<author>author name</author>
<section name = "section name">
    <page number = "1">
        <paragraph>
            <sentence-info>
                <sentence>sentence X</sentence>
                <manual-class>class name Y</manual-class>
                <manual-term>term 1; term 2; term n</manual-term>
                <automatic-class>class name Y</automatic-class>
                <automatic-term>term 1; term 2; term n</automatic-term>
            </sentence-info>
            ...
        </paragraph>
    </page>
    ...
</section>
```

Figura 57 – Exemplo de documento XML anotado.

- Investigação da possibilidade da aplicação da metodologia na identificação de efeitos positivos, além de outros termos como tratamento e fator de risco da doença Anemia Falciforme;
- Instanciação da metodologia proposta para identificar efeitos de outras doenças importantes, como câncer, mal de Alzheimer, mal de Parkinson e glaucoma;
- Investigação da aplicação da metodologia em outros domínios além do domínio biomédico, por exemplo, erupção de vulcão e poluição ambiental nos quais seria interessante identificar termos de efeitos ocasionados pela fuligem na erupção de vulcão e os efeitos originados da poluição ambiental;

- Utilização de recursos de Processamento de Língua Natural como a análise semântica, a fim de identificar se a sentença expressa um efeito positivo, um efeito negativo ou nenhum efeito. A análise semântica é útil em situações em que os termos não estão explicitamente escritos nas sentenças. Por exemplo, a seguinte sentença não contém um termo explícito, "*The recent availability of an oral iron chelator may render prolonged transfusion more acceptable.*";
- Investigação se a utilização de uma ontologia contribui com a identificação dos termos relevantes;
- Utilização de algoritmos de mineração de dados para fornecer associações e hipóteses de ocorrências não explícitas no banco de dados. Para isso, este banco de dados deve ser povoado com termos de interesse como efeitos, tratamento e quantidade de pacientes;
- Desenvolvimento de um sistema de Perguntas e Respostas, a fim de fornecer ao usuário uma informação específica, por exemplo, "Qual a faixa etária de crianças que têm mais efeitos negativos?";
- Desenvolvimento de uma ferramenta de visualização dos termos extraídos e as correlações entre eles.

8.4 Produção Científica e Técnica

A primeira discussão sobre a proposta deste trabalho foi realizada no *Workshop* de Teses e Dissertações em Banco de Dados. Como consequência do desenvolvimento do trabalho, foi publicada a visão geral do ambiente de análise de dados para o domínio biomédico com resultados sobre a classificação de sentenças em um evento internacional. As duas produções científicas são listadas a seguir:

MATOS, P. F.; LOMBARDI, L. O.; PARDO, T. A. S; CIFERRI, C. D. A. ; VIEIRA, M. T. P.; CIFERRI, R. R. An environment for data analysis in biomedical domain: information extraction for decision support systems. In: GARCÍA-PEDRAJAS, N. et al. (Ed.). **International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE)**. 23th. Heidelberg: Springer, 2010. p. 306-316. (Lecture Notes in Computer Science; v. 6096).

MATOS, P. F.; CIFERRI, R. R.; PARDO, T. A. S. Metodologia de pré-processamento textual para extração de informação em artigos científicos do domínio biomédico. In: WORKSHOP DE TESES E DISSERTAÇÕES EM BANCOS DE DADOS, VIII, 2009, Fortaleza, Ceará. **Anais...** Simpósio Brasileiro de Banco de Dados, 2009. p. 7-12.

Como resultado de desenvolvimento deste trabalho, também foram publicadas quatro produções técnicas: um pôster e três relatórios técnicos, a saber:

MATOS, P. F.; CIFERRI, R. R.; PARDO, T. A. S. Methodology of textual preprocessing for information extraction in scientific papers of the biomedical domain. In: WORKSHOP DE PÓS-GRADUAÇÃO SEMANA DE COMPUTAÇÃO, 3º, 2010, São Carlos. **Anais...** UFSCar, 2010. Pôster. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/poster.WPG.PPG-CC.pdf>. Acesso em: 30 ago. 2010.

MATOS, P. F.; LOMBARDI, L. O.; CIFERRI, R. R.; PARDO, T. A. S.; CIFERRI, C. D. A.; VIEIRA, M. T. P. **Relatório Técnico "Conceitos sobre Aprendizado de Máquina"**. São Carlos: Departamento de Computação, Universidade Federal de São Carlos, 2009. p. 23. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/ReportAM-MatosEtAl.pdf>. Acesso em: 30 ago. 2010.

MATOS, P. F.; LOMBARDI, L. O.; CIFERRI, R. R.; PARDO, T. A. S.; CIFERRI, C. D. A.; VIEIRA, M. T. P. **Relatório Técnico "Métricas de Avaliação"**. São Carlos: Departamento de Computação, Universidade Federal de São Carlos, 2009. p. 15. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/ReportMetrica-MatosEtAl.pdf>. Acesso em: 30 ago. 2010.

PINTO, A. C. S.; MATOS, P. F.; PERLIN, C. B.; ANDRADE, C. G.; CAROSIA, A. E. O.; LOMBARDI, L. O.; CIFERRI, R. R.; PARDO, T. A. S.; CIFERRI, C. D. A.; VIEIRA, M. T. P. **Technical Report "Sickle Cell Anemia".** São Carlos: Department of Computer Science, Federal University of São Carlos, 2009. p. 16. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/ReportSCA-PintoEtAl.pdf>. Acesso em: 30 ago. 2010.

Ademais, para dar suporte as etapas da metodologia proposta neste trabalho foram desenvolvidas três ferramentas utilizando a linguagem de programação Java, a saber:

MATOS, P. F.; CIFERRI, R. R.; PARDO, T. A. S. **SCA-TermManager:** a tool from the biomedical domain to assist the expert in term management. 2010. Software. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/SCA-TermManager.rar>. Acesso em: 30 ago. 2010.

MATOS, P. F.; CIFERRI, R. R.; PARDO, T. A. S. **SCA-Extractor:** a tool for information extraction in scientific papers of the biomedical domain. 2010. Software. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/SCA-Extractor.rar>. Acesso em: 30 ago. 2010.

MATOS, P. F.; CIFERRI, R. R.; PARDO, T. A. S. **SCA-Classifier:** a tool for sentence classification in scientific papers of the biomedical domain. 2010. Software. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/SCA-Classifier.rar>. Acesso em: 30 ago. 2010.

REFERÊNCIAS

AFANTENOS, S.; KARKALETSIS, V.; STAMATOPOULOS, P. Summarization from medical documents: a survey. **Artificial Intelligence in Medicine**, v. 33, n. 2, p. 157-177, 2005. Disponível em: <http://dx.doi.org/10.1016/j.artmed.2004.07.017>. Acesso em: 25 fev. 2010.

AGATONOVIC, M. et al. Large-scale, parallel automatic patent annotation. In: ACM WORKSHOP ON PATENT INFORMATION RETRIEVAL, 2008, Napa Valley, California. **Proceedings...** New York: ACM, 2008. p. 1-8. Disponível em: <http://doi.acm.org/10.1145/1458572.1458574>. Acesso em: 10 mar. 2010.

ANANIADOU, S.; FRIEDMAN, C.; TSUJII, J. I. (Ed.). Introduction: named entity recognition in biomedicine. **Journal of Biomedical Informatics**, v. 37, n. 6, p. 393-395, 2004. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2004.08.011>. Acesso em: 12 mar. 2010.

ANANIADOU, S.; MCNAUGHT, J. (Ed.). **Text mining for biology and biomedicine**. Norwood, MA: Artech House, 2006. 302 p.

ANANIADOU, S.; NENADIC, G. Automatic terminology management in biomedicine. In: ANANIADOU, S.; MCNAUGHT, J. (Ed.). **Text mining for biology and biomedicine**. Norwood, MA: Artech House, 2006. p. 67-98.

ARANHA, C. N. **Uma abordagem de pré-processamento automático para mineração de textos em português:** sob o enfoque da inteligência computacional. 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: <http://www.maxwell.lambda.ele.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=10081@1>. Acesso em: 19 abr. 2010.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20-29, 2004. Disponível em: <http://doi.acm.org/10.1145/1007730.1007735>. Acesso em: 18 maio 2010.

BREMER, E. G. et al. Text mining of full text articles and creation of a knowledge base for analysis of microarray data. In: LÓPEZ, J. A.; BENFENATI, E.; DUBITZKY, W. (Ed.). **Knowledge Exploration in Life Science Informatics (KELSI)**. Heidelberg: Springer, 2004. p. 84-95. (Lecture Notes in Computer Science; v. 3303). Disponível em: <http://dx.doi.org/10.1007/b103729>. Acesso em: 26 mar. 2010.

CAROSIA, A. E. O.; CIFERRI, C. D. A. **Ferramenta SCDtRanslator:** conversão do formato PDF para o formato XML aplicada ao domínio de artigos médicos sobre a Doença Anemia Falciforme. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2010. p. 40. Relatório Científico. Bolsa de Iniciação Científica – Processo 2008/10621-4. Disponível em: <http://sca.dc.ufscar.br/download/files/Report.SCDtRranslator.pdf>. Acesso em: 03 ago. 2010.

CARRILHO JUNIOR, J. R. **Desenvolvimento de uma metodologia para mineração de textos**. 2007. 96 f. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia Elétrica do Centro Técnico Científico, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: <http://www.maxwell.lambda.ele.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1>. Acesso em: 18 mar. 2010.

CHEN, H. **Knowledge management systems:** a text mining perspective. Tucson, AZ: University of Arizona, 2001. 50 p. Disponível em: <http://ai.bpa.arizona.edu/go/download/chenKMSi.pdf>. Acesso em: 14 maio 2010.

CHEUNG, C. F.; LEE, W. B.; WANG, Y. A multi-facet taxonomy system with applications in unstructured knowledge management. **Journal of Knowledge Management**, v. 9, n. 6, p. 76-91, 2005. Disponível em: <http://dx.doi.org/10.1108/13673270510629972>. Acesso em: 23 fev. 2010.

CHUN, H.-W. et al. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING (PSB), 11th, 2006, Hawaii. **Proceedings.....** 2006. p. 4-15. Disponível em: <http://psb.stanford.edu/psb-online/proceedings/psb06/chun.pdf>. Acesso em: 11 fev. 2010.

COHEN, A. M.; HERSH, W. R. A survey of current work in biomedical text mining. **Briefings in Bioinformatics**, v. 6, n. 1, p. 57-71, 2005. Disponível em: <http://dx.doi.org/10.1093/bib/6.1.57>. Acesso em: 11 fev. 2010.

COHEN, K. B.; HUNTER, L. Getting started in text mining. **PLoS Computational Biology**, v. 4, n. 1, p. 1-3, 2008. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pcbi.0040020>. Acesso em: 15 fev. 2010.

COLLIER, N.; NOBATA, C.; TSUJII, J.-I. Extracting the names of genes and gene products with a hidden Markov model. In: CONFERENCE ON COMPUTATIONAL LINGUISTICS - VOLUME 1, 18th, 2000, Saarbrücken, Germany. **Proceedings.....** Morristown, NJ: Association for Computational Linguistics, 2000. p. 201-207. Disponível em: <http://dx.doi.org/10.3115/990820.990850>. Acesso em: 11 mar. 2010.

CORNEY, D. P. A. et al. BioRAT: extracting biological information from full-length papers. **Bioinformatics**, v. 20, n. 17, p. 3206-3213, 2004. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bth386>. Acesso em: 27 fev. 2010.

CUNNINGHAM, H. Information extraction, automatic. In: KEITH, B. (Ed.). **Encyclopedia of language & linguistics**. 2nd. Oxford: Elsevier, 2006. p. 665-677. v. 5. Disponível em: <http://dx.doi.org/10.1016/B0-08-044854-2/00960-3>. Acesso em: 10 mar. 2010.

DÖRRE, J.; GERSTL, P.; SEIFFERT, R. Text mining: finding nuggets in mountains of textual data. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 5th, 1999, San Diego, California. **Proceedings.....** New York: ACM, 1999. p. 398-401. Disponível em: <http://doi.acm.org/10.1145/312129.312299>. Acesso em: 10 fev. 2010.

EGOROV, S.; YURYEV, A.; DARASELIA, N. A simple and practical dictionary-based approach for identification of proteins in MEDLINE abstracts. **Journal American Medical Informatics Association (JAMIA)**, v. 11, n. 3, p. 174-178, 2004. Disponível em: <http://dx.doi.org/10.1197/jamia.M1453>. Acesso em: 25 fev. 2010.

FAN, W. et al. Tapping the power of text mining. **Communications of the ACM**, v. 49, n. 9, p. 76-82, 2006. Disponível em: <http://doi.acm.org/10.1145/1151030.1151032>. Acesso em: 15 mar. 2010.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996. Disponível em: <http://www.aaai.org/AITopics/assets/PDF/AIMag17-03-2-article.pdf>. Acesso em: 20 mar. 2010.

FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases (KDT). In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD), 1995, Montréal, Québec. **Proceedings.....** Menlo Park, CA: AAAI Press, 1995. p. 112-117. Disponível em: <http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf>. Acesso em: 23 abr. 2010.

FELDMAN, R.; SANGER, J. **The text mining handbook:** advanced approaches in analyzing unstructured data. New York: Cambridge University Press, 2007. 391 p.

FRANZÉN, K. et al. Protein names and how to find them. **International Journal of Medical Informatics**, v. 67, n. 1-3, p. 49-61, 2002. Disponível em: <http://dx.doi.org/10.1016/S1386-5056(02)00052-7>. Acesso em: 12 mar. 2010.

FUKUDA, K. et al. Toward information extraction: identifying protein names from biological papers. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING (PSB), 3th, 1998, Hawaii. **Proceedings.....** 1998. p. 705-716. Disponível em: <http://psb.stanford.edu/psb-online/proceedings/psb98/fukuda.pdf>. Acesso em: 24 fev. 2010.

GANTZ, J. F. et al. **The expanding digital universe:** a forecast of worldwide information growth through 2010. IDC Whitepaper, 2007. Disponível em: <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf>. Acesso em: 14 maio 2010.

GARTEN, Y.; ALTMAN, R. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. **BMC Bioinformatics**, v. 10, p. S6, 2009. Suppl. 2. Disponível em: <http://dx.doi.org/10.1186/1471-2105-10-S2-S6>. Acesso em: 12 mar. 2010.

GHANEM, M. M. et al. Automatic scientific text classification using local patterns: KDD CUP 2002 (task 1). **ACM SIGKDD Explorations Newsletter**, v. 4, n. 2, p. 95-96, 2002. Disponível em: <http://doi.acm.org/10.1145/772862.772876>. Acesso em: 11 fev. 2010.

GUPTA, V.; LEHAL, G. S. A survey of text mining techniques and applications. **Journal of Emerging Technologies in Web Intelligence**, v. 1, n. 1, p. 60-76, 2009. Disponível em: <http://www.academypublisher.com/jetwi/vol1/no1/jetwi01016076.pdf>. Acesso em: 27 abr. 2010.

HALL, M. et al. The WEKA data mining software: an update. **SIGKDD Explorations**, v. 11, n. 1, p. 10-18, 2009. Disponível em: <http://www.kdd.org/explorations/issues/11-1-2009-07/p2V11n1.pdf>. Acesso em: 05 mar. 2010.

HAN, J.; KAMBER, M. **Data mining:** concepts and techniques. 2nd ed. San Francisco, CA: Morgan Kaufmann, 2006. 743 p.

HANISCH, D. et al. ProMiner: rule-based protein and gene entity recognition. **BMC Bioinformatics**, v. 6, p. S14, 2005. Suppl. 1. Disponível em: <http://dx.doi.org/10.1186/1471-2105-6-S1-S14>. Acesso em: 26 fev. 2010.

HEARST, M. A. Untangling text data mining. In: ANNUAL MEETING OF THE ASSOCIATION OF COMPUTATIONAL LINGUISTICS, 37th, 1999, College Park, Maryland. **Proceedings.....** Morristown, NJ: Association for Computational Linguistics, 1999. p. 3-10. Disponível em: <http://dx.doi.org/10.3115/1034678.1034679>. Acesso em: 23 abr. 2010.

HOTHO, A.; NÜRNBERGER, A.; PAASS, G. A brief survey of text mining. **LDV Forum - GLDV Journal for Computational Linguistics and Language Technology**, v. 20, n. 1, p. 19-62, 2005. Disponível em: <http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf>. Acesso em: 17 maio 2010.

HU, Z. Z. et al. Literature mining and database annotation of protein phosphorylation using a rule-based system. **Bioinformatics**, v. 21, n. 11, p. 2759-2765, 2005. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bti390>. Acesso em: 23 fev. 2010.

IKONOMAKIS, M.; KOTSIANTIS, S.; TAMPAKAS, V. Text classification using machine learning techniques. **WSEAS Transactions on Computers**, v. 4, n. 8, p. 966-974, 2005. Disponível em: <http://www.math.upatras.gr/~esdlab/en/members/kotsiantis/Text%20Classification%20final%20journal.pdf>. Acesso em: 13 fev. 2010.

IMAMURA, C. Y.-M. **Pré-processamento para extração de conhecimento de bases textuais**. 103 f. Dissertação (Mestrado em Ciência de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2001.

JACKSON, P.; MOULINIER, I. **Natural language processing for online applications:** text retrieval, extraction and categorization. John Benjamins, 2002. 223 p.

JENSEN, L. J.; SARIC, J.; BORK, P. Literature mining for the biologist: from information retrieval to biological discovery. **Nature Reviews Genetics**, v. 7, n. 2, p. 119-129, 2006. Disponível em: <http://dx.doi.org/10.1038/nrg1768>. Acesso em: 24 fev. 2010.

JUDE. **Jude Community**. 2010. Disponível em: <http://jude.change-vision.com/jude-web/product/community.html>. Acesso em: 06 ago. 2010.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing:** an introduction to natural language processing, computational linguistics and speech recognition. Englewood Cliffs, New Jersey: Prentice Hall, 2000. 950 p.

KAZAMA, J. I. et al. Tuning support vector machines for biomedical named entity recognition. In: WORKSHOP ON NATURAL LANGUAGE PROCESSING IN THE BIOMEDICAL DOMAIN, 2002, Phildadelphia, Pennsylvania. **Proceedings.....** Morristown, NJ: Association for Computational Linguistics, 2002. p. 1-8. Disponível em: <http://dx.doi.org/10.3115/1118149.1118150>. Acesso em: 24 mar. 2010.

KOU, Z.; COHEN, W. W.; MURPHY, R. F. High-recall protein entity recognition using a dictionary. **Bioinformatics**, v. 21, p. i266-273, 2005. Suppl. 1. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bti1006>. Acesso em: 23 mar. 2010.

KRAUTHAMMER, M.; NENADIC, G. Term identification in the biomedical literature. **Journal of Biomedical Informatics**, v. 37, n. 6, p. 512-526, 2004. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2004.08.004>. Acesso em: 25 fev. 2010.

KRAUTHAMMER, M. et al. Using BLAST for identifying gene and protein names in journal articles. **Gene**, v. 259, n. 1-2, p. 245-252, 2000. Disponível em: <http://dx.doi.org/10.1016/S0378-1119(00)00431-5>. Acesso em: 24 mar. 2010.

LEONARD, J. E.; COLOMBE, J. B.; LEVY, J. L. Finding relevant references to genes and proteins in Medline using a Bayesian approach. **Bioinformatics**, v. 18, n. 11, p. 1515-1522, Nov., 2002. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/18.11.1515>. Acesso em: 24 fev. 2010.

LUO, Q. Advancing knowledge discovery and data mining. In: INTERNATIONAL WORKSHOP ON KNOWLEDGE DISCOVERY AND DATA MINING, 2008, Adelaide, Australia. **Proceedings.....** IEEE Computer Society, 2008. p. 3-5. Disponível em: <http://dx.doi.org/10.1109/WKDD.2008.153>. Acesso em: 10 abr. 2010.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008. 482 p. Disponível em: <http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html>. Acesso em: 28 abr. 2010.

MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. Building a large annotated corpus of English: the penn treebank. **Computational Linguistics**, v. 19, n. 2, p. 313-330, 1993. Disponível em: <http://portal.acm.org/citation.cfm?id=972475#>. Acesso em: 28 abr. 2010.

MARTINS, C. A. **Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. 174 f. Tese (Doutorado em Ciência de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-08032004-164855/>. Acesso em: 09 fev. 2010.

MATHIAK, B.; ECKSTEIN, S. Five steps to text mining in biomedical literature. In: EUROPEAN WORKSHOP ON DATA MINING AND TEXT MINING IN BIOINFORMATICS, 2nd, 2004, Pisa, Italy. **Proceedings.....** 2004. p. 47-50. Disponível em: <http://www2.informatik.hu-berlin.de/Forschung_Lehre/wm/ws04/7.pdf>. Acesso em: 13 mar. 2010.

MATOS, P. F. et al. **Relatório Técnico "Conceitos sobre Aprendizado de Máquina"**. São Carlos: Departamento de Computação, Universidade Federal de São Carlos, 2009a. p. 23. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/ReportAM-MatosEtAl.pdf>. Acesso em: 03 ago. 2010.

_____. **Relatório Técnico "Métricas de Avaliação"**. São Carlos: Departamento de Computação, Universidade Federal de São Carlos, 2009b. p. 15. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/ReportMetrica-MatosEtAl.pdf>. Acesso em: 03 ago. 2010.

MCDONALD, R.; PEREIRA, F. Identifying gene and protein mentions in text using conditional random fields. **BMC Bioinformatics**, v. 6, p. S6, 2005. Suppl. 1. Disponível em: <http://dx.doi.org/10.1186/1471-2105-6-S1-S6>. Acesso em: 11 mar. 2010.

MCNAUGHT, J.; BLACK, W. J. Information extraction. In: ANANIADOU, S.; MCNAUGHT, J. (Ed.). **Text mining for biology and biomedicine**. Norwood, MA: Artech House, 2006. p. 143-178.

MIKA, S.; ROST, B. NLProt: extracting protein names and sequences from papers. **Nucleic Acids Research**, v. 32, p. 634-637, 2004a. Suppl. 2. Disponível em: <http://dx.doi.org/10.1093/nar/gkh427>. Acesso em: 25 fev. 2010.

_____. Protein names precisely peeled off free text. **Bioinformatics**, v. 20, p. i241-247, 2004b. Suppl. 1. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bth904>. Acesso em: 27 fev. 2010.

MOLLÁ, D.; VICEDO, J. L. Question answering in restricted domains: an overview. **Computational Linguistics**, v. 33, n. 1, p. 41-61, 2007. Disponível em: <http://www.ics.mq.edu.au/~diego/answerfinder/rdqa/CLQA07.pdf>. Acesso em: 16 mar. 2010.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Ed.). **Sistemas inteligentes:** fundamentos e aplicações. São Carlos: Manole, 2003. p. 89-114. cap. 4.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Linguisticae Investigationes**, v. 30, n. 1, p. 3-26, 2007. Disponível em: <http://nlp.cs.nyu.edu/sekine/papers/li07.pdf>. Acesso em: 24 fev. 2010.

NATARAJAN, J. et al. Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. **BMC Bioinformatics**, v. 7, n. 1, p. 373, 2006. Disponível em: <http://dx.doi.org/10.1186/1471-2105-7-373>. Acesso em: 16 mar. 2010.

_____. GetItFull - a tool for downloading and pre-processing full-text journal articles. In: BREMER, E. G. et al. (Ed.). **Knowledge Discovery in Life Science Literature (KDLL)**. Heidelberg: Springer, 2006. p. 139-145. (Lecture Notes in Computer Science; v. 3886). Disponível em: <http://dx.doi.org/10.1007/11683568_12>. Acesso em: 27 mar. 2010.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. **Entrez, the life sciences search engine**. 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/Entrez/>. Acesso em: 26 mar. 2010.

_____. **PubMed**. 2010. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/>. Acesso em: 16 mar. 2010.

NATIONAL LIBRARY OF MEDICINE. **Fact Sheet:** MEDLINE. 2008. Disponível em: <http://www.nlm.nih.gov/pubs/factsheets/medline.html>. Acesso em: 16 mar. 2010.

NOBATA, C.; COLLIER, N.; TSUJII, J.-I. Automatic term identification and classification in biology texts In: NATURAL LANGUAGE PACIFIC RIM SYMPOSIUM (NLPRS), 4th, 1999, Beijing, China. **Proceedings.....** 1999. p. 369-374. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.8384>. Acesso em: 26 fev. 2010.

ONO, T. et al. Automated extraction of information on protein-protein interactions from the biological literature. **Bioinformatics**, v. 17, n. 2, p. 155-161, 2001. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/17.2.155>. Acesso em: 23 mar. 2010.

PARK, J. C.; KIM, J.-J. Named entity recognition. In: ANANIADOU, S.; MCNAUGHT, J. (Ed.). **Text mining for biology and biomedicine**. Norwood, MA: Artech House, 2006. p. 121-142.

PINTO, A. C. S. et al. **Technical Report "Sickle Cell Anemia"**. São Carlos: Department of Computer Science, Federal University of São Carlos, 2009. p. 16. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/ReportSCA-PintoEtAl.pdf>. Acesso em: 03 ago. 2010.

RADEV, D. R.; HOVY, E.; MCKEOWN, K. Introduction to the special issue on summarization. **Computational Linguistics**, v. 28, n. 4, p. 399-408, 2002. Disponível em: <http://dx.doi.org/10.1162/089120102762671927>. Acesso em: 16 mar. 2010.

REBHOLZ-SCHUHMANN, D.; KIRSCH, H.; COUTO, F. Facts from text - is text mining ready to deliver? **PLoS Biology**, v. 3, n. 2, p. e65, 2005. Disponível em: <http://dx.doi.org/10.1371%2Fjournal.pbio.0030065>. Acesso em: 16 mar. 2010.

REZENDE, S. O. **Sistemas inteligentes:** fundamentos e aplicações. Manole, 2003. 525 p.

SAYERS, E. W. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 37, p. D5-15, 2009. Suppl. 1. Disponível em: <http://dx.doi.org/10.1093/nar/gkn741>. Acesso em: 12 mar. 2010.

SCHUEMIE, M. J. et al. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. **Journal of Biomedical Informatics**, v. 40,

n. 3, p. 316-324, 2007. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2006.09.002>. Acesso em: 25 fev. 2010.

_____. Distribution of information in biomedical abstracts and full-text publications. **Bioinformatics**, v. 20, n. 16, p. 2597-2604, 2004. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bth291>. Acesso em: 09 mar. 2010.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM Computing Surveys**, v. 34, n. 1, p. 1-47, 2002. Disponível em: <http://doi.acm.org/10.1145/505282.505283>. Acesso em: 17 fev. 2010.

SEKI, K.; MOSTAFA, J. An approach to protein name extraction using heuristics and a dictionary. In: ANNUAL CONFERENCE OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY (ASIST), 2003, Long Beach, CA. **Proceedings.....** 2003. p. 1-7. Disponível em: <http://www.ai.cs.kobe-u.ac.jp/~kseki/myarticles/seki2003asis.pdf>. Acesso em: 25 mar. 2010.

_____. A hybrid approach to protein name identification in biomedical texts. **Information Processing & Management**, v. 41, n. 4, p. 723-743, 2005. Disponível em: <http://dx.doi.org/10.1016/j.ipm.2004.02.006>. Acesso em: 24 mar. 2010.

SEKINE, S. **Named entity:** history and future. 2004. 5 p. Disponível em: <http://cs.nyu.edu/~sekine/papers/NEsurvey200402.pdf>. Acesso em: 23 mar. 2010.

SILVA, P. P. **ExtraWeb:** um sumarizador de documentos Web baseado em etiquetas HTML e ontologia. 158 f. Dissertação (Mestrado em Ciência de Computação) – Departamento de Ciência da Computação, Universidade Federal de São Carlos, São Carlos, 2006. Disponível em: <http://www.bdtd.ufscar.br/tde_busca/arquivo.php?codArquivo=1170>. Acesso em: 10 abr. 2010.

SPASIC, I. et al. Text mining and ontologies in biomedicine: making sense of raw text. **Briefings in Bioinformatics**, v. 6, n. 3, p. 239-251, 2005. Disponível em: <http://dx.doi.org/10.1093/bib/6.3.239>. Acesso em: 13 fev. 2010.

STAVRIANOU, A.; ANDRITSOS, P.; NICOLOYANNIS, N. Overview and semantic issues of text mining. **SIGMOD Record**, v. 36, n. 3, p. 23-34, 2007. Disponível em: <http://doi.acm.org/10.1145/1324185.1324190>. Acesso em: 23 abr. 2010.

SUN MICROSYSTEMS. **Java platform API specifications**. Disponível em: <http://java.sun.com/reference/api/>. Acesso em: 11 ago. 2010.

TAN, A.-H. Text mining: the state of the art and the challenges. In: KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES (KDAD), 1999, Beijing, China. **Proceedings.....** PAKDD, 1999. p. 71-76. Disponível em: <http://www3.ntu.edu.sg/home/asahtan/Papers/tm_pakdd99.pdf>. Acesso em: 23 abr. 2010.

TANABE, L.; WILBUR, W. J. Tagging gene and protein names in biomedical text. **Bioinformatics**, v. 18, n. 8, p. 1124-1132, 2002a. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/18.8.1124>. Acesso em: 11 fev. 2010.

_____. Tagging gene and protein names in full text articles. In: WORKSHOP ON NATURAL LANGUAGE PROCESSING IN THE BIOMEDICAL DOMAIN, 2002b, Phildadelphia, Pennsylvania. **Proceedings.....** Morristown, NJ: Association for Computational Linguistics, 2002b. p. 9-13. Disponível em: <http://dx.doi.org/10.3115/1118149.1118151>. Acesso em: 25 fev. 2010.

THE STANFORD NATURAL LANGUAGE PROCESSING GROUP. **Stanford log-linear part-of-speech tagger**. 2010. Disponível em: <http://nlp.stanford.edu/software/tagger.shtml>. Acesso em: 28 abr. 2010.

TOUTANOVA, K. et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOGY ON HUMAN LANGUAGE TECHNOLOGY (HLT-NAACL), 2003, Edmonton, Canada. **Proceedings.....** Association for Computational Linguistics, 2003. p. 173-180. Disponível em: <http://dx.doi.org/10.3115/1073445.1073478>. Acesso em: 28 abr. 2010.

TOUTANOVA, K.; MANNING, C. D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND VERY LARGE CORPORA (EMNLP/VLC), 38th, 2000, Hong Kong. **Proceedings.....** Association for Computational Linguistics, 2000. p. 63-70. Disponível em: <http://dx.doi.org/10.3115/1117794.1117802>. Acesso em: 28 abr. 2010.

TSURUOKA, Y.; TSUJII, J. I. Improving the performance of dictionary-based approaches in protein name recognition. **Journal of Biomedical Informatics**, v. 37, n. 6, p. 461-470, 2004. Disponível em: <http://dx.doi.org/10.1016/j.jbi.2004.08.003>. Acesso em: 25 fev. 2010.

VAN RIJSBERGEN, C. J. **Information retrieval**. 2nd ed. Butterworth-Heinemann, 1979. 224 p. Disponível em: <http://www.dcs.gla.ac.uk/Keith/Preface.html>. Acesso em: 10 mar. 2010.

ZHOU, G. et al. Recognition of protein/gene names from text using an ensemble of classifiers. **BMC Bioinformatics**, v. 6, p. S7, 2005. Suppl. 1. Disponível em: <http://dx.doi.org/10.1186/1471-2105-6-S1-S7>. Acesso em: 11 mar. 2010.

_____. Recognizing names in biomedical texts: a machine learning approach. **Bioinformatics**, v. 20, n. 7, p. 1178-1190, 2004. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bth060>. Acesso em: 25 fev. 2010.

ZWEIGENBAUM, P. et al. Frontiers of biomedical text mining: current progress. **Briefings in Bioinformatics**, v. 8, n. 5, p. 358-375, 2007. Disponível em: <http://dx.doi.org/doi:10.1093/bib/bbm045>. Acesso em: 13 fev. 2010.

GLOSSÁRIO

***Application Programming Interface* (API):** É uma interface de programação de aplicativos que oferece um conjunto de rotinas e padrões estabelecidos.

Apositivo: Frase que altera e completa a palavra ou frase.

Baseline: É um ponto de partida que serve como limite inferior de referência.

Biomédico: O termo biomédico refere-se ao estudo da ligação entre a química e o funcionamento do corpo humano. A ciência da biomedicina conduz estudos nas áreas da medicina e biologia que são direcionadas a pesquisa das doenças humanas, com intuito de encontrar causa, prevenção, diagnóstico e tratamento.

Farmacogenética: É o estudo das variações hereditárias em resposta a drogas e/ ou alterações metabólicas.

Farmacogenômica: Pode ser caracterizada como um estudo generalista de todos os genes que podem afetar o comportamento das drogas.

Gazetteers: Dicionário de termos.

Gold Standard: Limite máximo de referência.

***Medical Subject Headings* (MeSH):** Amplo vocabulário controlado para publicações de artigos e livros na ciência.

***National Center for Biotechnology Information* (NCBI)**: Instituto Nacional de Saúde criado em 1988 para desenvolver sistemas de informação para biologia molecular. Fornece sistemas de recuperação para pesquisar nas bases do PubMed e PubMed Central (PMC) (http://www.pubmedcentral.nih.gov/).

Ruído e Silêncio: É a informação desnecessária selecionada para treinamento que aumenta o erro do classificador, diferentemente de Silêncio que é a característica importante que não foi selecionada para treinamento.

Sentença e Frase: Sentença é uma palavra ou conjunto de palavras que constituem um enunciado de sentido completo. O ponto final delimita uma sentença. Frase é uma unidade bem menor do que a sentença. Pode ser sintagma (nominal, verbal, adverbial, etc.).

Termo Curado: É um termo não curado que foi definido (i.e., validado) pelo especialista como um termo curado. Isto significa que o termo curado passou pelo crivo do especialista e, portanto, é realmente um termo importante. Todo termo extraído automaticamente (i.e., pela abordagem de regra) é inserido no banco de dados como termo não curado. O dicionário utiliza para extrair informação somente o termo que foi curado.

APÊNDICE A – ESQUEMA CONCEITUAL EER

Na Figura 58 é mostrado o esquema conceitual do banco de dados (sem os atributos) desenvolvido e utilizado neste trabalho de mestrado. Na Figura 59 e na Figura 60 são mostrados, respectivamente, o tipo-relacionamento binário entre os efeitos negativos e o artigo, e o tipo-relacionamento binário entre o efeito positivo e o artigo.

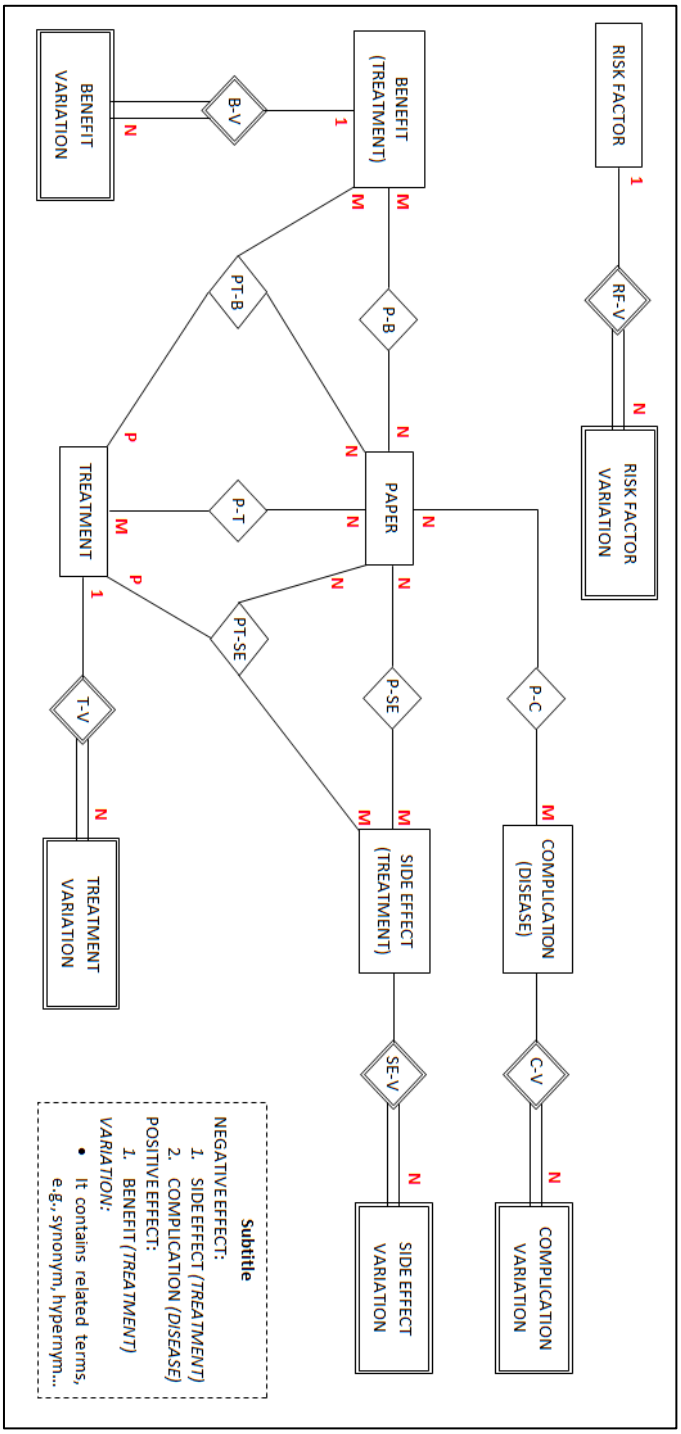

Figura 58 – Esquema conceitual sem atributos.

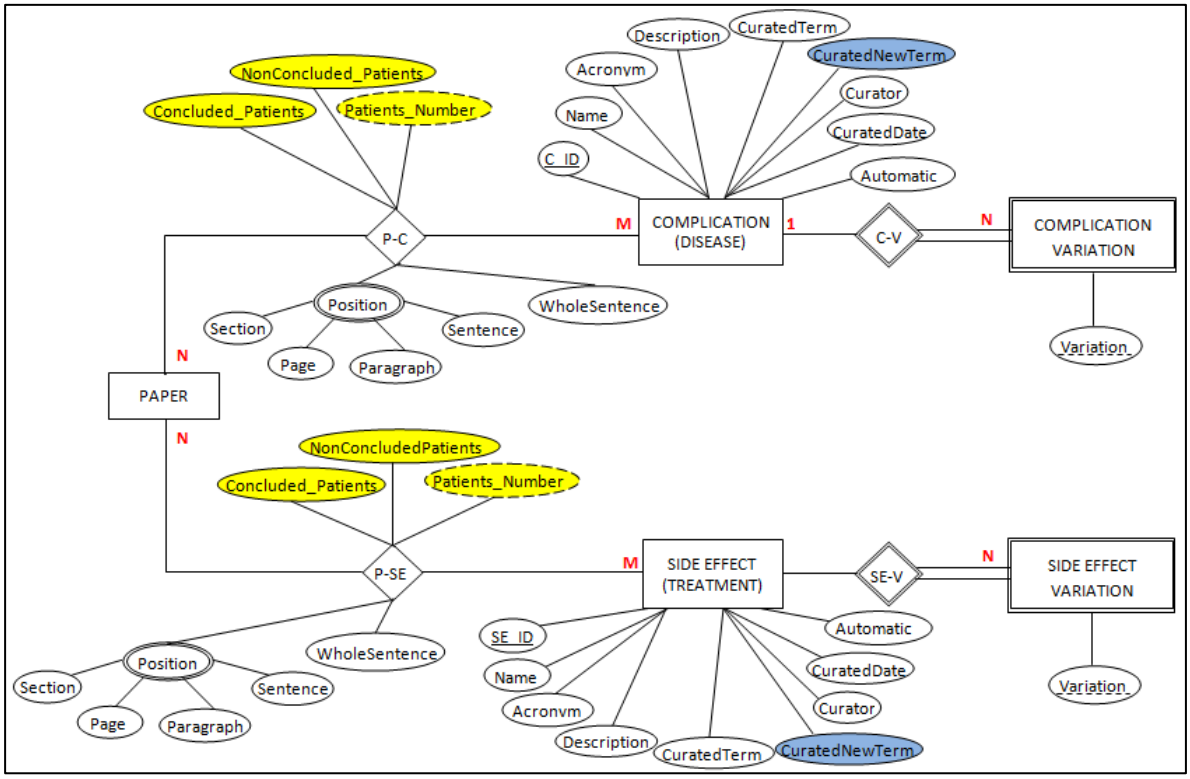

Figura 59 – Esquema conceitual dos tipos entidade "efeitos negativos" e do tipo entidade "artigo" com atributos.

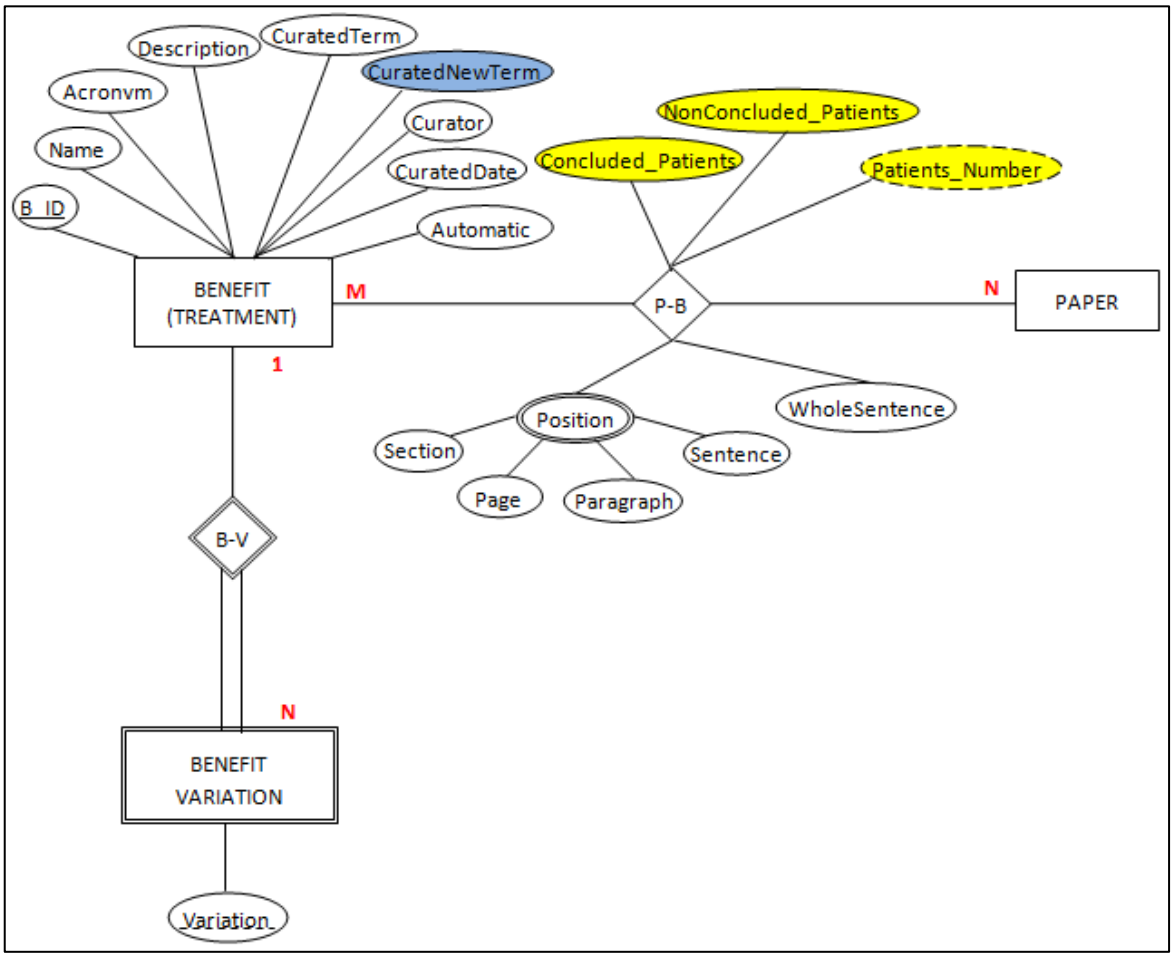

Figura 60 – Esquema conceitual do tipo entidade "efeito positivo" e do tipo entidade "artigo" com atributos.

APÊNDICE B – ESQUEMA LÓGICO RELACIONAL

A seguir encontra-se o modelo relacional referente ao esquema conceitual da Figura 58 do APÊNDICE A – ESQUEMA CONCEITUAL EER.

PAPER (Paper_ID[1], Title[2], Journal[2], Year[2], Authors, PDF, HTML, XML, Total_Patients[3], SC_Patients, UC_Patients, Paper_Curated, Paper_Curator, Paper_CuratedDate, Paper_Automatic)
SC_Patients = Satisfactory Concluded Patients, UC_Patients = Unsatisfactory Concluded Patients

COMPLICATION (Complication_ID[1], Complication_Name[4], Complication_Acronym, Complication_Desc, Complication_Curated, Complication _Curator, Complication _CuratedDate, Complication_Automatic)
PAPER_COMPLICATION (Paper_ID, Complication_ID, Section, Page, Paragraph, Sentence, WholeSentence, Concluded_Patients, NonConcluded_Patients, Patients_Number[3])
COMPLICATION_VARIATION (Complication_ID, Complication_Variation)

SIDE_EFFECT (SE_ID[1], SE_Name[4], SE_Acronym, SE_Desc, SE_Curated, SE_Curator, SE_CuratedDate, SE_Automatic)
PAPER_SIDE_EFFECT (Paper_ID, SE_ID, Section, Page, Paragraph, Sentence, WholeSentence, Concluded_Patients, NonConcluded_Patients, Patients_Number[3])
SIDE_EFFECT _VARIATION (SE_ID, SE_Variation)

BENEFIT (Benefit_ID[1], Benefit_Name[4], Benefit_Acronym, Benefit_Desc, Benefit_Curated, Benefit_Curator, Benefit_CuratedDate, Benefit_Automatic)
PAPER_BENEFIT (Paper_ID, Benefit_ID, Section, Page, Paragraph, Sentence, WholeSentence, Concluded_Patients, NonConcluded_Patients, Patients_Number[3])
BENEFIT _VARIATION (Benefit_ID, Benefit_Variation)

RISK_FACTOR (Risk_ID[1], Risk_Name[4], Risk_Acronym, Risk_Desc, Risk_Curated, Risk_Curator, Risk_CuratedDate, Risk_Automatic)
RISK_FACTOR_VARIATION (Risk_ID, Risk_Variation)

[1] Atributo **auto_increment**.
[2] Verificar inserção única de artigos (Title, Journal, Year).
[3] Atributo derivado.
[4] Atributo **unique**.

APÊNDICE C – EFEITOS NEGATIVOS CURADOS

Na Tabela 38 e na Tabela 39 mostram, respectivamente, os efeitos negativos da doença (complicação) e os efeitos negativos do tratamento (efeito colateral) que foram previamente cadastrados no dicionário pelo especialista.

Tabela 38 – Termos sobre complicação e suas variações curados pelo especialista.

Número #	Termos sobre Complicação	Variações
1	*acute chest syndrome*	*chest pain* *respiratory failure* *pulmonary failure* *pulmonary insufficiency* *respiratory anomalies* *shortness of breath* *tachypnea* *wheezing* *dyspnea*
2	*severe anemia*	*acute anemia* *chronic anemia* *worsening of anemia*
3	*cerebral vascular accident*	*stroke* *cerebrovascular* *cerebrovascular event* *seizure* *cognitive problem* *neurologic event* *neurologic complication* *neurologic problem* *neurological event* *neurological complication* *neurological problem*
4	*dysuria*	*nocturnal enuresis* *hematuria* *hipostenuria* *proteinuria*
5	*renal failure*	*chronic renal failure* *renal insufficiency*
6	*infection*	*infection for bacteria* *infection for legionella* *infection for mycoplasma* *infection for virus* *bacterial infection* *legionella infection* *mycoplasma infection* *virus infection* *sepsis* *parvovirus*

7	*pain*	*painful episode* *pain crises* *pain crisis*
8	*vasoocclusive*	*vasoocclusive pain episode* *vasoocclusive crises* *vasoocclusive crisis* *vascular occlusion* *vaso-occlusive* *voe*
9	*weakness*	*exhaustion* *fatigue* *reluctance*
10	*hemorrhage*	*central nervous system hemorrhage* *intracranial hemorrhage*
11	*infarct*	*infarction* *myocardial infarction* *pulmonary infarction*
12	*hypertension*	*pulmonary hypertension*
13	*splenic sequestration*	*acute splenic sequestration* *hypersplenism* *loss of spleen* *splenomegaly*
14	*stenoses*	*stenosis* *arterial stenosis* *arterial stenoses*
15	*cholelithiases*	*cholestatic syndrome* *cholelithiasis*
16	*aplastic crisis*	*aplastic crises*
17	*eye problem*	*sickle cell retinopathy* *retinopathy* *icterus* *jaundice* *retinal detachment*
18	*cardiac ischemia*	*cardiac abnormalities* *cardiac enlargement* *cardiac insufficiency* *cor pulmonale*
19	*ischemic bone necrosis*	*osteonecrosis*
20	*dactylitis*	*swelling*
21	*hepatic sequestration*	*acute hepatic sequestration* *hepatic insufficiency* *anoxic brain injury*
22	*hypoxia*	----------[1]
23	*nocturnal apnea*	----------[1]
24	*cough*	----------[1]
25	*delayed puberty in children*	----------[1]

26	*fever*	----------[1]
27	*headache*	----------[1]
28	*infiltrate in lung*	----------[1]
29	*inflammation*	----------[1]
30	*lack of calcification in bone*	----------[1]
31	*leg ulcer*	----------[1]
32	*neuromuscular abnormalities*	----------[1]
33	*pallor*	----------[1]
34	*phospholipase a2 elevation*	----------[1]
35	*priapism*	----------[1]
36	*pneumonia*	----------[1]
37	*fat embolism*	----------[1]
38	*vertigo*	----------[1]
[1] Significa que não tem nenhuma variação associada ao termo.		

Tabela 39 – Termos sobre efeito colateral e suas variações curados pelo especialista.

Número #	Termos sobre Efeito Colateral	Variações
1	*alloimmunization*	*red cell alloimmunization*
2	*death*	*died*
3	*hepatitis*	*hepatitis b* *hepatitis c* *hepatitides*
4	*cancer*	*acute promyelocytic leukemia* *leukemia* *lymphoma*
5	*toxicity*	*renal toxicity* *hepatic toxicity*
6	*hemochromatosis*	*iron overload*
7	*bacteremia*	----------[1]
8	*blood hypotension*	----------[1]
9	*depression*	----------[1]
10	*high WBC*	----------[1]
11	*human immunodeficiency virus*	----------[1]
12	*hyperviscosity*	----------[1]
13	*hypoventilation*	----------[1]
14	*iatrogenic overhydration*	----------[1]
15	*immunosuppression*	----------[1]
16	*infertility*	----------[1]

17	*neutropenia*	----------[1]
18	*thrombocytopenia*	----------[1]
19	*transfusion-transmitted infection*	----------[1]
[1] Significa que não tem nenhuma variação associada ao termo.		

APÊNDICE D – EXPRESSÕES REGULARES DA ESTRATÉGIA 1

A seguir são apresentadas as expressões regulares desenvolvidas para a Estratégia 1 (Verbo e Expressão com POS). Na Tabela 40 as expressões regulares em alto nível e em Java são destacadas para cada verbo e expressão composta representativos. Na Tabela 41 são apresentados os padrões POS.

As expressões regulares dos padrões apresentados a seguir podem ser testadas por meio de um testador de expressões regulares *on-line* de fácil uso. O testador pode ser acessado nesta URL: http://www.piazinho.com.br/exemplos.html#1. Para que as expressões regulares funcionem corretamente neste testador, é necessário adaptar a sintaxe da expressão regular para a sintaxe aceita pelo testador. Nas expressões regulares das tabelas a seguir, somente é necessário trocar todas as duas barras invertidas (“\\”) por somente uma barra invertida (”\”). Na Figura 61 é apresentado um exemplo de uma sentença etiquetada, sendo casada com o verbo “*to document*” da Tabela 40.

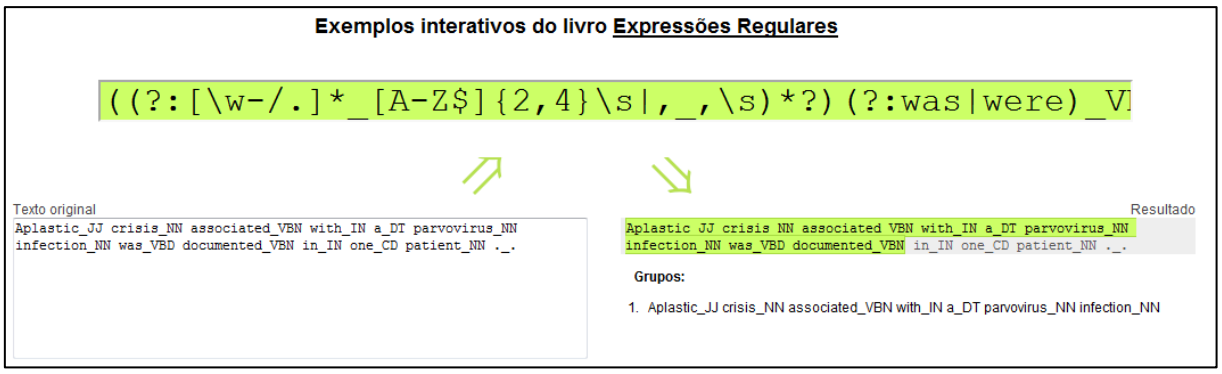

Figura 61 – Exemplo de expressão regular do verbo “*to document*” da Tabela 40.

Na Tabela 40 são apresentadas as expressões regulares dos verbos representativos e expressões representativas com a sintaxe da linguagem de programação Java. O atalho “\\w” na expressão regular de cada padrão significa que a expressão regular irá casar com letras, dígitos ou *underline* (i.e., “_”). A letra “X” nas expressões regulares corresponde a “Parte Específica” da sentença delimitada pelo verbo ou expressão. Os padrões POS da Tabela 41 são aplicados nesta parte específica, a fim de identificar um termo candidato.

Tabela 40 – Expressões regulares dos verbos e expressões compostas da Estratégia 1.

Verbo ou Expressão Composta	Expressão Regular em Alto Nível	Expressão Regular em Java
to observe, *to diagnose*, *to document* na voz passiva (termo antes)	X (was|were)_VBD (.*) (documented_VBN |diagnosed_VBN |observed_[A-Z]{2,3})	((?:[\\w-/.]*_[A-Z$]{2,4}\\s|,_,\\s)*?)(?:was|were)_VBD\\s(?:[\\w-/]*_[A-Z]{2,3}\\s)*?(?:documented_VBN|diagnosed_VBN|observed_[A-

	(of_IN X)?	Z]{2,3})(?:\\sof_IN\\s((?:[\\w-/]*_[A-Z]{2,3}\\s)*))?
to develop na voz ativa (termo depois)	developed_[A-Z]{2,3} (~IN |.) X (had_VBD |soon_RB |during_IN |.|,|:|;)	developed_[A-Z]{2,3}\\s(?![\\w-/]*_IN\\s|\\._)((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:had_VBD\\s|soon_RB\\s|during_IN\\s|\\.|,|:|;)
to develop na voz ativa (termo antes)	X developed_[A-Z]{2,3} (IN|.|,|:|;) (.* with_IN X (.|,|:|;))?	((?:[\\w-/]*_[A-Z]{2,3}\\s|,_,\\s)*)developed_[A-Z]{2,3}\\s?(?:\\s[\\w-/]*_IN|\\.|,|:|;)\\s?(?:(?:[\\w-/]*_[A-Z]{2,3}\\s)*?(?:with_IN\\s)((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:\\.|,|:|;))?
to occur na voz ativa (termo antes)	X (occurred_VBD|did_VBD occur_VB)	((?:[\\w-/]*_[A-Z]{2,3}\\s|,_,\\s)*)(?:occurred_VBD\\s|did_VBD occur_VB\\s)
prevalence of, diagnosis of e *was associated with* (termo depois)	prevalence_NN of_IN X (VB[DPZ]) diagnosis_NN of_IN X with_IN evidence_NN of_IN was_VBD associated_VBN with_IN X (CC X|.|,|:|;)	prevalence of ([\\w\\s-/]*?)(?:has\\s|is\\s) diagnosis of ([\\w\\s-/]*?)with evidence of was associated with ([\\w\\s-/]*?)\\s?(?:and([\\w\\s-/]*)|\\.|,|:|;)
at risk of (termo depois)	at risk of X (and |on |.|,|:|;)	at_IN risk_NN of_IN ((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:[\\w-/]*_CC\\s|[\\w-/]*_IN\\s|\\.|,|:|;)
to stop, to interrupt e *to initiate* na voz passiva, *to undergo* na voz ativa e *because of* (termo depois)	**if** primeira **then** segunda primeira: ((was|were)_VBD (stopped_VB[DN] |interrupted_VB[DN] |initiated_VB[DN])|underwent_VBD) segunda: (because_IN of_IN |for_IN) X (and_CC|.|,|:|;)	**if** primeira **then** segunda primeira: (?=(?:was|were)_VBD\\s((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:stopped_VB[DN]|interrupted_VB[DN]|initiated_VB[DN])|underwent_VBD) segunda:(?:because_IN of_IN|for_IN)\\s((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:and_CC|\\.|,|:|;|-LRB-|-RRB-)
caused by (termo antes e depois)	X caused_[A-Z]{2,3} by_IN X (VB[DPZ] |.|,|:|;)	((?:[\\w-/]*_[A-Z]{2,3}\\s|,_,\\s)*)caused_[A-Z]{2,3} by_IN ((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:[\\w-/]*_VB[DPZ]\\s|\\.|,|:|;)
died of e *died from* (termo depois)	died_[A-Z]{2,3} (.*)? from_IN X (.|,|:|;) died_[A-Z]{2,3} of_IN X (.|,|:|;)	died_[A-Z]{2,3} of_IN ((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:\\.|,|:|;) died_[A-Z]{2,3} (?:(?:[\\w-/]*_[A-Z]{2,3}\\s)*?)?from_IN ((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:\\.|,|:|;)
to have na voz ativa (termo depois)	(had_VBD|had_VBD had_VBN) (~VBN) X (.|,|:|;|( )	(?:had_VBD|had_VBD had_VBN)\\s(?![\\w-/]*_VBN)((?:[\\w-/]*_[A-Z]{2,3}\\s)*?)(?:\\.|,|:|;|-LRB)

Na Tabela 41 são apresentadas as expressões regulares dos padrões POS da Estratégia 1 com a sintaxe da linguagem de programação Java. O atalho "\\w" na expressão regular de cada padrão significa que a expressão regular irá casar com letras, dígitos ou *underline* (i.e., "_"). Estes padrões são aplicados na "Parte Específica" delimitada pelo verbo e expressão composta, a fim de identificar um termo candidato.

Tabela 41 – Expressões regulares dos padrões POS da Estratégia 1.

Número	Padrão	Expressão Regular em Java
1.0[1]	**(JJ_JJ_NN_NN_(NN)?)**	([\\w-/]*_JJ[RS]?\\s[\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[PS]?\\s[\\w-/]*_NN[PS]?(\\s[\\w-/]*_NN[PS]?)?)
1.1[1]	(~JJ)_**(JJ_NN_NN_(NN)?)**	(?:[\\w-/]*_(?:[^J]..?\\s|J[^J].?\\s)|^)([\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[PS]?\\s[\\w-/]*_NN[PS]?\\s(?:[\\w-/]*_NN[PS]?)?)
1.2[1]	**(JJ_JJ_NN)**_(~NN)	([\\w-/]*_JJ[RS]?\\s[\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[PS]?)\\s(?:\\._|,_|:_|;_|(?![\\w-/]*_NN.?))
1.3	(~JJ)_**(JJ_NN)**_(~NN)	(?:[\\w-/]*_(?:[^J]..?\\s|J[^J].?\\s)|,\\s|:\\s|;\\s|-LRB-\\s|-RRB-\\s|^)([\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[PS]?)\\s(?:\\._|,_|:_|;_|(?![\\w-/]*_NN.?))
1.4	((~NN)&(~JJ))_**(NN_NN)**_((~NN)&(~JJ))	(?:[\\w-/]*_(?:[^JN]..?\\s|[JN][^JN].?\\s)|,\\s|:\\s|;\\s|-LRB-\\s|-RRB-\\s|^)([\\w-/]*_NN[PS]?\\s[\\w-/]*_NN[PS]?)\\s(?:-LRB-|-RRB-|\\._|,_|:_|;_|[\\w-/]*_(?:[^JN]..?\\s|[JN][^JN].?\\s))
1.5	((~NN)&(~JJ))_**(NN)**_(~NN)	(?:[\\w-/]*_(?:[^JN]..?\\s|[JN][^JN].?\\s)|,\\s|:\\s|;\\s|-LRB-\\s|-RRB-\\s|^)([\\w-/]*_NN[PS]?)\\s(?:\\._|,_|:_|;_|[\\w-/]*_(?:[^N]..?\\s|[N][^N].?\\s))

[1] Padrão também utilizado na Estratégia 2.

APÊNDICE E – EXPRESSÕES REGULARES DA ESTRATÉGIA 2

A seguir são apresentadas as expressões regulares desenvolvidas para a Estratégia 2 (POS). Na Tabela 42 são apresentados os padrões POS.

As expressões regulares dos padrões apresentados a seguir podem ser testadas por meio de um testador de expressões regulares *on-line* de fácil uso. O testador pode ser acessado nesta URL: http://www.piazinho.com.br/exemplos.html#1. Para que as expressões regulares funcionem corretamente neste testador, é necessário adaptar a sintaxe da expressão regular para a sintaxe aceita pelo testador. Nas expressões regulares dos padrões apresentadas na Tabela 42, somente é necessário trocar todas as duas barras invertidas (“\\”) por somente uma barra invertida (”\”). Na Figura 62 é apresentado um exemplo de uma sentença etiquetada, sendo casada com o padrão 1.2 da Tabela 42.

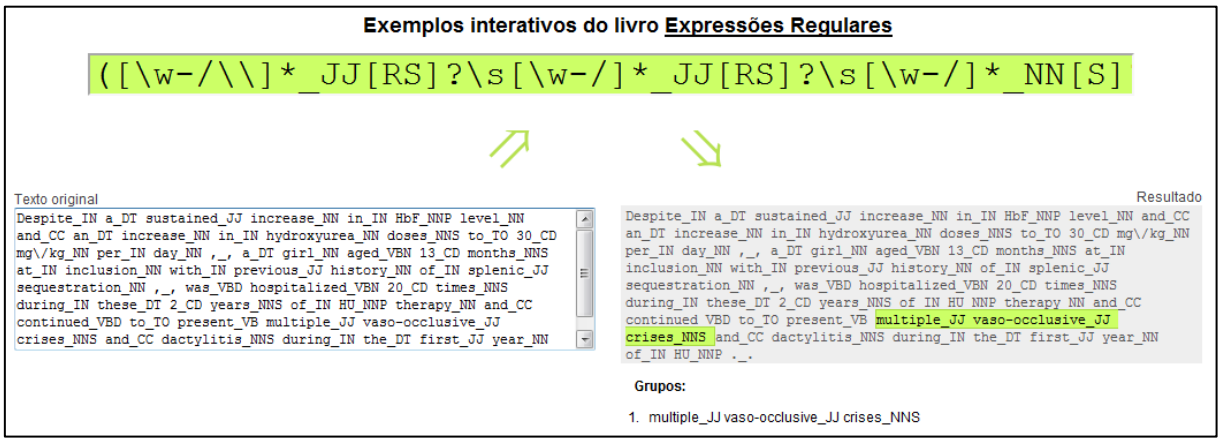

Figura 62 – Exemplo do padrão 1.2 da Tabela 42.

Na Tabela 42 são apresentadas as expressões regulares dos padrões POS da Estratégia 2 com a sintaxe da linguagem de programação Java. O atalho “\\w” na expressão regular de cada padrão significa que a expressão regular irá casar com letras, dígitos ou *underline* (i.e., “_”). Estes padrões são utilizados para identificar um termo candidato na sentença.

Tabela 42 – Expressões regulares dos padrões POS da Estratégia 2.

Número	Padrão	Expressão Regular em Java
1.0[1]	**(JJ_JJ_NN_NN_(NN)?)**	([\\w-/]*_JJ[RS]?\\s[\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?\\s[\\w-/]*_NN[S]?(\\s[\\w-/]*_NN[S]?)?)
1.1[1]	(~JJ)_**(JJ_NN_NN_(NN)?)**	(?:[\\w-/]*_(?:[^J]..?\\s|J[^J].?\\s)|^)([\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?\\s[\\w-/]*_NN[S]?\\s(?:[\\w-/]*_NN[S]?)?)
1.2[1]	**(JJ_JJ_NN)**_(~NN)	([\\w-/]*_JJ[RS]?\\s[\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?)\\s(?:\\._|,_|:_|;_|(?![\\w-/]*_NN.?))

2.0	(~JJ)_**(JJ_NN_IN_JJ_NN)**_(~NN)	Observação: Descartar prefixo JJ_NN_IN. (?:[\\w-/]*_(?:[^J].?.?\\s|J[^J].?\\s)|^)([\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?\\s[\\w-/]*_IN\\s[\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?)(?:\\s[\\w-/]*_(?:[^N]..?|N[^N].?)|\\s\\._|\\s,_|\\s:_|\\s;_|$)
2.1	**(**(~JJ)_**NN_IN)_(JJ_NN)**_(~NN)	(?:(?:[\\w-/]*_(?:[^J].?.?\\s|J[^J].?\\s)|^)[\\w-/]*_NN[S]?\\s[\\w-/]*_IN)(\\s[\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?)(?:\\s[\\w-/]*_(?:[^N]..?|N[^N].?)|\\s\\._|\\s,_|\\s:_|\\s;_|$)
3.0	(~JJ)_**(JJ_NN)_(IN_NN_NN_NN)**	(?:[\\w-/]*_(?:[^J].?.?|J[^J].?)\\s)([\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?)(?:\\s[\\w-/]*_IN\\s[\\w-/]*_NN[S]?\\s[\\w-/]*_NN[S]?\\s[\\w-/]*_NN[S]?)
3.1	(~JJ)_**(JJ_NN_IN_NN_NN)**_(~NN)	(?:[\\w-/]*_(?:[^J]..?|J[^J].?)\\s)([\\w-/]*_JJ[RS]?\\s[\\w-/]*_NN[S]?\\s[\\w-/]*_IN\\s[\\w-/]*_NN[S]?\\s[\\w-/]*_NN[S]?)(?:\\s[\\w-/]*_(?:[^N]..?|N[^N].?)|\\s\\._|\\s,_|\\s:_|\\s;_|$)
3.2	**(**(~JJ)_**JJ_NN_IN)_(NN)**_(~NN)	(?:(?:(?:[\\w-/]*_(?:[^J]..?|J[^J].?)\\s)[\\w-/]*_JJ[RS]?\\s)[\\w-/]*_NN[S]?\\s[\\w-/]*_IN)(\\s[\\w-/]*_NN[S]?)(?:\\s[\\w-/]*_(?:[^N]..?|N[^N].?)|\\s\\._|\\s,_|\\s:_|\\s;_|$)
[1] Padrão também utilizado na Estratégia 1.		

APÊNDICE F – DIAGRAMA DE CLASSES

Na Figura 63 são mostradas as principais classes da ferramenta SCA-Extractor.

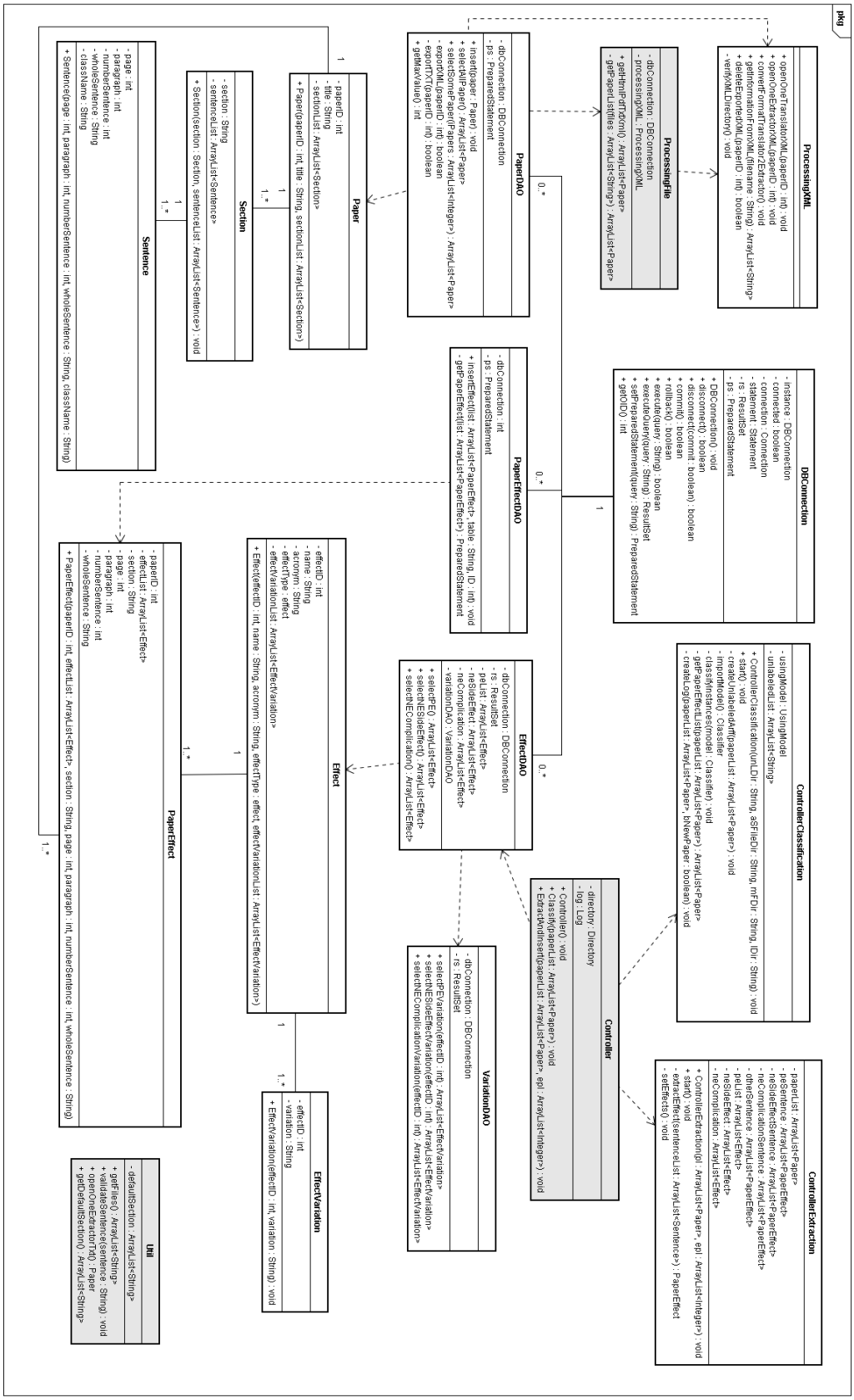

Figura 63 – Diagrama de classes da ferramenta SCA-Extractor.

APÊNDICE G – ENTRADA DE DADOS

Na Tabela 43 são listados os 10 artigos por título e ano de publicação que foram utilizados nos estudos de caso apresentados e discutidos no Capítulo 7. Estes 10 artigos contêm 901 sentenças identificadas nas seções *abstract*, *results* e *discussion*, e correspondem à entrada de dados da metodologia. Os artigos foram enumerados por um número cardinal qualquer. Estes artigos estão disponíveis nos formatos PDF, HTML e TXT e podem ser baixados na seguinte URL: http://gbd.dc.ufscar.br/~pablofmatos/files/sca-10-papers.rar.

Tabela 43 – Entrada de dados de 10 artigos científicos contendo ao todo 901 sentenças das seções *abstract*, *results* e *discussion*.

Número Artigo #	Título do Artigo	Ano
1	*Hydroxyurea for sickle cell disease in children and for prevention of cerebrovascular events: the Belgian experience*	2005
2	*Effect of hydroxyurea on the frequency of painful crises in sickle cell anemia*	1995
3	*Long-term hydroxyurea therapy for infants with sickle cell anemia: the HUSOFT extension study*	2005
5	*Effects of hydroxyurea on the membrane of erythrocytes and platelets in sickle cell anemia*	2004
6	*Stroke Prevention Trial in Sickle Cell Anemia (STOP): extended follow-up and final results*	2006
8	*Long-term hydroxyurea treatment in children with sickle cell disease: tolerance and clinical outcomes*	2006
11	*Causes and outcomes of the acute chest syndrome in sickle cell disease*	2000
26	*Asthma and acute chest in sickle-cell disease*	2004
27	*Temporal relationship of asthma to acute chest syndrome in sickle cell disease*	2007
29	*Physician-Diagnosed asthma and acute chest syndrome: associations with NOS polymorphisms*	2007