PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTRAÇÃO DE RELAÇÕES HIPONÍMICAS EM CORPORA DE LÍNGUA PORTUGUESA

PABLO NEVES MACHADO

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Vera Lucia Strube de Lima

Porto Alegre
2015

Dados Internacionais de Catalogação na Publicação (CIP)

M149e Machado, Pablo Neves

Extração de relações hiponímicas em corpora de língua portuguesa / Pablo Neves Machado.– Porto Alegre, 2015.
80 p.

Dissertação (Mestrado) – Faculdade de Informática, PUCRS.
Orientador: Profª. Drª. Vera Lucia Strube de Lima.

1. Informática. 2. Processamento da Linguagem Natural. 3. Análise Semântica (Programação). I. Lima, Vera Lucia Strube de. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Extração de Relações Hiponímicas em Corpora de Língua Portuguesa" apresentada por Pablo Neves Machado como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, aprovada em 26/03/2015 pela Comissão Examinadora:

Profa. Dra. Vera Lúcia Strube de Lima – PPGCC/PUCRS
Orientadora

Profa. Dra. Renata Vieira – PPGCC/PUCRS

Profa. Dra. Valéria Delisandra Feltrim – UEM

Homologada em 21/05/2015, conforme Ata No. 008 pela Comissão Coordenadora.

p/ Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

AGRADECIMENTOS

A todos os profissionais da área de processamento de linguagem natural com quem tive contato, por terem se empenhado em contribuir para este trabalho, mesmo com todos os seus deveres pessoais. Especialmente para Vera Strube, que orientou e contribuiu de forma ímpar para o desenvolvimento deste trabalho.

A minha namorada, pela paciência e apoio incondicional enquanto eu despendia meses de trabalho para realizar essa dissertação.

A todos os familiares, mas especialmente aos meus pais, por terem patrocinado meus estudos ao longo de anos, muitas vezes deixando de realizar seus sonhos para investir em minha educação.

Aos colegas que durante o decorrer do curso compartilharam sonhos, dedicação e horas de estudo formando uma relação de amizade e respeito.

EXTRAÇÃO DE RELAÇÕES HIPONÍMICAS EM CORPORA DE LÍNGUA PORTUGUESA

RESUMO

O Processamento da Linguagem Natural (PLN) é uma área da Ciência da Computação destacada por sua relevância para o desenvolvimento de aplicações em processamento de grandes quantidades de documentos textuais ou orais.

Neste trabalho focamos nos textos em língua portuguesa, deles extraindo relações hiponímicas entre entidades, usando uma abordagem baseada em regras adaptadas dos trabalhos de Hearst para o inglês, Freitas e Quental e Taba e Caseli para o português, aqui complementadas.

Para validar a proposta foi desenvolvido um protótipo que extrai relações hiponímicas de corpora em língua portuguesa. O protótipo foi executado sobre corpus de textos e os resultados obtidos foram analisados tanto por fonte de referência como por grupos de regras. O processo avaliativo seguiu o proposto por Freitas e Quental com avaliação humana, e as medidas obtidas são comparadas com as relatadas nas principais fontes de referência. A dissertação ainda estuda em detalhe os erros mais frequentes identificados.

Palavras-chave: Extrações de Relações; Extração de Informações; Relações Hiponímicas; Processamento de Linguagem Natural.

HYPONYMIC RELATIONS EXTRACTION IN PORTUGUESE LANGUAGE CORPORA

ABSTRACT

Natural Language Processing (NLP) is a Computer Science area featured by its relevance to the development of applications that process large amounts of text or speech.
In this paper we focus on texts in Portuguese, extracting from them hyponymic relations between entities, using a rules-based approach adapted from Hearst to English, and Freitas and Quental and Taba and Caseli to Portuguese. The prototype was executed over a corpus of Portuguese texts and the output was analyzed according to the reference author and rule sets. The evaluation process followed the one proposed by Freitas and Quental with human judgment, and the results are compared to those reported in the main references. The dissertation also studies in detail the most common errors identified.

Palavras-chave: Relation Extraction; Information Extraction; Hyponymic Relations; Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ClausIE	Clause-based Open Information Extraction
CORSA	Corpus da Saúde Publica
HTML	HyperText Markup Language
IE	Information Extraction
JSON	JavaScript Object Notation
LSA	Latent Semantic Analysis
Mb	Megabytes
NP	Noun Phrase
PLN	Processamento de Linguagem Natural
POS	Part Of Speech
SVM	Support Vector Machine
VISL	Visual Interactive Syntax Learning
WWW	World Wide Web

SUMÁRIO

1. INTRODUÇÃO

Um dos focos da área do Processamento da Linguagem Natural (PLN) é o processamento de grandes quantidades de documentos textuais. Existe escassez de estudos e ferramentas de PLN referentes à língua portuguesa [Bic00] e [Bas07], o que foi um dos principais fatores motivadores para a escolha do tema abordado. A extração de informações a partir de textos, especialmente extração de relações hiponímicas entre entidades na língua portuguesa, é o foco desta dissertação.

"Uma relação é um conjunto de tuplas que representam um relacionamento entre objetos no universo do discurso, onde cada tupla é uma sequência finita e ordenada de objetos." [Gru92] (tradução livre). Na definição de relação apresentada por Gruber, uma tupla é uma sequência ordenada e finita de objetos correspondendo aos argumentos da relação, podendo ser representados pela expressão (nome-da-relação arg_1 arg_2 ... arg_n), onde arg_i é um objeto na tupla. No presente trabalho apenas serão abordadas relações hiponímicas binárias, sendo representadas por "Hiponímia(arg_1,arg_2)".

Este trabalho reúne padrões propostos por diferentes autores, como [Hea92], [Fre07] e [Tab13], adaptando a escrita das regras num padrão único, para a criação de uma ferramenta de extração de relações em corpus de língua portuguesa. Para isto são realizadas adaptações do inglês para o português, assim como propostas melhorias, e é realizada a análise dos padrões criados. Complementarmente é realizada a avaliação e análise comparativa, como também a discussão dos resultados.

1.1 Motivação

O crescimento rápido da *World Wide Web* (WWW) teve como consequência um desafio na compreensão do conteúdo das informações. Existem diversas

tecnologias envolvidas e diferentes maneiras de difundir conteúdo. Hoje, o acesso a informações na web é realizado prioritariamente por meio da busca por palavras-chave, e essa busca é realizada por mecanismos de comparação lexical. Devido ao gigantesco tamanho que a web apresenta atualmente, e sua contínua expansão, quando são realizadas buscas por palavras-chave diversos conteúdos irrelevantes para o usuário são encontrados.

Diversas fontes de dados manualmente estruturados foram surgindo, mas devido à grande quantidade de conteúdo existente na rede e sua contínua expansão, fica evidente a importância de ferramentas para extração da informação disponível em língua natural, representando-a de forma mais estruturada. Apesar de ser possível encontrar ferramentas que se proponham a realizar essa tarefa de forma automática, é relevante salientar que a grande maioria destas são criadas para suportar apenas a língua inglesa. Esse fato provoca a necessidade de criação de ferramentas específicas para a língua portuguesa, motivando assim este trabalho, que pretende contribuir na solução aos desafios existentes na aplicação de técnicas de extração de relações nesta língua, assim como no modo de tratar as diferenças intrínsecas existentes entre o português e outras línguas estudadas no âmbito dessa pesquisa.

Em uma rápida comparação com o inglês, podemos observar algumas diferenças:

- No português, é possível que uma sentença não apresente pronome pessoal (pronome oculto), caso que ocorre com muita frequência na língua escrita, enquanto que, em outras línguas, o pronome pode ser necessário.
- No inglês, existe uma variedade menor de conjugações verbais, enquanto, no português, existem diversas formas de conjugação verbal.
- As perguntas em português são feitas, no caso da língua escrita, com um ponto de interrogação no final da frase. Já no inglês, quando ocorre uma pergunta existem mudanças na estrutura da frase.
- No inglês, a maioria dos compostos nominais apresenta o modificador à esquerda e o núcleo à direita. Já no português, a construção mais comum é com o núcleo à esquerda e o modificador à direita. Exemplificando, *apple pie* corresponde a torta de maçã.

1.2 O Trabalho Realizado

A presente dissertação endereça a extração de relações em língua portuguesa. A abordagem inicial se baseia na organização das contribuições presentes nos trabalhos de Hearst [Hea92], Freitas e Quental [Fre07] e Taba e Caseli [Tab13], sendo que a arquitetura da solução e a prototipação seguem organização própria. Também são aproveitados os esforços de outros pesquisadores, tais como Baségio em [Bas07], que realizou a adaptação de padrões existentes na língua inglesa para a portuguesa.

No presente trabalho foi desenvolvido um protótipo de extração de relações hiponímicas de corpora em língua portuguesa. Os resultados obtidos, após a execução do protótipo são analisados e avaliados de forma comparativa com os registrados na literatura. Também são discutidos o processo de avaliação manual e analisados os erros frequentes.

1.3 Organização do Texto

O restante desse trabalho é organizado da seguinte forma: O Capítulo 2 contém a fundamentação teórica na área da extração de relações hiponímicas. O Capítulo 3 descreve os trabalhos correlatos que foram de fundamental importância para o desenvolvimento da dissertação. O Capítulo 4 descreve o modelo proposto, as relações a serem extraídas assim como a estratégia utilizada. O Capítulo 5 apresenta o protótipo construído para a aplicação do arcabouço de regras propostas. Já o Capítulo 6 descreve o resultado dos testes decorrentes da aplicação do protótipo. O Capítulo 7 contém uma análise detalhada dos resultados assim como a avaliação realizada. Por fim no Capítulo 8 algumas conclusões são trazidas com intuito de possibilitar trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados conceitos que são de fundamental importância para o embasamento da dissertação. Primeiramente são abordados temas relevantes para o trabalho com corpus e o pré-processamento do corpus. A seguir são trazidos conceitos sobre relações semânticas e por fim apresentadas estratégias de extração de relações semânticas, especialmente hiponímicas.

2.1 Corpus e Processamento de Corpora

Como os textos podem ser obtidos de diferentes fontes e em diferentes formatos, é necessária uma conversão para um formato padrão. É relevante salientar a preocupação com os detalhes da etapa de pré-processamento, pois muitas vezes a forma como os dados são formatados na web contém significado que pode ser perdido durante esta etapa. Exemplificando, poderíamos considerar a exibição de dados em listas ou tabelas, e a retiradas das *tags* que estruturam os dados nesse formato sem ocorrer um cuidado em preservar essa formatação. O resultado do pré-processamento inadequado nesse caso, pode gerar dados irrelevantes devido à falta de significado.

Corpus é um conjunto de textos escritos em uma determinada língua que está organizado de maneira a facilitar o estudo da mesma. Pustejovsky e Stubbs em [Pus12] descrevem corpus como sendo uma coleção de textos legíveis por máquina que foram produzidos de maneira natural.

De posse de um corpus, os pesquisadores podem dispor de dados confiáveis para utilizar em suas pesquisas. Este ainda pode ser classificado de diversas formas de acordo com seu conteúdo. Um corpus oral é um conjunto de textos extraídos de conversas, enquanto que um corpus textual pode ser retirado de livros ou textos da internet.

Este conjunto de textos pode ter um tema ou natureza específicos como, por exemplo, medicina, textos históricos, literatura, entre outros. A escolha de um corpus adequado é de grande importância para o sucesso da pesquisa. Uma escolha inadequada pode prejudicar os resultados e assim levar a conclusões errôneas.

2.1.1 Tokenização

Um dos primeiros passos após o pré-processamento do texto é geralmente a tokenização. Esta etapa consiste em quebrar o texto em pequenas partes que são chamadas de *tokens*. Frequentemente elas representam apenas uma palavra. Uma forma simples de realizar essa atividade é a identificação dos *tokens* por espaços em branco existentes na sentença [Ing12].

Essa separação da oração em porções menores permite que uma máquina consiga interpretar o texto como sendo a composição de várias palavras e que possa ser realizada uma análise aprofundada de cada componente ou grupo de componentes da oração.

2.1.2 Anotações Linguísticas no Corpus

Um corpus pode ser anotado, ou seja, cada palavra pode ter anotações que aumentem sua expressividade. Algumas informações comumente providas pelo processo de anotação são:

- *Part-of-speech* (categoria gramatical).
- Informações morfológicas tais como flexão, derivação ou composição de uma palavra.
- Estrutura sintática da frase e da sentença.

As informações que a anotação linguística provê podem ser empregadas para aumentar a acurácia da interpretação da informação contida na sentença. Uma utilidade dessa informação seria a alimentação de sistemas de aprendizado de máquina, ou ainda a criação de padrões para extrair relações que levam em conta anotações como a classe gramatical de uma palavra.

2.2 Relações Semânticas

Nesta seção são abordados os fundamentos teóricos quem envolvem relações semânticas. Entre eles é contextualizado o termo relações semânticas, com enfoque para as relações hiponímicas.

2.2.1 Semântica

A semântica estuda o significado de palavras, símbolos e frases. Ela constitui elemento de qualquer tipo de linguagem como, por exemplo, linguagens de programação e linguagens formais, além das linguagens naturais. Esse estudo fará uso da semântica da língua portuguesa.

Enquanto a sintaxe se preocupa com a estrutura da sentença, a semântica foca no significado contido. Nas linguagens utilizadas entre seres humanos para a comunicação é comum à semântica se sobrepor à sintaxe, devido a o objetivo ser a transmissão da informação que está contida no significado da sentença.

2.2.2 Relações Semânticas

Uma relação semântica se refere ao significado da ligação entre determinadas palavras. Segundo Jurafsky e Martin em [Jur09] o significado de uma palavra pode ser expresso como sendo sua relação com outras palavras.

“Uma relação é um conjunto de tuplas que representam um relacionamento entre objetos no universo do discurso, onde cada tupla é uma sequência finita e ordenada de objetos.” [Gru92] (tradução livre). Na definição de relação descrita por Gruber uma tupla é uma sequência ordenada e finita de objetos correspondendo aos argumentos da relação, podendo ser representados pela expressão (nome-da-relação arg_1 arg_2 ... arg_n), onde arg_i é um objeto na tupla. No presente trabalho apenas serão abordadas relações hiponímicas binárias, sendo representadas por “Hiponímia(arg_1,arg_2)”.

Existem diferentes tipos de relações semânticas. A relação de sinonímia expressa equivalência de significado entre palavras. Já a relação de antonímia apresenta uma definição diretamente inversa à da sinonímia, transmitindo uma ideia de oposição entre as palavras pertencentes à relação.

Outra categoria de relações semânticas é a relação hierárquica. As relações hierárquicas são representadas pela hiperonímia e hiponímia. A hiperonímia expressa uma relação de significado geral, enquanto a hiponímia representa um significado hierárquico restrito. Alguns exemplos das relações citadas nesta seção podem ser vistos na Tabela 2.1.

Tabela 2.1- Exemplos de relações semânticas

Argumento 1	**Relação**	**Argumento 2**
Alvo	Sinonímia	Claro
Escuro	Antonímia	Claro
Cachorro	Hiperonímia	Animal
Animal	Hiponímia	Cachorro

Existem outras relações semânticas que ligam argumentos no texto, verbais ou não verbais. Por exemplo, da oração “Alexandre adora fritas“ pode ser extraída a tupla (Alexandre, adora, fritas) onde “Alexandre” e “fritas” são argumentos e “adora“ representa a relação.

Uma possível área de utilização de relações semânticas é a criação de ontologias [Bas07], [Gru92], [Lee07], [Mar08], [Xav13].

Na criação de uma ontologia é preciso selecionar quais tipos de relações serão utilizadas, assim como as entidades envolvidas. Estas decisões são tomadas com base no domínio, podendo tomar diversos formatos, mas existe um consenso quando discutida a importância das relações “é_um” e “parte_de” para a construção de ontologias. Estas relações ainda podem ser subdivididas em outros grupos. A relação “é_um”, por exemplo, pode ligar dois conceitos genéricos (“carro” “é_um” “meio de transporte”) assim como um conceito genérico com sua instância (“fusca” “é_um” “carro”). Na Tabela 2.2 podem ser vistos alguns exemplos de subcategorias da relação “parte_de”.

Tabela 2.2 – Exemplo de subrelações “parte_de” extraído de [Win87]

Relação	Exemplo
component-integral object	*pedal – bike*
member-collection	*ship – fleet*
portion-mass	*slice – pie*
stuff-object	*steel – car*
feature-activity	*paying – shopping*
place-area	*Everglades – Florida*

2.2.3 Relações Hiponímicas

Os primeiros trabalhos relacionados a extração automática de relações abordaram, principalmente, relações hiponímicas e meronímicas. Isso se deve ao fato de essas relações serem a base para a construção de ontologias. Será dado um foco maior nas relações hiponímicas, que são o principal objetivo deste trabalho.

As relações hiponímicas são comumente representadas por “é_um”. Isto se deve a expressarem relações entre instâncias e classes, como também entre classes. Quando falamos de relações hiponímicas é comum, na literatura, encontrarmos a expressão “relação hiperonímica”. Ambos os tipos estão associados. A diferença está na ordem dos termos. Por exemplo: “Hiponímia (sanduíche é um tipo de lanche), Hiperonímia (lanche tem sanduíche entre seus tipos)” [Nas13]. Neste exemplo a autora ilustra a relação de significado entre os dois tipos de relações.

2.3 Extração de Relações

Relações verbais são comumente representadas por verbos e seus argumentos. Um exemplo do formato de uma relação binária, possivelmente extraído de um corpus, pode ser dado por

(primeiro argumento, relação, segundo argumento).

Esse modelo de relação pode ser extraído de textos em linguagem natural com base no processamento de corpora.

Conforme [Ban07] os sistemas de extração de relações normalmente focam em satisfazer determinadas demandas pré-especificadas como, por exemplo, extrair o local e horário de um evento a partir de um conjunto de anúncios. Quando ocorre a necessidade de extrair relações de um novo domínio costuma ser necessário um retrabalho. Umas das tarefas que pode ser necessário refazer é o estabelecimento da heurística empregada na extração, como também a etiquetagem de um novo conjunto de treino. Para evitar problemas como estes, existem diferentes abordagens para a extração de relações semânticas.

2.3.1 Método Supervisionado de Extração de Relações

A Extração Supervisionada de Relações tem esse nome devido à necessidade de um supervisor, ou seja, uma intervenção humana que auxilie o método de extração de relações. Esta etapa é chamada de treinamento. O supervisor mapeia um conjunto de dados em suas saídas desejadas, então o método envolve a construção de uma função que, por aproximação, prevê a saída para qualquer entrada. Assim essa solução é generalizada para uma função que idealmente cobre todos possíveis dados de entrada.

A dificuldade da utilização do aprendizado Supervisionado é a necessidade de um grande número de exemplos rotulados, para que o método possa induzir um bom classificador. Essa tarefa não é simples, pois necessita que um operador humano (especialista na área) realize a rotulação manual.

2.3.2 Método Não Supervisionado de Extração de Relações

O aprendizado Não Supervisionado difere do Supervisionado devido a não existir necessidade de supervisão. Os métodos de extração de relações, nesse caso, precisam descobrir as relações existentes no corpus sem o auxílio humano.

[Fin99] afirma que sistemas automatizados de extração de relações usualmente são compostos por grupos de padrões pré-definidos, um procedimento de extração e um mecanismo de atribuição de pesos para as relações extraídas, com objetivo de filtrar os candidatos não relevantes.

A Extração Não Supervisionada de Relações apresenta vantagens e desvantagens se comparada à Extração Supervisionada. Uma vantagem seria a possibilidade de reconhecer uma relação sem o sistema ter sido anteriormente treinado para essa relação. Uma desvantagem do método Não Supervisionado seria a sua menor cobertura, já que métodos Supervisionados podem usar uma grande quantidade de dados como entrada e aprender diversos padrões.

2.3.3 Método de Extração de Relações por Regras

Métodos de Extração de Relações por Regras podem ser classificados como métodos supervisionados de aprendizagem de máquina, já que é necessário que regras específicas sejam fornecidas como entrada para o sistema. Estes métodos recebem maior atenção por serem de grande interesse para esta dissertação, uma vez que pode buscar-se o aproveitamento do arcabouço já disponível junto à literatura, sendo [Hea92] o principal trabalho usado como referência nesse contexto.

Outra característica dessa abordagem é a velocidade de processamento. O motivo é a execução baseada em regras previamente escritas com objetivo de extrair relações que normalmente se aplicam à língua específica do corpus.

Uma dificuldade na utilização deste método é a necessidade de construção manual de regras para extração de relações, já que esse processo envolve estudo detalhado e é custoso de ser realizado. Outra dificuldade vem da dependência do idioma, já mencionada. Regras escritas para sistemas que trabalham com outros idiomas podem ter de ser completamente reescritas.

3. TRABALHOS RELACIONADOS

No atual estado da arte existem trabalhos, principalmente para a língua inglesa, que abordam o tema da extração de relações. Existem também ferramentas e recursos disponíveis que são de interesse. Este capítulo introduz alguns desses trabalhos referentes ao tema.

Entre os trabalhos que estudam a extração de relações em corpora textuais, duas são as abordagens mais comuns: o aprendizado de máquina e a extração baseada em regras. Na exposição desses trabalhos será dada uma ênfase maior para a segunda abordagem, já que esta apresenta vínculo com o trabalho proposto.

Existe uma grande variedade de relações que podem existir entre conceitos ou entre conceitos e instáncias. Dentre estas, as mais abordadas são as relações hierárquicas. Um fator que contribui para tal pode ser o seu emprego na construção de ontologias, que contêm estruturas compostas por hierarquias de conceitos [Rui05].

3.1 Trabalhos com Foco em Língua Estrangeira

Em [Hea92], Hearst propõe um método de aquisição de relações hiponímicas, entre sintagmas nominais, para a língua inglesa, com base em 6 padrões simples que podem ser encontrados com frequência em textos. Estes podem ser vistos na Tabela 3.1.

Tabela 3.1 – Padrões extraídos de [Hea92]

i	NP such as {NP ,}* {(or | and)} NP
ii	such NP as {NP ,}* {(or | and)} NP
iii	NP {, NP}* {,} or other NP
iv	NP {, NP}*{,} and other NP
v	NP {,} including {NP ,}* {or | and} NP
vi	NP {,} especially {NP ,}* {or | and} NP

Um dos objetivos que conduziu Hearst a esta abordagem foi criar um método aplicável a grandes quantidades de textos. A importância do trabalho de Hearst se deve ao fato de ser um dos primeiros trabalhos encontrados na literatura a propor padrões lexicais na extração de relações semânticas, com grande aceitação acadêmica. Os padrões textuais criados por Hearst são utilizados em diversos trabalhos, como por exemplo em [Fre07], [Bas07], [Mae02] e [Deg04]. Um exemplo da aplicação destes padrões pode ser o retirado de [Hea92], no qual é mostrada uma aplicação prática do padrão (vi).

"...most European countries, especially France, England and Spain."

Aplicando o padrão "NP {,} especially {NP ,}* {or | and} NP" (vi), apresentado na Tabela 3.1, onde NP é uma *Noun Phrase*, as seguintes relações são extraídas:

Hiponímia ("France", "European country")
Hiponímia ("England", "European country")
Hiponímia ("Spain", "European country")

Hearst aplicou seus padrões em corpora enciclopédicos e jornalísticos avaliando que 63% das relações identificadas eram de boa qualidade.

Em [Ced03] os autores demonstram que a aplicação de informações linguísticas provenientes de modelos matemáticos para medir a similaridade semântica entre conceitos pode melhorar a cobertura e precisão de métodos automáticos de extração de relações hiponímicas de corpus em língua inglesa. São

utilizados os padrões propostos por Hearst [Hea92], e é aplicado um método denominado *latent semantic analysis* (LSA) para filtrar as relações incorretas, aumentando a precisão em 30%. Relações corretamente extraídas podem ser usadas como "semente" para a extração de diversas outras relações, assim aumentando a cobertura.

Em [Mor03] Morin e Jacquemin apresentam padrões para a aquisição de relações hiponímicas em corpora de língua francesa.

Tabela 3.2 - Regras para a língua francesa extraídas de [Mor03]

i	deux|trois...|2|3|4...} NP1 (LIST2)
ii	{certain|quelque|de autre...} NP1 (LIST2)
iii	{deux|trois...|2|3|4...} NP1: LIST2
iv	{certain|quelque|de autre...} NP1: LIST2
v	{de autre} NP1 tel que LIST2
vi	NP1, particulièrement NP2
vii	{de autre} NP1 comme LIST2
viii	NP1 tel LIST2
ix	NP2 {et|ou} de autre NP1
x	NP1 et notamment NP2

Na Tabela 3.2 são descritas as regras propostas por Morin e Jacquemin. O exemplo a seguir, dado pelos autores, demonstra como tais padrões se comportam. Se o padrão "{deux|trois...|2|3|4...} NP1 (LIST2)" é aplicado ao trecho:

"... analyse foliaire de quatre espèces ligneuses
(chêne, frêne, lierre et cornouiller) dans..."

... é possível identificar as seguintes relações:

Hiponímia ("chêne", "espèce ligneux")
Hiponímia ("frêne", "espèce ligneux")
Hiponímia ("lierre", "espèce ligneux")

Hiponímia (“cornouiller”, “espèce ligneux”)

Uma nova abordagem para a extração de relações é a *Open Information Extraction (OpenIE)*, que visa a extração aberta e em grande escala, sem se preocupar em tipificar as relações extraídas. [Cor13], em seu trabalho, propõe uma abordagem para extração aberta de relações, apresentando o sistema ClausIE (Clause-based *Open Information Extraction*). Os experimentos realizados sugerem que o sistema obtenha os melhores resultados entre os que realizam OpenIE, se tornando uma referência na área. Esse sistema difere dos demais por utilizar uma abordagem baseada em cláusulas (orações), de mais forte cunho linguístico. Ele identifica conjuntos de orações e o tipo destas (de acordo com a função gramatical do conteúdo). Uma oração expressa uma informação coerente composta por sujeito, verbo, e opcionalmente objeto indireto, objeto direto, complemento e advérbio [Abr13]. O sistema ClausIE é baseado em um *parser* de dependências e também em um pequeno conjunto de léxicos independentes de domínio. Essa abordagem permite ao sistema, segundo os autores, o processamento em paralelo, e assim o processamento de grandes coleções de conteúdo, de maneira escalável. Assim como o presente trabalho, ClausIE não necessita de pós processamento e de dados de treinamento (rotulados ou não-rotulados) para sua execução. Segundo Corro [Cor13] uma das principais fontes de incorreções nas relações extraídas são provenientes de erros de *parser*.

Em seu trabalho Gamallo e coautores [Gam12] descrevem um método que utiliza o paradigma OpenIE para a extração de triplas baseadas em verbos de corpora multilíngues. O método extrai relações em corpora nos idiomas português, inglês, espanhol e galego. Segundo os autores o método descrito apresenta resultados superiores aos alcançados pelos trabalhos no estado da arte, devido principalmente ao fato de o método utilizar análise sintática profunda e um *tokenizer* robusto e rápido.

[Xav13] relata o desenvolvimento de uma proposta para extração aberta de relações em textos de língua inglesa, pela aplicação de um conjunto de padrões sintáticos em um texto *POS-tagged*. Diferente do presente trabalho, os padrões utilizados se propõem a extrair outros tipos de relações além das hiponímicas. Os padrões propostos pela autora podem ser vistos na Tabela 3.3.

Tabela 3.3 - Exemplos de padrões de relações semânticas extraídos de [Xav13]

A NP OF NP IS NP
NP IS THE EXP OF NP
NP VERB (IN|AT) NP
NP (WAS|IS) (IN|AT) NP
(NP)? NP AND NP VERB (PREPOSITION/SUBORD. CONJ) (THE|A)? NP
NP (WORD)? VERB (WORD)? (A)? (WORD)? (ADJECTIVE)?NP
NP (MODAL)? VERB (PREPOSITION/SUBORD. CONJ)? (A)?(PREPOSITION/SUBORD. CONJ.)? (ADJECTIVE) (NP)?
NP (MODAL)? VERB (PREPOSITION/SUBORD. CONJ)? (A)? (ADJECTIVE) (NP)?
(ADJECTIVE) VERB (DETERMINER) NP (PREPOSITION/SUBORD. CONJ) NP
NP (TO TO) VERB (ADJECTIVE)?NP
NP VERB (FOR)? THE NP NP, VERB (A)? NP
NP VERB ADVERB (CARDINAL NUMBER)? NP
NP VERB WORD JJ (FOR|TO) NP
NP WORD VBD (VERB BE, PAST PARTICIPLE)? TO WORD VERB (THE)? NP
(NP (THAT|WHICH) (DETERMINER)) VERB ((PREPOSITION/SUBORD. CONJ)?(WORD DT)? NP)
NP WAS (VERB PAST) VERB (PREPOSITION/SUBORD. CONJ) NP

Na Tabela 3.3 apenas o primeiro padrão (“A NP OF NP IS NP”) busca extrair relações hiponímicas. Xavier também compara os resultados obtidos por um protótipo, com os resultados de outros dois sistemas de OpenIE (ReVerb [Fad11] e DepOE [Gam12]). A análise comparativa dos resultados sugere que o protótipo descrito atinja resultados superiores em alguns aspectos.

3.2 Trabalhos com Foco na Língua Portuguesa

O software PALAVRAS [Bic00] reúne diversas ferramentas para o processamento da linguagem natural que aceitam como entrada textos em língua

portuguesa e pode ser utilizado para etiquetação de corpus, processamento léxico-morfológico, geração de árvores sintáticas e reconhecimento de entidades nomeadas, entre outros. É relatada precisão maior que 97%, tanto em termos de morfologia quanto em sintaxe. O *parser* é um sistema baseado em regras e foi desenvolvido em 2000 por Bick. Está disponível através do projeto VISL [Ins15].

Em [Fre07] são adaptados dois padrões de Hearst para a língua portuguesa (“such as” e “and/or others”), e criados outros quatro padrões com base em análise de ocorrências no texto. Estes são capazes de identificar relações hiponímicas. O trabalho utiliza o *parser* PALAVRAS, com etapa de identificação de sintagmas nominais descrita em [San05]. As regras foram aplicadas ao corpus CORSA (corpus da Saúde Pública) que contém cerca de dois milhões de palavras. Os resultados foram compatíveis com os de Hearst, mostrando um percentual de 73% de relações consideradas de boa qualidade.

Tabela 3.4 - Padrões extraídos de [Fre07]

i.a	SN HHiper (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) Sni
i.b	SN Hiper, (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) Sni
ii	SN HHipo { ,SN Hipoi } * { , } e|ou outros SN Hiper
iii	tipos de SN Hiper: SN1 { , SN2 ... ,} (e | ou) Sni
iv	SN HHiper chamado/s/a/as (de) SN Hipo
v	SN Hiper conhecido/s/a/as como SN Hipo

A Tabela 3.4 ilustra os dois padrões de Hearst adaptados por Freitas e Quental (i.a, i.b e ii), assim como os três padrões propostos pelas autoras (iii, iv e v). O excerto de texto retirado de [Fre07] e reproduzido a seguir, demonstra a aplicação do padrão (iv):

“e nele existe uma [substância] chamada [benzopireno].”

Com a aplicação do padrão “SN HHiper chamado/s/a/as (de) SN Hipo” a seguinte relação deve ser extraída: “Hiponímia (benzopireno, substância)”. Segundo as autoras o símbolo HHiper representa o padrão onde o termo hiperônimo é o

primeiro substantivo à esquerda. Na Seção 4.2.3 o trabalho de Freitas e Quental será melhor detalhado, já que este é de fundamental importância para o trabalho corrente.

Conforme já relatado em [Oli09], para a língua portuguesa não havia, livremente disponível, um banco de dados lexical, como por exemplo existe, para a língua inglesa, a WordNet [Fel98]. Para a construção deste recurso lexical para a língua portuguesa, os autores propuseram o PAPEL, um recurso construído por relações entre termos extraídas de forma semiautomática de um dicionário geral da língua portuguesa. O processo de criação do PAPEL foi constituído pelas seguintes etapas:

- Criação dos padrões;
- Extração das relações,
- Análise manual dos resultados,
- Realização de ajustes nas relações.

A etapa de avaliação ocorreu de duas formas distintas. Para as relações de sinonímia foi realizada uma comparação com os dados existentes no Thesaurus Eletrônico para o Português do Brasil [Maz08], considerado como o *Gold Standard*. Já para as outras relações, foi utilizada uma abordagem onde as relações foram transformadas em padrões textuais e, a seguir, estes foram buscados no corpus CETEMPúblico [San01]. O trabalho apresentou o resultado de 63% de precisão para a extração de relações hiponímicas, enquanto que, para outras relações, os resultados variaram entre 35% e 59%.

O trabalho descrito em [Bas07] tem como objetivo a construção semiautomática de ontologias a partir de textos na língua portuguesa do Brasil. Para esse fim é empregada uma abordagem que inclui extração de relações hiponímicas, e para tal o autor traduziu para a língua portuguesa do Brasil relações propostas em outros trabalhos consolidados como, principalmente, [Hea92], como pode ser visto na Tabela 3.5.

Estas adaptações propostas por Baségio foram utilizadas como referência para a abordagem do presente trabalho.

Tabela 3.5 - Padrões de Hearts adaptados em [Bas07]

i	NP such as {(NP,)*(or|and)} NP	SUB como {(SUB,)*(ou|e)} SUB
		SUB tal(is) como {(SUB,)*(ou|e)} SUB
ii	such NP as {(NP,)*(or|and)} NP	tal(is) SUB como {(SUB,)*(ou|e)} SUB
iii	NP {, NP}* {,} or other NP	SUB {, SUB}* {,} ou outro(s) SUB
iv	NP {, NP}* {,} and other NP	SUB {, SUB}* {,} e outro(s) SUB
v	NP {,} including {NP,}*{or|and} NP	SUB {,} incluindo {SUB,}*{ou|e} SUB
vi	NP {,} especially {NP,}*{or|and} NP	SUB {,} especialmente {SUB,}*{ou|e} SUB
		SUB {,} principalmente {SUB,}*{ou|e} SUB
		SUB {,} particularmente {SUB,}*{ou|e} SUB
		SUB {,} em especial { SUB,}*{ou|e} SUB
		SUB {,} em particular { SUB,}*{ou|e} SUB
		SUB {,} de maneira especial { SUB,}*{ou|e} SUB
		SUB {,} sobretudo { SUB,}*{ou|e} SUB

Para atingir seu objetivo, Baségio implementou um processo de remoção de palavras pouco relevantes para o domínio. Este processo removeu cerca de 70% das palavras analisadas. O autor obteve resultados próximos a 55% de precisão em estudos de casos.

Em seu trabalho Gamallo e coautores [Gam12] extraem relações em corpora nos idiomas português, inglês, espanhol e galego. Como já exposto, segundo os autores o método descrito apresenta resultados superiores aos alcançados pelos trabalhos no estado da arte, devido principalmente ao fato de o método utilizar análise sintática profunda e um *tokenizer* robusto e rápido.

[Bat13] propõe um método para classificação de relações entre entidades mencionadas. Este método difere dos demais por utilizar uma abordagem que pesquisa pelos exemplos de treino mais próximos, utilizando o algoritmo *k-nearest neighbors*, como forma de fazer a classificação, aproveitando um método eficiente

baseado em valores mínimos de funções de dispersão como forma de medir a similaridade entre relações, para diferentes tipos de relações semânticas. O trabalho [Bat13] tem o objetivo de não necessitar de intervenção humana. Os exemplos de treino são recolhidos automaticamente da Wikipédia correspondendo a frases que expressam relações entre pares de entidades extraídas da DBPédia. Diferente de outros trabalhos na literatura, como [Hea92] e [Fre07], os padrões utilizados em [Bat13] não contêm palavras específicas (palavras-chave). Os padrões adotados baseiam-se principalmente nas classes gramaticais das palavras que ocorrem antes, depois e entre duas entidades mencionadas.

Em [Tab13] também foi investigado o modo como relações semânticas podem ser extraídas automaticamente de textos em português. Os autores utilizaram 2 corpora anotados pelo *parser* PALAVRAS, onde o primeiro é o CETENFolha, corpus de caráter jornalístico, composto por 24 milhões de palavras de artigos do jornal Folha de São Paulo, enquanto o segundo é de caráter científico, composto por 870 mil palavras, proveniente de textos de uma revista de divulgação científica (FAPESP). Os principais pontos investigados foram o aprendizado de máquina e padrões textuais, onde os autores buscam extrair os seguintes tipos de relações:

- *is-a*
- *part-of*
- *location-of*
- *effect-of,*
- *property-of*
- *made-of*
- *used-for*

Os resultados apresentados no artigo indicam que o aprendizado de máquina, é uma técnica promissora, mas obteve resultados inferiores à extração por padrões textuais em alguns casos investigados. Os padrões utilizados pelos autores podem ser vistos na Tabela 3.6. Onde o termo T1 representa o hiperônimo de uma relação, enquanto os termos T2, T3 representam possíveis hipônimos.

Tabela 3.6 - Padrões de relações semânticas extraídos de [Tab13]

Identificador	Relação	Padrão Textual
I	*is-a*	T1 (tais como|como) T2 {, T3}* (e|ou) TN
Ii		T2 {, T3}* ,? (e|ou) outros T1
Iii		tipos de T1: T2 {, T3}* (e|ou) TN
Iv		T1 chamad(o|a|os|as) de? T2
V		T2 {, T3}* ,? (qualquer|quaisquer) T1
Vi		T2 é (o|a|um|uma) T1
vii		T2 são T1
viii	*property-of*	de T1_ADJ T2 N
ix		T1_N T2_ADJ
x		T2_ADJ T1_N
xi		T1_N “ T2_ADJ ”
xii	*part-of*	T1 com T2
xiii		T1 {verbo fazer} parte de T2
xiv		T1 {verbo ser} parte de T2
xv	*made-of*	T1_N de T2_N
xvi		T1 (é|são)? feit(o|a|os|as) de T2
xvii	*location-of*	T1 chega a o T2
xviii		T1 em (o|a|os|as) T2
xix		T1 entrou em T2
xx		T1 ,? localizad(a|o) em T2
xxi	*effect-of*	T2_V .* devido=a T1
xxii		T2 V por=causa=de (a|o|as|os)? T1
xxiii	*used-for*	T1 para (o|a|os|as) T2_V (e|ou)
xiv		T1 (que podem ser)? usadas? para T2_V

Os autores Taba e Caseli utilizaram 24 padrões textuais que se propõem a extrair sete tipos de relações diferentes. Dentre os padrões utilizados, sete foram obtidos por meio da execução do algoritmo para descoberta de padrões textuais apresentado em [Hea92]. Outros 13 padrões foram manualmente definidos. Ainda completam o total os quatro padrões hiponímicos apresentados em [Fre07]. Na

Tabela 3.6 estes padrões podem ser vistos em detalhe. Apenas as regras de (i) até (vii) são de interesse para o trabalho corrente, visto que estas extraem relações hiponímicas. Dois métodos de classificação baseados em aprendizado de máquina supervisionado foram utilizados pelos autores: Árvore de Decisão e Máquinas de Vetores de Suporte (SVM). O método de avaliação empregado prevê a comparação de resultados obtidos automaticamente com resultados provenientes de extrações manuais.

4. MODELO PROPOSTO

Neste capítulo é apresentada a proposta que norteia esta dissertação de mestrado. Ao longo do capítulo é explicado o modo como o trabalho se organiza. São apresentadas as regras propostas e o formato como elas foram escritas. Ainda são descritos com maior detalhe os trabalhos que influenciaram a criação destas regras.

4.1 Descrição Geral

A dissertação tem como principal objetivo propor uma abordagem de extração de relações em corpora de língua portuguesa, partindo do trabalho de Hearst [Hea92] e mantendo, por princípio, a estratégia de extração baseada em regras. Entretanto, como o trabalho descrito em [Hea92] foi realizado especificamente para a língua inglesa, existem desafios não contemplados para a aplicação da abordagem utilizada, para a língua portuguesa do Brasil.

Quando se trabalha com o processamento da língua portuguesa um dos principais desafios enfrentados pelos pesquisadores é a escassez de recursos e ferramentas. Na língua inglesa existem diversas ferramentas e conjuntos de dados disponíveis para utilização nessa área, enquanto que na língua portuguesa o número de ferramentas e conjuntos de dados disponíveis é muito pequeno se considerada a importância dessa língua.

Neste trabalho é proposta uma abordagem de extração de relações hiponímicas em corpus de língua portuguesa. Esta tem como base a adaptação de trabalhos que já abordam o tema na língua inglesa, levando em conta diferenças eminentes entre as duas línguas. Também são incorporados trabalhos que já conduziram esforços para a adaptação desses padrões para a língua portuguesa.

4.2 Adaptação das Regras

Visando extrair relações hiponímicas em corpora de língua portuguesa, foram realizadas adaptações de padrões propostos por autores como Hearst [Hea92], Freitas e Quental [Fre07] e Taba e Caseli [Tab13]. As regras adaptadas foram inseridas num protótipo desenvolvido especialmente para esta dissertação.

Na Seção 4.2.1 é descrita detalhadamente a sintaxe utilizada para representar as regras adaptadas neste trabalho. São descritos os operadores de repetição, assim como as estruturas utilizadas para representar os Sintagmas Nominais.

Na Seção 4.2.2 são apresentadas cinco adaptações dos padrões sugeridos por Hearst [Hea92]. Nesta seção os padrões são detalhadamente descritos, assim como são relatadas alterações realizadas com o intuito de aumentar a cobertura.

Na Seção 4.2.3 são abordadas as adaptações realizadas com base nas regras propostas em [Fre07]. Das seis regras propostas pela autora, três tem caráter original. Estas foram adaptadas para o atual trabalho, sendo que foi criada uma regra para cada uma das originais.

Na Seção 4.2.4 são abordados os padrões textuais para extração de relações hiponímicas propostos em [Tab13]. O autor aplicou em corpora de língua portuguesa o algoritmo para descoberta de padrões textuais sugerido em [Hea92]. Durante o desenvolvimento do presente trabalho estas regras foram adaptadas, sendo que para a regra (vi) da Tabela 3.6 duas regras correspondentes foram criadas.

4.2.1 Formato das Regras

O formato das regras é muito semelhante à sintaxe de expressões regulares, e os sintagmas nominais são representados por “SN”. Assim como nas expressões regulares, os parênteses são utilizados para agrupar as expressões, enquanto que o “*” representa que uma expressão pode ocorrer nenhuma ou mais vezes. A interrogação significa nenhuma ou uma repetição. Outro símbolo comumente utilizado é o “|” que representa um “ou exclusivo”. Também foi utilizada a notação “<sn PALAVRA-CHAVE sn>” para identificar a ocorrência de uma palavra-chave que está contida dentro de um *chunk*. Como um Sintagma Nominal pode ser formado por

outros SNs, o símbolo “sn” (minúsculo) foi empregado para identificar um Sintagma Nominal que é um dos elemento de um “SN”, como ilustra a Figura 4.1.

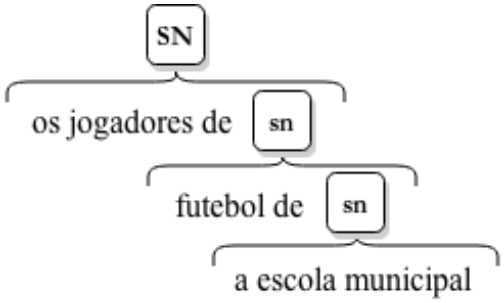

Figura 4.1 – Árvore sintática gerada pelo analisador sintático PALAVRAS

Caso a palavra-chave encontre-se diretamente após o símbolo “<” ou antes do símbolo “>”, significa que ela é respectivamente a primeira ou a última palavra do *chunk*, como está exemplificado em: “<outros sn>” (a palavra-chave é representada por “outros”).

4.2.2 Hearst

Os padrões propostos por Hearst (Tabela 3.1) foram criados com o intuito de extrair relações hiponímicas em corpus de língua inglesa. Para a utilização destas regras junto à língua portuguesa do Brasil, foi necessário um trabalho de tradução e contextualização das mesmas face à semântica da língua portuguesa. A seguir pode ser visualizada uma dessas regras:

(1) SN(,)? como (SN ,)*(SN (e|ou))*SN

Dado o excerto de texto: “Países como o Brasil, Equador e os EUA.”, o padrão representado acima pode extrair as seguintes relações:

Hiponímia (Brasil, País)

Hiponímia (Equador, País)

Hiponímia (EUA, País).

O padrão exemplificado é o referente ao “such as” proveniente dos estudos de Hearst. Este corresponde ao “como”’ em português, que pode exercer diversas funções sintáticas em uma sentença, o que causa dificuldade em obter altos níveis

de precisão, como já mencionado em [Fre07]. Outros trabalhos já realizaram esforços para adaptar o padrão “como” para a língua portuguesa, entre eles destacamos [Bas07], que entretanto considera apenas substantivos, simplificando a ideia de sintagma nominal presente nas regras em [Hea92]. Em nosso trabalho escolhemos utilizar SNs evitando essa simplificação e empregando padrões mais complexos. Já na adaptação em [Fre07], foram utilizadas regras levando em conta a existência de SNs, mas ocorreu, assim como em [Bas07], uma flexibilização, neste caso visando o uso apenas da palavra mais à direita, dentro do sintagma nominal. Uma melhoria introduzida em relação a [Fre07] foi o tratamento da vírgula, que pode ocorrer antes da palavra “como”, por exemplo, em:

“... [outras falhas] , como [dois nomes para um mesmo fator] ...”.

Como veremos em detalhe mais adiante, esta alteração aumentou em torno de 40% o número de relações extraídas com o padrão “como”, em relação aos resultados anteriores.

Utilizando uma abordagem semelhante, foi possível criar regras adaptadas dos padrões 2, 5 e 6 de Hearst [Hea92], apresentados no Capítulo 3:

(2) SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN
(3) SN(,)? incluindo (SN ,)*(SN (e|ou))*SN
(4) SN(,)? especialmente (SN ,)*(SN (e|ou))*SN

Observa-se que o padrão a seguir, inspirado nos padrões 3 e 4 de Hearst [Hea92], necessitou de uma implementação alternativa:

(5) (SN (ou|e|,))*<outr(a|o)(s)? sn>

O analisador sintático PALAVRAS, ao processar um texto como “Brasil, Equador, EUA e outros países”, identifica diversos SNs, um dos quais inclui o determinante “outros”:

“[Brasil], [Equador], [EUA] e [outros países]”

Foi proposta uma adaptação para encontrar SNs nessa situação. Com essas alterações, as relações que podem ser extraídas com o padrão para o texto do exemplo são: Hiponímia (Brasil, País), Hiponímia (Equador, País), Hiponímia (EUA, País).

A Tabela 4.1 associa as regras propostas por Hearst com as regras propostas no presente trabalho.

Tabela 4.1 - Associação entre padrões de Hearst e as regras propostas neste trabalho

Regra	Padrão de Hearst
1	NP such as {NP ,}* {(or | and)} NP
2	such NP as {NP ,}* {(or | and)} NP
3	NP {,} including {NP ,}* {or | and} NP
4	NP {,} especially {NP ,}* {or | and} NP
5A	NP {, NP}* {,} or other NP
5B	NP {, NP}*{,} and other NP

Na Tabela 4.1 pode-se observar que a regra **(5)** foi utilizada para expressar duas regras propostas por Hearst (5A e 5B). Isso se deve a tais regras apresentarem grande semelhança e poderem ser representadas por apenas uma expressão.

4.2.3 Freitas e Quental

Em [Fre07] foram utilizados padrões baseados em [Hea92], padrões (i.a), (i.b) e (ii) da Tabela 3.4. Estes padrões sofreram adaptações realizadas por Freitas e Quental. Os demais padrões de Hearst foram descartados, pois, segundo a autora, foram considerados pouco produtivos. Por meio da observação do corpus foram propostos outros três padrões capazes de identificar relações hiponímicas.

Com o objetivo de maximizar o número de relações identificadas pelo protótipo, foram adicionados os três padrões apresentados pela autora. Para isso ocorreu um esforço de adaptação para a realidade deste trabalho, com uma

proposta de alteração dos padrões no intuito de otimizar o número de relações extraídas.

4.2.3.1 Padrões Adaptados

Os três primeiros padrões mostrados na Tabela 3.4 (i.a, i.b e ii) são adaptações realizadas por Freitas e Quental de padrões propostos por Hearst. Como os padrões de Hearst já foram abordados na Seção 4.2.2, abordaremos nesta seção apenas os padrões (iii), (iv) e (v), originados de [Fre07]. O padrão (iii) da Tabela 3.4, também denominado pelas autoras de “tipos de”, busca extrair relações com base nas palavras-chave que dão origem ao seu nome. Com intuito de demonstrar as relações que a regra é capaz de extrair, será considerado o excerto de texto a seguir:

“desenvolver [dois tipos de dengue] : [dengue clássica] e [dengue hemorrágica]”

Desse trecho, a regra deve ser capaz de extrair as relações: Hiponímia (dengue clássica, dengue), Hiponímia (dengue hemorrágica, dengue). O resultado da adaptação criada para realizar tal tarefa é descrito abaixo:

(6) <... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN

Pode-se notar uma semelhança na escrita deste padrão com o padrão **(5)** proposto anteriormente. Essa semelhança se dá na utilização dos símbolos “<” e “>” para representar um sintagma nominal que contém em seu interior as palavras-chave da regra. Isso se deve ao fato de o analisador sintático PALAVRAS definir que a expressão “tipos de” faz parte de um *chunk* com outras palavras que podem vir antes ou depois do padrão, como por exemplo: “[todos os tipos de cortes]” e “[os principais tipos de tifo]”.

Para maximizar o número de relações extraídas, a regra foi flexibilizada para aceitar a expressão “tipo de”, sem a utilização do plural.

Esta regra apresenta um alto grau de confiança, como as autoras descrevem:

> "... o padrão 'tipos de' não apresenta problemas de ambiguidade relativos ao sintagma preposicionado, nem particularidades de natureza discursiva ou coesiva – o que significa que as relações identificadas são altamente confiáveis." [Fre07]

Outra adaptação realizada com base em [Fre07] foi a do padrão denominado "chamado/a/os/as". Este está representado como (iv) na Tabela 3.4. Este padrão deve extrair relações de excertos de texto como:

"... e nele existe uma [substância] chamada [benzopireno]."

Nesse caso a relação extraída seria Hiponímia (benzopireno, substância). A regra encarregada de tal tarefa pode ser visualizada a seguir:

(7) SN(,| é| são| foram)? chamad(o|a|os|as)(de)? (SN ,)*(SN (e|ou))*SN

Para maximizar o número de relações extraídas, foi flexibilizado o uso do verbo "ser" em quatro formas (é, são, foi, foram), assim como a utilização de vírgula. Foi também permitida a ocorrência de uma lista de sintagmas nominais após a palavra-chave "chamado". Este formato de lista já é presente em outras regras (1, 2, 3, 4) e permite a extração de relações de excertos de texto como:

"... vem estudando profundamente [o fenômeno] , chamado de
[sinantropia] ou [domiciliação] ..."

A regra (v) da Tabela 3.4, é a última regra adaptada do trabalho de Freitas e Quental. Esta foi denominada pelas autoras de "conhecido/a/os/as como", devendo extrair relações de excertos como:

"[vesículas esféricas de gordura] , conhecidas como [lipossomas]"

Obtendo a relação Hiponímia (lipossomas, vesículas esféricas de gordura). Após o processo de adaptação, a regra ganhou a seguinte representação:

(8) SN((,)? também)?(,|é|são|foram)? conhecid(o|a|os|as) como (SN ,)*SN (e|ou))*SN"

Para maximizar o número de relações extraídas, assim como na regra (7), foram realizadas alterações para permitir a presença de vírgula e das formas verbais “é”, “são”, “foi” e “foram” antes da expressão “conhecido como”, como também, a presença de uma lista de sintagmas nominais após a expressão. Ainda foi alterada a regra para permitir a presença da palavra “também” após o primeiro sintagma nominal.

4.2.3.2 Considerações

As regras propostas por Freitas e Quental extraem uma quantidade menor de relações se comparadas às regras propostas por Hearst, mas “apresentaram um alto índice de precisão”, conforme [Fre07].

Na Tabela 4.2 são associadas ao presente trabalho as regras propostas por Freitas e Quental.

Tabela 4.2 - Associação entre padrões de Freitas e Quental e os do presente trabalho

6	tipos de SN Hiper: SN 1 { , SN 2 ... ,} (e | ou) Sni
7	SN HHiper chamado/s/a/as (de) SN Hipo
8	SN Hiper conhecido/s/a/as como SN Hipo.

Na adaptação desenvolvida no trabalho corrente, foram criadas três regras, onde cada uma corresponde a uma regra presente no trabalho de Freitas e Quental. Analisando o trabalho de Freitas e Quental, é possível notar um formato de sintagma nominal que está ausente nas regras do presente trabalho: “SN HHiper”. Em [Fre07] foi utilizado este prefixo para os sintagmas nominais com o objetivo de melhorar a precisão das extrações. O SN HHiper é utilizado para identificar apenas a

primeira palavra encontrada mais à direita de um sintagma nominal. Exemplo: “[a administração de **medicamentos**]”. Já os “SN Hipo” e “SN Hiper” são utilizados para representar um sintagma nominal como elemento hiponímico ou hiperonímico da relação.

4.2.4 Taba e Caseli

O trabalho de Taba e Caseli [Tab13] assemelha-se com o presente trabalho por estudar o modo como relações semânticas podem ser automaticamente extraídas de corpora de língua portuguesa. Taba e Caseli estudam tanto a abordagem baseada em aprendizado de máquina quanto a abordagem baseada em regras. Durante sua pesquisa os autores utilizaram os padrões criados por Freitas e Quental assim como outros padrões de sua própria autoria. Destes padrões propostos, abordaremos apenas os padrões v, vi e vii da Tabela 3.6, pois estes realizam extração de relações hiponímicas (denominadas em [Tab13] de relações “is-a”) e foram propostos pelos autores.

O primeiro padrão adaptado foi o padrão (v). Este padrão busca extrair relações de excertos de texto como:

“... apresentar [febre] ou [qualquer outro sintoma da doença de Chagas] ...”

Este padrão obtem a relação Hiponímia (febre, sintoma da doença de Chagas). A representação da adaptação construída com base nesta regra pode ser vista a seguir:

(9) (SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>

Na regra original em [Tab13] eram permitidas apenas as palavras “outro” ou “outros” antes do último SN. No corrente trabalho foi flexibilizado esse modelo para que a palavra no gênero feminino também fosse válida (“outra”, “outras”). Assim como em outras regras, foram utilizados os sinais “>” e “<” para indicar que as palavras chaves são encontradas dentro de um *chunk*, e uma subparte deste *chunk* que é representada por “sn” será considerada nas relações extraídas.

Já a regra (vi) presente na Tabela 3.6 é capaz de extrair relações de sentenças como:

“por [a agência local de a Fundação Instituto Brasileiro de Geografia e Estatística] , **[** Pelotas **]** é **[** uma cidade **]** [cuja zona urbana comporta 297.825 habitantes]”

No caso, é obtida a relação Hiponímia (Pelotas, cidade). A regra (vi) foi subdividida em duas regras no momento da adaptação. Estas podem ser vistas a seguir:

(10.A) SN é < (o|a) sn>
(10.B) SN é < (um|uma) sn>

Como pode ser observado, as regras apresentam semelhanças. O motivo da criação de duas regras é o fato de elas serem generalistas. Como elas extraem um grande número de relações, foi realizada esta divisão para que futuras análises possam determinar a precisão das regras individualmente. Ambas as regras apresentam a estrutura que indica que as palavras chaves estão dentro do *chunk*.

A regra (iii) visa extrair relações de excertos tal como no exemplo a seguir:

“[as hemoglobinopatias **]** são **[** doenças geneticamente determinadas **]** e apresentam [morbidade significativa] em todo o mundo.”

Obtendo a relação Hiponímia(as hemoglobinopatias, doenças geneticamente determinadas).

A seguir podemos ver a última regra adaptada com base em [Tab13]:

(11) SN são SN

A construção dessa regra reflete basicamente a transcrição do padrão para a sintaxe utilizada neste trabalho. Isto se deve ao fato de a regra ser extremamente simples.

Tabela 4.3 - Relação entre padrões de Taba e Caseli e o presente trabalho

9	T2 {, T3}* ,? (e|ou) (qualquer|quaisquer) outro{s}? T1
10.A	T2 é (o|a|um|uma) T1
10.B	
11	T2 são T1

Na Tabela 4.3 podem ser vistas as 3 regras adaptadas de Taba e Caseli, com suas correspondências para 4 regras do presente trabalho. O motivo de a regra 10 ser subdividida em duas se deve a esta ter duas regras correspondentes no presente trabalho.

4.3 Resumo

Neste capítulo foi apresentada a proposta que norteia esta dissertação de mestrado. Foram apresentadas as regras propostas e o formato como elas foram escritas. Na Tabela 4.4 todas podem ser vistas, na ordem em que foram apresentadadas ao longo do capítulo.

Tabela 4.4 – Grupo de padrões propostos no presente trabalho

1	SN(,)? como (SN ,)*(SN (e|ou))*SN
2	SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN
3	SN(,)? incluindo (SN ,)*(SN (e|ou))*SN
4	SN(,)? especialmente (SN ,)*(SN (e|ou))*SN
5	(SN (ou|e|,))*<outr(a|o)(s)? sn>
6	<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN
7	SN(,| é| são| foram)? chamad(o|a|os|as)(de)? (SN ,)*(SN (e|ou))*SN
8	SN((,)? também)?(,|é|são|foram)? conhecid(o|a|os|as) como (SN ,)*SN (e|ou))*SN"
9	(SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>
10.A	SN é < (o|a) sn>
10.B	SN é < (um|uma) sn>
11	SN são SN

Na Tabela 4.4 encontram-se todos os padrões propostos neste capítulo. Nos proxímos capítulos se discutirá a utilização desses padrões na contrução de um protótipo. Por fim, as relações extraídas serão avaliadas e os resultados analisados.

5. PROTÓTIPO E APLICAÇÃO DAS REGRAS

Com o objetivo de implementar e testar um extrator de relações hiponímicas de textos em português com base nos padrões trabalhados, foi desenvolvido um protótipo funcional cuja arquitetura é descrita neste capítulo. Também serão descritas as etapas de processamento empregadas ao longo da execução do protótipo. Ainda neste capítulo é apresentado o corpus escolhido para a realização das extrações.

5.1 Arquitetura

A arquitetura proposta para a criação do protótipo consiste de um conjunto de etapas sequenciais, onde a saída gerada por uma etapa alimenta a próxima etapa.

O processo inicia pela inserção do corpus como um parâmetro de entrada. Logo o processo de formatação age sobre todo o corpus e retorna como parâmetro de saída um corpus em um formato mais adequado para as próximas etapas. Então o processo de aplicação de regras entra em ação, executando as regras criadas, sobre cada sentença. Como resultado este processo retorna todos os trechos de sentenças que foram identificados pelas regras. Na última etapa estes trechos são inseridos como parâmetro de entrada para o processo de extração.

Nesse processo as relações resultantes são criadas e então é retornada uma lista com todas as extrações obtidas pela execução do protótipo. Este processo é ilustrado na Figura 5.1.

Figura 5.1 – Ilustração da arquitetura utilizada na construção do protótipo

Cada etapa do processo será descrita com maior detalhe nas próximas seções.

5.2 Expressões Regulares

Expressão regular é uma composição de símbolos que, agrupados, provêm uma forma concisa de identificar cadeias de caracteres, palavras ou um padrão de texto. As expressões regulares são escritas em linguagem formal e podem ser interpretadas por um processador de expressões regulares. Este examina o texto e procura por trechos que atendam às regras determinadas pela expressão.

Expressões regulares são importantes para o atual trabalho por representarem as regras propostas e extrair as relações textuais. A escolha desse método se deu pela sua simplicidade e expressividade, assim como por estar disponível para uso em diversas linguagens de programação.

5.3 Corpus

O corpus utilizado como entrada para experimentar o protótipo desenvolvido foi o CORSA (Corpus de Saúde Pública, descrito em [Fre07]).

Este corpus é formado por 1.846.502 palavras dispostas em um arquivo de 11Mb. O CORSA foi criado com base em textos da área de saúde pública, incluindo artigos acadêmicos, cartilhas, manuais, textos divulgados, textos didáticos e também textos jornalísticos. A diversidade das fontes é proposital, com o objetivo de agregar variadas formas de escrita, assim como diferentes níveis de aprofundamento técnico.

Estes conjuntos de textos foram analisados previamente pelo *parser* PALAVRAS [Bic00]. Após a análise, os Sintagmas Nominais (SN) foram etiquetados de acordo com as indicações expostas em [San05]. A escolha deste método se deve a ele ter sido utilizado em um trabalho semelhante [Fre07], permitindo assim uma análise comparativa dos resultados.

No corpus, cada linha apresenta uma palavra com sua etiqueta POS. A palavra é separada de sua etiqueta pelo símbolo “_”. Ainda, no final de cada linha é encontrada uma etiqueta do tipo “BIO” que pode ser “I” para representar o início de um Sintagma Nominal, “O” para representar o fim, ou ainda “B” representando a ocorrência conjunta do fim do SN anterior e início de um novo.. Essa organização pode ser vista na Figura 5.2.

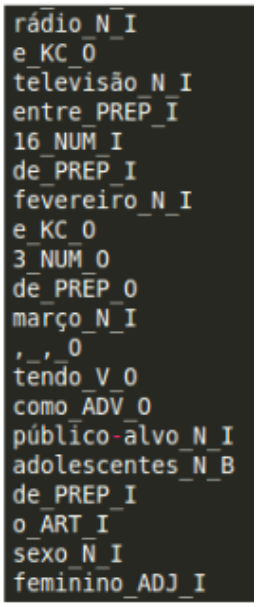

Figura 5.2 – Dados contidos no corpus CORSA

Neste formato de corpus não é possível que um Sintagma Nominal contenha outro, ou seja, aninhamentos de SNs não podem ser representados, nem podem, por consequência, ser empregadas regras recursivas.

A escolha de um corpus já etiquetado foi realizada com o intuito de diminuir a influência do erro na fase de pré-processamento. Assim, possíveis erros nesta fase não são propagados para a fase de avaliação das extrações, evitando o prejuízo à análise dos resultados.

5.4 Formatação do Corpus

Com objetivo de possibilitar o funcionamento com diferentes formatos de corpus e ainda facilitar a criação das regras, o corpus de entrada é convertido para um formato específico. Assim mesmo, é possível desenvolver conversores de formatos específicos para o formato padrão utilizado pelo software.

O formato adotado aceita sentenças descritas textualmente, com apenas um destaque para os sintagmas nominais. Estes estão entre colchetes, como pode ser visto a seguir:

“... entre **[** os municípios maiores **]** , **[** Cáceres **]** e **[**Rondonópolis **]** são ...”

Esse formato é aplicado a todo corpus, onde, após o processamento, cada sentença é adicionada a uma lista para se dar início à próxima etapa.

5.5 Aplicação das Regras

Após o pré-processamento do corpus, é iniciada a etapa de aplicação das regras. Nesta etapa a lista de sentenças é percorrida e, para cada sentença, todas as regras são aplicadas em forma de expressões regulares. Quando uma expressão “combina” (*matches*) com uma sentença, se dá início à etapa de Identificação dos termos da relação.

Ao longo do trabalho de prototipação foi preciso adicionar diversas regras e alterá-las. Foi percebido que era necessário simplificar este processo, já que, até então, era necessário escrever todo o código para a criação e aplicação de cada regra. Assim, foi adotado o conceito do armazenamento de regras em arquivo externo. As regras foram escritas em um arquivo externo, e este arquivo foi usado como entrada na etapa de aplicação das regras. O arquivo de entrada consiste de um documento JSON (*Java Script Object Notation*) com todas as regras listadas por autor. Este formato de documento foi adotado por ser um padrão leve, de simples implementação e alta expressividade.

5.6 Extração

Nesta etapa a relação já foi identificada na sentença, mas ainda é necessário identificar quais dos SNs compõem cada relação extraída, já que uma regra pode identificar mais de uma relação binária. Além disso, é necessário identificar qual sintagma nominal é o termo hiponímico e hiperonímico da relação.

Por fim é gerada uma lista com todas as relações encontradas, no seguinte formato:

Sentença:

{Sentença analisada}

Extrações:

{Autor}-{Padrão} {Nome da Relação}({Argumento1}, {Argumento2})

...

6. ANÁLISES COMPARATIVAS E AVALIAÇÃO

Neste capítulo será abordado o processo avaliativo desenvolvido de modo a analisar os resultados obtidos. Inicialmente serão apresentados desafios enfrentados na avaliação, seguindo-se um relato da metodologia de avaliação proposta por Freitas e Quental em [Fre07].

É então descrito o processo avaliativo aqui empregado, e é oferecida uma minuciosa análise preliminar comparativa. Os resultados da avaliação são discutidos no Capítulo 7.

6.1 Desafios da Avaliação

Durante a execução das etapas de avaliação, diversas dificuldades foram encontradas. Entre elas podemos destacar o grande número de relações extraídas pelo protótipo, que impossibilitou a análise manual de todas as extrações. Outro motivo que dificultou a execução da análise manual foi a falta de uma equipe que contasse com o número apropriado de avaliadores para realizar o processo avaliativo manual comum nessa área. Neste trabalho pudemos contar com dois avaliadores, ambos com dedicação parcial.

A possibilidade de avaliação automática foi descartada, pois esta se tornou inviável devido à indisponibilidade de um *Gold Standard* na língua portuguesa, com o qual os resultados poderiam ser comparados.

Durante o processo de avaliação de resultados torna-se necessário situar o trabalho perante a bibliografia, para isto é preciso comparar os resultados com os de outros autores. Na literatura encontramos poucos trabalhos que realizam a extração de relações em corpora de língua portuguesa e, dentre estes, não foi possível encontrar resultados que possam ser considerados um *Gold Standard*, a partir dos quais possam ser calculadas a precisão e a cobertura.

6.2 Metodologia de Avaliação Proposta por Freitas e Quental

Freitas e Quental [Fre07] realizaram a avaliação de seus resultados em dois formatos. No primeiro, as autoras analisaram os resultados dos padrões por elas propostos, individualmente, em busca de erros sintáticos.

O objetivo era a eliminação dos erros mais frequentes para cada padrão. Já no segundo formato de avaliação, que o presente trabalho toma como principal referência para o processo avaliativo empregado, foi realizada uma validação humana onde o foco era tornar os resultados “mais comparáveis” e “mais significativos”. As relações foram pontuadas com base nos critérios apresentados na Tabela 6.1.

Tabela 6.1 – Critérios de avaliação extraídos de [Fre07]

Nota	Descrição
3	A relação está correta da forma como foi extraída.
2	A relação está “um pouco” correta, isto é, o substantivo núcleo está correto, mas preposições, adjetivos, etc. que o acompanham deixam a relação estranha.
1	A relação está correta em termos gerais; isto é, é muito geral ou muito específica para ser útil.
0	A relação está errada.

Neste processo desenvolvido por Freitas e Quental três avaliadores realizaram a análise em conjunto, isto é, para cada uma das 436 relações avaliadas (cerca de 1/3 do total das extraídas), o resultado foi obtido com o consenso entre os três. Estes avaliadores tinham formação em biologia, educação física e direito, ou seja, bastante diversificada.

6.3 Descrição do Processo Avaliativo

Para realizar os testes e a avaliação, o corpus CORSA foi utilizado como parâmetro de entrada para o protótipo construído em nosso trabalho. A escolha desse corpus teve o intuito de permitir a comparação de resultados com os descritos em [Fre07], mesmo corpus utilizado por Freitas e Quental. Após a execução, onde todas regras propostas foram aplicados ao corpus, foi realizada uma etapa de avaliação das relações extraídas.

Para a realização da avaliação foi adotada a metodologia comparativa. Para isto foi selecionado um subgrupo do total de relações, composto por todas as extrações realizadas pelas regras 6, 7 e 8. Estas regras foram escolhidas por terem extraído uma quantidade aceitável de relações (218) e por pertencerem ao conjunto de regras adaptadas do trabalho de Freitas e Quental [Fre07]. Para este propósito dois juízes humanos, que não receberam treinamento prévio, analisaram 218 relações extraídas, sob os mesmos critérios utilizados no processo avaliativo usado por Freitas e Quental, e descrito na Seção 6.2.

O processo de análise de resultados do presente trabalho foi realizado individualmente por cada um dos dois avaliadores humanos. Estes atribuíram notas de zero a três às extrações e, calculada a média entre as avaliações, foi realizada a análise levando em consideração exclusivamente os resultados em que houve concordância entre os avaliadores.

6.4 Resultados Obtidos e Análise Detalhada

É descrita aqui uma primeira análise dos resultados obtidos. Após a aplicação das onze regras sobre o corpus CORSA, extraímos 8601 relações que foram subdivididas em três grupos, compostos respectivamente pelas relações obtidas aplicando as regras baseadas nas obras dos autores de referência: Hearst [Hea92], Freitas e Quental [Fre07] e Taba e Caseli [Tab13]. O número total e o percentual de relações obtidas em cada caso consta na Tabela 6.2.

Tabela 6.2 – Número de relações extraídas por autor de referência

Autor	Número de Relações	Percentual
Hearst	5936	69,02%
Freitas e Quental	218	2,53%
Taba e Caseli	2447	28,45%
Total	8601	100,00%

Conforme a Tabela 6.2, as regras provenientes de Hearst em [Hea92] foram as mais produtivas, gerando 69,2% das 8601 relações obtidas. Já Taba e Caseli motivaram o segundo grupo mais produtivos de regras com 28,45% do total de relações obtidas. Por fim as regras provenientes de Freitas e Quental [Fre07] geraram 2,53% do total.

Os dados representados na Tabela 6.2 demonstram que as regras baseadas em [Fre07] extraíram poucas relações, já as regras baseadas no trabalho de Taba e Caseli obtiveram maior número. Mas grande parte do total pertence aos grupos das relações extraídas pelos padrões propostos por Hearst [Hea92].

Tabela 6.3 – Número de relações extraídas por regras adaptadas de Hearst [Hea92]

Regras	Número de Relações	Percentual
1	4565	76,90%
2	351	5,91%
3	578	9,74%
4	376	6,33%
5	63	1,06%
Total	5936	100,00%

Na Tabela 6.3, é exibido o número de relações obtidas e o valor percentual em relação ao total de 5936 extrações. As regras referenciadas foram inspiradas em [Hea92] e apresentadas na Seção 4.2.2.

A regra número 1, que busca extrair relações por meio da palavra chave “como” extraiu um número grande de relações, representando 76,9% das extrações.

Este resultado já era esperado, pois a palavra chave em questão é comum na língua portuguesa. Este grande número de relações influenciou fortemente que as relações obtidas com base em padrões propostos por Hearst tenham apresentado o número maior de extrações (Tabela 6.2), em nossa análise.

Já as regras extraídas com base no trabalho de Freitas e Quental (vide Tabela 6.4) tiveram um número significativamente menor de relações extraídas, apenas 218. Isso se deve ao fato de estas relações serem mais específicas, ou seja, são baseadas em termos com menor frequência em textos em língua portuguesa. O número de relações extraídas para cada regra adaptada de Freitas e Quental na Tabela 6.4 mostra que a regra 6, representada como “<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN”, teve a melhor performance, extraindo 44,95% das relações provenientes de Freitas e Quental.

Tabela 6.4 – Número de relações extraídas por regras adaptadas de Freitas e Quental [Fre07]

Regras	**Número de Relações**	**Percentual**
6	98	44,95%
7	75	34,40%
8	45	20,64%
Total	218	100,00%

Com as regras adaptadas do trabalho de Taba e Caseli [Tab13] foi possível extrair 2447 relações cuja distribuição é apresentada na Tabela 6.5. Parte das regras são abrangentes, obtendo alto número de relações, principalmente as regras que baseiam-se em expressões como “é um” e “são”. Este comportamento, como visto na Tabela 6.5, leva a uma distribuição que é semelhante em percentual para as regras 10, 11 e 12.

Tabela 6.5 – Número de relações extraídas por regras adaptadas de Taba e Caseli [Tab13]

Regras	Número de Relações	Percentual
09	23	01,00%
10.A	920	37,59%
10.B	694	28,36%
11	810	33,10%
Total	2447	100,00%

Ainda analisando a quantidade de relações extraídas com base em [Tab13], a discrepância nessa quantidade fica evidente com relação à regra 9, que apresenta uma quantidade de extrações muito inferior. Esta regra baseia-se na combinação das palavras “qualquer” e “outros” que é menos comum na língua portuguesa, tornando-se uma regra específica, e menos produtiva.

7. AVALIAÇÃO E DISCUSSÃO DOS RESULTADOS

Neste capítulo será trazido em maior detalhe o processo de realização de testes, e serão apresentados e discutidos os resultados da avaliação. Também serão analisados os erros frequentes que foram identificados.

7.1 Análise dos Resultados

Para validar individualmente as regras propostas neste trabalho foi conduzido um processo de avaliação das relações extraídas. Devido ao grande número de relações e à dificuldade de encontrar um *Gold Standard,* para realizar uma comparação automatizada, foi utilizado o processo de avaliação manual dos resultados, assim como também é relatado na literatura.

Devido ao fato de o total de resultados ser superior a 8 mil relações, a análise manual tornou-se inviável no tempo disponível. Então foi estabelecido um subgrupo de relações. Foram escolhidas as relações extraídas com base nas regras adaptadas de [Fre07], e com o total formado por estas (218 extrações) foi possível realizar a avaliação manual. Os dados provenientes da avaliação estão disponíveis no Apêndice A. Estes são apresentados em uma tabela onde os parâmetros das relações, assim como as notas de cada avaliador, estão representados na forma de colunas. Outro motivo importante para a escolha das relações utilizadas nessa etapa foi a possível comparação de resultados com o trabalho de Freitas e Quental [Fre07], já que este utilizou o corpus CORSA, mesmo corpus do presente trabalho.

O Avaliador 1 classificou cada resultado em um de quatro grupos que são representados por notas que variam de 0 a 3, gerando os dados presentes na Tabela 7.1.

Tabela 7.1 – Resultado da Avaliação 1: Total de relações encontradas por nota de avaliação

Nota	Relações	Percentual
0	29	13,3%
1	41	18,8%
2	46	21,1%
3	102	46,8%

Analisando a Tabela 7.1 reparamos que um total de 46,8% de relações extraídas com 100% de correção não é um valor alto. Por outro lado, apenas 13,3% das relações foram consideradas totalmente erradas, o que é um resultado promissor.

Na segunda avaliação, feita pelo Avaliador 2, obtivemos resultados semelhantes, como mostra a Tabela 7.2.

Tabela 7.2 – Resultado da Avaliação 2: Total de relações encontradas por nota de avaliação

Nota	Relações	Percentual
0	26	11,9%
1	53	24,3%
2	41	18,8%
3	98	45,0%

No caso do Avaliador 2 os resultados se assemelham com os obtidos na avaliação 1, com um leve desvio nas relações classificadas com nota 1 e 2, o que pode demonstrar alguma dificuldade em trabalhar-se com a escala proposta por Freitas e Quental.

Para obter um resultado composto das avaliações, foi calculada a média aritmética entre valores obtidos pelos avaliadores para cada uma das quatro possíveis notas (Tabela 7.1 e 7.2). Assim foi calculado o resultado composto por ambas as avaliações. Esse resultado está disponível na Tabela 7.3, com o percentual referente à média aritmética.

Tabela 7.3 – Resultado da avaliação composta

Nota	Percentual
0	12,6%
1	21,6%
2	19,9%
3	45,9%

Também foi realizado o cálculo da média aritmética entre ambas as avaliações, para cada uma das regras cujas relações foram avaliadas. Esse processo obteve o seguinte resultado exposto na Tabela 7.4.

Tabela 7.4 – Percentual médio de relações encontradas por nota de avaliação e por regra

Regra\Nota	0	1	2	3
6	17,8%	31,1%	13,3%	37,8%
7	12,8%	16,8%	21,9%	48,5%
8	09,3%	22,0%	21,3%	47,3%

Analisando a Tabela 7.4 constata-se que a regra 6, que corresponde a “tipos de”, apresenta um desempenho consideravelmente inferior ao das outras regras, cerca de 10% menos relações corretas foram encontradas para esta regra.

Outro ponto que é interessante analisarmos é a diferença entre os julgamentos atribuídos por cada avaliador, como mostra a Tabela 7.5.

Tabela 7.5 – Comparação entre resultados de julgamento pelos avaliadores

Nota	Avaliações idênticas
0	13
1	14
2	13
3	69
Total	109

Analisando a Tabela 7.5 constata-se que o número de relações que receberam a mesma nota pelos avaliadores é consideravelmente baixo, 50% das relações avaliadas receberam uma nota diferente de cada um dos dois avaliadores. Este resultado demonstra a diferença nos critérios de cada avaliador ao determinar se uma relação está correta. Um exemplo dessa diferença entre critérios pode ser visualizada nas seguintes relações:

A. Hiponímia (técnicos de segurança de o trabalho; profissionais)

B. Hiponímia (transtorno de a compulsão alimentar periódica; transtorno alimentar)

C. Hiponímia (questionário individual de homens; questionários)

D. Hiponímia (questionário individual de mulheres; questionários)

E. Hiponímia (colinesterase verdadeira; colinesterases)

Todas estas relações foram avaliadas com nota 3 pelo processo de avaliação realizado em [Fre07], já no processo de avaliação realizado neste trabalho estas relações receberam notas distintas, como nos mostra a Tabela 7.6.

Tabela 7.6 – Comparação entre julgamentos para 5 relações específicas

Relação	Avaliador 1	Avaliador 2
A	3	3
B	3	1
C	2	3
D	2	3
E	3	1

Na Tabela 7.6 pode-se notar que apenas a relação A obteve o mesmo resultado nas três avaliações.

A discordância entre os avaliadores sugere que os critérios de julgamento são ambíguos. Na avaliação realizada em [Fre07] os resultados são obtidos por meio do consenso de três avaliadores. No corrente trabalho as avaliações foram realizadas de maneira independente. Seguindo este critério de consenso podemos prover uma

nova análise dos resultados, considerando apenas as ocorrência onde os autores obtiveram concordância. Esta é mostrada na Tabela 7.7.

Tabela 7.7 – Resultado da avaliação para os casos de concordância entre avaliadores

Nota	**Percentual**
0	11,9%
1	12,8%
2	11,9%
3	63,3%

Esta abordagem com relação à concordância permite ter uma confiança maior nos resultados obtidos, tornando-se um recurso para evitar erros individuais cometidos pelos avaliadores. Comparando a Tabela 7.3 com a Tabela 7.7 fica evidente um aumento no percentual de relações consideradas completamente corretas. Este fato pode ser atribuído à subjetividade dos critérios de avaliação que caracterizam os grupos de nota 1 e 2.

Outra forma utilizada para elucidar os resultados é a comparação relativa por regra, considerando apenas os resultados obtidos levando em conta a concordância entre as avaliações.

Tabela 7.8 – Percentual médio de relações encontradas por critério de avaliação e por regra, segundo critério de concordância entre avaliadores

Regra\Nota	**0**	**1**	**2**	**3**
6	11,1%	9.3%	14,8%	64.8%
7	8,1%	16,2%	10,8%	64,9%
8	22,2%	16,7%	6,25%	55,6%

Considerando os resultados mostrados na Tabela 7.8, as regras 6 e 7 apresentam resultados semelhantes. Já a regra 8 apresenta um resultado inferior. Este fato indica que a regra 8 apresenta uma precisão inferior, se comparada com as regras 6 e 7.

7.2 Análise dos Erros

Analisando as relações que obtiveram classificação zero levando em conta oresultado de ambos os avaliadores, podemos destacar alguns motivos de erros mais frequentes. Um destes é o erro de *chunking*, quando o *parser* realiza uma identificação incorreta. Este erro foi apontado por [Cor13] como uma das principais fontes de incorreções em seu trabalho. Este erro ocorre após a etapa de tokenização, quando o *chunker* identifica os sintagmas nominais. E é exemplificado a seguir, onde o *parser* identificou incorretamente a letra “o” como sendo um sintagma nominal.

“... [dois tipos de modelos] : **[** o **]** logístico e **[** o **]** hierárquico ...”

.

Em alguns casos um sintagma nominal pode ser subdividido em SNs menores sem de fato gerar um erro sintático (vide Figura Figura 4.1). Este comportamento não pode ser considerado uma falha no *chunker*, pois tecnicamente tanto a identificação de um Sintagma Nominal composto (formado por um grupo de SNs), quanto a identificação de apenas um subelemento desse conjunto estão corretas, mas este comportamento gera resultados incoerentes. Exemplos podem ser vistos nas sentenças a seguir.

“[o aparecimento de anticorpos] em **[** o sangue **]** ,
chamado de **[** janela imunológica **]**”

O *parser* identificou “[o aparecimento de anticorpos]” e “[o sangue]” como sendo dois SNs distintos, gerando uma possível extração errada: Hiponímia (o sangue, janela imunológica). Caso o *parser* identificasse ambos SNs como um só, uma relação mais precisa poderia ser extraída: Hiponímia (o aparecimento de anticorpos em o sangue, janela imunológica). Para corrigir esta falha seria preciso de um *chunker* que agrupasse os SNs nesses casos. Outra solução seria prover uma etapa de pré-processamento que unisse *chunks* em situações específicas.

Outro erro encontrado é o erro de correferência. Este acontece quando o sintagma nominal faz referência a outro SN que foi citado anteriormente na sentença. Um exemplo pode ser visto no trecho a seguir, onde o SN faz referência a “corpo”.

“tornar dócil [um corpo] não é [coisa simples] , pois ele , normalmente , está submetido a [seu chefe natural] , chamado [personalidade]”

Uma extração adequada para essa sentença seria Hiponímia (personalidade, chefe natural do corpo). Uma abordagem para solucionar este problema seria a utilização de métodos criados em trabalhos na área de resolução de correferência, como, por exemplo, [Sto10] e [Lee11].

Outro erro encontrado se refere à falta de contexto. Este erro ocorre quando o termo é extraído corretamente, mas ele só faz sentido quando está inserido em um determinado contexto. Abaixo segue um exemplo.

“... [a segunda fase] , chamada de [análise] ...”

A regra está correta em extrair a relação Hiponímia (a segunda fase, análise), mas como não sabemos a que entidade a palavra “fase” faz referência, a extração perde o significado, se analisada fora do seu contexto.

Outro erro encontrado está presente na expressão que explora relações formadas por listas de SNs. Esta expressão considera que todos os SNs seguidos por “e”, “ou” e “,” fazem parte da mesma lista, mas em determinados casos estes conectores podem apenas ligar duas sentenças, não tendo a função de criar lista de sintagmas nominais. Seguem alguns exemplos.

“[um gênero de vírus] conhecido como [flavivírus] , [a enfermidade] apresenta ...”
“[a bactéria] chamada [Rickettsia mooseri] e [os sintomas] são praticamente ...”

A relação Hiponímia (a enfermidade, um gênero de vírus) é extraída indevidamente, assim como Hiponímia (os sintomas, a bactéria). Apesar de, em ambas as sentenças, o padrão ser aplicado corretamente no primeiro SN, o segundo sintagma nominal é considerado indevidamente como parte da lista.

Já quando analisamos as relações apontadas por ambos os avaliadores como pertencendo ao grupo 1, o erro mais comum encontrado é a aparição de palavras desnecessárias para o significado da relação dentro de um dos sintagmas nominais. A seguir podem ser vistos exemplos deste fenômeno.

“[a ação de os vírus] conhecidos como [Influenza A]”
“[essas lesões] , chamadas de [isquemia]”

As relações extraídas nesse caso são Hiponímia (a ação de os vírus, Influenza A) e Hiponímia (essas lesões, isquemia). Caso as relações extraídas fossem respectivamente Hiponímia (influenza A, vírus) e Hiponímia (isquemia, lesões) as relações obteriam uma classificação melhor. Para solucionar este tipo de problema as autoras Freitas e Quental criaram uma etapa de pós-processamento automatizada, que aplica filtros para remover palavras dos sintagmas nominais que não agreguem significado à relação. Uma etapa semelhante poderia ser utilizada no trabalho atual com o objetivo de melhorar a precisão, mas para isso é necessário dispor de uma lista de palavras que frequentemente não agregam valor semântico, como por exemplo preposições e pronomes.

7.3 Discussão dos Resultados

Ao longo deste capítulo foram relatados os resultados encontrados em todos os testes realizados. Para fins comparativos será considerado que a precisão das relações extraídas pelas regras analisadas é 63,3%, com base nos resultados ilustrados na Tabela 7.7. Uma das etapas mais complexas e subjetivas é a comparação de resultados. Como não é possível obter um *Gold Standard* a comparação é feita com outros trabalhos. A dificuldade de avaliar os resultados por comparação está no uso de regras diferentes por cada autor, assim como a escolha de corpora distintos, e ainda etapas distintas de pré-processamento ou pós-processamento. Ocorre também uma discrepância entre os avaliadores, que são de áreas do conhecimento diferentes e de contextos culturais distintos. Estes elementos provocam incerteza nas avaliações manuais.

Para eliminar um dos elementos citados acima, a primeira comparação de resultados realizada será em relação ao publicado em [Fre07], onde é utilizado em uma das etapas o corpus CORSA. Em uma das etapas avaliativas a autora afirma obter 73,4% quando aplicou as regras “como/tais como”, “e outros”, “tipos de”, “chamado” e “conhecido como” sobre o corpus CORSA. O motivo de este resultado ser expressivamente superior ao do presente trabalho pode ser explicado pela primeira etapa de avaliação realizada por Freitas e Quental. Nesta etapa foi realizada uma análise manual sobre o resultado e foram removidas 726 relações consideradas sintaticamente erradas. É válido ressaltar que o resultado de 73,4% obtido por Freitas e Quental é um resultado parcial, já que este processo considera a segunda etapa de avaliação realizada pela autora sobre extrações no corpus CORSA. Na conclusão de seu trabalho Freitas e Quental consideram seu resultado final como sendo 75%, este calculado utilizando o corpus CETEN-Folha, sem a realização da primeira etapa onde são removidas manualmente relações sintaticamente errôneas, mas já utilizando os filtros propostos em [Fre07].O primeiro filtro proposto remove relações cujo argumento hiperonímico trata-se de substantivo com um alto grau de generalidade ou falta de especificidade. Outros dois filtros aplicados em [Fre07] buscam remover palavras que não agregam valor semântico. Com este objetivo o primeiro filtro remove pronomes dêiticos e o segundo remove alguns adjetivos.

Também é possível realizar comparações com outros autores como Hearst e Morin e Jacquemin, mas sempre levando em conta a diferença entre corpora, processo avaliativo, e também idioma.

Tabela 7.9 – Comparação dos resultados obtidos

	Corpus em Língua Portuguesa		Corpus em Língua Estrangeira		
	Presente Trabalho	Freitas e Quental (2007)	Morin e Jacquemin (2004)	Cederberg e Widdows (2003)	Hearst (1998)
Precisão	63%	73,4%	81%	64%	63%

Analisando a Tabela 7.9 é possível constatar que os resultados obtidos assemelham-se àqueles obtidos por outros trabalhos na área. Consideramos, deste modo, que o presente trabalho cumpre com o objetivo proposto de extração de relações hiponímicas em corpora de língua portuguesa. Ainda assim existem diversas técnicas que foram citadas neste trabalho e poderiam melhorar os resultados obtidos, permitindo atingir uma precisão semelhante à dos trabalhos de Freitas e Quental e Morin e Jacquemin, como por exemplo a filtragem de palavras que não agregam valor semântico ou ainda uma etapa de pré-processamento que una *chunks* em situações específicas (Vide Seção 7.2).

8. CONSIDERAÇÕES FINAIS

Neste capítulo são discutidas as contribuições oferecidas para a área de extração de relações em textos de língua portuguesa, assim como as perspectivas futuras para a continuidade deste trabalho.

8.1 Contribuições

Uma das principais contribuições do presente trabalho é a agregação, num único estudo, de regras elencadas por diferentes autores, como as encontradas em [Fre07], [Hea92] e [Tab13], produzindo um trabalho mais completo em termos de escopo e de quantidade de relações extraídas. Outra contribuição é a criação de um protótipo que recebe como entrada um corpus e as regras que devem ser aplicadas ao corpus. Assim, se outras regras precisarem ser implementadas, é apenas necessário inserir estas no arquivo de entrada. Toda a etapa de interpretação das regras, aplicação e extração é abstraída, evitando, em estudos futuros, a necessidade de programação.

Não menos importante, outra contribuição é a análise minuciosa dos resultados obtidos. Estes foram analisados segundo diferentes critérios tais como: por regras, por autor, por nota e por avaliador. Ainda foram discutidos os fatores que tornam subjetivo o processo de avaliação manual.

8.2 Perspectivas Futuras

Devido à restrição de tempo determinada pela duração do curso de mestrado, algumas melhorias idealizadas poderão ser implementadas em uma próxima etapa. Entre elas podemos destacar a criação de uma interface gráfica para simplificar

ainda mais a criação de padrões, contribuindo com trabalhos futuros que visem o uso do interpretador na condição de ferramenta para a extração de relações na língua portuguesa. Outra melhoria no protótipo seria a capacidade de trabalhar genericamente com diversos formatos de corpora. Assim, as mesmas regras poderiam ser facilmente aplicadas a diferentes corpora sem necessidade de retrabalho.

Durante o desenvolvimento deste trabalho ficou evidente a necessidade de criação de um *Gold Standard* para extração de relações hiponímicas na língua portuguesa. Este artefato contribuiria imensamente para o desenvolvimento das pesquisas na área, pois permitiria o cálculo de precisão e cobertura. A tarefa, entretanto, teria de contar com a condução de especialistas, que trabalhariam também questões de escopo, contexto e referência, bem além da etiquetagem de relações, esforço que também teria de ser amplamente registrado, formalizando critérios e condutas adotados.

Durante a etapa de avaliação não foi possível analisar muitas relações, por esse motivo optamos por focar em um grupo contendo apenas regras extraídas com base no trabalho de Freitas e Quental. O ideal seria dispor de um número maior de avaliadores dedicados ao processo, assim poderíamos ter uma cobertura de avaliação maior sobre as regras adaptadas.

8.3 Divulgação de Resultados

Resultados parciais do presente trabalho, na forma de artigo [Mac14], foram apresentados oralmente, como trabalho completo, no Encontro de Linguística de Corpus (ELC 2014), em Uberlândia. A publicação definitiva do evento ainda se encontra em preparação. Mais informações podem ser obtidas no site do evento http://www.elc-ebralc-2014.com.br.

REFERÊNCIAS BIBLIOGRÁFICAS

[Abr13] S. C. Abreu. “Extração de Relações do Domínio de Organizações para o Português”, Tese de Doutorado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2013, 106p.

[Ban07] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni. ”Open Information Extraction from the Web”. In: Proceedings of the Twentieth International Joint Conference, 2007, 7p.

[Bas07] T. L. Baségio. “Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil”, Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2007, 124p.

[Bat13] D. S. Batista, D. Forte, R. Silva, B. Martins, M. J. Silva. “Extração de Relações Semânticas de Textos em Português Explorando a DBpédia e a Wikipédia”, Linguamática: Revista para o Processamento Automático das Línguas Ibéricas, vol 5-1, Jul 2013, pp. 41-57.

[Bic00] E. Bick. “The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework”. In: Aarhus: Aarhus University Press, 2000, 505p.

[Ced03] S. Cederberg, D. Widdows. “Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction”. In: Proceedings of CoNLL-2003, 2003, pp. 111-118.

[Cor13] L. Corro, R. Gemulla. “ClausIE: clause-based open information extraction”. In: Proceedings of the 22th International Conference on World Wide Web, 2013, pp. 355-366.

[Deg04] M. Degeratu, V. Hatzivassiloglou. “An Automatic Method for Constructing Domain-Specific Ontology Resources”. In: Proceedings of

the Language Resources and Evaluation Conference (LREC2004), 2004, pp. 2001-2004.

[Fad11] A. Fader, O. Etzioni. ”Identifying Relations for Open Information Extraction”. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1535–1545.

[Fin99] M. Finkelstein-Landau, E. Morin. “Extracting semantic relationships between terms: Supervised vs. unsupervised methods”. In: Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure, 1999, pp. 71-80.

[Fel98] C. Fellbaum. “WordNet: An Electronic Lexical Database (Language, Speech, and Communication)”, A Bradford Book, 1998, 423p.

[Fre07] C. Freitas, V. Quental. “Subsídios para a Elaboração Automática de Taxonomias”. In: V Workshop de Tecnologia da Informação e da Linguagem Humana, 2007, pp. 1585-1594.

[Gam12] P. Gamallo, M. Garcia, S. Fernández-Lanza. “Dependency-based open information extraction”. In: Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, 2012, pp. 10-18.

[Gru92] T. Gruber. “Ontolingua: A mechanism to support portable ontologies“, Technical Report, Knowledge Systems Laboratory, Stanford University, 1992, 61p.

[Hea92] M. Hearst. “Automatic acquisition of hyponyms from large text corpora.” In: Proceedings of the 14th International Conference on Computational Linguistics, 1992, pp. 23–28.

[Ing12] G. Ingersoll, T. Morton, A. Farris. “Taming Text: How to Find, Organize, and Manipulate It”. Manning Publications Company, 2012, 289p.

[Ins15] Institute of Language and Communication. "Visual Interactive Syntax Learning (VISL)". Capturado em: http://beta.visl.sdu.dk/, Janeiro 2015.

[Jur09] D. Jurafsky, J. Martin. “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”. In: Pearson Education Inc., 2009. 950p.

[Lee07] C. Lee, Y. Kao, Y. Kuo, M. Wang. “Automated ontology construction for unstructured text documents”, Data and Knowledge Engineering, vol. 60-3, Mar 2007, pp. 547–566.

[Lee11] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, D. Jurafsky. “Stanford’s Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task”. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Association for Computational Linguistics, 2011, pp. 28-34.

[Mac14] P. Machado, V. L. Strube de Lima. “Extração de relações hiponímicas aplicada a corpus em língua portuguesa". In: XII Encontro de Linguística de Corpus-ELC, 2014, 6p.

[Mae02] A. Maedche, S. Staab. “Ontology Learning for the Semantic Web”. Massachusetts: Kluwer Academic Publishers, 2002, 272p.

[Mar08] M. S. Chaves. “Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem.” In: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, 2008, pp. 231–245.

[Maz08] E. Maziero, T. Pardo, A. Di Felipo, B. Dias-da-Silva. “A Base de Dados Lexical e a Interface Web do TeP 2 . 0 – Thesaurus Eletrônico para o Português do Brasil”. In: VI Workshop em Tecnologias da Informação e da Linguagem Humana (TIL), 2008, pp. 390-392.

[Mor03] E. Morin, C. Jacquemin. “Automatic acquisition and expansion of hypernym links”. Computer and the humanities, Kluwer Academic Press, vol. 38-4, Nov 2003, pp. 363-396.

[Nas13] V. Nastase, P. Nakov. D. O. Séaghdha, S. Szpakowicz. “Semantic Relations Between Nominals (Synthesis Lectures on Human Language Technologies)”. Morgan & Claypool, 2013, 119p.

[Oli09] G. Oliveira, D. Santos, P. Gomes. "Evaluating the Extraction of Semantic Relations between Portuguese Words by Means of a Dictionary”. In: Simpósio de Tecnologias da Informação e da Linguagem Humana (TIL) , 2009, pp. 8-11.

[Pus12] J. Pustejovsky, A. Stubbs, “Natural language annotation for machine learning”. O’Reilly Media, 2012, 350p.

[Rui05] M. Ruiz-Casado, E. Alfonseca, P. Castells. “Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia”. In: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, 2005, pp. 67-79.

[San01] D. Santos, P. Rocha. “Evaluating CETEMPblico, a free resource for Portuguese”. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 2001, pp. 442-449.

[San05] N. Santos, M. Oliveira. “Aplicação de aprendizado baseado em transformações na identificação de sintagmas nominais”. In: Anais do XXV Congresso da Sociedade Brasileira de Computação, 2005, pp. 2138-2147.

[Sto10] V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, D. Hysom. “Coreference resolution with reconcile”. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 156-161.

[Tab13] L. Taba, H. Caseli. “Automatic semantic relation extraction from Portuguese texts”. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2013, pp. 2739-2746.

[XIL01] XILINX INC. "Virtex Series Configuration Architecture User Guide". Capturado em: http://www.xilinx.com/xapp/xapp151.pdf, Maio 2002.

[Win87] M. Winston, R. Chaffin, D. Herrmann. “A Taxonomy of Part-Whole Relations” Cognitive Science, 1987, pp. 417-444.

[Xav13] C. Xavier, V. L. Strube de Lima, M. Souza. “Open Information Extraction Based on Lexical-Syntactic Patterns”. In: Brazilian Conference on Intelligent Systems (BRACIS), 2013, 6p.

APÊNDICE A - Resultado da avaliação manual

Parâmetro 1	Parâmetro 2	Nota 1	Nota 2
rede pública	segundo tipo de prestador	1	2
as burocracias	Organizações	0	1
as municipais e federais	maternidades	1	0
Pontos	Elementos	1	1
o Cartão da Criança	Registros	2	1
cinco gestores de o sistema municipal de saúde	Profissionais	3	1
quarenta profissionais	Profissionais	2	3
O	modelos	0	0
O	um mesmo tipo de inquirição subjacente	0	0
ficha de domicílio	questionários	3	1
questionário individual de mulheres	questionários	2	3
questionário individual de homens	questionários	2	3
interpretação errônea de as regras de codificação	esse tipo de estudo	1	2
dificuldade	esse tipo de estudo	0	0
o Necator americanus	vermes	3	3
a dengue clássica	Dengue	3	3
a dengue hemorrágica	Dengue	3	1
acetilcolinesterase	colinesterases	2	1
colinesterase verdadeira	colinesterases	3	1
o colesterol total	colesterol	3	3
o colesterol HDL	colesterol	2	1
o colesterol LDL	colesterol	2	1
anorexia nervosa	transtorno alimentar	3	3
bulimia nervosa	transtorno alimentar	3	3
transtorno de a compulsão alimentar periódica	transtorno alimentar	3	1
notificação espontânea de infestação humana por carrapatos	notificação	1	1
um erro	vírus causadores de a gripe	0	0
carne bovina	Cortes	1	1
Clostrídios	bactérias	3	3
o câncer nasofaringeo	câncer não muito comuns	1	3
o linfoma de Burkitt	câncer não muito comuns	1	3
tifo epidêmico	Tifo	3	3
o auditor de convênio	a implantação de um novo tipo de profissional médico	0	3
o pagante ou proveniente de um seguro de saúde particular	um novo tipo de paciente	2	2
adenina	Bases	3	3

cérebro	Tumor	0	1
cólon	Tumor	0	1
cabeça	Tumor	0	1
pescoço	Tumor	0	1
14 acidentes	infortúnios	1	0
empresa com política	limitações associadas com o tipo de contexto	3	1
práticas de segurança existentes ou inexistentes	limitações associadas com o tipo de contexto	1	3
fatores pessoais	Causas	3	2
engenheiros	profissionais	3	3
técnicos de segurança de o trabalho	profissionais	3	3
momento	dois momentos distintos	0	1
produção de a universalidade empírica 15	esta nova situação histórica	3	0
elemento	a lei	0	0
pesquisa sócio-antropológica	esse processo de desvelamento de a realidade	3	3
Posição de Atendimento	cada posto	2	2
personalidade '	seu chefe natural	0	0
análise	a segunda fase	0	0
influenza	um vírus	3	3
dengue hemorrágico	a forma mais grave de a doença	3	3
janela imunológica	o sangue	0	0
halteres	um par de pequenas estruturas	2	3
balancins	um par de pequenas estruturas	1	3
merozoíta	outra fase evolutiva	2	2
Morbillivirus	uma doença infecto-contagiosa causada por um vírus	3	3
Rubivirus rubella	um vírus	3	1
a rubéola	um vírus	1	3
tetanospasmina	uma poderosa toxina	3	3
Rickettsia mooseri	a bactéria	3	3
os sintomas	a bactéria	0	0
Mycobacterium tuberculosis	uma bactéria	3	3
herpes	este quadro	1	0
cisticercos	a ingestão de carne de porco contaminada com larvas	2	3
lactase	uma enzima	3	3
flavivírus	uma doença infecciosa causada por um tipo de vírus	3	3
cujo reservatório natural	uma doença infecciosa causada por um tipo de vírus	0	0
prostaglandinas	o aumento de a concentração de substancias	2	1
carcinoma in situ	uma forma localizada de câncer	3	3
células-tronco	curingas	3	1
blastocisto	cem células	3	1
grupos colaborativos	esses grupos	1	1

cefalotórax	a porção	2	0
Depressoras da Atividade do Sistema Nervoso Central	estas drogas	3	1
basukos	cigarros	3	1
rabdomiólise	uma degeneração irreversível de os músculos esqueléticos	3	3
esquizofrenia	a doença mental	3	3
psicoses	as doenças	3	1
mirações	as alucinações produzidas por a bebida	0	2
benzopireno	uma substância	3	3
síndrome amotivacional	este efeito crônico de a maconha	3	3
Papaver somniferum	uma planta	3	3
drogas opiáceas	estas substâncias todas	1	2
de hidrocarbonetos	um grupo químico	2	2
nicotina	uma substância	3	3
tranqüilizantes	estas drogas	3	2
meprobamato	uma droga	3	3
clordizepóxido	a substância	3	3
de obesologistas	os médicos	3	1
humor aquoso	um líquido transparente	3	1
Período de Incubação	o início de os sintomas	3	2
onicomicoses	as micoses de unha	3	3
síndrome retroviral aguda	uma síndrome semelhante a a mononucleose infecciosa	3	3
pixel	unidades	3	1
substância periarquedutal	uma região de o tronco cerebral	1	1
o neurotransmissor principal responsável	uma região de o tronco cerebral	1	1
perfusor	os testes	0	2
LDL	a participação de uma proteína	2	2
HSP	uma proteína	3	3
skank	laboratório	1	0
Síndrome de Marfan	uma proteína envolvida em uma doença	2	1
causadora de deformações cardiovasculares	uma proteína envolvida em uma doença	2	3
placa bacteriana	uma película muito fina	3	2
TFD	um direito	2	2
ambulatórios gerais	unidades especializadas	2	1
isquemia	essas lesões	1	1
patch clamp	o auxílio de um sofisticado aparelho	3	3
a administração de o salgadão	o auxílio de um sofisticado aparelho	2	0
POL	duas regiões de um importante gene de o vírus de a Aids	2	1
anti-retrovirais	um conjunto de medicamentos	3	3
multimistura	alguns componentes de um	3	2

	suplemento alimentar		
superóxido dismutase	níveis sangüíneos de uma enzima	2	0
ala desaminase	a deficiência em a produção de uma enzima	3	3
luciferina	a substância luminescente produzida por o vaga-lume	3	3
rizoma	um tipo de caule diferenciado	3	3
macrófagos	Células	3	3
fator estimulador de colônias de granulócitos	um composto	3	2
macrófagos	um composto	0	1
hidroxitolueno butilado	um outro ingrediente	1	1
praziquantel	um medicamento a a base de um fármaco	2	3
macrófagos	células imunológicas	3	2
trissomia livre	uma anomalia	3	3
transfecção	uma técnica	3	2
cinetoplasto menos volumosa	uma organela	3	3
sinantropia	o fenômeno	1	2
domiciliação	o fenômeno	2	2
Dicer	uma enzima	3	3
oligopepetidases	um grupo de enzimas	2	3
interferon gama	a produção de moléculas	2	2
pristane	óleo mineral	3	3
apicoplastos	o funcionamento de estruturas	1	0
dextrana	um tipo de açúcar	2	2
magnetotermocitólise	um processo	3	3
MSX 1	mutações em esse gene	1	2
de potencial evocado	o auxílio de um exame	2	0
noradrenalina	uma substância	2	3
Stop Huntingdon Animal Cruelty	protestos de um grupo	3	3
Casa Vital Brazil	uma fundação	3	3
braquiterapia	um tratamento	0	2
laringoscopia	um exame	3	3
Revolução Verde	importante pólo de aplicação de a nova dinâmica de produção agrícola	2	1
índice de Kessner	indicador composto	2	1
redes hierárquicas	redes em árvore	3	2
geografia teorética	a incorporação de o aporte teórico-metodológico de a denominada New Geography	2	3
Lei dos Genéricos	a lei 9.787 de 10 de fevereiro de 1999 3	3	3
tipo 1	três sorotipos	1	0
Brunhild	três sorotipos	0	0
tuberculose primária	esta fase de a infecção	3	1
os indivíduos acometidos geralmente	esta fase de a infecção	1	0

varicela hemorrágica	Forma	2	1
mesêntero	Parte	1	1
incubação	esse período	2	2
tripanossomíase por Trypanosoma cruzi	a doença de Chagas	3	3
tripanossomíase americana	a doença de Chagas	3	3
miracídio	a primeira forma larval de o S. mansoni	3	3
cercária	outra larva	2	3
flavivírus	gênero de vírus	3	3
a enfermidade	gênero de vírus	0	0
cirrose	cicatrizes irreversíveis	0	1
flebotomíneos	insetos vetores ou transmissores	2	2
macrófagos	o interior de células de defesa de o sangue	0	2
a doença de o beijo	angina monocítica	3	3
peste negra	a peste bubônica	3	3
refluxo gastroesofágico	Azia	2	3
Herpes-Zoster	Doença	3	3
síndrome de a dependência de o álcool	o alcoolismo	3	3
rinite alérgica	a inflação alérgica de a mucosa de o nariz	3	3
tosse comprida	a coqueluche	3	3
mal de os sete dias	o tétano neonataltétano	3	3
boneca de larvicida	este artifício	1	1
febre de o dengue	os dois quadros mais distintos	1	2
formas alternativas	outras formas de transmissão	1	3
alcoolismo	quadro de dependência	3	3
alcoolismo	condição esta	0	0
planorbídeos	o gênero Biomphalaria	1	2
febre de as montanhas rochosas	a doença	3	1
micuins	as formas jovens de o carrapato	3	3
Influenza A	a ação de os vírus	1	1
o H5 N1	a detecção de a cepa de alta patogenicidade	3	2
células de o plasma	seus descendentes diretos	0	3
cadeias pesadas	duas cadeias peptídicas mais longas	2	3
h	duas cadeias peptídicas mais longas	1	3
cadeias leves	duas cadeias peptídicas mais curtas	2	3
l	duas cadeias peptídicas mais curtas	1	3
PRP	polímero de d-ribose-ribosil-fosfato	1	3
bromélias	plantas de a família de as Bromeliáceas	3	3

gravatá	plantas de a família de as Bromeliáceas	3	3
caraguatá	plantas de a família de as Bromeliáceas	3	3
acesso malárico	o conjunto de sintomas e sinais	2	2
a doença de o beijo	a virose Mononucleose Infecciosa	3	3
solitária	a teníase	1	2
pediculose	suas cabeças invadidas por uma infestação de piolhos	3	3
bacilo de Koch	Mycobacterium tuberculosis	3	3
long survivers	as crianças	0	1
hibridização	o processo	2	1
Iluminismo	os movimentos culturais e econômicos	3	3
Revolução Industrial	os movimentos culturais e econômicos	2	3
Rede Brasileira de Laboratórios	o projeto	1	2
os sons musicais	todas as direções	2	2
deficiência androgênica parcial	esse processo	2	0
Ramal da Fome	o Vale do Ribeira	2	3
Hospital das Clínicas	o HC	1	3
eNOS	o óxido nítrico	1	2
desfibrilador	aparelho	3	2
estresse oxidativo	condição	1	1
taiuiá	trepadeira	3	2
Fator Potenciador da Bradicinina	a resposta a a bradicinina	3	3
polimorfismos de nucleotídeos únicos	esse tipo de substituição	1	1
SNPs	esse tipo de substituição	1	1
citocinas	proteínas	3	2
EP	endopeptidase neutra	0	3
estreptococo de o Grupo A	a Streptococus pyogenes	3	3
lipossomas	vesículas esféricas de gordura	3	3
estimulação elétrica neuromuscular	a metodologia usada por o pesquisador paulista	3	1
Fototrombose Mediada	o novo procedimento	3	2