UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM MÉTODO PARA DESCOBERTA DE RELACIONAMENTOS SEMÂNTICOS DO TIPO "CAUSA E EFEITO" EM SENTENÇAS DE ARTIGOS CIENTÍFICOS DO DOMÍNIO BIOMÉDICO

RICARDO BRIGATO SCHEICHER

ORIENTADOR: PROF. DR. RICARDO RODRIGUES CIFERRI

São Carlos – SP

Novembro/2013

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UM MÉTODO PARA DESCOBERTA DE RELACIONAMENTOS SEMÂNTICOS DO TIPO "CAUSA E EFEITO" EM SENTENÇAS DE ARTIGOS CIENTÍFICOS DO DOMÍNIO BIOMÉDICO

RICARDO BRIGATO SCHEICHER

Dissertação 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, área de concentração: Engenharia de Software, Banco de Dados e IHC.
Orientador: Prof. Dr. Ricardo Rodrigues Ciferri

São Carlos – SP

Novembro/2013

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S318md

Scheicher, Ricardo Brigato.
Um método para descoberta de relacionamentos semânticos do tipo “causa e efeito” em sentenças de artigos científicos do domínio biomédico / Ricardo Brigato Scheicher. -- São Carlos : UFSCar, 2015.
131 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2013.

1. Inteligência artificial. 2. Relações semânticas. 3. Rede semântica. 4. Extração de informação. 5. Mineração de textos. 6. Domínio biomédico. 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 método para Descoberta de Relacionamentos Semânticos do Tipo "Causa e Efeito" em Sentenças de Artigos Científicos do Domínio Biomédico"

Ricardo Brigato Scheicher

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:

Ricardo Rodrigues Ciferri

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

Prof. Dr. Thiago Alexandre Salgueiro Pardo
(ICMC/USP)

Profa. Dra. Ana Cristina Silva Pinto
(USP/RP)

São Carlos
Novembro/2013

AGRADECIMENTOS

Primeiramente, quero agradecer a Deus e à minha família, meu pai Jorge, minha mãe Magda, pelo apoio constante e por sempre acreditarem que eu atingiria mais essa etapa em minha vida.

Meu próximo agradecimento é para minha namorada, Aline, que sempre me fortalece nos momentos mais difíceis e recupera meu potencial, fazendo-me enxergar que a vida não era apenas acadêmica.

Agradeço de coração ao meu orientador, prof. Ricardo Rodrigues Ciferri. Mesmo passando por suas dificuldades, me ajudou, me ensinou e me colocou de cabeça erguida até a conclusão deste trabalho.

Agradeço aos grandes amigos que fiz durante todo o mestrado e que me ajudaram em todas as minhas dificuldades, Leonardo Taba, Roberto Gueleri, Fernando Proença, Shermila Guerra, Cesar Teixeira, Thiago Vieira, Augusto Cesar, Vinicius Ferraz, Elias Adriano, Debora Marrach, Rodrigo Barroso e Mayra Zegarra.

Aos professores Marilde Terezinha, Renato Bueno, Estevam Hruschka, Thiago Pardo, que, além do meu orientador fizeram uma diferença gigantesca em meu trabalho e em minha vida. Com vocês aprendi muito.

A CAPES pelo apoio financeiro.

A todos que fizeram a diferença em minha vida.

Obrigado!

RESUMO

Atualmente, existe uma enorme quantidade de material científico escrito em formato textual e publicado em meios eletrônicos (artigos em anais de eventos e periódicos). Na área biomédica, pesquisadores necessitam assimilar uma grande parte deste conteúdo com a finalidade de se atualizarem e, por conseguinte realizarem diagnósticos mais precisos e aplicar tratamentos mais modernos e eficazes. A tarefa de obtenção de conhecimento é bastante onerosa e o processo manual para anotar relacionamentos e propor novas hipóteses de tratamentos torna-se muito lento. Neste sentido, como resultado desta pesquisa de mestrado, foi proposto um método para a extração de relacionamentos semânticos do tipo "causa e efeito" em artigos científicos do domínio biomédico. Mais especificamente, o objetivo deste trabalho é propor e implementar uma solução para (1) extrair termos do domínio biomédico de documentos científicos (genes, componentes químicos, proteínas, estruturas e processos anatômicos, componentes e estruturas celulares e tratamentos), (2) identificar relacionamentos existentes nos textos, com base nos termos extraídos, e (3) sugerir uma rede de conhecimento baseada nos relacionamentos extraídos. Através de uma abordagem utilizando regras e padrões textuais, o método proposto extraiu relacionamentos semânticos com uma precisão de 94,83 %, cobertura de 98,10 % e Medida-F de 96,43 %.

Palavras-chave: Extração de Informação, Mineração de Textos, Relacionamentos Semânticos, Redes Semânticas, Domínio Biomédico, Anemia Falciforme

ABSTRACT

Recently, there is an enormous amount of scientific material written in textual format and published in electronic ways (paper on proceedings and articles on journals). In the biomedical field, researchers need to analyse a vast amount of information in order to update their knowledges, in order to get more precise diagnostics and propose more modern and effective treatments. The task of getting knowledge is extremely onerous and the manual process to annotate relationships and to propose novel hypothesis for treatments becomes very slow and error-prone. In this sense, as a result of this master's research it is proposed a method to extract "cause and effect" semantic relationships in sentences of scientific papers of the biomedical domain. The goal of this work is to propose and implements a solution for: (1) to extract terms from the biomedical domain (genes, proteins, chemical components, structures and anatomical processes, cell components and strutures, and treatmens), (2) to identify existing relationships on the texts, from the extracted terms, and (3) to suggest a knowledge network based on the relations of "cause and effect". Over the approach using textual patterns, our proposed method had extracted semantic relations with a precision of 94,83 %, recall of 98,10 %, F-measure of 96,43 %.

Keywords: Information Extraction, Text Mining, Semantic Relations, Semantic Networks, Biomedical Domain, Sickle Cell Anemia

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AF – Anemia Falciforme
ARS – Anotador de Relações Semânticas
ATR – *Automatic Term Recognition*
BD – Banco de Dados
CRF – *Condicional Random Fields*
EER – *Enhanced Entity-Relationship*
EI – Extração de Informação
F-Logic – *Frame Logic*
Fe – Ferro
GBD – Grupo de Banco de Dados
HbS – Hemoglobina S
HTML – *Hypertext Markup Language*
IA – Inteligência Artificial
JSON – *JavaScript Object Notation*
MD – Mineração de Dados
MT – Mineração de Textos
NER – *Named Entity Recognition*
PDF – *Portable Document Format*
PLN – Processamento de Língua Natural
POS – *Part-of-Speech* (Etiquetador Gramatical)
PS – PolySearch
RAT – Reconhecimento Automático de Termo
RDF – *Resource Description Framework*
REN – Reconhecimento de Entidades Nomeadas
RS – Redes Semânticas
SCA – *Sickle Cell Anemia*
SGBD – Sistema Gerenciador de Banco de Dados
SN – *Semantic Network*

SNePS – *Semantic Network Processing System*
SVM – *Suport Vector Machines*
TXT – Texto puro, sem formatações
UML – *Unified Modeling Language*
XML – *Extensible Markup Language*

SUMÁRIO

Capítulo 1

INTRODUÇÃO

*Este capítulo apresenta o **contexto** em que este trabalho está inserido e a **motivação** que deu origem a esta pesquisa de mestrado. Em seguida são discutidos os **objetivos** a serem alcançados, as **hipóteses** as serem validadas e finaliza-se com a descrição da **organização** desta monografia.*

1.1 Contexto do Trabalho

A Anemia Falciforme (AF) é uma doença genética recessiva e hereditária, caracterizada pela produção de hemoglobina anormal, denominada hemoglobina S (HbS) (SILVA_PINTO et al., 2009) (SILVA_PINTO, 2011). Hemoglobina é a proteína presente no interior do citoplasma das células do sangue, denominadas hemácias, responsável pela coloração vermelha do sangue. Como sua molécula contém ferro, este elemento permite a ligação e transporte de moléculas de oxigênio por todo o sistema circulatório.

A AF é caracterizada também pela anemia crônica e episódios de dor severa. Estes sintomas são consequência de uma alteração estrutural específica na molécula de hemoglobina S. A HbS possui um formato de foice, impedindo que a mesma circule livremente pelos capilares, podendo haver interrupção de fluxo sanguíneo e morte tissular.

Em portadores heterozigotos (aqueles que recebem o gene de um dos parentes) a doença se manifesta como forma de proteção. Eles levam uma vida normal embora suas hemácias tenham uma meia-vida mais curta. Em regiões endêmicas de malária, por exemplo, é encontrada alta incidência de indivíduos com AF. Geralmente, esses indivíduos são mais resistentes à malária do que as pessoas que não possuem a doença. Isso ocorre, porque os protozoários causadores da malária se reproduzem necessariamente no interior das hemácias humanas e as células modificadas do indivíduo falciforme não são adequadas a esse tipo de função.

A Figura 1 a) apresenta um exemplo com hemácias sadias em um indivíduo normal e a Figura 1 b) mostra um exemplo de como são as hemácias HbS, em formato de foice.

1.2 Motivação

A quantidade de documentos científicos diferentes (por exemplo, de artigos publicados em anais de eventos e em periódicos) escritos no formato textual e publicados em meios eletrônicos aumenta a cada dia em todas as áreas de pesquisa, gerando um enorme volume de dados. Com isso, a capacidade

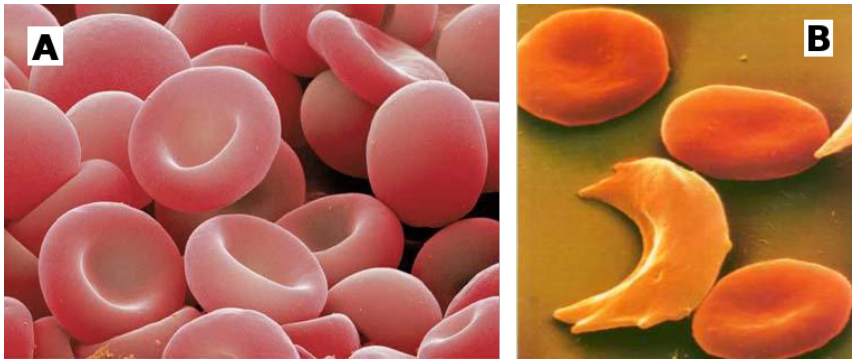

Figura 1: a) Hemácia de um indivíduo normal. b) Hemácia de um indivíduo falciforme.

dos seres humanos de absorver, processar e assimilar todo este conhecimento produzido fica cada vez mais limitado (JENSEN; SARIC; BORK, 2006; STAVRIANOU; ANDRITSOS; NICOLOYANNIS, 2007; LUO, 2008). Na área de medicina e biomedicina este fato não é diferente. Os profissionais necessitam investigar uma quantidade bastante grande de textos para conseguirem realizarem diagnósticos mais precisos, além de identificarem novos tratamentos mais modernos e eficazes e, muitas vezes, aplicados somente em outras doenças.

Durante a tarefa de descoberta de novos conhecimentos, os especialistas necessitam absorver todo o conteúdo de um conjunto de artigos científicos e ainda tentar relacionar as informações obtidas. Existem determinados relacionamentos que o cérebro humano não possui capacidade de reconhecer com facilidade e em um curto prazo, pois não estão associados de forma direta e clara. Um exemplo são os relacionamentos do tipo “causa e efeito” presentes em sentenças distintas de um artigo, que não são facilmente e rapidamente identificados pelos seres humanos. Para melhorar essa situação faz-se necessária a utilização de técnicas e ferramentas computacionais para analisar, processar e facilitar a absorção das informações presentes nos documentos de forma mais ágil.

No projeto de pesquisa sobre a Anemia Falciforme (SCA - *Sickle Cell Anemia*), do Grupo de Banco de Dados (GBD), da Universidade Federal de São Carlos, estudos voltados para a solução dessas necessidades já estão sendo realizados. Foram desenvolvidas algumas soluções para a extração de termos que compreendem efeitos negativos da doença (i.e. complicações), efeitos negativos do tratamento (i.e. efeitos colaterais), efeitos positivos do tratamento (i.e., benefícios), tratamentos e informações sobre os pacientes envolvidos nos estudos (e.g. número de pacientes, idade e sexo dos pacientes).

Em estudos subsequentes a serem desenvolvidos, além da extração de outras categorias de termos, tais como genes, componentes químicos, proteínas, estruturas e processos anatômicos, componentes e estruturas celulares e tratamentos, existe a necessidade da aplicação de um processo semântico, que por sua vez, permita associar e relacionar conjuntos de termos para que novos conhecimentos possam ser identificados.

1.3 Objetivo

O objetivo desta pesquisa de mestrado é propor um método computacional para solucionar as seguintes necessidades:

- Identificar, extrair e disponibilizar informações úteis, como termos de tratamentos, genes, componentes químicos, proteínas, estruturas e processos anatômicos, componentes e estruturas celulares, a partir de textos científicos, aos profissionais das áreas biomédicas.
- Relacionar semanticamente conjuntos de termos que possuam relações de associação do tipo "causa e efeito".
- Construir uma rede de conhecimento, por meio dos termos extraídos e relacionados, para facilitar o trabalho de análise dos artigos e descoberta de novos tratamentos e diagnósticos médicos pelo especialista do domínio.

Para a obtenção de informações são utilizadas técnicas de Processamento de Língua Natural (PLN), Mineração de Textos (MT) e Extração de Informação (EI). Para organizar o relacionamento entre as informações extraídas será utilizada uma representação por meio do conceito de Rede Semântica (RS).

Os estudos das necessidades indicadas acima são de grande importância para a comunidade médica que trabalha com o tratamento da doença Anemia Falciforme. Além disso, o desenvolvimento desta solução pode ser estendido e utilizado por médicos e especialistas para outros tipos de doenças. No âmbito computacional, este trabalho beneficia os estudos da comunidade de PLN, avançando o estado da arte em extração de informação e relacionamentos semânticos.

Um exemplo fictício do estudo citado no parágrafo anterior consiste em: Ao realizar a extração de termos em determinado conjunto de artigos A_i $(i = 1, 2, 3, \ldots)$, identificou-se em um artigo A_1 que a complicação X está diretamente relacionada ao aumento da produção da proteína P. Em outro artigo A_5, identificou-se que o medicamento M reduz a produção da proteína P. Portanto, podemos sugerir ao especialista do domínio, que pode existir uma relação de "causa e efeito" entre o medicamento M, a proteína P e a complicação C. A partir do pressuposto, o especialista poderá realizar testes que comprovem ou rejeitem tal afirmação.

1.4 Hipóteses

A partir do objetivo identificado na Seção 1.3 e de um estudo piloto realizado para entendimento do problema e que será apresentado na Seção 4.1, foram levantadas duas hipóteses a respeito da extração de termos relevantes, extração de relacionamento do tipo "causa e efeito" e construção de uma rede de conhecimento.

As hipóteses propostas são:

H1. É possível identificar relações semânticas entre termos do domínio biomédico em um mesmo artigo científico, utilizando técnicas de PLN e Redes Semânticas.

H2. É possível, a partir das relações semânticas R_1 $(r1 \rightarrow r2 \rightarrow rn)$ e R_2 $(rn \rightarrow rn+1 \rightarrow rk)$ obtidas de artigos científicos distintos, compor uma cadeia de relações semânticas do tipo R $(r1 \rightarrow r2 \rightarrow rn \rightarrow rn+1 \rightarrow rk)$ e, consequentemente, construir uma rede de conhecimentos.

H3. É possível extrair relações ternárias *cause-effect*(termo1 , termo2 , termo 3), indicando termos que possuem relação de causalidade com dois outros termos na mesma sentença. Por exemplo $<$*endothelial dysfunction*$>$ *and* $<$*end-organ disease*$>$ $\rightarrow$ $<$*renal dysfunction*$>$.

A validação destas hipóteses é mostrada na Seção 5.5.

1.5 Organização da Monografia

O conteúdo desta dissertação de mestrado está organizado dentro dos seguintes capítulos:

- **Capítulo 2 - Fundamentação Teórica**: São descritos os conceitos teóricos essenciais utilizados neste trabalho de pesquisa, como Mineração de Textos, Extração Automática e Redes Semânticas.
- **Capítulo 3 - Revisão da Literatura**: Estudo específico dos trabalhos já realizados no projeto SCA Anemia Falciforme, bem como dos trabalhos correlatos que serão utilizados para comparação da solução proposta.
- **Capítulo 4 - Método Proposto**: Apresenta o método proposto para extração de relacionamentos semânticos do tipo "causa e efeito" em artigos científicos do domínio biomédico. Descreve o estudo piloto realizado para o entendimento do problema, os conceitos utilizados no decorrer do trabalho, ferramentas utilizadas e desenvolvidas, os recursos utilizados e detalha cada etapa de execução do método.
- **Capítulo 5 - Validação e Testes**: São descritos os métodos de validação utilizados e os testes realizados, bem como os resultados obtidos e a validação das hipóteses levantadas na Seção 1.4.
- **Capítulo 6 - Conclusões**: Apresenta as conclusões sobre o trabalho desenvolvido, destaca as principais contribuições e possíveis trabalhos futuros que poderão incrementar este projeto.
- **Referências**: Encerra a monografia com a bibliografia consultada e referenciada neste documento.
- **Apêndice A: Estrutura JSON para codificação de sentenças**: Apresenta a estrutura em formato JSON construída para codificar as sentenças anotadas.
- **Apêndice B: Telas da Ferramenta JPdf2JSON**: Apresenta as telas principais da ferramenta JPdf2JSON desenvolvida para conversão de documentos PDF para TXT e JSON.
- **Apêndice C: Tela Principal da Ferramenta ARS**: Apresenta a tela principal da ferramenta ARS desenvolvida para anotação e a extração de termos e relacionamentos semânticos.
- **Apêndice D: *Script* Gerador de Resultados - Seleção de sentençaas**: Apresenta detalhes sobre o gerador de resultados numéricos para a fase de seleção de sentenças.
- **Apêndice E: *Script* Gerador de Resultados - Extração de Relacionamentos**: Apresenta detalhes sobre o gerador de resultados numéricos para a fase de extração de relacionamentos semânticos.
- **Apêndice F: Estrutura e Categorias do Dicionário de Termos do Domínio**: Apresenta a estrutura e as categorias que envolvem o Dicionário de Termos do Domínio.
- **Apêndice G: Estrutura e Significados do Dicionário de *Tip Words***: Apresenta a estrutura e os significados que envolvem o Dicionário de *Tip Words*.
- **Apêndice H: Estrutura do Dicionário *Blacklist***: Apresenta a estrutura que envolve o Dicionário *Blacklist*.
- **Apêndice I: MetaRegras**: Apresenta a estrutura que envolve a MetaRegra de Associação e a MetaRegra *Increase/Decrease*.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

*Este capítulo descreve os principais **conceitos** utilizados no desenvolvimento desta dissertação de mestrado. Entre os conceitos apresentados estão a Mineração de Textos, Processamento de Língua Natural, Extração de Informação, Mineração de Dados, Extração Automática, Reconhecimento Automático de Termo, Extração de Relacionamentos Semânticos, Abordagens para Extração de Informação e Redes Semânticas.*

2.1 Mineração de Textos

Considerada uma extensão da Mineração de Dados (MD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), a Mineração de Textos (MT) também é conhecida por Mineração de Dados Textuais (HEARST; HALL, 1999; REZENDE; MARCACINI; MOURA, 2011) ou Descoberta de Conhecimentos Textuais (TAN, 1999). Trata-se do processo de extração de informação relevante a partir de documentos no formato textual não estruturado por meio da identificação de padrões e, principalmente, da identificação de conhecimentos existentes no texto (FELDMAN; DAGAN, 1995). Já a Mineração de Dados busca extrair conhecimentos a partir de bancos de dados estruturados em um domínio específico, aplicando algoritmos de descoberta de padrões a partir do cruzamento de informações.

Na biologia e biomedicina, a MT compreende uma subárea do Processamento de Língua Natural (PLN), mais especificamente de PLN biomédico (ANANIADOU; MCNAUGHT, 2006). O principal objetivo é auxiliar especialistas a absorver grande quantidade de informação por meio da extração desta informação, recuperação de informação e descoberta de relacionamentos implícitos (SPASIC et al., 2005).

Este capítulo está dividido da seguinte forma: na Seção 2.1.1, são destacadas as áreas de conhecimento que estão envolvidas com a Mineração de Textos. Na Seção 2.1.2 são apresentados os conceitos e as etapas do processo de mineração. Em seguida, na Seção 2.2, são apresentadas as considerações finais.

2.1.1 Áreas de Conhecimento relacionadas à Mineração de Textos

A Mineração de Textos é uma área multidisciplinar que envolve Extração de Informação, Recuperação de Informação, Aprendizado de Máquina, Estatística, Processamento de Língua Natural e Mineração de Dados (HOTHO et al., 2005). Caracteriza-se, portanto, pelo uso de métodos automatizados para explorar uma enorme quantidade de conhecimento.

A seguir são apresentadas sucintamente algumas das áreas que fazem parte do desenvolvimento deste trabalho e que contribuem com a mineração textual: Processamento de Língua Natural, Extração de Informação e Mineração de Dados.

2.1.1.1 Processamento de Língua Natural

O Processamento de Língua Natural (PLN) ou Linguística Computacional é uma subárea da Inteligência Artificial e da Linguística. Tem como propósito desenvolver métodos e algoritmos computacionais que permitam compreender a língua natural, falada ou escrita, da mesma forma que um ser humano (JACKSON; MOULINIER, 2002).

Na língua escrita, PLN busca analisar e sintetizar textos em diversos níveis linguísticos. Segundo Jurafsky e Martin (2000) pode-se dividir a análise linguística em seis categorias diferentes:

- **Fonética e Fonológica:** estudo dos sons linguísticos;
- **Morfológica:** formação e construção das palavras;
- **Sintática:** analisa a relação entre as palavras em uma sentença;
- **Semântica:** analisa os significados das palavras e das sentenças;
- **Pragmática:** compreensão do uso da língua utilizando conhecimentos do mundo;
- **Discursiva:** interpretando a estrutura e o significado de um texto por completo.

No trabalho proposto abordaremos o PLN em língua escrita, nos níveis morfológico e semântico.

Em seguida, na Seção 2.1.1.2, introduzimos os conceitos sobre **extração de informação** em Mineração de Textos.

2.1.1.2 Extração de Informação

Técnicas de Extração de Informação (EI) são aquelas que permitem a localização e posterior extração de trechos de sentenças de um textos ou de elementos específicos de textos não estruturados e o seu armazenamento em formato estruturado.

Tais técnicas podem fazer parte da tarefa de Mineração de Textos, facilitando a extração de conhecimentos (FELDMAN; SANGER, 2007). Os resultados da EI, informações estruturadas, normalmente são armazenadas em um banco de dados para que, posteriormente, possam ser utilizados por algoritmos de Mineração de Dados para identificar padrões interessantes.

Na área biomédica a tarefa básica mais utilizada para extração de informação é o Reconhecimento de Entidades Nomeadas (REN), por isso, ela será mais bem detalhada na Seção 2.2 - Extração Automática.

Neste trabalho será utilizada a técnica de Extração de Informação, por meio de tarefas de REN, para identificação de termos relacionados com a área de medicina, como proteínas, efeitos negativos da doença, tratamentos da doença, estruturas celulares, entre outros.

A seguir, abordaremos sucintamente os conceitos sobre **Mineração de Dados**.

2.1.1.3 Mineração de Dados

Mineração de Dados (MD), é definida como o processo de identificar padrões consistentes (como regras de associação e sequências temporais) por meio da exploração de grandes quantidades de dados estruturados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; CLIFTON, 2010).

MD faz parte do processo de Descoberta de Conhecimento em Banco de Dados, do inglês Knowledge Discovery in Databases (KDD). O processo de KDD consiste nos seguintes passos: seleção dos dados, pré-processamento (limpeza dos dados), transformação dos dados, busca por padrão (mineração de dados) e interpretação e avaliação dos resultados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Durante o desenvolvimento deste trabalho a técnica de Mineração de Dados será utilizada na identificação de padrões textuais para identificar e selecionar sentenças que possuem relacionamentos do tipo “causa e efeito”.

Em seguida é apresentado o **Processo de Mineração de Textos**, base para implementações de sistemas MT.

2.1.2 Processo de Mineração de Textos

Para se implementar sistemas que realizam atividades de Mineração de Textos, primeiramente é necessário conhecer todo o processo que a envolve. O processo de MT é formado pelos passos apresentados na Figura 2 (ARANHA, 2007):

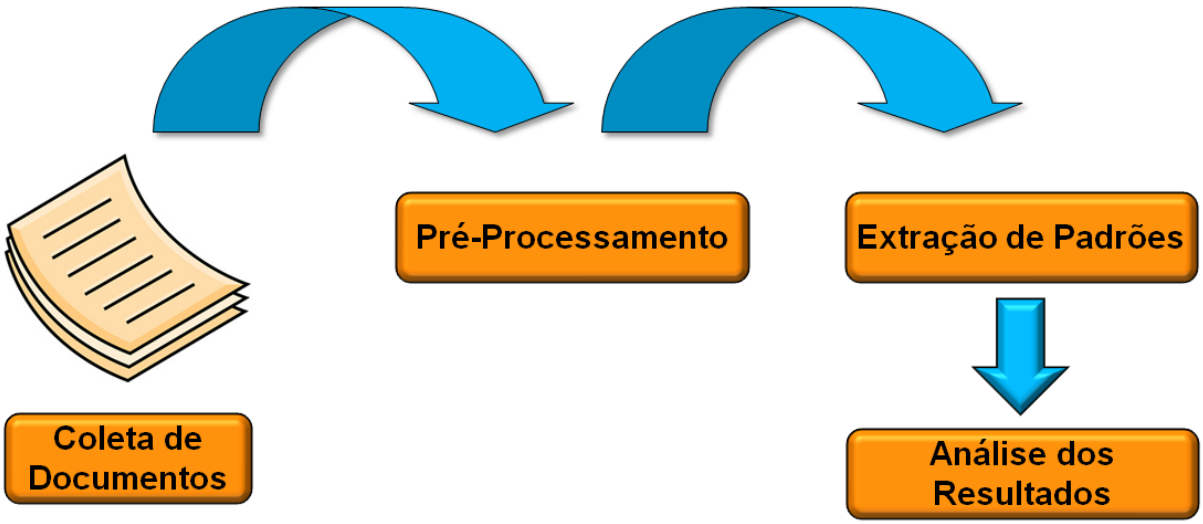

Figura 2: Etapas do processo de Mineração de Textos (ARANHA, 2007).

2.1.2.1 Coleta de Documentos

Na etapa inicial, “COLETA”, é realizada a captação de toda a base de documentos e textos a ser trabalhada nas etapas seguintes (ARANHA; PASSOS, 2006)(MATOS, 2010). Na área de linguística esta base é chamada de *corpus*.

Existem vários locais onde os textos podem ser encontrados. Estes locais podem ser bibliotecas de documentos impressos ou mídias digitais, computador em arquivos armazenados em discos rígidos e, em geral, na Internet. Neste último, podem ser encontrados diversos repositórios espalhados em servidores por todo o mundo. No domínio biomédico podemos obter documentos a partir de repositórios importantes como o PubMed, que contém mais de 21 milhões de artigos científicos.

Para auxiliar a coleta de dados nos repositórios e ambientes de armazenamento na Internet, estão disponíveis máquinas de busca (*search engines*) que indexam o conteúdo existente na *Web*, a partir de palavras-chave, e disponibiliza aos usuários. A mais conhecida delas, nos dias de hoje, é a máquina de busca do Google. Na área científica, podemos citar máquinas de busca como a *Scopus*, *Web of Science*, máquina de busca de periódicos da Capes, entre outras (MATOS, 2010).

2.1.2.2 Pré-Processamento

A etapa seguinte é denominada "PRÉ-PROCESSAMENTO", sendo considerada a fase mais onerosa do processo. É responsável pela estruturação dos documentos coletados para serem submetidos aos algoritmos de mineração de dados (FELDMAN; SANGER, 2007)(ARANHA, 2007).

Para que seja possível processar textos automaticamente, várias técnicas podem ser aplicadas durante a etapa de pré-processamento (SPASIC et al., 2005). É aconselhável forte análise sobre o texto para avaliar quais devem ser utilizadas.

Algumas das técnicas que podem ser aplicadas nesta fase são:

- **Tokenização:** Dividir o texto em unidades básicas conhecidas como *tokens*, utilizando delimitadores (espaço em branco ou pontuação);
- **Lematização:** Substitui uma palavra flexionada pela forma eliminando número e gênero (ex.: cantaremos $\Rightarrow$ cantar);
- ***Stemming*:** Reduz a palavra em seu radical (ex.: cantaremos $\Rightarrow$ cant);
- **Remoção de *Stopwords*:** Filtrar palavras ou tokens que aparecem repetidamente e acabam por perder sua utilidade dentro de uma análise textual ;
- **Etiquetador gramatical (*Part-Of-Speech* - POS):** identifica as classes gramaticais de cada palavra do texto, podendo ser em nível morfológico (substantivo, adjetivo, artigo) ou morfossintático (sujeito, predicado, aposto).

2.1.2.3 Extração de Padrões

A etapa de "EXTRAÇÃO DE PADRÕES" visa à aplicação de técnicas para extração de conhecimentos úteis. Para isso, são utilizadas combinações de algoritmos e técnicas de Mineração de Dados provenientes de diversas áreas do conhecimento (ARANHA, 2007), tais como: aprendizado de máquina, estatística e bancos de dados.

Segundo Camilo e Silva (2009), as principais tarefas de Mineração de Dados que podem ser aplicadas nesta etapa, são:

- **Classificação (*Classification*):** É uma das tarefas mais comuns. Visa identificar a qual classe um determinado registro pertence. Normalmente utilizam-se técnicas de aprendizado de máquina

supervisionado. Por exemplo, no domínio do projeto SCA, podemos classificar sentenças de um texto, como sendo de classes como “Sentenças de Tratamento”, “Sentenças de Complicação” ou “Sentenças Gerais”.

- **Regressão (*Regression*):** Similar à classificação, a regressão é usada quando o registro é identificado por um valor numérico e não por uma categoria. Assim, é possível estimar o valor de uma determinada variável analisando-se os valores das demais variáveis. Por exemplo, estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas.
- **Agrupamento (*Clustering*):** Essa tarefa visa identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados. Por exemplo, em uma auditoria, poderiam ser separados os comportamentos suspeitos.
- **Associação (*Association*):** A tarefa de associação consiste em identificar quais atributos estão relacionados. Por exemplo, identificar quais produtos são levados em conjunto pelos consumidores de um supermercado, ou seja, se um consumidor levou um produto a, será que ele leva b ou c ?

2.1.2.4 Análise dos Resultados

A última etapa corresponde à “ANÁLISE DOS RESULTADOS”. Nesta fase é necessário contar com o fator humano, preferencialmente um especialista no domínio, para avaliar se os resultados obtidos estão de acordo com algumas métricas. Por fim, o usuário é o responsável por julgar a aplicabilidade destes resultados.

Além disso, no processo de avaliação dos resultados são utilizadas algumas métricas. Elas provêm informações importantes que permitem uma avaliação dos dados gerados e, também, da qualidade de cada uma das etapas individualmente.

As principais métricas existentes para análise dos dados extraídos pelos processos de mineração de textos são: **Precisão**, **Cobertura** (ou Revocação), **Medida-F**.

A precisão e a cobertura são métricas padrão para qualidade para algoritmos de Recuperação de Informação (RI). A **precisão** é uma medida de fidelidade. Ela nos informa a taxa em que o algoritmo conseguiu recuperar dos dados, ou seja, é o número de elementos relevantes recuperados (também chamados Verdadeiros Positivos ou VP), dividido pelo número total de elementos recuperados. O número total de elementos recuperados corresponde à soma do número de elementos relevantes recuperados (VP) com o número de elementos recuperados que não são relevantes (Falsos Positivos ou FP):

$$Precisão = \frac{Número\ de\ elementos\ relevantes\ recuperados}{Número\ total\ de\ elementos\ recuperados}$$

ou

$$Precisão = \frac{VP}{VP + FP}$$

A cobertura (ou revocação) nos informa a taxa de acertos, tomados os valores já recuperados, ou seja, é definida como o número de elementos relevantes recuperados (Verdadeiros Positivos ou VP), dividido pelo número total de elementos relevantes existentes (que deveriam ter sido recuperados). O número total de elementos relevantes existentes corresponde à soma entre o número de elementos relevantes recuperados (VP) e o número de elementos relevantes que não foram recuperados (também chamados Falsos Negativos ou FN).

$$Cobertura = \frac{Número\ de\ elementos\ relevantes\ recuperados}{Número\ total\ de\ elementos\ relevantes}$$

ou

$$Cobertura == \frac{VP}{VP + FN}$$

Em RI, um valor para precisão e cobertura perfeita é sempre 1,0, equivalente a 100%. No caso da precisão significa que cada resultado obtido por uma pesquisa foi relevante. Para a cobertura significa que todos os elementos relevantes foram recuperados pela pesquisa.

A **Medida F** (*F Measure*) é média harmônica ponderada da precisão e cobertura. F_β mede a eficácia da recuperação em relação ao valor atribuído a Beta (β). Pesos comumente utilizados para β são: F_2 (cobertura é o dobro da precisão) e $F_{0,5}$ (precisão é o dobro de cobertura). A precisão tem peso maior para valores $\beta\ <\ 1$, enquanto que $\beta\ >\ 1$ favorece a cobertura.

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

Medida F Balanceada:

$$Medida\ F = \frac{2\ \times\ P\ \times\ R}{P\ +\ R}$$

Em seguida, serão apresentados os conceitos que envolvem **redes semânticas**.

2.2 Extração Automática

Nesta seção são abordados os conceitos, técnicas e métodos existentes na literatura para extração de termos e para extração de relações semânticas.

2.2.1 Reconhecimento Automático de Termo

Segundo o autor Roberto et al. (2011), a tarefa de Reconhecimento de Entidades Nomeadas (REN), do inglês *Named Entity Recognition* (NER), consiste em identificar e classificar elementos de um texto em categorias pré-definidas. Por exemplo, pessoa, organização, lugar, quantidade, dentre outras.

NER é uma das áreas de extração de informação mais estudadas e a tarefa mais utilizada em biomedicina (ANANIADOU; MCNAUGHT, 2006; PARK; JUNG-JAE, 2006).

Paralelamente, temos o Reconhecimento Automático de Termo (RAT), do inglês *Automatic Term Recognition* (ATR), que consiste na aplicação de técnicas de REN, associando técnicas de extração e classificação para identificar e extrair termos técnicos de um domínio específico dentro de um *corpus* de trabalho (SEKINE, 2004; ANANIADOU; MCNAUGHT, 2006; PARK; JUNG-JAE, 2006). No contexto biomédico, essas entidades são genes, proteínas, nomes de doenças, complicações, tratamentos, entre outros.

Na área biomédica, devido à quantidade de neologismos, sinônimos (i.e., conceito representado usando-se vários termos), homônimos (i.e., termos com vários significados) e ambiguidades, normalmente considerados como obstáculos na identificação dos termos, existe a necessidade do reconhecimento automático de termos. Tais termos representam conceitos de domínio utilizados pela comunidade científica. Sem a identicação prévia desses termos, torna-se impossível a compreensão ou extração de informações de um artigo científico (ANANIADOU; FREIDMAN; TSUJII, 2004).

Segundo Krauthammer e Nenadic (2004), o processo de identificação e extração de termos utilizando RAT pode ser dividido em três etapas, como mostra a Figura 3.

Figura 3: Etapas para identificação de termos. (KRAUTHAMMER; NENADIC, 2004).

O **Reconhecimento de Termo** diferencia os termos dos não termos, a **Classificação de Termo**, ou categorização de termo, classifica os termos reconhecidos em classes do domínio e o **Mapeamento de Termo** associa termos com conceitos bem definidos representados por vocabulários, dicionários, ontologias ou bases de dados. Cada passo não possui a obrigatoriedade de ser executado separadamente. Dependendo da necessidade podem ser implementados conjuntamente.

Neste trabalho o foco principal é o reconhecimento de relacionamentos semânticos, portanto o reconhecimento automático de termos foi utilizado como base para a extração das relações.

Em seguida serão apresentados os conceitos sobre **Extração de Relacionamentos Semânticos**.

2.2.2 Extração de Relacionamentos Semânticos

Um relacionamento semântico é todo tipo de associação que existe entre termos de um texto no nível de seus significados. Existem muitos tipos de relacionamentos semânticos, sendo que as mais frequentemente consideradas são aquelas apresentadas na Tabela 1.

Em seguida, iremos apresentar algumas abordagens existentes para extrair termos e relacionamentos semânticos.

2.2.3 Abordagens para Extração de Informação

Ainda hoje não existe uma definição exata das abordagens que podem ser utilizadas para extração de informação.

Tabela 1: Relacionamentos Semânticos mais comuns

	Relacionamento Semântico	**Sentença exemplo**	**Relacionamento extraído**
1.	hiponímia/hiperonímia: *is-a*(subclasse, superclasse)	Maçã é uma fruta	*is-a*(maçã, fruta)
2.	meronímia/holonímia: *part-of*(todo, parte)	Parafuso é uma parte de uma máquina	*part-of*(máquina, parafuso)
3.	causa e efeito: *effect-of*(ação/estado, consequência)	Gripe causa febre	*effect-of*(gripe, febre)

Tanto para extração de termos quanto para extração de relacionamentos semânticos, os métodos apresentados na literatura, de maneira geral, podem ser divididos em três grupos: abordagem baseada em dicionário, abordagem baseada em regras e abordagem baseada em aprendizado de máquina. Normalmente, elas são utilizadas para EI em textos não estruturados.

Cohen e Hunter (2008) apresentam duas abordagens: baseada em regras e baseada em aprendizado de máquina. Já Krauthammer e Nenadic (2004) e Ananiadou e Nenadic (2006) definem uma outra abordagem, chamada abordagem baseada em dicionário, além das abordagens anteriores.

Nas seções seguintes serão apresentadas essas três abordagens na seguinte ordem: abordagem baseada em dicionário (Seção 2.2.4), abordagem baseada em regras (Seção 2.2.5) e abordagem baseada em aprendizado de máquina (Seção 2.2.6).

2.2.4 Abordagem Baseada em Dicionário

A abordagem baseada em dicionário utiliza informações de um dicionário para auxiliar na identificação dos termos, relacionamentos ou das entidades no texto.

O dicionário consiste em uma lista, normalmente implementada em um arquivo de texto puro, em linguagem estruturada de marcação de dados, por exemplo XML e JSON, ou ainda, em uma tabela de um banco de dados. Essa lista contém um conhecimento prévio sobre o domínio do problema.

Na identificação de termos relacionados com o domínio biomédico, por exemplo, podemos construir um dicionário com os nomes de proteínas, doenças, genes, já conhecidos. Esses termos serão localizados nos textos.

Os bancos de dados são fontes de dados consideradas como abordagens baseadas em dicionários, que armazenam informações e conceitos do domínio. Segundo Rebholz-Schuhmann, Kirsch e Couto (2005), recursos terminológicos, como ontologias, podem ajudar a relacionar as informações citadas em publicações científicas com informações armazenadas em um banco de dados.

Neste tipo de abordagem, nos deparamos sempre com a questão de que as palavras de um texto possam ter sinônimos, homônimos ou variações léxicas. Essas construções linguísticas podem ser consideradas como obstáculos no reconhecimento de termos, dificultando a identificação dos mesmos. Por isso, com o objetivo de aumentar a precisão dos resultados das extrações é necessário que os sinônimos, homônimos e as variações sejam incluídos nos dicionários.

No desenvolvimento deste trabalho, a abordagem baseada em dicionário será utilizada na extração de termos relacionados com o domínio biomédico.

2.2.5 Abordagem Baseada em Regras

Uma das abordagens mais simples e antigas na extração de informação é a abordagem baseada em regras. Também chamada de abordagem baseada em padrões textuais ou abordagem baseada em conhecimento (COHEN; HUNTER, 2008), como o próprio nome indica, as regras são conhecimentos sobre padrões de escrita normalmente utilizados e localizados em um *corpus*, que servem como "pistas" para identificar termos ou relacionamentos.

Alguns exemplos de regras de identificação e extração de informações podem ser observadas a seguir:

- Esse padrão indica uma relação de hiponímia (*is-a*) (HEARST, 1992, 1998):

$$NP_0\ such\ as\ \{NP_1,\ NP_2,\ \ldots,\ (and \mid or)\}NP_n$$

onde, NP denota um sintagma nominal (*noun phrase*), { e } representam a repetição de 0 ou mais vezes do padrão entre as chaves e $|$ indica uma opção de escolha entre valores.

Exemplo:

Fruit such as a banana
is-a(banana, fruit)

- Dois padrões que localizam relacionamentos entre determinado gene e doença. Para identificação desses relacionamentos pode-se utilizar análise linguística e semântica (COHEN; HUNTER, 2008):

$$< gene >\ plays\ a\ role\ in\ < disease >$$

$$< disease >\ is\ associated\ with\ < gene >$$

Neste trabalho, a abordagem baseada em regras será utilizada na filtragem de sentenças e extração de relacionamentos semânticos do tipo causo-efeito.

2.2.6 Abordagem Baseada em Aprendizado de Máquina

Nesta abordagem, são utilizados algoritmos de aprendizado de máquina (AM) para construir sistemas que classificam dados em grupos específicos. No caso da extração de informação, os dados podem ser termos ou relacionamentos semânticos.

As técnicas que utilizam aprendizado de máquina necessitam de dados de treinamento para que o algoritmo possa aprender. O aprendizado acontece quando os dados são submetidos comparativamente com determinadas características úteis e relevantes, em inglês são chamadas *features*.

As *features* são atributos que descrevem uma instância. Elas são utilizadas por métodos de aprendizado de máquina como forma de generalizar e discriminar instâncias. Por exemplo, algumas *features* que podem descrever uma maçã são a sua cor, tamanho, e massa. Tais características são fornecidas aos classificadores que, através de métodos probabilísticos ou estatísticos, conseguem predizer em qual classe determinada informação pode ser inserida. Um dos maiores desafios dentro da área de aprendizado de máquina é a gerar um conjunto de características que possa representar significativamente os dados que estão sendo classificados. Neste projeto de mestrado não utilizaremos esta abordagem.

Os trabalhos que compõem o estado da arte nesta área para o idioma inglês, realizam extrações automáticas de relacionamentos semânticos de diversos tipos, porém sempre focando textos jornalísticos. As principais técnicas utilizadas são as abordagems baseada em regras e abordagem por aprendizado de máquina. Na utilização de extração de relacionamentos semânticos em textos científicos biomédicos, o trabalho mais atual (CHENG et al., 2008) utiliza apenas a abordagem por dicionários.

Nesta proposta temos como objetivo a extração automática de relacionamentos semânticos do tipo causa e efeito (*effect-of*). A partir dos textos biomédicos, podemos extrair relacionamentos entre doença-proteína, proteína-proteína, gene-doença, entre outras categorias que são definidas para os termos do domínio.

Em seguida, será introduzido o conceito de **Redes Semânticas**. Elas serão utilizadas para agrupar, organizar e disponibilizar o conhecimento proveniente da extração dos relacionamentos do tipo “causa e efeito”.

2.3 Redes Semânticas

Uma Rede Semântica (RS) é um tipo de notação gráfica que usa nós e arcos interconectados. Utilizada para a representação de conhecimento ou como ferramenta de suporte a sistemas automatizados de inferências sobre o conhecimento (SOWA, 2006). A forma mais genérica deste tipo de representação é o **grafo**.

Segundo Sowa (2006) podemos classificar as redes semânticas em seis tipos mais comuns. Apesar de existir esta diferenciação, devido à complexidade de alguns sistemas, muitas vezes fica difícil dizer com exatidão em qual grupo determinada rede faz parte.

Os tipos básicos são apresentados a seguir:

2.3.1 Redes de Definição

Enfatiza o subtipo ou um relacionamento de hiponímia e hiperonímia, **é um** (***is a***, em inglês) entre o tipo de conceito e o subtipo recentemente definido.

As primeiras implementações foram utilizadas para definição de conceitos e definição de padrões de relacionamentos para máquinas de tradução. Um exemplo deste tipo de rede semântica é a árvore de Porfírio, Figura 4, que define, em diversos níveis, diferentes subtipos para um mesmo supertipo.

Outro exemplo muito conhecido na área linguística de rede de definição para o idioma inglês é a WordNet, cuja principal relação entre as palavras é o sinônimo, porém também compõem relações de

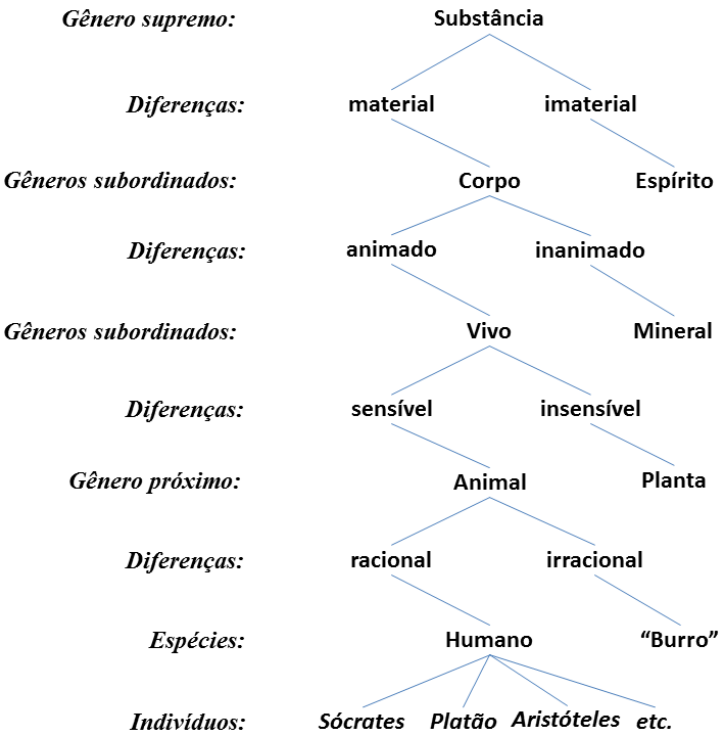

Figura 4: Árvore de Porfírio (SOWA, 2006).

hiponímia e hiperonímia (MILLER, 1995)(FELLBAUM, 1998).

A Figura 5 ilustra um pequeno trecho da WordNet. Nela podemos perceber que os nós correspondem aos sinônimos e as arestas indicam as relações.

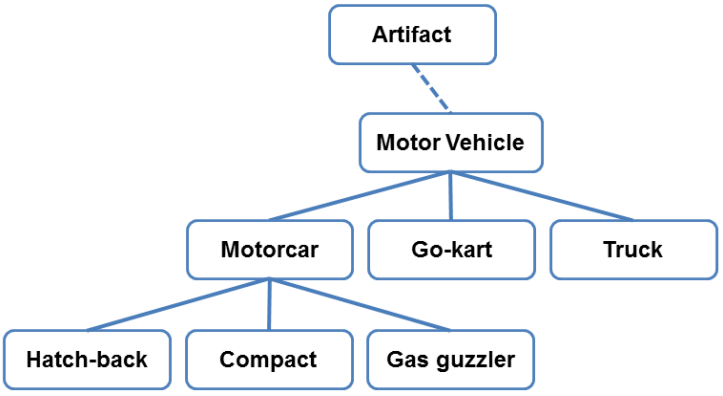

Figura 5: Fragmento da Hierarquia de Conceitos da WordNet (BIRD; KLEIN; LOPER, 2009).

2.3.2 Redes de Asserção

As redes de asserção foram desenvolvidas para a asserção de proposições lógicas. A notação gráfica, por grafos relacionais, introduzida em 1882, foi criada com base na notação utilizada pela química orgânica.

Diversas propostas foram feitas para melhorar a notação gráfica deste tipo de rede semântica, permitindo que outros operadores lógicos também pudessem ser utilizados. A primeira implementação

utilizada na área de inteligência artificial foi feita em 1971. Ela evoluiu para o chamado *Semantic Network Processing System* (SNePS).

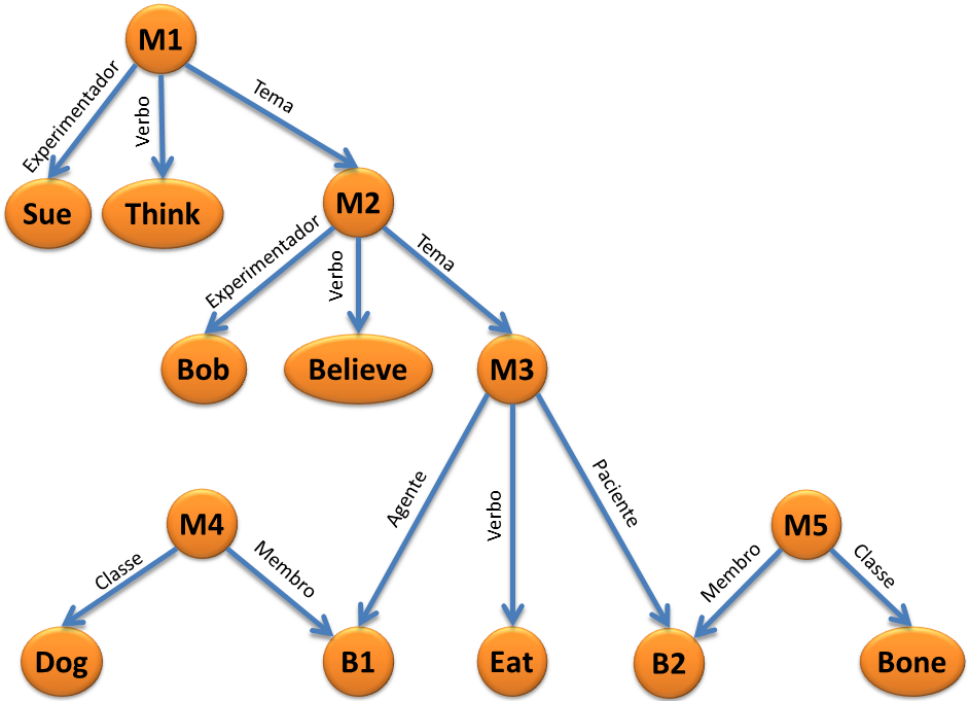

Figura 6: Proposições representadas em SNePS (SOWA, 2006).

A Figura 6 mostra um grafo SNePS que representa a sentença “Sue thinks that Bob believes that a dog is eating a bone”.

Cada um dos nós, rotulados M1 a M5, representam proposições distintas cujo conteúdo relacional está anexo ao nó proposicional.

A proposição M1 estabelece que “*Sue*” é o experimentador do verbo “*think*”, cujo tema é outra proposição M2. Para M2, o experimentador é “*Bob*”, o verbo é “*believe*” e o tema é a proposição M3. Em M3, o agente é alguma entidade B1, que é um membro da classe “*Dog*”, o verbo é “*eat*” e o paciente é uma entidade B2, que é um membro da classe “*Bone*”.

2.3.3 Redes de Implicação

Uma Rede de Implicação é um caso especial de rede semântica proposicional, cuja primeira relação é uma implicação. Elas estabelecem relações de implicação entre os nós.

Estas redes podem ser vistas como redes de crenças e redes Bayesianas, para isto seria necessário envolver valores de probabilidade nas relações de verdadeiro e falso.

Uma rede de implicação pode ser usada, conforme o exemplo da Figura 7, para representar o tipo de conhecimento necessário para inferir as possíveis causas de alguém ter escorregado na grama.

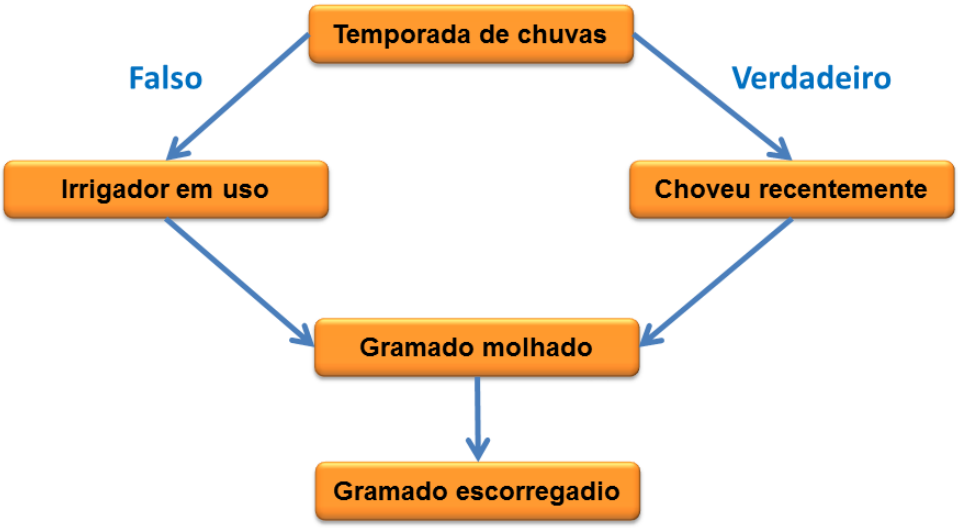

Figura 7: Rede de Implicação sobre a razão da grama molhada (SOWA, 2006).

2.3.4 Redes Executáveis

As Redes Executáveis são aquelas que possuem a capacidade de se modificar dinamicamente, devido ações de programas externos à rede.

Três tipos de mecanismos são comumente utilizados em redes executáveis:

- **Passagem de Mensagens**: São redes que podem transmitir mensagens de um nó para outro. Essas mensagens podem ser um simples *bit*, um "peso" numérico, ou uma mensagem mais comprida.
- **Procedimentos Anexos**: Pequenos programas contidos ou anexados em um nó. Esses programas realizam ações ou computam algum dado existente em determinado nó, ou em nós próximos.
- **Transformações de Grafo**: Combina, modifica, ou quebra grafos em tamanhos menores.

Os três mecanismos podem ser combinados livremente.

A Figura 8 nos mostra um exemplo das redes mais simples com procedimentos anexos. Este exemplo corresponde à assinatura da equação:

$$x = (A + B) \times S2N(C)$$

onde, S2N significa uma função que faz a conversão do tipo de dado *String* para Número.

Os chamados grafos de fluxo de dados possuem nós passivos (tipo caixas), que apenas mantém os dados, e nós ativos (tipo losangos), que recebem os dados dos nós de entrada e enviam os resultados para os nós de saída. Os rótulos das caixas indicam o tipo de dados (Número ou *String*) e os rótulos nos losangos indicam as funções ($+$, $\times$, ou conversão de valores *string* para numéricos).

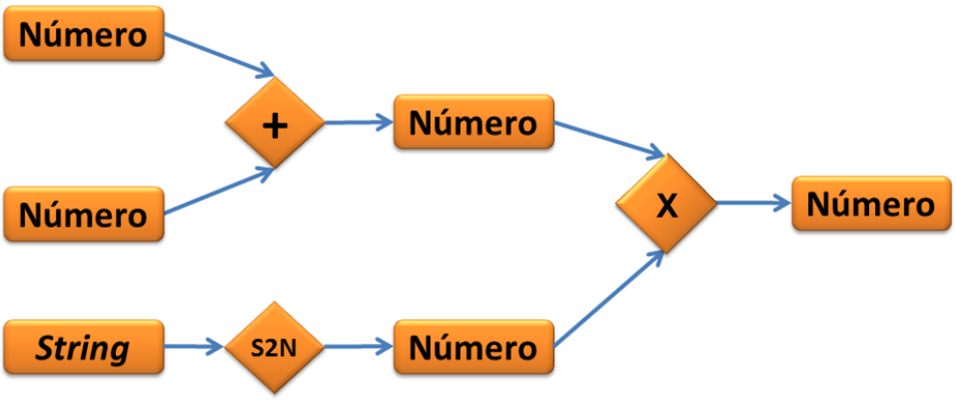

Figura 8: Grafo de Fluxo de Dados (SOWA, 2006).

2.3.5 Redes de Aprendizado

Um **Sistema de Aprendizado**, seja natural ou artificial, responde a novas informações por meio da alteração da representação interna do seu conhecimento, permitindo que o sistema passe a responder de forma mais eficiente.

Os sistemas que utilizam redes de aprendizado são baseados majoritariamente em três mecanismos para alteração de sua rede:

- ***Rote Memory***: Mais simples forma de aprendizado. Converte a nova informação em uma rede e adiciona à rede atual sem alterações adicionais.
- **Alteração de Pesos**: Realiza a alteração de valores, chamados "pesos", que a rede possui associados aos nós e aos arcos. Os pesos representam, por exemplo, probabilidades, que é incrementada a cada nova ocorrência de uma mesma rede.
- **Reestruturação**: Realiza uma reestruturação de todas as conexões entre os nós. Este mecanismo requer grande capacidade computacional.

Os sistemas que utilizam *Rote memory* são mais indicados para aplicações que necessitam da recuperação exata dos dados originais.

Já os sistemas que utilizam alteração de pesos (e.g. redes neurais), mostrados na Figura 9, são mais indicados para reconhecimento de padrões. Neste tipo de rede a estrutura de arcos e nós é fixa, as únicas mudanças são nos pesos dos arcos. A entrada é um conjunto de números que representam uma proporção relativa de alguma característica selecionada. A saída também é um conjunto de números que indica o conceito mais próximo, com relação à característica escolhida.

2.3.6 Redes Híbridas

Redes Híbridas utilizam dois ou mais tipos de redes semânticas, combinando-as em uma mesma rede, ou em redes separadas.

O exemplo mais comum, e amplamente utilizado, de rede semântica híbrida é a *Unified Modeling Language* (UML). Ela não é normalmente chamada de rede semântica, mas utiliza os conceitos de

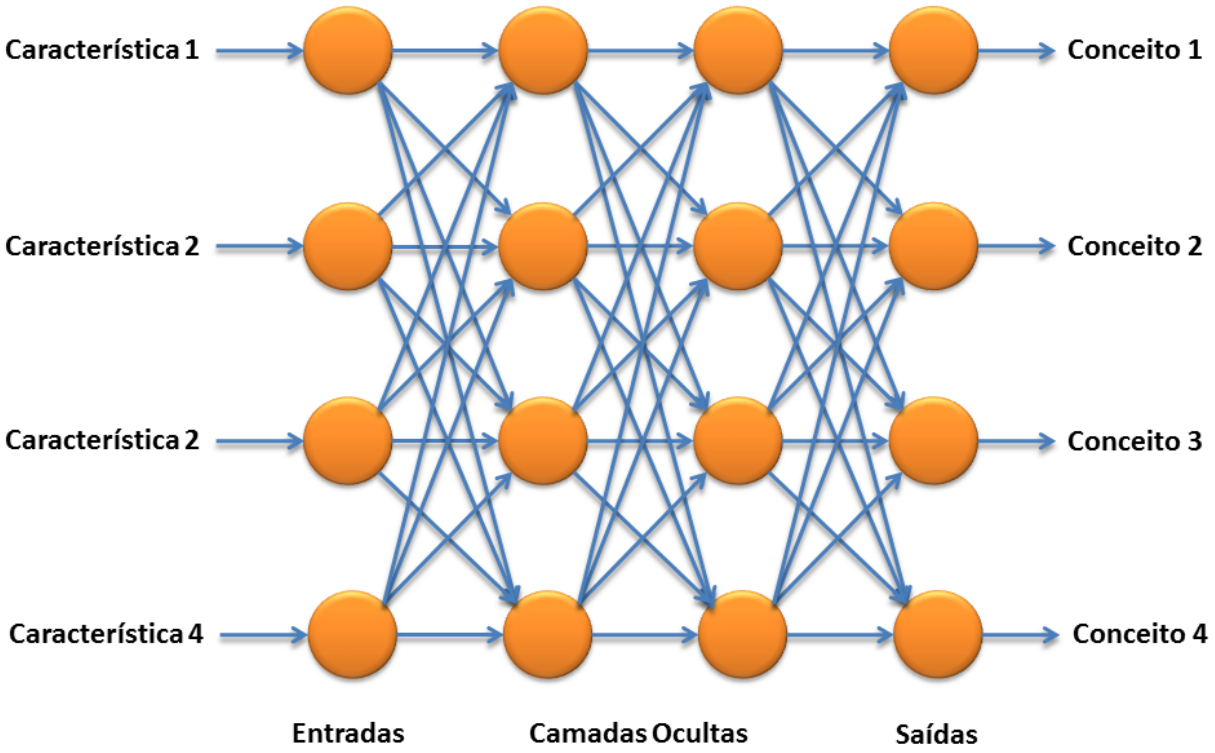

Figura 9: Uma Rede Neural (SOWA, 2006).

redes de definição para a definição de tipos de objetos. Podemos encontrar redes executáveis nas representações dos diagramas de estados e diagramas de classes.

Nesta pesquisa, será utilizada a rede de implicação para representar o conhecimento extraído dos artigos científicos por meio dos relacionamentos semânticos.

2.4 Considerações Finais

No Capítulo 2, **Fundamentação Teórica**, foram apresentados os conceitos fundamentais para o entendimento do método proposto para a extração de relacionamentos semânticos do tipo "causa e efeito" em artigos científicos do domínio biomédico.

Iniciando com o conceito de Mineração de Textos, Seção 2.1, apresentamos as áreas de conhecimento que contribuem para a atividade de MT e que utilizaremos como base para o desenvolvimento do projeto: Processamento de Língua Natural, Extração de Informação (área onde se encontra esta pesquisa) e Mineração de Dados. Mostramos ainda que existem outras áreas envolvidas mas que não farão parte deste estudo.

Na Seção 2.1.2, **Processo de Mineração de Textos**, é descrito o processo que um sistema de mineração de textos deve executar, segundo Aranha (2007). O processo possui quatro etapas que consistem em: Coleta de Dados, Pré-Processamento, Extração de Padrões e Análise dos Resultados.

Em seguida, na Seção 2.2, são apresentados os conceitos que envolvem **Extração Automática**. São apresentados também, os tipos de extração que serão utilizados no desenvolvimento desta pesquisa. São eles: **Reconhecimento Automático de Termos** e **Extração de Relacionamentos Semânticos**.

Juntamente com a apresentação dos conceitos sobre Extração Automática, são exibidas as abordagens utilizadas para extração de informação. Entre elas estão **Abordagem Baseada em Dicionário**, **Abordagem Baseada em Regras**, e por último, a **Abordagem Baseada em Aprendizado de Máquina**.

O último conceito apresentado na Seção 2.3 é o de **Redes Semânticas**. Nesta seção exploramos vários tipos de representações de conhecimento. Será utilizada uma representação de rede de implicação para representar o conhecimento extraído dos artigos científicos por meio dos relacionamentos semânticos.

No próximo capítulo, Revisão da Literatura, são apresentados os trabalhos complementares desenvolvidos no projeto SCA, além dos trabalhos correlatos a esta pesquisa de mestrado.

Capítulo 3

REVISÃO DA LITERATURA

Este capítulo descreve um conjunto de trabalhos correlatos a esta pesquisa de mestrado, separados por áreas que envolvem PLN. Mais especificamente, este capítulo descreve os trabalhos de Matos (2010) e Duque (2012), predecessores na investigação e uso de extração de termos no contexto do projeto de pesquisa SCA (Anemia Falciforme). São descritos também os trabalhos de Girju e Moldovan (2002), a ferramenta PolySearch de Cheng et al. (2008) e Taba (2013) que investigam relacionamentos semânticos em textos.

Devido ao fato deste trabalho de pesquisa ser bastante amplo e envolver diversas subáreas de PLN (extração de termos, extração de relações semânticas e rede semântica), é possível encontrar muitos trabalhos que podem ser considerados correlatos. Porém, nenhum deles foi usado para solucionar o problema apresentado nesta pesquisa, por diversos fatores. Muitos deles não são aplicáveis ao domínio biomédico, outros são aplicáveis apenas para idiomas diferentes do inglês, alguns não realizam determinadas tarefas necessárias no desenvolvimento do trabalho, muitos simplesmente não estavam disponíveis para utilização e outros não possuíam descrições suficientes para que pudessem ser reproduzidos.

Iniciando com os trabalhos que envolvem extração de termos na língua inglesa, podemos citar:

- Krauthammer et al. (2000) combinam nomes de proteínas e genes contidos em um dicionário com o BLAST.
- Ono et al. (2001), propõem um método para extrair informação de interações de proteína-proteína de resumos do MEDLINE utilizando um dicionário que contém nomes de proteínas, padrões de palavra e simples regras de POS.
- Corney et al. (2004) apresenta uma ferramenta que extrai termos diversos a partir de resumos e artigos completos, utilizando padrões textuais (regras).
- Müller, Kenny e Sternberg (2004) apresenta o *software* **Textpresso** que extrai termos do tipo genes e células. Utiliza padrões textuais (regras) e aprendizado de máquina.
- Settles (2004) apresenta o *software* **ABNER** que reconhece entidades nomeadas do tipo proteína, DNA, RNA, linhas celulares, tipos celulares.
- EGOROV (2004) com o sistema **ProtScan** que utiliza uma abordagem baseada em dicionário para identificação de nomes de proteínas da classe mamífero em resumos do MEDLINE.
- Tsuruoka e Tsujii (2004) com o objetivo reconhecer nomes de proteínas.

- Hirschman et al. (2005b), Hirschman et al. (2005a), Yeh et al. (2005) apresentam o **BioCreAtIvE** e suas tarefas. Um ambiente de avaliação crítica dos sistemas de extração de informação em biologia. Possui diversos sistemas, ferramentas e bases de dados para trabalhar com extração de informação em dados biológicos.
- Kou, Cohen e Murphy (2005) propõem um método de aprendizado denominado **Dict-HMMs** em que um dicionário é convertido para um modelo oculto de Markov (HMM) que reconhece frases do dicionário, assim como as variações destas frases.
- Schuemie et al. (2007) avaliaram algumas técnicas para aumentar a medida de cobertura na identificação de nomes de genes e proteínas, utilizando a combinação de um dicionário construído a partir de informações armazenadas em vários bancos de dados com regras para gerar variações de ortografia.
- Garten e Altman (2009) apresenta o *software* **Pharmspresso** que extrai conceitos farmacogênicos e alguns tipos de relacionamentos. Utiliza dicionários e regras.

Em um nível intermediário podemos apontar o trabalho de Swanson (1986), iniciando suposições e hipóteses sobre a extração de relacionamentos semânticos entre termos de um artigo científico e, também, entre artigos distintos.

Em Swanson e Smalheiser (1997) é apresentado um processo experimental, chamado *ARROWSMITH*, que auxilia e guia a busca por possíveis literaturas complementares. A abordagem é realizada por meio de observações humanas e sugestões.

Em seguida, apresentamos alguns trabalhos sobre extração de relacionamentos semânticos utilizando abordagem baseada em regras (padrões textuais):

- Hearst (1992) e Hearst (1998), busca extrair relações de hiponímia, também conhecida como relação *is-a* (é-um).
- Berland e Charniak (1999) utilizam o algoritmo de Hearst mas procuram por relações de meronímia (*part-of*) em um *corpus* do domínio jornalístico.
- Agichtein e Gravano (2000) descrevem o **Snowball**, um método que expande o DIPRE (um método para extrair conhecimento de milhares de documentos da *Web*). Ele faz uma etapa de reconhecimento de entidades nomeadas de forma que os padrões gerados tenham restrições semânticas quanto à classe dos termos.

Neste mesmo contexto, seguem alguns trabalhos sobre extração de relacionamentos semânticos utilizando abordagem baseada em aprendizado de máquina:

- Girju (2003) e Girju, Badulescu e Moldovan (2003) investigam, respectivamente, os relacionamentos de meronímia (*part-of*). Utilizando a *WordNet* como recurso de conhecimento semântico e árvores de decisão C4.5 como algoritmo de aprendizado e causalidade.
- Yang, Lin e Wu (2009) e Yang, Lin e Li (2010), apresentam o BioPPIExtractor e o BioPPISVMExtractor. Extraem relacionamentos proteína-proteína utilizando *Condicional Random Fields* (CRF) e *Suport Vector Machines* (SVM), auxiliando na construção de redes de relacionamentos.

Snow, Jurafsky e Ng (2005) e Yap e Baldwin (2009) realizam o aprendizado de padrões textuais por meio do uso de classificadores SVM.

Kok e Domingos (2008) apresenta um método de extração de conhecimento da *Web* e construção de uma rede, utiliando o conhecimento extraído.

A Tabela 2 mostra um resumo geral dos trabalhos que realizam extração de termos, todos para língua inglesa.

Tabela 2: Trabalhos correlatos gerais

	Tipo de dado extraído	**Técnica**
Krauthammer et al. (2000)	termos: proteínas e genes	dicionário
Ono et al. (2001)	interações de proteína-proteína	dicionário
Corney et al. (2004)	diversos	regras
Müller, Kenny e Sternberg (2004)	termos: genes e células	regras, ap. de máquina
Settles (2004)	termos: proteína, DNA, RNA, linhas celulares, tipos celulares	reconhec. entidades nomeadas
EGOROV (2004)	termos: proteínas da classe mamífero	dicionário
Tsuruoka e Tsujii (2004)	termos: proteínas	dicionário
Hirschman et al. (2005b), Hirschman et al. (2005a), Yeh et al. (2005)	dados biológicos	–
Kou, Cohen e Murphy (2005)	método de aprendizado	dicionário, modelo de Markov
Schuemie et al. (2007)	avaliaram técnicas termos: genes e proteínas	dicionário, regras
Garten e Altman (2009)	conceitos farmacogênicos e relacionamentos	dicionário, regras

A Tabela 3 mostra um resumo geral dos trabalhos que realizam extração de relacionamentos, todos para língua inglesa.

Na sequência são apresentados com maiores detalhes outros trabalhos relacionados com maior proximidade a esta pesquisa de mestrado.

3.1 (MATOS, 2010)

O trabalho de mestrado de Pablo Freire Matos (MATOS, 2010), intitulado “Metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico”, propõe um método para identificação e extração de informação sobre efeitos da doença Anemia Falciforme em textos científicos do domínio biomédico.

A área médica, mais especificamente a área clínica, possui algumas definições sobre “efeitos”, importantes para este trabalho:

Tabela 3: Trabalhos correlatos gerais

Swanson (1986)	relacionamentos	suposições, hipóteses
Swanson e Smalheiser (1997)	processo experimental	observações humanas e sugestões
Hearst (1992), Hearst (1998)	relações de hiponímia	regras
Berland e Charniak (1999)	relações de meronímia	regras
Agichtein e Gravano (2000)	reconhecimento de entidades nomeadas	regras
Girju (2003), Girju, Badulescu e Moldovan (2003)	relações de meronímia	ap. de Máquina
Yang, Lin e Wu (2009), Yang, Lin e Li (2010)	relações proteína-proteína	ap. de Máquina
Snow, Jurafsky e Ng (2005), Yap e Baldwin (2009)	aprendizado de padrões textuais	ap. de Máquina
Kok e Domingos (2008) extração de conhecimento da *Web*	construção de um rede	ap. de Máquina

- **Efeito negativo da doença, ou complicação**: São os fatores negativos que a doença causa no indivíduo. Por exemplo, a doença Anemia Falciforme causa uma complicação chamada “Hipóxia”, que significa baixo teor de oxigênio.
- **Efeito negativo do tratamento, ou efeito colateral**: São os fatores negativos que um determinado tratamento causa no indivíduo. Por exemplo, ao receber um determinado medicamento, o paciente reagiu com um problema estomacal.
- **Efeito positivo do tratamento, melhorias ou benefícios**: São os fatores positivos que um determinado tratamento causa ao indivíduo. Por exemplo, ao receber o determinado medicamento o paciente teve seu quadro de hipóxia normalizado.

O método proposto é aplicado na etapa de “Pré-Processamento” de um processo de Mineração de Textos. Ele é composto por quatro etapas bem definidas: 1. Entrada de Dados, 2. Classificação de Sentenças, 3. Identificação de Termos Relevantes e 4. Gerenciamento de Termos.

Um dos diferenciais deste trabalho é o fato de realizar a extração em todo o documento. Outros trabalhos realizavam a extração apenas nos resumos ou em locais específicos do texto.

3.1.1 Etapa 1: Entrada de Dados

Diferentemente da etapa de coleta de dados do processo de Mineração de Textos que tem como propósito obter os documentos de alguma fonte ou repositório, a etapa de entrada de dados do método proposto realiza o tratamento dos dados para que se possa fazer um processamento inicial.

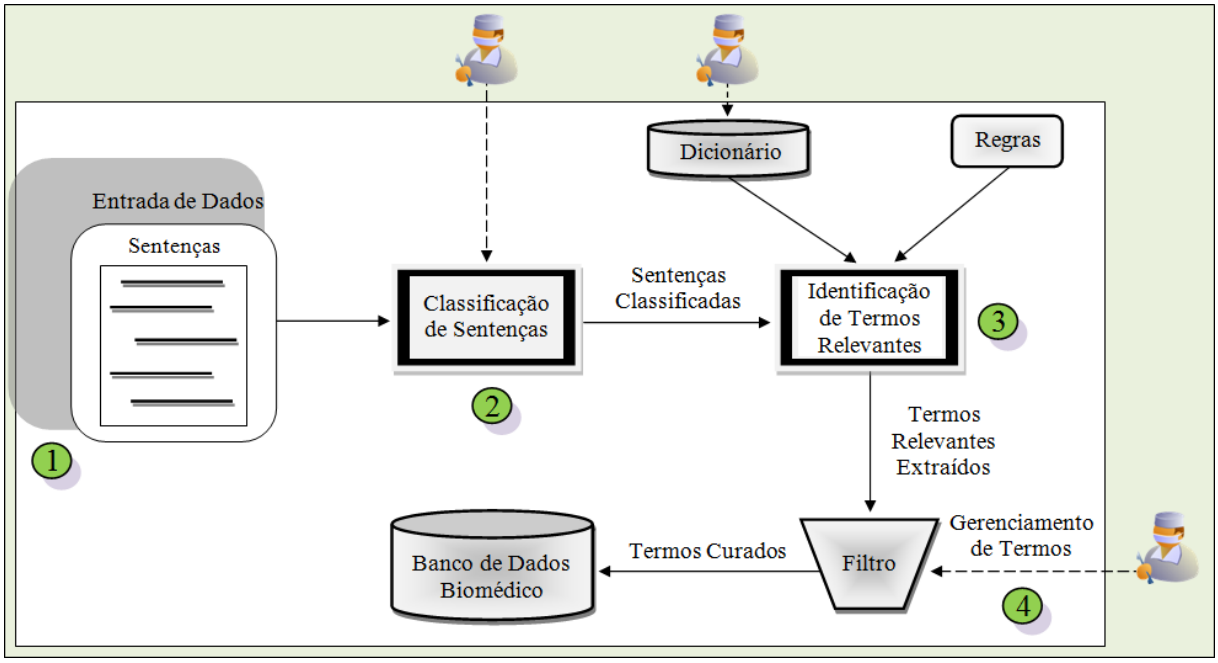

Figura 10: Método de Pré-Processamento Textual para Extração de Informações (MATOS, 2010).

Nesta etapa os especialistas do domínio fornecem ao sistema um conjunto de documentos específicos. Estes documentos são artigos científicos completos em formato PDF.

O próximo passo desta etapa é convertê-los para um formato TXT, ou um formato semiestruturado XML. O processo de conversão em TXT é totalmente manual. Na conversão para o formato XML, deve-se respeitar a hierarquia de etiquetas: seção >> página >> parágrafo >> sentença.

A Figura 11(a) apresenta o exemplo de um documento convertido para o formato XML, assumindo a hierarquia citada no parágrafo anterior. A Figura 11(b) mostra um exemplo do documento convertido em formato TXT.

3.1.2 Etapa 2: Classificação de Sentenças

A etapa de Classificação de Sentenças funciona com um filtro de sentenças. Após a conversão dos documentos, as sentenças são classificadas usando a técnica de Aprendizado de Máquina (AM) Supervisionado, ou seja, utiliza a "abordagem baseada em aprendizado de máquina". Esta técnica constrói um modelo de classificação que identifica sentenças de interesse do domínio e exclui aquelas que são irrelevantes.

A Figura 12 apresenta um exemplo de sentenças classificados por um modelo gerado pela técnica apresentada anteriormente. Neste exemplo as sentenças podem ser agrupadas em classes de domínio, como efeitos positivos e negativos.

3.1.3 Etapa 3: Identificação de Termos Relevantes

Ao iniciar a etapa 3, Identificação de Termos Relevantes, as sentenças são submetidas a um etiquetador POS (que consiste em rotular as palavras segundo a sua classe gramatical) e, em seguida,

Figura 11: Exemplo de um documento convertido (a) no formato XML e (b) no formato TXT (MATOS, 2010).

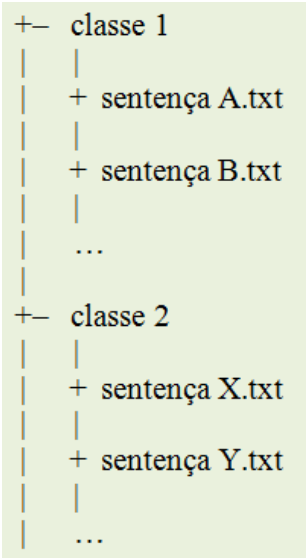

Figura 12: Exemplo de sentenças classificadas (MATOS, 2010).

cada sentença classificada e rotulada é analisada por “regras” escritas por expressões regulares. Essa é uma técnica de EI conhecida como “abordagem baseada em regras”.

O objetivo de se utilizar esta técnica é encontrar e separar termos relevantes baseando-se nos padrões de escrita dentro das sentenças. Todos os termos identificados são extraídos, inseridos em

um dicionário e designados como “termos não-curados”. Após este procedimento, especialistas do domínio avaliam tais termos, excluindo-os ou modificando seu *status* para “termos curados”.

O dicionário possui duas finalidades. A primeira finalidade, como citado anteriormente, é armazenar os novos termos não-curados encontrados. A segunda finalidade é auxiliar a identificação de sentenças que possuem os termos curados, aumentando a precisão dos resultados. Esta segunda função do dicionário engloba outra técnica de EI denominada “abordagem baseada em dicionário”.

Em Matos (2010) o dicionário é implementado por meio de um banco de dados, cujas tabelas armazenam informações tais como o tipo de termos encontrado (e.g. droga, tratamento, etc.), os nomes dos termos, variações para este termo (abreviações ou outras formas de escrita) e se o termo é curado ou não.

Figura 13: Dicionário de termos curados e suas variações (MATOS, 2010).

A Figura 13 apresenta um exemplo fictício de dicionário de termos, mostrando como é a tabela de termos curados e a tabela de variações dos termos.

3.1.4 Etapa 4: Gerenciamento de Termos

Na última etapa, Gerenciamento de Termos, o especialista possui fundamental importância na validação dos termos extraídos. Ele exclui termos irrelevantes ou transforma termos não-curados em curados. Por último os termos curados são inseridos manualmente no banco de dados biomédico (podendo ser o mesmo que o dicionário).

Outras funções que o especialista pode ter nesta etapa são:

- Inserir novos termos como curados;
- Hierarquizar termos;
- Mover termos extraídos, alterando sua categoria.

3.1.5 Resultados

Foram realizados diversos experimentos que avaliaram o método. Em geral, os melhores resultados para extração de termos atingiram 74,75 % de precisão, 87,06 % de cobertura e 80,43 % de medida-f.

3.2 (DUQUE, 2012)

No trabalho de mestrado, intitulado “Um Processo Eficiente para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico” (DUQUE, 2012), Juliana Lilian Duque estende o trabalho de Pablo Freire Matos.

Enquanto a proposta de Matos realiza a extração de informações sobre efeitos da doença AF, a proposta de Duque extrai informações sobre tratamentos da mesma doença.

Assim como Matos, Duque também utiliza as técnicas de EI, abordagem baseada em regras, abordagem baseada em aprendizado de máquina e abordagem baseada em dicionário. O grande diferencial é trabalhar com duas fases de classificação e, ainda, por meio de avaliações empíricas, sustentar a hipótese de que os termos de “Tratamento” estão próximos, ou, pelo menos, no mesmo parágrafo das sentenças de “Complicação”. Portanto, o parágrafo é considerado como uma unidade com conteúdo de informações centralizado, no qual se localiza a informação de interesse.

Na Figura 14, são apresentadas as etapas do processo de extração de informação sobre termos de tratamentos proposta neste projeto. O processo é composto por seis etapas: 1. Classificação de Sentenças de Complicação, 2. Agrupamento de Sentenças, 3. Classificador de Sentenças de Tratamento, 4. Etiquetador POS, 5. Extração e, por último, 6. Armazenamento em Banco de Dados.

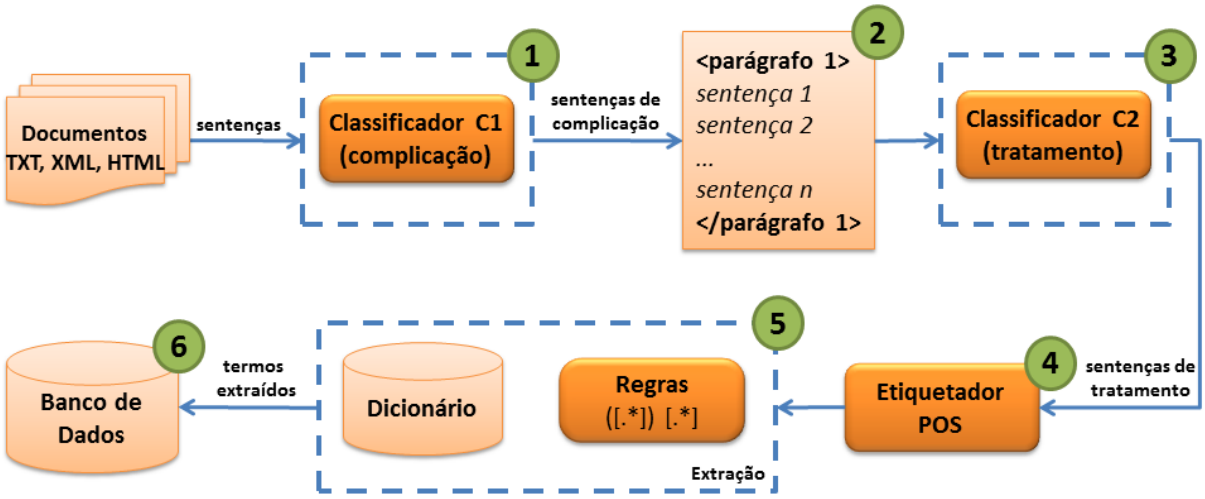

Figura 14: Etapas do processo para extração de tratamentos (DUQUE, 2012).

Entrada de dados

A entrada de dados acontece da mesma forma que no trabalho de Matos (2010). Os documentos, que estão originalmente em formato PDF sofrem conversão para XML ou TXT.

A conversão para o formato TXT é feita manualmente. Neste formato cada linha deve ser preenchida com uma sentença e no final de cada sentença deve constar o nome da seção entre parênteses. A Figura 15 exemplifica um documento TXT convertido manualmente.

Etapa 1: Classificação de Sentenças de Complicação

Na primeira etapa, com a utilização de uma abordagem por aprendizado de máquina, é gerado um modelo e realizada a classificação das sentenças de entrada nas classes **sentenças de complicação e outros**.

```
1  Patients were enrolled at 30 centers. (General)
2
3  The a globin genotype was unknown for most of the patients. (Results)
4
5  This represents 426 patient-years of follow-up. (Results)
6
7  Since its cause is largely unknown, therapy is supportive. (Abstract)
8
9  The use of HU at MTD may bring additional benefit. (Discussion)
10
11 The mean length of hospitalization was 10.5 days. (Abstract)
```

Figura 15: Exemplo de documento TXT (DUQUE, 2012).

As sentenças que não fazem parte do grupo de complicação são rejeitadas.

Etapa 2: Agrupamento de Sentenças em Parágrafos

Nesta etapa, as sentenças classificadas na Etapa1 como **sentenças de complicação**, são agrupadas com aquelas classificadas como **outras** sentenças que participam do mesmo parágrafo. Em seguida, as sentenças agrupadas são enviadas para o classificador de tratamentos. A Figura 16 mostra a classificação por complicação e o agrupamento das classes.

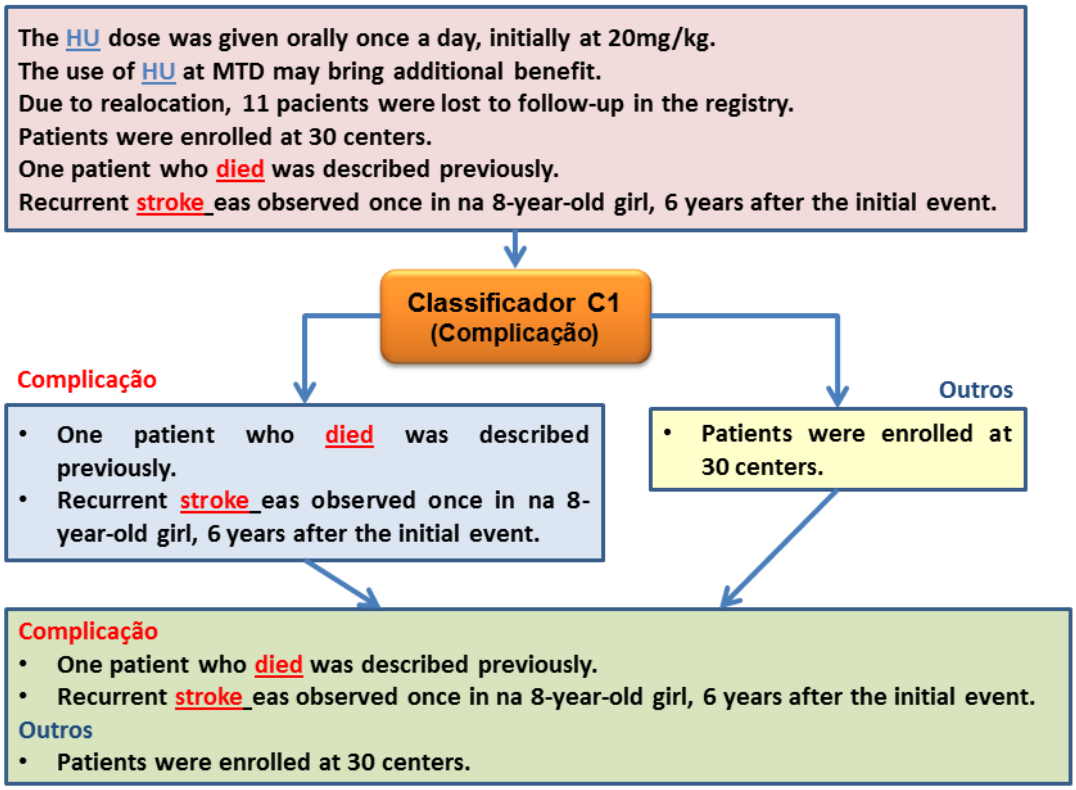

Figura 16: Exemplo de Classificação de Complicação e Agrupamento de Sentenças (DUQUE, 2012).

Etapa 3: Classificação de Sentenças de Tratamento

Na terceira etapa do processo, também utilizando a abordagem por aprendizado de máquina, é realizada a classificação em **sentenças de tratamento e outros**. A Figura 17 apresenta um exemplo

de sentenças classificadas.

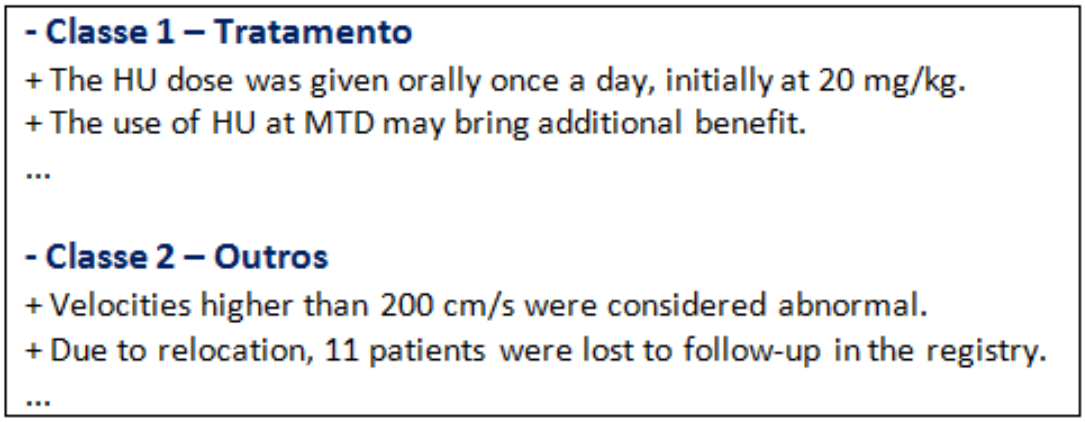

Figura 17: Exemplo de estrutura de Classificação de Tratamentos (DUQUE, 2012).

Etapa 4: Etiquetador POS

No passo 4, as sentenças com termos de tratamento são etiquetadas conforme a sua classe gramatical, pelo etiquetador POS. O objetivo deste procedimento é auxiliar a abordagem baseada em regras a encontrar padrões de escrita de textos. A Figura 18 apresenta um exemplo de sentença após a aplicação do etiquetados POS.

The_DT mean_JJ HU_NNP dose_NN per_IN kilo_NN was_VBD calculated_VBN each_DT year_NN ._.

Figura 18: Exemplo de sentença etiquetada pelo POS (DUQUE, 2012).

Etapa 5: Extração

Na etapa de extração, são utilizadas a abordagem baseada em regras e a abordagem baseada em dicionário. Assim como no método de Matos (2010), na abordagem por regras são utilizadas expressões regulares para encontrar padrões nas sentenças. A abordagem por dicionários auxilia a extração identificando termos existentes no dicionário.

Etapa 6: Armazenamento em Banco de Dados

Ao final do processo, na etapa 6, os termos relevantes e interessantes que foram extraídos são armazenados em um banco de dados relacional.

3.2.1 Resultados

Foram realizados experimentos que avaliaram o método. Em geral os melhores resultados para extração de termos atingiram 96 % de precisão e 19 % de cobertura. Apesar da baixa cobertura, em outros experimentos obteve-se uma taxa de 100 % de cobertura na identificação de termos disitintos. Isso representa a extração completa dos termos distintos relevantes existentes no *corpus* testado.

3.3 (GIRJU; MOLDOVAN, 2002)

Girju possui vários trabalhos na área de extração de relacionamentos semânticos. Em Girju e Moldovan (2002) é apresentado um estudo da extração de relacionamentos semânticos de causalidade (causa-efeito).

Utiliza-se um processo semelhante ao de Hearst (1998), cujos padrões são construídos manualmente, por observações do *corpus* utilizado. Porém, em Girju e Moldovan (2002) é utilizada a WordNet versão 1.7 (FELLBAUM, 1998) para a obtenção da lista de termos relacionados (passo 2) e para desambiguação de verbos polissêmicos, aqueles que assumem um mesmo significado. Por exemplo, em “Tinha muita gente na sala”, o verbo ter assume o mesmo significado do verbo haver, “Havia muita gente na sala”.

3.3.1 Algoritmo

O algoritmo de extração semiautomática utilizado por Girju e Moldovan (2002) é descrito em Hearst (1998) e apresentado a seguir:

Algoritmo 1. Extração de Relações Girju e Moldovan 2002.

Entrada: Sentenças selecionadas.
Saída: Relacionamentos identificados.

01 **INÍCIO**

02 Identifica-se uma relação semântica de interesse. (ex. hiponímia, meronímia, causalidade, etc.);

03 Constrói-se uma lista de termos para os quais se sabe que a relação é válida;

04 Procura-se no *corpus* por sentenças em que esses termos ocorrem sintaticamente próximos e armazena-se o contexto (palavras ao redor ou sentença) em que eles aparecem;

05 Procura-se por similaridades entre esses contextos e hipotetiza-se que contextos comuns indicam a relação de interesse;

06 Quando um padrão é identificado positivamente, ele é utilizado para encontrar mais instâncias da relação alvo e volta-se à linha 03.

07 **FIM**

Figura 19: Algoritmo 1: Extracao de Relações Girju e Moldovan (2002).

3.3.2 Avaliação e Resultados

Foi utilizado um *corpus* de 3GB de textos jornalísticos, provenientes do *Wall Street Journal*, *Financial Times*, entre outros. Utilizou-se também uma lista de 60 verbos que exprimem causalidade. Foram selecionadas 50 sentenças que continham cada verbo da lista, resultando em 3000 sentenças que foram analisadas sintaticamente e marcadas com etiquetas de *part-of-speech*.

A partir de 1321 relacionamentos obtidos, foi selecionada uma amostra de 300 para serem avaliados manualmente por dois anotadores humanos. A precisão média obtida foi de 65,6%, um pouco melhor que a de Hearst (1998).

No entanto, a comparação deve ser cuidadosa, pois os dois trabalhos têm foco em relações distintas (causalidade e hiponímia), e Girju e Moldovan (2002) sugerem que relações distintas têm dificuldades diferentes para serem extraídas, além dos métodos de avaliação serem distintos.

3.4 PolySearch

Este trabalho apresenta um sistema *Web*, *PolySearch* (PS), que realiza mineração de textos e extração de relacionamentos entre doenças humanas, genes, mutações, drogas e metabólitos (CHENG et al., 2008).

Os autores do PS procuraram solucionar um problema bastante parecido com o objetivo desta pesquisa de mestrado, a dificuldade que pesquisadores da área biomédica encontram para absorver todo o conteúdo existente em documentos científicos e bancos de dados biomédicas.

O sistema PS extrai informações e analisa relacionamentos entre componentes de diversas fontes de informações, como:

- **PubMed**: Base de dados de artigos científicos da biblioteca nacional dos Estados Unidos (PUBMED, 2013);
- **OMIM (*Online Mendelian Inheritance in Man*)**: Um catálogo de genes humanos e distúrbios genéticos (OMIM, 2013);
- **GAD (*Genetic Association Database*)**: Um repositório de estudos de associação genética de doenças complexas e distúrbios (GAD, 2013);
- **SwissProt**: Um banco de dados curado de sequências de proteínas que se esforça para fornecer um elevado nível de anotações (tais como a descrição da função de uma proteína, a sua estrutura de domínios, modificações pós-translacionais, variantes, etc.), um nível mínimo de redundância e alto nível de integração com outras bases de dados (SWISS-PROT, 2013);
- **HPRD (*Human Protein Reference Database*)**: Uma plataforma centralizada para representação visual e integração das informações relativas ao domínio da arquitetura, modificações pós-traducionais, redes de interação e de associação da doença para cada proteína no proteoma humano (HPRD, 2013);
- **DrugBank**: Um recurso de bioinformática e quimioinformática único que combina detalhes de drogas (ou seja, químico, farmacológico e farmacêutico) com informações de dados alvo da droga abrangente (e.g. sequência, estrutura e caminhos) (DRUGBANK, 2013);
- **HMDB (*Human Metabolome Database*)**: Um banco de dados disponível livremente de forma eletrônica, contendo informações detalhadas sobre metabólitos de moléculas pequenas encontradas no corpo humano (HMDB, 2013);
- **HapMap**: Um recurso disponível gratuitamente que contém informações relativas ao mapa de alótipos do genoma humano. Descreve os padrões comuns de variação humana da sequência de DNA (HAPMAP, 2013);

- **Entrez SNP (dbSNP)**: Um repositório central para ambas as substituições de bases simples de nucleotídeo único (SNPs), deleções curtas e polimorfismos de inserção no genoma humano (SNP, 2013);
- **CGAP SNP500Cancer Database**: Parte do Projeto Genoma do Câncer, é projetado especificamente para conter dados sobre a variação genética em genes importantes no câncer (CGAP, 2013);
- **Human Genome Mutation Database**: Um banco de dados que compreende vários tipos de mutação dentro das regiões de codificação e regiões reguladoras de genes nucleares humanos que causam doenças hereditárias (HGMD, 2013);

As consultas permitidas pelo PS respondem a questões do tipo: "Dado um X, encontre todos os Y associados", onde X pode ser uma simples doença humana, nome de gene ou proteína, droga, metabólito, SNP, sequência, ou palavra do texto fornecida pelo usuário. Y pode ser qualquer doença humana, nome de gene ou proteína, drogas, metabólitos, localizações sub-celulares, SNP, PCR *primers* ou palavra retornada ao usuário.

A Figura 20 apresenta uma tabela com todas as possíveis relações que o sistema pode identificar. Portanto, nem todas as combinações de busca podem ser feitas.

	Dado								
Encontre	Doença	Gene/ proteína	Droga	Metabólito	Palavra do texto	Caminho	Tecido	SNP	Sequência Gene/ proteína
Doença	✓	✓	✓	✓	✓	✓	✓		✓
Gene/proteína	✓	✓	✓	✓	✓	✓	✓	✓	✓
Droga	✓	✓	✓	✓	✓	✓	✓		✓
Metabólito	✓	✓	✓		✓	✓	✓		✓
Tecido	✓	✓	✓	✓	✓	✓	✓		✓
Órgão	✓	✓	✓	✓	✓	✓	✓		✓
Localização Sub celular	✓	✓	✓	✓	✓	✓	✓		✓
Caminho	✓	✓	✓	✓	✓				
Palavra do texto					✓				
SNP		✓							✓
PCR *primers*							✓		✓

Figura 20: Lista detalhada sobre as consultas básicas no PolySearch (CHENG et al., 2008).

3.4.1 Base de dados e Algoritmo

Baseado no tipo de consulta descrito anteriormente, "Dado um X, encontre todos os Y associados", onde X e Y são termos biológicos ou relacionados à saúde humana, *PolySearch* gera uma lista dos possíveis Y (mais seus sinônimos e abreviações) que se relacionam com X.

No caso de uma busca nos repositórios do PubMed, por exemplo, o sistema gera uma consulta formal utilizando os termos X e Y, com os respectivos operadores booleanos. Após a consulta retornar os artigos relacionados, o sistema realiza o *download* dos resumos. Em seguida é feita uma varredura de termos, sentença por sentença. As ocorrências de associações são agrupadas em categorias e pontuadas de acordo com sua relevância. Um *ranking* é formado para que os resultados possam ser apresentados aos usuário.

PolySearch não utiliza classificador POS, mas utiliza abordagem por regras, dicionário e *'bag-of-words'* para identificar relacionamentos relevantes. O software mantém uma coleção de nove bases de dados de palavras e sinônimos para genes humanos, proteínas, doenças, drogas, metabólitos, caminhos de proteínas, tecidos e localizações sub-celulares. Estas bases de dados são formadas a partir de bancos de dados públicos curados como *SwissProt, Entrez Gene, HGNC*, etc. Cópias destes bancos citados também estão armazenadas nas bases de dados do *PolySearch* para melhorar o desempenho das consultas, reduzindo o tempo de busca.

A premissa central para a estratégia de pesquisa do *PolySearch* é o pressuposto de que, quanto maior a frequência com que um X e Y se associam dentro de uma coleção de resumos ou bancos de dados, mais significativa esta associação é provável ser. Exemplo, se **COX2** aparece 510 vezes associado ao **câncer de cólon** e **tioredoxina** aparece apenas 1 vez, entende-se que a associação **COX2** $=>$ **câncer de cólon** é mais confiável.

A abordagem baseada em regras é utilizada como auxílio ao método estatístico de frequência explicado anteriormente. Neste caso, padrões de escrita baseado são localizados, procurando identificar relações, entre elas interações proteína-proteína. As regras aplicadas às sentenças seguem o seguinte padrão abaixo:

```
‘‘Palavra-chave de busca + termo de associação + Termo do dicionário + Termo qualquer’’
```

No padrão aplicado é verificada a quantidade de palavras entre o conjunto de termos do padrão. Pode-se observar que o padrão determinado pelos autores possui dependência com as palavras-chave de busca, efetuadas por meio do ambiente *web*.

3.4.2 Limitações

A principal limitação decorre na utilização de uma abordagem simples por dicionários para identificar associações biológicas e biomédicas. Esta abordagem é capaz de recuperar apenas relacionamentos conhecidos e registrados no dicionário. O algoritmo não consegue identificar novos termos para doenças, tipos celulares, genes, drogas, etc. Uma abordagem por regras auxilia o método estatístico de frequência, porém auxilia a verificação apenas de interações proteína-proteína. Outra limitação acontece com a inabilidade de extrair contexto ou significado de sentenças ou termos. Métodos que utilizam Inteligência Artificial (IA), palavras de contexto ou Aprendizado de Máquina (AM) poderiam incrementar potencialmente o sistema de identificação de termos existente.

O algoritmo e o sistema *web* desenvolvidos possuem mais duas limitações importantes que dificultam sua utilização no trabalho proposto nesta dissertação. Uma delas está relacionada à busca de artigos em repositórios online, sem permitir a possibilidade de inserção de um conjunto específico de documentos. Esse problema ainda é agravado pela falta de disponibilização dos códigos-fonte ou detalhes do algoritmo para alteração do sistema atual. Outra limitação verificada acontece no padrão

textual aplicado na abordagem por regras. O padrão utilizado para selecionar os melhores resultados traz consigo uma dependência com as palavras-chave de busca que são informadas via *interface web*.

Nesta pesquisa de mestrado, a principal necessidade é que sejam descobertos todos os relacionamentos semânticos existentes, com o grande diferencial de poder encontrar relacionamentos desconhecidos para os especialistas. Por meio do agrupamento desses relacionamentos em uma rede de conhecimentos, possam ser descobertos novos tratamentos para doenças.

3.4.3 Resultados

Foram realizados experimentos que avaliaram o método. Foram executadas duas tarefas: identificação de interações proteína-proteína e identificação de genes de doenças. Os resultados nessas tarefas atingem 88,81 % e 79 % de medida-f, respectivamente.

3.5 (TABA, 2013)

Com o trabalho de mestrado intitulado "Extração automática de relações semânticas a partir de textos escritos em português do Brasil", Taba busca extrair um conjunto de relacionamentos semânticos binárias a partir de textos jornalísticos escritos no português do Brasil. Utiliza tanto a abordagem baseada em regras, utilizando padrões textuais, quanto a abordagem baseada em aprendizado de máquina, englobando o uso de técnicas como classificadores probabilísticos e estatísticos e métodos de *kernel*.

3.5.1 Tipos de relações semânticas extraídas

Os tipos de relações semânticas extraídas por Taba (2013) podem ser visualizados na Tabela 4.

Tabela 4: Relações semânticas investigadas por Taba (2013).

	Relação semântica	**Sentença exemplo**	**Relação extraída**
1	location-of(algo/alguém, local)	Uma secretária pode ser encontrada em um escritório	location-of(secretária, escritório)
2	is-a(subclasse, superclasse)	Maçã é uma fruta	is-a(maçã, fruta)
3	property-of(algo/alguém, característica)	O prédio é alto	property-of(prédio, alto)
4	part-of(todo, parte)	Parafuso é uma parte de uma máquina	part-of(máquina, parafuso)
5	made-of(produto, substância)	Cacau é utilizado para fazer chocolate	made-of(chocolate, cacau)
6	effect-of(ação/estado, consequência)	Gripe causa febre	effect-of(gripe, febre)
7	used-for(entidade, função)	Pás são usadas para cavar	used-for(pás, cavar)

3.5.2 Recursos

Taba (2013) utilizou como recursos dois *corpora* de treinamento anotados manualmente com as relações semânticas existentes entre os termos de cada sentença.

O primeiro *corpus* é composto por 646 artigos científicos da revista Pesquisa FAPESP, escritos em português do Brasil, que possui um total de 17397 sentenças (cerca de 870 mil palavras). Foi originalmente utilizado e enriquecido com informações linguísticas no projeto ReTraTos (CASELI, 2007) e o processamento morfossintático do *parser* PALAVRAS (BICK, 2000).

O segundo, é o *corpus* do jornal *Folha de São Paulo*, do ano de 1994, CETENFolha, composto por cerca de 24 milhões de palavras. Esse *corpus* foi também enriquecido com informações morfossintáticas pelo *parser* PALAVRAS (BICK, 2000).

3.5.3 Ferramentas

O projeto contou com a utilização de ferramentas existentes na literatura para anotação e execução de diversas tarefas. São eles: anotador morfossintático PALAVRAS (BICK, 2000), o etiquetador morfossintático derivado no projeto ReTraTos (CASELI, 2007), o identificador de sintagmas nominais descrito em (SANTOS, 2005; OLIVEIRA, 2005).Para os experimentos com aprendizado de máquina foram utilizados o pacote de algoritmos WEKA (HALL et al., 2009), e o software SVM Light (JOACHIMS, 1998).

Durante o desenvolvimento do trabalho foi desenvolvida uma ferramenta específica para auxiliar a tarefa de anotação manual de relações semânticas em *corpora*: a ARS (Anotador de Relações Semânticas). Desenvolvida em Java, manipula sentenças codificadas em formato JSON, permitindo a marcação visual de termos e de relações entre eles.

3.5.4 Desenvolvimento

O desenvolvimento do trabalho foi realizado em duas abordagens:

1. Abordagem de padrões textuais;
2. Abordagem de aprendizado de máquina com diferentes classificadores (árvores de decisão C4.5 e *Support Vector Machines*)

Na abordagem de padrões textuais foram implementados os padrões de Hearst (1992) e Freitas e Quental (2007).

Na abordagem de aprendizado de máquina, foram desenvolvidas um total de 288 *features*. Estas foram testadas nos *corpora* apresentados na Seção 3.5.2.

Vários tipos de experimentos foram realizados para se testar a melhor combinação de algoritmos e *features*.

3.5.5 Resultados

Foram realizados diversos experimentos que avaliaram os métodos. Não houve uma abordagem para extração de relacionamentos de causalidade utilizando padrões textuais. Para extrações de relações de causalidade foi utilizada a abordagem por aprendizado de máquina. Os melhores resultados atingem 34 % de precisão, 10,4 % de cobertura e 16 % de medida-f utilizando o algoritmo de árvores de decisão e 45,5 % de precisão, 16,9 % de cobertura e 24,6 % de medida-f utilizando o algoritmo *Suport Vector Machine* (SVN).

3.6 Avaliação das abordagens investigadas

As abordagens descritas nas Seções 3.1 a 3.5 possuem algumas características que se aproximam muito do trabalho realizado nesta pesquisa. Tais trabalhos não podem ser diretamente comparados a este, pois são aplicados a domínios e idiomas diferentes. Porém, os resultados apresentados possuem grande proximidade com os resultados obtidos neste trabalho.

A Figura 21 apresenta as principais caracteríscas de cada abordagem.

	Idioma alvo	**Bases de Dados**	**Abordagem**	**Termos Extraídos**	**Relações Extraídas**	***Corpus* de Testes**	**Resultados**
Matos (2010)	Inglês	Qualquer	Dicionário, Regras e Aprendizado de Máquina	Termos de Efeitos	Não realiza	10 artigos científicos médicos	Prec.: 74,75% Cob.: 87,06% M-F: 80,43%
Duque (2012)	Inglês	Qualquer	Dicionário, Regras e Aprendizado de Máquina	Termos de Tratamento	Não realiza	10 artigos científicos médicos	Prec.: 96% Cob.: 19% M-F: 80,43%
Girju (2002)	Inglês	Qualquer	Regras	Não realiza	**X → Y** effect-of	3Gb de textos jornalísticos	Prec.: 65,6%
PolySearch (2008)	Inglês	PubMed, OMIM, GAD, SwissProt, HPRD, DrugBank, HMDB, HapMap, Entrez, CGAP e HGMDB	Dicionário, Regras e Comparação numérica em resumos	doença humana, gene, proteína, droga, metabólito, SNP, sequência, palavra do texto	**X → Y** effect-of	artigos científicos médicos	M-F: 88,81%
Taba (2013)	Português	Qualquer	Regras e Aprendizado de Máquina	Não realiza	**X → Y** location-of, is-a, property-of, part-of, made-of, effect-of, used-for	Textos jornalísticos	Prec.: 45,5% Cob.: 16,9% M-F: 24,6%

Figura 21: Características dos trabalhos correlatos.

3.6.1 Considerações Finais

Neste capítulo foram apresentados diversos trabalhos cujos temas estão relacionados a esta pesquisa de mestrado. Em seguida, foram detalhados alguns trabalhos que possuem maior proximidade com esta pesquisa de mestrado. Nenhum destes trabalhos porém, serão comparados com esta proposta de pesquisa.

Dos trabalhos com maior proximidade, algumas ideias foram utilizadas no desenvolvimento deste trabalho de mestrado. São elas:

Baseando-se nos trabalhos de Matos (2010), Duque (2012) e Cheng et al. (2008), utilizamos dicionários para extração de termos do domínio biomédico e abordagem por regras para extração dos relacionamentos. Dicionários são abordagens mais simples, porém bastante útil devido ao fato de não termos esta tarefa como foco principal.

Utilizando algumas ideias de Hearst (1998) e, consequentemente, (GIRJU; MOLDOVAN, 2002), e de Matos (2010), utilizamos um dos passos do algoritmo correspondente à construção e utilização de uma lista de palavras que indicam uma possível relação de causalidade. Nesta pesquisa, foi construído um dicionário, nomeado "*Tip Words Dictionary*".

Baseando-se também em Matos (2010), Duque (2012), foi utilizada o conceito de classificação de sentenças, que melhora o processo de extração tanto para termos como para relacionamentos semânticos. Nesta pesquisa, as MetaRegras classificam sentenças que possuem relacionamentos semânticos de Associação e *Increase/Decrease*.

Com base em Taba (2013), utilizamos a ferramenta desenvolvida, ARS (Anotador de Relações Semânticas) como base para implementação. Porém, diversas funções foram modificadas, além do idioma que foi convertido para o inglês.

Utilizando ainda algumas ideias extraídas de Taba (2013), realizamos a extração de relações semânticas utilizando abordagem baseada em regras (padrões textuais).

Contudo, a proposta de mestrado apresentada não se trata de adaptações dos trabalhos anteriores para um domínio específico. Nele, foram reutilizadas ideias importantes dos trabalhos citados para gerar novos algoritmos que extraem relacionamentos do tipo "causa e efeito" e os agrupem em uma rede de conhecimento possibilitando encontrar novos tratamentos para doenças.

No próximo capítulo é apresentado o método de extração de relacionamentos semânticos, do tipo "causa e efeito" para o domínio biomédico.

Capítulo 4

MÉTODO PARA A EXTRAÇÃO DE RELACIONAMENTOS SEMÂNTICOS

*Este capítulo apresenta um **estudo piloto** realizado no início dos trabalhos para o entendimento do problema e para a definição da proposta. Detalha a pesquisa de mestrado, mais especificamente, apresenta **definições**, **recursos** utilizados e detalha a **arquitetura do método proposto** para a extração de relacionamentos semânticos do tipo "causa e efeito" em artigos científicos do domínio biomédico.*

4.1 Estudo Piloto

Durante o período de revisão bibliográfica da literatura, um trabalho muito importante foi realizado pelo aluno com o objetivo de entender profundamente o problema a ser atacado e identificar nos textos científicos como as relações semânticas entre os termos e entre as sentenças acontecem. Foram feitas anotações sobre os relacionamentos existentes nos textos e quais os tipos de informação que podem ser extraídas. Este trabalho serviu como base para propor um método para a extração de relacionamentos semânticos em artigos científicos do domínio biomédico. O método será apresentado na Seção 4.6.

O estudo foi realizado por meio da leitura de um conjunto de artigos científicos relacionados à doença Anemia Falciforme (AF), mais especificamente sobre o ciclo de absorção do ferro (Fe) pelo organismo, todos escritos em língua inglesa.

A partir da leitura de quatro artigos relacionados à AF e ao ciclo de absorção do ferro, pôde-se extrair, por meio de obeservações e um processo manual, relações entre informações dos textos buscando construir cadeias de relações.

Alguns exemplos de relacionamentos extraídos são mostrados abaixo. O símbolo ($=>$) indica uma relação de causa e efeito. A causa são as palavras que antecedem o símbolo $=>$ e o efeito as palavras que sucedem o símbolo $=>$, ou seja, a ocorrência do antecessor (causa) leva a ocorrência do sucessor (efeito):

- **Artigo 1:**

Inflamação $=>$ Aumento da hepcidina $=>$ Diminuição da ferroportina $=>$ Diminuição do ferro $=>$ Anemia por deficiência de ferro $=>$ hipóxia (falta de oxigênio).

Sentenças:

1. Inflammation causes an increase of production of hepcidin. (Inflammation => Aumento da hepcidina)
2. Ferroportin is upregulated by the amount of available iron and downregulated through its interaction with hepcidin. (Hepcidina => Diminuição da ferroportina)
3. Hepcidin controls intestinal iron absorption by regulating ferroportin expression on the basolateral membrane of enterocytes. (Hepcidina => Aumento/Diminuição da ferroportina)
4. The exit of iron from macrophages is controlled by ferroportin, which is regulated by hepcidin. (Hepcidina => Aumento/Diminuição da ferroportina => Diminuição de ferro)
5. When iron supply to the plasma from macrophages and other storage sites is reduced, i.e. iron deficiency anemia, anemia of chronic inflammation (disease), and in some cases of ferroportin mutations. (Diminuição de ferro => Anemia por deficiência de ferro)
6. Many enzymes in oxygen-utilizing pathways are iron-dependent: thus, low iron content in the organism mimics hypoxia. (Diminuição de ferro => hipóxia)

- **Artigo 2:**

Aumento da hepcidina => Diminuição da absorção de ferro => Anemia por deficiência de ferro.

Sentenças:

1. Increased expression of hepcidin leads to decreased iron absorption and iron deficient anemia. (Aumento da hepcidina => Diminuição da absorção de ferro) e (Aumento da hepcidina => Anemia por deficiência de ferro)
2. Excessive decrease in iron absorption causes iron deficient anemia. (Diminuição da absorção de ferro => Anemia por deficiência de ferro)

- **Artigo 3:**

Excesso de hepcidina => Aumento de ferritina => Aumento de ferro armazenado nos macrófagos => Diminuição da saturação de ferro/transferrina => Diminuição de ferro reciclados.

Sentenças:

1. The increased serum ferritin indicative of increased macrophage iron stores and decreased serum iron/transferrin saturation indicative of decreased macrophage iron recycling typical of this disorder suggest a condition of hepcidin excess, as is seen in several murine models of hepcidin overexpression.
   (Excesso de hepcidina => Aumento de ferritina => Aumento de ferro armazenado nos macrófagos => Diminuição da saturação de ferro/transferrina => Diminuição de ferro reciclados.)

- **Artigo 4:**

Aumento de ferro $=>$ Aumento de hepcidina $=>$ Degradação da ferroportina $=>$ Diminuição de ferro no plasma.

Sentenças:

1. Hepcidin synthesis is stimulated by increased plasma iron and tissue iron stores, and hepcidin in turn decreases the release of iron into plasma, both from macrophages and from absorptive enterocytes in the duodenum.
2. Upon reaching its target tissues hepcidin binds to ferroportin and causes its internalization and degradation.
3. Removal of ferroportin from the cells surface decreases the efflux of iron from cells into plasma. (Aumento de ferro $=>$ Aumento de hepcidina $=>$ Degradação da ferroportina $=>$ Diminuição de ferro no plasma.)

Em todos os artigos lidos é possível encontrar outros relacionamentos com menor número de ligações. Os relacionamentos apresentados acima ilustram a ideia principal do texto. Cadeias com menos relacionamentos são de grande importância, pois podem se relacionar com outros relacionamentos oriundos de artigos diferentes, complementando-as e incrementando-as.

Além dos relacionamentos semânticos, também são encontrados conceitos, por exemplo, sobre funções moleculares, além de siglas que representam moléculas ou genes.

A partir dos estudos preliminares realizados e detalhados nesta seção e dos objetivos apresentados na Seção 1.3, pudemos inferir as hipóteses que foram apresentadas na Seção 1.4.

A seguir serão apresentadas algumas definições essenciais para o entendimento da solução apresentada.

4.2 Definições

Antes de iniciar a descrição do desenvolvimento do trabalho de pesquisa é interessante apresentar algumas definições que facilitarão o entendimento do método.

- ***Token*** – É toda sequência de caracteres com exceção do espaço em branco;
- **Termo** – É toda sequência de *tokens* que representa uma entidade ou tem algum significado específico em uma sentença;
- ***Tip Word*** – Possui a mesma definição de **Termo**. A diferença é que o *Tip Word* é um termo, normalmente, mas não obrigatoriamente, um verbo, que sugere ao algoritmo de extração de relacionamentos que em determinada sentença podem existir relacionamentos de causalidade. Os *tip words* se distinguem em:
  - ***Increase*/*Decrease***: São aqueles que indicam que um termo aumenta ou diminui. Por exemplo: *[increased]1 [macrophage iron stores]2*.
    A anotação marcada com o número 1 determina um ***tip word*** que indica **aumento**, enquanto a anotação marcada com o número 2 determina um **termo**.

  - ***Association***: São aqueles que indicam que um termo possui associação com outro. Por exemplo: *[inflammation]1 [causes]2 an [increase of]3 [production of hepcidin]4*.
    As anotações marcadas com os números 1 e 4 determinam **termos**. A anotação marcada com o número 3 determina um **aumento**, como no exemplo anterior. A anotação marcada com o número 2 determina um ***tip word*** que indica uma **associação** entre os termos 1 e 4.
  - ***Negative***: São aqueles que indicam ***tip words*** de ***negação***. Por exemplo: *levels of [sICAM-1]1 [was not]2 significant*.
    A anotação marcada com o número 1 determina um **termo**. A anotação marcada com o número 2 determina um ***tip word*** que indica uma **negação**.
  - ***Possibility***: São aqueles que indicam ***tip words*** de ***possibilidades***. Por exemplo: *[excessive]1 [endothelial activation]2 and [vaso-constriction]3 because of impaired NO bioavailability [may]4 [contribute to]5 [vascular instability]6*.
    As anotações marcadas com os números 2, 3 e 6 determinam **termos**. A anotação marcada com o número 1 determina um **aumento**. A anotação marcada com o número 5 determina um ***tip word*** que indica uma **associação** entre os termos. A anotação marcada com o número 4 determina um ***tip word*** que indica uma **possibilidade**.

- **Relacionamento semântico** – Usualmente, a definição adotada na literatura para expressar um relacionamento semântico possui o mesmo formato utilizado por Taba (2013), no qual um relacionamento é uma tripla $<$*relação, termo1, termo2*$>$, onde *relação* é a denominação de um relacionamento semântico (por exemplo is-a, made-of, part-of) e *termo1* e *termo2* são dois termos distintos em uma sentença. Nesta pesquisa realizamos uma extensão da notação usual, pois a mesma não exprime todo o conteúdo que este trabalho pretende representar nos relacionamentos semânticos extraídos.
  Na extensão utilizada definiremos duas representações. A primeira, utilizada em relacionamentos que não possuem associação, representaremos um **relacionamento semântico** como uma tripla $<$*relação, termo_estendido1, termo_estendido2*$>$. Na segunda, utilizada em relações que possuem associação, representaremos por meio de uma quádrupla $<$*relação, termo_estendido1, tip_word,*
  termo_estendido2$>$, onde, ***relação*** é a denominação de um relacionamento semântico semelhante à notação usual (por exemplo is-a, made-of, part-of), ***tip_word*** é a *tip word* que participa de um relacionamento de associação, negação ou possibilidade e ***termo_estendido1*** e ***termo_estendido2*** são dois termos distintos em uma sentença, os quais possuem as seguintes variações:

  - apenas o *termo*. Por exemplo: [inflamation];
  - *tip word increase* $+$ *termo*. Por exemplo: [increase] [inflamation];
  - *tip word decrease* $+$ *termo*. Por exemplo: [decrease] [inflamation].

  Com um exemplo mais completo, a partir das sentenças abaixo podemos extrair algumas relações:

```
Sentença tipo Increase/Decrease: With [endothelial dysfunction]1 and [vascular
 injury]2 , the levels of endothelial bound and [soluble adhesion molecules]3
                              [increase]4 .
```

```
Relação 1: cause-effect([endothelial dysfunction], [increase] [soluble adhesion molecules])

  Relação 2: cause-effect([vascular injury], [increase] [soluble adhesion molecules])

                                    ----------

  Sentença de Associação: Levels of [soluble endothelium-derived adhesion molecules]1
in patients with sickle cell disease are [associated with]2 [pulmonary hypertension]3 ,
                        [organ dysfunction]4 , and [mortality]5 .

Relação 1: cause-effect([soluble endothelium-derived adhesion molecules], [associated with],
                                [pulmonary hypertension])

Relação 2: cause-effect([soluble endothelium-derived adhesion molecules], [associated with],
                                [organ dysfunction])

Relação 3: cause-effect(soluble endothelium-derived adhesion molecules, [associated with],
                                [mortality])
```

Como o trabalho desenvolvido busca apenas relacionamentos de causalidade, os relacionamentos encontrados serão nomeadas como *(cause-effect(termo_estendido1, termo_estendido2))* ou *(cause-effect(termo_estendido1, tip_word, termo_estendido2))*. Na literatura, podemos encontrar a forma usual escrita como *(cause-effect(termo1,termo2))* e também no formato *(effect-of(termo1,termo2))*.

4.3 Recursos

Para o desenvolvimento desta pesquisa de mestrado foi necessária a utilização de alguns recursos textuais. Com a ajuda do especialista de domínio da área biomédica, foram obtidos dois *corpora* contendo artigos científicos da mesma área.

O primeiro *corpus*, o qual chamaremos durante esta dissertação de *Corpus* Estudo Piloto, é constituído de 17 artigos que envolvem o tema da Anemia Falciforme (AF), relacionados ao problema do **priapismo** e da **cadeia de metabolismo do ferro**. Este *corpus* foi utilizado como base para o estudo piloto apresentado na Seção 4.1. A partir deste *corpus* foram geradas as regras, baseadas em padrões textuais, que realizam a seleção de sentenças e extração de relacionamentos do tipo “causa e efeito”.

O segundo *corpus* é constituído de 30 artigos. 15 deles envolvem o tema da AF e estão relacionados ao problema da **hipertensão pulmonar**. Os outros 15 artigos também estão relacionados ao problema da **hipertensão pulmonar**, porém estão ligados a outras doenças diferentes da AF.

Neste segundo *corpus*, o qual chamaremos durante esta dissertação de *Corpus* de Trabalho, destacamos a escolha de dois subconjuntos de artigos. Ambos os subconjuntos estão relacionados à doenças distintas, porém tratando de um mesmo efeito negativo. Esta estratégia serve para demonstrarmos que os artigos podem se relacionar e ideias extraídas de documentos que envolvem doenças diferentes daquelas pesquisadas pelo especialista pode ajudar diretamente nos estudos por novos tratamentos e curas.

Outro recurso utilizado neste trabalho foi um conjunto de ontologias. Extraídas do *The Open Biological and Biomedical Ontologies* (OBO, 2013), foram utilizadas como base para a construção de dicionários de termos do domínio biomédico, utilizados na extração de Termos. Mais detalhes estarão descritos na Seção 4.6.2 - Etapa 2: Extração de Termos - e na Seção 4.6.2.1 - Dicionários de Termos e *Tip Words*.

4.4 Anotação do *Corpus* de Trabalho

Nas tarefas de Extração Automática existe a necessidade de realizar anotações manuais no *corpus* de trabalho para que os resultados possam ser avaliados. Desta forma, é possível realizar uma comparação entre a abordagem de extração automática e documentos anotados manualmente pelos especialistas.

A anotação foi feita em três níveis distintos:

1. **Anotação de termos de domínio:** Consiste na anotação dos termos importantes do texto relacionados com a área biomédica.

   ```
   Levels of [soluble endothelium-derived adhesion molecules] in patients with
       [sickle cell disease] are associated with [pulmonary hypertension] ,
                   [organ dysfunction] , and [mortality] .
   ```

   Os termos anotados entre colchetes são bastante relevantes na área biomédica.

2. **Anotação de termos especiais:** Consiste na anotação dos termos, denominados neste trabalho como ***tip words***, que podem indicar que determinada sentença possui um relacionamento do tipo causa e efeito. Matos (2010) e Duque (2012) utilizam um conceito semelhante que chamam de "verbos representativos". Utilizando a mesma sentença acima, temos:

   ```
   Levels of soluble endothelium-derived adhesion molecules in patients with
       sickle cell disease are [associated with] pulmonary hypertension ,
                   organ dysfunction , and mortality .
   ```

   O termo ***associated with***, anotado entre colchetes, nos traz a indicação de que a sentença pode possuir algum tipo de relacionamento.

   Nos níveis 1 e 2 foram anotados os 30 artigos relacionados com o problema de **hipertensão pulmonar**. Desses, 15 artigos relacionam o problema como efeito negativo da Anemia Falciforme e 15 relacionam o problema a outras doenças.

3. **Anotação de relacionamentos semânticos de "causa e efeito":** Após a identificação de sentenças que possuam relacionamentos, são anotados o par de termos que se relacionam.

   ```
   Levels of [soluble endothelium-derived adhesion molecules]1 in patients with
     [sickle cell disease]2 are [associated with]3 [pulmonary hypertension]4 ,
   ```

```
[organ dysfunction]5 , and [mortality]6 .

cause-effect(1,3,4) ; cause-effect(1,3,5) ; cause-effect(1,3,6)
```

Os termos de interesse anotados anteriormente são identificados entre colchetes e pelos números de 1 a 6 e as relações semânticas *cause-effect* indicam as relações de causa e efeito que existem na sentença.

Neste nível, foi selecionado e anotado um subconjunto de 10 artigos do conjunto anotado nos níveis 1 e 2. A partir deles os testes foram realizados.

Além da anotação manual, utilizou-se um etiquetador morfossintático *Part-of-Speech* (*POS*). Esse tipo de anotação é bastante comum no processo de extração de informação. Neste trabalho de pesquisa não foi necessária a utilização das anotações *Part-of-Speech*, porém foi mantida nos *corpora* anotados para o caso de futuros trabalhos necessitarem do recurso.

Mesmo para a anotação manual foi utilizada uma ferramenta bastante útil, desenvolvida por Taba (2013) (e denominada Anotador de Relações Semânticas) e totalmente adaptada para o idioma e o domínio do Projeto SCA. A ferramenta será apresentada com maiores detalhes na Seção 4.5.

4.5 Ferramentas

Durante o processo de anotação e desenvolvimento da proposta, foram utilizadas e desenvolvidas algumas ferramentas.

Durante o processo de anotação morfossintática do *corpus*, apresentado na Seção 4.3, foi utilizado um etiquetador Part-of-Speech (POS), *Stanford POS Tagger* (TOUTANOVA et al., 2003). Este etiquetador foi extraído do Grupo de Processamento de Língua Natural da Universidade de Stanford nos Estados Unidos. Sua função é reconhecer e adicionar etiquetas morfossintáticas no texto.

Neste trabalho de mestrado foram desenvolvidas duas importantes ferramentas. A primeira delas, nomeada **JPdf2JSON**, é uma ferramenta para converter documentos do formato PDF para o formato TXT e JSON. O processo de conversão ocorre em três passos básicos:

1. Seleção dos arquivos que serão convertidos: Neste passo podem ser selecionados múltiplos arquivos;
2. Conversão para o formato TXT e limpeza do texto: A conversão do formato PDF para o formato TXT ocorre com o utilização da biblioteca PDFBox (PDFBOX, 2013). É possível aplicar alguns filtros para limpeza dos textos, como remoção de hifenizações, remoção de quebras de linha, quebrar sentenças linha a linha, colocar todas as palavras em letras minúsculas e inserir quebras de linha de modo a deixar o texto visualmente legível;
3. Aplicação do etiquetador POS e Geração do formato JSON: Neste último passo existe a opção de aplicação do etiquetador morfossintático *Stanford POS Tagger* citado no parágrafo anterior. Por último, é gerado um arquivo no formato JSON. A estrutura do arquivo JSON pode ser visualizada no Apêndice A.

As telas da ferramenta JPdf2JSON podem ser visualizadas no Apêndice B.

A segunda ferramenta, nomeada **Anotador de Relações Semânticas** (ARS), foi totalmente modificada e adaptada ao problema à partir de Taba (2013). Taba desenvolveu a ferramenta para o idioma português, buscando encontrar diversos tipos de relacionamentos semânticos, utilizando abordagem por regras e abordagem por aprendizado de máquina. Neste trabalho a ferramenta foi inteiramente modificada para atender a extração de relacionamentos em artigos científicos, em idioma inglês.

Com a ferramenta é possível anotar manualmente e extrair automaticamente os termos de um domínio dentro das sentenças dos artigos, utilizando a abordagem por dicionário. É possível, também, anotar manualmente e extrair automaticamente relacionamentos semânticos do tipo "causa e efeito". O sistema possui ainda a função de visualização de uma possível rede de conhecimentos, construída a partir dos relacionamentos extraídos.

A tela principal da ferramenta ARS pode ser visualizada no Apêndice C.

Por último, foram desenvolvidos alguns *scripts*, utilizando linguagem de programação Perl, para contagem de resultados e dados de testes. Eles fazem uma contagem automática dos dados e emitem resultados numéricos, incluindo as métricas padrão utilizadas neste trabalho como forma de avaliação. Detalhes sobre os *scripts* podem ser visualizados no Apêndice D e Apêndice E.

Todo o conteúdo produzido durante o trabalho de pesquisa, nos quais incluem os artigos utilizados, os artigos anotados (com todos os níveis de anotações apresentados), as ferramentas desenvolvidas e os dados dos resultados, estão disponíveis no *website* do **Grupo de Banco de Dados da UFSCar (GBD)** (UFSCAR, 2013), por meio do seguinte endereço: `http://gbd.dc.ufscar.br/sicklecellanemia/`.

4.6 Arquitetura do Método Proposto

Como descrito na Seção 1.3, o objetivo desta pesquisa de mestrado foi relacionar informações de artigos científicos, extraindo relacionamentos semânticos do tipo "causa e efeito" em sentenças, para auxiliar nas necessidades dos médicos e pesquisadores que trabalham na descoberta de novos tratamentos e curas para doenças.

A Figura 22 nos apresenta uma visão em alto nível das principais etapas do método de extração de relacionamentos semânticos proposto. No total tem-se um processo em 4 etapas distintas: 1. Entrada e Preparação de Dados, 2. Extração de Termos, 3. Identificação de Relacionamentos Semânticos do tipo "causa e efeito" e 4. Construção de uma Rede Semântica de Conhecimentos.

Nas seções subsequentes será detalhada cada etapa descrita na arquitetura apresentada.

4.6.1 Etapa 1: Entrada e Preparação de Dados

A Figura 23 destaca os passos de execução que ocorrem na Etapa 1, Entrada e Preparação de Dados. Passo 1: Entrada de documentos PDF. Passo 2: Conversão para TXT e Limpeza do texto. Passo 3: POS *tagger* e Geração do arquivo JSON.

Todo o processo da Etapa 1 pode ser feito de forma manual, porém para facilitar o trabalho dos usuários foi desenvolvida a ferramenta JPdf2JSON, apresentada na Seção 4.5.

A ferramenta foi implementada utilizando a linguagem de programação Java, a biblioteca de con-

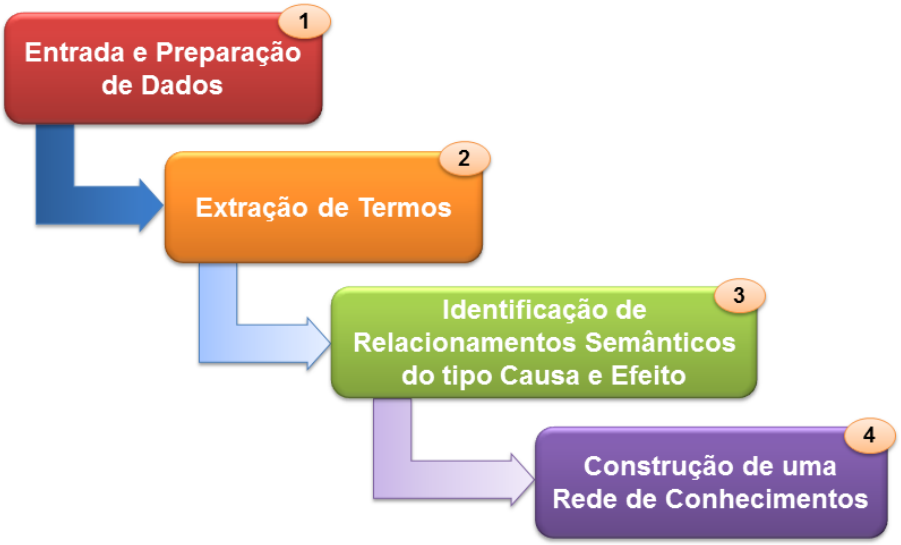

Figura 22: Etapas do método de extração de relacionamentos semânticos.

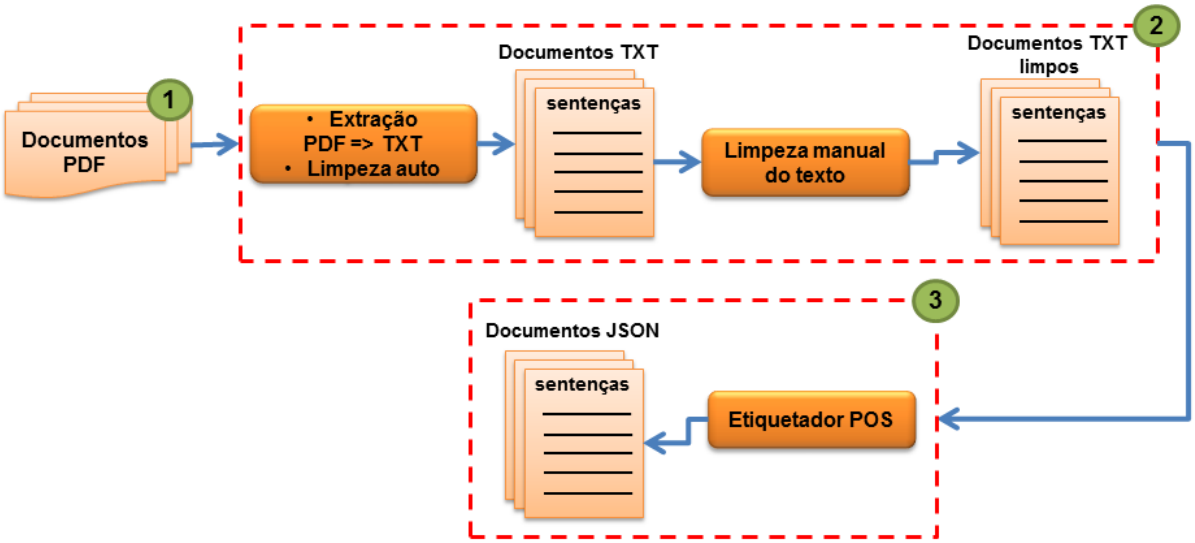

Figura 23: Etapa 1: Entrada e Preparação de Dados.

versão de formato PDF para TXT chamada PDFBox (PDFBOX, 2013) e o etiquetador morfossintático *Stanford POS Tagger* (TOUTANOVA et al., 2003).

No **passo 1**, Entrada de documentos PDF, os especialistas do domínio fornecem ao sistema um conjunto de documentos específicos. Estes documentos são artigos científicos completos, em formato PDF.

No **passo 2**, Conversão para TXT e Limpeza do texto, os documentos em formato PDF são convertidos para o formato TXT. Se a conversão for feita utilizando-se a ferramenta JPdf2JSON, e consequentemente a biblioteca PDFBox, será necessária a utilização dos algoritmos de limpeza automática do texto. Esses algoritmos estão presentes na ferramenta JPdf2JSON. Com eles é possível remover hifenizações, remover de quebras de linha, quebrar sentenças linha a linha, colocar todas as palavras em letras minúsculas e inserir quebras de linha de modo a deixar o texto visualmente legível;

Por outro lado, a biblioteca PDFBox não consegue excluir, por exemplo, caracteres especiais como

α e β, dados oriundos de tabelas (que podem confundir o algoritmo de extração de relacionamentos), após a geração do arquivo TXT, ainda no passo 2, uma boa prática é a limpeza manual de alguns dados do texto que não podem ser filtrados automaticamente. Esta limpeza faz com que trechos do texto que não são úteis não comprometam futuramente a extração de termos e relacionamentos semânticos.

Abaixo, destacam-se possíveis trechos, normalmente encontrados na maioria dos textos científicos, que podem ser excluídos manualmente:

- Nomes de autores e localidades;
- Palavras-chave;
- Dados provenientes de tabelas;
- Seções de agradecimentos;
- Notas de rodapé;
- Referências bibliográficas.

Como a biblioteca PDFBox não consegue converter de forma perfeita e sem erros, o processo de limpeza manual dos dados é bastante importante para não refletir problemas para a etapa de extração de relacionamentos semânticos. Durante este processo, é importante:

- Separar palavras do texto que aparecem unidas erroneamente, por exemplo, *sicklecell disease*, quando o correto é *sickle cell disease*;
- Corrigir parágrafos quebrados.

Enquanto não houver disponível alguma ferramenta que realize perfeitamente o processo de conversão dos formatos PDF para TXT, o processo de limpeza manual terá que ser feita, caso contrário, poderá transmitir erros para as etapas seguintes.

Por último, no **passo 3**, POS *tagger* e Geração do arquivo JSON, é aplicado o etiquetador *Part-of-Speech* ao texto e gerado um arquivo no formato JSON. A estrutura do arquivo JSON pode ser visualizada no Apêndice A.

4.6.2 Etapa 2: Extração de Termos

A Figura 24 destaca os passos de execução que ocorrem dentro da Etapa 2, Extração de Termos. Passo 1: Entrada de documentos JSON. Passo 2: Extração de Termos do Domínio. Passo 3: Extração de *Tip Words*.

No **passo 1**, os dados em formato JSON gerados manualmente ou pela ferramenta JPdf2JSON serão utilizados como entrada de dados na Etapa 2. A partir desta etapa os processos também podem ser executados manualmente, porém foi desenvolvida a ferramenta **Anotador de Relações Semânticas (ARS)** para facilitar este trabalho.

Nos **passos 2 e 3**, são extraídos dois tipos de termos: Termos do domínio e *Tip Words*. Os termos de domínio são aqueles considerados relevantes dentro do domínio biomédico, por exemplo, genes, doenças, proteínas, tratamentos, etc. Detalhes sobre o algoritmo aplicado a essas etapas podem ser vistos na Seção 4.6.2.2.

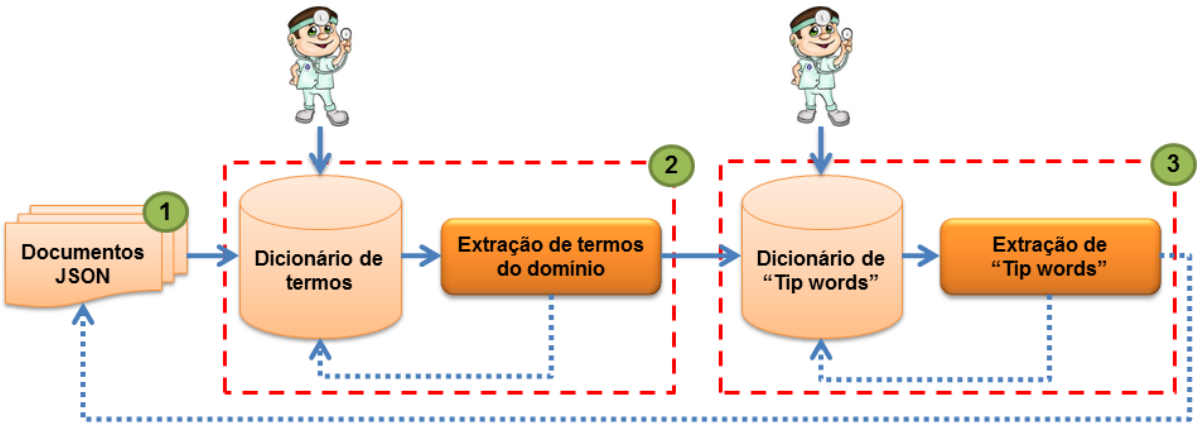

Figura 24: Etapa 2: Extração de Termos.

O papel do especialista do domínio é de fundamental importância nesta etapa. É ele quem irá validar se os termos extraídos são adequados ou não e em quais categorias cada um pode ser inserido.

A ferramenta ARS provê recursos importantes nesta etapa. Ela permite tanto a extração manual dos termos e *tip words* quanto a extração automática. Todos os dados extraídos são salvos no próprio arquivo JSON.

Na extração automática de termos e *tip words* é utilizada a abordagem baseada em dicionários, portanto, dois dicionários foram construídos para dar suporte à tarefa.

4.6.2.1 Dicionários de Termos e *Tip Words*

Para realizar a anotação do *corpus* de trabalho e a extração automática de termos e *tip words* foi necessária a construção de dois dicionários: **Dicionário de Termos** e **Dicionário de Tip Words**.

Um dicionário é um banco de dados que armazena nomes e outras informações a respeito desses nomes. Neste caso, os nomes são termos utilizados na área médica.

Para a construção do Dicionário de Termos, foi realizado o mapeamento de um conjunto ontologias do domínio biomédico. Elas serviram de base para o reconhecimento dos termos. O mapeamento, neste caso, é extrair da ontologia os nomes dos termos, descrições, categorias e sinônimos e copiá-los para o arquivo referente ao dicionário. Durante o mapeamento foi feito o trabalho de eliminar nomes e sinônimos duplicados.

Abaixo segue a lista de ontologias utilizadas. Elas foram extraídas do repositório The Open Biological and Biomedical Ontologies (OBO, 2013):

- Uberon Ontology (MUNGALL et al., 2012) - *Anatomial Structures*;
- Cell Type (BARD; RHEE; ASHBURNER, 2005) - *Cellular Structures*;
- ChEBI Ontology (DEGTYARENKO et al., 2008) - *Chemical Entities*;
- Human Disease Ontology (OSBORNE et al., 2009) - *Diseases*;
- Genes (CTD, 2013) - *CTD Genes Vocabulary*;
- Protein Ontology (WU, 2003) - *Protein and Cellular Components*.

Além das ontologias, durante o trabalho de anotação do *corpus* de trabalho, cada termo encontrado que não existia no dicionário foi incluído e categorizado, incrementando ainda mais o banco de dados. O trabalho de incluir, excluir e validar termos deve ser feito pelo especialista do domínio.

Os detalhes sobre a estrutura e as categorias do dicionário de termos do domínio estão disponíveis no Apêndice F.

Os *tip words*, como apresentado na Seção 4.2, são termos encontrados no *corpus* que servem como sugestão para que o algoritmo de extração de relacionamentos semânticos possa reconhecer sentenças que possuam relações de causalidade. No trabalho de Matos (2010) e Duque (2012) existe um conceito semelhante, cuja nomenclatura é "verbos representativos", porém, no caso dos *tip words*, os termos não são compostos apenas de verbos podem conter palavras de negação (por exemplo, *not, don't*), palavras de possibilidade (por exemplo, *could, maybe*).

O Dicionário de *Tip Words* foi construído à partir do trabalho de anotação do *corpus* de trabalho. Os detalhes da estrutura e das categorias do dicionário estão disponíveis no Apêndice G.

4.6.2.2 Algoritmo de Extração de Termos

Como discutido anteriormente, os processos dos passos 2: Extração de Termos e 3: Extração de *Tip Words* estão conectados. O algoritmo escrito em pseudo-código pode ser entendido com mais detalhes a seguir:

O algoritmo da Figura 25 é detalhado da seguinte forma:

Entrada: O algoritmo recebe como entrada o arquivo JSON com sentenças extraídas dos artigos.

Saída: O algoritmo retorna o mesmo arquivo JSON com termos e tip words anotados nas sentenças.

Variáveis:

- **qtde_sentenca:** Variável que refere-se à quantidade de sentenças;
- **point_sentenca:** variável que refere-se ao ponteiro de sentenças;
- **qtde_tokens:** variável que refere-se a quantidade de *tokens*;
- **point_tokens:** variável que refere-se ao ponteiro de *tokens*;

Descrição:

- Da linha 02 a 05, inicializam-se as variáveis;
- Na linha 06 inicia-se um laço contador de sentenças, que se repete até que todas as sentenças sejam lidas (instrução da linha 28);
- Na linha 07 realiza-se a leitura da sentença marcada pela variável **point_sentenca**.
- Na linha 08 inicia-se o processo do passo 2: Extração de Termos. Inicia-se um laço contador de *tokens*, que se repete até que todos os *tokens* da sentença atual sejam lidos (instrução da linha 16);
- Na linha 09 é realizada a montagem de um conjunto de *tokens*. O conjunto é montado tomando-se como base o primeiro *token* da sentença.
  São montados 10 conjuntos, sendo que o primeiro contém os 10 primeiros *tokens* da sentença,

Algoritmo 1. Passo 2 e 3: Extração de Termos e Tip Words.

```
Entrada: Arquivo JSON com sentenças extraídas dos artigos
Saída: Arquivo JSON com termos e tip words anotados nas sentenças.
Dados: qtde_sentenca: variável que refere-se a quantidade de sentenças;
       point_sentenca: variável que refere-se ao ponteiro de
       sentenças;
       qtde_tokens: variável que refere-se a quantidade de tokens;
       point_tokens: variável que refere-se ao ponteiro de tokens;
01  INÍCIO
02     qtde_sentenca ← quantidade de sentenças;
03     point_sentenca ← 0;
04     qtde_tokens ← quantidade de tokens da sentença atual;
05     point_tokens ← 0;

06     Repita {
07        Leia a sentença(point_sentenca);

08        Repita {
09           Montar 10 conjuntos (10, 9, 8, ... , 1) tokens;
10           Comparar o conjunto de tokens com o Dic. de Termos e seus
             Sinônimos;

11           Se (conjunto de tokens existe na sentença) então {
12              Marcar o conjunto de tokens como um termo no arquivo
                JSON;
13              point_tokens ← posição do ultimo tokens encontrado;
14           Senão
15              point_tokens ← point_tokens + 1;
16        até (point_tokens > qtde_tokens)

17        point_tokens ← 0;
18        Repita {
19           Montar 10 conjuntos (10, 9, 8, ... , 1) tokens;
20           Comparar o conjunto de tokens com o Dic. de Tip Words e
             seus Sinônimos;

21           Se (conjunto de tokens existe na sentença e nenhum termo
             marcado no conjunto de tokens) {
22              Marcar o conjunto de tokens como uma tip word no
                arquivo JSON;
23              point_tokens ← posição do ultimo tokens encontrado;
24           Senão
25              point_tokens ← point_tokens + 1;
26        até (point tokens > qtde tokens)

27        point_sentenca ← point_sentenca + 1;
28     até (point_sentenca > qtde_sentenca)
29  FIM
```

Figura 25: Algoritmo 1: Extracao de Termos e *Tip Words*.

o segundo contém os 9 primeiros *tokens*, sucessivamente até obter 10 conjuntos e o último conjunto possuir apenas o primeiro *token*. Lembrando que vírgulas, pontos, números, etc. também são considerados *tokens*. Veja o exemplo abaixo para a sentença:

Sentença: [With]1 [endothelial]2 [dysfunction]3 [and]4 [vascular]5 [injury]6 [,]7 [the]8 [levels]9 [of]10 [endothelial]11 [bound]12 [and]13 [soluble]14 [adhesion]15 [molecules]16 [increase]17 [.]18

Os tokens estão marcados entre colchetes. Abaixo são apresentados os 10 primeiros conjuntos de *tokens* para esta sentença:

Conjunto 1: With endothelial dysfunction and vascular injury , the levels of
Conjunto 2: With endothelial dysfunction and vascular injury , the levels

```
Conjunto 3: With endothelial dysfunction and vascular injury , the
...
Conjunto 10: With
```

- Após a montagem, na linha 10 dos conjuntos de *tokens*, eles são comparados com o dicionário de termos e seus sinônimos. O objetivo é verificar se os conjuntos existem como um termo do dicionário. O processo começa pelo conjunto com maior número de *tokens*. Este representa um termo cujo nome é mais específico;
- Na linha 11, é feita uma verificação se algum termo foi encontrado com o conjunto de tokens montado.
  - Caso verdadeiro, na linha 12 aquele conjunto de *tokens* é marcado como um termo no arquivo JSON e os dados como sinônimos, descrição, classificação POS, são atribuídos a ele. Na linha 13, o primeiro *token* passará a ser o *token* seguinte ao último *token* daquele termo encontrado. Por último voltar à linha 09 até terminar os *tokens* da sentença. Novos conjuntos passarão a ser formados.
  - Caso contrário (instrução **senão** na linha 14), na linha 15 a variável **point_tokens** é incrementada e o algoritmo volta à linha 09 até terminarem os *tokens* da sentença. O *token* inicial passará a ser o seguinte àquele considerado na última iteração:

```
Sentença: [With]1 [endothelial]2 [dysfunction]3 [and]4 [vascular]5 [injury]6 [,]7
[the]8 [levels]9 [of]10 [endothelial]11 [bound]12 [and]13 [soluble]14 [adhesion]15
                        [molecules]16 [increase]17 [.]18
```

```
Na segunda iteração do laço, o token inicial será: [endothelial]2

Futuro Conjunto 1: endothelial dysfunction and vascular injury , the levels of
endothelial
Futuro Conjunto 2: endothelial dysfunction and vascular injury , the levels of
Futuro Conjunto 3: endothelial dysfunction and vascular injury , the levels
...
Futuro Conjunto 10: endothelial
```

- Ao término da verificação dos conjuntos de *tokens* da sentença, o processo do passo 2: Extração de Termos é finalizado para aquela sentença. Neste momento, a sentença do exemplo anterior terá as seguintes características:

```
Sentença: With [endothelial dysfunction]1 and [vascular injury]2 , the levels of
endothelial bound and [soluble adhesion molecules]3 increase .

onde: 1, 2 e 3 são termos de domínio.
```

- Na linha 17, a variável **point_tokens** é zerada, para início do processo do passo 3;
- Na linha 18 inicia-se o processo do passo 3: Extração de *Tip Words*, para a sentença atual.
  É iniciado um novo laço contador de *tokens*, que se repete até que todos os *tokens* da sentença atual sejam lidos (instrução da linha 26);

- Na linha 19, é executado novamente a montagem de um conjunto de *tokens*. O conjunto é montado, tomando-se como base o primeiro *token* da sentença. Esse processo é semelhando àquele apresentado na linha 09;
- Após a montagem dos conjuntos de *tokens*, na linha 20 eles são comparados com o dicionário de *Tip Words* e seus sinônimos. O objetivo é verificar se os conjuntos existem como um *tip word* do dicionário. O processo começa pelo conjunto com maior número de *tokens*.
- Na linha 21, é feita uma verificação se alguma *tip word* foi encontrada naquele conjunto de tokens montado;
  - Caso verdadeiro, na linha 22 aquele conjunto de *tokens* é marcado como uma *tip word* no arquivo JSON e os dados como significado e categoria, são atribuídos a ele. Na linha 23 O primeiro *token* passará a ser o *token* seguinte ao último *token* daquela *tip word* encontrada. Por último voltar à linha 19 até terminar os *tokens* da sentença. Novos conjuntos passarão a ser formados.
  - Caso contrário (instrução **senão** na linha 24), na linha 25 a variável **point_tokens** é incrementada e o algoritmo volta à linha 19 até terminar os *tokens* da sentença. O *token* inicial passará a ser o seguinte àquele considerado na última iteração, de forma semelhante ao laço anterior.
- Por último, na linha 27, incrementa-se a variável **point_sentenca**.
  Neste momento, se existir outra sentença a ser lida, volta-se à linha 07 e o processo continua.
  Se não existir mais sentenças a serem lidas (instrução **até** na linha 28), segue-se para a linha 29 e o processo é finalizado.
  O exemplo a seguir mostra como os termos de uma sentença ficam identificados:

  `Sentença: With [endothelial dysfunction]1 and [vascular injury]2 , the levels of endothelial bound and [soluble adhesion molecules]3 [increase]4 .`

  `onde: 1, 2 e 3 são termos. 4 é uma` *`tip word`*`.`

4.6.3 Etapa 3: Identificação de Relacionamentos Semânticos do tipo “Causa e Efeito”

Na Etapa 2 foi realizada a extração de termos do domínio biomédico e *tip words*. A partir da Etapa 3, Identificação de Relacionamentos Semânticos do tipo “Causa e Efeito”, são aplicadas algumas técnicas que selecionam sentenças que contém relacionamentos semânticos de causalidade e, ainda, realizam a extração dessas relações. Esta etapa é aquela em que esta pesquisa de mestrado possui maior foco.

A Figura 26 destaca os passos realizados na Etapa 3. Passo ou Fase 1: Entrada de sentenças com termos e *tip words* anotadas. Passo ou Fase 2: Seleção de sentenças com relacionamentos de causalidade. Passo ou Fase 3: Extração de relações semânticas de causalidade. Nessas fases, um conjunto de filtros é aplicado para auxiliar na seleção das sentenças e uma abordagem baseada em regras realiza a extração das relações semânticas.

No estudo dos artigos científicos da área biomédica, descrito na Seção 4.1, foi verificado empiricamente que as sentenças que possuem relações de causalidade, geralmente, apresentam duas características:

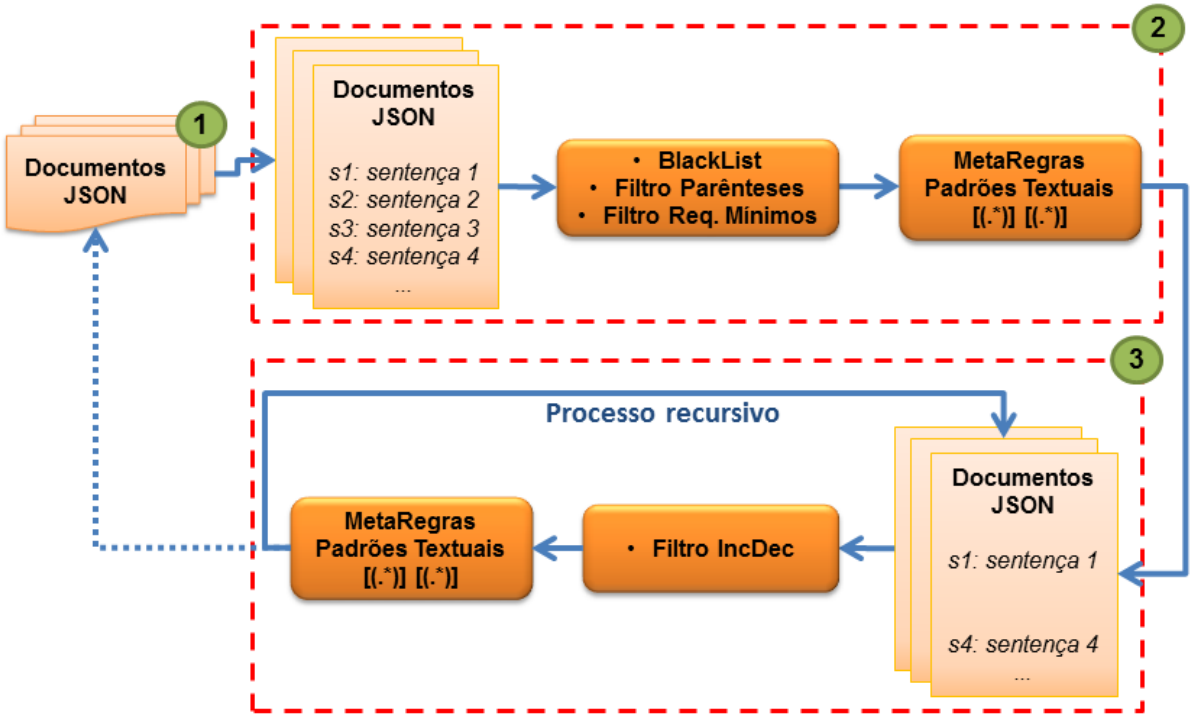

Figura 26: Etapa 3: Identificação de Relacionamentos Semânticos do tipo "Causa e Efeito".

- Possuem ao menos dois termos de domínio e uma *tip word* cujo significado é uma associação, ou
- Possuem ao menos dois termos de domínio e, pelo menos uma *tip word* cujo significado é aumento (*Increase*) ou diminuição (*Decrease*).

Exemplo:

`Associação: Levels of [soluble endothelium-derived adhesion molecules]1 in patients with [sickle cell disease]2 are [associated with]3 [pulmonary hypertension]4 , [organ dysfunction]5 , and [mortality]6 .`

`onde: 1, 2, 4, 5 e 6 são termos de domínio e 3 é uma` *`tip word`* `com significado associação.`

`Increase/Decrease``: With [endothelial dysfunction]1 and [vascular injury]2 , the levels of endothelial bound and [soluble adhesion molecules]3 [increase]4 .`

`onde: 1, 2 e 3 são termos de domínio e 4 é uma` *`tip word`* `com significado aumento.`

Existem casos que possuem relações de associação e *increase/decrease* em uma única construção.

Exemplo:

```
[E-selectin]1 and [P-selectin]2 [induce]3 [specific inflammatory cells]4 to slowly
roll across the [endothelial surface]5 , until firm adhesive interactions [develop
on]6 [endothelial vascular cell adhesion molecule-1]7 .

onde:  1 e 2 possuem relação de associação com 4 e
       5 possui relação de increase/decrease com 7 .
```

A **fase 1** consiste na entrada dos dados. Os dados são as sentenças com termos e *tip words* anotadas dentro de arquivos que representam um artigo científico. Esses arquivos com sentenças anotadas foram produzidos na Etapa 2 do método de extração de relações semânticas.

A **fase 2** irá selecionar sentenças que possuem algum tipo de relacionamento de causalidade. Para isso, há a necessidade da construção de um novo dicionário, denominado *Blacklist*. Nas **fase 2** e **fase 3** serão utilizadas também as MetaRegras para identificação das sentenças. A definição da *Blacklist* e MetaRegras serão apresentados a seguir.

Blacklist

A *Blacklist* consiste em um conjunto expressões que devem ser eliminadas das sentenças na fase de seleção e na fase de extração de relacionamentos. Tais expressões prejudicam a qualidade da seleção pelos padrões textuais pelo fato de possuírem termos de domínio e *tip words* mas não serem úteis ou participarem de um relacionamento semântico. Neste caso, os termos que constituem essas expressões não se assemelham aos padrões textuais dos relacionamentos semânticos.

Um exemplo de expressão contida na *Blacklist* é:

```
Antes da Blacklist:

Sentença: Levels of [soluble endothelium-derived adhesion molecules] in patients
with [sickle cell disease] are [associated with] [pulmonary hypertension] , [organ
dysfunction] , and [mortality] .

Expressão a ser removida: in patients with [sickle cell disease]


Após a Blacklist:

Sentença: Levels of [soluble endothelium-derived adhesion molecules] are [associated
with] [pulmonary hypertension] , [organ dysfunction] , and [mortality] .
```

Detalhes sobre a estrutura da *Blacklist* construída neste trabalho pode ser verificada no Apêndice H.

MetaRegras

As MetaRegras constituem-se pelos padrões textuais. São chamadas assim, pois são construídas com base na estrutura de cada uma das sentenças. Utilizando partes fixas e partes dinâmicas, elas agregam os termos pertencentes à sentenças, as categorias desses termos e, ainda, as *tip words*.

A partir do *Corpus do Estudo Piloto* dois tipos de MetaRegras foram construídos, a MetaRegra de Associação (*Association*) e a MetaRegra *Increase/Decrease*. A MetaRegra de Associação possui

a função de identificar uma sentença como **Associação** e a MetaRegra *Increase/Decrease* de identificar uma sentença como ***Increase/Decrease***. Outra função é dividir a sentença em grupos principais referente às partes que se associam.

A seguir pode-se verificar um exemplo de duas sentenças e da aplicação das MetaRegras para cada sentença.

MetaRegra de Associação:

```
Sentença: Levels of [soluble endothelium-derived adhesion molecules] in
patients with [sickle cell disease] are [associated with] [pulmonary
hypertension] , [organ dysfunction] , and [mortality] .

MetaRegra para esta sentença:
((?:and|or| , | .)?(?:<.*>)(?:<protein>|<disease>|<sca complication>).*)
((?:<associated with>)<tip word>) ((?:and|or| , |.*| .)?(?:<.*>)
(?:<protein>|<disease>|<sca complication>).*)

Grupo 1 (1a parte da associação): [soluble endothelium-derived adhesion
molecules] in are
Grupo 2 (tip word da associação): [associated with]
Grupo 3 (2a parte da associação): [pulmonary hypertension] , [organ
dysfunction] , and [mortality] .
```

MetaRegra *Increase/Decrease*:

```
Sentença: With [endothelial dysfunction] and [vascular injury] , the levels of
endothelial bound and [soluble adhesion molecules] [increase] .

MetaRegra para esta sentença: ((?:and|or| , | .)?(?:<increase><tip word> )?(?:<.*>)
(?:<sca complication>|<protein>)(?: <increase><tip word>)?).*?((?:and|or| , | .)?
(?:<increase><tip word> )?(?:<.*>)(?:<sca complication>|<protein>)(?: <increase><tip word>)?)

Grupo 1 (1a parte da associação): [endothelial dysfunction] and [vascular injury]
Grupo 2 (2a parte da associação): [soluble adhesion molecules] [increase]
```

Detalhes sobre a forma e estrutura das MetaRegras podem ser verificados no Apêndice I. Detalhes sobre a execução das MetaRegras e sobre os grupos serão apresentadas nas seções seguintes, **Algoritmo de Seleção de Sentenças** e **Algoritmo de Extração de Relacionamentos de Causalidade - Fase 3**.

O algoritmo referente à fase 2 é apresentado a seguir.

4.6.3.1 Algoritmo de Seleção de Sentenças - Fase 2

Na **fase 2**, podem ser aplicados quatro filtros, ***Blacklist***, **Filtro de Parênteses**, **Filtro de Requisitos Mínimos** e o **Filtro de Requisitos Mínimos *Increase/Decrease***. A *Blacklist*, como apresentada

anteriormente, irá auxiliar o algoritmo a remover expressões que prejudiquem a qualidade do processo de seleção de sentenças. O Filtro de Parênteses irá remover conteúdos inseridos dentro de parênteses, que normalmente prejudicam a qualidade do processo. O Filtro de Requisitos Mínimos irá eliminar as sentenças que não possuem as características mínimas daquelas com relações de causalidade. O Filtro de Requisitos Mínimos *Increase/Decrease* irá verificar se uma sentença possui os requisitos mínimos para ser do tipo *Increase/Decrease*, ou seja, se possui ao menos uma *tip word* do tipo *increase/decrease* e dois termos de domínio.

Algoritmo 2. Fase 2: Seleção de Sentenças.

```
Entrada: Sentença com termos e tip words anotadas, oriunda de um
processo iterativo de leitura das sentenças do arquivo JSON.
Saída: grupos ou nulo.
Dados: grupos: lista ou array que refere-se aos grupos das MetaRegras;
01  INÍCIO
02     Aplicar Blacklist à sentença;
03     Aplicar Filtro de Parênteses à sentença;
04     Aplicar Filtro de Requisitos Mínimos à sentença;

05     Se (possuir os requisitos mínimos) então
06        grupos ← Aplicar a MetaRegra de Associação à sentença;

07        Se (grupos != vazio) então
08           Retornar grupos;
09        Senão
10        Aplicar o Filtro de Requisitos Mínimos Increase/Decrease à
          sentença;

11        Se (possuir os requisitos mínimos) então
12           grupos ← Aplicar a MetaRegra Increase/Decrease à sentença;
13           Se (grupos != vazio) então
14              Retornar grupos;
15     Retornar nulo;
16  FIM
```

Figura 27: Algoritmo 2: Fase 2 - Seleção de Sentenças.

O algoritmo da Figura 27 é detalhado da seguinte forma:

Entrada: Esta fase se inicia dentro de um processo iterativo. Nesse processo as sentenças do arquivo JSON são lidas e passadas como entrada para o algoritmo.

Saída: Como saída, o algoritmo retorna os grupos extraídos a partir da aplicação das MetaRegras ou nulo, caso a sentença não se aplique a nenhuma MetaRegra.

Variáveis:

- **grupos:** lista ou array que refere-se aos grupos das MetaRegras.

Descrição:

- Na linha 02, aplica-se a ***Blacklist*** à sentença de entrada para remover expressões que possam prejudicar a qualidade da seleção das sentenças;
- Na linha 03, aplica-se o **Filtro de Parênteses** à sentença, que elimina conteúdos dentro de parênteses, pois também prejudicam a qualidade da seleção das sentenças;

Antes do Filtro de Parênteses:

`Sentença: [Expression of] vascular [VCAM-1] , [ICAM-1] and [P-selectin] are [associated with] [sickle retinopathy] ( Kunz et al. , 2002 ).`

`Expressão a ser removida: ( Kunz et al. , 2002 )`

`Após o Filtro de Parênteses:`

`Sentença: [Expression of] vascular [VCAM-1] , [ICAM-1] and [P-selectin] are [associated with] [sickle retinopathy] .`

- Na linha 04 é aplicado o **Filtro de Requisitos Mínimos** à sentença.
  O requisito mínimo para uma sentença é possuir ao menos, uma *tip word* e, ao menos dois termos de domínio;

  `Elimina sentenças do tipo:`

  `Sentença: Our data are consistent with [steady state] levels of [soluble adhesion molecules] as markers of [pulmonary hypertension] and [risk of death] .`

  `Motivo: Não possui nenhuma` *`tip word`*

- Na linha 05 é realizada uma verificação se os requisitos mínimos foram atingidos;
  - Caso verdadeiro, o algoritmo segue para a linha 06.
  - Caso contrário, o algoritmo segue para a linha 15.
- Na linha 06 é aplicada a **MetaRegra de Associação** à sentença. O resultado é armazenado na variável **grupos**;
- Na linha 07 realiza-se a verificação se a regra conseguiu ser a aplicada. A verificação é feita por meio do conteúdo da variável **grupos**. Verifica-se se **grupos** é diferente de vazio;
  - Caso verdadeiro, a regra conseguiu ser a aplicada e significa que a sentença é do tipo Associação. Neste caso, na linha 08 o conteúdo da variável **grupos** é retornado.
  - Caso contrário (instrução da linha 09), segue-se para a linha 10.
- Na linha 10, aplica-se o **Filtro de Requisitos Mínimos *Increase/Decrease*** à sentença;
  O requisito mínimo para uma sentença *Increase/Decrease* é possuir ao menos, uma *tip word* do tipo *Increase/Decrease* e, ao menos dois termos de domínio;
- Na linha 11 realiza-se a verificação se os requisitos mínimos foram atingidos;
  - Caso verdadeiro, a sentença possui os requisitos mínimos *Increase/Decrease*, deve-se seguir para a linha 12.
  - Caso contrário, a sentença não possui os requisitos mínimos *Increase/Decrease*, seguir para a linha 15.
- Na linha 12 é aplicada a **MetaRegra *Increase/Decrease*** à sentença. O resultado é armazenado na variável **grupos**;
- Na linha 13 realiza-se a verificação se a regra conseguiu ser a aplicada. A verificação é feita por meio do conteúdo da variável **grupos**. Verifica-se se **grupos** é diferente de vazio;

  - Caso verdadeiro, a regra conseguiu ser a aplicada e significa que a sentença é do tipo *Increase/Decrease*. Neste caso, na linha 14 o conteúdo da variável **grupos** é retornado.
  - Caso contrário, significa que e a sentença não é do tipo *Increase/Decrease*. Então, seguir para a linha 15.

- Caso o algoritmo atinja a linha 15, significa que a sentença não foi selecionada em nenhuma categoria. Portanto retorna-se o valor nulo.

Em seguida, apresentaremos o Algoritmo de Extração de Relacionamentos de Causalidade que trabalha em conjunto com o algoritmo anterior.

4.6.3.2 Algoritmo de Extração de Relacionamentos de Causalidade - Fase 3

Na **fase 3**, Extração de Relacionamentos de Causalidade, apresentamos o Algoritmo de Extração de Relacionamentos de Causalidade. Este algoritmo trabalha em conjunto com o algoritmo anterior.

Um fato importante de se relembrar é que as sentenças que foram selecionadas como do tipo Associação, podem conter outras relações do tipo *Increase/Decrease*. Em relações selecionadas como do tipo *Increase/Decrease* não existe a possibilidade de encontrar outras relações de associação, devido à forma como os filtros foram aplicados.

Algoritmo 3. Fase 3: Extração de Relacionamentos.

extraiRelacionamentos(grupos)

```
Entrada: grupos (partes) extraídos das sentenças selecionadas.
Saída: Arquivo JSON com os relacionamentos anotados.
Dados: grupos: lista ou array que refere-se aos grupos das MetaRegras.
01  INÍCIO
02     Se (grupos for do tipo Associação) então
03        grupos ← Aplicar a MetaRegra de Associação ao grupo;

04        Se (grupos != vazio) então
05           extraiRelacionamentos(grupos);
06        Senão
07           grupos ← Aplicar a MetaRegra de Increase/Decrease ao grupo;

08           Se (grupos != vazio) então
09              extraiRelacionamentos(grupos);
10           Senão
11              Parser que extrai os termos e tip words de cada parte do
                grupo;
12              ligar tip words increase/decrease aos termos;
13     Senão
14        grupos ← Aplicar a MetaRegra de Increase/Decrease ao grupo;
15        Se (grupos != vazio) então
16           extraiRelacionamentos(grupos);
17        Senão
18           Parser que extrai os termos e tip words de cada parte do
             grupo;
19           ligar tip words increase/decrease aos termos;

20     Construir relação semântica cause-effect;
21     Armazenar relação semântica no arquivo JSON;
22  FIM
```

Figura 28: Algoritmo 3: Fase 3 - Extração de Relacionamentos Semânticos.

O algoritmo da Figura 28 é detalhado da seguinte forma:

Entrada: Assim como na fase anterior, esta se inicia dentro de um processo iterativo. Nesse processo as sentenças do arquivo JSON são lidas e passadas como entrada para o algoritmo.
O grupo extraído de cada sentença selecionada no algoritmo 2 também é utilizado como entrada.

Saída: Como saída, o algoritmo retorna o arquivo JSON com os relacionamentos semânticos de causa e efeito anotados nas sentenças .

Variáveis:

- **grupos:** lista ou array que refere-se aos grupos das MetaRegras.

Descrição:

- Na linha 02, realiza-se a verificação se o grupo de entrada pertence a uma sentença do tipo associação;
  - Caso verdadeiro, na linha 03 aplica-se a **MetaRegra de Associação** em cada parte grupo selecionado.
  - Na linha 04, realiza-se a verificação se a regra conseguiu ser a aplicada. A verificação é feita por meio do conteúdo da variável **grupos**. Verifica-se se **grupos** é diferente de vazio;
    - Caso verdadeiro, a regra conseguiu ser a aplicada e significa que existem mais casos de associação na sentença. Na linha 05, é feita a chamada para este mesmo algoritmo, fazendo com que o processo se torne recursivo.
    - Caso contrário (instrução **senão** da linha 06), na linha 07 aplica-se a **MetaRegra *Increase*/*Decrease*** em cada parte grupo selecionado.
    - Na linha 08, realiza-se a verificação se a regra conseguiu ser a aplicada. A verificação é feita por meio do conteúdo da variável **grupos**. Verifica-se se **grupos** é diferente de vazio;
      - Caso verdadeiro, a regra conseguiu ser a aplicada e significa que existem mais casos de *increase/decrease* na sentença. Na linha 09, é feita a chamada para este mesmo algoritmo, fazendo com que o processo se torne recursivo.
      - Caso contrário (instrução **senão** da linha 10), na linha 11 aplica-se um *parser* em cada grupo extraindo apenas o termo de domínio e os *tip words* de associação e *increase/decrease*. Caso o grupo possua mais de um termo, separados por vírgula, palavra *and* ou palavra *or*, os termos serão separados, porém considerados em um mesmo nível na ordem das relações. *Tip Words* que indiquem negações e possibilidades também são extraídas e armazenadas.

        ```
        Grupo: [endothelial dysfunction] and [vascular injury]

        Extração 1: endothelial dysfunction
        Extração 2: vascular injury
        ```

      - Na linha 12, *tip words increase/decrease* são ligadas diretamente aos termos próximos. Em todos os termos extraídos foi padronizada a seguinte ordem: primeiro *Tip Word increase/decrease* em seguida o termo de domínio;

        ```
        Grupo: [soluble adhesion molecules] [increase]

        Extração: increase soluble adhesion molecules
        ```

  - Caso contrário (instrução **senão** da linha 13), na linha 14 aplica-se a MetaRegra *Increase/Decrease* buscando selecionar grupos do tipo *increase/decrease* internamente a estes.
  - Na linha 15, realiza-se a verificação se a regra conseguiu ser a aplicada. A verificação é feita por meio do conteúdo da variável **grupos**. Verifica-se se **grupos** é diferente de vazio;
    * Caso verdadeiro, a regra conseguiu ser a aplicada e significa que existem mais casos de *increase/decrease* na sentença. Na linha 16, é feita a chamada para este mesmo algoritmo, fazendo com que o processo se torne recursivo.
    * Caso contrário (instrução **senão** da linha 17), na linha 18 aplica-se um *parser* em cada grupo extraindo apenas o termo de domínio e os *tip words* de associação e *increase/decrease*. Caso o grupo possua mais de um termo, separados por vírgula, palavra *and* ou palavra *or*, os termos serão separados, porém considerados em um mesmo nível na ordem das relações. *Tip Words* que indiquem negações e possibilidades também são extraídas e armazenadas.
    * Na linha 19, *tip words increase/decrease* são ligadas diretamente aos termos próximos. Em todos os termos extraídos foi padronizada a seguinte ordem: primeiro *Tip Word increase/decrease* em seguida o termo de domínio;

- Na linha 20, constrói-se uma relação semântica entre termo (ou *tip word* + termo) extraído de um grupo com o conjunto semelhante ao do grupo anteriormente armazenado, unindo-os pelo termo de associação, *tip words* que indicam negação e possibilidade, caso existam. Neste caso, o termo armazenado anteriormente será o antecessor na relação e o outro termo o sucessor. Relações que, por algum motivo, ligarem dois termos iguais, serão rejeitadas:

```
Exemplo Associação:

Grupo 1: [soluble endothelium-derived adhesion molecules] in are
Grupo 2: [associated with]
Grupo 3: [pulmonary hypertension] , [organ dysfunction] , and [mortality] .

Relação 1: cause-effect(soluble endothelium-derived adhesion molecules,
associated with, pulmonary hypertension)
Relação 2: cause-effect(soluble endothelium-derived adhesion molecules,
associated with, organ dysfunction)
Relação 2: cause-effect(soluble endothelium-derived adhesion molecules,
associated with, mortality)

----------

Exemplo Increase/Decrease:

Grupo 1: [endothelial dysfunction] and [vascular injury]
Grupo 2: [soluble adhesion molecules] [increase]

Relação 1: cause-effect(endothelial dysfunction, increase soluble adhesion
molecules)
Relação 2: cause-effect(vascular injury, increase soluble adhesion molecules)
```

```
----------

Exemplo de relação inválida:

Grupo 1: [expression] is [modulated by] [nitric oxide] , and in , these
levels are
Grupo 2: [inversely associated with]
Grupo 3: measures of [nitric oxide] bioavailability .

Relação 1: cause-effect(nitric oxide, inversely associated with, nitric oxide)

Problema: nitric oxide associando-se com o próprio termo nitric oxide.

Portanto, RELAÇÃO REJEITADA.
```

- Na linha 21, armazena-se no arquivo JSON os relacionamentos semânticos extraídos.

Sentenças na voz passiva não conseguem ser tratadas por estes algoritmos. Estas sentenças requerem tratamentos mais específicos, tanto na categorização dos termos e *tip words*, quanto na forma como os termos são relacionados.

Testes foram realizados nesta etapa e podem ser visualizados na Seção 5.

4.6.4 Etapa 4: Construção de uma Rede Semântica de Conhecimentos

Na última etapa do método, Etapa 4: Construção de uma Rede Semântica de Conhecimentos, uma rede semântica é formada unindo a maior quantidade possível de relações para formar uma rede de conhecimentos.

A rede é uma estrutura de grafo. Possui informações tanto nos nós quanto nas arestas. Nos nós são armazenadas informações como termo, se há e qual a *tip word increase/decrease* associada, o número e o nome do artigo a qual foi extraída. Os nós possíveis são nós do tipo ***tip word increase/decrease* + termo** ou apenas **termo**. Resumidamente, seguem os mesmo conceitos apresentados na Seção 4.2, no item Relação Semântica.

Na Figura 29 são apresentados os tipos de nós possíveis na rede de conhecimentos.

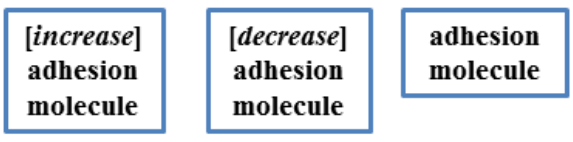

Figura 29: Tipos de nós possíveis na rede de conhecimento.

A aresta liga um nó a outro, indicando que um termo possui relação com outro. Na aresta são armazenadas informações como *tip words association, negation* e *possibility* que fazem parte daquela relação.

Como exemplo, a partir dos artigos do *Corpus* de trabalho, foram extraídas algumas relações de causalidade de quatro artigos diferentes. As relações podem ser verificadas a seguir e os artigos, simbolicamente nomeados de 1 a 4, indicam a origem de onde os relacionamentos foram extraídos:

```
artigo 1: cause-effect(adhesion molecule, associated with, pulmonary hypertension)
artigo 1: cause-effect(adhesion molecule, associated with, organ dysfunction)
artigo 1: cause-effect(adhesion molecule, associated with, mortality)
artigo 2: cause-effect(sickle cell anemia, associated, pulmonary hypertension)
artigo 2: cause-effect(sickle cell anemia, associated, organ dysfunction)
artigo 3: cause-effect(vascular injury, increase adhesion molecule)
artigo 3: cause-effect(endothelial dysfunction, increase adhesion molecule)
artigo 4: cause-effect(increase adhesion molecule, correlated with severity of, mortality)
```

Podemos observar que foram extraídas oito relações. As relações oriundas do artigo 1, possuem apenas um relacionamento de associação representado pelo *tip word associated with*, assim como no artigo 2 que possui apenas um relacionamento de associação representado pelo *tip word associated*. Já no artigo 3, foi encontrada uma relação do tipo *increase/decrease* ligada ao termo *adhesion molecule*. Por último, no artigo 4, encontramos um relacionamento de associação, representado pelo *tip word correlated with severity of*, seguido pelo termo *increase*.

Podemos verificar também, que os termos de cada relacionamento podem se ligar uns aos outros tecendo a rede semântica de conhecimento. Na Figura 30 é possível visualizar essa construção.

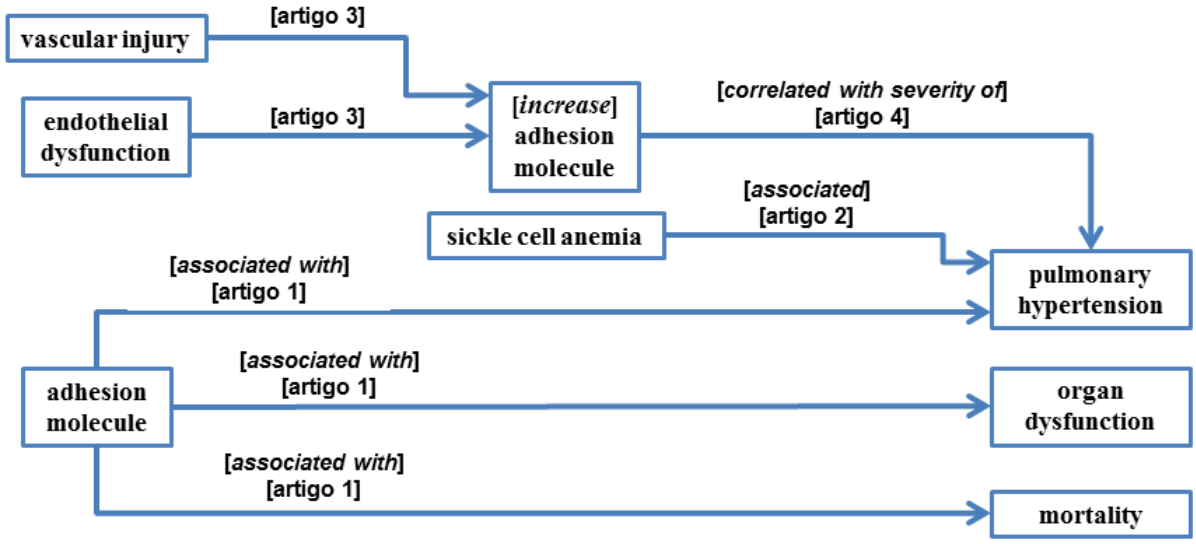

Figura 30: Etapa 4: Construção de uma Rede Semântica de Conhecimentos.

A ideia principal que envolve a construção da rede semântica de conhecimentos é que, por meio de relacionamentos parciais como A => B e B => C, podemos encontrar uma cadeia de relações, como A => B => C e concluir que A => C. Com isso, os especialistas na área biomédica podem sugerir novas hipóteses de tratamentos, por meio de relacionamentos que antes poderiam ser desconhecidos sem uma leitura minuciosa de vários artigos científicos que envolvem esse conhecimento.

4.7 Considerações Finais

Ao longo do Capítulo 4, Método Proposto, foi possível conhecer o método desenvolvido para extração de relacionamentos semânticos do tipo “causa e efeito” em artigos científicos do domínio biomédico.

Na Seção 4.1, Estudo Piloto, foi apresentado o estudo realizado previamente ao exame de qualificação, cujo objetivo foi o entendimento mais aprofundado do problema em questão e das necessidades que os pesquisadores da área biomédica possuem quando enfrentam esse tipo de problema.

No estudo, foram lidos alguns artigos da área biomédica, relacionados às complicações, específicos da Anemia Falciforme. Foram extraídas manualmente sentenças que continham relações de causalidade e, em seguida, tais relações foram associadas entre si formando pequenas redes de conhecimento.

Na Seção 4.2, Definições, foram apresentados alguns conceitos necessários para o entendimento do trabalho de pesquisa, tais como: ***Token***, ***Termo***, ***Tip Word*** e **Relação Semântica**. Redefinimos a representação semântica adotada neste trabalho de pesquisa, apresentando as principais diferenças da representação usualmente adotada na literatura.

Na Seção 4.3, Recursos, foram apresentados os *corpora* utilizados tanto no estudo piloto quanto nos testes (apresentados no Capítulo 5) executados com o método proposto. Além disso, destaca-se também a utilização de dados oriundos de ontologias da área biomédica que foram utilizados para a construção dos dicionários de termos de domínio.

Na Seção 4.4, Anotação do *Corpus* de Trabalho, foi apresentada a forma na qual os textos foram anotados.

Na Seção 4.5, Ferramentas, foram apresentadas as ferramentas desenvolvidas durante o trabalho, JPdf2JSON e ARS. Apresentamos também os *softwares* e as bibliotecas utilizadas, como etiquetador *Stanford POS Tagger*. Por último, foram citados os *scripts* de testes desenvolvidos para geração dos resultados.

Na Seção(4.6), Arquitetura do Método Proposto, foi apresentado o método de extração de relacionamentos semânticos de causalidade, composto por 4 etapas principais: 1. Entrada e Preparação de Dados, 2. Extração de Termos, 3. Identificação de Relacionamentos Semânticos do tipo “causa e efeito” e 4. Construção de uma Rede Semântica de Conhecimentos.

A seguir são apresentados os métodos de validação e testes do método proposto, bem como os resultados conquistados e a validação das hipóteses levantadas no início dos trabalhos.

Capítulo 5

VALIDAÇÃO E TESTES

*Este capítulo apresenta o processo de **validação** do método proposto, além da **validação das hipóteses**.*

A validação do método de extração de relacionamentos semânticos do tipo “causa e efeito” em artigos científicos do domínio biomédico foi realizada em algumas etapas. Nessas etapas comparamos a utilização do método proposto com outros métodos já existentes e aplicados em situações semelhantes. Além disso, os resultados obtidos pela aplicação do método proposto foram comparados com os resultados obtidos de forma manual pelos humanos.

As etapas executadas durante a fase de testes estão divididas entre (1) comparação com a anotação manual, (2) a implementação do algoritmo *PolySeach*, (3) a implementação do algoritmo de Girju e Moldovan (2002), (4) e experimentos utilizando técnicas de aprendizado de máquina.

5.1 Anotação manual

Como apresentado na Seção 4.4, foi realizada a anotação de um *corpus* para testes, chamado na Seção 4.3 de *Corpus* de Trabalho. Possui um conjunto de 5 artigos que tratam do problema da hipertensão pulmonar no contexto da AF. Possui outro conjunto de 5 artigos que também tratam do problema da hipertensão pulmonar em outras doenças distintas. A anotação manual foi validada pelo especialista do domínio para garantir melhor qualidade dos resultados.

O foco principal do trabalho está na etapa de extração de relacionamentos semânticos, Seção 4.6.3, cujo método é constituído por duas fases. Ambas foram avaliadas e seus resultados serão apresentados nesta seção.

A primeira fase consiste na seleção de sentenças que possuem relacionamentos semânticos do tipo “causa e efeito”. A segunda fase corresponde à extração dos relacionamentos a partir das sentenças selecionadas na primeira fase. Na avaliação de ambas as fases foram utilizadas métricas padrão amplamente difundidas na literatura, a saber: precisão, cobertura e medida-F.

Com as ferramentas JPdf2JSON e ARS, foram produzidos e anotados dois conjuntos de dados utilizando o *Corpus* de Trabalho. Esses conjuntos de dados são constituídos de arquivos em formato textual (TXT) com saídas específicas para cada uma das fases e individuais para cada artigo científico analisado. Esses dados foram comparados com outro conjunto semelhante produzido e anotado de forma manual.

Como exemplo do conjunto de dados produzidos na segunda fase, extração dos relacionamentos a partir das sentenças selecionadas na primeira fase, foram gerados manualmente e com a ferramenta ARS um conjunto de arquivos em formato TXT, para cada artigo científico, com sentenças que poderiam possuir relacionamentos de causalidade e os possíveis relacionamentos de causalidade que cada sentença selecionada na primeira fase poderiam possuir. A amostra de uma sentença selecionada pelo sistema pode ser verificada a seguir:

```
SENTENCE 1000 ORIGINAL: Levels of soluble endothelium-derived adhesion molecules
in patients with sickle cell disease are associated with pulmonary hypertension ,
organ dysfunction , and mortality .

SENTENCE 1000 ANNOTATED: levels of <soluble endothelium-derived adhesion
molecules><protein> in are <associated with><tip word> <pulmonary
hypertension><disease> , <organ dysfunction><sca complication> ,  and
<mortality><sca complication> .

REGEX: ((?:and|or| , | .)?(?:<.*>)(?:<protein>|<disease>|<sca complication>).*)
((?:<associated with>)<tip word>) ((?:and|or| , |.*| .)?(?:<.*>)(?:<protein>|
<disease>|<sca complication>).*)

1st PHASE

RESULT 1 : <soluble endothelium-derived adhesion molecules><protein> in are
<associated with><tip word> <pulmonary hypertension><disease> , <organ dysfunction>
<sca complication> , and <mortality><sca complication> .

RESULT 2 : <soluble endothelium-derived adhesion molecules><protein> in are

RESULT 3 : <associated with><tip word>

RESULT 4 : <pulmonary hypertension><disease> , <organ dysfunction>
<sca complication> , and <mortality><sca complication> .

2nd PHASE

RELACAO 1 :: RELACAO VALIDA :: adhesion molecule => pulmonary hypertension

RELACAO 2 :: RELACAO VALIDA :: adhesion molecule => organ dysfunction

RELACAO 3 :: RELACAO VALIDA :: adhesion molecule => mortality

> END SENTENCE
```

No primeiro conjunto foram anotadas manualmente todas as sentenças que possuíam algum tipo de relacionamento e todas os relacionamentos semânticos de causalidade (representando as duas fases da etapa 3 do método). Esse passo foi realizado com a ajuda do especialista do domínio. O

segundo conjunto foi executado utilizando o método proposto nesta pesquisa de mestrado e foram extraídas automaticamente sentenças e os relacionamentos. A partir dos conjuntos, as informações extraídas foram comparadas e contabilizadas.

A Figura 31 destaca o método utilizado para colher os resultados dos testes por meio das métricas de precisão, cobertura e medida-F.

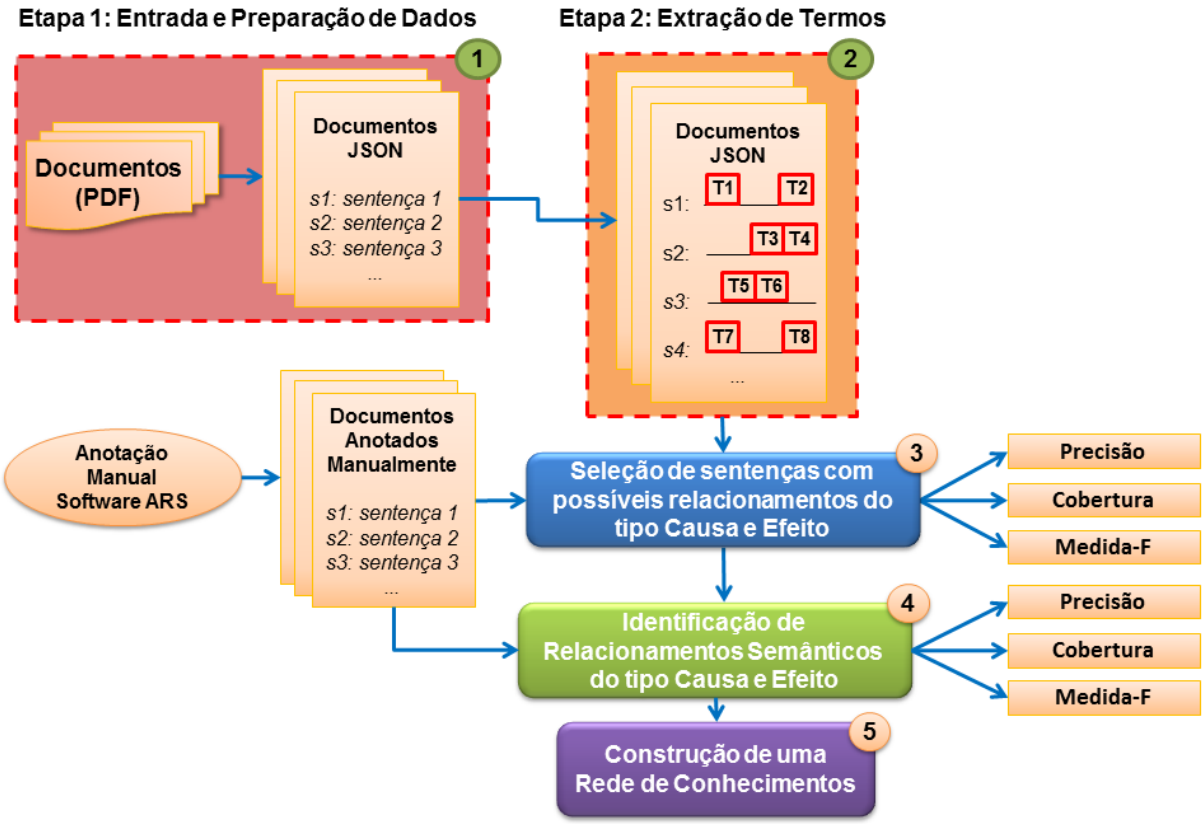

Figura 31: Avaliação das Etapas.

Na etapa 1, os documentos foram convertidos do formato PDF para o formato TXT, foram limpos, e convertidos para JSON. Na etapa 2, por meio dos dicionários, os termos foram extraídos em cada sentença. A etapa 3, avaliada pelas métricas padrão, ocorre a seleção das sentenças com possíveis relacionamentos semânticos do tipo "causa e efeito". Na etapa 4, também avaliada pelas métricas padrão, ocorre a identificação de relacionamentos semânticos do tipo "causa e efeito". Por último, na etapa 5, uma rede semântica de conhecimento é construída utilizando os relacionamentos extraídos na etapa anterior. Esta última etapa não foi avaliada, uma vez que constrói uma forma de representação computacional do conhecimento extraído na etapa 4.

Na avaliação por meio da comparação com a anotação manual do corpus são utilizadas as métricas padrão (Precisão, Cobertura e Medida-F), apresentadas na Seção 2.1.2.4. Para cada conjunto de dados, em cada fase, foram definidos os valores: verdadeiros positivos (VP), falsos positivos (FP), falsos negativos (FN) e verdadeiros negativos (VN).

Em dados gerais, nos 10 artigos selecionados temos:

- Total de sentenças existentes : 1509
- Total de *tokens* existentes : 42999

Na fase 1, seleção de sentenças que possuem relacionamentos de causalidade, temos os resultados aplicados para duas classificações, associação e *increase/decrease*:

- Total de sentenças selecionadas manualmente: 572 sentenças
- Total de sentenças selecionadas automaticamente: 432 sentenças

Associação:

- Total de sentenças anotadas manualmente e que foram encontradas pelo algoritmo (Verdadeiro Positivo) $=$ 411
- Total de sentenças não anotadas manualmente e que foram encontradas pelo algoritmo (Falso Positivo) $=$ 8
- Total de sentenças anotadas manualmente e que não foram encontradas pelo algoritmo (Falso Negativo) $=$ 0
- Total de sentenças não anotadas manualmente e que não foram encontradas pelo algoritmo (Verdadeiro Negativo) $=$ 1090
- Precisão : 98,09 %
- Cobertura : 98,13 %
- Medida-F : 99,03 %

***Increase/Decrease*:**

- Total de sentenças anotadas manualmente e que foram encontradas pelo algoritmo (Verdadeiro Positivo) $=$ 158
- Total de sentenças não anotadas manualmente e que foram encontradas pelo algoritmo (Falso Positivo) $=$ 26
- Total de sentenças anotadas manualmente e que não foram encontradas pelo algoritmo (Falso Negativo) $=$ 3
- Total de sentenças não anotadas manualmente e que não foram encontradas pelo algoritmo (Verdadeiro Negativo) $=$ 1322
- Precisão : 85,86 %
- Cobertura : 98,13 %
- Medida-F : 90,67 %

Na fase 2, extração dos relacionamentos de causalidade, a partir das 432 sentenças selecionadas automaticamente, temos:

- Total de relacionamentos anotados manualmente: 2596 relações
- Total de relacionamentos extraídos automaticamente: 2399 relações

———

- Total de relacionamentos anotados manualmente e encontrados pelo sistema (Verdadeiro Positivo): 2275
- Total de relacionamentos não anotados manualmente e encontrados pelo sistema (Falso Positivo): 124
- Total de relacionamentos anotados manualmente e não encontrados pelo sistema (Falso Negativo): 321
- Total de relacionamentos não anotados manualmente e que não foram encontrados pelo algoritmo (Verdadeiro Negativo) $= 0$
- Precisão : 94,83 %
- Cobertura : 87,63 %
- Medida-F : 91,08 %

Na Seção 5.2 será descrito o experimento executado à partir da implementação do algoritmo PolySearch.

5.2 *PolySearch* (CHENG et al., 2008)

Nessa etapa, serão apresentadas comparações entre os resultados da utilização do modelo proposto neste trabalho e do algoritmo *PolySearch* (PS), apresentado com maiores detalhes na Seção 3.4.

Uma grande dificuldade encontrada nessa fase foi a falta de disponibilização do algoritmo PS para testes. Atualmente, existe um ambiente *Web* com a implementação do algoritmo. Porém, a implementação disponível realiza a busca de dados diretamente em bases de dados *online*, como *PubMed*, não sendo possível adaptá-la para testes nos conjuntos de dados propostos. Devido a esse fato, foi realizada a implementação de uma versão do algoritmo que pudesse ser aplicada no mesmo conjunto de dados utilizado na proposta deste trabalho, o ***Corpus* de Trabalho**.

A implementação foi baseada tanto na descrição do algoritmo apresentada em Cheng et al. (2008), quanto em informações existentes no *site* onde o sistema está hospedado. Algumas modificações foram feitas visando contemplar as necessidades descritas anteriormente, dentre elas a remoção do processo de filtragem dos resultados por palavras-chave. Na versão implementada para testes essa funcionalidade não foi necessária, pois o resultado esperado deveria se basear em todas as combinações de termos possíveis e não apenas nas combinações de um conjunto específico de termos.

Um dicionário de termos foi obtido do *site* do projeto PS e utilizado na identificação dos termos e suas categorias.

Após a implementação, o conjunto de dados foi submetido ao algoritmo e os seguintes dados foram obtidos:

- Total de sentenças existentes: 1.509
- Total de *tokens* existentes : 42.999

Após a aplicação do algoritmos verificou-se a extração dos relacionamentos de causalidade:

- Total de relacionamentos anotados manualmente: 2.596 relações
- Total de relacionamentos extraídos automaticamente com o PS: 96 relações

———

- Total de relacionamentos anotados manualmente e encontrados pelo PS (Verdadeiro Positivo): 27
- Total de relacionamentos não anotados manualmente e encontrados pelo PS (Falso Positivo): 69
- Total de relacionamentos anotados manualmente e não encontrados pelo PS (Falso Negativo): 2.569
- Total de relacionamentos não anotados manualmente e que não foram encontrados pelo PS (Verdadeiro Negativo) $= 0$
- Precisão : 28,12 %
- Cobertura : 1,04 %
- Medida-F : 2,01 %

Os resultados alcançados atingem valores baixos para as medidas utilizadas, quando comparamos com os resultados conquistados pelos autores em Cheng et al. (2008). Após uma análise sobre o experimento realizado, pode-se verificar que esses valores aconteceram devido a um conjunto de fatores:

- Os testes foram realizados utilizando um conjunto de documentos mais específicos sobre um assunto e sobre uma área;
- As modificações citadas anteriormente sobre a implementação do algoritmo, necessárias para a adaptação do sistema para a recepção de novos conjuntos de documentos, podem ter influenciado o resultado dos testes;
- Algum detalhe sobre a implementação do algoritmo pode não ter sido bem interpretado;
- Os dicionários de termos disponibilizados no *site* do projeto PS pode não estar suficientemente completo, não sendo possível identificar termos e, consequentemente, reconhecer os relacionamentos;
- O dicionário de relacionamentos disponibilizados no *site* do projeto PS podem estar limitando a localização dos relacionamentos.

Na Seção 5.3 serão apresentados os experimentos executados utilizando técnicas de Aprendizado de Máquina.

5.3 Aprendizado de Máquina (AM)

O Aprendizado de Máquina (AM) é uma subárea da inteligência artificial que trabalha com algoritmos que aprendem com experiências existentes, para realizarem uma determinada tarefa. As experiências são exemplos de situações que o algoritmo deve aprender. De maneira geral, quanto mais exemplos são apresentados ao algoritmo, melhor ele aprende uma tarefa.

Na implementação de modelos utilizando a tecnologia de Aprendizado de Máquina (AM) são definidas algumas informações:

- **Instâncias**: A instância representa o tipo de dado que será fornecido ao algoritmo e, a partir dela, deseja-se obter uma resposta. Nestes experimentos, as instâncias serão representadas por pares de termos existentes em uma sentença. Nos conjuntos de dados de treinamento, forneceremos aos algoritmos todas as combinações de termos possíveis dentro de uma sentença.

  Por exemplo:

  Sentença: With endothelial dysfunction and vascular injury , the levels of endothelial bound and soluble adhesion molecules increase .

  Termos encontrados:
  1. endothelial dysfunction;
  2. vascular injury;
  3. soluble adhesion molecules;
  4. increase.

  Instâncias geradas:
  1. endothelial dysfunction e vascular injury;
  2. endothelial dysfunction e soluble adhesion molecules;
  3. endothelial dysfunction e increase;
  4. vascular injury e soluble adhesion molecules;
  5. vascular injury e increase.

- **Classe**: A classe representa o tipo de informação que pretende-se obter após executarmos as técnicas de AM. Nos experimentos que serão apresentados a seguir, fixamos a classe como um valor binário que represente que uma instância (par de termos) forma uma relação de causa e efeito.

  Considerando a mesma sentença e as mesmas instâncias do exemplo anterior, temos:

  1. Instância: endothelial dysfunction e vascular injury;
  Classe: É um relacionamento do tipo causa e efeito.

  2. Instância: endothelial dysfunction e soluble adhesion molecules;
  Classe: Não é um relacionamento do tipo causa e efeito.

  3. Instância: endothelial dysfunction e increase;
  Classe: Não é um relacionamento do tipo causa e efeito.

  4. Instância: vascular injury e soluble adhesion molecules;
  Classe: É um relacionamento do tipo causa e efeito.

  5. Instância: vascular injury e increase.
  Classe: Não é um relacionamento do tipo causa e efeito.

- **Conjunto de Dados**: O conjunto de dados foi gerado pelo agrupamento de instâncias, características e classes extraídas do *Corpus* de Trabalho, definido na Seção 4.3, anotado manualmente e validado pelo especialista do domínio.

Técnicas baseadas em AM necessitam, no mínimo, de dois conjuntos de dados. O primeiro é utilizado para treinamento dos modelos. O segundo é utilizado para validação do modelo gerado. Como discutido anteriormente, neste trabalho será utilizado apenas um conjunto de dados (*Corpus* de Trabalho). Nesses casos, uma técnica muito conhecida na literatura e que será utilizada é chamada *K-Fold Cross Validation*.

K-Fold Cross Validation é uma técnica de validação cruzada utilizada para avaliar a capacidade de generalização de um modelo, à partir de um conjunto de dados. Seu funcionamento consiste em dividir o conjunto de dados em K partes. Com um subconjunto de $K-1$ partes o algoritmo de AM realiza o treinamento de um modelo e a validação com a parte restante. Em seguida, o algoritmo seleciona outro subconjunto de $K-1$ partes para treinamento e 1 parte para validação. Este procedimento é repetido até que todas as partes foram utilizadas tanto para treinamento quanto para validação. Ao final, gera-se um modelo médio entre todos.

De forma geral, o valor 10 para K é bastante utilizado na literatura como um valor que divide bem o conjunto de dados, gerando resultados mais satisfatórios. Neste caso, a técnica passa a ser mais conhecida como *10-Fold Cross Validation*. Nos experimentos apresentados neste trabalho será utilizada a técnica *10-Fold Cross Validation* como validação dos modelos gerados.

Além das informações e técnicas definidas, existe a necessidade de definir um conjunto de características (*features*) que auxiliam os algoritmos a construírem modelos. A descrição das características escolhidas será detalhada na Seção 5.3.1

5.3.1 Características *(features)*

Na implementação de modelos utilizando a tecnologia de AM foram desenvolvidos um conjunto de características. Tais características são utilizadas pelos algoritmos para treinamento e construção dos modelos. Não existem regras pré-estabelecidas ou modelos para extrair características. A forma e os tipos de características dependem de uma análise de cada conjunto de dados. Portanto, características definidas para artigos científicos médicos e com objetivo de extrair relações semânticas certamente não serão úteis para outros tipos de textos (como textos jornalísticos, por exemplo) ou com objetivo de obter outros tipos de informações.

Nessa fase dos testes são sugeridas algumas características e executados diversos experimentos utilizando técnicas diferentes. As características foram propostas após observações do próprio *corpus*. Em cada técnica são testados alguns parâmetros diferentes. O conjunto de características desenvolvidas nessa etapa será descrito pela Tabela 5.

Na fase de experimentos, descritas a partir da Seção 5.3.3 serão considerados dois formatos para o conjunto de dados. O primeiro será o formato completo, utilizando todas as instâncias geradas, incluindo aquelas que possuem um dos termos do tipo *tip word*. O segundo formato considera-se a remoção das instâncias que possuem um dos termos do tipo *tip word*. Nos experimentos cujas instâncias contendo *tip words* foram removidas, o conjunto de características será definido pela

Tabela 5: Características (*features*) desenvolvidas.

	Característica	**Tipo**
1.	Distância entre termos	Numérica
2.	Número de *tip words* na sentença	Numérica
3.	Existe *tip word* entre o Termo 1 e o Termo 2	Binária
4.	Algum termo da instância é uma *tip word*	Binária
5.	Termo 1 possui um substantivo	Binária
6.	Termo 2 possui um substantivo	Binária
7.	Termo 1 possui um adjetivo	Binária
8.	Termo 2 possui um adjetivo	Binária

Tabela 6. A diferença entre elas está na remoção da característica “Algum termo da instância é uma *tip word*”.

Tabela 6: Características (*features*) desenvolvidas.

	Característica	**Tipo**
1.	Distância entre termos	Numérica
2.	Número de *tip words* na sentença	Numérica
3.	Existe *tip word* entre o Termo 1 e o Termo 2	Binária
4.	Termo 1 possui um substantivo	Binária
5.	Termo 2 possui um substantivo	Binária
6.	Termo 1 possui um adjetivo	Binária
7.	Termo 2 possui um adjetivo	Binária

A partir das características e do conjunto de dados desenvolvidos, foram aplicados alguns algoritmos de AM. Dentre as técnicas existentes foram utilizadas nos experimentos Naive Bayes, Árvore de Decisão e Redes Neurais. Os principais parâmetros referentes à utilização de cada técnica foram modificados e executados com um conjunto variado de valores. A descrição das aplicações e discussão dos resultados obtidos nos experimentos podem ser verificados nas seções seguintes.

Na Seção 5.3.2 será apresentada a análise do conjunto de dados que será aplicado nos experimentos executados utilizando técnicas de Aprendizado de Máquina.

5.3.2 Discussão sobre o Conjunto de Dados

Antes de executar experimentos utilizando as técnicas de AM é importante realizar uma análise dos dados que serão utilizados como treinamento e testes.

Como apresentado na Seção 5.3, nos experimentos a seguir será utilizado o conjunto de dados nomeado como *Corpus* de Trabalho, composto por 10 artigos científicos da área médica. Foram extraídas as instâncias a partir de cada uma das sentenças dos artigos e identificados os valores para as respectivas características, apresentadas na Seção 5.3.1.

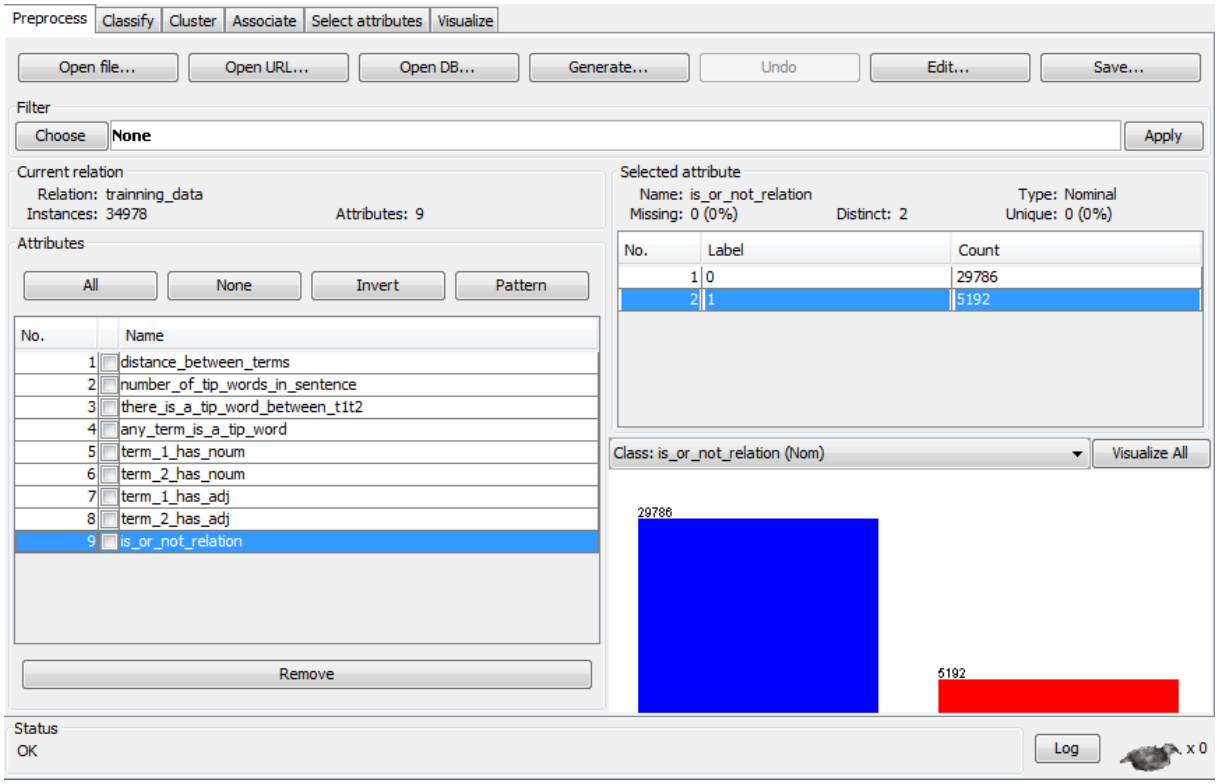

Figura 32: Distribuição de instâncias por classe.

A Figura 32 mostra a distribuição da quantidade de instâncias por classe. Ela apresenta um total de 34.978 instâncias extraídas, considerando pares de termos que continham *tip words*. Desse total, 29.786 instâncias pertencem à classe negativa,"*Não é um relacionamento do tipo causa e efeito*", e 5.192 instâncias pertencem à classe positiva,"*É um relacionamento do tipo causa e efeito*". Nessa situação diz-se que as classes estão desbalanceadas.

Quando as classes estão desbalanceadas, podemos ter um problema na fase treinamento dos modelos. O algoritmo selecionado pode gerar um modelo que aponte com maior frequência os resultados para a classe majoritária, neste caso a classe negativa, ocasionando um conjunto de medidas (precisão, cobertura e medida-F) com valores altos, entretanto, esse modelo não poderá ser considerado satisfatório. Portanto, um modelo gerado por um algoritmo que esteja sendo testado, deve obrigatoriamente ter um desempenho maior do que o modelo que aponte com maior frequência os resultados na classe majoritária.

A Figura 33 apresenta uma relação entre a quantidade de instâncias existentes quando observamos apenas a característica 4, "Algum termo da instância é uma *tip word*". Dos 34.978 instâncias, 20.838 possuem ao menos um tempo do tipo *tip word*. Pode-se observar que essas instâncias certamente não irão gerar resultados na classe positiva e, caso fossem removidas, as instâncias poderiam ficar mais próximas de um balanceamento perfeito.

Ao longo das execuções dos experimentos, serão verificados os resultados utilizando todas as instâncias geradas e os resultados cujas instâncias contendo *tip words* foram removidas.

Nas próximas seções serão apresentados os experimentos executados com a aplicação de técnicas de AM. As técnicas aplicadas são: Naive Bayes, Árvore de Decisão e Redes Neurais.

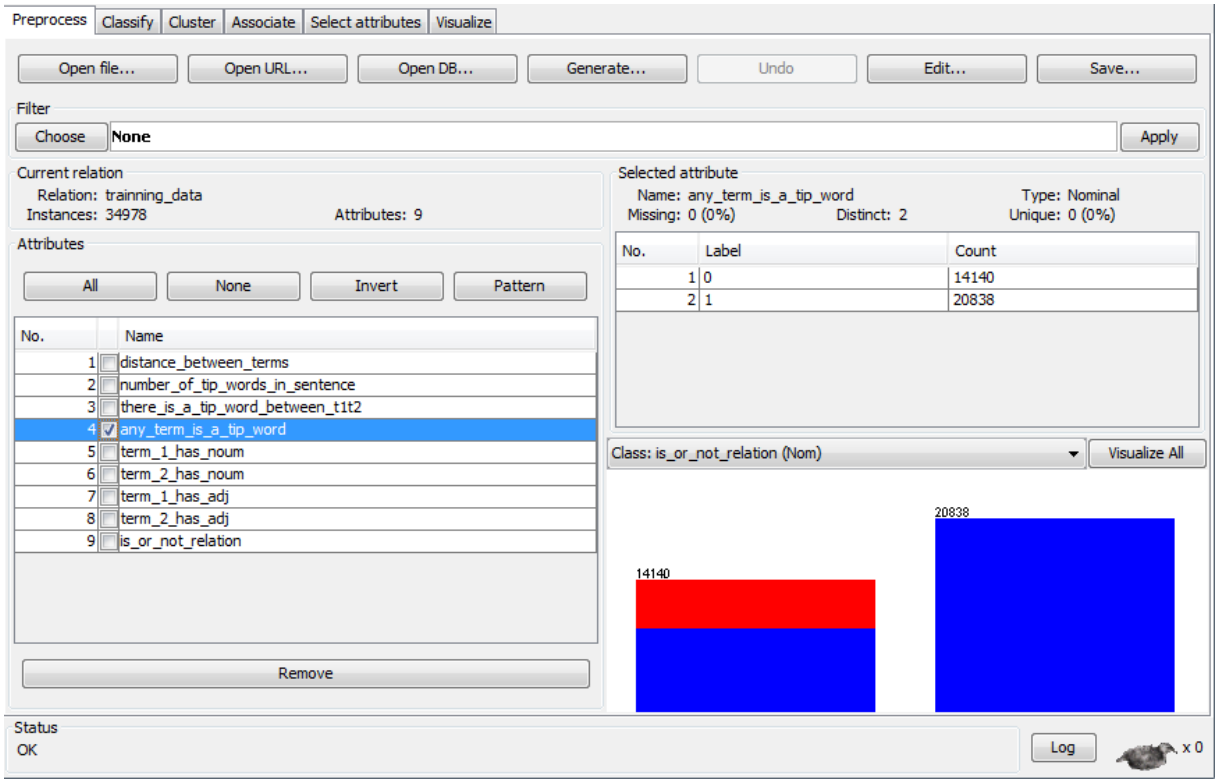

Figura 33: Distribuição de instâncias pela característica 4.

5.3.3 Naive Bayes

Nos primeiros experimentos executados foi utilizada a técnica **Naive Bayes**. É um algoritmo de AM do tipo probabilístico baseado no **Teorema de Bayes** que permite forte independência entre as características.

Na utilização deste algoritmo não são necessários especificar parâmetros. Foi utilizada a técnica *10-Fold Cross Validation* para divisão do conjunto de dados e validação do modelo.

No **Experimento 1** aplicamos o conjunto de dados completo, incluindo instâncias que continham *tip words*. Tomaremos a definição de **classe A**, a classe que representa as instâncias que não são do tipo causa e efeito. **classe B**, a classe que representa as instâncias que são do tipo causa e efeito. Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 34.978
- Total de instâncias na **classe A**: 29.786
- Total de instâncias na **classe B**: 5.192

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 23.518
- Total de instâncias da **classe A** classificadas incorretamente: 1.404

- Precisão : 94,4 %
- Cobertura : 79,2 %
- Medida-F : 86,2 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 3.788
- Total de instâncias da **classe B** classificadas incorretamente: 6.268
- Precisão : 37,7 %
- Cobertura : 73,4 %
- Medida-F : 49,7 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 27.306
- Total de instâncias classificadas incorretamente: 7.672
- Precisão : 86,5 %
- Cobertura : 78,1 %
- Medida-F : 80,6 %

No **Experimento 2** aplicamos o conjunto de dados com a remoção de instâncias nas quais ao menos um dos termos era uma *tip word*. Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 14.140
- Total de instâncias na **classe A**: 8.960
- Total de instâncias na **classe B**: 5.180

———

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 8.233
- Total de instâncias da **classe A** classificadas incorretamente: 727
- Precisão : 64,9 %
- Cobertura : 91,9 %
- Medida-F : 76,1 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 736
- Total de instâncias da **classe B** classificadas incorretamente: 4.444

- Precisão : 50,3 %
- Cobertura : 14,2 %
- Medida-F : 22,2 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 8.969
- Total de instâncias classificadas incorretamente: 5.171
- Precisão : 59,6 %
- Cobertura : 63,4 %
- Medida-F : 56,3 %

Na Seção 5.3.4 serão abordados os experimentos envolvendo algoritmos de Árvore de Decisão.

5.3.4 Árvore de Decisão

Árvore de Decisão é um dos algoritmos mais utilizados em AM devido à sua simplicidade e bons resultados. Seu funcionamento se baseia na criação de árvores em formato de grafos onde cada nó representa uma tomada de decisão sobre uma variável da representação do problema, com as folhas da árvore indicando a classificação de uma dada instância. Essa simplicidade estrutural facilita a leitura de um humano ao observar uma árvore de decisão e compreender seu funcionamento, diferentemente de outros modelos de AM cuja interpretação é mais complexa.

A Figura 34 apresenta um exemplo de árvore de decisão para a pergunta "É um bom dia para jogar tênis?" baseada em duas variáveis (tempo e umidade).

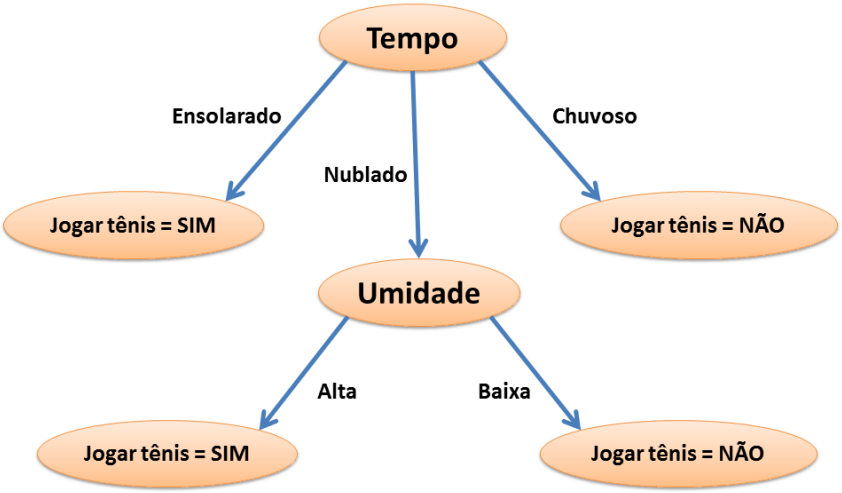

Figura 34: Exemplo de Árvore de Decisão.

O algoritmo de árvore de decisão utilizado neste trabalho foi o C4.5 (QUINLAN, 1993), que baseia a construção das árvores sobre a noção de entropia da teoria da informação. Descreveremos agora cada um dos experimentos realizados utilizando esta técnica. Em todos os experimentos foram utilizadas a técnica *10-Fold Cross Validation* para divisão do conjunto de dados e validação do modelo. Tomaremos

a definição de **classe A**, a classe que representa as instâncias que não são do tipo causa e efeito. **classe B**, a classe que representa as instâncias que são do tipo causa e efeito.

No **Experimento 3** aplicou-se o **conjunto de dados completo** ao algoritmo C4.5. Foi utilizado um tipo de filtro para que as instâncias fossem organizadas de forma aleatória. Nesse experimento selecionamos o parâmetro **não efetua a poda da árvore**.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 34.978
- Total de instâncias na **classe A**: 29.786
- Total de instâncias na **classe B**: 5.192

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 28.583
- Total de instâncias da **classe A** classificadas incorretamente: 1.203
- Precisão : 88,5 %
- Cobertura : 96,3 %
- Medida-F : 92,1 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.477
- Total de instâncias da **classe B** classificadas incorretamente: 3.715
- Precisão : 55,1 %
- Cobertura : 26,1 %
- Medida-F : 35,3 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 30.060
- Total de instâncias classificadas incorretamente: 4.918
- Precisão : 83,3 %
- Cobertura : 85,9 %
- Medida-F : 83,6 %

———

No **Experimento 4** foi aplicado o mesmo teste do Experimento 3, com os mesmos filtros e parâmetros, porém aplicou-se o **conjunto de dados com *tip words* removidas**.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 14.140
- Total de instâncias na **classe A**: 8.960
- Total de instâncias na **classe B**: 5.180

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 7.852
- Total de instâncias da **classe A** classificadas incorretamente: 1.108
- Precisão : 67,2 %
- Cobertura : 87,6 %
- Medida-F : 76,1 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.350
- Total de instâncias da **classe B** classificadas incorretamente: 3.830
- Precisão : 54,9 %
- Cobertura : 26,1 %
- Medida-F : 35,3 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 9.202
- Total de instâncias classificadas incorretamente: 4.938
- Precisão : 62,7 %
- Cobertura : 65,1 %
- Medida-F : 61,2 %

———

No **Experimento 5** aplicou-se o **conjunto de dados completo** ao algoritmo C4.5. Foi utilizado um tipo de filtro para que as instâncias fossem organizadas de forma aleatória. Nesse experimento selecionamos o parâmetro que **permite efetuar podas da árvore**. Outro parâmetro analisado foi o **fator de confiança**= 0,25 , no qual valores mais próximos de 0 indicam ao algoritmo que devem ocorrer mais podas, pois os ramos mais extremos da árvore não trazem confiança).

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 34.978
- Total de instâncias na **classe A**: 29.786
- Total de instâncias na **classe B**: 5.192

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 28.852
- Total de instâncias da **classe A** classificadas incorretamente: 934
- Precisão : 88,2 %
- Cobertura : 96,9 %
- Medida-F : 92,3 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.315
- Total de instâncias da **classe B** classificadas incorretamente: 3.877
- Precisão : 58,5 %
- Cobertura : 25,3 %
- Medida-F : 35,3 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 30.167
- Total de instâncias classificadas incorretamente: 4.811
- Precisão : 83,7 %
- Cobertura : 86,2 %
- Medida-F : 83,8 %

———

No **Experimento 6** foi aplicado o mesmo teste do Experimento 5, com os mesmos filtros e parâmetros, porém aplicou-se o **conjunto de dados com *tip words* removidas**.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 14.140
- Total de instâncias na **classe A**: 8.960
- Total de instâncias na **classe B**: 5.180

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 8.069
- Total de instâncias da **classe A** classificadas incorretamente: 891
- Precisão : 67,2 %
- Cobertura : 90,1 %

- Medida-F : 77,1 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.243
- Total de instâncias da **classe B** classificadas incorretamente: 3.937
- Precisão : 58,2 %
- Cobertura : 23,9 %
- Medida-F : 34,1 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 9.312
- Total de instâncias classificadas incorretamente: 4.828
- Precisão : 63,9 %
- Cobertura : 65,9 %
- Medida-F : 61,2 %

———

No **Experimento 7** aplicou-se o **conjunto de dados completo** ao algoritmo C4.5. Foi utilizado um tipo de filtro para que as instâncias fossem organizadas de forma aleatória. Nesse experimento selecionamos o parâmetro que **permite efetuar podas da árvore** e **fator de confiança** com valores 1.00, 10.00 e 100.00.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 34.978
- Total de instâncias na **classe A**: 29.786
- Total de instâncias na **classe B**: 5.192

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 28.644
- Total de instâncias da **classe A** classificadas incorretamente: 1.142
- Precisão : 88,3 %
- Cobertura : 96,2 %
- Medida-F : 92,1 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.383

- Total de instâncias da **classe B** classificadas incorretamente: 3.809
- Precisão : 54,8 %
- Cobertura : 26,6 %
- Medida-F : 35,8 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 30.027
- Total de instâncias classificadas incorretamente: 4.951
- Precisão : 83,3 %
- Cobertura : 85,8 %
- Medida-F : 83,7 %

———

No **Experimento 8** foi aplicado o mesmo teste do Experimento 7, com os mesmos filtros e parâmetros, porém aplicou-se o **conjunto de dados com *tip words* removidas**.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 14.140
- Total de instâncias na **classe A**: 8.960
- Total de instâncias na **classe B**: 5.180

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 7.851
- Total de instâncias da **classe A** classificadas incorretamente: 1.109
- Precisão : 67,2 %
- Cobertura : 87,6 %
- Medida-F : 76,1 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.354
- Total de instâncias da **classe B** classificadas incorretamente: 3.826
- Precisão : 55,0 %
- Cobertura : 26,1 %
- Medida-F : 35,1 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 9.205
- Total de instâncias classificadas incorretamente: 4.935
- Precisão : 62,7 %
- Cobertura : 65,1 %
- Medida-F : 61,2 %

Na Seção 5.3.5 serão abordados os experimentos envolvendo técnicas de Redes Neurais.

5.3.5 Redes Neurais

Redes Neurais Artificiais (RNAs) são outra subárea da Inteligência Artificial, mais especificamente da Aprendizagem de Máquina, que buscam construir modelos computacionais inspirados no sistema nervoso animal, em particular o cérebro humano. Geralmente, as RNAs são apresentadas como sistemas de neurônios interconectados que podem computar valores de entrada e emitir valores de saída.

Nos experimentos que serão apresentados a seguir será utilizada uma implementação do algoritmo chamado ***Perceptron* Multicamadas** (*MultiLayer Perceptron*). É uma extensão do *Perceptron* de camada única, esta arquitetura apresenta uma camada com unidades de entrada, conectada a uma ou mais unidades intermediárias, chamadas camadas ocultas, e uma camada de unidades de saída. Utiliza o processo de Aprendizado Supervisionado, sendo mais comum a utilização do algoritmo "*Back-propagation*".

Algumas características importantes devem ser ressaltadas:

- As unidades da rede utilizam uma função de ativação não-linear, em geral a função Sigmoide;
- A rede possui uma ou mais camadas ocultas, que lhe permite solucionar problemas complexos, extraindo as características mais significativas dos padrões de entrada;
- A rede possui alto grau de conectividade, o que permite interação entre as unidades.

A Figura 35 apresenta um exemplo gráfico de um *Perceptron* Multicamadas.

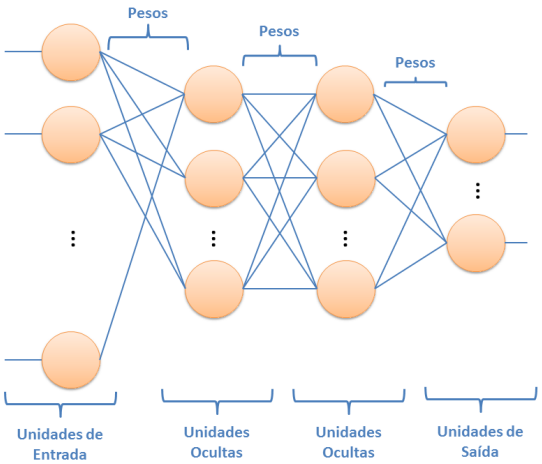

Figura 35: Exemplo gráfico de rede neural do tipo *Perceptron* Multicamadas.

Em todos os experimentos que serão apresentados a seguir foram utilizadas a técnica *10-Fold Cross Validation* para divisão do conjunto de dados e validação do modelo. Tomaremos a definição de **classe A**, a classe que representa as instâncias que não são do tipo causa e efeito. **classe B**, a classe que representa as instâncias que são do tipo causa e efeito.

No **Experimento 9** aplicou-se o **conjunto de dados completo** ao algoritmo *Perceptron* Multicamadas. Foi utilizado um tipo de filtro para que as instâncias fossem organizadas de forma aleatória. Nesse experimento selecionamos os parâmetros:

- Número de camadas ocultas: $\frac{(quantidade\ de\ atributos+classes)}{2}$
- Taxa de aprendizado (quantidade de pesos que são atualizados): 0,3
- Tempo de treinamento (Número de épocas): 500

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 34.978
- Total de instâncias na **classe A**: 29.786
- Total de instâncias na **classe B**: 5.192

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 28.980
- Total de instâncias da **classe A** classificadas incorretamente: 806
- Precisão : 87,7 %
- Cobertura : 97,3 %
- Medida-F : 92,3 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.141
- Total de instâncias da **classe B** classificadas incorretamente: 4.051
- Precisão : 58,6 %
- Cobertura : 22,1 %
- Medida-F : 32,1 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 30.121
- Total de instâncias classificadas incorretamente: 4857
- Precisão : 83,4 %
- Cobertura : 86,1 %
- Medida-F : 83,3 %

No **Experimento 10** foi aplicado o mesmo teste do Experimento 9, com os mesmos filtros e parâmetros, porém aplicou-se o **conjunto de dados com *tip words* removidas**.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 14.140
- Total de instâncias na **classe A**: 8.960
- Total de instâncias na **classe B**: 5.180

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 7.952
- Total de instâncias da **classe A** classificadas incorretamente: 1.008
- Precisão : 67,4 %
- Cobertura : 88,8 %
- Medida-F : 76,6 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.330
- Total de instâncias da **classe B** classificadas incorretamente: 3.850
- Precisão : 56,9 %
- Cobertura : 25,7 %
- Medida-F : 35,4 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 9.282
- Total de instâncias classificadas incorretamente: 4.858
- Precisão : 62,7 %
- Cobertura : 65,1 %
- Medida-F : 61,2 %

No **Experimento 11** aplicou-se o **conjunto de dados completo** ao algoritmo *Perceptron* Multicamadas. Foi utilizado um tipo de filtro para que as instâncias fossem organizadas de forma aleatória. Nesse experimento selecionamos os parâmetros:

- Número de camadas ocultas: $\frac{(quantidade\ de\ atributos+classes)}{2}$

- Taxa de aprendizado (quantidade de pesos que são atualizados): 0,3
- Tempo de treinamento (Número de épocas): 2.000

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 34.978
- Total de instâncias na **classe A**: 29.786
- Total de instâncias na **classe B**: 5.192

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 28.956
- Total de instâncias da **classe A** classificadas incorretamente: 830
- Precisão : 87,8 %
- Cobertura : 97,2 %
- Medida-F : 92,3 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.163
- Total de instâncias da **classe B** classificadas incorretamente: 4.029
- Precisão : 58,4 %
- Cobertura : 22,4 %
- Medida-F : 32,4 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 30.119
- Total de instâncias classificadas incorretamente: 4.859
- Precisão : 83,4 %
- Cobertura : 86,1 %
- Medida-F : 83,3 %

———

No **Experimento 12** foi aplicado o mesmo teste do Experimento 11, com os mesmos filtros e parâmetros, porém aplicou-se o **conjunto de dados com *tip words* removidas**.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 14.140
- Total de instâncias na **classe A**: 8.960

- Total de instâncias na **classe B**: 5.180

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 7.795
- Total de instâncias da **classe A** classificadas incorretamente: 1.165
- Precisão : 67,6 %
- Cobertura : 87,0 %
- Medida-F : 76,1 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.443
- Total de instâncias da **classe B** classificadas incorretamente: 3.737
- Precisão : 55,3 %
- Cobertura : 27,9 %
- Medida-F : 37,1 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 9.238
- Total de instâncias classificadas incorretamente: 4.902
- Precisão : 63,1 %
- Cobertura : 65,3 %
- Medida-F : 61,8 %

———

No **Experimento 13** aplicou-se o **conjunto de dados completo** ao algoritmo *Perceptron* Multicamadas. Foi utilizado um tipo de filtro para que as instâncias fossem organizadas de forma aleatória. Nesse experimento selecionamos os parâmetros:

- Número de camadas ocultas: $\frac{(quantidade\ de\ atributos+classes)}{2}$
- Taxa de aprendizado (quantidade de pesos que são atualizados): 1,0
- Tempo de treinamento (Número de épocas): 1.000

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 34.978
- Total de instâncias na **classe A**: 29.786
- Total de instâncias na **classe B**: 5.192

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 28.193
- Total de instâncias da **classe A** classificadas incorretamente: 1.593
- Precisão : 88,1 %
- Cobertura : 94,7 %
- Medida-F : 91,4 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.400
- Total de instâncias da **classe B** classificadas incorretamente: 3.792
- Precisão : 46,8 %
- Cobertura : 27,0 %
- Medida-F : 34,2 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 29.593
- Total de instâncias classificadas incorretamente: 5.385
- Precisão : 82,1 %
- Cobertura : 84,6 %
- Medida-F : 82,8 %

———

No **Experimento 14** foi aplicado o mesmo teste do Experimento 13, com os mesmos filtros e parâmetros, porém aplicou-se o **conjunto de dados com *tip words* removidas**.

Os resultados podem ser verificados a seguir:

- Total de instâncias existentes: 14.140
- Total de instâncias na **classe A**: 8.960
- Total de instâncias na **classe B**: 5.180

Para a **Classe A:**

- Total de instâncias da **classe A** classificadas corretamente: 7.308
- Total de instâncias da **classe A** classificadas incorretamente: 1.652
- Precisão : 66,8 %
- Cobertura : 81,6 %

- Medida-F : 73,5 %

Para a **Classe B:**

- Total de instâncias da **classe B** classificadas corretamente: 1.550
- Total de instâncias da **classe B** classificadas incorretamente: 3.630
- Precisão : 48,4 %
- Cobertura : 29,9 %
- Medida-F : 37,1 %

Média entre as classes:

- Total de instâncias classificadas corretamente: 8.858
- Total de instâncias classificadas incorretamente: 5.282
- Precisão : 60,1 %
- Cobertura : 62,6 %
- Medida-F : 60,1 %

Na Seção 5.4 serão discutidos os resultados alcançados durante os experimentos apresentados nas seções anteriores.

5.4 Discussão sobre os resultados

A partir dos resultados apresentados nos diversos experimentos realizados podem-se extrair algumas observações e conclusões. Em todos os experimentos foi utilizado como base de comparação a anotação manual do *Corpus* de Trabalho.

A anotação manual é considerada uma base de comparação pois executa a tarefa proposta pelo trabalho, neste caso a identificação de relacionamentos do tipo causa e efeito em artigos científicos médicos, sob a forma mais precisa possível. Teoricamente, o especialista do domínio consegue extrair todos os relacionamentos existentes nos documentos. Porém, em uma quantidade grande de artigos, a ação manual necessita de um tempo muito longo para ser executada.

Para auxiliar o trabalho de anotação e descoberta de conhecimento a partir dos documentos científicos, foi proposta neste trabalho uma solução utilizando técnicas de extração automática de informação. Essa solução utiliza as técnicas de extração baseada em dicionários e regras (padrões textuais). Segundo os experimentos realizados, ela obteve um rendimento alto, com aproximadamente 94 % de precisão e 87 % de cobertura, quando aplicado em um conjunto de artigos selecionados dentro do domínio do problema.

Uma solução existente na literatura para solucionar um problema semelhante àquele proposto neste trabalho é a implementação do sistema PolySearch. Para verificar a eficácia da utilização do sistema e do algoritmo PolySearch na identificação e extração de relacionamento de causa e efeito em textos científicos médicos, foram realizados experimentos aplicando-se o conjunto de documentos *Corpus* de Trabalho.

A principal dificuldade encontrada nos experimentos com o PolySearch foi na disponibilização dos códigos-fonte. Esse fato implicou na necessidade de implementação de uma versão baseada apenas nos algoritmos descritos no artigo e na página *web* do PolySearch. A versão desenvolvida foi submetida ao *Corpus* de Trabalho e atingiu resultados pouco satisfatórios, com 28 % de precisão e 1 % de cobertura.

Segundo a literatura existem outras técnicas que podem ajudar a solucionar um problema de extração de informação. A técnica considerada como estado-da-arte é a utilização de algoritmos de aprendizado de máquina. Sendo assim, foram propostos diversos experimentos na tentativa de gerar modelos utilizando os algoritmos Naive Bayes, Árvore de Decisão e Redes Neurais.

Para os experimentos com técnicas de Aprendizado de Máquina foram definidos um conjunto de características, definidas as classes e extraídos do *Corpus* de Trabalho as instâncias para treinamento. Foi utilizada também a técnica *10-Fold Cross Validation* para validação dos resultados. Cada experimento foi identificado para que seja possível comentá-los com maior precisão. Em cada experimento os resultados foram separados em **Classe A** (classe negativa, aquela cujos termos não formam um relacionamento de causa e efeito) e **Classe B** (classe positiva, aquela cujos termos formam relacionamentos de causa e efeito).

Os experimentos 1 e 2 foram realizados utilizando a técnica baseada em aprendizado estatístico, denominada Naive Bayes. Os experimentos se diferenciaram apenas nas instâncias utilizadas. O experimento 1 utilizou todas as instâncias geradas, incluindo combinações que continham *tip words*. O experimento 2 utilizou pares de instâncias que não continham *tip words*. Em ambos os resultados foram semelhantes, não atingindo valores satisfatórios. Para a Classe B (positiva), a precisão ficou entre 37 % e 50 %, a cobertura entre 14 % e 73 %, sendo que 73 % seria para o experimento 1 com maior quantidade de dados.

Nos experimentos de 5 a 8 foi utilizada a técnica de Árvore de Decisão, mais especificamente o algoritmo C4.5. Nos testes foram consideradas instâncias com e sem presença de *tip words*, além da variação de valores para os parâmetros **podas na árvore** e **fator de confiança**. Os experimentos também não obtiveram resultados satisfatórios. O melhor resultado foi o teste do experimento 8, que considerou exclusão de *tip words* e podas na árvore. Na classe B obteve precisão de 55 % e cobertura de 26 %.

Por último, os experimentos de 9 a 14 utilizaram a técnica de Redes Neurais, mais especificamente *Perceptron* Multicamadas. Nos testes foi efetuada a variação de alguns parâmetros como **número de camadas ocultas**, **taxa de aprendizado** e **tempo de treinamento**. Novamente, os resultados apresentaram valores baixos para a classe de interesse, com precisão entre 46 % e 58 %, cobertura entre 22 % e 29 %.

Em todos os testes utilizando técnicas de aprendizado de máquina, pode-se perceber que mesmo sendo o estado-da-arte na literatura para extração de informação, necessita-se de um conjunto de dados com uma quantidade grande de instâncias, necessita-se também que essas instâncias estejam o mais balanceadas possível. Outra grande necessidade é que o conjunto de características escolhidas deve ser representativo aos dados que deseja-se classificar. A tarefa de identificação das características é a mais complexa e onerosa quando se trabalha com desenvolvimento em aprendizado de máquina. Ao utilizar a abordagem por dicionários e regras para solução do problema proposto nesse trabalho, foram obtidos bons resultados para o domínio aplicado.

Todos os experimentos foram executados em um computador marca Dell, com processador Intel Core i5 dois núcleos de 2.40GHz cada, 4Gb de memória RAM, Sistema Operacional Windows 7 Home

Basic Service Pack 1, 64 bits. Os softwares de apoio utilizados nos experimentos foram Adobe Acrobat Reader (leitura de arquivos PDF), Notepad++ (leitura e edição de arquivos TXT). No desenvolvimento das ferramentas foi utilizada linguagem de programação Java, versão $1.7.0\_07$, 64 bits e a IDE NetBeans versão 7.2. No desenvolvimento dos scripts de leitura dos resultados foi utilizada linguagem de programação Perl, versão $5.16.3$, 32 bits. Nos experimentos com técnicas de Aprendizado de Máquina foi utilizado o software Weka.

5.5 Validação das Hipóteses

Após a realização de testes e apresentação de resultados, podemos validar as hipóteses formuladas no início dos trabalhos:

H1. É possível identificar relações semânticas entre termos do domínio biomédico em um mesmo artigo científico, utilizando técnicas de PLN e Redes Semânticas.

Resposta: Sim. Mesmo já tendo sido alcançado anteriormente por outros autores, com o método proposto podemos apresentar um exemplo extraído do *corpus* de trabalho.

Exemplo:

```
Artigo: Epidemiology and Clinical Management of Pulmonary Hypertension in Children.

Sentença: [Pediatric PH] is not very common , but is a greatly hazardous disease
that [leads to] a [high] [mortality] rate .

Relações: cause-effect(Pediatric PH , leads to, increase mortality)
```

H2. É possível, a partir das relações semânticas $R_1\ (r1 \rightarrow r2 \rightarrow rn)$ e $R_2\ (rn \rightarrow rn+1 \rightarrow rk)$ obtidas de artigos científicos distintos, compor uma cadeia de relações semânticas do tipo $R\ (r1 \rightarrow r2 \rightarrow rn \rightarrow rn+1 \rightarrow rk)$ e, consequentemente, construir uma rede de conhecimentos.

Resposta: Sim. Aplicando-se as regras utilizadas no método e considerando as "tip words" que indicam aumento ou diminuição de um componente biológico, podemos construir uma cadeia de relações e chegar a uma rede de conhecimento.

Exemplo:

```
Artigo 1: Levels of soluble endothelium-derived adhesion molecules in patients with
sickle cell disease are associated with pulmonary hypertension, organ dysfunction, and
mortality.

Sentença: Thus , [excessive] [endothelial activation] and [vaso-constriction because] of
[impaired] [NO] bioavailability [may] [contribute to] [vascular instability] in patients
with SCD ( Reiter et al , 2002 ; Reiter & Gladwin , 2003 ; Nath et al , 2004 ) .

Relação r1: cause-effect(decrease nitric oxide (NO) , increase endothelial activation)

----------
```

Artigo 2: Epidemiology and Clinical Management of Pulmonary Hypertension in Children.

Sentença: Given that all [SCD] patients have a marked [shortening of] [red cell life] span , those with especially [severe] [hemolysis] will have particularly [high] [levels of plasma hemoglobin] and thus more [NO] [scavenging] .

Relação r2: cause-effect(increase hemolysis , increase levels of plasma hemoglobin)

Relação r3: cause-effect(increase levels of plasma hemoglobin , decrease nitric oxide (NO))

Cadeia r4 será r2 => r3 => r1: increase hemolysis => increase levels of plasma hemoglobin => decrease nitric oxide (NO) => increase endothelial activation

H3. É possível extrair relações ternárias *cause-effect*(termo1 , termo2 , termo 3), indicando termos que possuem relação de causalidade com dois outros termos na mesma sentença. Por exemplo, <*endothelial dysfunction*> *and* <*end-organ disease*> => <*renal dysfunction*> .

Resposta: Não. Utilizando o método proposto é possível apenas extrair relações binárias de causalidade. No exemplo sugerido na hipótese 3, serão necessários dois relacionamentos para representar a ideia transmitida na sentença.

Exemplo:

<endothelial dysfunction> and <end-organ disease> => <renal dysfunction>

Pelo método proposto extrai-se:

cause-effect(endothelial dysfunction , renal dysfunction) e
cause-effect(end-organ disease , renal dysfunction) .

Capítulo 6

CONCLUSÕES

*Este capítulo apresenta as **conclusões** a respeito desta pesquisa de mestrado, mostrando as principais **contribuições** e possíveis **trabalhos futuros**.*

Neste trabalho foi proposto um método para a extração de relacionamentos semânticos de causa e efeito em artigos científicos do domínio biomédico. Este método busca reduzir o grande trabalho realizado por médicos e pesquisadores da área biomédica na busca, por exemplo, de tratamentos para efeitos negativos de doenças. Neste trabalho foi utilizado o domínio do Projeto Anemia Falciforme (*Sickle Cell Anemia*).

O método foi testado e validado por meio de experimentos que comparam a extração de relacionamentos de causa e efeito realizada de forma automática com a mesma tarefa realizada de forma manual por especialistas das áreas aplicadas. Mesmo não sendo utilizadas as técnicas que compõem o estado-da-arte em extração de relacionamentos semânticos, foi possível atingir resultados bastante satisfatórios.

Além disso, foram realizados experimentos com o algoritmo *PolySearch* que busca encontrar relacionamentos semânticos em domínio biomédico. Experimentos com técnicas de Aprendizado de Máquina apresentaram a dificuldade de conseguir características significativas que permitam representar de forma abrangente *corpus* com documentos científicos biomédicos.

Em seguida são destacadas as principais contribuições do trabalho desenvolvido.

6.1 Contribuições

Em um primeiro momento podemos considerar como grande contribuição ao Projeto SCA a anotação e disponibilização de um *corpus* de trabalho, anotado em formato digital, na qual pode ser utilizado por outros pesquisadores que futuramente trabalharem no projeto. Como apresentado na Seção 4.4, este *corpus* foi anotado em três níveis: o primeiro sendo anotação de termos relacionados à área biomédica, o segundo sendo a anotação de termos, nomeados *tip words* que indicam sentenças com possíveis relacionamentos semânticos de causalidade e, por último, a anotação dos próprios relacionamentos de "causa e efeito".

Em seguida, podemos considerar como contribuição o desenvolvimento de duas ferramentas. A primeira ferramenta, nomeada "JPdf2JSON" possibilita a conversão de documentos em formato PDF para os formatos TXT e JSON, sendo também possível a aplicação de um etiquetador POS. A segunda

ferramenta possibilita ler pacotes, que representam artigos, em formato JSON, contendo sentenças. Permite ainda a anotação manual e extração automática de termos do domínio, além da anotação manual e extração automática de relacionamentos semânticos de causalidade.

A principal contribuição destaca-se com o próprio método proposto para extração de relacionamentos do tipo "causa e efeito", tanto nos textos relacionados com a Anemia Falciforme quanto em textos relacionados a outras doenças. Foi gerado um modelo inicial de rede de conhecimento que demonstra a grande utilidade deste artefato na geração de novas hipóteses de pesquisa sobre tratamentos, por pesquisadores especialistas da área biomédica.

Podemos destacar também a disponibilização de *scripts* para geração de resultados numéricos dos testes. Desenvolvidos em linguagem de programação Perl, eles realizam a contagem dos resultados automaticamente e são de grande importância para a reprodução dos testes aplicados e de continuidade dos trabalhos por novos pesquisadores. Detalhes sobre os *scripts* podem ser visualizados no Apêndice D e Apêndice E.

São disponibilizados nesse trabalho a implementação do algoritmo *PolySearch*, bem como os resultados dos experimentos realizados com técnicas de aprendizado de máquina.

Todo o conteúdo produzido durante o trabalho de pesquisa, nos quais incluem os artigos utilizados, os artigos anotados (com todos os níveis de anotações apresentados), as ferramentas desenvolvidas e os dados dos resultados, estão disponíveis no *website* do **Grupo de Banco de Dados da UFSCar (GBD)** (UFSCAR, 2013), por meio do seguinte endereço: `http://gbd.dc.ufscar.br/sicklecellanemia/`.

Na sequência são apresentados os trabalhos futuros, que poderão incrementar ainda mais este projeto.

6.2 Trabalhos Futuros

O trabalho apresentado visa auxiliar os profissionais da área médica na busca por novas hipóteses de tratamentos para os efeitos das doenças. Dentro do Projeto SCA muitos trabalhos foram realizados e vários métodos propostos para que este objetivo seja alcançado com cada vez mais eficiência.

Dentre as continuações e melhorias para este trabalho, podemos citar:

- Aumento na quantidade de artigos e documentos anotados nos *corpora* existentes no momento. A quantidade de textos anotados no domínio biomédico ainda é pequena perto de *corpora* oriundos de textos jornalísticos.
- A inserção dos métodos de extração de termos biomédicos desenvolvidos por Matos (2010), Duque (2012), além de outros métodos mais eficientes do que a abordagem de extração utilizando dicionários, para extração de diferentes categorias de termos.
- O desenvolvimento de novos conjuntos de características para uma abordagem utilizando técnicas de aprendizado de máquina.
- O desenvolvimento ou implementação de métodos já existentes para extração de outros tipos de relacionamentos semânticos que podem auxiliar ainda mais a obtenção de conhecimentos existentes nos textos. Por exemplo, em Hearst (1992, 1998) extraindo relacionamentos de hiponímia,

Berland e Charniak (1999), Girju (2003), Girju, Badulescu e Moldovan (2003) extraindo relacionamentos de meronímia.

- A melhoria do método de extração de relacionamentos, diferenciando com maior eficiência, relacionamentos extraídos de sentenças na voz passiva e na voz ativa.

Acredita-se que com as continuações sugeridas, o estado da arte dentro deste domínio será fortemente ampliado trazendo grandes contribuições à área de PLN e à computação aplicada à área biomédica.

REFERÊNCIAS

AGICHTEIN, E.; GRAVANO, L. Snowball : Extracting Relations from Large Plain-Text Collections. 2000.

ANANIADOU, S.; FREIDMAN, C.; TSUJII, J. Introduction: named entity recognition in biomedicine. *Journal of Biomedical Informatics*, v. 37, n. 6, p. 393–395, 2004. Disponível em: <http://dl.acm.org/citation.cfm?id=1053008>.

ANANIADOU, S.; MCNAUGHT, J. *Book Reviews Text Mining for Biology and Biomedicine*. 1. ed. [S.l.: s.n.], 2006. 135–140 p. ISBN 158053984X.

ANANIADOU, S.; NENADIC, G. Automatic terminology management in biomedicine. In: HOUSE, A. (Ed.). *Text mining for biology and biomedicine*. [S.l.: s.n.], 2006. p. 67–98.

ARANHA, C.; PASSOS, E. A Tecnologia de Mineração de Textos. *Revista Elerônica de Sistemas de Informação*, p. 1–8, 2006. Acesso em: 10 out. 2013. Disponível em: <http://www.facecla.com.br/revistas/resi/edicoes/ed8tut01.pdf>.

ARANHA, C. N. *Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português : Sob o Enfoque da Inteligência Computacional*. Tese (Tese de Doutorado em Engenharia Elétrica) — Pontifícia Universidade Católica do Rio de Janeiro, 2007. Disponível em: <http://www.maxwell.lambda.ele.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=10081@1>.

BARD, J.; RHEE, S. Y.; ASHBURNER, M. An ontology for cell types. *Genome biology*, v. 6, n. 2, p. R21, jan. 2005. ISSN 1465-6914. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=551541&tool=pmcentrez&rendertype=abstract>.

BERLAND, M.; CHARNIAK, E. Finding Parts in Very Large Corpora. v. 1910, n. c, p. 57–64, 1999. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.4257>.

BICK, E. THE PARSING SYSTEM "PALAVRAS "Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. Disponível em: <http://visl.sdu.dk/ eckhard/pdf/PLP20-amilo.ps.pdf>.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. [S.l.: s.n.], 2009.

CAMILO, C. O.; SILVA, J. a. C. da. *Mineração de Dados : Conceitos , Tarefas , Métodos e Ferramentas*. [S.l.], 2009. 28 p. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>.

CASELI, H. D. M. *Indução de léxicos bilíngües e regras para a tradução automática*. Tese (Doutorado), 2007. Disponível em: <http://www.nilc.icmc.usp.br/nilc/projects/retratos.htm>.

CGAP. *CGAP SNP500Cancer Database*. 2013. Disponível em: <http://variantgps.nci.nih.gov/cgfseq/pages/home.do;jsessionid=DD79D6669B0F59D911492558DC49D7C1>.

CHENG, D. et al. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic acids research*, v. 36, n. Web Server issue, p. W399–405, jul. 2008. ISSN 1362-4962. Acesso em: 10 out. 2013. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447794&tool=pmcentrez&rendertype=abstract>.

CLIFTON, C. *Definition of Data Mining*. 2010.

COHEN, K. B.; HUNTER, L. Getting started in text mining. *PLoS computational biology*, v. 4, n. 1, p. 1–3, jan. 2008. ISSN 1553-7358. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2217579&tool=pmcentrez&rendertype=abstract>.

CORNEY, D. P. a. et al. BioRAT: extracting biological information from full-length papers. *Bioinformatics (Oxford, England)*, v. 20, n. 17, p. 3206–13, nov. 2004. ISSN 1367-4803. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15231534>.

CTD. *CTD*. 2013. Disponível em: <http://ctdbase.org/downloads/>.

DEGTYARENKO, K. et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, v. 36, n. Database issue, p. D344–50, jan. 2008. ISSN 1362-4962. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238832&tool=pmcentrez&rendertype=abstract>.

DRUGBANK. *Drugbank*. 2013. Disponível em: <http://www.drugbank.ca/>.

DUQUE, J. L. *Um Processo Baseado em Parágrafos para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico Um Processo Baseado em Parágrafos para a Extração de Tratamentos em Artigos Científicos do Domínio Biomédico*. 1–117 p. Tese (Mestrado em Ciência da Computação) — Universidade Federal de São Carlos, 2012. Acesso em: 10 out. 2013.

EGOROV, S. A Simple and Practical Dictionary-based Approach for Identification of Proteins in Medline Abstracts. p. 174–178, 2004. Disponível em: <www.ncbi.nlm.nih.gov/pubmed/14764613>.

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. Acesso em: 10 out. 2013. Disponível em: <http://www.aaai.org/AITopics/assets/PDF/AIMag17-03-2-article.pdf>.

FELDMAN, R.; DAGAN, I. Knowledge Discovery in Textual Databases (KDT). In: PARK, M. (Ed.). *INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD)*. Montréal, Québec: CA: AAAI Press, 1995. p. 112–117. Acesso em: 10 out. 2013. Disponível em: <http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf>.

FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches in analyzing unstructured data.* Cabridge University Press, 2007. 391 p. Acesso em: 10 out. 2013. ISBN 978-0-511-33507-5. Disponível em: <http://wtlab.um.ac.ir/parameters/wtlab/filemanager/E-library/Text Mining/The Text Mining HandBook.pdf>.

FELLBAUM, C. WordNet: An Electronic Lexical Database. *Cambridge, MA: MIT Press*, 1998. Disponível em: <http://wordnet.princeton.edu/wordnet/>.

FREITAS, M. C. D.; QUENTAL, V. Subsídios para a Elaboração Automática de Taxonomias. n. Yarowsky, 2007. Disponível em: <http://www.linguateca.pt/Repositorio/Til07_MCFreitas.pdf>.

GAD. *Genetic Association Database (GAD)*. 2013. Disponível em: <http://geneticassociationdb.nih.gov/>.

GARTEN, Y.; ALTMAN, R. B. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC bioinformatics*, v. 10 Suppl 2, p. S6, jan. 2009. ISSN 1471-2105. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2646239&tool=pmcentrez&rendertype=abstract>.

GIRJU, R. Automatic detection of causal relations for Question Answering. *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering -*, Association for Computational Linguistics, Morristown, NJ, USA, v. 12, p. 76–83, 2003. Disponível em: <http://portal.acm.org/citation.cfm?doid=1119312.1119322>.

GIRJU, R.; BADULESCU, A.; MOLDOVAN, D. Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, Association for Computational Linguistics, Morristown, NJ, USA, v. 1, p. 1–8, 2003. Disponível em: <http://portal.acm.org/citation.cfm?doid=1073445.1073456>.

GIRJU, R.; MOLDOVAN, D. Text Mining for Causal Relations. p. 360–364, 2002. Disponível em: <http://secs.ceas.uc.edu/ mazlack/dbm.w2010/Causal Text Networks/Girju.2002.Text.pdf>.

HALL, M. et al. The WEKA Data Mining Software : An Update. v. 11, n. 1, p. 10–18, 2009.

HAPMAP. *International HapMap Project*. 2013. Disponível em: <http://hapmap.ncbi.nlm.nih.gov/>.

HEARST, M. A. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the 14th International Conference on Computational linguistics*, v. 2, p. 539—-545, 1992. Disponível em: <http://acl.ldc.upenn.edu/C/C92/C92-2082.pdf>.

HEARST, M. A. Automated Discovery of WordNet Relations. *WordNet: An electronic lexical database*, p. 131–151, 1998. Disponível em: <http://www.icst.pku.edu.cn/course/mining/11-12spring/%E5%8F%82%E8%80%83%E6%96%87%E7%8C%AE/13-01 WordNet98.pdf>.

HEARST, M. A.; HALL, S. Untangling Text Data Mining. *Proceeding ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 3–10, 1999.

HGMD. *Human Genome Mutation Database (HGMD)*. 2013. Disponível em: <http://www.hgmd.org/>.

HIRSCHMAN, L. et al. Overview of BioCreAtIvE task 1B : normalized gene lists. v. 10, p. 1–10, 2005.

HIRSCHMAN, L. et al. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics*, v. 6 Suppl 1, p. S1, jan. 2005. ISSN 1471-2105. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1869002&tool=pmcentrez&rendertype=abstract>.

HMDB. *Human Metabolome Database (HMDB)*. 2013. Disponível em: <http://www.hmdb.ca/>.

HOTHO, A. et al. A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, v. 20, p. 1–37, 2005. Disponível em: <http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf>.

HPRD. *Human Protein Reference Database (HPRD)*. 2013. Disponível em: <http://www.hprd.org/>.

JACKSON, P.; MOULINIER, I. *Natural Language Processing for Online Applications: Text retrieval, extraction and categorization.* [S.l.: s.n.], 2002. 223 p.

JENSEN, L. J.; SARIC, J.; BORK, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews. Genetics*, v. 7, n. 2, p. 119–29, fev. 2006. ISSN 1471-0056. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16418747>.

JOACHIMS, T. Making Large-Scale SVM Learning Practical. 1998. Disponível em: <http://www.cs.cornell.edu/People/tj/publications/joachims_99a.pdf>.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. p. 950, 2000.

KOK, S.; DOMINGOS, P. Extracting Semantic Networks from Text Via Relational Clustering. p. 1–16, 2008.

KOU, Z.; COHEN, W. W.; MURPHY, R. F. High-recall protein entity recognition using a dictionary. *Bioinformatics (Oxford, England)*, v. 21 Suppl 1, n. 2002, p. i266–73, jun. 2005. ISSN 1367-4803. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2857312&tool=pmcentrez&rendertype=abstract>.

KRAUTHAMMER, M.; NENADIC, G. Term identification in the biomedical literature. *Journal of biomedical informatics*, v. 37, n. 6, p. 512–526, dez. 2004. ISSN 1532-0464. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1532046404000826>.

KRAUTHAMMER, M. et al. Using BLAST for identifying gene and protein names in journal articles. *Gene*, v. 259, n. 1-2, p. 245–52, dez. 2000. ISSN 0378-1119. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11163982>.

LUO, Q. Advancing Knowledge Discovery and Data Mining. *First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008)*, Ieee, p. 3–5, jan. 2008. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4470338>.

MATOS, P. F. *Metodologia de pré-processamento textual para extração de informação sobre efeitos de doenças em artigos científicos do domínio biomédico*. 1–161 p. Tese (Mestrado em Ciência da Computação) — Universidade Federal de São Carlos, 2010. Acesso em: 10 out. 2013. Disponível em: <http://gbd.dc.ufscar.br/ pablofmatos/files/DissPFM.set2010.pdf>.

MILLER, G. A. WordNet: A Lexical Database for English. *Communications of the ACM*, v. 38, n. 11, p. 39–41, 1995. Disponível em: <http://wordnet.princeton.edu/wordnet/>.

MüLLER, H.-M.; KENNY, E. E.; STERNBERG, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, v. 2, n. 11, p. e309, nov. 2004. ISSN 1545-7885. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=517822&tool=pmcentrez&rendertype=abstract>.

MUNGALL, C. J. et al. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, BioMed Central Ltd, v. 13, n. 1, p. R5, 2012. ISSN 14656906. Disponível em: <http://genomebiology.com/2012/13/1/R5>.

OBO. *The Open Biological and Biomedical Ontologies*. 2013. Disponível em: <http://www.obofoundry.org/>.

OLIVEIRA, H. G. Avaliação da Extração de Relações Semânticas entre palavras portuguesas a partir de um dicionário. 2005. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/stil/2009/002.pdf>.

OMIM. *Online Mendelian Inheritance in Man (OMIM)*. 2013. Disponível em: <http://www.omim.org/>.

ONO, T. et al. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics (Oxford, England)*, v. 17, n. 2, p. 155–61, fev. 2001. ISSN 1367-4803. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11238071>.

OSBORNE, J. D. et al. Annotating the human genome with Disease Ontology. *BMC genomics*, v. 10 Suppl 1, p. S6, jan. 2009. ISSN 1471-2164. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2709267&tool=pmcentrez&rendertype=abstract>.

PARK, J. C.; JUNG-JAE, K. Named Entity Recognition. In: ARTECH, H. B.; ANANIADOU, S.; MCNAUGHT, J. (Ed.). *Text Mining for Biology and Biomedicine*. [S.l.: s.n.], 2006. p. 121–142.

PDFBOX. *PDF Box*. 2013. Disponível em: <http://pdfbox.apache.org/>.

PUBMED. *PubMed*. 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed>.

QUINLAN, J. C4.5: programs for machine learning. *USA: Morgan Kaufmann Publishers Inc.*, 1993.

REBHOLZ-SCHUHMANN, D.; KIRSCH, H.; COUTO, F. Facts from text–is text mining ready to deliver? *PLoS biology*, v. 3, n. 2, p. e65, fev. 2005. ISSN 1545-7885. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=548955&tool=pmcentrez&rendertype=abstract>.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. *Revista de Sistemas de Informacao da FSMA*, v. 7, p. 7–21, 2011.

ROBERTO, P. et al. U SE OF M ACHINE L EARNING TECHNIQUES IN RECOGNITION OF. p. 73–81, 2011. Disponível em: <http://revistas.unibh.br/index.php/dcet/article/view/305/164>.

SANTOS, C. N. dos. Aplicação de Aprendizado Baseado em Transformações na Aplicação de Sintagmas Nominais. *XXV Congresso da Sociedade Brasileira de Computação*, p. 2138–2147, 2005. Disponível em: <http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Trabalho?id=11225>.

SCHUEMIE, M. J. et al. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *Journal of biomedical informatics*, v. 40, n. 3, p. 316–24, jun. 2007. ISSN 1532-0480. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17079192>.

SEKINE, S. Named Entity: History and Future. p. 5, 2004. Disponível em: <http://cs.nyu.edu/ sekine/papers/NEsurvey200402.pdf>.

SETTLES, B. Biomedical named entity recognition using conditional random fields and rich feature sets. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04*, Association for Computational Linguistics, Morristown, NJ, USA, p. 104, 2004. Disponível em: <http://portal.acm.org/citation.cfm?doid=1567594.1567618>.

SILVA_PINTO, A. C. *A hidroxicarbamida atua sobre componentes do metabolismo da adenosina em células sanguíneas de pacientes com anemia falciforme*. 1–98 p. Tese (Tese de Doutorado em Ciências Médicas) — UNIVERSIDADE DE SÃO PAULO, 2011.

SILVA_PINTO, A. C. et al. *Relatório Técnico "Doença Anemia Falciforme"*. [S.l.], 2009. 1–17 p. Disponível em: <http://gbd.dc.ufscar.br/ pablofmatos/files/ReportSCA-PintoEtAl.pdf>.

SNOW, R.; JURAFSKY, D.; NG, A. Y. Learning syntactic patterns for automatic hypernym discovery. 2005.

SNP, E. *Entrez SNP*. 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/projects/SNP/>.

SOWA, J. F. *Semantic Networks*. 2006. 1–32 p. Acesso em: 10 out. 2013. Disponível em: <http://www.jfsowa.com/pubs/semnet.htm>.

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, set. 2005. ISSN 1467-5463. Acesso em: 10 out. 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16212772>.

STAVRIANOU, A.; ANDRITSOS, P.; NICOLOYANNIS, N. Overview and semantic issues of text mining. *ACM SIGMOD Record*, v. 36, n. 3, p. 23, set. 2007. ISSN 01635808. Disponível em: <http://portal.acm.org/citation.cfm?doid=1324185.1324190>.

SWANSON, D. R. Fish Oil, Raynoud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, v. 30, p. 7–18, 1986.

SWANSON, D. R.; SMALHEISER, N. R. An interactive system for finding complementary literatures : a stimulus to scientific discovery. *Artificial Intelligence*, v. 91, p. 183–203, 1997.

SWISS-PROT. *Swiss-Prot*. 2013. Disponível em: <http://web.expasy.org/docs/swiss-prot_guideline.html>.

TABA, L. S. *Extração automática de relações semânticas a partir de textos escritos em português do Brasil*. 92 p. Tese (Doutorado) — Universidade Federal de São Carlos, 2013.

TAN, A.-h. Text Mining : The state of the art and the challenges Concept-based. In: *KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES (KDAD)*. [s.n.], 1999. p. 71–76. Acesso em: 10 out. 2013. Disponível em: <http://www3.ntu.edu.sg/home/asahtan/Papers/tm_pakdd99.pdf>.

TOUTANOVA, K. et al. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL 2003*, p. 252–259, 2003. Disponível em: <http://nlp.stanford.edu/ manning/papers/tagging.pdf>.

TSURUOKA, Y.; TSUJII, J. Improving the performance of dictionary-based approaches in protein name recognition. *Journal of biomedical informatics*, v. 37, n. 6, p. 461–70, dez. 2004. ISSN 1532-0464. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15542019>.

UFSCAR, G. *GBD UFSCar*. 2013. Disponível em: <http://gbd.dc.ufscar.br/site/>.

WU, C. H. The Protein Information Resource. *Nucleic Acids Research*, v. 31, n. 1, p. 345–347, jan. 2003. ISSN 13624962. Disponível em: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg040>.

YANG, Z.; LIN, H.; LI, Y. BioPPISVMExtractor: a protein–protein interaction extractor for biomedical literature using SVM and rich feature sets. *Journal of biomedical informatics*, Elsevier Inc., v. 43, n. 1, p. 88–96, fev. 2010. ISSN 1532-0480. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19706337>.

YANG, Z.; LIN, H.; WU, B. BioPPIExtractor: A protein–protein interaction extraction system for biomedical literature. *Expert Systems with Applications*, Elsevier Ltd, v. 36, n. 2, p. 2228–2233, mar. 2009. ISSN 09574174. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0957417407006410>.

YAP, W.; BALDWIN, T. Experiments on pattern-based relation learning. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, ACM Press, New York, New York, USA, p. 1657, 2009. Disponível em: <http://portal.acm.org/citation.cfm?doid=1645953.1646197>.

YEH, A. et al. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC bioinformatics*, v. 6 Suppl 1, p. S2, jan. 2005. ISSN 1471-2105. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1869012&tool=pmcentrez&rendertype=abstract>.

APÊNDICE A – Estrutura JSON para codificação de sentenças

Este apêndice apresenta a estrutura JSON para codificação de sentenças utilizada pela ferramenta de auxílio à anotação de relações semânticas, a ARS.

O formato JSON se baseia em duas estruturas básicas, objetos e vetores. Objetos (definidos entre "{" e "}") são conjuntos de pares chave:valor, similares a vetores associativos, onde a chave é uma *string*; e vetores (definidos entre "[" e "]") são sequências ordenadas de valores. Valores podem ser *strings*, números, objetos, vetores, `true`/`false` (valores booleanos) ou `null` (valor inexistente). Baseado nessas estruturas, uma sentença é um objeto que contém outros objetos (*tokens*, termos, relações) e alguns campos (id, texto, entre outros).

A seguir é mostrado um exemplo completo de uma sentença codificada nesse formato, retirada dos *corpus* de trabalho.


```
{
    "id": 1004,
    "ignored": false,
    "text": "With endothelial dysfunction and vascular injury , the levels of endothelial
bound and soluble adhesion molecules increase .",
    "relations": [
        { "t2": 2, "t1": 0, "r": "cause-effect" },
       { "t2": 2, "t1": 1, "r": "cause-effect" }
    ],
    "associationTermTipWord": [],
    "terms": [
        { "of": 1, "category": "sca complication", "meaning": "", "until": 2,
"dictionaryTerm": "endothelial dysfunction" },
        { "of": 4, "category": "sca complication", "meaning": "", "until": 5,
"dictionaryTerm": "vascular injury" },
        { "of": 13, "category": "protein", "meaning": "", "until": 15,
"dictionaryTerm": "adhesion molecule" },
        { "of": 16, "category": "tip word", "meaning": "increase", "until": 16,
"dictionaryTerm": "increase" }
    ],
    "annotated": false,
    "tokens": [
        { "t": "With", "sin": null, "l": null, "pos": "IN" },
```

```
        { "t": "endothelial", "sin": null, "l": null, "pos": "JJ" },
        { "t": "dysfunction", "sin": null, "l": null, "pos": "NN" },
        { "t": "and", "sin": null, "l": null, "pos": "CC" },
        { "t": "vascular", "sin": null, "l": null, "pos": "JJ" },
        { "t": "injury", "sin": null, "l": null, "pos": "NN" },
        { "t": ",", "sin": null, "l": null, "pos": "," },
        { "t": "the", "sin": null, "l": null, "pos": "DT" },
        { "t": "levels", "sin": null, "l": null, "pos": "NNS" },
        { "t": "of", "sin": null, "l": null, "pos": "IN" },
        { "t": "endothelial", "sin": null, "l": null, "pos": "JJ" },
        { "t": "bound", "sin": null, "l": null, "pos": "VBN" },
        { "t": "and", "sin": null, "l": null, "pos": "CC" },
        { "t": "soluble", "sin": null, "l": null, "pos": "JJ" },
        { "t": "adhesion", "sin": null, "l": null, "pos": "NN" },
        { "t": "molecules", "sin": null, "l": null, "pos": "NNS" },
        { "t": "increase", "sin": null, "l": null, "pos": "VBP" },
        { "t": ".", "sin": null, "l": null, "pos": "." }
    ],
    "annotators": [
        "Ricardo"
    ],
    "comments": ""
}
```


Essa sentença possui três termos marcados (“*endothelial dysfunction*”, “*vascular injury*” e “*adhesion molecule*”), uma “*tip word*” (“*increase*”) e as relações: *cause-effect(endothelial dysfunction, adhesion molecule)* e *cause-effect(vascular injury, adhesion molecule)*.

As estruturas definidas para a codificação de sentenças, *tokens*, termos e relações são mostradas nas Tabelas 7 a 10, a seguir.

Tabela 7: Campos do objeto JSON que representam uma sentença

Campo	Tipo	Descrição
id	Numérico	Número que identifica a sentença
text	*String*	Texto da sentença
comments	*String*	Comentários adicionados pelos anotadores
annotated,	Booleano	Se a sentença já foi anotada por algum anotador. Utilizada para anotação de dados de treinamento para Aprendizado de Máquina.
ignored	Booleano	Se a sentença foi marcada como ignorada (porque contém algum erro) pelos anotadores. Utilizada para anotação de dados de treinamento para Aprendizado de Máquina.
tokens	Vetor	Vetor de objetos que representam cada *token* da sentença
terms	Vetor	Vetor de objetos que representam os termos marcados na sentença
relations	Vetor	Vetor de objetos que representam as relações semânticas marcadas na sentença
annotators	Vetor	Vetor de *strings* que armazenam o nome dos anotadores que trabalharam na sentença

Tabela 8: Campos do objeto JSON que representam um *token*

Campo	Tipo	Descrição
t	*String*	Forma superficial do *token*
l	*String*	Lema (forma base) do *token*. Não preenchido nesta versão do sistema.
pos	*String*	Etiquetas *part-of-speech* do *token* (cf. anotado pelo Stanfford Log-linear POS Tagger (TOUTANOVA et al., 2003)
sin	*String*	Papel sintático do *token* na árvore de dependências da sentença. Não preenchido nesta versão do sistema.

Tabela 9: Campos do objeto JSON que representam um termo

Campo	Tipo	Descrição
of	Numérico	Índice do *token* (no vetor de *tokens* da sentença) onde o termo se inicia
until	Numérico	Índice do *token* (no vetor de *tokens* da sentença) onde o termo termina
dictionaryTerm	*String*	Nome oficial dado ao termo pelos especialistas
category	*String*	Categoria que o termo foi marcado no dicionário de termos
meaning	*String*	Significado atribuído apenas aos termos "*tip word*".

Tabela 10: Campos do objeto JSON que representam um relacionamento

Campo	Tipo	Descrição
r	*String*	Qual relação semântica está marcada. Nesta versão do sistema teremos apenas "*cause-effect*"
t1	Numérico	Índice do termo que é o primeiro participante da relação
t2	Numérico	Índice do termo que é o segundo participante da relação

APÊNDICE B – Telas da Ferramenta JPdf2JSON

Abaixo apresentamos as telas da ferramenta JPdf2JSON:

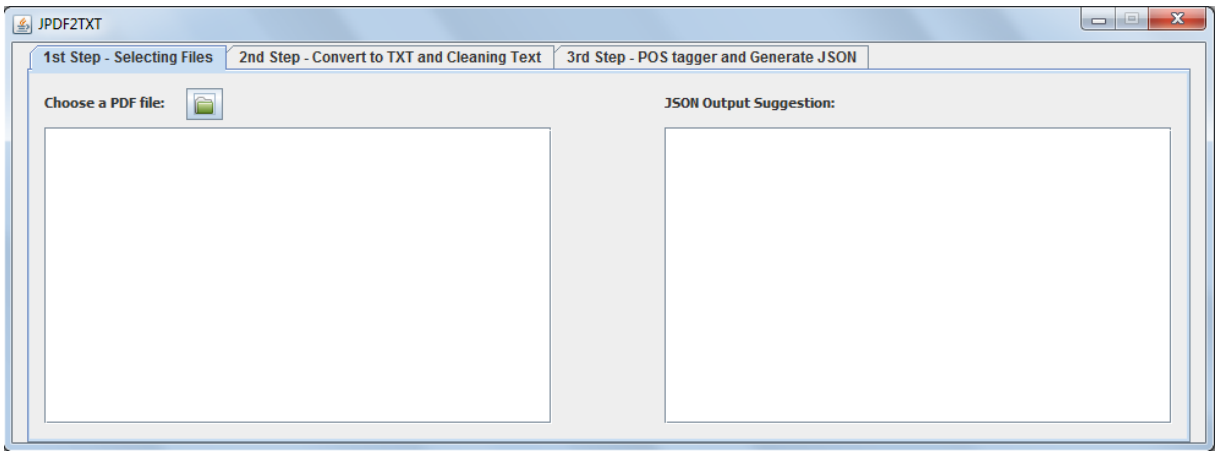

Figura 36: Passo 1: Seleção de arquivos PDF.

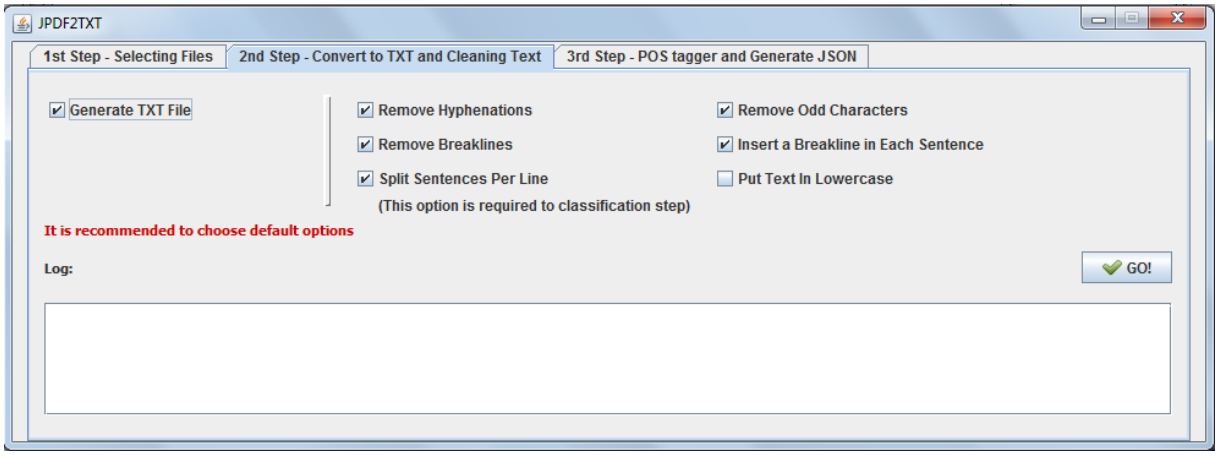

Figura 37: Passo 2: Conversão para TXT e Limpeza do texto.

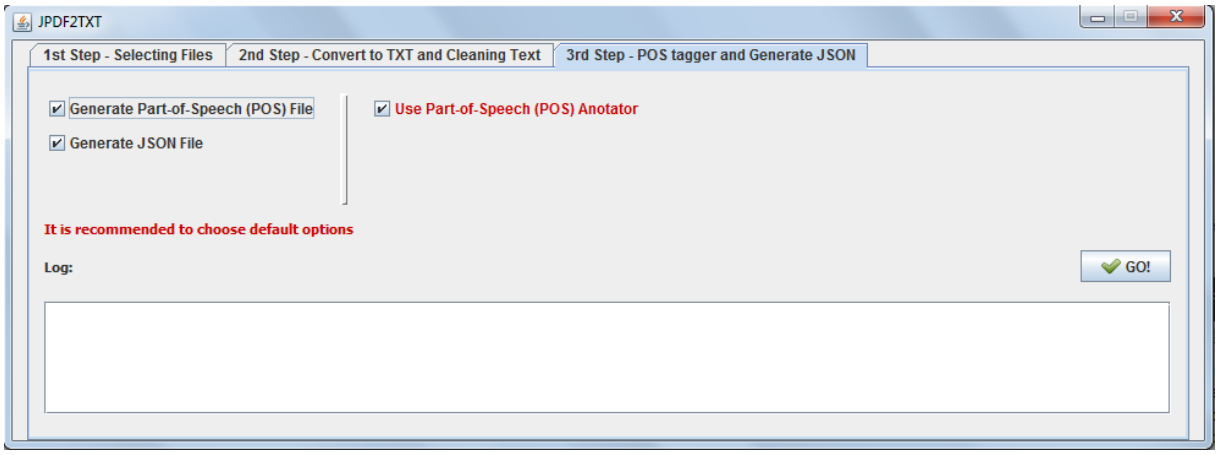

Figura 38: Passo 3: POS *tagger* e Geração do arquivo JSON.

APÊNDICE C – Tela Principal da Ferramenta ARS

Abaixo apresentamos a tela principal da ferramenta ARS:

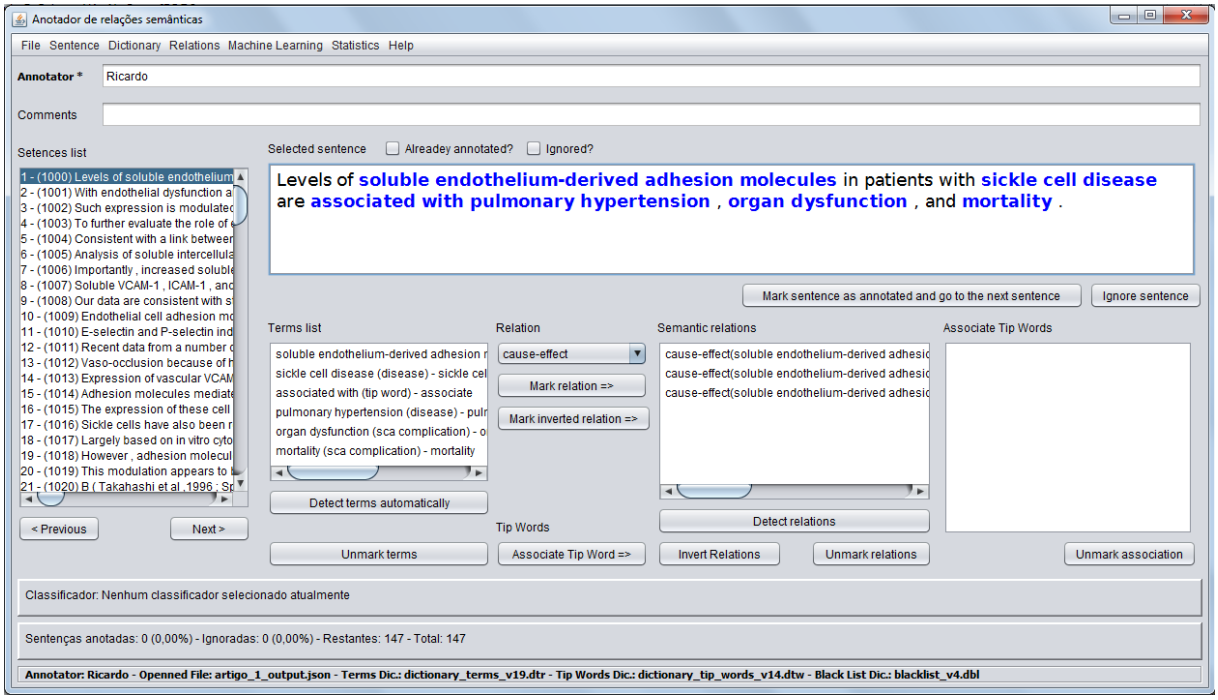

Figura 39: Tela principal da ferramenta ARS.

APÊNDICE D – *Script* Gerador de Resultados - Seleção de sentenças

Dois *scripts* geram os resultados numéricos para a fase de seleção de sentenças. O primeiro, denominado “generate_numerical_results.pl” gera arquivos para contagem de resultados. O segundo, denominado “result_counter.pl” executa a contagem dos resultados.

Estes códigos fazem a leitura dos arquivos de saída, emitidos pela ferramenta ARS, após a execução da primeira fase de extração de relacionamentos. Os arquivos de saída da ferramenta ARS são nomeados como:

- MANUAL_ASSOC_OUTPUT_TOTAL.txt : arquivo que registra as sentenças anotadas manualmente como relação de associação.
- MANUAL_INC_DEC_OUTPUT_TOTAL.txt : arquivo que registra as sentenças anotadas manualmente como relação *increasedecrease*.
- MANUAL_ASSOC_NOT_OUTPUT.txt : arquivo que registra as sentenças anotadas manualmente como relação de associação e também os resultados da seleção automática.
- MANUAL_INC_DEC_NOT_OUTPUT.txt : arquivo que registra as sentenças anotadas manualmente como relação de associação e também os resultados da seleção automática .
- SYSTEM_ASSOC_OUTPUT.txt : arquivo que registra as sentenças selecionadas e as relações extraídas automaticamente pelo sistema como relação de associação.
- SYSTEM_INC_DEC_OUTPUT.txt : arquivo que registra as sentenças selecionadas e as relações extraídas automaticamente pelo sistema como relação de *increase/decrease*.

Para executar os *scripts*, primeiramente é necessária a instalação da linguagem Perl. Caso esteja utilizando sistema operacional baseado em Unix (Linux, MacOS, FreeBSD) basta executar diretamente. Caso esteja utilizando sistema operacional Windows, os *scripts* podem ser executados por meio dos arquivos de extensão .bat que recebem o mesmo nome. Os arquivos citados acima devem estar no mesmo diretório dos *scripts*.

Como resultado da contagem, o *script* gera um arquivos denominado “RESULT_SENTENCES_FILTERING.txt”. Abaixo, um exemplo de dos resultados apresentados neste arquivo:

```
--------------------------------
CONTANDO SENTENCAS ASSOC
--------------------------------
Total de sentenças ASSOC anotadas MANUALmente e que foram encontradas pelo SISTEMA
(Verdadeiro Positivo) : 411
Total de sentenças ASSOC NAO anotadas MANUALmente e que foram encontradas pelo SISTEMA
(Falso Positivo) : 8
Total de sentenças ASSOC anotadas MANUALmente e que NAO foram encontradas pelo SISTEMA
(Falso Negativo) : 0
Total de sentenças ASSOC NAO anotadas MANUALmente e que NAO foram encontradas pelo SISTEMA
(Verdadeiro Negativo) : 1091

PRECISAO_ASSOC : 98.090692124105 %
COBERTURA_ASSOC (REVOCACAO) : 98.1366459627329 %
MEDIDA-F_ASSOC : 99.0361445783133 %

--------------------------------
CONTANDO SENTENCAS INC_DEC
--------------------------------
Total de sentenças INC_DEC anotadas MANUALmente e que foram encontradas pelo SISTEMA
(Verdadeiro Positivo) : 158
Total sentenças INC_DEC NAO anotadas MANUALmente e que foram encontradas pelo SISTEMA
(Falso Positivo) : 26
Total sentenças INC_DEC anotadas MANUALmente e que NAO foram encontradas pelo SISTEMA
(Falso Negativo) : 3
Total sentenças INC_DEC NAO anotadas MANUALmente e que NAO foram encontradas pelo SISTEMA
(Verdadeiro Negativo) : 1323

PRECISAO_INC_DEC : 85.8695652173913 %
COBERTURA_ASSOC (REVOCACAO) : 98.1366459627329 %
MEDIDA-F_ASSOC : 90.6759652755076 %
```

APÊNDICE E – *Script* Gerador de Resultados - Extração de Relacionamentos

Um *script*, denominado “2nd_phase_result_counter.pl” gera os resultados numéricos para a fase de extração de relacionamentos semânticos.

Estes códigos fazem a leitura dos arquivos de saída, emitidos pela ferramenta ARS, após a execução da primeira fase de extração de relacionamentos. Os arquivos de saída da ferramenta ARS são nomeados como:

- SYSTEM_ASSOC_OUTPUT.txt : arquivo que registra as sentenças selecionadas e as relações extraídas automaticamente pelo sistema como relação de associação.
- SYSTEM_INC_DEC_OUTPUT.txt : arquivo que registra as sentenças selecionadas e as relações extraídas automaticamente pelo sistema como relação de *increase/decrease*.
- RELATIONS_OUTPUT_manual.txt : arquivo que registra as relações extraídas em cada sentença de forma manual.

Para executar o *script*, primeiramente é necessária a instalação da linguagem Perl. Caso esteja utilizando sistema operacional baseado em Unix (Linux, MacOS, FreeBSD) basta executar diretamente. Caso esteja utilizando sistema operacional Windows, o *script* pode ser executado por meio do arquivo de extensão .bat que recebem o mesmo nome. Os arquivos citados acima devem estar no mesmo diretório dos *scripts*.

Como resultado da contagem, o *script* gera um arquivos denominado “RELATIONS_RESULTS_OUTPUT.txt”. Abaixo, um exemplo de dos resultados apresentados neste arquivo:

```
-----------------
  ESTATISTICAS
-----------------

TOTAL DE SENTENCAS EXISTENTES: 1509
TOTAL DE SENTENCAS VALIDAS: 432

TOTAL DE RELACOES ANOTADAS MANUALMENTE: 2596
TOTAL DE RELACOES ANOTADAS AUTOMATICAMENTE E VALIDAS: 2399

TOTAL DE RELACOES ANOTADAS MANUALMENTE E ENCONTRADAS PELO SISTEMA
(Verdadeiro Positivo): 2275
TOTAL DE RELACOES NAO ANOTADAS MANUALMENTE E ENCONTRADAS PELO SISTEMA
(Falso Positivo): 124
TOTAL DE RELACOES ANOTADAS MANUALMENTE E NAO ENCONTRADAS PELO SISTEMA
(Falso Negativo): 44

-----------------

PRECISAO: 98.83 %
COBERTURA: 98.10 %
MEDIDA_F: 96.43 %
```

APÊNDICE F – Estrutura e Categorias do Dicionário de Termos do Domínio

Este apêndice apresenta a estrutura para codificação de termos do domínio utilizada no Dicionário de Termos e pela ferramenta de auxílio à anotação de relações semânticas, a ARS.

A estrutura foi desenvolvida pelo próprio autor. Ela foi construída em um arquivo texto, com caracteres para delimitar as informações, para que possa ser mapeado por um *parser* bastante simples embutido no código da ferramenta ARS.

O exemplo abaixo destaca um termo registrado no dicionário. O delimitador de informações é um par de caracteres *pipe* (||):

```
name: sickle cell anemia || category: disease || def: is a hereditary blood disorder,
    characterized by red blood cells that assume an abnormal, rigid, sickle shape ||
                    synonym: Hb SC disease || synonym: Hb S
```

A Tabela 11 destaca cada campo:

Tabela 11: Estrutura de um termo do Dicionário de Termos.

Campo	Obrigatório	Descrição
name	Sim	Nome que identifica um termo.
category	Sim	Categoria em que o termo foi classificado.
def	Não	Definição ou significado do termo.
synonym	Não	Nomes sinônimos que determinado termo pode receber.

A Tabela 12 destaca cada categoria de um termo do Dicionário de Termos:

Tabela 12: Categorias de um termo do Dicionário de Termos.

Categoria	**Descrição**
Anatomical Process	Processos anatômicos que acontecem no corpo humano.
Anatomical State	Possíveis estados físicos do corpo humano ou parte do corpo. Ex.: fraco, sadio.
Anatomical Structure	Estruturas anatômicas. Ex.: Órgãos.
Cellular Structure	Estruturas celulares. Ex.: Organelas.
Chemical Entity	Entidades químicas. Ex.: C - *Carbon*.
Disease	Doenças ou efeitos negativos. Ex.: *Anemia*.
Gene	Genes. Ex.: HAMP.
Measure	Medidas. Ex.: *mortalidade rate*.
Protein	Proteínas. Ex.: *Ferroportin*.
SCA Complication	Complicações específicas da SCA.
Treatment	Tratamentos.

APÊNDICE G – Estrutura e Significados do Dicionário de *Tip Words*

Este apêndice apresenta a estrutura para codificação de *tip words* utilizada no Dicionário de *Tip Words* e pela ferramenta de auxílio à anotação de relações semânticas, a ARS.

A estrutura também foi desenvolvida pelo próprio autor, já levando em consideração a estrutura dos dicionários de termos. Construída em um arquivo texto, com caracteres para delimitar as informações, para que possa ser mapeado por um *parser* bastante simples embutido no código da ferramenta ARS.

O exemplo abaixo destaca um *tip word* registrado no dicionário. O delimitador de informações é um par de caracteres *pipe* (||). :

```
name: associate || category: tip word || meaning: association || synonym:
associates || synonym: associated || synonym: associating || synonym: positively
associated || synonym: negatively associated || synonym: significantly associated ||
synonym: independently associated || synonym: association || synonym: inversely
associated || synonym: associations || synonym: associated predominantly
```

A Tabela 13 destaca cada campo:

Tabela 13: Estrutura de um *tip word* do Dicionário de *Tip Words*.

Campo	Obrigatório	Descrição
name	Sim	Nome que identifica um *tip word*.
category	Sim	Categoria em que o *tip word* foi classificado. Só existe *tip word*.
meaning	Sim	Significado que o *tip word* possui.
synonym	Não	Nomes sinônimos que determinado *tip word* pode receber.

A Tabela 14 destaca cada significado de um *tip word* do Dicionário de *Tip Words*:

Tabela 14: Significados de um *tip word* do Dicionário de *Tip Words*.

Significado	**Descrição**
Association	Associação entre termo e outro.
Increase	Indica o aumento ou crescimento da quantidade de determinado componente (termo). Ex.: *ferroportin increased*.
Decrease	Indica a diminuição ou decrescimento da quantidade de determinado componente (termo). Ex.: *ferroportin decreased*.
Negative	Indicam sentido negativo ou contrário à relação. Ex.: *Not, don't*.
Possibility	Indicam sentido de possibilidade à relação. Ex.: *could, should*.

APÊNDICE H – Estrutura do Dicionário *Blacklist*

Este apêndice apresenta a estrutura utilizada no Dicionário *Blacklist*.

A estrutura também foi desenvolvida pelo próprio autor. Construída em um arquivo texto, com caracteres para delimitar as informações, para que possa ser mapeado por um *parser* bastante simples embutido no código da ferramenta ARS.

A Tabela 15 destaca cada campo:

Tabela 15: Estrutura de um *tip word* do Dicionário de *Tip Words*.

Campo	**Descrição**
Term	Expressão a ser removida das sentenças.
Annotated	Mesma expressão do campo *Term*, com anotação de termos.

O exemplo abaixo destaca uma expressão registrada na *Blacklist*. O delimitador de informações é um par de caracteres *pipe* (||). :

```
Term: in adults with sickle cell disease || Annotated: in <adults><anatomical structure>
with <sickle cell disease><disease>
```

APÊNDICE I – MetaRegras

Abaixo se destacam as estruturas das MetaRegras:

Association:

```
((?:and|or| , | .)?(?:<.*>)(?:<CATEGORIES>).*)(?:(?:<TIP WORDS OTHER>)<tip word>)
                ((?:and|or| , | .)?(?:<.*>)(?:<CATEGORIES>).*)
```

Increase/Decrease:

```
    ((?:and|or| , | .)?(?:<INCREASE/DECREASE><tip word> )?(?:<.*>)(?:  *
 <CATEGORIES>)(?:<INCREASE/DECREASE><tip word>)?).*?((?:and|or| , | .)?(?:  *
      <INCREASE/DECREASE><tip word> )?(?:<.*>)(?:<CATEGORIES>)(?:  *
                    <INCREASE/DECREASE><tip word>)?)
```

Onde, CATEGORIES representam as categorias das *tip words*, TIP WORDS OTHER representam as *tip words* cujo significado é *other*, INCREASE/DECREASE representam as *tip words* cujo significado é *increase* ou *decrease*.