UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Um Processo Baseado em Parágrafos para a Extração de Tratamentos de Artigos Científicos do Domínio Biomédico

JULIANA LILIAN DUQUE

São Carlos - SP
Fevereiro/2012

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JULIANA LILIAN DUQUE

Um Processo Baseado em Parágrafos para a Extração de Tratamentos de 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: Prof. Dr. Ricardo Rodrigues Ciferri
Coorientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo

São Carlos - SP
Fevereiro/2012

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

D946pb

Duque, Juliana Lilian.
Um processo baseado em parágrafos para a extração de tratamentos de artigos científicos do domínio biomédico / Juliana Lilian Duque. -- São Carlos : UFSCar, 2012.
121 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2012.

1. Inteligência artificial. 2. Banco de dados. 3. Mineração de textos. 4. Reconhecimento de padrões. 5. Extração de informação. 6. Anemia falciforme. I. Título.

CDD: 006.3 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"Um Processo Baseado em Parágrafos para a Extração de Tratamentos de Artigos Científicos do Domínio Biomédico"

JULIANA LILIAN DUQUE

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)

Prof. Dr. Thiago Alexandre Salgueiro Pardo
(Co-orientador – ICMC/USP)

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

Prof. Dr. João Eduardo Ferreira
(IME/USP)

São Carlos
Fevereiro/2012

À Deus, por iluminar meus passos e pelas bênçãos realizadas em minha vida.

À minha família que se faz presente, mesmo distante.

AGRADECIMENTO

À Deus em primeiro lugar, pela força e por permitir concluir este trabalho.

Ao meu pai querido, que não está mais entre nós, que só deixou saudade e um grande vazio na minha vida, mas sei que sente muito orgulho de mim.

À minha mãe Fátima, meu padrasto Júnior, minha filha Caroline e minhas irmãs Kátia e Mariana pela paciência e compreensão em todos os momentos em que estive ausente.

Ao meu namorado Eduardo pela paciência, apoio e pelas palavras de conforto.

À minha avó Maria pelas orações direcionadas a conclusão deste projeto.

A todos os meus amigos de mestrado, em especial ao meu amigo-irmão Pablo Matos que foi uma pessoa indispensável para a conclusão deste trabalho; ao Arthur Emanuel, Alexandre Bellini, Carlos Eduardo Cirilo, Maísa Duarte e Gilmar Favarin.

Aos meus amigos Reinaldo Lima, Leonardo França, Ricardo Alves, Ricardo Scheicher e Matheus Fontes que me ajudaram ao longo do Mestrado.

Ao Grupo Provac e ao Universo Online pela liberação para cursar o Mestrado.

Ao professor e orientador Prof. Dr. Ricardo Rodrigues Ciferri pelos ensinamentos, compreensão, críticas, incentivos e confiança depositada em mim.

Ao professor Prof. Dr. Mauro Biajiz, em memória, pela oportunidade de cursar o Mestrado na Universidade Federal de São Carlos, um sonho concretizado.

Ao docente Prof. Dr. Thiago Alexandre Salgueiro Pardo que me coorientou com dedicação, sempre tão solícito e compreensivo.

À docente Prof.ª. Drª. Cristina Dutra de Aguiar Ciferri pelas correções, dicas e apoio nas reuniões e na escrita de artigos.

Por fim, a todos os amigos e familiares que me apoiaram nesta longa caminhada.

"Eu sei que não sou nada e que talvez nunca tenha tudo. Aparte isso, eu tenho em mim todos os sonhos do mundo."

(Fernando Pessoa)

"Os dias prósperos não vêm por acaso, nascem de muita fadiga e persistência."
(Henri Ford)

"Noventa por cento do sucesso se baseia simplesmente em insistir."
(Woody Allen)

RESUMO

Atualmente na área médica existe uma grande quantidade de informações não estruturadas (i.e., em formato textual) sendo produzidas na literatura médica. Com o grande volume de dados, torna-se impossível que os médicos e especialistas da área analisem toda a literatura de forma manual, exigindo técnicas para automatizar a análise destes documentos. Com o intuito de identificar as informações relevantes, estruturar e armazenar estas informações em um banco de dados, para posteriormente identificar relacionamentos interessantes entre as informações extraídas, nesta dissertação é proposto um processo baseado em parágrafos para a extração de tratamentos de artigos científicos do domínio biomédico. A hipótese é que a busca inicial de sentenças que possuem termos de complicação melhora a eficiência na identificação e na extração de termos de tratamento. Isso acontece porque tratamentos ocorrem principalmente na mesma sentença de complicação ou em sentenças próximas no mesmo parágrafo. Esta metodologia utiliza três abordagens de extração de informação encontradas na literatura: abordagem baseada em aprendizado de máquina para classificar as sentenças de interesse; abordagem baseada em dicionário com termos validados pelo especialista da área e abordagem baseada em regras. A metodologia foi validada como prova de conceito, utilizando artigos do domínio biomédico, mais especificamente da doença Anemia Falciforme. A prova de conceito foi realizada na classificação de sentenças e identificação de termos relevantes. O valor da acurácia obtida na classificação de sentenças foi de 79% para o classificador de complicação e 71% para o classificador de tratamento. Estes valores condizem com os resultados obtidos com a combinação do algoritmo de aprendizado de máquina *Support Vector Machine* juntamente com a aplicação do filtro Remoção de Ruído e Balanceamento das Classes. Na identificação de termos relevantes, os resultados da metodologia proposta obteve percentual superior de 42% de medida-F comparado à classificação manual (31%) e comparado ao processo parcial, ou seja, sem utilizar o classificador de complicação (36%). Mesmo com a baixa revocação, foi possível obter 100% de revocação para os termos distintos de tratamento, não impactando o processo de extração, e portanto a hipótese considerada neste trabalho foi comprovada.

Palavras-chave: Extração de Informação, Tratamentos, Mineração de Textos, Pré-Processamento, Domínio Biomédico, Doença Anemia Falciforme.

ABSTRACT

Currently in the medical field there is a large amount of unstructured information (i.e., in textual format). Regarding the large volume of data, it makes it impossible for doctors and specialists to analyze manually all the relevant literature, which requires techniques for automatically analyze the documents. In order to identify relevant information, as well as to structure and store them into a database and to enable future discovery of significant relationships, in this paper we propose a paragraph-based process to extract treatments from scientific papers in the biomedical domain. The hypothesis is that the initial search for sentences that have terms of complication improves the identification and extraction of terms of treatment. This happens because treatments mainly occur in the same sentence of a complication, or in nearby sentences in the same paragraph. Our methodology employs three approaches for information extraction: machine learning-based approach, for classifying sentences of interest that will have terms to be extracted; dictionary-based approach, which uses terms validated by an expert in the field; and rule-based approach. The methodology was validated as proof of concept, using papers from the biomedical domain, specifically, papers related to Sickle Cell Anemia disease. The proof of concept was performed in the classification of sentences and identification of relevant terms. The value obtained in the classification accuracy of sentences was 79% for the classifier of complication and 71% for the classifier of treatment. These values are consistent with the results obtained from the combination of the machine learning algorithm Support Vector Machine with the filter Noise Removal and Balancing of Classes. In the identification of relevant terms, the results of our methodology showed higher F-measure percentage (42%) compared to the manual classification (31%) and to the partial process, i.e., without using the classifier of complication (36%). Even with low percentage of recall, there was no impact observed on the extraction process, and, in addition, we were able to validate the hypothesis considered in this work. In other words, it was possible to obtain 100% of recall for different terms, thus not impacting the extraction process, and further the working hypothesis of this study was proven.

Keywords: Information Extraction, Treatments, Text Mining, Preprocessing, Biomedical Domain, Sickle Cell Anemia.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

AF - Anemia Falciforme

AM - Aprendizado de Máquina

BD - Banco de Dados

EI - Extração de Informação

HMM - *Hidden Markov Model*

HU – Hidroxiureia

IA - Inteligência Artificial

IDC - *International Data Corporation*

IM - Informação Mútua

KDD - *Knowledge Discovery in Database*

MD - Mineração de Dados

MT - Mineração de Textos

MUC - *Message Understanding Conference*

NB - *Naïve Bayes*

PDF - *Portable Document Format*

PLN - Processamento de Língua Natural

POS - *Part-of-Speech*

RAT - Reconhecimento Automático de Termo

REN - Reconhecimento de Entidade Nomeada

RI - Recuperação de Informação

SCA - *Sickle Cell Anemia*

SCD - *Sickle Cell Disease*

SVM - *Support Vector Machines*

DICIONÁRIO DE CONCEITOS

Tratamento: drogas, terapias e procedimentos usados para tratar uma doença.

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 Anemia Falciforme. 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 causados por um tratamento. O uso de certas drogas ou terapias pode causar nos pacientes com Anemia Falciforme os seguintes efeitos colaterais: leucemia (*leukemia*), contágio por vírus devido à transfusão de sangue (HIV) e depressão (*depression*);

Efeito positivo do tratamento (ou benefício): melhorias ou benefícios ocasionados por estímulos do tratamento, ou seja, são os efeitos positivos proporcionados por um tratamento, tais como remissão da doença (*disease remission*), melhora clínica (*clinical improvement*) e redução no tempo de internação (*reduction in hospitalization time*).

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Neste capítulo é apresentado o problema de pesquisa que foi investigado ao longo do mestrado, o contexto envolvido, a motivação e os desafios que deram origem ao desenvolvimento deste projeto de pesquisa. Os principais objetivos são discutidos e as contribuições mais importantes são apresentadas, finalizando com a descrição da organização da dissertação.

1.1 Considerações Iniciais

Atualmente na área médica uma grande quantidade de dados tem sido produzida e armazenada em documentos no formato textual não estruturado, proporcionando a criação de volumosos conjuntos de documentos digitais (STAVRIANOU; ANDRITSOS; NICOLOYANNIS, 2007; LUO, 2008; AGARWAL; YU; KOHANE, 2011). O grande volume de dados armazenados ultrapassa em muito as habilidades humanas para interpretá-los manualmente, exigindo técnicas computacionais para automatizar a análise desses documentos digitais de forma ágil e preferencialmente de forma automática ou semiautomática. Devido à alta taxa de crescimento da quantidade de documentos textuais, a qual é medida no contexto desta dissertação de mestrado em termos do número de publicações de artigos em periódicos e em eventos científicos, torna-se impossível que os médicos e especialistas da área médica analisem toda a literatura relevante de forma manual,

mesmo quando se restringe a artigos em tópicos específicos (JENSEN; SARIC; BORK, 2006).

Assim, abordagens de **extração de informação** são utilizadas como recursos para estruturar as informações relevantes do texto, com o objetivo de permitir a descoberta futura de relacionamentos interessantes entre as informações extraídas (JENSEN; SARIC; BORK, 2006; AGARWAL; YU; KOHANE, 2011). Com o objetivo de identificar as informações relevantes, além de estruturar e armazenar estas informações em um banco de dados, nesta dissertação é proposto um **processo baseado em parágrafos para a extração de tratamentos de artigos científicos do domínio biomédico**.

Este trabalho foi desenvolvido em parceria com a Universidade de São Paulo e a Universidade Metodista de Piracicaba, as quais participaram conjuntamente com a UFSCar no projeto de pesquisa intitulado "*An Environment for Analyzing Data of Sickle Cell Disease*", cujo projeto objetiva definir e implantar um ambiente para a análise de dados da doença Anemia Falciforme. Este ambiente é constituído por dois sistemas principais: DORS-SCA (*Data Organizing and Recovering System for Sickle Cell Anemia*) e DSS-SCA (*Decision Support System for Sickle Cell Anemia*). O primeiro sistema visa extrair informações de artigos científicos originalmente no idioma inglês sobre a doença Anemia Falciforme e armazená-las em um banco de dados. O segundo sistema objetiva identificar padrões e permitir a predição de fatos futuros por meio da aplicação de técnicas de mineração de dados em um data warehouse.

O projeto "*An Environment for Analyzing Data of Sickle Cell Disease*" originou algumas pesquisas em nível de mestrado, sendo que MATOS (2010) se preocupou em extrair informações de efeitos negativos da doença (ou complicações) e efeitos negativos do tratamento (ou efeitos colaterais); LOMBARDI (2011) destacou a importância de extrair padrões numéricos sobre pacientes, tais como: informações de quantidade de pacientes e quantificação de eventos aos quais os pacientes foram submetidos nos estudos para análises estatísticas, e GOMIDE (2011) enfocou a mineração de dados baseada em grafos no contexto da Anemia Falciforme.

Complicação entende-se por qualquer efeito negativo inerente da doença, ou seja, decorrentes das hemácias falciformes, independente do uso de um determinado tratamento. Síndrome torácica aguda, sequestro esplênico e falha renal crônica são alguns exemplos de complicações da doença AF. Sintomas da doença

também são considerados complicações, como febre, hemorragia e inflamação dos dedos do pé e da mão (*dactylitis*). Já efeito negativo do tratamento (ou efeito colateral) são problemas ocasionados por estímulos do tratamento, ou seja, são os efeitos negativos de um tratamento, como exemplo, contágio por vírus devido à transfusão de sangue (HIV) e depressão (SILVA; RAMALHO; CASSORLA, 1993; SILVA_PINTO et al., 2009).

1.2 Motivação

O surgimento da Mineração de Textos (MT) foi motivado pela necessidade de se descobrir de forma semiautomática informações e conhecimento em textos. A utilização das ferramentas de Mineração de Textos torna-se indispensável neste cenário, possibilitando o processamento de uma grande quantidade de textos, permitindo recuperar informações relevantes, possibilitando a extração de informação e o reconhecimento de padrões (EBECKEN; LOPES; COSTA, 2003; GUPTA; LEHAL, 2009).

Este trabalho utiliza a MT para identificar e extrair informações úteis, novas e interessantes em artigos científicos do domínio biomédico, os quais estão no formato PDF e escritos em inglês, mais especificamente aplicada como prova de conceito em artigos completos da doença Anemia Falciforme. A Anemia Falciforme (AF), ou *Sickle Cell Anemia* (SCA), é uma doença hematológica e hereditária, a qual causa a destruição crônica das células vermelhas do sangue, afetando principalmente a população negra e considerada como um problema de saúde pública no Brasil (SILVA_PINTO et al., 2009) .

Neste contexto, esta pesquisa em nível de mestrado tem como objetivo extrair as informações em artigos científicos completos sobre tratamentos de doenças relacionados ao domínio biomédico.

1.3 Objetivos e Hipóteses

A extração de informação enfocará especificamente em termos de "tratamentos" (i.e. drogas, terapias e procedimentos usados para tratar uma doença).

Foi realizada uma análise empírica de artigos da doença Anemia Falciforme, para os quais se descobriu que: (i) ~10% dos tratamentos apareceram na sentença anterior ou posterior à sentença na qual uma complicação foi encontrada; (ii) ~90% dos tratamentos apareceram na mesma sentença na qual a complicação foi encontrada; e (iii) muitos poucos tratamentos ocorreram em sentenças mais distantes.

A partir desta análise empírica, esta pesquisa se baseia nas seguintes hipóteses:

H1. É possível extrair termos de tratamento de forma semiautomatica.

H2. É possível alcançar uma alta precisão e revocação de termos distintos de tratamentos conhecidos utilizando um dicionário estendido com variações de termos e siglas.

H3. É possível alcançar uma precisão e revocação satisfatória na identificação de novos termos de tratamento utilizando regras especificas desenvolvidas para um domínio biomédico.

H4. A busca inicial de sentenças que possuem termos de complicação melhora a eficiência na identificação e extração de termos de tratamento. Considera-se como hipótese que na maioria dos casos os termos de tratamento ocorrem em mesma sentença que possui um termo de complicação (i.e. de um efeito negativo da doença) ou em sentenças próximas de uma complicação em um mesmo parágrafo. Sendo assim, o parágrafo é considerado neste processo como uma unidade com conteúdo de informações centralizado, no qual se localiza a informação de interesse (i.e. termos de tratamento).

Portanto, o objetivo desta pesquisa é extrair informações de artigos científicos do domínio biomédico, especificamente a extração de tratamentos, e para alcançar este objetivo é proposto um processo que combina três abordagens para a extração de informação: aprendizado de máquina, regras e dicionário. A técnica de aprendizado de máquina é utilizada especificamente na classificação de sentenças, visando filtrar o conjunto de sentenças para um subconjunto de sentenças de

interesse, enquanto as regras e o dicionário são usados para a identificação e a posterior extração de termos de tratamento nas sentenças de interesse.

O processo proposto para a extração de termos de tratamento é independente da doença e os passos do processo não dependem de uma doença em si, mas foram aplicados para uma doença particular (i.e. Anemia Falciforme).

1.4 Organização do Trabalho

O conteúdo desta dissertação está organizado em seis capítulos:

- Capítulo 1: é abordado o problema que foi investigado, o contexto da pesquisa, a motivação para a definição do tema, as hipóteses e os objetivos;
- Capítulo 2: descreve a fundamentação teórica que será usada no desenvolvimento deste trabalho a qual possui conceitos importantes para o entendimento do problema e da solução proposta;
- Capítulo 3: descreve e analisa os trabalhos relacionados a este projeto;
- Capítulo 4: detalha o trabalho desta pesquisa em nível de mestrado, descrevendo o ambiente proposto para a extração de tratamentos;
- Capítulo 5: descreve a prova de conceito para analisar as principais etapas da metodologia (i.e., classificação de sentenças e identificação de termos relevantes);
- Capítulo 6: sumariza as principais contribuições, produções científicas e conclusões obtidas a partir do desenvolvimento do processo proposto, além de propor sugestões de trabalhos futuros.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo descreve os conceitos e os fundamentos teóricos necessários para compreender a proposta da metodologia de pré-processamento textual para extração de informação de termos de tratamento de artigos científicos do domínio biomédico. Isto compreende os conceitos de extração de informação, as suas abordagens e as principais tarefas de extração. Compreende também conceitos sobre mineração de textos, e ainda, fundamentos usados em mineração de textos, tais como processamento de língua natural e aprendizado de máquina.

2.1 Mineração de Textos

A Mineração de Textos (MT), também conhecida como Mineração de Dados Textual ou Descoberta de Conhecimento Textual, refere-se ao processo de obter conhecimento relevante, útil e interessante de bases textuais, ou seja, de dados não estruturados. Apoia-se em alguns conceitos de Mineração de Dados (MD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), o qual pretende extrair regularidades, padrões ou tendências de grandes volumes de textos em língua natural para um domínio específico (EBECKEN; LOPES; COSTA, 2003; MHAMDI; ELLOUMI, 2008; GUPTA; LEHAL, 2009).

Mineração de textos é uma área multidisciplinar que utiliza técnicas das áreas de Recuperação de Informação, Processamento de Língua Natural, Extração de Informação, juntamente com algoritmos e métodos de KDD (*Knowledge Discovery in Database*), Aprendizado de Máquina e Estatística (HOTHO; NÜRNBERGER; PAASS, 2005; GUPTA; LEHAL, 2009).

Atualmente, diversas áreas têm armazenado um grande volume de documentos em formato textual (JENSEN; SARIC; BORK, 2006; AGARWAL; YU; KOHANE, 2011). A informação, que é indiretamente medida em termos do número de artigos e revistas que são publicados, está aumentando a uma taxa considerável, de modo que não é mais possível analisar toda a literatura científica relevante de forma manual, mesmo em temas especializados (JENSEN; SARIC; BORK, 2006; THOMPSON et al., 2011). A quantidade de informação *on-line* atualizada em 2009 pela pesquisa do IDC (*International Data Corporation*) foi de 800 *exabytes* de dados, que compreende o valor de 8×10^{20} bytes (800 quintilhões de bytes).

Devido a essa taxa de crescimento de documentos textuais, ferramentas de mineração tornam-se essenciais, as quais possibilitam extrair as informações de modo semiautomático e promover o reconhecimento de padrões. Ademais, para que toda esta informação não estruturada armazenada em documentos textuais seja processada é necessário utilizar métodos e algoritmos de pré-processamento para extrair padrões úteis (JENSEN; SARIC; BORK, 2006; AGARWAL; YU; KOHANE, 2011).

Existem na literatura algumas variações do processo de mineração de textos. Originalmente foi proposta uma divisão do processo em nove etapas em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). No geral, o processo contém quatro etapas principais (EBECKEN; LOPES; COSTA, 2003; FELDMAN; SANGER, 2007; MHAMDI; ELLOUMI, 2008): (i) coleta de documentos, (ii) pré-processamento, (iii) extração de padrões; e (iv) análise e avaliação dos resultados, conforme ilustrado na Figura 2.1.

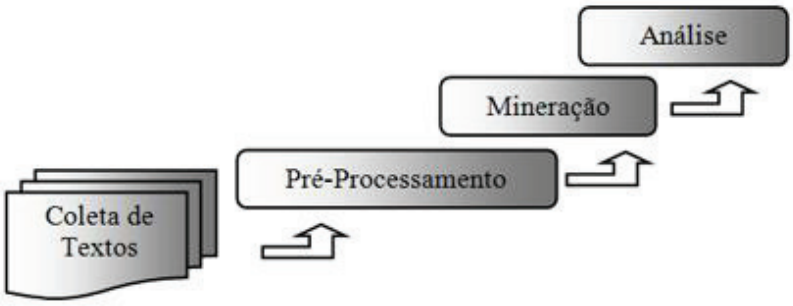

Figura 2.1 - Processo de mineração de texto em quatro etapas.

A fase inicial tem por objetivo a coleta dos dados que vão constituir a base textual, ou seja, determinar e selecionar o domínio de abrangência das técnicas de mineração de texto. A segunda etapa, denominada de pré-processamento, é

responsável por obter uma representação estruturada dos documentos (FELDMAN; SANGER, 2007). Após os documentos serem representados em um formato adequado, é possível realizar a extração de conhecimento utilizando técnicas de mineração de textos de forma similar ao processo tradicional de mineração de dados (MHAMDI; ELLOUMI, 2008). Finalmente, na última etapa, deseja-se avaliar o resultado gerado a partir dos passos anteriores.

Rezende et al. (2003) apresentam uma abordagem que divide o processo em um ciclo que pode ser repetido várias vezes. Esta abordagem é dividida em cinco grandes etapas, como pode ser observado na Figura 2.2: (i) identificação do problema (fase anterior ao processo de MD), etapa no qual o especialista do domínio identifica e define o problema, determina os requisitos, objetivos e metas a serem atingidas; (ii) pré-processamento, etapa de extração e integração, transformação, limpeza, seleção e redução dos dados; (iii) extração de padrões, entende-se como a aplicação de algoritmos de mineração de dados para a extração de conhecimento; (iv) etapas de pós-processamento e utilização do conhecimento (fase posterior ao processo de MD), compostas pelas fases de validação e visualização dos resultados.

Figura 2.2 - Etapas do processo de mineração de dados.

***Fonte*: Adaptado de Rezende et al. (2003).**

O enfoque desta pesquisa em nível de mestrado é a fase de pré-processamento do processo de mineração de textos, a qual corresponde à segunda

etapa do processo de MT. Para esta fase de pré-processamento, uma metodologia de extração de informação é proposta para tratar especificamente do assunto "tratamento" encontrado em artigos científicos do domínio biomédico. A prova de conceito irá abordar o assunto "tratamento" no domínio da doença Anemia Falciforme, para o qual a metodologia foi aplicada e cujo resultado é descrito no Capítulo 4.

A fase de pré-processamento é composta de um conjunto de transformações a serem executadas em dados crus não estruturados, com o objetivo de preparar, organizar e transformar estes dados para um formato adequado para a operação de mineração (ARANHA, 2007; FELDMAN; SANGER, 2007). O formato escolhido para classificação de sentença geralmente é representado em uma tabela atributo-valor, esta que constitui os documentos e tem como característica valores dispersos dos dados e uma alta dimensionalidade (ARANHA, 2007). A tabela atributo-valor é representada em um modelo espaço vetorial, no qual cada documento é representado por um vetor (d_i) e cada posição deste vetor equivale a um atributo (dimensão) do documento (t_i). A classe dos documentos (c_i) é ilustrada na última coluna da matriz, caso estes documentos sejam rotulados. Caso contrário, esta coluna é eliminada da matriz atributo-valor (EBECKEN; LOPES; COSTA, 2003). O modelo *bag-of-words* (saco de palavras) é um dos formatos de representação de documentos que utiliza a definição do modelo espaço vetorial. A abordagem *bag-of-words* ignora a ordem das palavras assim como qualquer informação de pontuação ou estrutural, mas retém o número de vezes que uma palavra aparece (ANTHONY; LASHKIA, 2003; EBECKEN; LOPES; COSTA, 2003).

Na Tabela 2.1 é apresentado um exemplo do preenchimento de uma tabela atributo-valor, com as células com valores binários e os atributos relacionados ao domínio da Anemia Falciforme.

As sentenças e atributos (*Hydroxyurea, HU, therapy) s*ão pertinentes às sentenças a seguir:

1. Treatment: *"**Hydroxyurea (HU)** is considered to be the most successful drug **therapy** for severe sickle cell disease (SCD)."*

2. Other: *"As a whole, 34 patients were considered at risk of primary stroke on the basis of abnormal TCD, and 7 of the 21 explored by MRI/MRA had moderate/severe arterial stenosis."*

Tabela 2.1 - Exemplo de uma tabela atributo-valor definida por duas classes.

	Atributos				
Sentenças		Hydroxyurea	HU	Therapy	Classe
	Sentença1	1	1	1	Treatment
	Sentença2	0	0	0	Other
Legenda: 1 = contém; 0 = não contém					

As transformações a serem aplicadas aos dados não estruturados constituem em identificar, ajustar e tratar os dados corrompidos, atributos irrelevantes e valores desconhecidos. Estas técnicas de pré-processamento são tratadas na Seção 2.6.

Depois da aplicação das técnicas de pré-processamento são aplicadas as tarefas de mineração de texto, tais como agrupamento e regras de associação, que torna explícito o relacionamento entre os documentos; classificação que identifica os tópicos de um documento e ainda sumarização que objetiva produzir automaticamente resumos, sem perder as características-chave (EBECKEN; LOPES; COSTA, 2003).

A etapa de análise e validação dos resultados obtidos é realizada na etapa final. A qualidade e o desempenho dos resultados podem ser avaliados utilizando medidas padrão da Recuperação de Informação, tais como Precisão (medida de fidelidade), Revocação (medida de completude), e ainda, medida-F (média harmônica ponderada entre a Precisão e a Revocação) (CLEVERDON; MILLS; KEEN, 1966; EBECKEN; LOPES; COSTA, 2003).

2.2 Tarefas de Mineração de Textos

Cada tipo de tarefa extrai um tipo diferente de conhecimento dos textos e a escolha pela tarefa é feita de acordo com o objetivo final do processo de descoberta de conhecimento. As tarefas de mineração de textos mais frequentes podem ser divididas, como observado na Figura 2.3 em tarefas preditivas e descritivas (REZENDE et al., 2003).

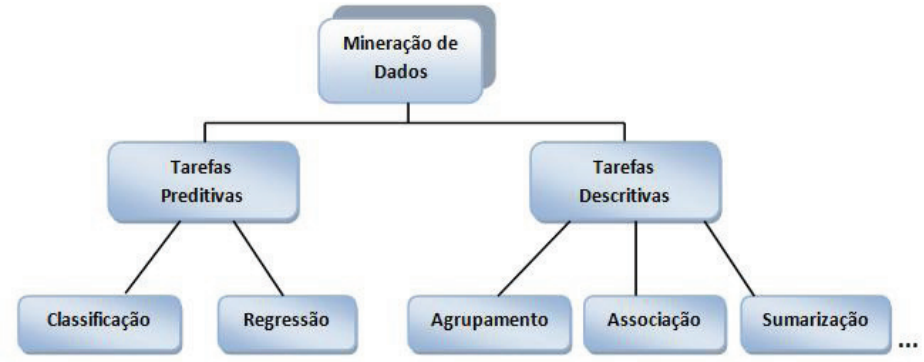

Figura 2.3 - Tarefas de mineração de textos.

Fonte: **Rezende et al. (2003).**

Tarefas preditivas compõem-se da generalização de exemplos ou experiências passadas com respostas conhecidas em uma linguagem capaz de identificar a classe de um novo exemplo. Os tipos de tarefas para predição são: classificação, que é baseado na predição de um valor categórico; e regressão, no qual o atributo a ser predito é baseado em um valor contínuo. Estas tarefas preditivas utilizam os modelos de aprendizado de máquina supervisionado, uma vez que as categorias são sempre pré-conhecidas e disponíveis junto aos dados, denominados exemplos rotulados (MONARD; BARANAUSKAS, 2003).

As tarefas descritivas utilizam modelos de aprendizado de máquina não supervisionado e baseiam-se na identificação de comportamentos particulares da coleção de dados, nos quais estes dados são exemplos não rotulados ou tratados como não rotulados (REZENDE et al., 2003). Os tipos de tarefas para descrição são: agrupamento, que visa agrupar os dados de acordo com alguma semelhança entre eles; regras de associação, que são relações lógicas inferidas entre dados correlacionados analisados conjuntamente (AGRAWAL; SRIKANT, 1994); e sumarização, que visa obter a produção de resumos (EBECKEN; LOPES; COSTA, 2003). A seguir é apresentada a tarefa de classificação que é utilizada nesta dissertação na fase de classificação de sentenças.

2.2.1 Classificação

Classificação é uma tarefa de aprendizado supervisionado que possui a função de classificar um documento (ou uma sentença no contexto desta dissertação) em categorias predefinidas (YANG; PEDERSEN, 1997; DÖRRE;

GERSTL; SEIFFERT, 1999; SEBASTIANI, 2002; IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005). Dado um conjunto pré-definido de categorias (ou classes), o objetivo da classificação é criar um classificador que possa predizer se um novo documento pertence ou não a uma categoria (YANG; PEDERSEN, 1997). Na comunidade científica a abordagem predominante para classificação é baseada em técnicas de aprendizado de máquina (SEBASTIANI, 2002).

A Figura 2.4 apresenta o processo de classificação. Um conjunto de documentos pré-classificados em categorias é considerado para treinamento (a). Este é analisado com o objetivo de prover um esquema de classificação (b). Assim, o esquema de classificação pode ser utilizado para a classificação de outros documentos (c), classificando o documento (d) nas categorias definidas anteriormente (a) (DÖRRE; GERSTL; SEIFFERT, 1999).

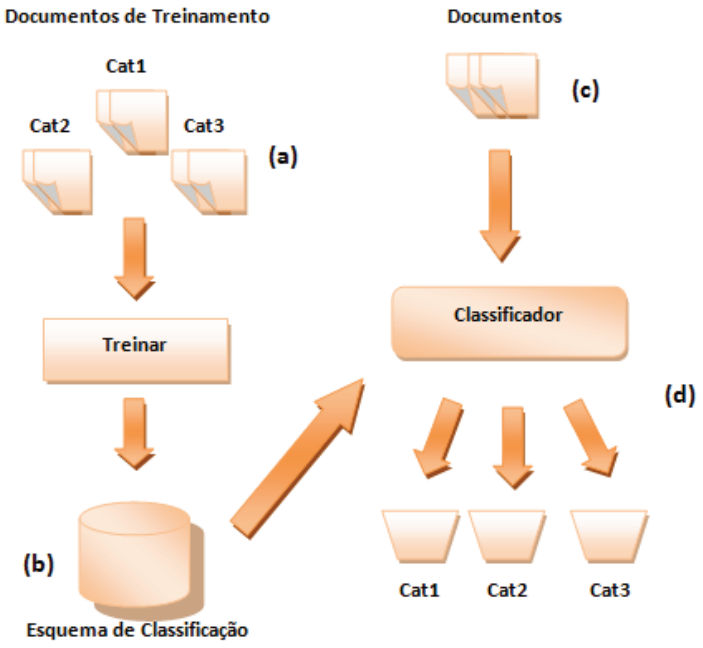

Figura 2.4 - Classificação de documentos.

***Fonte*: Adaptado de Dörre, Gerstl e Seiffert (1999).**

No âmbito deste projeto, esta tarefa será usada para classificar as sentenças de cada documento em classes pré-definidas, tais como a classe “tratamento “e a classe “outros” para o classificador de tratamento, e a classe “complicação” e a classe “outros” para o classificador de complicação. Esta etapa ocorrerá antes da etapa de extração de informação, a qual usará das técnicas de dicionário e regras para localizar e extrair partes relevantes das frases pré-classificadas em tratamento. Portanto, a classificação será usada na metodologia proposta como um filtro que

selecionará apenas as sentenças de interesse, diminuindo o custo de análise das sentenças de um artigo.

2.3 Extração de Informação

Extração de Informação (EI) é uma subárea do Processamento de Língua Natural (PLN) que se concentra em reconhecer e extrair trechos relevantes em documentos não estruturados ou semiestruturados, para posteriormente serem armazenados em um formato estruturado, tipicamente em um banco de dados (FELDMAN; SANGER, 2007). Cada documento é processado para extrair as entidades e os relacionamentos relevantes (i.e. fatos ou eventos que envolvem certas entidades) (FELDMAN; SANGER, 2007). Após o armazenamento dos dados extraídos, estes dados são submetidos a algoritmos de mineração de dados, integrados em bases de conhecimentos para permitir o raciocínio ou apenas apresentados diretamente para os usuários (JACKSON; MOULINIER, 2002; MOONEY; BUNESCU, 2005; ANANIADOU; KELL; TSUJII, 2006; ANANIADOU; NENADIC, 2006; CUNNINGHAM, 2006; FELDMAN; SANGER, 2007; GUPTA; LEHAL, 2009).

No contexto desta dissertação de mestrado, EI é caracterizada como um instrumento essencial inserido em um ambiente de análise dos dados de doenças no domínio biomédico. Este ambiente (tratado na Seção 1.1) visa extrair informações médicas de forma semiautomática de um conjunto de artigos científicos sobre uma doença escritos no idioma inglês e armazenados no formato PDF (*Portable Document Format*). Mais especificamente, este trabalho enfocará especificamente o assunto “tratamento” como objetivo da extração de informação.

Existem três abordagens principais aplicadas na EI: aprendizado de máquina (KOU; COHEN; MURPHY, 2005; COHEN; HUNTER, 2008), regras (ANANIADOU; MCNAUGHT, 2006) e dicionário (KRAUTHAMMER; NENADIC, 2004). Essas abordagens são detalhadas a seguir na seção 2.4.

2.4 Abordagens para extração de informação

Kou, Cohen e Murphy (2005) e Cohen e Hunter (2008) descrevem duas abordagens para a extração de informação: abordagem baseada em regras, utilizada para identificar padrões de extração com o uso de expressões regulares; e abordagem baseada em aprendizado de máquina, que utiliza classificadores para separar ou identificar sentenças de interesse. Além dessas, Krauthammer e Nenadic (2004) apresentam uma terceira abordagem para o reconhecimento automático de termos: abordagem baseada em dicionário, a qual utiliza informações de um dicionário para auxiliar na identificação dos termos ou das entidades no texto. Essas abordagens são as três predominantes na literatura e essenciais para a extração de conhecimento no domínio biomédico, sendo detalhadas nas próximas seções.

2.4.1 Abordagem Baseada em Aprendizado de Máquina

Técnicas de aprendizado de máquina são utilizadas em Reconhecimento Automático de Termo (RAT), que são projetadas para atender a uma classe específica de entidades, e usam dados de treinamento para aprender as características que são úteis e relevantes para o reconhecimento e a classificação de termos (KRAUTHAMMER; NENADIC, 2004). Várias técnicas de aprendizado de máquina têm sido utilizadas para identificação e classificação de termos, incluindo *Hidden Markov Model* (HMM), *Naïve Bayes*, *Support Vector Machines* (SVM) e árvores de decisão (EBECKEN; LOPES; COSTA, 2003).

Os principais problemas relacionados aos algoritmos de aprendizado de máquina são a necessidade de grandes quantidades de dados de treinamento e o fato que a classificação é prejudicada quando o conjunto de dados de uma classe é pequeno em relação a outras classes (BATISTA; CARVALHO; MONARD, 2000; ANANIADOU; MCNAUGHT, 2006).

O objetivo desta dissertação de mestrado é extrair informações de artigos científicos do domínio biomédico, especificamente a extração de dados sobre tratamentos. Portanto, esta abordagem será utilizada para construir um classificador destinado a classificar as sentenças nas classes “complicação” e “outros” (sentenças que não são de complicação) e posteriormente um classificador com o propósito de

classificar as sentenças pré-selecionadas nas classes “tratamento” e “outros” (sentenças que não são de tratamento). A construção destes dois classificadores se faz necessário para conferir a hipótese deste trabalho, que considera que na maioria dos casos os termos de tratamento ocorrem em uma mesma sentença que possui um termo de complicação ou em sentenças próximas em um mesmo parágrafo.

2.4.2 Abordagem Baseada em Regras

A abordagem baseada em regras utiliza termos padrão de formação. Esta abordagem baseia-se no desenvolvimento e na aplicação de regras que descrevem estruturas de nomes comuns para certas classes de termos, usando ortografia léxica descrita por expressão regular, ou recursos morfossintáticos mais complexos (ANANIADOU; MCNAUGHT, 2006).

Um exemplo de padrões extraídos a partir de regras é ilustrado abaixo, o qual permite encontrar o relacionamento entre hidroxiureia (droga) e um tratamento:

< hidroxiureia> desempenha um papel no <tratamento>
<tratamento> está associado com < hidroxiureia>

Para identificação destes relacionamentos**,** pode-se utilizar (SILVA et al., 2007):

- Análise linguística (ou análise sentencial): a estrutura de um texto é formada de sentença a sentença, sendo assim a primeira e a menor unidade do processamento. Uma sentença pode ser definida como uma unidade de comunicação, uma vez que se apresenta como uma declaração dotada de expressão completa de sentido, por exemplo, sentenças constituídas de uma palavra “Atenção!” ou “Perigo!”;
- Análise Semântica: extrair um significado completo da sentença a partir dos significados das palavras ou grupos de palavras, e das relações entre elas, e neste caso, é necessário o conhecimento particular do domínio, por exemplo, para distinguir a interpretação correta do termo “banco” (se é uma instituição financeira ou um assento em uma cadeira).

Segundo Ananiadou e McNaught (2006), esta abordagem é normalmente difícil de se ajustar a diferentes domínios ou classes, uma vez que as regras são específicas do domínio. Outra desvantagem desta abordagem é o tempo significativo para a definição e para a validação das regras (COHEN; HUNTER, 2008).

2.4.3 Abordagem Baseada em Dicionário

A abordagem baseada em dicionário dispõe de uma lista de termos para localizar as ocorrências no texto. Considera-se um termo ocorrência de cada sequência de palavras no texto que corresponder a uma entrada no recurso terminológico; apenas cadeias de caracteres são tratadas como tais termos (ANANIADOU; MCNAUGHT, 2006). Neste contexto, as informações armazenadas são pertinentes ao domínio biomédico, e estas informações promovem o reconhecimento de termos tais como genes, proteínas, doenças, tratamentos, efeitos negativos de tratamentos (efeitos colaterais), efeitos positivos de tratamentos (benefícios), e efeitos negativos de doenças (complicações) e ainda, a combinação entre eles. O casamento de padrão geralmente é utilizado entre as entradas contidas no dicionário e as palavras encontradas nas sentenças (MATOS, 2010).

Neste trabalho, termo refere-se a uma palavra. Um atributo composto por um único termo é conhecido como unigrama, e um atributo composto por n termos é chamado de n-grama (ARANHA, 2007). Uma n-grama de letras é uma sequência de n letras da uma dada palavra, por exemplo: neste contexto considera a combinação "*sickle cell*", na qual representa 2-gramas e "*bone marrow transplantation"* que é representada por 3-gramas.

Uma desvantagem da abordagem de dicionário é a restrição de nomes que estão presentes no dicionário, sendo assim indispensável o armazenamento de palavras com variações, tais como palavras no plural, palavras com variação de gênero e sinônimos. Por exemplo, variações de nomes da proteína "*NF-kappa B*" podem ser encontradas na literatura: "*NF kappa B*", "*NF-kappa-B*", "*NF-Kappa B*", "*NF-Kappa-B*". (TSURUOKA; TSUJII, 2004). Tsuruoka e Tsujii (2004) também alertam sobre o uso desta abordagem**,** apresentando dois problemas fundamentais: falso reconhecimento causado principalmente por nomes curtos e baixa revocação em sistemas de extração de informação devido a variações de ortografia. Na prova de conceito desta dissertação, o esquema do banco de dados do Projeto da Anemia

Falciforme utiliza uma tabela adicional nomeada de "*variation"*, para armazenar as variações e os sinônimos, de forma a reduzir os problemas da técnica de dicionário.

Segundo Kou, Cohen e Murphy (2005), extratores baseado em dicionário**,** ao extrair nomes de proteína**,** geralmente tem uma baixa revocação, exceto se lidar com as diversas variações de nome. Uma alternativa de se trabalhar com essas variações é utilizar técnicas como aproximação de cadeias de caracteres (distância de edição) (LEVENSHTEIN, 1966), (TSURUOKA; TSUJII, 2004). Esta técnica substitui, apaga e insere caracteres e dígitos que podem ser usados para implementar strings (cadeias de caracteres) mais flexíveis combinando sobre um dicionário de termos de proteínas (ANANIADOU; MCNAUGHT, 2006). Por exemplo, a distância de edição entre as palavras "*kitten*" e "*sitting*" é 3, pois com apenas 3 edições é capaz de converter uma palavra na outra: *kitten* – 1) *sitten* (substituição de 'k' por 's'); 2) *sittin* (substituição de 'e' por 'i') e 3) *sitting* (inserção de 'g' no final). Neste trabalho, não optamos pelo uso da técnica de aproximação de cadeias de caracteres.

2.5 Tarefas da Extração de Informação

Ananiadou e McNaught (2006) caracterizam extração de informação como uma representação de cada fato como um *template* cujos *slots* são preenchidos com base no que foi localizado no texto. A extração de informação está subdividida em cinco tarefas, conforme mostrados na Tabela 2.2, exemplificado por Cunningham (2006).

Tabela 2.2 - Cinco tarefas de extração de informação.

Tarefa	Descrição
Entidade Nomeada	Refere-se a uma entidade que possui um nome próprio. Extrai nome, pessoa, organização e localização. (e.g., Isabelle e Dominique como pessoas).
Correferência	Identifica relações entre entidades. As correferências são utilizadas quando a definição de um objeto tem uma relação de dependência conceitual com um objeto já instanciado (e.g., Comprei uma casa. Esta casa será sempre minha).
Template	Uma lista de entidades com seus atributos associados, tais como formas

Element	alternativas de um nome (e.g., o sistema acrescenta um *alias* alternativo quando nota que a 'administração do Lula-PT' também se refere 'oficial do governo').
Template Relation	Identificação das propriedades dos *Template Elements* ou relações entre eles (e.g., relação entre o funcionário-presidente e a organização-governo).
Scenario Template	Extrai eventos. Um ou mais *slots* são preenchidos com *template element* ou *template relation* para cada tipo de evento extraído (e.g., *template element* pode ter identificado Isabelle e Dominique como pessoas, entidades presentes na edição das cartas de amor de Napoleão).

O Reconhecimento de Entidade Nomeada (REN) envolve identificar referências para tipos de objetos particulares, tais como nomes de pessoas, empresas e localizações (MOONEY; BUNESCU, 2005). Esta tarefa tem sido utilizada em diversos domínios, inclusive para extrair informação de dados biológicos e de documentos médicos (KRAUS; BLAKE; WEST, 2007; LEE; WU; YANG, 2007).

Entidade mencionada entende-se uma entidade referenciada em um determinado contexto, podendo assim assumir papéis semânticos diferentes em função deste mesmo contexto (ARANHA, 2007).

Na área biomédica as entidades são tipicamente genes, proteínas, tratamentos, efeitos e doenças. Nesta área, é possível identificar uma proteína que interage com outra proteína, ou que uma proteína está localizada em uma parte específica da célula (MOONEY; BUNESCU, 2005). No contexto da Anemia Falciforme, exemplos de entidades são: complicações da doença (dor, tosse, febre e dispneia), tratamentos (hidroxiureia, acido fólico, antibióticos e transfusão de sangue), benefício do tratamento (redução do número de internação e remissão da doença).

Já o Reconhecimento Automático de Termo (RAT) refere-se ao processo de extrair sistematicamente termos técnicos pertinentes e suas variantes de uma coleção de documentos. Seu principal objetivo é distinguir os termos de um campo de assunto a partir de não termos, associando os termos extraídos com um conceito em um *framework* semântico bem definido (ANANIADOU; MCNAUGHT, 2006).

Na área biomédica, são envolvidos os dois domínios de pesquisa – REN (MOONEY; BUNESCU, 2005) e RAT (ANANIADOU; MCNAUGHT, 2006), pois há uma relação entre a pesquisa de entidade nomeada e terminologia. Por exemplo, uma entidade nomeada é qualquer nome próprio, e alguns destes nomes podem ser termos técnicos (e.g., Penicilina), que neste caso, são reconhecidos via RAT,

enquanto outros não (e.g., Hidroxiureia). Da mesma forma, pode haver termos técnicos que não são nomes próprios, por exemplo, a palavra “doença” é um termo técnico da medicina, mas não é um nome próprio. Tanto REN quanto RAT utilizam das mesmas abordagens (aprendizado de máquina, regras e dicionário) para a extração de informação (ANANIADOU; MCNAUGHT, 2006).

2.5.1 Reconhecimento Automático de Termo

Reconhecimento automático de termo é uma classificação geral binária que organiza unidades lexicais do texto em dois grupos: termos e não termos. Denota um conjunto de procedimentos que são utilizados sistematicamente para reconhecer os termos pertinentes na literatura, ou seja, destacar unidades lexicais que são relacionados com conceitos relevantes do domínio (KRAUTHAMMER; NENADIC, 2004).

Os termos encontrados na literatura biomédica (tais como, genes, proteínas, organismos, drogas e produtos químicos) constituem conhecimento de domínio utilizado pela comunidade científica e seria impossível compreender ou extrair informações de um artigo sem o reconhecimento e a associação correta desses termos (KRAUTHAMMER; NENADIC, 2004). Neste contexto, na Tabela 2.3 é apresentado um exemplo de termos sobre tratamento (em negrito) relacionado à doença Anemia Falciforme (e.g., termo encontrado: HU = hidroxiureia (Droga)).

Tabela 2.3 - Exemplo de sentença com termos sobre tratamento da AF.

*One large randomized trial tested the efficacy of **HU** in adults with SCD and found that after 2 years of treatment, Hb F% increased by 3.2 percent and hemoglobin increased by 0.6 g/dl.*

Fonte: **Segal et al. (2008).**

Reconhecimento Automático de Termo é composto de três passos (ANANIADOU; MCNAUGHT, 2006). O primeiro passo é o reconhecimento de termo que diferencia os termos dos não termos (KRAUTHAMMER; NENADIC, 2004); no segundo passo, os termos reconhecidos são classificados nas classes mais amplas do domínio, tais como genes, proteínas ou tecidos; e o passo final, que é o mapeamento de termos, que associa automaticamente termos com novos conceitos

representados por uma ontologia. Na Figura 2.5 são ilustrados os passos para identificação de termo no texto.

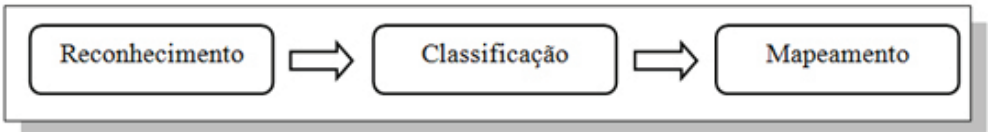

Figura 2.5 - Passos para identificação de termo no texto.
Fonte: **Adaptado de Krauthammer e Nenadic (2004).**

Segundo Ananiadou e McNaught (2006), a maioria das abordagens no âmbito biomédico integra o reconhecimento de termo e a classificação de termo em uma única etapa, por exemplo, para identificar os termos e associá-los a classes pré-definidas do domínio biomédico, como genes, proteínas ou doenças.

Esta dissertação objetiva aplicar RAT para reconhecer e classificar os termos, integrando as três abordagens utilizadas para extração de informação.

2.6 Processamento de Língua Natural

O termo Processamento de Língua Natural (PLN) (ou Linguística Computacional) é usado para descrever a função de software ou hardware de um sistema de computador que analisa e sintetiza a língua falada ou escrita (JACKSON; MOULINIER, 2002; PARDO, 2002; ARANHA, 2007).

Segundo Aranha e Passos (2006), PLN é uma técnica chave para a mineração de textos, na qual utiliza conhecimentos da área da linguística e permite aproveitar ao máximo o conteúdo do texto, extraindo entidades, seus relacionamentos, detectando sinônimos, corrigindo palavras escritas de forma errada, e ainda, desambiguizando-as.

A análise de documentos em linguagem natural é realizada em algumas etapas, conforme divisão das camadas de processamento (JURAFSKY; MARTIN, 2000): fonético-fonológico: estudo dos sons linguísticos; morfológico: estudo dos componentes significativas de palavras; sintático: estudo das relações estruturais

entre as palavras; semântico: estudo do significado; pragmático: estudo de como a linguagem é usada para realizar objetivos; e discursivo: estudo das unidades linguísticas maiores do que um discurso único.

Algumas das principais tarefas de PLN são: reconhecimento de contexto, analise sintática, semântica, léxica e morfológica, sumarização e tradução de textos (MANNING; SCHÜTZE, 1999).

Em mineração de textos, os métodos para analisar textos de PLN escritos são usados na etapa de pré-processamento de forma a melhor representar o texto e aproveitar mais o conteúdo (ARANHA, 2007). Estes métodos combinam análise sintaxe e semântica (JENSEN; SARIC; BORK, 2006). O principal objetivo de PLN para esta etapa baseia-se em reconhecer e classificar as entidades mencionadas (ARANHA, 2007).

Segundo Spasic et al. (2005), o passo inicial para o processamento de texto automático é aplicação da **tokenização**, que identifica as unidades básicas do texto conhecidas como *tokens*, utilizando demarcadores explícitos, tais como espaço em branco ou pontuação. Após a tokenização, pode ser realizado o processamento léxico ou sintático, descritos a seguir:

Léxico: inclui **lematização**: processo que substitui a palavra flexionada pela forma básica sem número e sem gênero; ***stemming*:** processo que reduz a palavra ao seu radical; ***Part-of-Speech* (POS):** classificação das palavras segundo a sua classe gramatical que fornece informações sobre o conteúdo morfológico de uma palavra (i.e., artigo, substantivo, verbo, adjetivo, preposição, número e nome próprio) ou morfossintático (identifica as funções sintáticas como sujeito, predicado, aposto). (EBECKEN; LOPES; COSTA, 2003; KOU; COHEN; MURPHY, 2005; SPASIC et al., 2005; ARANHA, 2007; FELDMAN; SANGER, 2007). A Tabela 2.4 apresenta um exemplo de frase etiquetada com *Part-of-Speech*.

Tabela 2.4 – Exemplo de *Part-of-Speech*

Part-of-Speech
<Prop>Mr. Eskew</Prop> <Verb>was</Verb> <Prop>VicePresident</Prop> <Prep>of</Prep>....

Sintático: envolve a análise da estrutura sintática de uma sentença, inclui: ***shallow parser*** que em vez de fornecer uma análise completa de uma sentença,

shallow parser analisa apenas partes que são mais fáceis e sem ambiguidades (FELDMAN; SANGER, 2007) e ***deep parser*** que gera a representação completa da estrutura sintática de uma sentença (FELDMAN; SANGER, 2007).

Remoção de *stopwords:* palavras consideradas não relevantes na análise de textos e que podem ser descartadas, tais como preposições, pronomes e artigos.

Essas técnicas podem ser aplicadas algumas conjuntamente e outras de forma independente. Cada técnica de pré-processamento inicia com um documento parcialmente estruturado até refinar a estrutura (por exemplo, usar *stemming* ou lematização) ou enriquecer a estrutura (por exemplo, valorizando as palavras com o uso de POS) (FELDMAN; SANGER, 2007). A Tabela 2.5 ilustra exemplos de lematização e *stemming*.

Tabela 2.5 - Exemplos de Lematização e *Stemming*.

Lematização	*Stemming*
Cantaremos-> Cantar	Cantaremos -> Cant

As técnicas de *stemming* e de remoção de *stopwords* não são aplicadas no contexto do domínio biomédico (por exemplo, para a doença Anemia Falciforme) porque geram perda de informação importante que pode prejudicar a extração de informação do assunto “tratamento”. O uso de *stopwords*, por exemplo, deve-se ao fato de que algumas *stopwords* ajudam na formação de regras para a extração dos termos de tratamento.

2.7 Aprendizado de Máquina

Aprendizado de Máquina (AM) é uma área da Inteligência Artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir novos conhecimentos, aperfeiçoando-se automaticamente com a sua experiência, produzindo hipóteses úteis (MITCHELL, 1997; MONARD; BARANAUSKAS, 2003).

AM é utilizada no suporte à tarefa de classificação, porém não se restringe à construção de classificadores de texto, mas também pode ser aplicado a uma ampla gama de tarefas de PLN para aplicações online, como exemplo, abordagens em que são necessárias aplicações de corretores ortográficos, *part-of-speech* ou *parsing* (JACKSON; MOULINIER, 2002).

Abordagens de aprendizado de máquina têm comprovado ser muito úteis para a extração de informação, incluindo abordagens que aprendem a extrair diversas categorias de entidades de textos estruturados e não estruturados (ZELENKO et al., 2003; CARLSON et al., 2010).

No contexto desta dissertação de mestrado, o aprendizado de máquina será utilizado para classificar as sentenças dos artigos científicos, e neste projeto não será explorado e considerado que o aprendizado também pode ser dedutivo, portanto, o mesmo utilizará a procedência da indução.

O aprendizado utiliza do princípio da indução, que é sua forma de inferência lógica, com o propósito de obter conclusões genéricas a partir de um conjunto de exemplos. Um conceito é aprendido executando a inferência indutiva sobre os exemplos apresentados. Para a indução originar conhecimento novo representativo, os exemplos das classes devem estar bem definidos e ter uma quantidade satisfatória de exemplos, adquirindo assim hipóteses convenientes para um determinado tipo de problema (MONARD; BARANAUSKAS, 2003). O objetivo do algoritmo (ou indutor) é construir um classificador que possa definir adequadamente a classe de novos dados ainda não rotulados.

O aprendizado indutivo pode ser dividido em supervisionado e não supervisionado, ilustrados na Figura 2.6. No aprendizado supervisionado é fornecido um algoritmo de aprendizado e um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. No aprendizado não supervisionado o indutor analisa os exemplos e tenta determinar se alguns deles podem ser agrupados de alguma maneira formando agrupamentos.

Figura 2.6 - Hierarquia do aprendizado

Fonte: **Adaptado de Monard e Baranauskas (2003).**

Monard e Baranauskas (2003) classificam AM em alguns paradigmas, que compreendem:

- **Simbólico,** buscam aprender construindo representações simbólicas de um conceito 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**, baseia-se em utilizar modelos estatísticos para encontrar uma boa aproximação do conceito induzido. Exemplo: *Support Vector Machines* (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 dos K-vizinhos mais próximos (K-*Nearest Neighbor,* K-*NN*);
- **Conexionista**, são construções matemáticas simplificadas inspiradas no modelo biológico do sistema nervoso. Exemplo: Redes Neurais;
- **Evolutivo**, modelo biológico de aprendizado. Exemplo: Analogia com a teoria de Darwin.

Detalhes sobre AM podem ser encontrados em (MITCHELL, 1997).

2.7.1 Métodos de Particionamento

Métodos de particionamento são utilizados para fazer a avaliação dos algoritmos de aprendizado de máquina supervisionados, conjuntamente com uma

medida de desempenho (geralmente a medida de precisão ou de erro). Alguns desses métodos de particionamento de amostragem randômico são: *Holdout*, Amostra Aleatória, *Cross-Validation* e *Bootstrap* (KOHAVI, 1995; MANNING; SCHÜTZE, 1999; CHEN et al., 2005).

Holdout: o estimador *holdout* divide os exemplos em uma porcentagem fixa de exemplos p de treinamento e $(1-p)$ para teste, considerando normalmente $\left(p > \frac{1}{2}\right)$. Valores típicos são $p = 2/3$ e $(1-p) = 1/3$ (MONARD; BARANAUSKAS, 2003).

Amostra Aleatória: baseia-se na múltipla aplicação do método *holdout.* Em cada iteração, os exemplos são particionados em conjuntos de treinamento e teste. Após o treinamento é obtida a taxa de erro do conjunto de teste (BATISTA; MONARD, 1998). Amostra aleatória pode produzir melhores estimativas de erro que o estimador *holdout* (MONARD; BARANAUSKAS, 2003).

Cross-Validation (Validação Cruzada): uso dos mesmos dados, repetidas vezes, divididos diferentemente. Em *k-fold cross-validation* o conjunto de dados (os exemplos) é aleatoriamente dividido em k partições reciprocamente exclusivas (*folds*). De tamanho aproximadamente igual a $\frac{n}{k}$ exemplos. As $(k-1)$ *folds* são usadas para treinamento e o *fold* restante para teste. Este processo é repetido k vezes, cada vez considerando um *fold* diferente para teste. O erro é a média dos erros calculados em cada um dos k *folds* (MONARD; BARANAUSKAS, 2003).

Stratified Cross-Validation: o estimador *stratified cross-validation* é similar à *cross-validation*, mas ao gerar os *folds* mutualmente exclusivos, a distribuição de classes é considerada em cada amostragem (MONARD; BARANAUSKAS, 2003).

Leave-One-Out: é um caso especial de *cross-validation* quando *k* for igual à quantidade de amostras. Considerando 150 exemplos, a quantidade de *folds* seria então 150. Para treinamento são os mesmos $(k-1)$ exemplos e um *fold* para teste.

Bootstrap: baseia-se em reproduzir o processo de classificação várias vezes. Os exemplos de treinamento são separados do conjunto de exemplo, mas os elementos selecionados se mantêm no conjunto de exemplos, na qual um mesmo elemento possa ser escolhido diversas vezes aleatoriamente (MONARD; BARANAUSKAS, 2003).

Detalhes peculiares sobre métodos de particionamento podem ser obtidos em (KOHAVI, 1995).

2.7.2 Seleção de Atributos

A seleção de atributos é o processo de selecionar um subconjunto de termos do conjunto de treinamento e usá-lo na classificação de texto (MANNING; RAGHAVAN; SCHÜTZE, 2008). O objetivo dos métodos de seleção de atributos é reduzir a dimensionalidade do conjunto de dados, removendo as características que são consideradas irrelevantes ou menos importantes para a classificação (IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005).

Métodos de extração de características são divididos em dois passos distintos (EBECKEN; LOPES; COSTA, 2003): a extração de termos pode ocorrer com base em informação linguística estruturada; e a seleção dos termos ocorre com base em alguma métrica estatística como a frequência ou informação mútua. No primeiro passo, algoritmos de extração de características podem utilizar dicionários para identificar alguns termos e padrões linguísticos para detectar outros termos. No segundo passo, são aplicados métodos para redução de características, que são: frequência de documento, ganho de informação, informação mútua e estatística X^2 (qui-quadrado).

Frequência de Documento (DF): técnica mais simples de redução de termos, a frequência de documentos é o número de documentos no qual um termo ocorre. A suposição é que termos raros não são importantes para a predição da categoria e não afeta o desempenho global. Não selecionando estes termos raros, reduz-se a dimensionalidade do espaço de característica (SEBASTIANI, 2002).

A Equação (2-1) apresenta o número de documentos da classe c que contém o termo t.

$$FD(t,c) = P(t|c) \tag{2-1}$$

Ganho de Informação (GI): frequentemente aplicado como critério de importância do termo no campo do aprendizado de máquina (MITCHELL, 1997), ganho de informação mede o número de bits de informação obtido por uma predição de categoria conhecendo a presença ou ausência do termo em um documento (EBECKEN; LOPES; COSTA, 2003). Dado um conjunto de documentos, o ganho de informação é calculado para cada termo, e os termos cujos ganhos de informação

foram menores que um determinado limite são removidos do espaço das características.

A complexidade do tempo é $O(N)$ e a complexidade do espaço é $O(VN)$, onde N é o número de documento de treinamento e V é o tamanho do vocabulário. A computação da entropia tem um tempo de complexidade de $O(Vm)$. O ganho de informação do termo t com a classe c_i variando de $1 \leq i \leq m$ é determinada pela Equação (2-2):

$$\begin{aligned} GI(t) = &-\sum_{i=1}^{m} P(c_i)\ log\ P(c_i) \\ &+P(t)\ \sum_{i=1}^{m} P(c_i|t)\ log\ P(c_i|t) \\ &+P(\bar{t})\ \sum_{i=1}^{m} P(c_i|\bar{t})\ log\ P(c_i|\bar{t}) \end{aligned} \tag{2-2}$$

$P(t)$ é a probabilidade que o termo t ocorre e $\bar{t}$ é a probabilidade que o termo t não ocorre. $P(c_i|t)$ é a probabilidade condicional da ocorrência de um termo na classe c_i e $P(c_i|\bar{t})$ é a probabilidade condicional de não ocorrer o termo na classe c_i.

Informação Mútua (IM): é um critério comumente usado em modelagem estatística de associação de palavras (EBECKEN; LOPES; COSTA, 2003). Considera o termo t e a categoria c, sendo que A é o número de vezes que t e c coocorrem, B é o número de vezes que t ocorre sem c, C é o número de vezes que c ocorre sem t e N é o número total de documentos (EBECKEN; LOPES; COSTA, 2003). A hipótese do termo t e categoria c é apresentada na Equação (2-3). O tempo de complexidade é $O(Vm)$, similar ao ganho de informação. É uma medida da quantidade de informação que uma variável contém sobre outra. A informação mútua é maior quando todas as ocorrências de dois termos são adjacentes umas às outras, deteriorando-se em baixa frequência.

$$IM(t,c) \cong log\ \frac{A \times N}{(A+C) \times (A+B)} \tag{2-3}$$

Estatística (X^2): Mede a falta de independência do termo t e da categoria c. A medida X^2 tem valor zero se t e c são independentes (EBECKEN; LOPES; COSTA, 2003). A computação tem complexidade quadrática, similar a informação mútua e ao ganho de informação. Considera o significado de A, B e C explicado na medida anterior e, D é o número de vezes que não ocorrem nem c e t. A medida é definida pela Equação (2-4).

$$X^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (2\text{-}4)$$

Dentre estes métodos para redução de características, apenas a Frequência de Documento será utilizada nesta dissertação na fase de classificação de sentenças.

2.8 Métricas de Avaliação de Desempenho

As medidas de precisão e revocação são as medidas de avaliação de desempenho padrão adotadas da área da Recuperação de Informação (RI) (CLEVERDON; MILLS; KEEN, 1966). Estas medidas são as principais métricas utilizadas na avaliação da eficiência de sistemas tanto para busca quanto para aprendizado, e aplicadas igualmente para analisar os resultados gerados a partir da Mineração de Textos.

Precisão (Equação (2-5)) é uma medida de fidelidade, no qual avalia o quanto o modelo acerta; revocação (Equação (2-6)) (também conhecida como cobertura ou sensibilidade) é uma medida de completude, no qual avalia o quanto o modelo contabiliza (EBECKEN; LOPES; COSTA, 2003).

$$Precisão = \frac{Número\ de\ elementos\ relevantes\ recuperados}{Número\ total\ de\ elementos\ recuperados} \quad (2\text{-}5)$$

$$Revocação = \frac{Número\ de\ elementos\ relevantes\ recuperados}{Número\ total\ de\ elementos\ relevantes} \quad (2\text{-}6)$$

Muitas vezes existe uma relação inversa entre precisão e revocação, onde é possível aumentar a qualidade de uma medida ao custo de reduzir a qualidade de outra medida (e.g., um sistema de RI pode aumentar a revocação recuperando mais elementos, ao custo de um número crescente de elementos irrelevantes recuperados e, portanto diminuindo a precisão).

Na Tabela 2.6 é apresentada a matriz de confusão para duas classes (Tratamento/Não Tratamento) da Anemia Falciforme, em que *P* constitui o valor positivo (compreende que a palavra-chave extraída é tratamento); *N* constitui o valor negativo (compreende que não é tratamento); *p* (Extração de Tratamento) representa valor positivo da extração e *n* (Extração de Não Tratamento) representa valor negativo da extração.

Verdadeiro Positivo (VP) define que uma quantidade *N* de tratamentos pertinentes à Anemia Falciforme extraídas do documento é tratamento e foi extraída adequadamente. Logo, Verdadeiro Negativo (VN) é o inverso, não é tratamento e não foi extraída. Falso Positivo (FP) não é tratamento, mas foi erroneamente extraído do documento e classificado como tratamento e Falso Negativo (FN) é tratamento, mas não foi corretamente extraído.

Tabela 2.6 - Matriz de confusão de duas classes (Tratamento/Não Tratamento).

Condição Atual (Avaliação Especialista) / Resultado da Extração Automática	*Tratamento (P)*	*Não Tratamento (N)*
Extração de Tratamento (p)	VP (Verdadeiro Positivo)	FP (*não tratamento, mas é extraído*)
Extração de não Tratamento (n)	FN (*tratamento, mas não é extraído*)	VN (Verdadeiro Negativo)

A Equação (2-7) representa o cálculo da porcentagem de acerto a partir dos tratamentos e não tratamentos que foram extraídas; e a Equação (2-8) representa o cálculo da percentagem dos tratamentos que foram extraídos em relação ao total dos tratamentos:

$$Precisão = \frac{VP}{VP + FP} \tag{2-7}$$

$$Revocação = \frac{VP}{VP + FN} \quad (2\text{-}8)$$

Precisão e revocação são medidas utilizadas no projeto da Anemia Falciforme para avaliar a classificação das sentenças e a extração dos termos.

Outras medidas de desempenho utilizadas na classificação e extração: ***Medida-F*** e Acurácia, que segue:

Medida-F (*F-Measure*): baseia-se da média harmônica ponderada entre a Precisão e a Revocação (Equação (2-9)). F_β mede a eficácia da recuperação em relação ao valor atribuído a Beta (β). Pesos frequentemente utilizados para β são: F_2 (revocação, que é o dobro da precisão) e $F_{0,5}$ (precisão, que é o dobro de revocação). A precisão tem peso maior para valores $\beta < 1$, enquanto que $\beta > 1$ favorece a revocação. Nas Equações (2-9) e (2-10) a seguir considera-se P = Precisão e R = Revocação.

$$Medida\ F_\beta = \frac{(1+\beta) \times (P \times R)}{(\beta \times P + R)}, onde\ \beta = \frac{1-\alpha}{\alpha} \quad (2\text{-}9)$$

A relação entre a $Medida\text{-}F_\beta$ e a medida de eficiência é: $F_\beta = 1 - E$. Quando a precisão e revocação têm o mesmo peso (β = 1) a medida é $Medida\text{-}F_1$, também conhecida como $Medida\text{-}F$ tradicional ou $F\text{-}Score$ balanceada, Equação (2-10).

$$Medida\ F = \frac{2 \times P \times R}{P + R} \quad (2\text{-}10)$$

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

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (2\text{-}11)$$

Será aplicada a $Medida\text{-}F$, acurácia, precisão e revocação para medir o desempenho global no âmbito deste projeto para avaliar a classificação e a extração.

Maiores detalhes sobre estas medidas pode ser encontradas em MATOS et al. (2009).

2.9 Considerações Finais

Neste capítulo foram apresentados os conceitos sobre Mineração de Textos divididos em quatro etapas principais: (i) coleta de documentos, que vão constituir a base textual, ou seja, determinar e selecionar o domínio de abrangência das técnicas de MT; (ii) pré-processamento, etapa responsável por obter uma representação estruturada dos documentos; (iii) extração de padrões, fase em que é possível aplicar técnicas de extração de conhecimento utilizando técnicas de forma semelhante ao processo tradicional de mineração de dados; e (iv) análise e avaliação dos resultados, etapa de avaliação do resultado gerado a partir dos passos anteriores.

Foram discutidos ainda, fundamentos de Extração de Informação (EI) para obter informações relevantes em dados não estruturados. Conhecimento este que será aplicado para extrair informações de artigos científicos. Para a extração e a identificação de termos são utilizadas três abordagens: (i) abordagem baseada em aprendizado de máquina, que utiliza classificadores para separar ou identificar sentenças de interesse; (ii) abordagem baseada em regras, que é utilizada para identificar padrões de extração com expressões regulares; e (iii) abordagem baseada em dicionário, que utiliza informações de um dicionário para auxiliar na identificação dos termos ou das entidades no texto.

Por conseguinte foram apresentadas técnicas de Processamento de Língua Natural (PLN) que são extremamente úteis para aplicação em processamento de textos, e ainda conhecimentos sobre Aprendizado de Máquina, em que foi enfocado o aprendizado indutivo, mais especificamente em aprendizado supervisionado, que classifica novos exemplos a partir do treinamento de expressivos exemplos. AM será utilizado para classificar as sentenças dos artigos científicos (tarefa de classificação).

Para finalizar foram apresentadas as principais métricas de avaliação para analisar os resultados gerados a partir da Mineração de Textos e outras medidas de avaliação e desempenho.

Para atender e ou validar a hipótese deste trabalho, o presente envolve técnicas que utilizam abordagens para extração de informação.

Na próxima seção são apresentados os trabalhos relacionados que são baseados nestas técnicas.

Capítulo 3

TRABALHOS RELACIONADOS

Neste capítulo são descritos os trabalhos científicos envolvendo o conceito de extração de informação de artigos científicos do domínio biomédico. Os principais trabalhos relacionados a este projeto de mestrado serão detalhados e comparados.

Na Tabela 3.1 são resumidos os trabalhos identificados como relacionados ao enfoque deste trabalho de pesquisa em nível de mestrado. São trabalhos encontrados na literatura que extraem informação de artigos completos do domínio biomédico, os quais são detalhados nas próximas seções.

Nota-se que a maioria dos trabalhos que utiliza as abordagens baseadas em dicionário e regras, possui o enfoque de extrair entidades de genes e proteínas e utilizam precisão e revocação como medidas de avaliação. Porém, nenhum destes trabalhos relacionados trata da extração de termos relacionados a “tratamentos” no domínio biomédico tal como no domínio da doença Anemia Falciforme ou de qualquer outra doença.

O trabalho de Matos (2010) adotou a estratégia de extrair informação de efeitos negativos (da doença e do tratamento) e efeitos positivos (do tratamento) utilizando três abordagens: aprendizado de máquina, dicionário e regras. A primeira abordagem foi utilizada para classificar as sentenças em sentenças que possuíam algum efeito positivo ou negativo, das sentenças que não tinham nenhum efeito. As sentenças que possuíam algum indício de ter efeito foram posteriormente utilizadas a fim de identificar termos relacionados a efeitos positivos e negativos. Para isso, foram utilizadas as outras duas abordagens: dicionário e regras. Com o uso destas duas abordagens de extração de informação, Matos (2010) obteve na extração de

informação os seguintes desempenhos: 74,7% de precisão, 87,06% de revocação e 80,43% de medida-F.

O trabalho proposto nesta dissertação utiliza as mesmas abordagens de Matos (2010), a saber: aprendizado de máquina, dicionário e regras. Contudo, foi utilizada uma estratégia diferente de extração de informação. Este trabalho se concentrou em extrair informação de tratamentos, informação esta não extraída por Matos (2010). As informações sobre tratamento encontram-se concentradas em determinados parágrafos, diferentemente dos efeitos negativos e positivos abordados por Matos (2010) que se encontram espalhados por todo o texto. Foi observado que nos parágrafos que continham algum tratamento sempre existiam complicações (seja efeitos positivo ou negativo). Nesse sentido, a estratégia adotada foi classificar informação não por sentença como realizado por Matos (2010), mas por parágrafo. O critério para classificar os parágrafos é a presença das complicações (efeitos negativos e positivos) nos parágrafos. A presença das complicações nos parágrafos é um indicio que o parágrafo provavelmente tenha também um termo de tratamento. Em seguida, os tratamentos são extraídos somente nos parágrafos classificados como tendo um possível tratamento.

Para entendimento da Tabela 3.1 foi utilizada a seguinte terminologia: D significa Dicionário; R significa Regras; AM significa Aprendizado de Máquina; e POS significa etiquetador *Part-of-Speech.*

Tabela 3.1 - Trabalhos relacionados que extraem informação de artigos.

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)
Yang et al. (2009)		x[3]		Proteína	BioPPIExtractor	Extrair informação	Sim	Resumos Prec. 55,4% Rev. 41,6%
Yang et al. (2009)			x[3]	Proteína	BioPPISVMExtractor	Extrair informação	Sim	Resumos Prec. 49,2% Rev. 71,8%
MATOS (2010)	x	x	x	Complicação e Benefício da Anemia Falciforme	SCAeXtractor	Povoar um banco de dados	Sim	Artigos Acurácia 62,33% Prec. 74,75% Rev. 87,06% Med.F 80,43%

[1] Ontologia e expressões regulares, respectivamente, do sistema Textpresso.

[2] Prec. significa Precisão e Rev. significa Revocação.

[3] Método baseado em *Conditional Random Fields* (CRF).

3.1 ABGene

O ABGene é um sistema treinado em resumos de artigos do banco de dados do MEDLINE e testado em uma coleção de artigos completos do domínio biomédico selecionados aleatoriamente para reconhecer especificamente nomes de gene e proteína. É utilizado um etiquetador POS baseado em transformação que treina sentenças de resumos com ocorrência de gene marcada manualmente para induzir regras. Após isso, regras e dicionário são aplicados como pós-processamento.

O ABGene recebeu duas adequações para extrair informações de artigos completos (TANABE; WILBUR, 2002b). Inicialmente aplicou-se um classificador para operar na classificação em nível de sentença em artigos. Posteriormente realizou-se um pós-processamento com o intuito de extrair supostos grupos de nomes de gene e proteína.

O treinamento foi realizado com um conjunto de 1.000 artigos selecionados aleatoriamente do PubMed Central, totalizando 7.000 sentenças que foram selecionadas manualmente nos artigos. O experimento foi realizado com um conjunto de 2.600 sentenças, a fim de estimar como a conformidade de artigos completos afeta o desempenho do ABGene.

Alguns problemas na extração em artigos completos foram mencionados, tais como: falsos positivos como nomes de reagentes químicos são mais limitados em resumos, e vários falsos negativos que são observados em figuras e tabelas (TANABE; WILBUR, 2002a). Para resolver este problema e filtrar os falsos positivos e falsos negativos, foram utilizadas algumas técnicas:

Falsos positivos: dicionário (contendo termos biológicos e termos não biológicos) e regras foram utilizados para eliminar os falsos positivos. Expressões regulares foram produzidas para remover drogas com sufixos comuns;

Falsos negativos: dicionário (nomes simples e compostos) foi construído a partir do banco de dados LocusLink e do *Gene Ontology*, aprendizado de máquina (nomes com baixa frequência de trigramas foram selecionados; palavra de contexto foi gerada automaticamente por um algoritmo de probabilidade, que indica a probabilidade de nomes de genes adjacentes aparecerem no texto), e regras (expressões regulares adicionais foram criadas para permitir o casamento de padrões, as quais foram utilizadas para recuperar os falsos negativos).

Também foi utilizado o aprendizado Bayesiano para encontrar a probabilidade de um documento conter um nome de gene ou proteína, podendo assim não extrair informação de documentos que não contêm nomes relacionados. Para isso, documentos que contêm nomes de gene e proteína foram usados para treinamento. Na classificação de novos documentos, documentos com valores de similaridade abaixo de um limiar foram eliminados.

3.2 BioRAT

BioRAT (***Bio****logical* ***R****esearch* ***A****ssistant for* ***T****ext mining,* http://bioinf.cs.ucl.ac.uk/biorat/) é uma ferramenta para extração de informação que recupera e analisa informação de resumos e artigos completos na área biomédica

(CORNEY et al., 2004). A pesquisa por resumos e artigos completos é realizada a partir de uma consulta determinada pelo usuário no banco de dados da PubMed, e após recuperar os documentos relevantes, o sistema extrai ocorrências interessantes, que possam ser posteriormente armazenadas automaticamente em um banco de dados.

A extração de informação do BioRAT é baseada na coleção de ferramentas nomeada como GATE (***G****eneral* ***A****rchitecture for* ***T****ext* ***E****ngineering*), desenvolvida pela Universidade de Sheffield. O GATE é utilizado para etiquetar as palavras (POS) para posteriormente serem aplicados filtros para excluir verbos que não são proteínas. Dois componentes do GATE são utilizados: *gazetteers (*permite identificar palavras ou frases relacionadas a genes e proteínas*)* e *templates* (permite extrair informação automaticamente a partir de padrões textuais).

No experimento realizado, BioRAT foi comparado com o sistema de extração de informação SUISEKI (BLASCHKE; VALENCIA, 2002). O sistema SUISEKI utiliza conhecimento estatístico como a frequência de palavras que ocorrem em uma frase. Os *frames* de SUISEKI, comparáveis aos *templates* do BioRAT, contêm padrões relacionados a substantivos e verbos, porém não reconhecem conjunções, adjetivos ou outras classes de palavras.

Para avaliar o BioRAT foi utilizado o DIP (*Database of Interacting Proteins*) (XENARIOS, 2000) 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 *benchmark* 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 os 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 25,06% do corpo do artigo e 18% do resumo. O sistema BioRAT obteve maior precisão (55,07%) nos resumos e obteve 51,25% em artigos completos. Este fato ocorreu devido as deficiências no conjunto de *templates* usado pelo BioRAT (CORNEY et al., 2004).

3.3 Bremer et al. (2004)

Bremer et al. (2004) desenvolveram um sistema integrado que combina dicionários (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 (PubMed ID) e de quatros seções (resumo, introdução, materiais e métodos, resultados e discussão) para identificar o artigo selecionado e a seção de onde as informações foram extraídos.

Dois dicionários foram 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 e proteína. O dicionário de gene e proteína foi construído a partir de vários bancos de dados existentes como o LocusLink, o SWISS-PROT, dentre outros. O dicionário de sinônimos contém variações de sinônimos (e.g., *inhibit → inhibits, inhibition, inhibited*), informações contextuais, tais como prefixos e sufixos (e.g., *kinase, phosphate, receptor*) e verbos de interação que foram desenvolvidos 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 e proteína. A partir das sentenças reconhecidas, um conjunto de regras padrão foi construído para extrair genes. As regras foram baseadas na combinação de nomes de gene e proteína, preposições e palavras-chave que indicam o tipo de relação entre genes. Foram desenvolvidos também padrões usando substantivos e verbos na forma passiva e ativa.

A extração de informação dividiu-se em quatro passos: (i) toquenizar o texto em sentenças; (ii) analisar sentenças para identificar frases com substantivo e verbo; (iii) selecionar sentenças que contêm genes usando dicionários de nome de gene e proteína, e sinônimos; (iv) extrair genes 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.

O software GetItRight comercial (disponível em http://www.cthtech.com/) foi utilizado para auxiliar no desenvolvimento de *scripts* para conectar e baixar artigos completos automaticamente no formato HTML. Realizou-se um pré-processamento

para converter o arquivo HTML para o formato XML, e no XML (Figura 3.1) 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.

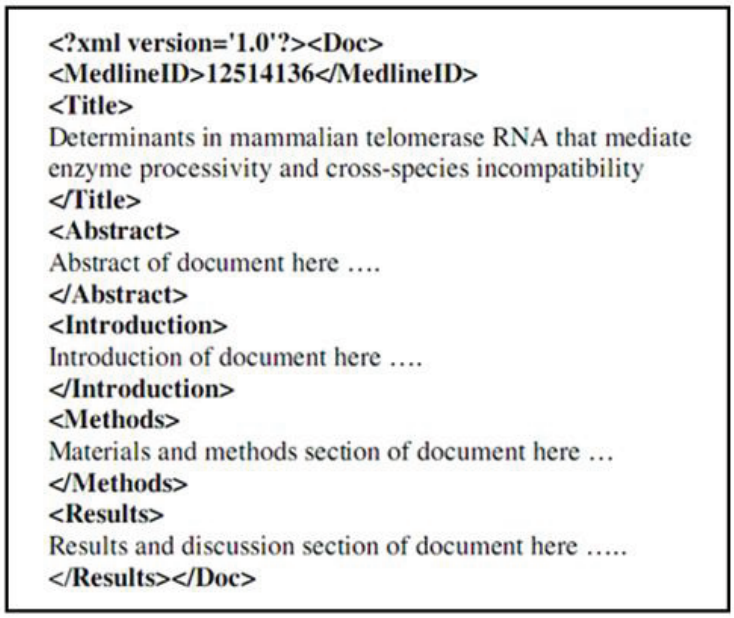

Figura 3.1 - Exemplo de um documento XML com etiquetas de quatro seções.

Fonte: **Bremer et al. (2004).**

Foram selecionados artigos no domínio da biologia molecular e da biomedicina, mais particularmente 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 obtidas foram, respectivamente, 63,5% e 37,3%. A baixa precisão foi devido aos erros de padrão na identificação de nomes de gene e proteína em algumas sentenças e na falta de padrões com palavras compostas para identificar sentenças complexas. A baixa revocação foi devido à variedade da coleção de 20 artigos completos publicados em 20 jornais diferentes.

3.4 Continuação do Trabalho de Bremer et al. (2004)

Natarajan et al. (2006) implementaram um processo de mineração de textos, a partir da extração de informação de artigos científicos desenvolvida por Bremer et al. (2004). Conforme pode ser observado na Figura 3.2, os artigos são descarregados no formato HTML sem imagem e convertidos para o formato XML, utilizando a ferramenta para download GetItFull (NATARAJAN et al., 2006); termos são extraídos do LexiQuestMine utilizando padrões; posteriormente, no módulo Curador, os termos são normalizados utilizando um dicionário de sinônimos para serem adiante armazenados em um *data warehouse*. Os dados armazenados são utilizados em uma rede de interação para visualizar as interações de gene e proteína.

Natarajan et al. (2006) concluíram que a extração automática de informações a partir de literatura biológica assegura desempenhar um papel cada vez mais importante na descoberta de conhecimento biológico.

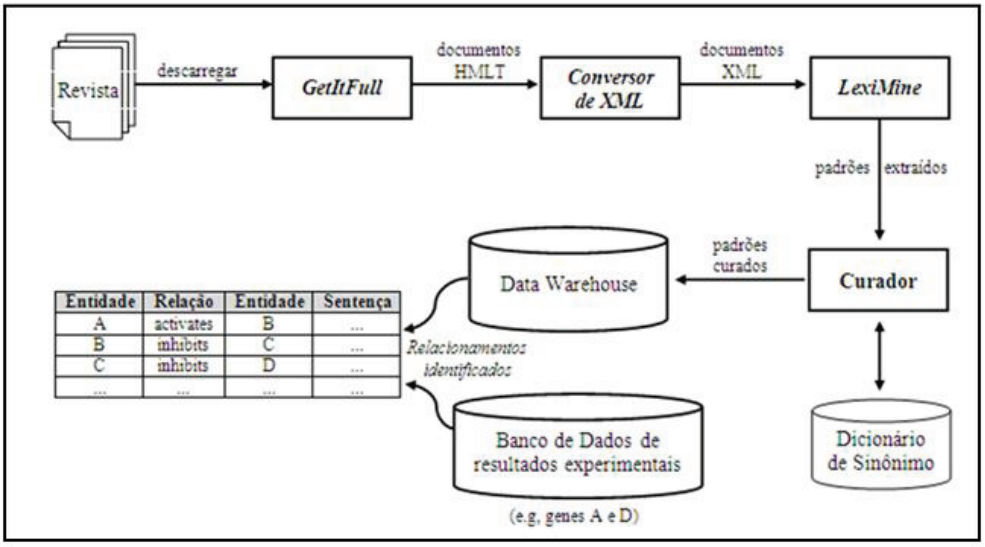

Figura 3.2 - Processo de extração de padrão e *data warehouse*.

Fonte: **Adaptado de Natarajan et al. (2006).**

3.5 Pharmspresso

O sistema Pharmspresso (GARTEN; ALTMAN, 2009) (http://pharmspresso.stanford.edu) extrai informação sobre genes, drogas e polimorfismos de artigos completos da literatura referente à área da farmacogenômica. É um sistema de recuperação de informação que utiliza da extração de informação para recuperar as informações de acordo com a consulta determinada pelo usuário. Os principais pontos fortes do Pharmspresso são a sua capacidade de processar artigos em texto completo em formato PDF utilizando expressões regulares, e o índice de seu conteúdo é baseado em uma ontologia de conceitos-chave. Fornece um motor de busca entidades importantes e relações semânticas entre eles.

O Pharmspresso é baseado no sistema Textpresso (http://www.textpresso.org), pacote de código aberto desenvolvido por Müller, Kenny e Sternberg (2004) (GARTEN; ALTMAN, 2009).

Textpresso (MULLER HM, 2004) é um sistema de pesquisa baseado em um conjunto de expressões regulares para encontrar informação a partir da consulta do usuário em artigos fornecidos no formato PDF. Utiliza uma ontologia que contém categorias de frases e palavras de interesse biológico que foi construída com ajuda de especialista. A ontologia inclui 35 categorias de dois tipos: (1) entidades biológicas e (2) relações entre entidades. As categorias foram importadas de *Gene Ontology*.

Na Figura 3.3 é mostrado o processo de recuperação e extração de informação realizada pelo sistema. Primeiramente os artigos PDF são baixados, depois convertidos em formato textual e toquenizado em palavras e sentenças individuais. Após, o texto é analisado para identificar palavras ou frases que são membros de categorias específicas de uma ontologia. Estas são marcadas e indexadas para serem utilizadas em pesquisas futuras realizadas por palavras-chave definidas pelo usuário.

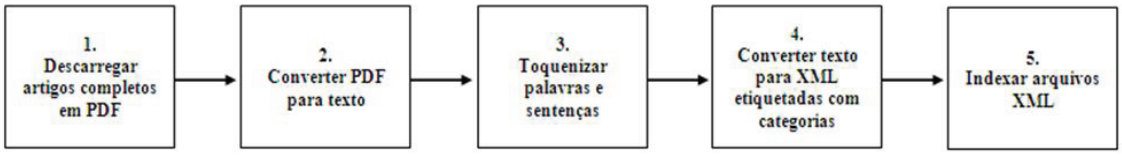

Figura 3.3 - Processo para recuperar e extrair informação do Pharmspresso.
Fonte: Adaptado de Garten e Altman (2009).

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

A avaliação do sistema foi realizada por 11 avaliadores da literatura farmacogenética e observou a capacidade de extrair informações sobre genes, drogas e polimorfismos de 45 artigos. Nestes artigos, constavam 178 genes, 191 drogas e 204 polimorfismos, e o Pharmspresso 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 de acerto é de somente 50,3% dessas associações.

Problemas com variações de nomes de gene foram encontrados, causando falsos positivos. Uma de suas limitações é que o Pharmspresso só funciona em um corpus de artigos relevantes pré-definido, e não em toda a literatura existente. Em um trabalho futuro, o autor comenta que o Pharmspresso poderá incluir um corpus maior e o resumo poderá ser utilizado quando o texto completo não estiver disponível. Ademais, permitir que o sistema recupere automaticamente a literatura referente, usando a ontologia proposta, extraindo os fatos de interesse e usá-las para preencher um banco de dados de interações.

3.6 BioPPIExtractor

BioPPIExtractor é um sistema de extração de interação proteína-proteína para literatura biomédica desenvolvido por (YANG; LIN; WU, 2009). Este aplica o modelo *Conditional Random Fields* (CRF) para marcar os nomes de proteínas no texto biomédico, em seguida, usa um *link grammar parsing* para identificar as funções sintáticas em sentenças, e em seguida extrai interações destas funções sintáticas.

O sistema é composto de seis passos principais para extrair informações de interação das sentenças de entrada: *"pronoun resolution", "protein name recognition", "interaction word recognition", "link grammar parsing", "complex sentence processing",* e *"interaction extraction"*, conforme apresentado na Figura 3.4.

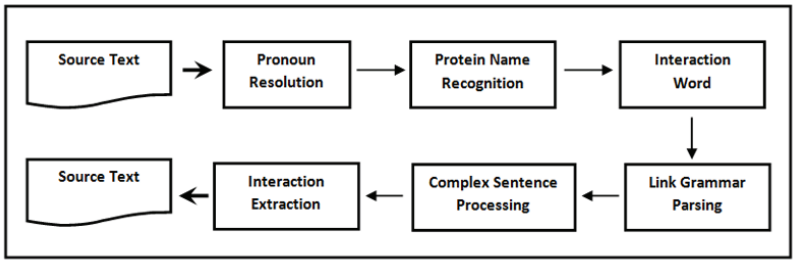

Figura 3.4 - Passos do sistema BioPPIExtractor.

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

- "*Pronoun Resolution*": substantivo e frase nominal no texto são identificados usando GENIA *Tagger* (TSURUOKA; TSUJII, 2004), que é ajustado especificamente para textos biomédicos, tais como resumos da MEDLINE;
- "*Protein name recognition*": é utilizado um método baseado em *Conditional Random Fields* (CRF) sendo um tipo de modelo probabilístico discriminativo mais frequentemente utilizado para a etiquetagem ou de análise de dados sequenciais, tais como texto em linguagem natural ou sequências biológicas. Esses têm sido recentemente aplicados à tarefa de descoberta de genes e proteínas. Maiores detalhes sobre CRF pode ser encontrados em (LAFFERTY; MCCALLUM; PEREIRA, 2001).

No modelo baseado em CRF são utilizados alguns recursos, tais como: todas as palavras são escritas em minúscula para que a dimensão dos recursos possa ser diminuída e a perda de informação pode ser compensada por meio de sua combinação com outras funções; *Part-of-speech features*, (aqui GENIA *Tagger* é aplicado novamente); dentre outros.

- "*Interaction word recognition*": no sistema BioPPIExtractor, uma sentença é considerada para incluir uma *protein–protein interaction* (PPI) somente se a sentença tem pelo menos dois nomes de proteínas e uma palavra de interação (por exemplo, "*bind*", "*down-regulate"*, "*interact"* e assim por diante). O dicionário para reconhecimento de palavras de interação contém um total de aproximadamente 150 entradas, incluindo verbos e suas variantes de

interação (por exemplo, o verbo interação "*bind*" tem variações como "*binding"* e "*bound"*).

Em BioPPIExtractor é utilizado um *link grammar parser*, e no módulo de extração de interação, este extrai interações de sentenças simples produzido pelo módulo "*complex sentence processing*".

- "*Interaction extraction"*: o *link grammar* – identifica interações entre proteínas, e sua abordagem baseia-se em links e caminhos entres várias entidades nomeadas como genes e nomes de proteínas (DING et al., 2003). *Link grammar* considera um caso de profunda análise baseada do conteúdo das diversas funções sintáticas das frases como seus sujeito (S), verbos (V), objetos (O) e modificando frases (M), bem como suas combinações linguísticas significativas, como a S-V-O, S-V-M, para encontrar e extrair interações proteína-proteína. Apenas no caso de uma função sintática (ou combinação significativa) ter pelo menos dois nomes de proteína e uma palavra interação é possível uma interação proteína-proteína ser extraída. Contudo, BioPPIExtractor não considera extrair a interação a partir de combinações de S-O e S-M desde que o autor descobriu que iria introduzir muitos erros de extração.

O sistema BioPPIExtractor foi testado apenas em resumos de artigos do MEDLINE e sua avaliação experimental foi comparada com outros sistemas do estado arte: BioRAT (CORNEY et al., 2004) e IntEx (AHMED et al., 2005). Esta indica que sistema BioPPIExtractor alcança melhor desempenho.

A Tabela 3.2 apresenta a avaliação de interação de 229 resumos do MEDLINE, e compara o BioPPIExtractor com o sistemas BioRAT e IntEx.

Tabela 3.2 – Avaliação - BioPPExtrator.

Sistemas	**Revocação**	**Precisão**
BioPPIExtractor	41.62%	55.41%
BioRAT	20.31%	55.07%
IntEx	26.94%	65.66%

3.7 BioPPISVMExtractor

Sistema que também extrai informação sobre interação proteína-proteína para literatura biomédica, desenvolvido por (ZHIHAO YANG; HONGFEI LIN; LI, 2009) o mesmo autor do BioPPIExtractor. Este é baseado em *Support Vector Machines* (SVM) e utiliza alguns recursos ricos como palavras chaves, características chaves, característica de distância em nome de proteínas e um caminho para a classificação SVM. Além disso, utiliza o *link grammar* para identificar interações entre proteínas.

Neste sistema, o corpus IEPA (J. DING, 2002) é utilizado como o conjunto de treinamento para o classificador SVM e o corpus DIP é utilizado como conjunto de teste. O classificador SVM treinado é utilizado para identificar pares de proteínas em uma sentença que tem biologicamente relação relevante entre eles. A Figura 3.5 ilustra a arquitetura do BioPPISVMExtractor.

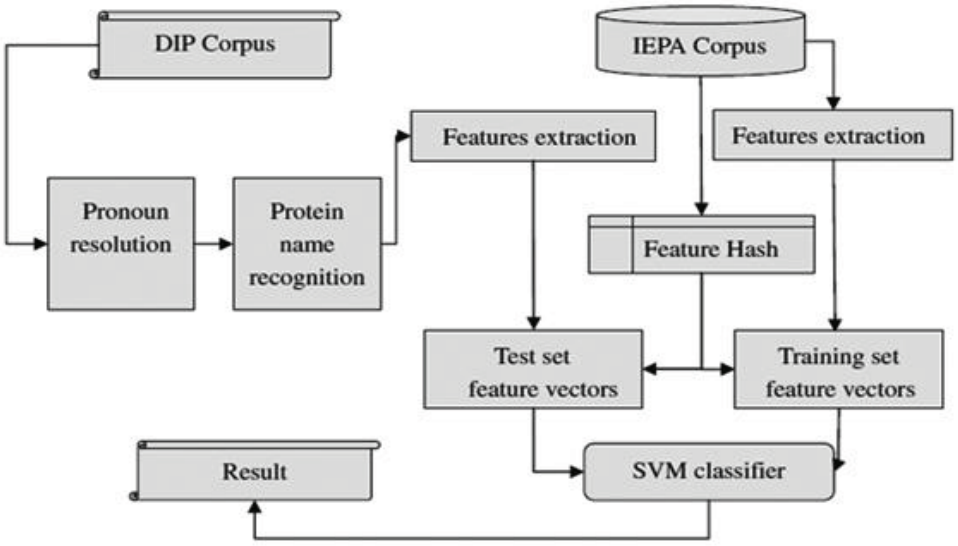

Figura 3.5 - Arquitetura do BioPPISVMExtractor.

***Fonte*: Adaptada de Yang, et al.(2009).**

- "*Pronoun Resolution*": substantivo e frase nominal no texto são identificados usando GENIA *Tagger* (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA) que é ajustado especificamente para textos biomédicos, tais como resumos da MEDLINE;
- "*Protein name recognition*": é aplicado o método baseado em *Conditional Random Fields* (CRF) e suas características são comumente utilizadas no sistema do mesmo autor, o BioPPIExtractor;

- Modelo SVM: um classificador SVM é treinado para reconhecer interações proteína-proteína em textos biomédicos. O SVM é um classificador binário desenvolvido por VAPNIK (1995). Neste experimento foi utilizado o pacote SVM-*Light* (JOACHIMS, 1999). A penalidade do parâmetro C na definição do SVM é um parâmetro muito importante, uma vez que controla a troca entre o erro e a margem de treinamento. Este parâmetro foi configurado como valor padrão. O pacote SVM-*Light* contribuiu significativamente na criação do valor padrão para este parâmetro.
- "*Feature selection*": as seguintes características são exploradas para o classificador SVM:
    - Palavras: palavras de dois nomes de proteínas, palavras entre dois nomes de proteínas, palavras envolvendo dois nomes de proteínas;
    - Distância do nome da proteína: quanto menor a distância (número de palavras) entre dois nomes de proteína, é o mais provável que as duas proteínas têm relação de interação. Portanto, a distância entre dois nomes de proteína é escolhida como um recurso.
    - Palavra-chave: para identificar as palavras-chave em textos, foi construído manualmente um dicionário para reconhecimento de palavras de interação com cerca de 500 entradas, que incluem os verbos interação e suas variantes (por exemplo, o verbo interação "*bind*" tem variantes como "*binding*" e "*bound*").
    - *Link path*": a ideia básica do *Link path* é conectar pares de palavras em uma sentença com vários links. Existem vários tipos de conectores, e conectores podem apontar para a direita ou para a esquerda. A sentença válida é aquela em que todas as palavras estão ligadas de alguma forma. Se existe um caminho de ligação entre as dois nomes de proteínas, o valor da característica do caminho das duas proteínas é definido como "*Link_YES*", caso contrário, "*Link_NO*". O *Link Grammar parser* usada em BioPPISVMExtractor foi desenvolvido por (GRINBERG; LAFFERTY; SLEATOR, 1995).

O experimento do BioPPISVMExtractor foi comparado com os sistemas BioRAT (CORNEY et al., 2004), IntEx (AHMED et al., 2005) e BioPPIExtractor (YANG; LIN; WU, 2009). Na avaliação de interação foram utilizados 229 resumos do MEDLINE. A Tabela 3.3 apresenta o resultado.

Tabela 3.3 - Avaliação – BioPPISVMExtractor.

Sistemas	Revocação	Precisão	Medida-F
BioPPISVMExtractor	71.83%	49.28%	58.46%
BioPPIExtractor	41.62%	55.41%	47.53%
BioRAT	20.31%	55.07%	29.68%
IntEx	26.94%	65.66%	38.20%

O autor comenta que, como pode haver muitos falsos positivos introduzidos pelo método baseado em SVM, o resultado do BioPPISVMExtractor de 49,28% é uma precisão bem aceitável.

3.8 Considerações Finais

Na literatura são encontrados vários trabalhos relacionados que extraem informação, alguns com objetivos diferentes, tais como: (i) povoar um banco de dados, (ii) destacar as sentenças de acordo com a consulta do usuário ou (iii) extrair informação. A maioria destes trabalhos são baseados em entidades de genes e proteínas, utilizam precisão e revocação como medidas de avaliação e utilizam-se de aprendizado de máquina, regras ou dicionário como abordagem para extração de informação.

Apesar dos diferentes trabalhos que extraem informação, nenhum dos trabalhos trata da extração de termos relacionados a “tratamentos” do domínio biomédico. A utilização de qualquer trabalho existente se torna inviável pelo fato de que não possuem enfoque de (i) extrair informação sobre tratamento de doenças, (ii) utilização de dicionário para garantir alta revocação de termos conhecidos, a (iii) extração de novos termos e (iv) utilização das três abordagens de informação conjuntamente.

Para preencher esta lacuna, este projeto de mestrado possui o objetivo de propor um processo baseado em parágrafos para a extração de tratamentos de artigos científicos do domínio biomédico.

Capítulo 4

PROCESSO PROPOSTO PARA EXTRAÇÃO DE TERMOS DE TRATAMENTO

Neste capítulo é apresentado e ilustrado o processo proposto nesta dissertação de mestrado para a extração de termos de tratamento e as abordagens utilizadas neste processo, a saber: abordagem de aprendizado de máquina na fase de classificação de sentenças por agrupamento de parágrafos, e abordagens de dicionário e regras na fase de extração de termos.

4.1 Visão Geral do Processo de Extração

Neste capítulo é apresentado e ilustrado o processo de pré-processamento de informações não estruturadas que visa extrair informações relevantes sobre termos de tratamento de doenças de artigos científicos do domínio biomédico. Nesta dissertação, o conceito de processo refere-se a um conjunto de passos que são utilizados para se alcançar um objetivo comum, ou seja, para se obter o resultado final desejado de extração de informação. Em cada um dos passos são empregadas técnicas que permitem atingir um resultado parcial para se obter a extração de informação sobre termos de tratamento.

A Figura 4.1 exemplifica o propósito geral deste processo, no qual é apresentado o processo de extração de informação sobre o tratamento do domínio biomédico. Como observado, dado um fragmento, a extração baseia-se em

identificar a sentença que contém a informação de interesse e posteriormente obter somente as partes desta sentença que indicam um tratamento, estruturando-as até o seu armazenamento final no banco de dados.

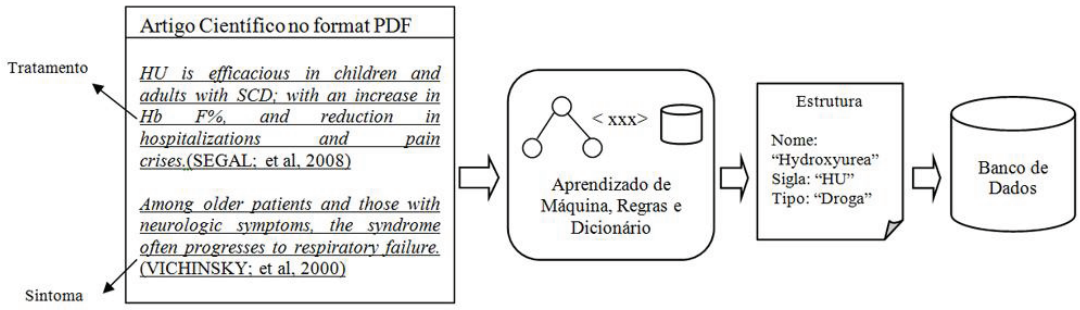

Figura 4.1 - Processo de extração de informação no domínio biomédico.

Considera-se como hipótese deste trabalho que na maioria dos casos os termos de tratamento ocorrem em uma mesma sentença que possui um termo de complicação (i.e. de um efeito negativo da doença, como exemplo dor e febre) ou em sentenças próximas de uma complicação em um mesmo parágrafo. Portanto, o parágrafo é considerado neste processo como uma unidade com conteúdo de informações centralizado, no qual se localiza a informação de interesse (i.e. termos de tratamento). Esta hipótese foi baseada em uma análise empírica de artigos da doença Anemia Falciforme, para os quais se descobriu que: (i) ~10% dos tratamentos apareceram na sentença anterior ou posterior à sentença na qual uma complicação foi encontrada; (ii) ~90% dos tratamentos apareceram na mesma sentença na qual a complicação foi encontrada; e (iii) muitos poucos tratamentos ocorreram em sentenças mais distantes.

O processo proposto para a extração de termos de tratamento é independente da doença e os passos do processo não dependem de uma doença em si, mas foram aplicados a uma doença particular (i.e. Anemia Falciforme).

Na Figura 4.2 são apresentadas as etapas do processo proposto para efetuar a extração de informação sobre "tratamentos" em artigos científicos do domínio biomédico. Essencialmente, utiliza-se de dois classificadores, chamados de C1 (Classificador 1) e de C2 (Classificador 2), ambos com o objetivo de separar as sentenças de interesse que provavelmente terão, respectivamente, termos de complicação e termos de tratamento, das sentenças que possivelmente não terão nenhum destes termos. Classificação será usada no processo proposto como um

filtro que selecionará apenas as sentenças de interesse, diminuindo o custo de análise das sentenças de um artigo na fase posterior de extração de informação. Usa-se também dicionários e regras para identificar e extrair as partes de interesse dentro das sentenças pré-selecionadas. Uma visão geral do processo é apresentada na Figura 4.2:

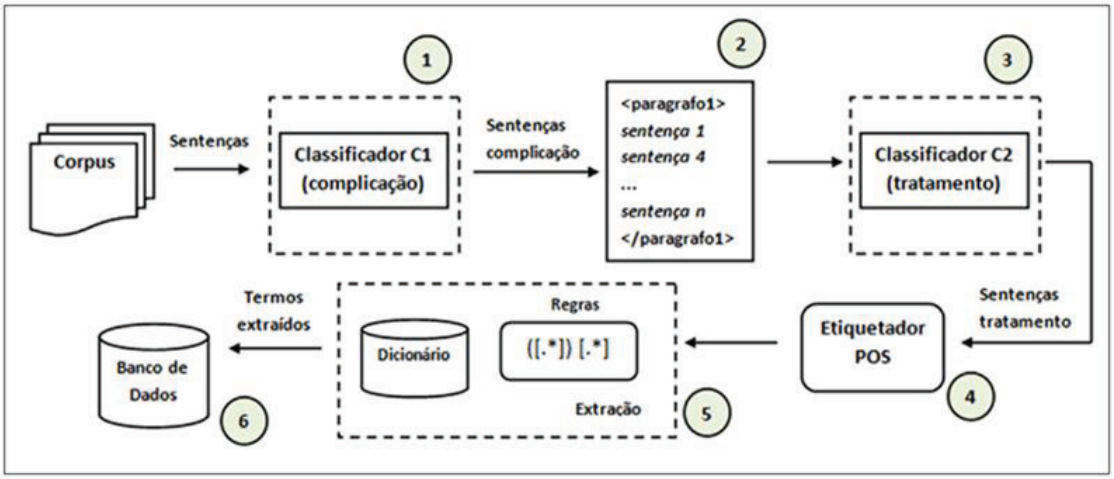

Figura 4.2 - Processo proposto para extração de tratamentos.

O processo proposto é composto por seis passos. Com o intuito de encontrar termos de tratamento, primeiramente as sentenças são classificadas em sentenças com termos de complicação e em sentenças que não possuem complicação (passo 1). No passo 2, as sentenças de interesse são agrupadas adicionando-se as demais sentenças que participam do mesmo parágrafo de uma sentença previamente selecionada e no passo 3 as sentenças são classificadas em sentenças com termos de tratamento e em sentenças que não possuem tratamento. No passo 4, as sentenças com termos de tratamento são etiquetadas conforme a sua classe gramatical. Após esta etapa, é realizado o processo de extração de informação dos termos de tratamento (passo 5), utilizando as abordagens de dicionário e regras. Ao final do processo, os termos relevantes e interessantes que foram extraídos são armazenados em um banco de dados relacional (passo 6).

4.2 Entrada de dados

O processo proposto lida com artigos científicos que estão originalmente no formato não estruturado PDF e escritos no idioma inglês. Nesta fase, a entrada de

dados é baseada no fornecimento de documentos no formato PDF. É necessário efetuar a conversão deste formato para um formato que permita o processamento textual. Dois formatos semiestruturados são aceitos para dar suporte ao processo de extração de informação: XML e TXT. A conversão do formato PDF para o formato XML é realizada usando a ferramenta SCA-Translator desenvolvida por Carosia e Ciferri (2010). Na conversão, 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. Essa 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.

Em contrapartida, a conversão do formato PDF para o formato TXT é realizada de forma manual. A limitação deste formato é que cada linha deve formar uma sentença e no final da sentença deve ser informado o nome da seção entre parênteses a qual a sentença pertence. Na Figura 4.3 é ilustrado um exemplo de documento TXT em que atende o requisito para o processamento textual.

```
Values within 3 months after a blood transfusion were excluded. (General)
Oxygenation significantly improved with transfusion. (Results)
The HU dose was given orally once a day, initially at 20 mg/kg. (General)
Daily hydroxyurea doses reported each year are shown in Table 1. (Results)
```

Figura 4.3 - Exemplo de documento TXT.

4.3 Classificação de Sentenças

O primeiro estágio da extração de informação é a fase da classificação de sentenças, cujo propósito é a construção de um modelo de classificação adequado que melhor represente as particularidades das sentenças de treinamento, e com isso possa predizer qual a classe de uma nova sentença.

A classificação de sentenças é constituída por três fases: treinamento (Fase 1), teste (Fase 2) e uso do modelo (Fase 3). A classificação de sentenças é supervisionada, no qual os rótulos das classes são previamente conhecidos. Na

Fase 1, o classificador (ou modelo de classificação) é construído com o objetivo de descrever o conjunto de sentenças. O classificador é criado a partir da avaliação do conjunto de treinamento, e este conjunto é rotulado em classes predefinidas. Na Figura 4.4 são apresentadas as classes predefinidas relacionadas à doença Anemia Falciforme e suas respectivas sentenças.

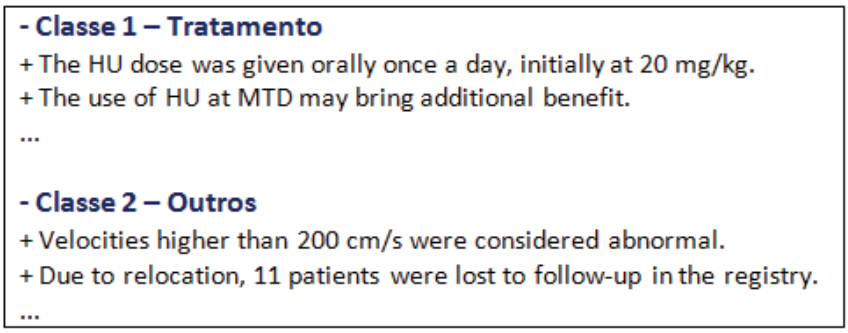

Figura 4.4 - Exemplo da estrutura dos arquivos de treinamento.

Após o modelo construído, é necessário avaliar se o modelo gerado é indicado para ser utilizado em sentenças em que o rótulo não é conhecido. Para isso, na Fase 2, sentenças que não foram utilizadas na fase de treinamento, são avaliadas com base na medida de desempenho acurácia. Para realizar o cálculo da acurácia, o rótulo da sentença testada é comparado com o rótulo da sentença classificada. O método de particionamento *10-Fold Cross-Validation* é utilizado para estimar a acurácia do classificador. Após a avaliação e análise das sentenças, o modelo criado é utilizado na Fase 3 para classificar as novas sentenças dos artigos, com o intuito de extrair informações relevantes do domínio biomédico.

Para atingir o objetivo do processo proposto, foram criados dois classificadores: C1 e C2, ambos com o objetivo de separar as sentenças de interesse que provavelmente terão, respectivamente, termos de complicação e termos de tratamento, das sentenças que possivelmente não terão nenhum destes termos.

Na Figura 4.5 é apresentado o processo de classificação de sentenças supervisionado em que é constituído por três fases: Coleta dos Dados (Fase 1), Pré-processamento (Fase 2) e Classificação (Fase 3).

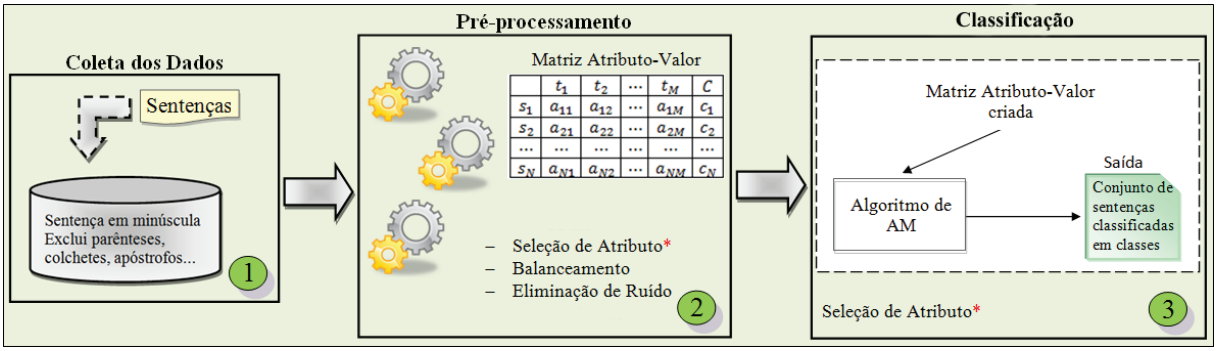

Figura 4.5 - Processo de classificação de sentenças.

Fonte: Adaptado de Matos (2010).

A Fase 1 constitui na aquisição de sentenças a serem utilizadas para o treinamento e o teste do classificador. Alguns procedimentos são empregados no conjunto de sentenças na Fase 1, tais como: todos os caracteres das sentenças são transformados em letras minúsculas; são removidos vírgula, ponto e vírgula, dois pontos, parênteses, colchetes, apóstrofos, sinal de mais ou menos (±) e excessos de espaço em branco.

As sentenças são estruturadas utilizando o modelo *bag-of-words* na Fase 2. Este comportamento é indicado para que as sentenças possam ser manipuladas por algoritmos de aprendizado de máquina. A matriz atributo-valor é elaborada utilizando a frequência mínima igual a dois para filtrar os atributos que ocorreram no mínimo duas vezes na sentença. Os atributos são compostos de 1 a 3 gramas. Também utiliza-se a medida binária, para a qual considera-se que o valor 1 representa a ocorrência do *n-grama* na sentença e o valor 0 a não ocorrência do *n-grama* na sentença. Técnicas de balanceamento das sentenças e de remoção de ruído, ainda, devem ser utilizadas, para respectivamente, balancear a distribuição das sentenças entre as classes e remover as sentenças que estejam dificultando o aprendizado.

Finalmente na Fase 3 é realizada a classificação das sentenças. Dois algoritmos estatísticos clássicos de aprendizado de máquina foram designados para serem avaliados na classificação das sentenças: *Support Vector Machine* (SVM) e *Naïve Bayes* (NB). Os modelos criados para cada algoritmo foram avaliados também com a medida de desempenho acurácia. Os modelos foram utilizados para classificar novas sentenças na Fase 3.

Na Figura 4.6 é mostrado um exemplo de sentenças da Anemia Falciforme que foram classificadas nas respectivas classes: classe "complicação" e classe

"outros" para o Classificador C1, e classe "tratamento" e classe "outros" para o Classificador C2.

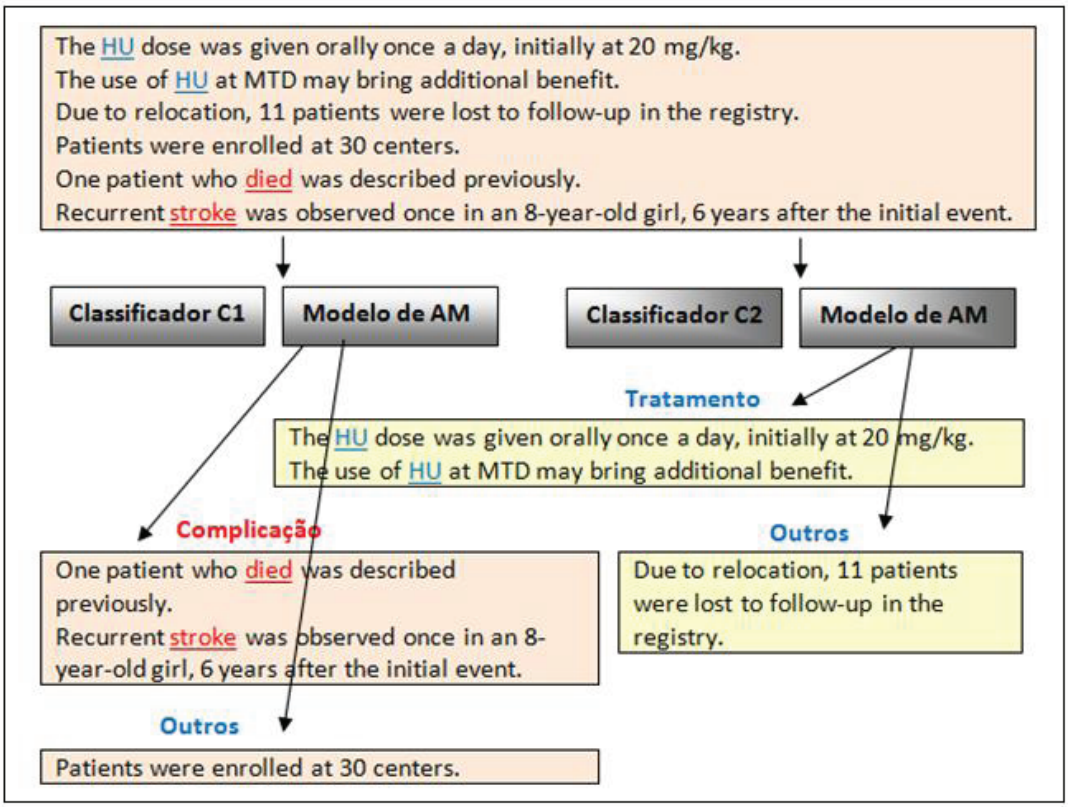

Figura 4.6 - Exemplo de sentenças da AF e as suas respectivas classificações.
Fonte: Adaptado de Matos (2010).

4.4 Identificação de termos relevantes

A partir das sentenças classificadas por um classificador, posteriormente, é possível realizar a identificação de termos relevantes em cada uma das sentenças de interesse (i.e., especificamente nas sentenças da classe tratamento). Para este propósito, duas abordagens para a extração de informação são utilizadas: dicionário e regras. O dicionário tem a função de identificar os termos validados e previamente conhecidos nas sentenças de interesse, com a finalidade de completar o tipo-relacionamento artigo/termo do banco de dados, ou seja, com a finalidade de armazenar os dados que relacionam um artigo científico e a ocorrência de um termo conhecido de tratamento. Por outro lado, a finalidade das regras é extrair automaticamente novos termos das sentenças de interesse e armazená-los no

dicionário. Os termos já existentes no dicionário não são armazenados novamente, pois a inserção de termos apenas é realizada com termos inexistentes no dicionário.

São apresentados a seguir exemplos destas duas abordagens de extração de informação a serem aplicadas no domínio biomédico.

4.4.1 Abordagem de extração de informação baseada em regras

A abordagem baseada em regras é utilizada para extrair automaticamente termos relevantes, mediante padrões encontrados nas sentenças de interesse (i.e., sentenças de tratamento). A técnica de processamento de língua natural *Part-Of-Speech* (POS) é utilizada para etiquetar as palavras em suas respectivas classes gramaticais (i.e., classificar nas classes substantivo, adjetivo, verbo, dentre outras). Nesta dissertação utiliza-se de expressões regulares e POS para formar as regras e com isso permitir uma melhor precisão no reconhecimento e na extração de informação em documentos não estruturados do domínio biomédico. O padrão das etiquetas utilizado foi o padrão Penn Treebank (MARCUS; MARCINKIEWICZ; SANTORINI, 1993). Na Tabela 4.1 é apresentado um exemplo de uma sentença etiquetada. As etiquetas DT significa artigo, JJ adjetivo, NN substantivo, NNP significa nome próprio, IN significa preposição e, VBD, VBN e VBG significam verbos.

Tabela 4.1 - Exemplo de sentença etiquetada.

The_**DT** mean_**JJ** HU_**NNP** dose_**NN** per_**IN** kilo_**NN** was_**VBD** calculated_**VBN** each_**DT** year_**NN** ._.

O etiquetador POS utilizado foi o etiquetador 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. Maiores detalhes podem ser encontrados em (TOUTANOVA; MANNING, 2000) e (TOUTANOVA et al., 2003).

Para descobrir padrões manualmente no conjunto de sentenças do domínio biomédico, mais especificamente da Anemia Falciforme, inicialmente foi necessário

analisar um subconjunto de 394 sentenças, pertencentes a 8 artigos, nas quais continham um requisito principal: termos de tratamento ou palavras chaves que indicassem tratamentos, tais como: *study, trial e experiment*. Ainda, foram removidas as sentenças que não continham palavras chaves que indicassem ou que continham o termo relevante. As sentenças remanescentes continham exatamente os termos de tratamento relevantes, termos estes que já haviam sido avaliados e considerados pelo especialista da área, no caso uma especialista da doença AF. O subconjunto resultante de 123 sentenças continha 168 termos de tratamento. Alguns termos estavam presentes em várias sentenças, enquanto alguns termos de tratamento foram encontrados numa mesma sentença mais que uma vez, tais como os termos *hydroxyurea* e *transfusion*. Já outros termos foram encontrados apenas uma vez, tais como: *bone marrow transplantation* e *penicillin*. Logo, o subconjunto resultante de sentenças foi processado, etiquetados pelo padrão POS, e ainda, para facilitar, foram agrupadas pelo seu grau de similaridade. Com toda esta análise foi possível identificar padrões para serem usados na formação das regras.

No processo proposto, duas estratégias de padrões POS são aplicadas para extrair informação das sentenças: 1) Verbo ou palavra representativa com POS e 2) somente POS. Entende-se por verbo ou palavra representativa uma informação que pode identificar um termo relevante na sentença. O motivo de criar duas estratégias de regras, é que para a primeira, o processamento é realizado em parte da sentença, ou seja, a expressão regular somente casará com a sentença se, e somente se, existir o termo representativo na sentença. Caso o termo representativo existir, padrões POS são utilizados para extrair o termo relevante somente em uma parte específica da sentença. 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. Para a segunda estratégia, o uso somente de POS, o processamento é realizado na sentença por completo, fazendo com que o padrão POS case com algum padrão POS descoberto e criado por meio da análise realizada previamente no subconjunto de sentenças. A seguir são detalhadas as duas estratégias de regras.

4.4.1.1 Estratégia 1

As regras da estratégia 1 baseada em verbo ou palavra representativa com POS foram subdivididas em 2 conjuntos: a) conjunto amplo contendo 9 regras,

algumas regras desenvolvidas genericamente e outras mais específicas. O desenvolvimento de regras mais específicas foi idealizada pela dificuldade em diminuir a quantidade de falsos positivos e melhorar a precisão; e b) conjunto enxuto contendo 2 regras mais genéricas. O desenvolvimento de regras mais genéricas para o conjunto enxuto foi efetuado agrupando e generalizando regras a partir de algumas regras do conjunto amplo. A divisão dos dois conjuntos foi criada para analisar se o conjunto amplo (formado por regras mais específicas) consegue extrair menos falsos positivos em relação ao conjunto enxuto (formado por regras mais genéricas).

Na Tabela 4.2 e Tabela 4.3 é apresentado o conjunto de regras da estratégia 1, desmembrando o conjunto de regras amplo e enxuto respectivamente. Os verbos e as palavras representativas são evidenciados em destaque. As regras a seguir foram desenvolvidas neste trabalho, e considera-se que as regras mais específicas são utilizadas para o domínio da doença da Anemia Falciforme, e as regras mais genéricas além de atender a doença da AF, também podem ser reutilizadas para a extração de informação com outras doenças.

As etiquetas com o símbolo \w significa uma sequência de letras, números e underline, o símbolo til (~) significa negação e o símbolo de interrogação (?) significa optativo(i.e., a etiqueta pode estar presente ou não). As etiquetas JJ significa adjetivo, IN preposição, NN substantivo, NNP significa nome próprio, NNS significa substantivo no plural, CC conjunção e, VBD, VBN e VBG significam verbos.

Tabela 4.2 - Verbo ou palavra representativa com POS – Conjunto Amplo.

Número	Expressão Regular
1	(?:[\w]*_IN) (?:[\w\-/\\]*)?([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS) (?:**treatment**_NN|**therapy**_NN)
2	(?:**were**_VBD|**was**_VBD|**while**_IN|**went**_VBD|**while**_NN)(?:[\w\-/\\]* **on**_IN) ([\w\-]*)
3	(?:**studies**_NNS|**study**_NN|**studied**_VBD|**studied**_VBN|**trial**_NN) (?:[\w\-/\\]*)?([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS) (?:**treatment**_NN|**therapy**_NN)
4	(?:**received**_VBN|**received**_VBD|**receiving**_VBG) (?:[\w\-/\\]*)?([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS) (?:**treatment**_NN|**therapy**_NN|**dose**_NN|**doses**_NNS)
5	(?:**doses**_NNS|**dosing**_VBG) **of**_IN ([\w\-]*)
6	(?:**treated**_VBN|**treatment**_NN|**therapy**_NN) (?:**with**_IN|**by**_IN)

	([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS)
7	(?:**receive**|.*)_VB[P]?.*NN[SP]?\s[a-z]{1,}_CC\s([a-z]{6}_JJ\s[a-z]{5}_NN|[a-z]{7,12}_NN)
8	(?:**required**_VBD|**with**_IN|[\w\-]*_CC) ([a-z]{7,12}_JJ\s[a-z]{7,12}_NN)
9	(?:**required**_*|[\w]*_CC) ([a-z]{4}_NN\s[a-z]{6}_NN\s[a-z]{15}_NN)

Tabela 4.3 - Verbo ou palavra representativa com POS – Conjunto Enxuto.

Número	Expressão Regular
1	([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS) (?:**treatment**_NN|**therapy**_NN|**dose**_NN|**doses**_NNS)
2	(?:**were**_VBD|**was**_VBD|**while**_IN|**went**_VBD|**while**_NN|**doses**_NNS|**dosing**_VBG|**treated**_VBN|**treatment**_NN|**therapy**_NN) (?:[\w\-]*_IN) ([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS)

4.4.1.2 Estratégia 2

Os padrões POS criados e utilizados nas regras da estratégia 2 podem ser vistos na Tabela 4.4.

Considere o padrão 7 a título de exemplo: a expressão regular casará com este padrão se o termo for um adjetivo (JJ), seguido de um substantivo (NN), terminado de outro substantivo (NN).

Tabela 4.4 - Padrão POS criados para a estratégia 2.

Número	Padrão
1	(NNS) (IN|VBD|VBN) (NN|NNP|NNS) (NN)?
2	(NN|NNS)? (IN) (VBG)? (NN|NNP|NNS) (NN)
3	(IN) (JJ) (NN) (~JJ)
4	(VBD|VBN) (IN)? (NN|NNS|NNP)
5	(NN) (IN) (NN|NNP|NNS)
6	(NN) (NN) (NN)
7	(JJ) (NN) (NN)

O conjunto de regras da Estratégia 2 (i.e. Somente POS) contém 7 regras mais específicas, e foram desenvolvidas a partir do padrão POS apresentados na Tabela 4.4, no entanto, para melhor entendimento, os padrões da Tabela 4.4 são

mostrados numa forma simplificada. O conjunto de 7 regras pode ser observado na Tabela 4.5 a seguir:

Tabela 4.5 - Estratégia 2 – Somente POS.

Número	Expressão Regular
1	(?:[\w\-]*_NNS) (?:[\w\-]*_IN|[\w\-]*_VBD|[\w\-]*_VBN)* ([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS) (?:[\w\-]*_NN)?
2	(?:[\w\-/\\]*(?:_NN))?(?:[\w\-/\\]*(?:_NNS))?(?:[\w\-]*_IN)(?:\s)?(?:[\w\-/\\]*(?:_VBG))?([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS) (?:[\w\-]*_NN)
3	(?:[\w\-]*_IN) ([a-z]{7,12}_JJ\s[a-z]{7,12}_NN) (?:[\w\-]*_[^J]{1,})
4	(?:[\w\-]*_VBD|[\w\-]*_VBN) (?:[\w\-]*_IN)?([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS)
5	(?:[\w\-]*_NN) (?:[\w\-]*_IN) ([a-z]{7,12}_NN|[a-zA-Z]{2,3}_NNP|[a-z]{7,12}_NNS)
6	([a-z]{4}_NN\s[a-z]{6}_NN\s[a-z]{15}_NN)
7	([a-z]{7,12}_JJ\s[a-z]{7,12}_NN) (?:[\w\-]*_NN)

A validação das duas estratégias de regras criadas foram avaliadas por meio da aplicação dos conjuntos de regras no subconjunto de 123 sentenças citadas anteriormente. Neste trabalho, foi desenvolvido um *script* na Linguagem Perl para automatizar a aplicação das regras nas sentenças. Medidas de precisão, revocação e medida-F foram utilizadas para avaliar o desempenho das regras. Foram realizados 5 avaliações no conjunto de regras: a) Conjunto enxuto de 2 regras (2 regras com verbo+POS); b) Conjunto amplo de 9 regras (9 regras com verbo+POS); c) Conjunto Somente POS (7 regras Somente POS); d) Conjunto enxuto + Somente POS (2 regras com verbo+POS + 7 regras somente POS); e) Conjunto amplo + Somente POS (9 regras com verbo+POS + 7 regras somente POS). Um exemplo de aplicação de uma regra em uma sentença do subconjunto é ilustrado na Figura 4.7.

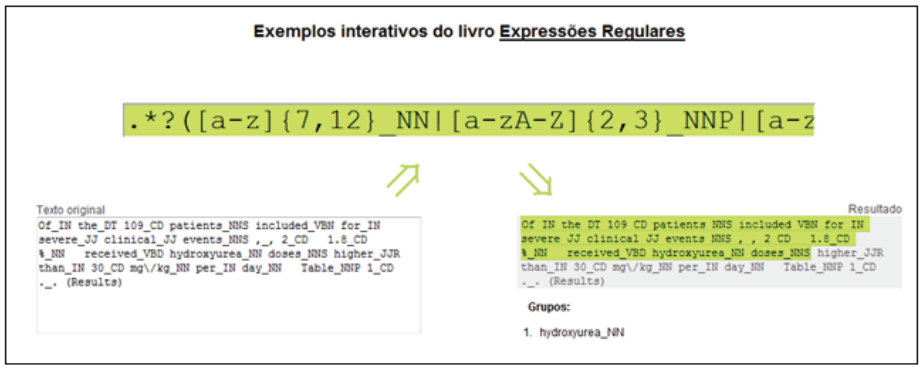

Figura 4.7 - Exemplo de aplicação de uma regra em uma sentença.
Fonte: Adaptado de Matos (2010).

Na Tabela 4.6 é apresentado o resultado da avaliação do conjunto de regras nas 5 formas de avaliação.

Tabela 4.6 - Resultado da avaliação do conjunto de regras.

Regras	Precisão	Revocação	Medida-F
2 regras com verbo+POS	97%	59%	73%
9 regras com verbo+POS	84%	57%	68%
7 regras Somente POS	60%	65%	62%
2 regras com verbo+POS e 7 regras somente POS	64%	78%	70%
9 regras com verbo+POS e 7 regras somente POS	59%	75%	66%

Estes resultados foram obtidos para a extração de novos tratamentos desconhecidos, e para os tratamentos conhecidos, o processo proposto utiliza-se de um dicionário estendido com variações e siglas visando melhorar a eficiência da extração de tratamentos nas sentenças.

Os valores apresentados na Tabela 4.6 são resultado da aplicação dos conjuntos de regras nas mesmas 123 sentenças em que foram criados os padrões de regras. Os mesmos conjuntos de regras serão utilizados em novas sentenças (outros artigos) a serem realizados em todos os experimentos desta dissertação.

O Algoritmo 1 descreve o pseudocódigo da extração de termos com o uso de regras. Para cada uma das sentenças classificadas e para o conjunto de regras (linhas 1 e 2), o algoritmo faz o seguinte: É inicializado o vetor MatrizResultado que irá armazenar a posição da sentença classificada e a regra do conjunto de regras (linha 3). Posteriormente, são selecionados o conjunto de regras e as sentenças classificadas (linhas 4 e 5). Em seguida, é aplicado o conjunto de regras nas

sentenças. Se a regra casar com sentença, então extrai o termo (linhas 6 a 12) e atualiza o vetor MatrizResultado com o valor 1 na posição sentença-regra (linha 10).

Algoritmo 1 - Extração de termos com o uso de regras.

```
1   for (i = 0; i < quantidade das sentenças;i++)
2         for (j = 0; j<= quantidade de regras; j ++)
3               MatrizResultado[i][j] = 0;
4   regra[] <- getRegras(); //termo representativo+POS e somente POS
5   sentença[] <- getSentençasClassificadas();
6   for (i = 0; i <= quantidade das sentenças; i++)
7     |     for (j = 0; j<= quantidade de regras; j ++)
8     |       |     if (regra[j] casou com sentença[i])
9     |       |           print ("Termo extraído");
10    |       |           MatrizResultado[i][j] = 1;
11    |     end
12  end
```

O Algoritmo 2 descreve o pseudocódigo da extração de termos com o uso de regras e dicionário. Para cada uma das sentenças classificadas, termos do dicionário e conjunto de regras (linhas 1 e 2; linhas 4 e 5), o algoritmo faz o seguinte: É inicializado o vetor MatrizDicionário que irá armazenar a posição da sentença classificada e o termo do dicionário (linha 3). É também inicializado o vetor MatrizRegras que irá armazenar a posição da sentença classificada e a regra do conjunto de regras (linha 6).

Posteriormente, são selecionados os termos do dicionário, o conjunto de regras e as sentenças classificadas (linhas 7 a 9). Em seguida, é realizada uma verificação em cada sentença a fim de encontrar um termo validado pertencente ao dicionário (linhas 10 a 16). Se existe termo validado na sentença, então atualiza o vetor MatrizDicionário com o valor 1 na posição sentença-dicionário (linha 15).

Por fim, é aplicado o conjunto de regras nas sentenças. Se a regra casar com sentença e o termo extraído não existir no dicionário, então insere-se no banco de dados o termo novo como "não validado" e atualiza-se o vetor MatrizRegras com o valor 1 na posição sentença-regra (linha 24); caso contrário, informa que o termo já existe no dicionário (linhas 17 a 30).

Algoritmo 2 - Extração de termos com o uso de regras e dicionário.

```
1   for (i = 0; i < quantidade das sentenças;i++)
2         for (j = 0; j<= quantidade de termos no dicionário; j ++)
3               MatrizDicionario[i][j] = 0;
4   for (i = 0; i < quantidade das sentenças;i++)
5         for (j = 0; j<= quantidade de regras; j ++)
6               MatrizRegras[i][j] = 0;
7   dicionário[] <- getTermos();
8   regra[] <- getRegras(); //termo representativo+POS e somente POS
9   sentença[] <- getSentençasClassificadas();
10  for (i = 0; i <= quantidade das sentenças; i++)
11        for (k = 0; k<= quantidade de termos no dicionário; k ++)
12              if (existe termo validado em dicionário[k] em
13              sentença[i])
14                    print ("Termo existente na sentença");
15              MatrizDicionario[i][j] = 1;
16        end
17        for (j = 0; j<= quantidade de regras; j ++)
18              if (regra[j] casou com sentença[i])
19                    Termo <- termoIdentificado(regra[j]);
20                    if (não existe termo em dicionário)
21
22                          insere (termo, dicionário, "Não Validado");
23                                print ("Termo novo extraído");
24                                MatrizRegras[i][j] = 1;
25                    end
26                    else print ("Termo extraído já existe no
27                    dicionário");
28              end
29        end
30  end
```

4.4.2 Abordagem de extração de informação baseada em dicionário

A abordagem de extração de informação baseada em dicionário tem o propósito de reconhecer em quais sentenças os termos validados ocorrem e, portanto, completar (i.e., preencher) o tipo-relacionamento termo/artigo do banco de dados. Os termos validados são aqueles que foram consolidados manualmente por um especialista. O dicionário não tem a função de identificar novos termos.

O dicionário possui duas características: a primeira é possuir a função de armazenar os termos em que são extraídos dos artigos como resultado do processo de extração de informação e a segunda é identificar sentenças que possuem termos comparando-se com os termos validados já existentes no dicionário.

O carregamento dos dados no dicionário pode ser efetuado de forma manual ou automática. No carregamento manual, os termos são inseridos com a

participação do especialista do domínio. No carregamento automático, os termos são identificados nos artigos científicos por meio da abordagem de regras. O dicionário de tratamentos foi construído por meio de uma carga inicial dos dados, com o auxílio de especialista do domínio. O papel do especialista do domínio nesta etapa foi muito importante para a inserção dos dados e o contato próximo com a especialista facilitou a construção do dicionário. O dicionário biomédico contém 26 termos de tratamento válidos e consolidados sobre a doença da Anemia Falciforme.

Além do dicionário servir como um recurso terminológico, ele também é utilizado como técnica alternativa para reduzir problemas de restrição de nomes com o uso de variações de termos, sinônimos e siglas, de forma a reduzir os problemas da técnica de dicionário e com isto melhorar a precisão da extração de termos conhecidos de tratamentos.

Um termo pode ser descrito de várias formas, isto é, pode ter variações, e para tratar o problema de variações de nomes, neste projeto, utiliza-se de uma entidade fraca denominada "*Treatment Variation*" para armazenar as variações dos nomes de cada termo.

Na Figura 4.8 é apresentado o esquema conceitual entre as entidades "*Treatment*" e "*Treatment Variation*".

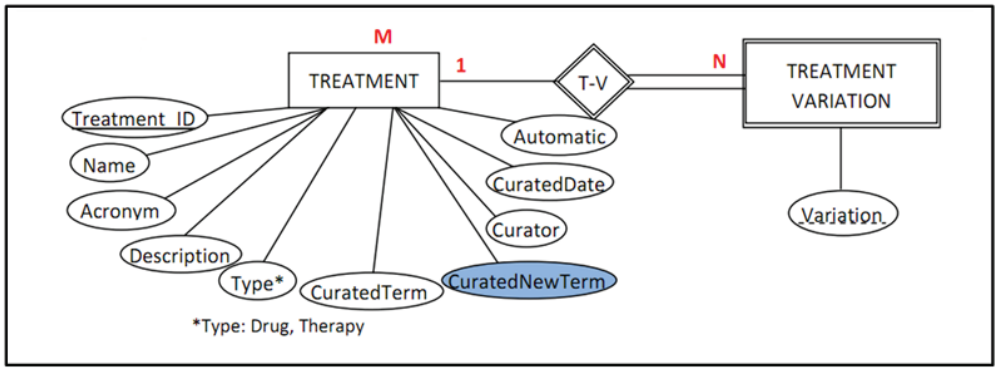

Figura 4.8 - Entidades *Treatment* e *Treatment Variation*.

O dicionário também serviu para auxiliar na qualidade da formação das regras. Foi realizada uma observação empírica sobre os termos relevantes de tratamentos consolidados armazenados no dicionário e constatou-se que a maioria dos termos existentes continha uma quantidade semelhante de caracteres, e a partir desta observação, para auxiliar no desenvolvimento das regras, foi inferida uma heurística sobre a quantidade de caracteres dos termos. Todos os termos foram contados, somados e feito à média da quantidade dos caracteres. A média resultou entre 7 e 12 caracteres para termos, e para siglas de 2 a 3 caracteres. O valor

resultante foi aproveitado e apoiado na formação das regras. Esta observação não foi validada pelo especialista do domínio, mas foi utilizada neste trabalho, pois o resultado desta observação fez com que diminuíssem os falsos positivos para termos já consolidados e ainda, obteve-se uma melhoria na precisão.

O Algoritmo 3 descreve o pseudocódigo da extração de termos com o uso do dicionário. Para cada uma das sentenças classificadas e termos do dicionário (linhas 1 e 2), o algoritmo faz o seguinte: É inicializado o vetor MatrizResultado que irá armazenar a posição da sentença classificada e o termo do dicionário (linha 3). Posteriormente, são selecionados os termos do dicionário e as sentenças classificadas (linhas 4 e 5). Em seguida, é realizada uma verificação em cada sentença a fim de encontrar um termo pertencente ao dicionário. Se existe o termo na sentença, então atualiza o vetor MatrizResultado com o valor 1 na posição sentença-termo e imprime o termo (linhas 6 a 12).

Algoritmo 3 - Extração de termos com o uso do dicionário.

```
1   for (i = 0; i < quantidade das sentenças;i++)
2         for (j = 0; j<= quantidade de termos no dicionário; j ++)
3               MatrizResultado[i][j] = 0;
4   dicionário[] <- getTermos(); // termo básico+ variações
5   sentença[] <- getSentençasClassificadas();
6   for (i = 0; i <= quantidade das sentenças; i++)
7     |     for (j = 0; j<= quantidade de termos no dicionário; j ++)
8     |       |     if (existe dicionário[j] em sentença[i])
9     |       |           MatrizResultado[i][j] = 1;
10    |       |     print ("Termo existente na sentença");
11    |     end
12  end
```

4.5 Considerações Finais

Neste trabalho foram utilizados exemplos da doença Anemia Falciforme para melhor contextualizar e explicar os conceitos envolvidos no processo proposto para a extração de informação de termos de tratamento de artigos científicos do domínio biomédico.

O processo proposto para a extração de tratamentos do domínio biomédico possui um conjunto de regras específicas (conjunto amplo da estratégia 1 e conjunto POS da estratégia 2) e genéricas (conjunto enxuto da estratégia 1). As regras específicas são dependentes da doença da Anemia Falciforme e as regras genéricas podem ser adaptadas com algum esforço para extrair informações de outras doenças.

A entrada de dados e a classificação de sentenças mencionadas neste capítulo é adequada para qualquer doença específica. O dicionário de dados foi construído e carregado com termos de tratamento específicos da doença Anemia Falciforme, porém esta abordagem está apta a ser aplicada a qualquer doença do domínio biomédico.

Na próxima seção, é discutida a prova de conceito, visando analisar a eficiência do processo proposto. Será realizada no capítulo 5 a análise separada da etapa de classificação de sentenças utilizando as medidas de precisão, revocação, acurácia e medida-F e da etapa de identificação de termos relevantes utilizando as medidas de precisão, revocação e medida-F.

Capítulo 5

AVALIAÇÃO DA METODOLOGIA PROPOSTA

Neste capítulo são apresentados e discutidos os resultados dos experimentos realizados com o objetivo de avaliar a metodologia proposta de extração de termos de tratamento, a qual foi apresentada no capítulo 4.

5.1 Considerações Iniciais

Neste capítulo é apresentada a prova de conceito correspondente a duas etapas da metodologia proposta, visando verificar se as hipóteses consideradas nesta pesquisa são válidas: (i) fase de classificação de sentenças por agrupamento de parágrafos, e (ii) fase de identificação de termos relevantes (i.e. fase específica para a extração de informação). O objetivo da prova de conceito é validar o processo proposto aplicado no domínio biomédico, em particular, aplicado em artigos científicos da doença da Anemia Falciforme.

As fases de classificação de sentenças e identificação de termos relevantes são etapas que contribuem para a extração de informação, que é a finalidade do processo completo proposto nesta dissertação. Estas fases foram avaliadas separadamente para permitirem atingir um resultado parcial do processo, alcançando o objetivo comum que é o resultado final desejado da extração de informação.

Inicialmente foi realizada uma prova de conceito na fase de classificação de sentenças, cujo objetivo foi utilizar um algoritmo de aprendizado de máquina supervisionado para classificar as sentenças sobre a doença da Anemia Falciforme em suas respectivas classes. Para o processo proposto foram utilizados dois classificadores ambos com o mesmo indutor: Classificador C1 para sentenças de complicação e Classificador C2 para sentenças de tratamento, ambos com o objetivo de separar as sentenças de interesse que provavelmente tinham, respectivamente, termos de complicação e termos de tratamento, das sentenças que possivelmente não tinham nenhum destes termos. O motivo de construir dois classificadores foi para comprovar a hipótese deste trabalho, em que na maioria dos casos os termos de tratamento ocorrem em uma mesma sentença que possui uma complicação ou em sentenças próximas em um mesmo parágrafo.

Portanto, a classificação foi usada no processo proposto como um filtro que selecionou apenas as sentenças de interesse, diminuindo o custo de análise na fase de identificação de termos relevantes das sentenças de um artigo.

As classes foram divididas em “complicação” e “outros” para o Classificador de Complicação (C1) e as classes de “tratamento” e “outros” para o Classificador de Tratamento (C2). Nesta fase, dois experimentos foram realizados: 1) fase de treinamento e fase de testes, para a qual os modelos de classificação foram criados e testados com sentenças que não foram treinadas; e 2) fase de uso dos modelos que teve como objetivo avaliar os classificadores com novas sentenças. As medidas de precisão, revocação, acurácia e medida-F foram utilizadas para avaliar a classificação.

A segunda prova de conceito realizada foi na fase de identificação de termos, cujo objetivo foi extrair os termos relevantes das sentenças classificadas na etapa anterior. As medidas de precisão, revocação e medida-F foram utilizadas para avaliar o percentual de termos extraídos. As abordagens baseadas em dicionário e regras foram utilizadas para identificar os termos relevantes nas sentenças relacionadas à classe de tratamentos.

Na Tabela 5.1 são sumarizados os objetivos de cada experimento.

Tabela 5.1 - Objetivos dos experimentos.

Experimento	Objetivo	Fase	Resultado esperado	Validação
Experimento1	Realizar o treinamento e o teste para construir os classificadores C1 e C2	Classificação de Sentenças (treinamento e teste)	Construir os classificadores a partir do melhor algoritmo de aprendizado de máquina	Medidas: Acurácia, Precisão, Revocação, Medida-F com Classificador SCA (API do Weka)
Experimento2	Avaliar os classificadores C1 e C2 na classificação de novas sentenças	Classificação de Sentenças (uso do modelo de classificação)	Obter valores satisfatórios na classificação de novas sentenças	Medidas: Acurácia, Precisão, Revocação, Medida-F
Experimento3	Identificar e extrair os termos relevantes de tratamentos	Identificação de Termos Relevantes	Extrair termos conhecidos utilizando o dicionário e identificar novos termos a partir de regras	Medidas: Precisão, Revocação, Medida-F

Nas próximas seções são apresentados os resultados dos experimentos na fase de classificação de sentenças e na fase de identificação de termos relevantes.

5.2 Classificação de Sentenças

Inicialmente foram selecionados 8 artigos científicos sobre a doença Anemia Falciforme nos quais foram identificadas 765 sentenças de todas as seções dos artigos. A escolha da amostra de artigos foi de forma estática e definida pelo especialista da área. Os artigos podem ser encontrados no apêndice digital desta dissertação e também se encontram disponíveis para download vinculado na Tabela 5.4.

Foi realizada a classificação manual neste conjunto de 765 sentenças existente no corpus e essas sentenças foram examinadas de duas formas: 1) 765 sentenças analisadas e compreendidas como sendo sentenças de “complicação” e “outros” e; 2) as mesmas 765 sentenças analisadas e mencionadas como sendo sentenças de “tratamento” e “outros”. Para a primeira representação, este cenário foi utilizado para preparar o desenvolvimento do Classificador C1 (Complicação) e a segunda para o Classificador C2 (Tratamento). Todas as sentenças foram utilizadas para o treinamento e o teste dos classificadores usando o método de particionamento *10-fold cross validation*. A tarefa de anotação manual foi realizada pelo especialista e demandou muito tempo e esforço. A porcentagem da distribuição das classes pode ser vista na Tabela 5.2.

Tabela 5.2 - Porcentagem da distribuição das classes.

Classificador	Quantidade	Sentenças	Porcentual
C1	337	Complicação	44%
	428	Outros	56%
C2	394	Tratamento	51%
	371	Outros	49%

Nas próximas seções são discutidos os resultados da classificação nas fases de treinamento e teste e da classificação na fase de uso do modelo de classificação.

5.2.1 Experimento 1: Fase de treinamento e de teste

O experimento é descrito como segue: primeiramente foi realizada uma limpeza no conjunto de treinamento, removendo das sentenças pontuação, parênteses, colchetes e apóstrofes. Após, a matriz de atributo-valor foi construída usando frequência mínima igual a 2 para a seleção de atributo. A seleção de atributo foi composta de 1 a 3 gramas. Não foram usadas a técnica de *stemmer* ou eliminação de *stopwords*. O uso de *stopwords* deve-se ao fato de que alguns termos ajudam na formação de regras para a extração dos termos.

Dois algoritmos de aprendizado de máquina foram utilizados neste experimento: *Support Vector Machines* (SVM) e *Naïve Bayes*. Também foram utilizadas quatro configurações de pré-processamento, as quais geraram, portanto 8 configurações. Os filtros utilizados foram: 1) *No Filter*; 2) *Randomize*; 3) *Remove Misclassified,* para remover ruído; e 4) *Resample*, método utilizado para balancear as classes de interesse.

Os testes foram executados usando o classificador SCA-Classifier do ambiente do Projeto SCA (MATOS, 2010) desenvolvido na linguagem de programação Java com o uso da API do ambiente Weka (WITTEN; FRANK, 2005). O algoritmo SVM apresentou o melhor resultado para os classificadores C1 e C2, conforme observado na Figura 5.1 (a) e (b), respectivamente. Assim, o Classificador C1 (Figura 5.1 (a)) e o Classificador C2 (Figura 5.1 (b)) foram construídos a partir do melhor resultado apresentado. Note que o modelo criado para os classificadores C1 e C2 usou os filtros *Remove Misclassified* e *Resample.*

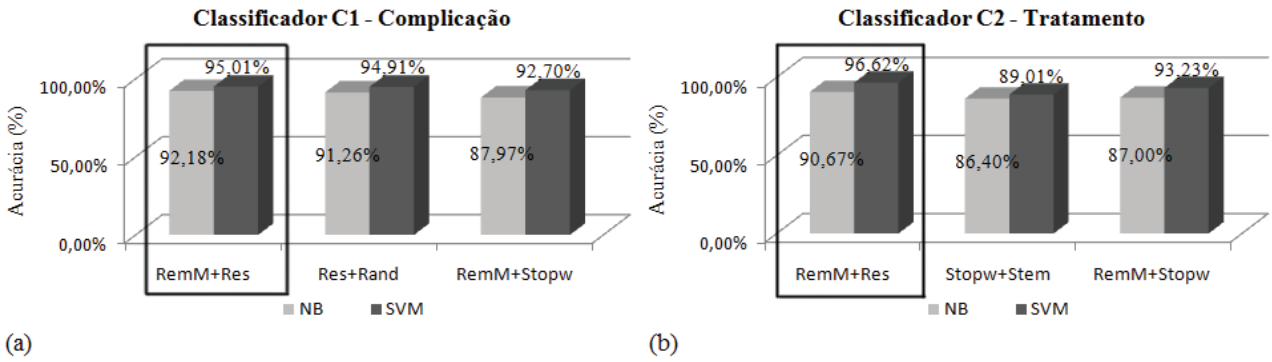

Figura 5.1 - Melhores resultados para os classificadores C1 (a) e C2 (b)

Na próxima seção é descrita a fase de uso dos modelos de classificação criados com as configurações conforme apresentado na Figura 5.1.

5.2.2 Experimento 2: Fase de uso do modelo de classificação

Este experimento tem o objetivo de avaliar qual a maior acurácia obtida dentre o algoritmo SVM na classificação de novas sentenças. Na Tabela 5.3 é apresentado o resultado da classificação automática das 359 novas sentenças usadas para o Classificador C1 e Classificador C2, ambos construídos a partir do mesmo indutor. Foram utilizadas as principais métricas usadas em sistemas de extração de informação como precisão, revocação e medida-F, e em sistemas de aprendizado de máquina como acurácia.

Tabela 5.3 - Resultado da classificação automática.

Classificador	Qtde*	Sentenças	Precisão	Revocação	Acurácia	Medida-F
C1	220	Complicação	85%	64%	79%	73%
C2	107	Tratamento	88%	51%	71%	64,5%

*Qtde = Quantidade

Os resultados mostram que as classificações de sentenças com complicação e com tratamento obtiveram uma boa precisão compatível com valores de precisão obtidos em outros trabalhos que realizam extração de informação de artigos completos. Apesar da revocação não ter sido alta, a repetição de termos de tratamento nas sentenças faz com que a perda de algumas sentenças não tenha tanto impacto no processo de extração de informação, conforme mostrado nos próximos experimentos.

5.3 Identificação de Termos Relevantes

O objetivo desta prova de conceito é avaliar a identificação de termos relevantes (i.e. extração da informação dos termos de interesse de tratamento) nas sentenças que foram classificadas como sendo de tratamento.

Nesta seção são realizados três tipos de experimentos, que são: (i) “classificação automática de sentenças de tratamento com complicação e extração com regras”, que é o experimento principal deste trabalho, que executa todos os passos da metodologia para possível comprovação da hipótese; (ii) “classificação automática de sentenças de tratamento sem complicação e extração com regras”, ou seja, o experimento que executa a metodologia parcialmente, desconsiderando o classificador de complicação; e (iii) “classificação manual de sentenças de tratamento com complicação e extração com regras”, ou seja, o experimento que executa todos os passos da metodologia variando a classificação automática por classificação manual. Os três tipos de experimentos cobrem todo o escopo e indica adaptabilidade da metodologia proposta.

A seguir são apresentados todos os experimentos realizados neste trabalho e seus respectivos resultados.

5.3.1 Classificação automática de sentenças de tratamento com complicação e extração com regras

Este estudo tem como objetivo avaliar a quantidade de termos relevantes de tratamentos nas sentenças fornecidas na etapa da entrada de dados. A hipótese desta dissertação de mestrado é que as sentenças que possuem termos de tratamento ocorrem em uma sentença que possui um termo de complicação ou ocorre em sentenças próximas em um mesmo parágrafo. Portanto, para validar esta hipótese, este experimento tem o propósito de validar se os termos de tratamento estão presentes em um mesmo parágrafo que contenha sentenças de complicação.

Essa hipótese foi baseada em uma análise empírica de artigos da doença Anemia Falciforme, conforme levantamento realizado em um conjunto de 10 artigos. A Tabela 5.4 ilustra o resultado deste levantamento, em que mostra que os termos de tratamento estão em sua maioria nos parágrafos de complicação (coluna Parágrafos de Complicação e Tratamento).

Tabela 5.4 - Análise empírica de parágrafos de artigos da doença da AF.

Artigo	Parágrafos de Complicação	Parágrafos de Tratamento	Outros Parágrafos	Parágrafos de Complicação e Tratamento	Parágrafos de Complicação e Tratamento (%)	Total
Artigo2	0	18	2	21	51%	41
Artigo3	4	18	2	13	35%	37
Artigo5	1	7	13	13	38%	34
Artigo6	1	1	1	25	89%	28
Artigo7	10	2	4	6	27%	22
Artigo8	0	5	1	11	64%	17
Artigo10	1	1	1	7	70%	10
Artigo14	1	1	2	14	77%	18
Artigo1	3	9	1	17	56%	30
Artigo11	18	7	4	17	36%	46

Os artigos 2 a 14 são utilizados para treinamento e teste na fase de classificação, e os artigos 1 e 11 são utilizados como entrada de novos dados para validar a metodologia proposta.

Na próxima seção é realizada a extração manual em todo artigo com a intenção de comprovar a hipótese que foi considerada neste trabalho.

5.3.1.1 Extração manual em todo artigo

O objetivo deste experimento é reforçar a hipótese deste trabalho que na maioria dos casos os termos de tratamento ocorrem na mesma sentença de uma complicação ou em sentenças próximas de uma complicação em um mesmo parágrafo. Para isso, foi feito uma análise nos artigos de entrada e realizado uma extração manual na entrada de dados, considerando dois novos artigos (artigo 1 e artigo 11), totalizando 359 sentenças.

Inicialmente foram assinalados todos os termos de tratamento presentes nas 359 sentenças, independente dos termos serem derivados de complicação, e foram encontrados 156 termos relevantes, em sua maioria, duplicados. A Tabela 5.5 apresenta a quantidade detalhada destes termos.

Tabela 5.5 - Quantidade de termos detalhados.

Termos de Tratamento	Ocorrências
Transfusion	48
Nitric Oxide	2
Mechanical Ventilation	9
Placebo	2
Bone Marrow Transplantation	2
Antibiotic	6
HU	77
Hydroxyurea	10
Quantidade de Termos	156

Após efetuar a extração manual dos termos em todas as sentenças, foi realizada a classificação manual dos parágrafos. A classificação foi dividida nas classes “parágrafos de complicação” e “outros parágrafos”. É considerado como sendo um parágrafo de complicação o parágrafo que contém pelo menos uma sentença de complicação.

Após a classificação dos parágrafos, foi anotada manualmente a quantidade de termos relevantes de tratamento que as classes de parágrafos apresentavam. A anotação manual dos textos foi uma tarefa muito custosa e exigiu esforço e tempo do especialista. A Tabela 5.6 apresenta o resultado da classificação dos parágrafos e a distribuição dos termos.

A quantidade total de termos relevantes distintos presentes no documento, ou seja, nas 359 sentenças é de 8 termos de tratamento, conforme apresentado na Tabela 5.5 (coluna Termos de Tratamento) e como pode ser observado, todos os 8 termos distintos estão presentes nos parágrafos de complicação com pelo menos uma ocorrência, conforme apresentado na Tabela 5.6 (coluna Parágrafos de Complicação). Sendo assim, apenas selecionando os parágrafos de complicação é possível obter 100% de revocação para os termos distintos, e consequentemente, é comprovada a hipótese desta dissertação, em que os termos de tratamento estão em uma sentença que possui um termo de complicação ou ocorre em sentenças próximas em um mesmo parágrafo. Ressalta-se que o conjunto de termos de tratamentos para a doença da Anemia Falciforme é pequeno e limitado, e para esta

prova de conceito são utilizados artigos científicos nos quais estão presentes uma pequena quantidade (~30%) do total de termos existentes de tratamentos para esta doença.

Tabela 5.6 - Classificação manual de parágrafos.

Termos de Tratamento	Total de termos no Artigo	Parágrafos de Complicação	Outros Parágrafos
Transfusion	48	32	16
Nitric Oxide	2	1	1
Mechanical Ventilation	9	4	5
Placebo	2	2	0
Bone Marrow Transplantation	2	1	1
Antibiotic	6	3	3
HU	77	55	22
Hydroxyurea	10	6	4
Quantidade de Termos	156 termos	104 termos	52 termos
Porcentagem	100% termos	67%	33%
Quantidade de Sentenças	359 sentenças	275 sentenças	84 sentenças
Quantidade de Parágrafos	76 parágrafos	55 parágrafos	21 parágrafos

Nas próximas seções são apresentados os experimentos com extração manual e automática.

5.3.1.2 Extração manual

Este experimento tem como objetivo realizar a classificação automática de sentenças de tratamento, agregado com a classificação automática de complicação e extração manual de termos de tratamento.

Primeiramente o conjunto de 359 novas sentenças contendo 156 termos relevantes de tratamento foi remetido ao classificador automático de complicação (C1), resultando como saída 120 sentenças classificadas como complicação. Estas 120 sentenças foram processadas pelo agrupador de parágrafos, estendendo-se em um total de 283 sentenças. O agrupador de parágrafos tem o intuito de agrupar as sentenças que pertencem a um mesmo parágrafo, desde que a sentença principal seja de complicação. Neste agrupamento, a sentença de complicação que pertencer a um determinado parágrafo, automaticamente, eleva todas as sentenças deste

mesmo parágrafo independente se a sentença se referir à classe “outros”, e assim, todas as sentenças são dispostas a um novo conjunto estendido.

Estas 283 sentenças foram enviadas ao classificador automático de tratamento (C2), classificando-se em 74 sentenças como sendo sentenças de tratamento. Estas 74 sentenças continham 59 termos relevantes de tratamentos. O subconjunto resultante de 74 sentenças foi processado e etiquetado com POS conforme sua classe gramatical. A partir das sentenças etiquetadas foi realizada a extração manual de termos. Com a extração manual de termos foi possível extrair 59 termos relevantes de tratamento, conforme pode ser observado na Tabela 5.7.

Tabela 5.7 – Termos extraídos manualmente a partir da classificação.

Termos	Qtde
Transfusion	24
Antibiotic	1
HU	29
Hydroxyurea	5
Total	59

Como se pode observar, houve perda de algumas sentenças no processo de classificação e consequentemente a perda de alguns termos relevantes, pois conforme citado anteriormente, o conjunto inicial continha 156 termos de tratamento e após a classificação automática de complicação e tratamento, restaram apenas 59 destes termos, e logo, estes termos foram extraídos manualmente. Na Tabela 5.8 é apresentado o resultado da extração manual.

Tabela 5.8 - Extração manual a partir da classificação automática.

Termos – Todos			Termos Distintos		
Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
100%	37,82%	54,88%	100%	50%	66,66%

A perda de algumas sentenças no processo de classificação automática não influencia na qualidade da extração, ou seja, na identificação dos termos relevantes, conforme pode ser observado no resultado da extração manual e automática. A prova de conceito na fase de classificação de sentenças foi validada com um conjunto de 8 artigos e o treinamento dos classificadores foram realizados com uma

quantidade pequena de sentenças. Com este cenário, perder algumas sentenças, ou seja, algumas sentenças não serem classificadas corretamente, e ainda, considerando-se que os termos de tratamento se repetem ao longo do artigo, esta perda se torna aceitável.

O resultado da Tabela 5.7 mostra-se a quantidade exata de termos relevantes de tratamento existentes no subconjunto de 74 sentenças a partir de uma classificação automática. Neste trabalho e em todo capítulo, extração manual é considerada a extração realizada manualmente por humanos. O intuito do próximo experimento é realizar a extração de termos automaticamente, ou seja, realizar a extração de termos a partir dos conjuntos de regras desenvolvidas e mencionadas no Capítulo 4, e ainda, avaliar o resultado dos mesmos. Para validação, foram utilizadas as medidas de precisão, revocação e medida-F.

5.3.1.3 Extração automática

Este experimento tem como objetivo realizar a extração automática de termos a partir da aplicação das regras no subconjunto de 74 sentenças classificadas como tratamento (contendo 59 termos relevantes) explicado no experimento anterior.

Neste estudo, consideram-se igualmente todos os passos descritos na seção 5.3.1.2, com exceção da extração manual, ou seja, as mesmas 359 sentenças foram encaminhadas ao classificador automático de complicação (C1), o subconjunto resultante de 120 sentenças foi processado e as sentenças foram agrupadas por parágrafos. Adiante o subconjunto estendido contendo 283 sentenças foi enviado ao classificador automático de tratamento (C2) e o subconjunto procedente de 74 sentenças foram etiquetas com POS para posteriormente serem enviadas para o processo de extração.

Após rotular as sentenças, foram aplicadas as regras do conjunto amplo, enxuto (verbo ou palavra representativa+POS) e Somente POS descritas no Capítulo 4, e ainda, para obter uma visão mais ampla dos resultados, a extração automática foi dividida em 2 etapas: 1) Aplicação das regras para obter o resultado da extração de todos os termos, ou seja, termos distintos e termos repetidos; e 2) Aplicação das regras para obter o resultado da extração apenas nos termos distintos.

Na Tabela 5.9 é apresentado o resultado da extração automática aplicada nas 74 sentenças de tratamento para todos os termos (distintos e repetidos - coluna Termos-Todos) e somente para os termos distintos (coluna Termos Distintos).

Tabela 5.9 - Extração automática a partir da classificação automática.

Regra	Termos – Todos			Termos Distintos		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
2 regras com verbo+POS	100%	27%	42%	100%	100%	100%
9 regras com verbo+POS	62%	25%	35%	33%	100%	50%
7 Somente POS	41%	49%	44%	13%	100%	23%

Como se pode observar, todos os conjuntos de regras obtiveram 100% de revocação para termos distintos, e somente o conjunto de 2 regras com verbo ou palavra representativa+POS obteve 100% de precisão para todos os termos.

A aplicação dos conjuntos de regras nas sentenças citadas neste experimento pode ser vistos no apêndice digital.

5.3.1.4 Considerações Finais

Nesta seção foi descrita a avaliação realizada na metodologia proposta neste projeto de mestrado. O objetivo do experimento foi de (i) reforçar a hipótese considerada neste trabalho conforme discutida na seção 4.1 e, por conseguinte (ii) executar todos os passos da metodologia conforme ilustrado na Figura 4.2.

Primeiramente foi realizada uma análise empírica em um conjunto de 10 artigos da doença Anemia Falciforme com o intuito de conferir a hipótese no qual se determina que: sentenças que possuem termos de tratamento ocorrem em uma sentença que possui um termo de complicação ou ocorre em sentenças próximas em um mesmo parágrafo. A prova de conceito foi realizada com um conjunto pequeno de artigos, apesar do ideal ser usar um conjunto maior.

O resultado desta análise mostra que termos de tratamento em sua maioria ocorrem nos parágrafos de complicação de acordo com os valores apresentados na

Tabela 5.4; e todos os termos relevantes distintos de tratamento estão presentes nos parágrafos de complicação (Tabela 5.6). Portanto obtêm-se 100% de revocação para os termos distintos, fazendo com que a hipótese seja reconhecida.

Ademais, para avaliar o processo proposto que contemplam todos os passos da metodologia, foi necessário dividi-lo em duas fases: (i) as sentenças foram classificadas em sentenças com termos de complicação e em sentenças que não possuíam termos de complicação (classificador de complicação C1); as sentenças de interesse foram agrupadas adicionando-se as demais sentenças que participavam do mesmo parágrafo de uma sentença previamente selecionada (agrupamento de parágrafos); as sentenças foram classificadas em sentenças com termos de tratamento e em sentenças que não possuíam tratamento (classificador de tratamento C2); as sentenças com termos de tratamento foram etiquetadas conforme a sua classe gramatical; e realizado a extração manual de termos de tratamento; (ii) classificação automática das sentenças de complicação (C1); agrupamento de sentenças por parágrafos; classificação automática das sentenças em tratamento (C2); etiquetagem das sentenças com POS; e aplicação do conjunto de regras para efetuar a extração automática de termos de tratamento.

Na extração manual foram encontrados 59 termos relevantes de tratamento a partir da classificação automática de complicação e tratamento. O resultado mostra-se que com a classificação automática eliminou algumas sentenças e alguns termos de tratamento. Com isso, houve uma perda de 50% para termos distintos (termos não repetidos), resultando numa revocação de 50% para termos distintos. Dos 8 termos distintos presentes no conjunto inicial, 4 termos foram perdidos.

A fase de classificação de sentenças foi validada com um conjunto de 8 artigos e o treinamento dos classificadores foram realizados com uma quantidade pequena de sentenças. A classificação pode ser melhorada utilizando um conjunto maior de dados para treinamento e/ou utilizando outros classificadores. Os termos distintos perdidos, ou seja, que não foram identificados na classificação, não prejudicaram o processo da extração, que é o objetivo desta prova de conceito, ademais, conforme citado anteriormente, a validação das etapas da metodologia são avaliadas separadamente.

Na avaliação de identificação de termos relevantes, a extração automática obteve 100% de precisão e 27% de revocação para todos os termos utilizando o conjunto de 2 regras com verbo ou palavra representativa+POS. Considera-se que

os termos de tratamento se repetem do longo do artigo, portanto, obter uma revocação baixa não impacta o processo de extração de informação. Ainda, o mesmo conjunto de regras atingiu 100% de revocação para os termos distintos, consequentemente, extraíram-se todos os termos não repetidos existentes no artigo comparado com a extração manual.

A título de comparação, a seguir são realizados dois experimentos da metodologia, com passos parciais variados, que são: (i) experimento que executa parcialmente a metodologia, sem considerar o classificador de complicação; (ii) experimento que executa a metodologia por completa, porém utiliza a classificação manual ao invés da classificação automática.

5.3.2 Classificação automática de sentenças de tratamento sem complicação e extração com regras

5.3.2.1 Extração manual

Este experimento tem como objetivo realizar a classificação automática de sentenças de tratamento e extração manual de termos de tratamento, sem considerar a classificação automática de complicação e, portanto sem o agrupamento de parágrafos.

Inicialmente o conjunto de 359 novas sentenças contendo 156 termos relevantes de tratamento foi enviado ao classificador automático de tratamento (C2), resultando como saída 107 sentenças classificadas como sendo sentenças de tratamento. Estas 107 sentenças foram processadas, etiquetadas com POS e realizado a extração manual dos termos relevantes de tratamento. O resultado da extração manual obteve um total de 95 termos relevantes, conforme é apresentado na Tabela 5.10.

Como se pode observar, houve perda de algumas sentenças no processo de classificação e consequentemente a perda de alguns termos relevantes, pois de 156 termos de tratamento que continha no conjunto inicial, após a classificação automática de tratamento foi possível extrair apenas destes 95 termos restantes. Na Tabela 5.11 é apresentado o resultado da extração manual.

Tabela 5.10 – Termos extraídos manualmente a partir da classificação.

Termos	Qtde
Transfusion	44
Antibiotic	2
HU	35
Hydroxyurea	10
Mechanical Ventilation	2
Bone Marrow Transplantation	2
Total	95

Tabela 5.11 - Extração manual a partir da classificação automática.

Termos – Todos			Termos Distintos		
Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
100%	60,90%	75,70%	100%	75%	85,71%

A perda de algumas sentenças no processo de classificação automática não influencia na qualidade da extração, ou seja, na identificação dos termos relevantes, conforme pode ser observado no resultado da extração manual e automática. A prova de conceito na fase de classificação de sentenças foi validada com um conjunto de 8 artigos e o treinamento dos classificadores foram realizados com uma quantidade pequena de sentenças. Com este cenário, perder algumas sentenças, ou seja, algumas sentenças não serem classificadas corretamente, e ainda, considerando-se que os termos de tratamento se repetem ao longo do artigo, esta perda se torna aceitável.

O resultado da Tabela 5.10 mostra a quantidade exata de termos relevantes de tratamento existentes no subconjunto de 107 sentenças a partir de uma classificação automática. O intuito do próximo experimento é realizar a extração de termos automaticamente, ou seja, realizar a extração de termos a partir dos conjuntos de regras desenvolvidas e mencionadas no Capítulo 4, e ainda, avaliar o resultado dos mesmos. Para validação, foram utilizadas as medidas de precisão, revocação e medida-F.

5.3.2.2 Extração automática

Este experimento tem como objetivo realizar a extração automática de termos a partir da aplicação das regras no subconjunto de 107 sentenças classificadas como tratamento (contendo 95 termos relevantes) explicado no experimento anterior.

Neste estudo, considera igualmente todos os passos descritos na seção 5.3.2.1, com exceção da extração manual, ou seja, as mesmas 359 sentenças foram encaminhadas ao classificador automático de tratamento (C2) e o subconjunto resultante de 107 sentenças foram etiquetados com POS para posteriormente serem enviadas para o processo de extração.

Após rotular as sentenças, foram aplicadas as regras do conjunto amplo, enxuto (verbo ou palavra representativa+POS) e Somente POS descritas no Capítulo 4, e ainda, para obter uma visão mais ampla dos resultados, a extração automática foi dividida em 2 etapas: 1) Aplicação das regras para obter o resultado da extração de todos os termos, ou seja, termos distintos e termos repetidos; e 2) Aplicação das regras para obter o resultado da extração apenas nos termos distintos.

Na Tabela 5.12 é apresentado o resultado da extração automática aplicada nas 107 sentenças de tratamento para todos os termos (distintos e repetidos - coluna Termos-Todos) e somente para os termos distintos (coluna Termos Distintos).

Tabela 5.12 - Extração automática a partir da classificação automática.

Regra	Termos – Todos			Termos Distintos		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
2 regras com verbo+POS	91%	23%	36%	80%	66%	72%
9 regras com verbo+POS	67%	24%	35%	37%	100%	54%
7 Somente POS	42%	48%	45%	14%	100%	24%

Neste experimento, o conjunto de 2 regras com verbo ou palavra representativa+POS continua sendo o que obteve maior precisão, mas foi o único

que não obteve 100% de revocação para termos distintos. O conjunto de 9 regras com verbo ou palavra representativa+POS obteve o percentual de 24% de revocação para todos os termos, mas foi melhor para termos distintos, obtendo 100% de revocação.

A aplicação dos conjuntos de regras nas sentenças citadas neste experimento pode ser vistos no apêndice digital.

5.3.2.3 Considerações Finais

Nesta seção foi descrita a avaliação realizada na metodologia de uma forma parcial, ou seja, não considerando o classificador de complicação. O objetivo do experimento foi de classificar automaticamente as sentenças de complicação, agrupar as sentenças por parágrafo, classificar automaticamente as sentenças em tratamento, etiquetar as sentenças com POS e realizar a extração automática a partir do conjunto de regras.

Para avaliar este experimento, foi necessário dividi-lo em duas fases: (i) classificação automática das sentenças em tratamento (classificador C2), etiquetagem das sentenças com POS e extração manual de termos de tratamento realizado por humano (especialista); (ii) classificação automática das sentenças em tratamento (classificador C2), etiquetagem das sentenças conforme sua classe gramatical e extração automática de termos de tratamento.

Na extração manual foram encontrados 95 termos relevantes de tratamento a partir da classificação automática de tratamento. O resultado mostra que com a classificação automática eliminou-se algumas sentenças e alguns termos de tratamento. Com isso, houve uma perda de 25% para termos não repetidos (ou seja, termos distintos), resultando numa revocação de 75% para termos não repetidos. Dos 8 termos distintos presentes no conjunto inicial, 2 termos foram perdidos. A fase de classificação de sentenças foi validada com um conjunto de 8 artigos e o treinamento dos classificadores foram realizados com uma quantidade pequena de sentenças. A classificação pode ser melhorada utilizando um conjunto maior de dados para treinamento e/ou utilizando outros classificadores. Os termos distintos perdidos, ou seja, que não foram identificados na classificação, não prejudicou o processo da extração, que é o objetivo desta prova de conceito, ademais, conforme

citado anteriormente, a validação das etapas da metodologia são avaliadas separadamente.

Na avaliação de identificação de termos relevantes a extração automática obteve 91% de precisão para todos os termos, porém obteve revocação de 66% para termos distintos utilizando o conjunto de 2 regras. Já o conjunto de 9 regras obteve precisão de 67% para todos os termos e 100% de revocação para termos distintos.

5.3.3 Classificação manual de sentenças de tratamento com complicação e extração com regras

5.3.3.1 Extração manual

Este experimento tem como objetivo realizar a classificação manual de sentenças de complicação, posteriormente classificar manualmente as sentenças de tratamento e por fim, realizar extração manual de termos.

Para iniciar, o conjunto de 359 novas sentenças contendo 156 termos relevantes de tratamento foi classificado manualmente como sendo sentenças de complicação, gerando como saída um subconjunto de 163 sentenças. Após a classificação manual, este subconjunto foi processado pelo agrupador de parágrafos, adicionando-se as demais sentenças que participam do mesmo parágrafo, e assim, este subconjunto foi estendido em um novo subconjunto contendo 325 sentenças.

Logo, foi realizada a classificação manual no subconjunto de 325 sentenças para classificar em sentenças de tratamento. Com a classificação manual remanesceram 112 sentenças de tratamento, nas quais foram etiquetadas com POS e realizado a extração manual dos termos relevantes de tratamento. O resultado da extração manual obteve um total de 141 termos relevantes, conforme é apresentado na Tabela 5.13.

Tabela 5.13 – 141 Termos extraídos manualmente a partir da classificação.

Termos	Qtde
Transfusion	43
Antibiotic	4
HU	72
Hydroxyurea	7
Mechanical Ventilation	9
Bone Marrow Transplantation	2
Nitric Oxide	2
Placebo	2
Total	141

Como se pode observar, houve uma pequena perda de algumas sentenças no processo de classificação e, portanto a perda de alguns termos relevantes, porém os termos que foram perdidos são considerados repetidos, pois conforme pode ser visto na Tabela 5.13, mostra que esta perda não houve impacto para os termos distintos e também não prejudicou o processo de identificação de termos relevantes. Na Tabela 5.14 é apresentado o resultado da extração manual.

Tabela 5.14 - Extração manual a partir da classificação manual.

Termos – Todos			Termos Distintos		
Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
100%	90,38%	94,94%	100%	100%	100%

O resultado da Tabela 5.13 mostra a quantidade exata de termos relevantes de tratamento existentes no subconjunto de 112 sentenças a partir de uma classificação manual. O intuito do próximo experimento é realizar a extração de termos automaticamente, ou seja, realizar a extração de termos a partir dos conjuntos de regras desenvolvidas e mencionadas no Capítulo 4, e ainda, avaliar o resultado dos mesmos. Para validação, foram utilizadas as medidas de precisão, revocação e medida-F.

5.3.3.2 Extração automática

Este experimento tem como objetivo realizar a extração automática de termos a partir da aplicação das regras no subconjunto de 112 sentenças classificadas como tratamento (contendo 141 termos relevantes) explicado no experimento anterior.

Neste estudo, considera igualmente todos os passos descritos na seção 5.3.3.1, com exceção da extração manual, ou seja, as mesmas 359 sentenças foram remetidas ao processo de classificação manual de complicação resultando em um subconjunto de 163 sentenças de complicação. Após a classificação manual, este subconjunto foi processado pelo agrupador de parágrafos, estendendo-se em um novo subconjunto resultando um total de 325 sentenças. Ademais, foi realizada a classificação manual no subconjunto de 325 sentenças para classificar em sentenças de tratamento. Com a classificação manual remanesceram 112 sentenças de tratamento, nas quais foram etiquetadas com POS e enviadas para o processo de extração.

Após rotular as sentenças foram aplicadas as regras do conjunto amplo, enxuto (verbo ou palavra representativa+POS) e Somente POS descritas no Capítulo 4, e ainda, para obter uma visão mais ampla dos resultados, a extração automática foi dividida em 2 etapas: 1) Aplicação das regras para obter o resultado da extração de todos os termos, ou seja, termos distintos e termos repetidos; e 2) Aplicação das regras para obter o resultado da extração apenas nos termos distintos.

Na Tabela 5.15 é apresentado o resultado da extração automática aplicada nas 112 sentenças de tratamento para todos os termos (distintos e repetidos - coluna Termos-Todos) e somente para os termos distintos (coluna Termos Distintos).

Neste experimento, o conjunto de 9 regras com verbo ou palavra representativa+POS obteve 73% de precisão para todos os termos, ficando inferior ao conjunto de 2 regras (que obteve 96% de precisão), porém foi o único que atingiu 100% de revocação para termos distintos.

A aplicação dos conjuntos de regras nas sentenças citadas neste experimento pode ser vistos no apêndice digital.

Tabela 5.15 - Extração automática a partir da classificação manual.

Regra	Termos – Todos			Termos Distintos		
	Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
2 regras com verbo+POS	96%	19%	31%	80%	50%	61%
9 regras com verbo+POS	73%	24%	36%	42%	100%	59%
7 Somente POS	49%	45%	47%	14%	75%	24%

5.3.3.3 Considerações Finais

Nesta seção foi descrita a avaliação realizada na metodologia proposta neste projeto de mestrado, porém utilizando classificação manual. O objetivo do experimento foi de classificar manualmente as sentenças de complicação, agrupar as sentenças por parágrafo, classificar manualmente as sentenças em tratamento, etiquetar as sentenças conforme sua classe gramatical e efetivar a extração automática a partir do conjunto de regras.

Para avaliar este experimento, foi necessário dividi-lo em duas fases: (i) classificar as sentenças em complicação em um processo manual, realizar o agrupamento de sentenças por parágrafos, classificar manualmente as sentenças em tratamento, etiquetar as sentenças com POS e realizar extração manual de termos de tratamento; (ii) classificar as sentenças em complicação manualmente, agrupar as sentenças por parágrafos, classificar manualmente as sentenças em tratamento, etiquetar as sentenças com POS e aplicar o conjunto de regras para efetivamente efetuar a extração automática de termos de tratamento.

Na extração manual foram encontrados 141 termos relevantes de tratamento partir da classificação manual de complicação e tratamento. O resultado mostra que a classificação manual não eliminou qualquer sentença/termo não repetido, portanto, obteve 100% de revocação para termos distintos.

Na avaliação de identificação de termos relevantes, a extração automática obteve 96% de precisão e 19% de revocação para todos os termos utilizando o conjunto de 2 regras, e 50% de revocação para termos distintos. Considera-se que os termos de tratamento se repetem ao longo do artigo, portanto o fato da revocação

ter sido baixa não impacta o processo de extração. Ademais, o conjunto de 9 regras obteve 100% de revocação para os termos distintos, promovendo então, a extração por completa dos termos distintos relevantes existentes no artigo comparado com a extração manual.

5.3.4 Dicionário

Nomeado como Dicionário, o banco de dados biomédico contém termos de tratamento relevantes avaliados pelo especialista da área. Conforme já citado, usamos uma técnica alternativa para reduzir os problemas de restrição de nomes em abordagens baseada em dicionário. Para tanto, com o uso de variações de termos e siglas, foi possível identificar 100% das ocorrências de tratamentos já conhecidas. Exemplos de termos armazenados no dicionário são: *hydroxyurea* e sua variação *HU*, *placebo* e *antibiotic*.

O dicionário biomédico contém 26 termos relevantes de tratamento consolidados pelo especialista da área.

O conjunto de regras permite encontrar novos tratamentos desconhecidos e obteve a extração em 31% de revocação no que é predito no dicionário. Apesar das regras identificarem novos termos, o uso do dicionário é necessário e imprescindível para garantir uma alta revocação dos termos conhecidos.

5.3.5 Considerações Finais

Neste capítulo foram apresentados os experimentos correspondentes a duas etapas da metodologia: fase de classificação de sentenças e fase de identificação de termos relevantes. O objetivo deste capítulo foi realizar duas provas de conceito para validar o processo proposto aplicado no domínio biomédico, em particular, aplicado em artigos científicos da doença da Anemia Falciforme.

As fases de classificação de sentenças e identificação de termos relevantes são etapas que contribuem para a extração de informação. Estas fases foram avaliadas separadamente para permitirem atingir um resultado parcial do processo, alcançando o objetivo comum que é o resultado final desejado da extração de informação.

A primeira prova de conceito foi realizada na fase de classificação de sentenças, cujo objetivo foi utilizar um algoritmo de aprendizado de máquina supervisionado para classificar as sentenças sobre a doença da Anemia Falciforme em suas respectivas classes. A classificação foi usada no processo proposto como um filtro que selecionou apenas as sentenças de interesse, diminuindo o custo de análise na fase de identificação de termos relevantes das sentenças de um artigo.

A segunda prova de conceito realizada foi na fase de identificação de termos, cujo objetivo foi extrair os termos relevantes das sentenças classificadas na etapa anterior. As medidas de precisão, revocação e medida-F foram utilizadas para avaliar o percentual de termos extraídos. As abordagens baseadas em dicionário e regras foram utilizadas para identificar os termos relevantes nas sentenças relacionadas à classe de tratamentos.

Na Tabela 5.16 são sintetizados os resultados de todos os experimentos apresentados neste capítulo e na próxima seção são apresentadas as conclusões deste trabalho de mestrado, as contribuições e indicações para trabalhos futuros.

Tabela 5.16 - Resultados dos experimentos.

Experimento	Objetivo			Precisão	Revocação	Acurácia	Medida-F
Fase de uso do modelo de classificação	Classificação automática para novas sentenças	Complicação (C1)		85%	64%	79%	73%
		Tratamento (C2)		88%	51%	71%	64,5%
Identificação de Termos Relevantes	Classificação automática de sentenças de tratamento com complicação e extração com regras	Extração manual a partir da classificação automática		Termos – Todos			Termos Distintos
				Precisão	Revocação	Medida-F	Precisão	Revocação	Medida-F
				100%	37,82%	54,88%	100%	50%	66,66%
		Extração automática a partir da classificação automática	2 regras	100%	27%	42%	100%	100%	100%
			9 regras	62%	25%	35%	33%	100%	50%
			7 regras	41%	49%	44%	13%	100%	23%
	Classificação automática de sentenças de tratamento sem complicação e extração com regras	Extração manual a partir da classificação automática		100%	60,90%	75,70%	100%	75%	85,71%
		Extração automática a partir da classificação automática	2 regras	91%	23%	36%	80%	66%	72%
			9 regras	67%	24%	35%	37%	100%	54%
			7 regras	42%	48%	45%	14%	100%	24%
Identificação de Termos Relevantes	Classificação manual de sentenças de tratamento com complicação e extração com regras	Extração manual a partir da classificação manual		100%	90,38%	94,94%	100%	100%	100%
		Extração automática a partir da classificação manual	2 regras	96%	19%	31%	80%	50%	61%
			9 regras	73%	24%	36%	42%	100%	59%
			7 regras	49%	45%	47%	14%	75%	24%

Capítulo 6

CONCLUSÃO

Neste capítulo são apresentadas as conclusões deste trabalho de pesquisa em nível de mestrado, as suas principais contribuições e também são indicados algumas sugestões de trabalhos futuros.

6.1 Considerações Iniciais

Nesta dissertação foi proposta uma metodologia de pré-processamento textual para extrair informação de termos de tratamentos de artigos científicos do domínio biomédico. A metodologia discutida no Capítulo 5 é composta por seis passos: **Classificação de Sentenças em Complicação** (passo 1), **Agrupamento por Parágrafos** (passo 2), **Classificação de Sentenças em Tratamento** (passo 3), **Etiquetagem com POS** (passo 4), **Identificação dos Termos Relevantes** (passo 5) e **Armazenamento no Banco de Dados** (passo 6).

Considera-se como hipótese deste trabalho que na maioria dos casos os termos de tratamento ocorrem em uma mesma sentença que possui um termo de complicação ou em sentenças próximas em um mesmo parágrafo. A busca inicial de sentenças que possuem complicações melhora a eficiência na identificação e extração de termos de tratamento. Isso acontece porque tratamentos ocorrem principalmente na mesma sentença de complicação ou em sentenças próximas no mesmo parágrafo. Ademais, a filtragem das sentenças na fase de classificação reduz o custo da fase posterior de identificação de termos relevantes (ou seja, da extração de informação propriamente dita).

O motivo de utilizar classificação de sentenças é que a classificação serve como um filtro que seleciona apenas as sentenças de interesse, auxiliando no processo de extração de informação porque o termo de tratamento a ser extraído está localizado na sentença previamente selecionada, e ainda, diminui-se o custo de análise das sentenças de um artigo. O dicionário é fundamental em auxiliar na extração de termos conhecidos, e a justificativa para utilizar regras é que com esta abordagem é possível extrair novos termos ainda não descobertos na área biomédica.

A metodologia foi validada separadamente por meio de provas de conceito com base em artigos científicos relacionados à doença Anemia Falciforme. Na fase de classificação de sentenças, o experimento teve como objetivo criar e testar o modelo de classificação. Para isso, foi utilizado um conjunto de 765 sentenças classificadas manualmente, no qual foram examinadas e analisadas para representar dois cenários: (i) o conjunto de 765 sentenças classificadas em classes de “complicação” e “outros”, para posteriormente desenvolver o classificador de complicação (C1); e (ii) o mesmo conjunto de sentenças classificadas em classes de “tratamento” e “outros” para desenvolver o classificador de tratamento (C2). Todas as sentenças foram utilizadas para o treinamento e o teste dos classificadores usando o método de particionamento *10-fold cross validation.* Os modelos de classificação foram criados utilizando o algoritmo SVM juntamente com a combinação dos filtros *Remove Misclassified* e *Resample.* Posteriormente, foi realizado o experimento na fase de uso do modelo de classificação com o objetivo de avaliar os classificadores na classificação de novas sentenças utilizando a medida acurácia. Para isso, foi utilizado um conjunto de 359 novas sentenças. O percentual de acurácia obtido foi de 79% para o classificador de complicação (C1) e 71% para o classificador de tratamento. Os resultados obtidos neste trabalho não estão longe dos valores comumente encontrados na literatura, apresentando inclusive melhores resultados comparados com alguns trabalhos que extraem informações de resumos e artigos completos, como pode ser visto na Tabela 6.1.

Na fase de identificação de termos relevantes, foram realizados três tipos experimentos a fim de avaliar a extração de termos de tratamento, a partir das sentenças que foram classificadas como sendo sentenças de tratamento. Os experimentos são: (i) o experimento principal e essencial que executa todos os passos da metodologia, sem nenhuma modificação, ou seja, o experimento

adequado para comprovar a hipótese deste trabalho; (ii) experimento que executa parcialmente a metodologia, sem considerar o classificador de complicação; (iii) experimento que executa todos os passos da metodologia, porém utiliza a classificação manual. O motivo de efetuar os dois últimos experimentos foi para promover uma comparação entre classificação manual e classificação automática, e principalmente para validar a hipótese deste trabalho, ou seja, se classificando inicialmente as sentenças em sentenças de complicação, tenderia a ter um resultado favorável na extração de termos de tratamento.

Para o experimento principal (i), foram utilizadas as abordagens de regra e dicionário em um conjunto de 359 novas sentenças. Na extração automática de termos foram aplicadas as regras do conjunto amplo e enxuto, ambos utilizando a estratégia 1 (verbo ou palavra representativa+POS) e a estratégia 2 (regras com Somente POS). O resultado da extração automática obteve precisão de 100%, revocação de 27%, medida-F de 42% para todos os termos e revocação de 100% para termos distintos, utilizando o conjunto enxuto de 2 regras (verbo ou palavra representativa+POS). O percentual de precisão, revocação e medida-F para o mesmo experimento utilizando o conjunto amplo de 9 regras (verbo ou palavra representativa+POS) foi de 62%, 25% e 35% respectivamente, para todos os termos. Já para os termos distintos, o percentual de revocação também foi de 100%. Considera-se que os termos de tratamento se repetem do longo do artigo, portanto, o baixo percentual de revocação em ambos os conjuntos de regras não impactou o processo de extração de informação.

O conjunto de regras que melhor representou foi o conjunto enxuto de 2 regras, no entanto, o conjunto de regras que obteve 100% de revocação para termos distintos em todos os experimentos realizados foi o conjunto amplo de 9 regras.

Já o conjunto de 7 regras Somente POS obteve maior percentual de revocação para todos os termos em todos os experimentos, mas conclui-se que não é preciso aplicar este conjunto para extrair mais termos repetidos, pois o conjunto amplo foi capaz de extrair todos os termos não repetidos que continham nas sentenças, fazendo com que seja desnecessária a aplicação das regras Somente POS, logo, diminui-se o custo da extração.

Os resultados dos dois últimos experimentos foram úteis para comprovar que: (i) classificando as sentenças automaticamente em complicação e tratamento, ou seja, executando todos os passos da metodologia (experimento principal), e

utilizando o conjunto de 2 regras com verbo ou palavra representativa+POS, a extração de termos de tratamento obtém um percentual de medida-F maior (42% para todos os termos e 100% para termos distintos), comparado à classificação manual (31% para todos os termos e 61% para termos distintos) e também comparado ao processo parcial, ou seja, sem utilizar o classificador de complicação (36% para todos os termos e 72% para termos distintos). Portanto, comprova-se a hipótese desde trabalho que os termos de tratamento estão essencialmente nas sentenças de complicação ou próximas em um mesmo parágrafo, e ademais, que a classificação automática obteve resultado superior à classificação manual. Logo, comprova-se a hipótese que é possível extrair termos de tratamento de forma semiautomatica e ainda, alcançar uma precisão e revocação aceitável na identificação de novos termos utilizando regras especificas desenvolvidas para um domínio biomédico.

Para o experimento com extração manual em todo o artigo, conclui-se que apenas selecionando os parágrafos de complicação é possível obter 100% de revocação para os termos distintos. Sendo assim, é comprovada a hipótese deste trabalho, em que os termos de tratamento estão em uma sentença que possui um termo de complicação ou ocorre em sentenças próximas em um mesmo parágrafo.

Ademais, para o experimento com dicionário, com o uso de variações de termos e siglas, foi possível identificar 100% das ocorrências de tratamentos já conhecidas, e o uso desta abordagem se torna necessário para garantir uma alta revocação dos termos de tratamento conhecidos. Sendo assim, comprova-se a hipótese de que é possível alcançar uma alta precisão e revocação de termos distintos de tratamentos conhecidos utilizando um dicionário estendido com variações de termos e siglas.

A maioria dos trabalhos relacionados que extraem informação possuem objetivos diferentes, porém, os valores obtidos por este trabalho de mestrado não estão longe dos valores comumente encontrados na literatura, apresentando inclusive melhores resultados comparados com alguns trabalhos que extraem informações de resumos e artigos completos, conforme pode ser observado na Tabela 6.1.

Tabela 6.1 - Trabalhos relacionados.

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)
Yang et al. (2009)		x[3]		Proteína	BioPPIExtractor	Extrair informação	Sim	Resumos Prec. 55,4% Rev. 41,6%
Yang et al. (2009)			x[3]	Proteína	BioPPISVMExtractor	Extrair informação	Sim	Resumos Prec. 49,2% Rev. 71,8%
MATOS (2010)	x	x	x	Complicação e Benefício da Anemia Falciforme	SCAeXtractor	Povoar um banco de dados	Sim	Artigos Acurácia 62,33% Prec. 74,75% Rev. 87,06% Med.F 80,43%
DUQUE (2012)	x	x	x	Tratamentos da Anemia Falciforme	SCAeXtractor	Povoar um banco de dados	Sim	Artigos Acurácia 79% <u>Conjunto Enxuto de 2 regras</u>: Prec. 100% Rev. 27% Med.F 42%

								Conjunto Amplo de 9 regras: Prec. 62% Rev. 25% Med.F 35%

[1] Ontologia e expressões regulares, respectivamente, do sistema Textpresso.

[2] Prec. significa Precisão e Rev. significa Revocação.

[3] Método baseado em *Conditional Random Fields* (CRF).

Este projeto de mestrado tem também uma conotação social, pois auxilia no processo de combate a uma importante doença considerada problema de saúde pública no Brasil, e entende-se que a metodologia proposta neste trabalho favorece os médicos e especialistas da área a terem acesso prático e rápido a pesquisas em artigos científicos sobre a doença Anemia Falciforme.

A seguir são destacadas as contribuições deste trabalho, a adaptabilidade da metodologia proposta, os trabalhos futuros e, por fim, as produções científicas e técnicas desenvolvidas durante o mestrado.

6.2 Contribuições

A essência deste projeto é a proposta de uma metodologia para extrair informação sobre tratamentos de doenças de artigos completos do domínio biomédico. Esta metodologia parte da hipótese que a maioria dos casos os termos de tratamento ocorrem na mesma sentença de uma complicação ou em sentenças próximas em um mesmo parágrafo. Para isto, aplicamos uma combinação de técnicas variadas com características distintas do que comumente são encontrados nos trabalhos relacionados, e ainda, nossa metodologia possui a vantagem de realizar a extração em artigos científicos completos, o que é efetuado por poucos trabalhos identificados na literatura.

São evidenciadas as contribuições teóricas e práticas:

Teóricas:

- Metodologia para Extração de Informação

Práticas:

- 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;

- Criação e disponibilização de ferramentas para extração de informação e agrupamento de parágrafos, ambos implementados em Perl.

6.3 Adaptabilidade da Metodologia Proposta

A metodologia proposta possui seis passos, conforme ilustrado na Figura 4.1. Todos os passos da metodologia com exceção do passo 5 são independentes do domínio e pode ser aplicado a qualquer domínio sem modificações. A entrada de dados tem a restrição dos artigos científicos estarem nos formatos XML ou TXT e no passo 5 é realizado o processo de extração de informação dos termos de tratamento. O passo 5 possui a dependência das abordagens de dicionário e regras para extração de informação. Como a extração de informação realizada pelo dicionário é o casamento exato dos termos armazenados no dicionário, o custo de adequação depende da existência de um novo dicionário do domínio em particular, sendo que o custo de sua criação pode não ser desprezível e em alguns casos alto. Já a extração de informação efetuada pela regra impõe uma adaptação maior, pois as regras são criadas a partir de uma análise no conjunto de sentenças do domínio específico, no qual são encontrados termos relevantes que inclui um conjunto de padrões *Part-Of-Speech* que são dependentemente dos termos a serem extraídos. As regras específicas (conjunto amplo da estratégia 1 e conjunto POS da estratégia 2) são dependentes da doença da Anemia Falciforme e as regras genéricas

(conjunto enxuto da estratégia 1) podem ser adaptadas com algum esforço para adequar a outros tipos de doenças do âmbito médico.

Contudo, as duas estratégias para identificar os termos relevantes descritas neste trabalho empregando a abordagem de regra podem ser utilizadas e adaptadas para a extração de informação de termos que não sejam tratamento de doenças.

6.4 Trabalhos Futuros

A seguir são enumerados os trabalhos futuros:

- Investigar a identificação de tratamentos e informações de sintomas em artigos científicos de outras doenças;
- Utilização de índices para acelerar a identificação de termos;
- Identificar relacionamentos semânticos entre termos de diversos textos científicos da área biomédica;
- Utilizar um modelo *bag-of-related-words* como um formato de representação de documentos (ROSSI; REZENDE, 2011);
- Outras áreas biomédicas podem também beneficiar da abordagem de mineração de textos.

6.5 Produção Científica e Técnica

A discussão sobre a proposta deste trabalho de mestrado foi publicada em um evento nacional, apresentando-se a visão geral do processo com resultados sobre a classificação, dicionários e regras. A produção científica é listada a seguir:

DUQUE, J. L. et al. Um processo baseado em parágrafos para a extração de tratamentos em artigos científicos do domínio biomédico. In: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL), 8th, 2011, Cuiabá, Mato Grosso. **Proceedings...** Symposium in Information and Human Language Technology, 2011. p. 124-133.

Foram desenvolvidos scripts em Perl para automatizar a extração com regras e o agrupamento de sentenças por parágrafo, a saber:

DUQUE, J. L.; CIFERRI, R. R.; PARDO, T. A. S. **Ferramenta de Extração de Informação - Regras 2011**. Disponível em: <http://gbd.dc.ufscar.br/~julianalduque/files/perl.SCD.2011.08.31.rar>. Acesso em: 25 out. 2011.

DUQUE, J. L.; CIFERRI, R. R.; PARDO, T. A. S. **Processamento por Parágrafo. 2011**. Disponível em: <http://gbd.dc.ufscar.br/~julianalduque/files/perl.Paragraph.2011.08.31.rar>. Acesso em: 25 out. 2011.

Ainda, foi realizada uma produção científica que não tem relação com este trabalho, mas foi desempenhada ao longo do mestrado, a saber:

BELLINI, A ; CIRILO, C. E ; FERRAZ, V. R. T ; DUQUE, J. L. ; ANNIBAL, L. P. ; ARAUJO, J. G ; DURELLI, R. S. ; MARCONDES, C. A. C . A Low Cost Positioning and Visualization System using Smartphones for Emergency Ambulance Service. In: Proceedings of the 2010 ICSE Workshop on Software Engineering in Health Care, 2010, Cape Town, South Africa. **Proceedings…** ACM, 2010. p.12-18.

REFERÊNCIAS

AGARWAL, S.; YU, H.; KOHANE, I. BioNOT: A searchable database of biomedical negated sentences. **BMC Bioinformatics**, v. 12, n. 1, p. 420, 2011. Disponível em: <http://www.biomedcentral.com/1471-2105/12/420>. Acesso em: 30 out. 2011.

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, 1994, **Proceedings...** Morgan Kaufmann Publishers Inc., 1994. p. 487-499.

AHMED, S. T. et al. IntEx: a syntactic role driven protein-protein interaction extractor for bio-medical text. In: Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, 2005, Detroit, Michigan. **Proceedings...** Association for Computational Linguistics, 2005. p. 54-61.

ANANIADOU, S.; KELL, D. B.; TSUJII, J.-I. Text mining and its potential applications in systems biology **Trends in Biotechnology**, v. 24, n. 12, p. 571-579, 2006.

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.

ANTHONY, L.; LASHKIA, G. V. Mover: a machine learning tool to assist in the reading and writing of technical papers. **IEEE Transactions on Professional Communication**, v. 46, n. 3, p. 185-193, 2003. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1227591>. Acesso em: 23 fev. 2010.

ARANHA, C. N. **Uma abordagem de pré-processamento automático para mineração de textos em português:** sob o enfoque da inteligência computacional. 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: 11 jan. 2010.

ARANHA, C. N.; PASSOS, E. P. L. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação (RESI)**, v.8, n. 2, p. 1-8, 2006. Disponível em: <http://www.facecla.com.br/revistas/resi/edicoes/ed8tut01.pdf>. Acesso em: 27 nov. 2009.

BATISTA, G. E. A. P. A.; CARVALHO, A. C. P. L. F.; MONARD, M. C. Applying One-Sided Selection to Unbalanced Datasets. In: Proceedings of the Mexican International Conference on Artificial Intelligence: Advances in Artificial Intelligence, 2000, **Proceedings...** Springer-Verlag, 2000. p. 315-325.

BATISTA, G. E. A. P. A.; MONARD, M. C. Utilizando métodos estatísticos de resampling para estimar a performance de sistemas de aprendizado. In: III Workshop de Dissertações Defendidas, 1998, São Carlos. **Proceedings...** ICMC-USP, 1998. p. 173-184. Disponível em: <http://www.icmc.usp.br/~mcmonard/public/icmcwshG1998.pdf>. Acesso em: 17 fev. 2010.

BLASCHKE, C.; VALENCIA, A. The frame-based module of the SUISEKI information extraction system. **IEEE Intelligent Systems**, v. 17, n. 2, p. 14-20, 2002.

BREMER, E. G. et al. Text mining of full text articles and creation of a knowledge base for analysis of microarray data. In: Knowledge Exploration in Life Science Informatics, International Symposium, 2004, Milan, Italy. **Proceedings...** 2004. p. 84-95.

CARLSON, A. et al. Coupled Semi-Supervised Learning for Information Extraction. In: Proceedings of the third ACM international conference on Web search and data mining, 2010, New York, NY, USA. **Proceedings...** 2010. p. 101-110.

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: 10 set. 2011.

CHEN, H. et al. **Medical informatics:** knowledge management and data mining in biomedicine. Berlin: Springer, 2005. 624 p. Disponível em: <http://ai.arizona.edu/hchen/chencourse/MedBook/>. Acesso em: 23 out. 2009.

CLEVERDON, C. W.; MILLS, J.; KEEN, M. **Factors determining the performance of indexing systems**. College of Aeronautics, 1966. 376 p. Disponível em: <https://dspace.lib.cranfield.ac.uk/bitstream/1826/863/2/1966e.pdf>. Acesso em: 29 jan. 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 dez. 2009.

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 dez. 2009.

DING, J. et al. Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser. In: In the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03), 2003, Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2003. p. 467.

DÖRRE, J.; GERSTL, P.; SEIFFERT, R. Text mining: finding nuggets in mountains of textual data. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 5th, 1999, San Diego, California, United States. **Proceedings...** New York: ACM, 1999. p. 398-401. Disponível em: <http://doi.acm.org/10.1145/312129.312299>. Acesso em: 20 fev. 2010.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. D. A. Mineração de textos. In: REZENDE, S. O. (Ed.). **Sistemas inteligentes:** fundamentos e aplicações. São Carlos: Manole, 2003. p. 337-370. cap. 13.

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: 14 fev. 2010.

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

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. Disponível em: <http://dx.doi.org/10.1186/1471-2105-10-S2-S6>. Acesso em: 12 mar. 2010.

GOMIDE, R. S. **ADI-Minebio: Mineração Baseada em Grafos Aplicada à Área Biomédica.** Dissertação (Mestrado em Ciência da Computação) – Faculdade Metodista de Piracicaba, Piracicaba, 2011.

GRINBERG, D.; LAFFERTY, J.; SLEATOR, D. **A Robust Parsing Algorithm for Link Grammars**. 1995.

GUPTA, V.; LEHAL, G. 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.

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: 23 nov. 2009.

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: 20 fev. 2010.

J. DING, D. B., D. NETTLETON, E. S. WURTELE Mining medline: Abstracts, sentences, or phrases? **In Proceedings of the Pacific Symposium on Biocomputing**, p. p. 326-337, 2002.

JACKSON, P.; MOULINIER, I. **Natural language processing for online applications:** text retrieval, extraction and categorization. Amsterdam: John Benjamins, 2002. 225 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 nov. 2009.

JOACHIMS, T. Making large-scale support vector machine learning practical. In: (Ed.). **Advances in kernel methods: support vector learning**. Cambridge, MA, USA: MIT Press, 1999. p. 169-184.

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.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, 14th, 1995, Montréal, Québec. **Proceedings...** Morgan Kaufmann, 1995. p. 1137-1145. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.529>. Acesso em: 13 fev. 2010.

KOU, Z.; COHEN, W. W.; MURPHY, R. F. High-recall protein entity recognition using a dictionary. **Bioinformatics**, v. 21, p. i266-273, 2005. Disponível em: <http://dx.doi.org/10.1093/bioinformatics/bti1006>. Acesso em: 23 nov. 2009.

KRAUS, S.; BLAKE, C.; WEST, S. L. Information extraction from medical notes. In: Word Congress on Health (Medical) Informatics, 12th, 2007, Brisbane, Australia. **Proceedings...** Medinfo, 2007. p. 1913-1915. Disponível em: <http://ils.unc.edu/~cablake/Papers/KrausBlakeWestMEDINFO2007.pdf>. Acesso em: 28 nov. 2009.

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 nov. 2009.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, 2001, **Proceedings...** Morgan Kaufmann Publishers Inc., 2001. p. 282-289.

LEE, C.-H.; WU, C.-H.; YANG, H.-C. Text mining of clinical records for cancer diagnosis. In: Proceedings of the Second International Conference on Innovative Computing, Information and Control, 2007, Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2007. p. 172-175.

LEVENSHTEIN, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. **Soviet Physics Doklady**, v. 10, p. 707-+, 1966.

LOMBARDI, L. O. **Extração de Padrões Numéricos em Artigos Científicos do Domínio Biomédico.** Dissertação (Mestrado em Ciência da Computação) – Faculdade Metodista de Piracicaba, Piracicaba, 2011.

LUO, Q. Advancing knowledge discovery and data mining. In: Proceedings of the First International Workshop on Knowledge Discovery and Data Mining, 2008, Adelaide, Australia. **Proceedings...** IEEE Computer Society, 2008. p. 3-5.

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: 13 jan. 2010.

MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. London, England: MIT Press, 1999. 680 p.

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: 15 mar. 2011.

MATOS, P. F. **Metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico**. 159 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de São Carlos, São Carlos, 2010. Disponível em: <http://gbd.dc.ufscar.br/~pablofmatos/files/DissPFM.set2010.pdf>. Acesso em: 30 set. 2010.

MATOS, P. F. et al. **Relatório Técnico "Métricas de Avaliação"**. São Carlos: Departamento de Computação. Universidade Federal de São Carlos, 2009, p. 15. Disponível em: <http://sca.dc.ufscar.br/download/files/report.metricas.pdf>. Acesso em: 10 fev. 2010.

MHAMDI, F.; ELLOUMI, M. A new survey on knowledge discovery and data mining. In: Research Challenges in Information Science, 2008. RCIS 2008. Second International Conference on, 2nd, 2008, Marrakech, Morocco. **Proceedings...** 2008. p. 427-432

MITCHELL, T. M. **Machine learning**. Boston: McGraw-Hill, 1997. 414 p.

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.

MOONEY, R. J.; BUNESCU, R. Mining knowledge from text using information extraction. **SIGKDD Explor. Newsl.**, v. 7, n. 1, p. 3-10, 2005. Disponível em: <http://delivery.acm.org/10.1145/1090000/1089817/p3-mooney.pdf?key1=1089817&key2=1641795521&coll=GUIDE&dl=&CFID=57239718&CFTOKEN=55078684>. Acesso em: 14 out. 2009.

MULLER HM, K. E., STERNBERG PW Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. **PLoS Biol** p. 2004. Disponível em: <http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0020309>. Acesso em: 22 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.

PARDO, T. A. S. **DMSumm:** um gerador automático de sumários. 103 f. Dissertação (Mestrado em Ciência da Computação) – Departamento de Ciência da Computação, Universidade Federal de São Carlos, São Carlos, 2002. Disponível em: <http://www.icmc.usp.br/~taspardo/DISSERTATION-Pardo.pdf>. Acesso em: 26 fev. 2010.

REZENDE, S. O. et al. Mineração de Dados. In: REZENDE, S. O. (Ed.). **Sistemas inteligentes:** fundamentos e aplicações. São Carlos: Manole, 2003. p. 307-335. cap. 12.

ROSSI, R. G.; REZENDE, S. O. Building a topic hierarchy using the bag-of-related-words representation. In: Proceedings of the 11th ACM symposium on Document engineering, 2011, Mountain View, California, USA. **Proceedings...** ACM, 2011. p. 195-204.

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: 20 fev. 2010.

SEGAL, J. B. et al. **Hydroxyurea for the treatment of sickle cell disease**. Rockville, MD: Agency for Healthcare Research and Quality, 2008, p. 1-298. (Evidence Report/Technology Assessment, n. 165). Disponível em: <http://www.ahrq.gov/downloads/pub/evidence/pdf/hydroxyurea/hydroxscd.pdf>. Acesso em: 26 nov. 2009.

SILVA, B. C. D. D. et al. **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. São Carlos: Núcleo Interinstitucional de Linguística Computacional - NILC - ICMC-USP, 2007, p. 121.

SILVA, R. B. P.; RAMALHO, A. S.; CASSORLA, R. M. S. A anemia falciforme como problema de saúde pública no Brasil. **Revista de Saúde Pública**, v. 27, n. 1, p. 54-58, 1993. Disponível em: <http://dx.doi.org/10.1590/S0034-89101993000100009>. Acesso em: 29 jan. 2010.

SILVA_PINTO, A. C. et al. **Technical Report "Sickle Cell Anemia"** Technical Report "Sickle Cell Anemia". São Carlos: Federal University of São Carlos, 2009, p. 16. Disponível em: <http://sca.dc.ufscar.br/download/files/report.sca.pdf>. Acesso em: 19 nov. 2009.

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 nov. 2009.

STAVRIANOU, A.; ANDRITSOS, P.; NICOLOYANNIS, N. Overview and semantic issues of text mining. **SIGMOD Rec.**, v. 36, n. 3, p. 23-34, 2007. Disponível em: <http://delivery.acm.org/10.1145/1330000/1324190/p23-stavrianou.pdf?key1=1324190&key2=5355506521&coll=portal&dl=ACM&CFID=57248742&CFTOKEN=91634109>. Acesso em: 03 out. 2009.

TANABE, L.; WILBUR, W. J. Tagging gene and protein names in biomedical text. **Bioinformatics**, v. 18, n. 8, p. 1124-1132, 2002a.

_____. Tagging gene and protein names in full text articles. In: Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain - Volume 3, 2002b, Phildadelphia, Pennsylvania. **Proceedings...** Morristown, NJ: Association for Computational Linguistics, 2002b. p. 9-13.

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: 15 mar. 2011.

THOMPSON, P. et al. The BioLexicon: a large-scale terminological resource for biomedical text mining. **BMC Bioinformatics**, v. 12, n. 1, p. 397, 2011.

TOUTANOVA, K. et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Human Language Technology Conference (HLT-NAACL 2003), 2003, Edmonton, Canada. **Proceedings...** Association for Computational Linguistics, 2003. p. 173-180. Disponível em: <http://dx.doi.org/10.3115/1073445.1073478>. Acesso em: 15 mar. 2011.

TOUTANOVA, K.; MANNING, C. D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-2000), 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: 15 mar. 2011.

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: 12 jan. 2010.

VAPNIK, V. N. **The nature of statistical learning theory**. New York, NY, USA: Springer-Verlag New York, Inc., 1995. p.

WITTEN, I. H.; FRANK, E. **Data mining:** pratical machine learning tools and techniques with Java implementations. 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005. 525 p.

XENARIOS, I., FERNANDEZ, E., SALWINSKI, L., DUAN, X., THOMPSON, M., MARCOTTE, E. AND EISENBERG, D DIP: the database of interacting proteins. **Nucl. Acid Res**, v. v-29, p. 239-241, 2000.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML), 14th, 1997, San Francisco, CA. **Proceedings...** Morgan Kaufmann Publishers, 1997. p. 412-420. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.9956>. Acesso em: 21 fev. 2010.

YANG, Z.; LIN, H.; WU, B. BioPPIExtractor: A protein-protein interaction extraction system for biomedical literature. **Expert Syst. Appl.**, v. 36, n. 2, p. 2228-2233, 2009.

ZELENKO, D. et al. Kernel methods for relation extraction. **Journal of Machine Learning Research**, v. 3, p. 36, 2003.

ZHIHAO YANG; HONGFEI LIN; LI, Y. BioPPISVMExtractor: A protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. **Journal of Biomedical Informatics**, v. 43, p. 88-96, 2009.