UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Geociências e Ciências Exatas – IGCE
Curso de Bacharelado em Ciências da Computação

HELEN DE CÁSSIA SOUSA DA COSTA

ESTUDO DE MÉTODOS PARA ANOTAÇÃO LINGUÍSTICA E EXTRAÇÃO DE CONCEITOS NA AQUISIÇÃO DE ONTOLOGIAS A PARTIR DE TEXTOS

Trabalho realizado sob orientação do Prof. Dr. Ivan Rizzo Guilherme,
DEMAC/IGCE
Período: 02.08 a 04.12.2010

Rio Claro – SP
2010

ESTUDO DE MÉTODOS PARA ANOTAÇÃO LINGUÍSTICA E EXTRAÇÃO DE CONCEITOS NA AQUISIÇÃO DE ONTOLOGIAS A PARTIR DE TEXTOS

Trabalho de Conclusão do Curso, modalidade Trabalho de Graduação, apresentado, no 2º semestre de 2010, à disciplina ES/TG do Curso de Bacharelado em Ciências da Computação, período Integral, do Instituto de Geociências e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", campus de Rio Claro, para apreciação segundo as normas estabelecidas pelo Conselho do Curso, em 27.11.2007.

Aluno: Helen de Cássia Sousa da Costa

Orientador: Prof. Dr. Ivan Rizzo Guilherme
DEMAC – IGCE

Rio Claro – SP
2010

Agradecimentos

Em primeiro lugar agradeço a Deus, que me capacitou para a realização deste trabalho.

Aos meus pais, Paulo e Edna, pelo amor, compreensão e por não medirem esforços pela minha educação.

A meu noivo, Marcos, por sua dedicação, companheirismo e amor, que foram fundamentais para que eu chegasse até aqui.

A meu orientador, professor Ivan Rizzo Guilherme, pelos ensinamentos, conselhos e por saber o momento certo de cobrar e também de incentivar.

A Lucelene Lopes, pela disposição em me ajudar quanto à realização deste trabalho, sua parceira foi fundamental.

A Igreja Batista Maranata, pelo carinho e orações, principalmente ao pastor Idalmar e sua esposa Luciene, que em todo tempo fizeram as vezes de meus pais neste período de faculdade, em que nem sempre pude tê-los por perto.

Aos amigos que fiz durante o curso, principalmente ao Felipe, João e Caio, com os quais compartilhei muitas noites mal dormidas, em prol de uma boa conversa ou de um trabalho de faculdade.

A FAPESP, pelo apoio financeiro durante minha iniciação científica.

LISTA DE FIGURAS

Página

LISTA DE TABELAS

Página

Sumário

Página

1 INTRODUÇÃO

A Web Semântica, que pode também ser chamada como a segunda geração da Web atual, foi proposta por Tim Berners-Lee, o mesmo cientista que inventou a WWW – *World Wide Web* em 1989 (BERNERS-LEE *et al.,* 2001). Tem como ideia principal o uso intensivo de metadados e ontologias na organização semântica de documentos e serviços distribuídos pela rede, utilizando formatos comuns de integração e combinação de dados a partir de diversas fontes, permitindo assim que os serviços tenham a habilidade de reusar e integrar dados, navegar pela rede e automatizar algumas tarefas. Um dos objetivos ao adicionar semântica aos dados é tornar o conteúdo da Web processável pela máquina (BERNERS-LEE *et al.,* 2001).

Uma das bases tecnológicas para a construção da Web Semântica é a utilização das ontologias. Segundo (GUARINO; GIARETTA, 1995), o termo ontologia refere-se a um artefato de engenharia que, em uma visão simplista, pode ser descrito como uma hierarquia de conceitos relacionados entre si através de uma classificação de parentesco, também chamada de taxonomia.

Considerando a importância das ontologias no desenvolvimento da Web Semântica, a construção das mesmas torna-se fundamental. Entretanto, a construção de ontologias não é uma tarefa fácil, sendo adotadas duas abordagens principais: a adoção de metodologias para sua construção manual ou a aquisição a partir de páginas Web ou textos, sendo a última abordagem realizada através de processos automáticos ou semi-automáticos. Há grandes dificuldades em qualquer uma das abordagens de construção de ontologias, pois é um processo complexo, e por ser extremamente artesanal é também propenso a erros. Para a realização da tarefa, geralmente é necessária a presença de um especialista no domínio a ser representado e, em muitos casos, essa pessoa precisa ser treinada na metodologia de construção, na estrutura de representação ou na utilização da ferramenta para a aquisição a partir de textos.

A atividade de aquisição de ontologias a partir de textos requer a utilização de técnicas de Processamento de Linguagem Natural (PLN), que integra conceitos da Linguística Computacional e da Computação no desenvolvimento de programas para tratar os problemas da geração e compreensão automática de textos escritos em um idioma humano. Segundo Ribeiro (RIBEIRO JUNIOR, 2008), a descrição da linguagem natural compõe-se basicamente de três elementos: a descrição formal, a descrição semântica e o sistema que relaciona ambos os planos. A primeira

compreende os elementos fonológicos, morfológicos e sintáticos; enquanto a segunda se relaciona a todos os elementos anteriores através das regras de interpretação semântica. As regras fonológicas, morfológicas e sintáticas definem as construções possíveis da língua, enquanto as semânticas relacionam as construções e seus significados.

Na primeira etapa da aquisição de ontologias a partir de textos, ou seja, a extração de termos, é necessário que os textos utilizados tenham anotações linguísticas que disponibilizem as informações morfológicas e sintáticas dos elementos. Em seguida, regras que expressem padrões encontrados nessas informações são utilizadas para gerar conceitos referentes ao texto. Conceitos que, posteriormente, podem ser organizados em uma estrutura hierárquica (taxonomia).

Uma das vantagens da aquisição de ontologias a partir de textos é a possibilidade de criar ou incrementar ontologias em um período de tempo menor do que o gasto em construções puramente manuais e obter uma qualidade tão elevada quanto possível.

Vários trabalhos têm sido desenvolvidos relacionados à aquisição semiautomática de ontologias a partir de um conjunto relevante de textos de um domínio específico (SANTANA, 2009). Dentro deste contexto, este trabalho tem como objetivo dar continuidade ao desenvolvimento do PhDic (GUILHERME *et al.,* 2006) - ferramenta computacional utilizada na construção do conhecimento e ontologias a partir de relatórios técnicos de perfuração e produção de petróleo. Em particular, no aprimoramento dos métodos de anotação linguística e extração de conceitos. No desenvolvimento do trabalho foram analisados trabalhos correlatos, visando adotar novas abordagens.

2 CONCEITOS E FERRAMENTAS UTILIZADAS

2.1 ANALISADOR SINTÁTICO: PALAVRAS

É um *parser* que processa e disponibiliza textos anotados com informações linguísticas, divididas em duas estruturas: PoS (*part-of-speech*) e sintagmas. A primeira estrutura é composta por etiquetas morfológicas, gramaticais, semânticas e lema (forma canônica) da palavra. Já a segunda, é composta por elementos (sintagmas) que constituem uma unidade significativa dentro das sentenças, e que mantém entre si relação de dependência e ordem (SILVA ; KOCH, 2001). O PALAVRAS fornece três tipos diferentes de saídas: um formato visual, um formato próprio do parser (VISL) e o formato TigerXML.

No contexto deste trabalho, as anotações de *part-of-speech* foram de primordial importância para a extração de conceitos. A Tabela 2.1 mostra a representação do PALAVRAS para etiquetas de categorias gramaticais. E a Tabela 2.2 mostra a representação de etiquetas semânticas. Segundo Bick (BICK, 2000), o PALAVRAS possui cerca de 160 tags semânticas que agregam significado a uma determinada palavra. Por exemplo, se a tag “H” é atribuída ao substantivo “pesquisador”, isso indica que a palavra pertence ao grupo “Humano”, e mais especificamente, o mesmo substantivo também pode receber a tag “Hprof”, indicando que pertence ao grupo “Profissional Humano”, que é subgrupo de “Humano”.

2.2 FORMATO DE ANOTAÇÃO LINGUÍSTICA TIGER-XML

Tiger-XML (KONIG *et al.*, 2003) é o formato de saída do PALAVRAS, um arquivo XML que contém todas as palavras do texto anotadas conforme suas características morfológicas, semânticas e funções sintáticas. Basicamente, os dados disponíveis no formato Tiger-XML são separados em sentenças, representadas pelo elemento “*s*”, e cada sentença é dividida entre os elementos *terminals* e *nonterminals*. No primeiro elemento estão informações de PoS da palavra:

etiquetas morfológicas, gramaticais, semânticas e lema. No segundo são armazenadas informações relacionadas à estrutura sintagmática da sentença. Os elementos *terminals* possuem um ou mais elementos filhos "*t*" (palavras), composto pelos seguintes atributos:

- *id:* índice da palavra;
- *word:* a palavra na forma como é lida da sentença;
- *lemma:* lema da palavra. Por exemplo, as palavras "sou" e "é" possuem lema "ser" e, as palavras "cientista" e "cientistas" possuem lema "cientista";
- *pos:* classe gramatical da palavra (Tabela 2.1);
- *morph:* informações morfológicas da palavra, como gênero e número;
- *sem*: informações semânticas da palavra (Tabela 2.2).

Na Figura 2.1 é apresentado um exemplo de saída do Tiger-XML para a frase "Aguardando condições de mar".

Tabela 2.1: Representação das classes gramaticais no PALAVRAS.

Categoria Gramatical		Tag "pos"
substantivo		n
nome próprio		prop
pronome		pron
artigo		art
adjetivo		adj
advérbio		adv
Verbo	indicativo / subjuntivo / imperativo	v-fin
	infinitivo	v-inf
	particípio	v-pcp
	gerúndio	v-ger
preposição		prp
Conjunção	coordenativa	co
	subordinativa	ks
numeral		num
interjeição		in
abreviação		abb

Tabela 2.2: Alguns exemplos de representação semântica do PALAVRAS.

Tag "sem"	Descrição	Exemplo
H	**ser humano (subclasses abaixo)**	**inimigo, morador**
HM	entidade mística ou religiosa	anjo, duende
Hprof	profissional	advogado, filósofo
Hfam	membro de família	pai, mãe
top	**topológico (imóvel)**	**Brasília, monte**
topagua	lugar aquático	rio, mar
topejo	lugar funcional	quarto, banheiro
toparea	região, área	área, terreno
V	**veículo (concreto, móvel)**	**carro, bicicleta**
VV	grupo de veículos	armada, comboio
Vfly	veículo aéreo	avião, paraquedas
Vor	máquina	britadeira
ac	**abstrato (contável)**	**método, módulo**
acanfeatc	característica anatômica	cabelo, verruga
acreg	regra	regra, lei
acgeom	formas geométricas	elipse, retângulo

```
<?xml version="1.0" encoding="iso-8859-1"?>
<xml>
<corpus>

        <body>
<s id="s1" ref="1" source="Running text" forest="1" text="Aguardando condições de mar.">
        <graph root="s1_500">
                <terminals>
                        <t id="s1_1" word="Aguardando" lemma="aguardar" pos="v-ger" morph="--" sem="--" extra="clb
vH mv"/>
                        <t id="s1_2" word="condições" lemma="condição" pos="n" morph="F P" sem="sit sem-c state"
extra="--"/>
                        <t id="s1_3" word="de" lemma="de" pos="prp" morph="--" sem="--" extra="np-close"/>
                        <t id="s1_4" word="mar" lemma="mar" pos="n" morph="M S" sem="Lwater" extra="--"/>
                        <t id="s1_5" word="." lemma="--" pos="pu" morph="--" sem="--" extra="--"/>
                </terminals>
                <nonterminals>
                        <nt id="s1_500" cat="s">
                                <edge label="fA" idref="s1_501"/>
                        </nt>
                        <nt id="s1_501" cat="advp">
                                <edge label="P" idref="s1_1"/>
                                <edge label="Od" idref="s1_502"/>
                                <edge label="PU" idref="s1_5"/>
                        </nt>
                        <nt id="s1_502" cat="np">
                                <edge label="H" idref="s1_2"/>
                                <edge label="DN" idref="s1_503"/>
                        </nt>
                        <nt id="s1_503" cat="pp">
                                <edge label="H" idref="s1_3"/>
                                <edge label="DP" idref="s1_4"/>
                        </nt>
                </nonterminals>
        </graph>
</s>
        </body>
</corpus>
</xml>
```

Figura 2.1: Exemplo de saída do Tiger-XML.

2.3 MEDIDAS ESTATÍSTICAS

2.3.1 Cálculo de Relevância

Medidas estatísticas são utilizadas para fazer o cálculo da relevância de termos e conceitos extraídos para um determinado domínio. Neste trabalho foram aplicadas duas medidas: Frequência Relativa (FR) e TFIDF (*Term Frequence-Inverse Document Frequence*).

A medida *FR(t)* considera o número de vezes que o termo '*t*' aparece em um documento dividido pelo total de palavras do documento '*N*'.

$$FR(t) = \frac{F(t)}{N}$$

E a medida $tfidf_{l,d}$ considera que termos que possuem alta frequência de ocorrência em um número limitado de documentos são relevantes para o domínio.

$$tfidf_{l,d} = lef_{l,d} * log\left(\frac{|D|}{df_l}\right)$$

Onde:

- $lef_{l,d}$ é a frequência da entrada do termo '*l*' em um documento '*d*';
- df_l é o número de documentos do conjunto '*D*' em que '*l*' ocorre.

A medida $tfidf_{l,d}$ retorna a relevância de um termo relacionado a um único documento de '*D*'. Para obter um resultado que considere todo o conjunto de documentos é preciso fazer um somatório dos valores obtidos em cada documento.

$$tfidf_l = \sum_{d \in D} tfidf_{l,d}$$

2.3.2 Métricas de Comparação de Listas

Para avaliação de desempenho dos métodos desenvolvidos foram aplicadas métricas utilizadas principalmente na área de Recuperação de Informação (RI). Através dessas métricas, são feitas comparações das listas de termos e conceitos extraídos com listas de referência, contendo os conceitos extraídos manualmente por um especialista no domínio específico. As medidas usadas foram: Precisão (P), Abrangência (A) e *F-measure* (F).

A medida Precisão (P) calcula a intercessão entre a lista de referência (LR) e a lista de extraídos (LE), ou seja, os termos corretos extraídos, pelo total de termos extraídos:

$$P = \frac{|LR \cap LE|}{|LE|}$$

Abrangência (A), calcula a intercessão entre a lista de referência (LR) e a lista de extraídos (LE), ou seja, os termos corretos extraídos, pelo total de termos da lista de referência:

$$A = \frac{|LR \cap LE|}{|LR|}$$

E a medida *F-measure* (F), calcula a média harmônica entre a precisão e abrangência.

$$F = \frac{2 \times P \times A}{P + A}$$

2.4 ANÁLISE E PROJETOS ORIENTADO A OBJETOS (A/POO) E PROCESSO UNIFICADO (PU)

No desenvolvimento do software criado, foi utilizada uma metodologia de Análise e Projeto Orientado a Objeto (A/POO) e processo de desenvolvimento iterativo e ágil, chamado Processo Unificado (PU). Foi investigado o domínio do problema, definindo claramente o que deveria ser feito e as necessidades requeridas.

O ciclo de vida de um projeto desenvolvido em PU é organizado em uma série de miniprojetos curtos, chamado iterações. Cada iteração inclui suas próprias atividades de análise de requisitos, projetos, implementação e teste.

Um projeto PU organiza o trabalho e as iterações em quatro fases principais (LARMAN, 2007):

- **Concepção**: visão aproximada, casos de negócio, escopo e estimativas vagas;
- **Elaboração**: visão refinada, implementação iterativa da arquitetura central, resolução de altos riscos, identificação da maioria dos requisitos e do escopo e estimativas mais realistas;
- **Construção**: implementação iterativa dos elementos restantes de menor risco e mais fáceis e preparação para a implantação;
- **Transição**: teste beta e implantação.

2.5 PARADIGMA DE PROGRAMAÇÃO ORIENTADO A OBJETOS

Paradigma baseado na composição e interação entre diversas unidades de software chamadas de objetos, na qual implementa-se um conjunto de classes que definem os objetos, determinando seus comportamentos (métodos) e estados possíveis (atributos), assim como relacionamento com outros objetos.

2.6 LINGUAGEM JAVA

Linguagem de programação orientada a objetos desenvolvida na empresa *Sun Microsystems* na década de 90. Seu código é compilado para um *bytecode* que é executado por uma máquina virtual (*Java Virtual Machine – JVM*), específica para cada sistema operacional.

2.7 NETBEANS

Ambiente integrado de desenvolvimento de software (*Integrated Development Environment – IDE*) empregado especialmente para a plataforma Java. É um projeto *open source* feito para auxiliar os desenvolvedores na criação de aplicativos multiplataformas.

2.8 JDOM

JDOM (*Java Document Object Model*) é uma biblioteca *open source* utilizada para otimizar manipulações de dados XML em Java. É projetada e desenvolvida de forma colaborativa, com mais de 3.000 associados e foi aceita pela *Java Community Process* (JCP) como uma especificação Java (*Java Specification Request*).

2.9 EXPRESSÕES REGULARES (REGEX)

Expressões regulares ou *regex* (abreviação para *regular expression*) são um meio flexível de combinar sequências de texto, como palavras ou padrões de caracteres. Uma expressão regular é escrita em uma linguagem formal e pode ser usada para buscar, editar e manipular textos ou dados. No contexto deste trabalho foi utilizada a biblioteca *regex* do próprio Java.

As expressões regulares fazem uso de metacaracteres, que diferente dos literais, possuem um significado diferenciado no contexto de uma expressão. Alguns exemplos de metacaracteres são:

- \d – representa números;
- \s – representa um espaço em branco;
- \w – representa letras, números ou o "_" (sublinhado);
- . – representa qualquer digito;
- [] – representa uma cadeia de valores. Ex: [a-c] buscaria a ou b ou c;
- ? – representa zero ou uma ocorrência;
- * – representa zero ou mais ocorrências;
- + – representa uma ou mais ocorrências;
- ^ – representa negação;
- () – agrupa os padrões, usado com os metacaracteres acima. Ex: (.)* retornaria o texto inteiro.

3 O PROCESSO DE AQUISIÇÃO DE ONTOLOGIAS A PARTIR DE TEXTOS

Uma pesquisa bibliográfica foi realizada e os trabalhos selecionados foram analisados de acordo com as metodologias que utilizam e, as etapas que cumprem do processo de aquisição de ontologias a partir de textos. Na seção 3.1 é apresentada uma visão geral das etapas que compõe um processo de aquisição de ontologias a partir de textos. Na seção 3.2 são apresentados os trabalhos relacionados. E na seção 3.3 são apresentadas as considerações do capítulo.

3.1 CAMADAS ONTOLÓGICAS

No processo de aquisição de Ontologias a partir de textos pode-se observar que existem várias técnicas e métodos envolvidos. Porém, segundo (TELINE *et al.*, 2003), o processo de aquisição pode ser tradicionalmente classificado conforme a metodologia que utiliza para reconhecer termos e extrair conceitos, a saber: sistemas que utilizam apenas métodos baseados em conhecimento estatístico; sistemas que utilizam apenas métodos baseados em conhecimento linguístico; e, sistemas que utilizam métodos baseados em conhecimento estatístico e linguístico, os chamados híbridos.

Segundo (BUITELAAR *et al.*, 2003), as etapas de um processo de aquisição podem ser classificadas em camadas ontológicas, que partem da aquisição de termos e podem chegar até a aquisição de regras, como mostra a Figura 3.1:

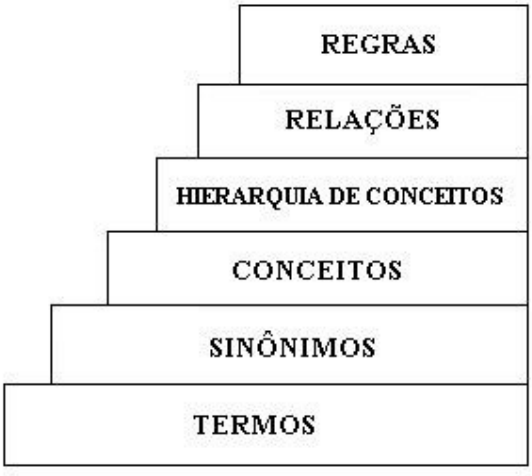

Figura 3.1: Camadas ontológicas.

3.1.1 Termos

Existem muitas metodologias que são usadas para extração de termos, que é o primeiro passo no processo de aquisição de ontologia a partir de textos. Geralmente, é feito um pré-processamento, como segmentação em frases e *tokenização* (separação em palavras) dos textos de domínio utilizado e, em seguida, podem ser usados padrões morfossintáticos para etiquetar os termos encontrados. Adicionalmente, e no intuito de identificar a relevância dos candidatos a termos, uma fase de processamento estatístico pode ser incluída, que compara a frequência de termos entre textos.

3.1.2 Sinônimos

Nesta fase é feita a identificação do sentido apropriado do termo em questão, que determina o conjunto de sinônimos que devem ser extraídos. No contexto de aquisição de ontologia, pesquisadores têm explorado o fato de termos ambíguos terem significados muito específicos em determinado domínio, permitindo uma abordagem integrada para a desambiguação de sentido e extração de sinônimos. Muito do trabalho desenvolvido nesta área tem foco na integração do *WordNet* para a aquisição de sinônimos em Inglês, e do *EuroWordNet* para os sinônimos bilíngues ou multilíngues e para traduções de termos.

3.1.3 Conceitos

Nesta fase são definidas as relações entre termos, os chamados conceitos. A maioria das pesquisas feitas com relação à extração de conceitos aborda a questão da linguística ou perspectiva textual. Porém, também existem abordagens puramente estatísticas, como o método N-grama, onde "N" indica a quantidade de palavras que podem constituir um conceito. A ideia do método é percorrer um documento extraindo "N" palavras de cada vez, calculando alguma medida estatística para cada N-grama extraído.

Aprendizagem de conceitos também inclui a extração de propriedades do conceito, que é a extração das relações entre um conceito e outros conceitos.

3.1.4 Hierarquia de Conceitos

Segundo (BUITELAAR *et al.*, 2003), existem pelo menos dois principais paradigmas explorados na construção de taxonomias a partir de textos. O primeiro é a aplicação de padrões morfossintáticos para detectar relações de hiponímia (relação em que um conceito é identificado como subclasse do outro, onde o conceito mais específico é chamado hipônimo do mais genérico). Também existem abordagens baseadas na estrutura de Sintagmas Nominais (SN), que, segundo (SILVA ; KOCH, 2001), são formados por um conjunto de elementos que constituem uma unidade significativa dentro da oração, e que mantêm entre si relação de dependência e ordem. Esse conjunto está organizado em torno de um elemento fundamental chamado núcleo, que pode por si só, constituir um sintagma. Explorando então, a estrutura interna dos sintagmas nominais, podem-se derivar relações taxonômicas entre classes (núcleo do sintagma) e suas subclasses (que podem ser uma combinação do núcleo com seus modificadores).

O segundo paradigma é a exploração de algoritmos de agrupamento hierárquico para gerar automaticamente hierarquias a partir de textos.

3.1.5 Relações (não hierárquicas)

Muito do trabalho desenvolvido nessa área tem sido feito em conjunto com o ramo biomédico, área que possui muitos textos disponíveis para esse tipo de pesquisa. O objetivo é descobrir novas relações entre conceitos conhecidos (sintomas, drogas, doenças, etc.) através da análise de grandes quantidades de artigos científicos biomédicos.

3.1.6 Regras

A extração de regras é provavelmente a área menos abordada nas pesquisas de aprendizagem de ontologia. O foco principal tem sido aprender vínculos léxicos para aplicação em sistemas de perguntas e respostas. Com relação à aquisição de Ontologias a partir de textos em Língua Portuguesa, não foi encontrada nenhuma pesquisa relacionada.

3.2 TRABALHOS RELACIONADOS

3.2.1 A abordagem de Teline

Em (TELINE *et al.*, 2003) são descritos os passos para extração de termos e conceitos de textos em português da área de Revestimentos Cerâmicos, com objetivo principal de avaliar o desempenho de medidas estatísticas no processo de extração.

Neste trabalho são abordadas várias técnicas para se alcançar algumas das camadas ontológicas. Todo o processo de extração de termos e conceitos é baseado no uso de uma ferramenta auxiliar, chamada NSP (*N-gram Statistics Package*), constituída por um conjunto de programas que auxilia na análise de N-gramas em arquivos texto.

Na fase de extração de termos são feitas algumas restrições na própria ferramenta antes da geração das listas de N-gramas. Devido ao fato da língua padrão da ferramenta ser a língua inglesa, o processo de geração de N-gramas não reconhece acentuações encontradas no texto. Por isso, foram feitas regras de formação de tokens para que a ferramenta pudesse reconhecer acentuação. Outra restrição feita foi a construção de uma lista de *stopwords*, composta por palavras comuns que possuem significado limitado e, portanto, não são relevantes para o domínio. Esta lista, que neste trabalho é composta de preposições, artigos, conjunções e alguns advérbios, é usada para excluir essas palavras do texto antes da geração de N-gramas.

Na fase de extração de conceitos há duas abordagens: uma manual e uma automática. A manual consiste na obtenção de uma lista de referência, composta por conceitos pré-definidos pelo especialista do domínio. Essa lista é utilizada como referência para comparação com os conceitos gerados automaticamente.

A extração automática de conceitos é uma etapa que acaba se confundindo com a extração de termos nesta abordagem, pois o processo de geração de N-gramas faz, ao mesmo tempo, a *tokenização* do texto e a geração de unigramas (termos simples), bigramas e trigramas.

Após o processo de geração das listas de N-gramas, utilizando as restrições citadas anteriormente, são aplicadas medidas estatísticas nas mesmas. Foram aplicadas quatro medidas estatísticas com o auxílio do pacote NSP: Frequência, *Log-likelihood*, Informação Mútua e *Dice*. Na lista de unigramas, foi aplicada somente a medida de frequência e na de bigramas, foram aplicadas as medidas de coeficiente *Log-Likehood*, Informação Mútua e coeficiente de *Dice*, já na lista de trigramas, foram aplicadas coeficiente *Log-Likehood* e Informação Mútua.

Em seguida, foi feita uma análise das medidas mais eficientes, utilizando uma lista de referência para a comparação do desempenho.

Segundo (TELINE *et al.*, 2003), para unigramas, não foi possível afirmar que quanto maior a frequência, maior a probabilidade dos conceitos aparecerem no texto específico deles. Já para bigramas, não foi possível escolher um dos métodos estatísticos dentre Frequência, Informação Mútua, *Log-likelihood* e *Dice* com melhor desempenho, pois seus resultados apresentaram-se bastante semelhantes. Já para o caso de trigramas, a Frequência apresentou um resultado melhor do que as medidas Informação Mútua e *Log-likelihood*.

3.2.2 A abordagem de Baségio

Em (BASÉGIO, 2006) são abordadas principalmente as camadas de conceitos e hierarquia de conceitos, de modo a semi-automatizar os passos da construção de ontologias a partir de textos em português do Brasil. Para isso, foi desenvolvida uma aplicação que foi utilizada no estudo de textos do domínio de Turismo.

Com relação à extração de termos, o autor optou por não fazer um pré-processamento do texto. Assim, assumiu como ponto de partida um texto anotado linguisticamente, com as seguintes informações associadas a cada palavra do documento: a palavra no seu formato original; o lema da palavra original, ou seja, a palavra em sua forma singular e masculina e; a etiqueta gramatical da palavra (exemplo: substantivo, adjetivo, etc.).

Assim como na abordagem anterior, aqui também são eliminados termos que não representam conceitos de domínio, através de uma lista de *stopwords*. Também são removidos do texto todos os termos contendo caracteres não alfabéticos como números e símbolos.

Outra etapa feita ainda na extração de termos é a identificação da ordem de relevância dos termos do domínio. Para isso, são utilizadas duas medidas estatísticas: *Log-Likelihood* e TFIDF (*term frequency x inverted document frequency*). A primeira medida é usada para selecionar apenas termos considerados relevantes para o domínio. A segunda, para organizar os termos por ordem de relevância e posteriormente apresentar os termos ao engenheiro de ontologia. Na medida TFIDF, um limiar mínimo pode ainda ser definido pelo próprio engenheiro e termos que estiverem abaixo do ponto de corte são desconsiderados.

Após essas etapas, a lista resultante é apresentada ao engenheiro de ontologia, possibilitando que ele inclua ou exclua termos de acordo com seu julgamento.

Na extração de conceitos, é feita a identificação de termos compostos. Essa seleção é realizada com base em regras expressas por sequências de etiquetas que, quando encontradas no texto, podem representar termos compostos. A Tabela 3.1 apresenta as regras utilizadas por Baségio.

Tabela 3.1: Regras para extração de termos compostos.

Nro.	Regra
1	_SU _AJ _PR _AD _SU _AJ
2	_SU _AJ _PR _AD _SU
3	_SU _PR _AD _SU _AJ
4	_SU _PR _AD _SU
5	_SU _AJ _PR _SU _AJ
6	_SU _AJ _PR _SU
7	_SU _PR _SU _AJ
8	_SU _PR _SU
9	_SU _AJ

Onde:

_SU: substantivos;
_AJ: adjetivos;
_PR: preposições;
_AD: advérbios.

As regras de mapeamento de conceitos são bastante genéricas, para possibilitar a geração de estruturas ontológicas para diferentes domínios.

A extração de hierarquia de conceitos consiste das seguintes etapas:

1. Identificar relações taxonômicas com base em termos compostos: é a identificação de relações a partir do núcleo de um termo composto. Por exemplo, se foram identificados o termo relevante “contrato” e o termo composto “contrato de venda”, a ideia é identificar que “contrato de venda” é um tipo de “contrato”.

2. Identificar relações taxonômicas através dos padrões de Hearst: Baségio propõe uma adaptação dos padrões léxico sintáticos propostos por Hearst. A Tabela 3.2 apresenta os padrões originais e as traduções/adaptações realizadas.

Tabela 3.2: Padrões de Hearst adaptados para o português por Baségio.

	Padrão Original	Tradução/Adaptação
1	**NP such as {(NP,) * (or | and)} NP**	SUB como {(SUB,) * (ou | e)} SUB
		SUB tal(is) como {(SUB,) * (ou | e)} SUB
2	**such NP as {(NP,) * (or | and)} NP**	tal(is) SUB como {(SUB,) * (ou | e)} SUB
3	**NP {,NP} * {,} or other NP**	SUB {,SUB} * {,} ou outro(s) SUB
4	**NP {,NP} * {,} and other NP**	SUB {,SUB} * {,} e outro(s) SUB
5	**NP {,} including {NP,} * {or | and} NP**	SUB {,} incluindo {SUB,} * {ou | e} SUB
6	**NP {,} especialy {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

Onde:

SUB: Substantivo;

NP: Sintagma Nominal.

Como exemplo, segue um trecho de texto e descrição de como esses padrões são encontrados:

- *"...foram analisadas muitas classes gramaticais, principalmente, substantivo e adjetivo."*

Podemos observar que o trecho se encontra no sexto padrão da Tabela 3.2 (*SUB {,} principalmente {SUB,} * {ou | e} SUB*), tendo *classes gramaticais*, *substantivo* e *adjetivo* como SUB, gerando uma relação de hiponímia, onde o SUB mais genérico é *classes gramaticais* e os mais específicos são *substantivo* e *adjetivo*.

3. Identificar relações taxonômicas através dos padrões de Morin e Jacquemin: assim como na etapa anterior, foi feita uma adaptação dos padrões léxico sintáticos de Morin e Jacquemin. Os padrões com suas adaptações são apresentados na Tabela 3.3.

Tabela 3.3: Padrões de Morin/Jacquemin adaptados para o português por Baségio.

	Padrão Original	Tradução/Adaptação
1	**{deux | trois...| 2 | 3 | 4...} NP1 (LIST2)**	{dois | três | 2 |3 | 4...} SUB1 (LIST_SUB2)
2	**{certain | quelque | de autre...}NP1 (LIST2)**	{certos | quaisquer | de outro(s)...} SUB1 (LIST_SUB2)
3	**{deux | trois...| 2 | 3 | 4...} NP1: LIST2**	{dois | três | 2 |3 | 4...} SUB1 : LIST_SUB2
4	**{certain | quelque | de autre...}NP1: LIST2**	{certos | quaisquer | de outro(s)...} SUB1 : LIST_SUB2
5	**{de autre} NP1 tel que LIST2**	{de outro(s)}* SUB1 {tal(is)}* como LIST_SUB2
6	**NP1, particulièrement NP2**	SUB1, {particularmente | especialmente} SUB2
7	**{de autre}NP1 comme LIST2**	{de outro(s)}* SUB1 como LIST_SUB2
8	**NP1 tel LIST2**	SUB1 como LIST_SUB2
9	**NP2 {et | ou} de autre NP1**	SUB2 {e | ou} de outro(s) SUB1
10	**NP1 et notamment NP2**	SUB1 e (notadamente | em particular) SUB2

Onde:
SUB1, SUB2: Substantivos;

NP1, NP2: Sintagmas Nominais;

LIST_SUB: refere-se a uma lista de substantivos.

Como exemplo desses padrões, observemos o trecho:

- *"Foram usadas três medidas estatísticas (Frequência, Log-likelihood e Informação Mútua) que serviram..."*

O padrão *{dois | três | 2 |3 | 4...} SUB1 (LIST_SUB2)* da Tabela 3.3 é encontrado no trecho acima, onde *medidas estatísticas* é identificado como SUB1 e *(Frequência, Log-likelihood e Informação Mútua)* como LIST_SUB2, gerando uma relação de hiponímia, onde o termo genérico é *medidas estatísticas* e os termos específicos são *Frequência, Log-likelihood* e *Informação Mútua*.

Vale salientar que Hearst e Morin/Jacquemin trabalham com sintagma nominal (*noun phrase* – NP) em seus padrões, em ambas as adaptações os sintagmas foram substituídos por substantivo (SUB).

Após esse processo, o engenheiro de ontologia pode gerar a estrutura ontológica na linguagem de representação OWL com as seguintes informações: termos simples; termos compostos; relações baseadas em termos compostos; relações baseadas nos padrões de Hearst; relações baseadas nos padrões de Morin e Jacquemin. Assim, é criado um arquivo OWL que pode ser utilizado em editores de ontologias como Protégé, permitindo ao especialista do domínio continuar o desenvolvimento da ontologia.

De acordo com Baségio, os resultados da avaliação da metodologia indicam que a utilização da medida *Log-Likelihood*, utilizada para exclusão de termos irrelevantes, foi muito importante para

os resultados obtidos nos estudos de caso. Se tivesse sido utiliza apenas a medida TFIDF, teriam sido retornados 3.308 candidatos a termos relevantes ao invés dos 412 retornados com o uso da medida *Log-Likelihood*.

A utilização da medida TFIDF para apresentar os candidatos a termos relevantes do domínio em ordem de relevância obteve um bom resultado. Mais de 50% dos termos selecionados pelo especialista estavam entre os 100 termos mais relevantes e 80% dos termos selecionados encontravam-se na primeira metade dos termos apresentados ao especialista.

A identificação de termos compostos obteve um bom resultado, 57% dos termos foram indicados como relevantes. Dentre as regras utilizadas, a regra "_SU _AJ" foi responsável por mais da metade dos termos compostos extraídos pela ferramenta e também pelo maior número de termos selecionados (57,41% do total de termos selecionados). Por outro lado, algumas regras não tiveram nenhum termo extraído, ou tiveram poucos termos extraídos e nenhum termo selecionado.

Com relação à extração de relações hierárquicas, o método com base em termos compostos foi o que obteve o melhor resultado, onde foram selecionadas 152 relações, o que representa 53,52% do total extraído.

3.2.3 A abordagem de Guilherme

Em (GUILHERME *et al.,* 2006) é apresentada uma metodologia para extração de termos e conceitos. A ferramenta desenvolvida é chamada PhDic (*Phrase Dictionary*) e os textos de domínio são uma coleção de documentos que relatam as anormalidades ocorridas nas operações de perfuração em plataformas de petróleo.

A primeira etapa está relacionada com a extração de termos, é a geração de um dicionário, composto de palavras encontradas nos arquivos textos. A ideia é usar esse conjunto de textos, chamado conjunto de treinamento, como ponto inicial para geração de uma coleção de termos básicos.

O processo adotado na obtenção do dicionário de termos consiste das seguintes etapas:

1. **Disponibilizar a palavra no seu formato original**: envolve o processo de *tokenização* do texto;

2. **Eliminar termos que não representam conceitos de domínio**: também é utilizada uma lista de *stopwords* para excluir do texto palavras que possuem significado semântico limitado.

3. **Reduzir cada palavra ao seu radical**: Basicamente, essa técnica procura reduzir palavras que pertencem a uma mesma família para sua forma raiz. Assim, as palavras *andar*,

andando e *andado* são reduzidas à forma raiz *and.* Esse processo é muito semelhante ao processo de lematização;

4. **Definição de limiar mínimo para termos do dicionário**: consiste em definir uma frequência mínima aceitável para um termo no dicionário ser considerado relevante ao domínio. Com a definição desse limiar, os termos com frequência abaixo são excluídos.

5. **Aplicação de medidas estatísticas**: é feita a identificação da ordem de relevância dos termos do domínio, através da medida estatística TFIDF.

A partir do dicionário gerado, o engenheiro de ontologia pode avaliar os termos gerados e excluir termos que considerar irrelevantes. Em seguida, fazer uma etiquetagem manual dos termos do dicionário, associando a cada termo uma sintaxe pré-definida. As etiquetas utilizadas são um conjunto de palavras definidas pelo próprio usuário e são baseadas nos rótulos utilizados na estrutura de representação do conhecimento adotada ou no vocabulário da ontologia.

Após a associação de sintaxe, ocorre a extração de conceitos, onde uma lista de conceitos é gerada, baseada nas informações contidas no dicionário. Os conceitos gerados são apresentados ao engenheiro e o mesmo pode então verificar, para cada conceito, as frases em que os mesmos aparecem e também excluir termos que considerar irrelevantes.

Tendo gerado os conceitos, o usuário pode reassociar esses conceitos com uma nova coleção de textos. Caso o conceito seja encontrado, então ele é reassociado à frase em que aparece e é recalculada sua frequência.

3.2.4 A abordagem de Almeida

Em (ALMEIDA;VALE, 2008) é focada a utilidade do conhecimento linguístico, em particular da morfologia, na identificação de candidatos a conceitos. Para isso, foi utilizado como ferramenta auxiliar o software Unitex para processamento de textos do domínio da Nanotecnologia.

O Unitex é um software que faz o processamento de texto com base em dicionários eletrônicos de cada uma das línguas que o integram. Para o português do Brasil, o Unitex traz um dicionário eletrônico bastante extenso – cerca de 67.500 formas canônicas (ou lemas), 880 mil formas flexionadas e 4.500 formas compostas com hífen. O software ainda permite que qualquer usuário crie seus próprios dicionários, integrando novas unidades lexicais (termos) ou, ainda, acrescentando novas informações morfológicas, sintáticas e semânticas ao léxico já existente ou ainda gerando novas formas a partir de uma forma canônica. Além disso, também é disponibilizado

para desenvolvedores, podendo ser compilado como uma biblioteca dinâmica que contém todas as funções Unitex.

Com relação à camada de extração de termos, neste projeto foi utilizada uma lista de frequência, gerada pelo próprio Unitex, que continha os itens léxicos (termos) mais frequentes. Em seguida, foram excluídos os termos que de fato não eram relevantes, tais como artigos, preposições, conjunções, pronomes, advérbios, nomes próprios, determinados substantivos (país, instituto, exemplo, etc), determinados adjetivos (novo, bom, etc.) e determinados verbos (sobretudo os modais), ficando assim, somente os itens léxicos mais frequentes.

A extração de conceitos é feita utilizando as funcionalidades do próprio Unitex, que permite realizar expressões de busca. Partindo da lista de frequência, é feita uma busca a um termo específico da lista ou a um segmento terminológico, baseado no conhecimento prévio do usuário sobre o domínio. Por exemplo, dado que o domínio dos textos é a nanotecnologia, fazer uma busca que retorne todas as ocorrências de termos que contenham o segmento terminológico "nano" tem grandes chances de retornar termos que sejam realmente relevantes para a ontologia. Além de retornar os termos específicos da busca, o Unitex também mostra as frases em que ocorrem. Assim, com base na observação do termo nas frases, o usuário pode fazer buscas cada vez mais refinadas. Por exemplo, buscando pelo termo "material" (substantivo ou N), observou-se que ele ocorre seguido de um adjetivo (A) ou de um sintagma preposicionado. Com base nessa observação, outras expressões foram formuladas para gerar novas buscas ao termo, por exemplo, a expressão "<material><A>", que recupera o item léxico 'material' lematizado, com formas no singular e no plural, seguido de adjetivo.

Sendo assim, a partir dessa análise dos termos no texto, é possível descrever a sua morfologia e, a partir dessa descrição morfológica, extrair mais termos, de forma cada vez mais eficiente.

Segundo (ALMEIDA ; VALE, 2008), esse tipo de busca, como se observou, pode ser uma boa ferramenta para a listagem de candidatos a conceitos. Entretanto, o sucesso dessa busca pode ser determinado pela qualidade dos recursos linguísticos que servem de base para ela.

3.2.5 A abordagem de Ribeiro Junior

Em (RIBEIRO JUNIOR, 2008) é apresentada uma metodologia para extração de termos, conceitos e hierarquia de conceitos, e também é desenvolvida uma ferramenta chamada OntoLP, implementada como um plug-in para o ambiente de auxílio a criação de ontologias Protégé.

A primeira etapa deste trabalho, relacionada com a extração de termos, é a anotação linguística do texto. Nesta tarefa, é utilizada uma ferramenta auxiliar, o analisador sintático PALAVRAS, descrito na seção 2.1.

Neste trabalho, foi usada uma biblioteca desenvolvida pelo Laboratório de Engenharia da Linguagem da UNISINOS (LEL) que converte o formato TigerXML (seção 2.2) para o formato XCES/PLN-BR. E este último é o adotado como padrão de entrada para o OntoLP. A escolha desse formato é decorrente da facilidade de processamento e entendimento da estrutura de armazenamento dos dados e também, da divisão das informações linguísticas em diferentes níveis, possibilitando carregar somente os arquivos necessários para a execução dos métodos.

O processo de extração de termos abrange tanto a extração de termo simples (unigramas) quanto à extração de conceitos. Neste processo são aplicados diferentes níveis de conhecimento linguístico e métodos. A Tabela 3.4 apresenta as etapas do processo de extração.

Tabela 3.4: Etapas do processo de Extração de Termos de Ribeiro Junior.

Processo de Extração de Termos
Extração dos Grupos Semânticos; (opcional)
Filtragem dos Grupos Semânticos irrelevantes, feita pelo engenheiro; (opcional)
Extração de Termos Simples considerando apenas aqueles pertencentes aos Grupos Semânticos selecionados;
Exclusão de termos simples irrelevantes, feita pelo engenheiro; (opcional)
Extração dos Termos Complexos considerando apenas aqueles que possuem no mínimo uma palavra presente na lista final de termos simples e que pertençam a um Grupo Semântico selecionado;
Exclusão dos termos complexos irrelevantes, feita pelo engenheiro. (opcional)

1. Seleção de Grupos Semânticos: etapa proposta para substituir a utilização de listas de *stopwords*. Utilizando as informações semânticas disponibilizadas pelo PALAVRAS, todas as *tags* semânticas presentes no corpus de entrada são extraídas. Em seguida, é aplicado um cálculo de FR (Frequência Relativa) à lista de *tags* para avaliar a relevância dos grupos semânticos extraídos. O resultado é então apresentado ao engenheiro de ontologia, que pode excluir os grupos semânticos que considerar irrelevantes para o domínio. Sendo assim, quando excluir um determinado grupo semântico, estará automaticamente desprezando todos os termos pertencentes somente àquele grupo.

Cabe salientar que a seleção de grupos semânticos é opcional, podendo ser desabilitada da ferramenta.

2. Extração de Termos Simples: utilizando a lista de Grupos Semânticos gerada na etapa anterior, os métodos de extração percorrem o corpus, selecionando os termos considerados aptos e que

pertençam a pelo menos um grupo semântico presente na lista de entrada, caso o método Filtro por Grupos Semânticos tenha sido habilitado.

A extração dos termos é baseada em dois métodos: Classe Gramatical e Núcleo do Sintagma Nominal. O primeiro possibilita que o engenheiro selecione quais classes gramaticais deseja extrair do corpus. O segundo extrai apenas termos considerados núcleo de um sintagma nominal. Em seguida, a lista de termos extraída pelos métodos é submetida às medidas de relevância: FR, TFIDF e NC-Value.

Após o cálculo, os termos são reorganizados em ordem decrescente conforme essas medidas e depois, são apresentados ao engenheiro, que exclui o que considera irrelevante.

3. Extração de Termos Complexos : etapa utilizada para extrair conceitos formados por dois ou mais termos, utilizando como entrada a lista final do processo de extração de termos e, caso o método Filtro por Grupos Semânticos tenha sido habilitado, a lista de grupos semânticos também é utilizada como entrada.

Para a extração de termos complexos foram implementados três métodos: N-grama, Padrões Morfossintáticos e Sintagma Nominal. O primeiro método é geralmente aplicado em textos sem informações linguísticas, mas foi adaptado neste trabalho, onde são extraídos somente conceitos pertencentes às classes gramaticais definidas pelo engenheiro de ontologia.

O segundo método utiliza regras formadas por padrões morfológicos, onde são utilizadas as regras propostas por Baségio para extração de conceitos.

O terceiro método extrai apenas conceitos que compõem todo ou parte de um sintagma nominal.

Assim como na extração de termos simples, aqui também são utilizadas medidas estatísticas para o cálculo de relevância: FR, TFIDF, C-Value e NC-Value.

O processo que envolve a camada de hierarquia de conceitos é chamado organização hierárquica dos termos. Para este processo são usados como entrada as listas de termos simples e complexos extraídas no processo anterior. Nesta etapa o engenheiro pode também editar as taxonomias geradas, melhorando o resultado final. Para a realização da tarefa foram implementados três métodos: Termos Complexos, Padrões de Hearst e Padrões de Morin/Jacquemin.

O primeiro método recebe como entrada uma lista de termos simples e uma lista de termos complexos. Sua execução busca ocorrências de um termo simples dentro dos complexos. E quando alguma é encontrada, o termo complexo selecionado é organizado como hipônimo do termo simples. O segundo e o terceiro são os padrões adaptados por Baségio. A diferença é que neste trabalho são feitas duas restrições para a utilização dos padrões. Uma é a restrição por termos, quando habili-

tada são extraídas somente relações onde no mínimo um dos conceitos está presente nas listas de termos simples e complexos. A outra é a restrição por Grupos Semânticos, quando habilitada são extraídas somente relações onde os conceitos são pertencentes a um mesmo grupo semântico.

Ao final deste processo, o engenheiro pode exportar as taxonomias inferidas para a interface principal de construção de ontologias do Protégé.

Na avaliação do processo de extração de termos, o método Filtro por Grupos Semânticos foi indicada por usuários como o que mais ajudou na extração de termos simples e complexos, pois o uso de informações semânticas durante o processo de extração melhora a execução da tarefa. E as combinações Classe Gramatical/TFIDF (unigramas) e Padrões Morfossintáticos/TFIDF (bigramas e trigramas) foram as que obtiveram melhores resultados na avaliação geral.

E no que diz respeito à Organização Hierárquica dos Termos, o método que obteve melhor resultado foi o baseado em Termos Complexos.

3.2.6 A abordagem de Lopes

Em (LOPES *et al.,* 2009) é apresentada uma metodologia para extração de termos e conceitos, e também é desenvolvida uma ferramenta chamada ExATOlp – Extrator Automático de Termos para Ontologias em Língua Portuguesa. A ferramenta recebe um conjunto de documentos anotados sintaticamente e extrai todos os sintagmas nominais (SN) do texto, classificando-os segundo o número de palavras e em seguida os salva em listas que podem conter tanto os SN na sua forma original no texto como em sua forma canônica, ou seja, os termos sem alterações de gênero, número ou conjugações verbais. A ferramenta ainda oferece algumas opções como aplicação de ponto de corte, comparação de listas e cálculo de medidas usuais de precisão e abrangência.

Na abordagem utilizada por Lopes, a primeira etapa a ser realizada é a anotação linguística dos textos que compõem um determinado domínio, realizada pelo mesmo parser utilizado por Ribeiro Junior, o PALAVRAS. A diferença é que, enquanto o ONTOLP utiliza como entrada o formato XCES/PLN-BR, nesta abordagem, o formato utilizado como entrada é o TIGER-XML.

O processo de extração de termos abrange tanto a extração de termo quanto à extração de conceito, pois a principal funcionalidade da ferramenta é a extração de SN. Segundo Kuramoto (apud. Lopes), ao contrário de palavras isoladas cujo significado depende do contexto, os SN são os melhores candidatos a conceitos, pois quando extraídos de um texto, seus significados permanecem os mesmos.

Durante a extração de SN, a ferramenta utiliza um conjunto de heurísticas para refinar o processo. As heurísticas aplicadas aos termos identificados como SN pelo PALAVRAS são:

- são eliminados SN que possuem números, por exemplo, “20 anos”, “seis meses”;
- são aceitos apenas sintagmas que possuem letras (acentuadas ou não) ou hífen, ou seja, SN que contém caracteres especiais são eliminados, por exemplo, “dupla mãe/neonato”;
- termos identificados como SN que iniciam com pronomes, “estas condições” e “todas as crianças”, são armazenados sem o pronome;
- termos identificados como SN que terminam com conjunções, por exemplo, “baixo peso e” e “leite materno ou” são armazenados sem a conjunção;
- termos identificados como SN que terminam com preposição, por exemplo, “criança acrescida de” e “dosagem diária para” são armazenados sem a preposição;
- termos identificados como SN que contém artigos são armazenados sem estes artigos, “a cicatriz renal” é armazenado apenas como “cicatriz renal”.

Opcionalmente, ainda é possível escolher armazenar apenas alguns SN sendo critérios o número de palavras que o compõem, a sua classe gramatical e a classe sintática do núcleo do SN. Estas opções são:

- é possível selecionar para extrair apenas SN compostos de números específicos de palavras, por exemplo, pode-se escolher extrair apenas sintagmas compostos de uma, duas e três palavras, ou seja, desprezar sintagmas compostos de quatro ou mais palavras;
- é possível extrair somente SN que aparecem como sujeitos, ou somente SN que aparecem como complementos das orações;
- é possível extrair somente SN que possuem como núcleo substantivos próprios, só substantivos comuns, só adjetivos, só verbos no particípio passado, ou qualquer combinação entre estas.

Em seguida, os candidatos a conceitos extraídos são salvos em dez listas que contém respectivamente os sintagmas compostos por 1 a 9 palavras e a última lista contém sintagmas compostos por 10 ou mais palavras. Cada uma das listas contém os termos em ordem decrescente de frequência no corpus.

Após a geração das listas, a ferramenta disponibiliza três opções de manipulação das mesmas: aplicação de ponto de corte, comparação de listas e cálculo de medidas usuais de precisão e abrangência.

A aplicação de ponto de corte é definir a partir de que ponto desprezar os termos menos frequentes no corpus. Por exemplo, desprezar todos os termos em que a frequência absoluta seja menor que 4 ou ainda, manter os 20% primeiros termos da lista ordenada.

A comparação de listas é uma opção que recebe como entrada duas listas, LR (lista de referência) e LE (lista de extraídos), retornando qualquer uma das seguintes opções:

- a interseção entre listas (LR ∩ LE);
- a união entre listas (LR ∪ LE);
- os termos de LR ausentes em LE (LR − (LR ∩ LE));
- os termos de LE ausentes em LR (LE − (LR ∩ LE)).

O cálculo de medidas de precisão e de abrangência também tem o objetivo de comparar a lista de referência com a lista de termos extraídos. As medidas utilizadas são: precisão, abrangência e *F-measure,* descritas na seção 2.3.2.

A ferramenta ExATOlp foi utilizada em dois tipos de domínio: um conjunto de textos com 54 teses e 89 artigos científicos da área de Geologia e outro com 283 artigos do Jonal Brasileiro de Pediatria.

Para avaliar o desempenho da ferramenta, além de comparação dos resultados com a lista de referência, foi feita uma comparação com outra ferramenta, a NSP (*N-gram Statistics Package*). Utilizando o corpus de Pediatria e uma lista de referência composta por bigramas e trigramas. Os resultados obtidos são apresentados na Tabela 3.5.

Tabela 3.5: Comparação entre ExATOlp e NSP (LOPES *et al.*, 2010).

	Termos	|LE|	|LR|	|LE∩LR|	P	A	F
ExATOlp	bigramas	1309	1404	702	**53,63%**	50,00%	**51,75%**
	trigramas	644	731	285	44,25%	38,99%	41,45%
NSP	bigramas	3709	1404	1230	33,16%	**87,61%**	48,11%
	trigramas	2550	731	556	21,80%	76,16%	33,90%

A ferramenta ExATOlp apresentou precisão maior à NSP, porém a abrangência foi menor. Apesar disso, a combinação destas métricas expressa pela *f-measure* foi superior ao NSP.

Cabe salientar que a ferramenta ExATOlp se insere em uma tese de doutorado ainda em curso e a inclusão de novas heurísticas de extração de termos bem como funcionalidades mais avança-

das, como por exemplo, construção automática de hierarquias de conceitos, estão sendo desenvolvidas. Logo, os resultados referentes a ferramenta ExATOlp apresentados refletem a aplicação da sua versão de novembro de 2010.

3.3 CONSIDERAÇÕES

A primeira etapa, no desenvolvimento deste trabalho, foi fornecer uma visão geral das abordagens existentes para aquisição de ontologias a partir de textos em Língua Portuguesa, com o intuito de estabelecer uma metodologia para aprimorar a ferramenta PhDic, com base nessas metodologias já existentes. Na Tabela 3.6 são apresentadas as características gerais das abordagens estudadas.

A primeira observação feita na análise dos trabalhos é que abordagens que utilizam somente métodos estatísticos obtiveram os piores resultados. A abordagem de Lopes exemplifica claramente essa observação, pois a ferramenta desenvolvida no projeto, que utiliza informações linguísticas, obteve melhores resultados quando comparada a ferramenta NSP, que é puramente estatística.

Em métodos puramente estatísticos, um documento é tratado como um simples vetor de termos e suas frequências. Portanto, é possível aplicá-los sem a necessidade de anotar os textos. O método N-grama é um exemplo de extração de conceitos puramente estatístico e dentre as abordagens estudadas, só apresentou resultados significativos quando adaptado por Ribeiro Junior, onde são extraídos somente os termos pertencentes às classes gramaticais que geralmente constituem conceitos de uma ontologia.

Em abordagens puramente estatísticas, ainda não existe uma com resultados que retornem poucos termos indesejáveis. Por esse motivo, essas técnicas vêm sendo aplicadas em conjunto com outras metodologias, visando obter melhores resultados.

Nas extração de conceitos, verificou-se que tanto Guilherme quanto Almeida adotaram dicionários de termos que podem ser incrementados após o processo de extração. A diferença é que Guilherme gera o dicionário de termos na própria aplicação, com base nos textos de treinamento, e este mesmo dicionário pode ser reassociado a novos textos, modificando assim, as informações de frequência dos termos. Já Almeida, utiliza o dicionário eletrônico disponibilizado pelo software Unitex para reconhecer nos textos de entrada termos do dicionário, porém, o software permite que

estes dicionários eletrônicos possam ser incrementados, nos quais qualquer usuário pode integrar novas unidades lexicais (termos) ou, ainda, acrescentar novas informações morfológicas, sintáticas e semânticas a léxicos já existentes, criando assim, dicionários personalizados.

É importante frisar também que Básegio, Guilherme, Ribeiro Junior e Lopes permitiram que em cada etapa da metodologia o engenheiro de ontologia interviesse no processo, aprimorando a saída de cada etapa. Esta estratégia foi adotada porque não existe atualmente técnica de extração e/ou organização de termos que não necessite da intervenção do usuário para obtenção de resultados mais precisos.

Autor	Principais Etapas	Principais técnicas	Camadas Identificadas	Intervenção do Usuário	Ferramentas Associadas
(TELINE *et al.*, 2003)	Pré-processamento do texto; Obtenção de lista de referência de conceitos (manualmente); Extração de termos e conceitos; Comparação dos conceitos com a lista de referência.	*Tokenização*; Uso de lista de *stopwords*; Método N-grama para extração de conceitos.	Termos; Conceitos.	O usuário valida a lista de conceitos gerados.	NSP (N-gram Statistics Package) – ferramenta auxiliar.
(BASÉGIO, 2006)	Importação de textos anotados linguisticamente; Extração de termos; Extração de conceitos; Identificação de relações taxonômicas; Geração de estrutura ontológica em OWL.	Uso de lista de *stopwords*; Abordagens estatísticas; Padrões Morfossintáticos para extração de conceitos; Padrões de Termos Compostos, Hearst e Morin/Jacquemin na extração de relações taxonômicas.	Termos; Conceitos; Hierarquia de Conceitos.	O usuário define limiares; E valida o resultado de cada etapa do processo.	Protótipo de software desenvolvido pelo autor.
(GUILHERME *et al.,* 2006)	Pré-processamento do texto; Geração de dicionário de termos; Associação manual de sintaxe aos termos; Extração de conceitos; Reassociação de conceitos a novos textos.	*Tokenização*; Uso de lista de *stopwords*; Abordagens estatísticas; Identificação de conceitos baseada em padrões sintáticos.	Termos; Conceitos.	O usuário faz a associação de sintaxe a cada termo gerado; Define limiares; E valida o resultado de cada etapa do processo.	PhDic – ferramenta desenvolvida no projeto.
(ALMEIDA ; VALE, 2008)	Pré-processamento do texto; Extração de termos; Geração de um novo dicionário de termos ou modificação do atual (opcional); Extração de conceitos;	*Tokenização*; Uso de lista de *stopwords*; Pré-processamento do texto baseado em dicionários prontos; Abordagens estatísticas; Extração de conceitos baseada em expressões de busca;	Termos; Conceitos	O usuário define regras de mapeamento de conceitos; E valida a lista de conceitos gerada.	Unitex – ferramenta auxiliar.
(RIBEIRO JUNIOR, 2008)	Importação de textos anotados linguisticamente; Identificação de termos; Extração de conceitos; Identificação de relações taxonômicas; Geração de estrutura ontológica em OWL.	Extração de termos, conceitos e hierarquia baseada em Grupos Semânticos; Abordagens estatísticas; N-grama, Padrões Morfossintáticos e Sintagma Nominal para extração de conceitos; Padrões de Termos Compostos, Hearst e Morin/Jacquemin na extração de relações taxonômicas.	Termos; Conceitos; Hierarquia de Conceitos.	O usuário define alguns limiares; E valida o resultado de cada etapa do processo.	PALAVRAS – ferramenta auxiliar; OntoLP – ferramenta desenvolvida no projeto.
(LOPES *et al.*, 2009)	Importação de textos anotados linguisticamente; Extração de SN (Sintagmas Nominais).	Refinamento da extração de SN baseada em heurísticas; Abordagens estatísticas.	Termos; Conceitos.	O usuário define alguns pontos de corte. E valida o resultado das listas de conceito geradas.	PALAVRAS – ferramenta auxiliar; ExATOlp – ferramenta desenvolvida no projeto.

Tabela 3.6: Características gerais das abordagens estudadas.

4 METODOLOGIA PARA AQUISIÇÃO DE CONCEITOS

Neste capítulo é apresentada a metodologia desenvolvida, com base na combinação de técnicas estudadas no capítulo 3, para aprimoramento dos métodos de anotação de textos e extração de conceitos da ferramenta PhDic - ferramenta computacional utilizada na construção do conhecimento e ontologias a partir de relatórios técnicos de anormalidades na perfuração e produção de petróleo. Esta nova abordagem para extração de conceitos utiliza uma metodologia baseada em informações linguísticas, mais consistente do que a usada atualmente pelo PhDic.

4.1 FORMATO DOS TEXTOS DE ENTRADA

O primeiro ponto a ser considerado é o formato dos textos de entrada que serão utilizados na aplicação. Optou-se por adotar um formato em que os textos já estivessem anotados com informações linguísticas. Por ter sido a ferramenta mais utilizada nos trabalhos estudados, o pré-processamento dos relatórios técnicos da perfuração e produção de petróleo foi feito através do analisador sintático PALAVRAS (BICK, 2000), uma ferramenta paga, que neste trabalho foi utilizado através de uma parceria com o grupo de pesquisa do Laboratório de PLN da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS).

O formato de saída do PALAVRAS que será utilizado como entrada para a aplicação é o Tiger-XML (KONIG *et al.*, 2003), descrito em detalhes na seção 2.2.

4.2 EXTRAÇÃO DE TERMOS

4.2.1 Extração Inicial de Termos

Esta fase é responsável pela extração inicial de termos dos documentos, considerando as classes gramaticais que eles pertencem. São considerados termos relevantes somente palavras que pertençam às classes gramaticais que normalmente representam algum tipo de conceito, como substantivo, adjetivo e advérbio. Além da classe gramatical, também é levado em consideração se o termo possui alguma informação semântica. Caso o termo não possua nenhuma informação semântica, ele não é extraído. Cabe salientar que essas informações semânticas também foram consideradas por Ribeiro Junior, como visto na seção 3.2.5.

Utilizando esta técnica, automaticamente, são desconsiderados termos que possuem significado limitado, como artigos, preposições e conjunções, pois estas classes gramaticais não são extraídas. Assim, torna-se desnecessário o uso de *stopwords*. Contudo, ainda são eliminados termos considerados irrelevantes para o domínio.

4.2.2 Aplicação de Restrições

Algumas restrições foram consideradas na extração de termos: a forma canônica, o tamanho e o tipo da palavra. Estas restrições foram usadas também por Ribeiro Junior e por gerar bons resultados foram adotadas neste trabalho.

Para que as flexões das palavras não interferissem no cálculo da relevância estatística, foi levado em consideração a forma canônica (lema) dos termos, sem flexão de gênero e número. Por exemplo, as palavras “sou” e “é” possuem lema “ser” e, as palavras “cientista” e “cientistas” possuem lema “cientista”.

Também não são consideradas palavras que possuam tamanho menor que três. Entretanto, é importante citar que este tipo de restrição pode influenciar negativamente os resultados de extração de termos em domínios caracterizados por palavras pequenas, como na Química, onde os elementos da tabela periódica podem ser considerados termos relevantes mesmo com comprimento igual a um (RIBEIRO JUNIOR, 2008). Por fim, são retirados da lista de termos palavras que contenham caracteres especiais (com exceção de hífen e sublinhado), como “@”, “/”, “&”, e também palavras

que contenham números em sua composição, como “222m”, “1º” e “3/4”. Esta restrição é usada porque o parser PALAVRAS comete alguns erros ao classificar caracteres não alfanuméricos e também numerais coletivos, fracionários e multiplicativos, considerando-os como substantivos comuns. A mesma estratégia também é utilizada no trabalho de Lopes, como visto na seção 3.2.6.

4.2.3 Aplicação de Medidas Estatística

Após as restrições, para cálculo de relevância de termos são aplicados os métodos Frequência Relativa (FR) e TFIDF (*Term Frequence-Inverse Document Frequence*), descritos na seção 2.3.1.

Em seguida, os termos são reorganizados em ordem decrescente de relevância. E disponibilizados para o usuário, engenheiro da ontologia, que pode excluir da lista apresentada termos que considerar desnecessários ou incorretos. Esta listagem final é utilizada como entrada para a próxima etapa, descrita a seguir.

4.3 EXTRAÇÃO DE CONCEITOS

4.3.1 Extração Inicial de Conceitos

Nesta fase são extraídos padrões morfossintáticos presentes nas sentenças. Os padrões adotados são as regras propostas por Baségio, como visto na Tabela 3.1. Porém, só serão extraídos termos compostos que tenham pelo menos uma palavra pertencente à lista de termos gerada na etapa anterior.

4.3.2 Aplicação de Restrições

Nesta etapa também são consideradas as restrições feitas na extração de termos: forma canônica das palavras, seu tamanho e tipo. A única observação é que na restrição de tamanho da palavra não é levado em consideração palavras que pertençam as classes de preposição ou artigo. Logo, conceitos que possuem palavras como “de”, que tem tamanho menor que três, não devem ser excluídos.

Outra restrição usada somente nesta etapa é “preposição + artigo”, proposta primeiramente por Ribeiro Junior. Devido ao fato do parser PALAVRAS fazer a separação de contrações como “do” em duas palavras “de” e “o”, esta restrição permite que uma preposição seguida de um artigo seja considerada como uma única palavra.

4.3.3 Aplicação de Medidas Estatísticas

Após este processo, também são aplicadas as medidas estatísticas FR e TFIDF para avaliação da relevância de conceitos extraídos.

Por fim, a lista gerada é reorganizada em ordem decrescente de relevância, e em seguida, o usuário também pode excluir conceitos que considerar desnecessários ou irrelevantes para o domínio.

5 APLICAÇÃO LPhDic

Neste capítulo é apresentado em detalhes a aplicação LPhDic (*Linguistic* PhDic), desenvolvida utilizando uma metodologia orientada a objeto, visando implementar os métodos de extração de termos e conceitos apresentados no capítulo 4.

No desenvolvimento do LPhDic foi utilizada uma metodologia de Análise e Projeto Orientado a Objeto (A/POO) e processo de desenvolvimento iterativo e ágil, chamado Processo Unificado (PU). Foi investigado o domínio do problema, definindo claramente o que deveria ser feito e as necessidades requeridas. Para documentação da aplicação foi utilizada a notação UML de diagramação.

Por se tratar inicialmente de um sistema pequeno, o processo de desenvolvimento foi feito em apenas uma iteração, composta pelas fases de concepção e elaboração.

5.1 CONCEPÇÃO

É uma fase que deve ser curta, com propósito de estabelecer uma visão inicial e um escopo básico do projeto. Nesta fase foram criados os casos de uso da aplicação.

5.1.1 Casos de Uso

Inicialmente, foram definidos os casos de uso, requisitos funcionais ou comportamentais que indicam o que o sistema fará (LARMAN, 2007).

Nesta seção são mostrados os principais casos de uso, com os requisitos básicos da aplicação, a saber, “Configurar Sistema”, “Extrair Termos” e “Extrair Conceitos”.

As Tabelas 5.1, 5.2 e 5.3 representam os casos de uso do sistema e a Figura 5.1 apresenta o diagrama que os representa.

Tabela 5.1: Caso de uso "Configurar Sistema".

CDU1: Configurar Sistema
Escopo: Aplicação LPhDic
Nível: Objetivo do Usuário
Ator Principal: Usuário
Interessados e Interesses: Usuário deseja informar ao sistema as configurações necessárias para o restante das operações que o mesmo vier a realizar.
Pré-Condição: A aplicação deve estar executando.
Pós-Condição: Configurações foram feitas com sucesso.
Fluxo Básico: 1. Usuário fornece o diretório que contém os documentos que serão utilizados pela aplicação. 2. Usuário fornece o tipo de regra que deseja aplicar na extração de termos. 3. Usuário fornece o tipo de regra que deseja aplicar na extração de conceitos.
Fluxo Alternativo: 1a. Sistema detecta que o formato do documento não é Tiger-XML. 1. Sistema avisa usuário sobre o erro. 2. Usuário, provavelmente, fornece um novo diretório de documentos.

Tabela 5.2: Caso de uso "Extrair Termos".

CDU2: Extrair Termos
Escopo: Aplicação LPhDic
Nível: Objetivo do Usuário
Ator Principal: Usuário
Interessados e Interesses: Usuário deseja extrair dos documentos os termos que são considerados relevantes para o domínio.
Pré-Condição: Os documentos devem ter sido importados para a aplicação.
Pós-Condição: Uma lista de termos é gerada e apresentada para o usuário.
Fluxo Básico: 1. Sistema identifica um documento. 2. Para cada palavra do documento, o sistema verifica a tag “pos” e caso esta informação represente classes gramaticais relevantes (substantivo, adjetivo e advérbio), a palavra é selecionada. 3. Para cada palavra selecionada, o sistema verifica seu tamanho, caso seja menor que três essa palavra é desconsiderada. 4. Para cada palavra que não foi desconsiderada em 3, o sistema identifica o lema da palavra através da tag “lemma”. 5. Para cada lema selecionado, o sistema calcula medidas estatísticas de frequência e armazena essas informações. *Sistema repete passos de 1 a 5 até que acabe os documentos.* 6. O sistema apresenta a lista gerada para o usuário. 7. O usuário, se quiser, pode excluir algum termo da lista. 8. O sistema salva a lista final em um arquivo de texto “lista de termos”.
Fluxo Alternativo: 2a. Sistema não encontra a informação da tag “pos”. 1. Usuário é alertado. 2. Sistema aborta o processo (Pós-Condição:A lista de termos não é gerada).

Tabela 5.3: Caso de uso "Extrair Conceitos".

CDU3: Extrair Conceitos
Escopo: Aplicação LPhDic
Nível: Objetivo do Usuário
Ator Principal: Usuário
Interessados e Interesses: Usuário deseja extrair dos documentos os conceitos que são considerados relevantes para o domínio.
Pré-Condição: Os documentos devem ter sido importados para a aplicação e a lista de termos deve ser sido gerada.
Pós-Condição: Uma lista de conceitos é gerada e apresentada para o usuário.
Fluxo Básico: 1. Sistema identifica um documento. 2. O sistema verifica a tag "pos" presente em cada palavra do documento e caso seja encontrada a sequência "preposição + artigo", o sistema considera como se fosse uma única palavra e a tag considerada é "preposição". 3. Considerando a tag "pos" de cada palavra, o sistema busca por padrões morfossintáticos (regras de Baségio) e caso um padrão seja identificado, o conjunto de palavras é selecionado. 4. Para cada conjunto selecionado, o sistema verifica o tamanho de palavras que não sejam "preposições" ou "artigos", caso o tamanho de uma dessas palavras seja menor que três esse conjunto de palavras é desconsiderado. 5. Para cada conjunto que não foi desconsiderada em 4, o sistema identifica o lema da palavra através da tag "lemma" e caso pelo menos um dos lemas pertença a listagem final de termos gerados, o conjunto de palavras é selecionado. 6. Para cada conjunto selecionado em 5, o sistema calcula medidas estatísticas de frequência e armazena essas informações. *Sistema repete passos de 1 a 6 até que acabe os documentos.* 7. O sistema apresenta a lista de conceitos gerada para o usuário. 8. O usuário, se quiser, pode excluir algum conceito da lista. 9. O sistema salva a lista final de conceitos em um arquivo de texto "lista de conceitos".

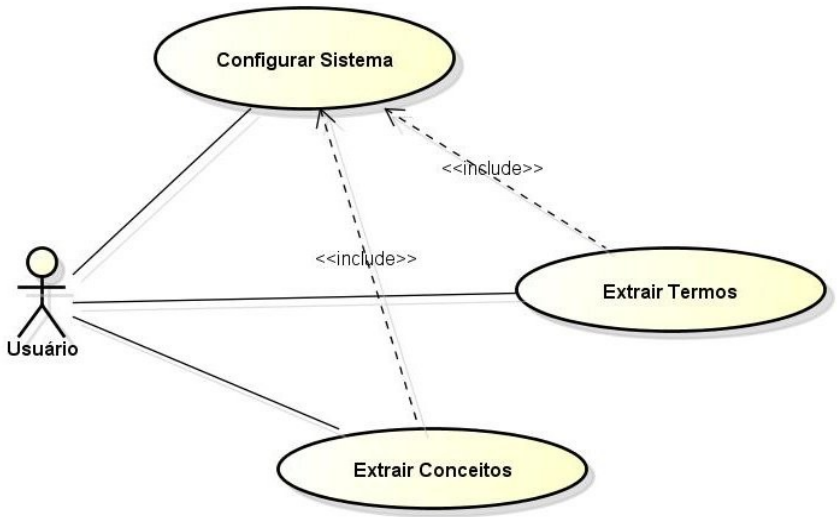

Figura 5.1: Diagrama de caso de uso do LPhDic.

5.2 ELABORAÇÃO

A elaboração é a série inicial de iterações durante a qual, em um projeto padrão (LARMAN, 2007):

- A arquitetura central e de alto risco do software é programada e testada;
- A maioria dos requisitos é descoberta e estabilizada;
- Os principais riscos são mitigados ou retirados.

Nesta fase foram criados os artefatos modelo de domínio, diagramas de sequência, diagrama de classes e implementação da arquitetura central.

5.2.1 Modelo de Domínio

O modelo de domínio é o primeiro passo para definir conceitos ou objetos que são de interesse do domínio. É a representação visual de classes conceituais.

Aplicando a notação UML, um modelo de domínio é representado com um conjunto de diagramas de classes em que nenhuma operação (assinatura de método) é definida. Ele fornece uma perspectiva conceitual, podendo mostrar: objetos de domínio ou classes conceituais; associações entre classes conceituais; e atributos de classes conceituais (LARMAN, 2007). Na Figura 5.2 é apresentado o diagrama do modelo de domínio da aplicação.

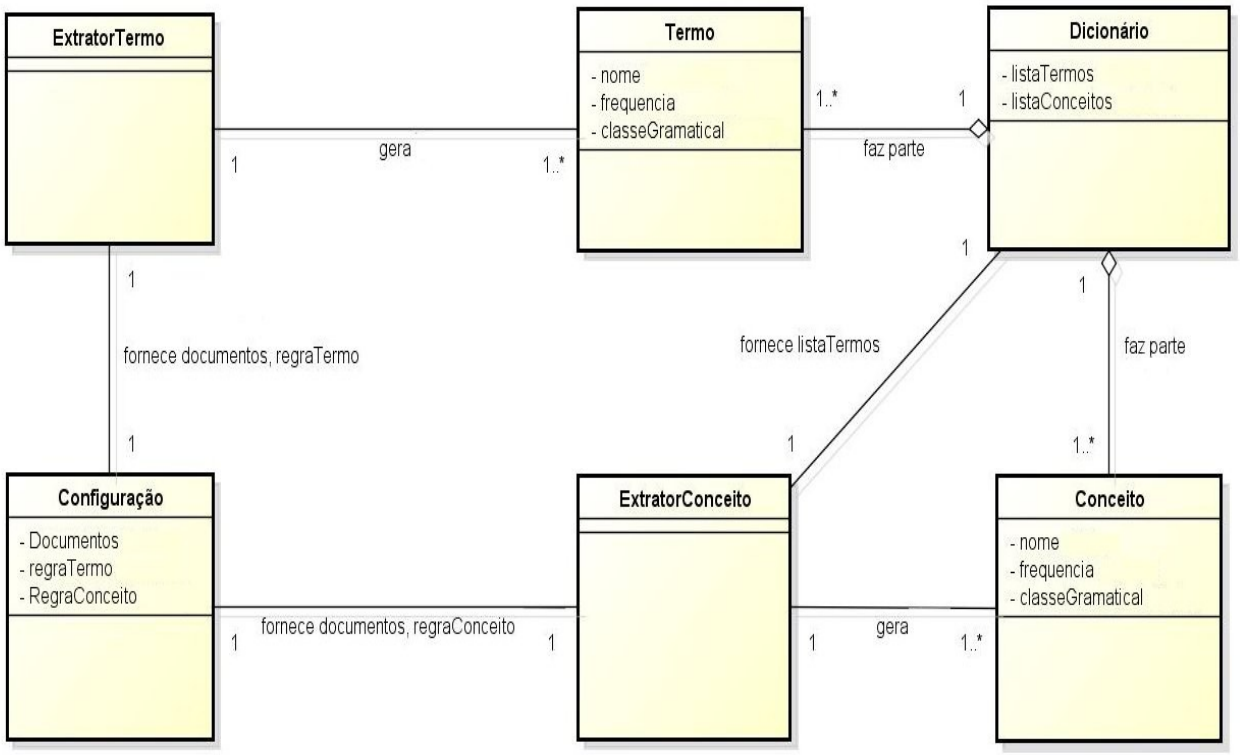

Figura 5.2: Diagrama do modelo de domínio do LPhDic.

5.2.2 Diagramas de Sequência

Os casos de uso descrevem como os atores externos interagem com o sistema a ser criado. Durante essa interação, um ator gera eventos para o sistema, geralmente solicitando alguma operação que trate o evento. Sendo assim, um diagrama de sequência ilustra tais eventos de entrada e saída relacionados com o sistema.

Enquanto no modelo de domínio são definidos os objetos, nos diagramas de sequência são feitas atribuições de responsabilidades e colaborações desses objetos. Nas Figuras 5.3, 5.4 e 5.5 são apresentados os diagramas de sequência para cada caso de uso da aplicação.

5.2.3 Diagrama de Classe

Os diagramas de classes são usados para modelagem estática de objetos, ilustrando classes com seus métodos e atributos, interfaces e associações.

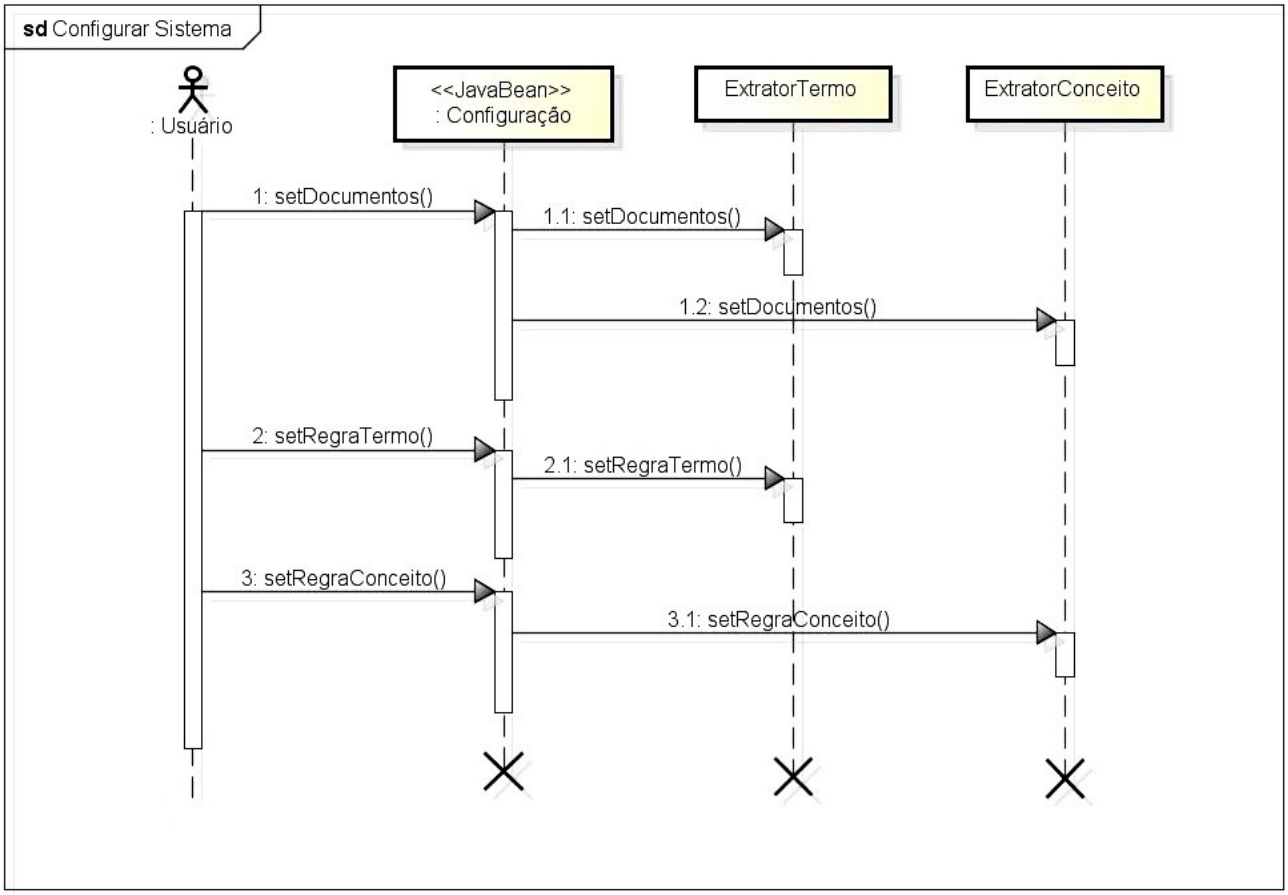

Figura 5.3: Diagrama de sequência "Configurar Sistema".

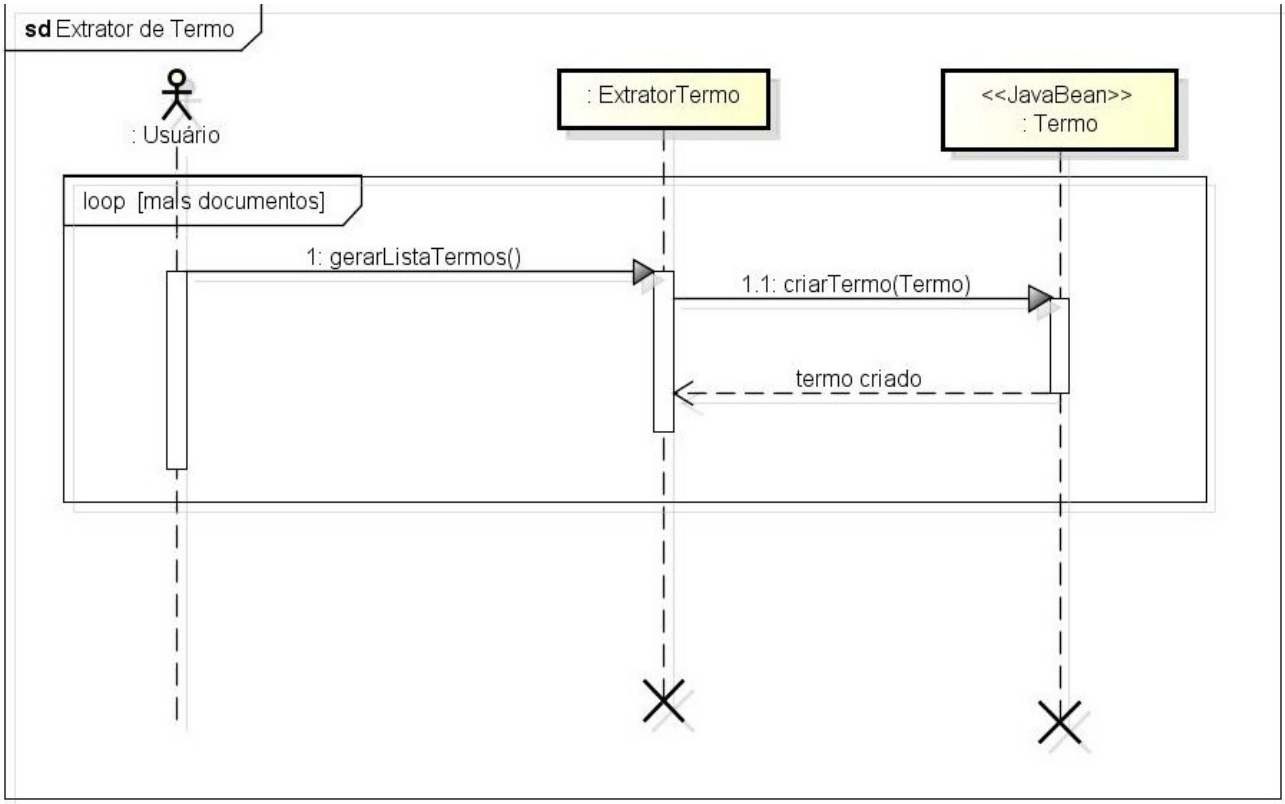

Figura 5.4: Diagrama de sequência "Extrair Termos".

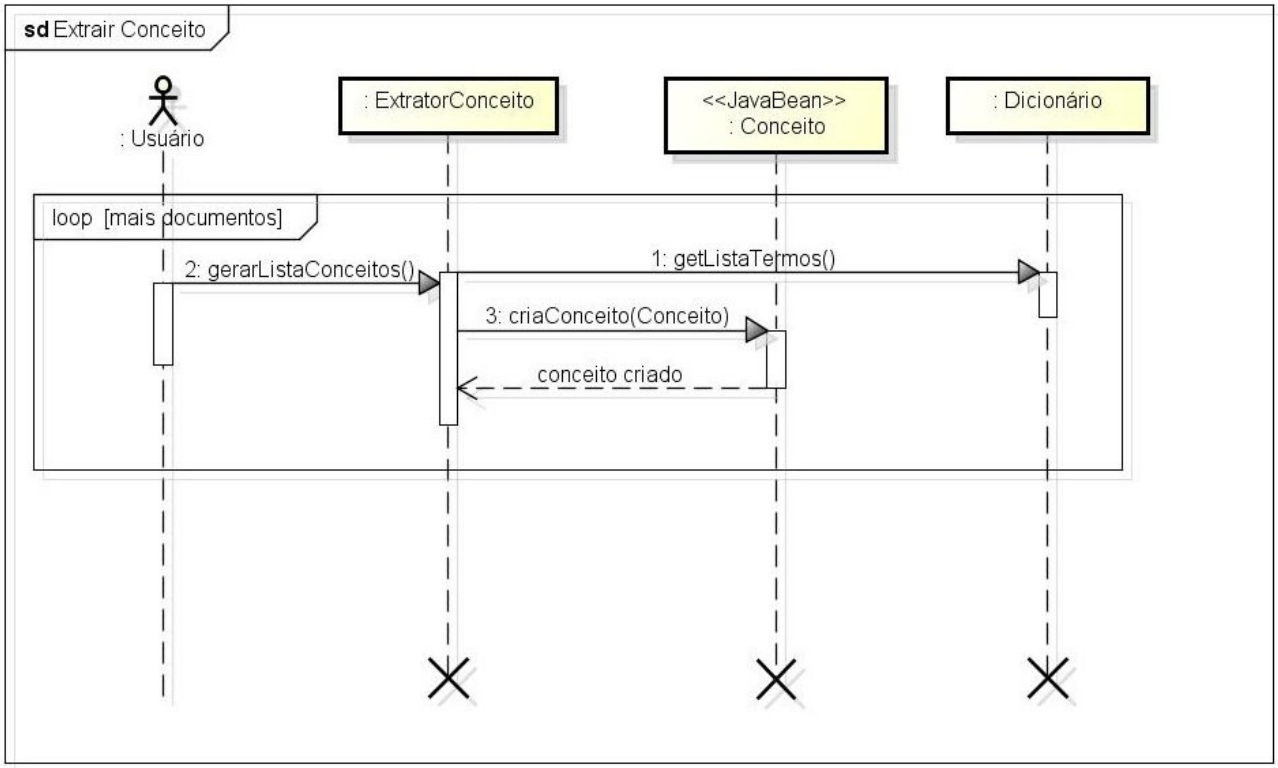

Figura 5.5: Diagrama de sequência "Extrair Conceitos".

A construção de diagrama de classes é feita a partir do modelo de domínio e dos diagramas de sequência, na qual os conceitos do domínio são listados como classes e os eventos como métodos de classes.

O diagrama de classes é usado em uma perspectiva do software, na qual questões de implementação já são consideradas. E neles já são usadas setas de navegação nas associações das classes. A Figura 5.6 apresenta o diagrama de classes da aplicação.

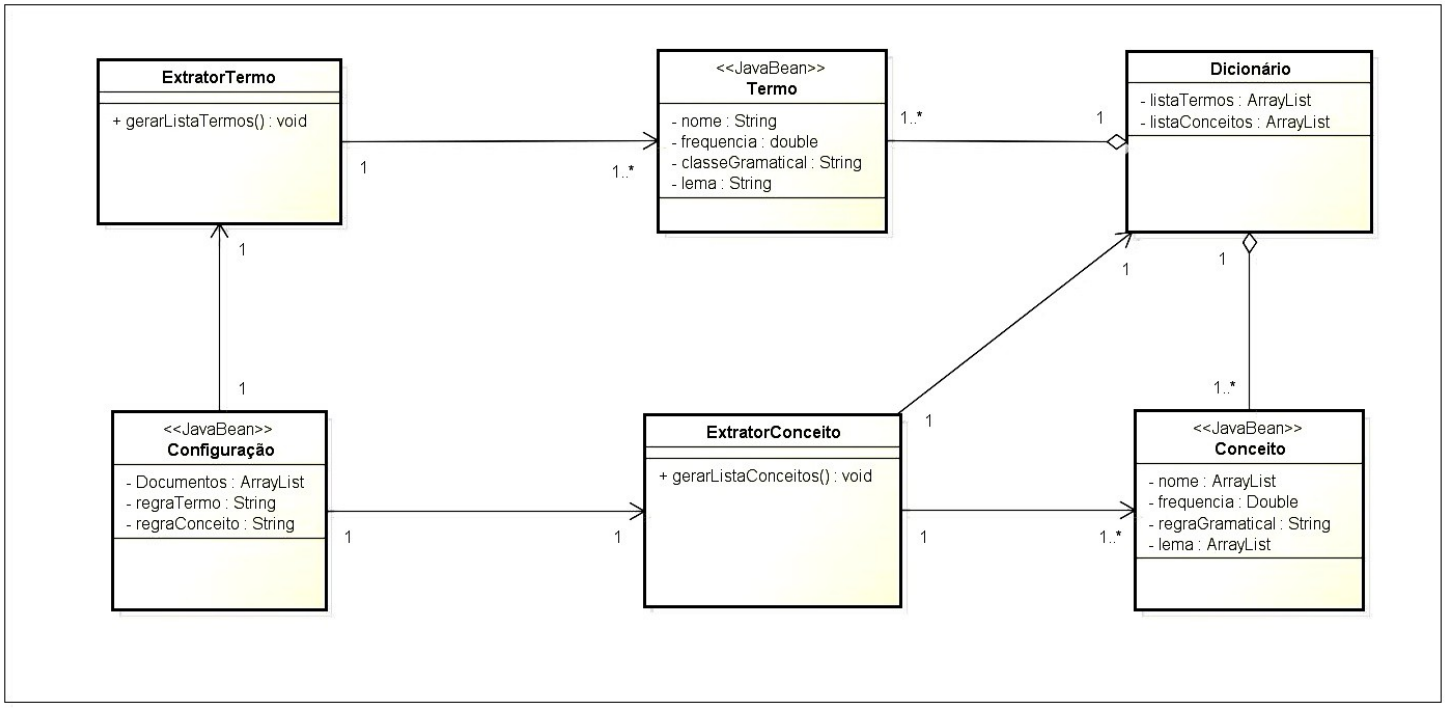

Figura 5.6: Diagrama de classes do LPhDic.

5.2.4 Implementação

Nesta seção são abordados os algoritmos desenvolvidos para os métodos de extração de termos e conceitos.

Extrator de Termos

Este método é responsável pela busca e extração de termos simples (unigramas) em cada um dos textos do domínio, que já devem ter sido processados pelo PALAVRAS e estão em formato Tiger-XML (Figura 2.1).

O método é dividido em três partes: extração inicial de termos, verificação de restrições e cálculo de frequências.

- **Extração Inicial de Termos**

O algoritmo apresentado na Figura 5.7 consiste em buscar em cada arquivo por sentenças do texto. E para cada sentença, são buscadas as palavras que a compõe, juntamente os seus atributos: palavra em sua forma original, forma canônica ou lema, classe gramatical e informação semântica.

```
Para cada documento
.  Para cada "s" (sentença)
.  .  Para cada "t" (palavra)
.  .  .  Para cada termo da lista (TermList)
.  .  .  .  Se o valor do atributo "word" for igual ao nome do termo
.  .  .  .  .  Adicionar 1 a frequência do termo no documento atual
.  .  .  Se o valor do atributo "word" não foi encontrado na lista de termos
.  .  .  .  Se o atributo "pos" for igual a "n" (substantivo) ou igual a
                "adj" (adjetivo) ou igual a "adv" (advérbio) e o atributo
                "sem" for diferente de "--" (ou seja, tem semântica)
.  .  .  .  .  Criar um novo termo com informações de nome, frequência,
                  classe gramatical e lema
.  .  .  .  .  Adicionar o termo a TermList
```

Figura 5.7: Algoritmo para extração de termos.

Em seguida, para cada palavra, é verificado se esta já não pertence a lista de termos que está sendo gerada. Se sim, então, somente as informações de frequência da palavra no texto atual são alteradas. Senão, então é verificado se a palavra é um substantivo, adjetivo ou advérbio e, se possui alguma informação semântica. Caso essas informações sejam verdadeiras, então a palavra é considerada como um novo termo, que é adicionado à lista de termos com todas as informações referentes a ele.

- **Verificação de Restrições**

Em seguida, após ter verificado todos os textos e gerado a lista de termos, são feitas algumas manipulações para aprimoramento dos termos gerados.

A primeira manipulação é fazer a verificação das restrições de tamanho e tipo da palavra, descritas na seção 4.2.2. Foi utilizada uma expressão regular para descrever o padrão que deveria ser encontrado, ou seja, para cada termo da lista de termos, se o padrão não é encontrado, então o termo é excluído. A Figura 5.8 mostra a expressão regular utilizada.

```
^([a-zA-ZÀ-ú]{3,}|([a-zA-ZÀ-ú]+[_//-][a-zA-ZÀ-ú]+)+)$
```

Figura 5.8: *Regex* para restrições.

A lógica da expressão criada é a seguinte:

- ^ - começo da palavra;
- **[a-zA-ZÀ-ú]{3,}** - a palavra é uma sequência de 3 ou mais caracteres (**{3,}**), que podem ser letras maiúsculas (**A-Z**) ou minúsculas (**a-z**), acentuadas ou não (**À-ú**);
- | - ou a palavra é uma sequência de ;
- **[a-zA-ZÀ-ú]+** - caracteres (**[a-zA-ZÀ-ú]**) que ocorrem pelo menos uma vez (+);
- [_//-] - seguido de sublinhado (_) ou hífen (-);
- **[a-zA-ZÀ-ú]+** - seguido de caracteres que ocorrem pelo menos uma vez;
- **()+** - ocorrendo pelo menos uma vez, ou seja, não é aceito pelo padrão palavra vazia;
- **$** - final da palavra.

- **Cálculo de Frequências**

Outra manipulação é calcular a frequência relativa e TFIDF para cada termo, pois no processo de extração somente são armazenadas as frequências absolutas do termo para cada documento do domínio. Neste ponto é levado em consideração a forma canônica (lema) da palavra para o cálculo da frequência, restrição descrita na seção 4.2.2. E antes de aplicar os cálculos é feita uma ordenação na lista, através de um comparador que ordena a lista considerando o atributo lema. A Figura 5.9 apresenta o algoritmo utilizado para fazer este cálculo.

E por fim, os termos são reorganizados em ordem decrescente de relevância. Está função é bastante simples, feita através de um comparador que ordena a lista de termos a partir do atributo frequência.

```
Criar uma lista auxiliar de termos (AuxList)
Adicionar na AuxList o primeiro elemento da lista de termos (TermList)
Para cada elemento da TermList a partir do segundo
.  Se valor do atributo lema do elemento anterior for igual ao valor de lema
        do elemento atual
.  .  Adiciona elemento atual na AuxList
.  Senão, se forem diferentes
.  .  Somar as frequências dos elementos da AuxList
.  .  Calcular a frequência relativa
.  .  Calcular a frequência TFIDF
.  .  Armazenar os valores calculados no atributo frequência de cada termo de
           TermList que pertença a AuxList
.  .  Limpar a AuxList
.  .  Adicionar na AuxList o elemento atual da TermList
```

Figura 5.9: Algoritmo para cálculo de frequências.

Extrator de Conceitos

Este método é responsável pela busca e extração de conceitos (termos compostos) em cada um dos textos do domínio, que estão no formato Tiger-XML. A busca é feita baseada nas regras propostas por Baségio, descritas na seção 3.2.2 (Tabela 3.1).

O método é dividido em quatro partes: representação das regras, extração inicial de conceitos, manipulações na lista de conceitos e cálculo de frequências.

- Representação das Regras

O primeiro passo é representar na forma de expressões regulares cada uma das regras gramaticais de Baségio. Para entender o contexto das expressões criadas, é preciso saber que para cada sentença do arquivo, é atribuído a uma variável de texto o índice de cada palavra seguido de sua classe gramatical. Assim, o padrão definido deve ser buscado nessas variáveis que representam sentenças.

Para o padrão "**[0-9]+n[0-9]+prp[0-9]*(art)?[0-9]+n**", por exemplo, é buscado na sentença:

- **[0-9]+** - qualquer sequência de pelo menos um caractere numérico (índice da palavra);
- **n** - seguido de "n" (classe da palavra que representa substantivo);
- **[0-9]+** - seguido de outra sequência numérica (índice de outra palavra);
- **prp -** seguido de "prp" (classe de outra palavra que representa preposição);

- **[0-9]*** - seguido de zero ou mais caracteres numéricos (outro índice);
- **(art)?** - seguido de “art” uma ou nenhuma vez (outra classe que representa artigo);
- **[0-9]+** - seguido de outro índice;
- **n –** e de outro substantivo.

Através do exemplo é possível observar que a restrição “preposição + artigo” é verificada nas sentenças, descrita na seção 4.3.2. Ou seja, uma preposição pode aparecer sozinha ou seguida de um artigo. A Tabela 5.4 apresenta exemplos de outras regras usadas e sua representação em *regex*.

Tabela 5.4: Exemplos de regras e sua representação em *regex*.

Regra Gramatical	Representação em *regex*
_SU _AJ	[0-9]+n[0-9]+adj
_SU _PR _SU	[0-9]+n[0-9]+prp[0-9]*(art)?[0-9]+n
_SU _PR _AD _SU	[0-9]+n[0-9]+prp[0-9]*(art)?[0-9]+adv[0-9]+n"
_SU _AJ _PR _AD _SU	[0-9]+n[0-9]+adj[0-9]+prp[0-9]*(art)?[0-9]+adv[0-9]+n

- **Extração Inicial de Conceitos**

O algoritmo apresentado na Figura 5.10 consiste em buscar em cada sentença de cada arquivo, as ocorrências que satisfazem cada padrão definido pelas expressões regulares. Em seguida, para cada ocorrência é verificado se ela já não pertence a lista de conceitos que está sendo gerada. Se sim, então, somente as informações de frequência do conceito no texto atual são alteradas. Senão, utilizando a listagem gerada na extração de termos, é verificado se pelo menos uma palavra da ocorrência pertence a lista de termos e caso essa informação seja verdadeira, a ocorrência é considerada como um novo conceito, que é adicionado à lista de conceitos com todas as informações referentes a ele.

```
Para cada regra de extração
.  Para cada documento
.  .  Para cada "s" (sentença)
.  .  .  Criar uma variável de texto vazia (string)
.  .  .  Para cada "t" (palavra)
.  .  .  .   Adicionar em string o índice da palavra seguido de sua classe
                  classe gramatical
.  .  .  Encontrar todas as ocorrências da regra atual
.  .  .  Para cada ocorrência encontrado
.  .  .  .  Recuperar as informações de "word" da ocorrência na sentença
.  .  .  .  Para cada conceito da lista (ConceptList)
.  .  .  .  .  Se os valores do atributo "word" da ocorrência forem iguais
.  .  .  .  .  .  Adicionar 1 a frequência do conceito no documento atual
.  .  .  .  Se a ocorrência não foi encontrada na ConceptList
.  .  .  .  .  Verificar se pelo menos uma palavra da ocorrência pertence a
                   TermList
.  .  .  .  .  Se sim,
.  .  .  .  .  .  Criar um novo conceito com informações de nome, frequência,
                      regra encontrada e lema
.  .  .  .  .  .  Adicionar o conceito a ConceptList
```

Figura 5.10: Algoritmo para extração de conceitos.

- **Cálculo de Frequência**

Após as manipulações, são calculadas as frequências relativas e TFIDF, da mesma maneira que é feita na extração de termos, só que aplicada a lista de conceitos.

E por fim, os conceitos são reorganizados em ordem decrescente de relevância, através de um comparador que ordena a lista de conceitos a partir do atributo frequência.

- **Manipulações na Lista de Conceitos**

Nesta etapa, após ter verificado todos os textos e gerado a lista inicial de conceitos, são feitas manipulações para aprimoramento dos conceitos gerados.

A primeira manipulação é fazer a verificação das restrições de tamanho e tipo das palavras que pertencem ao conceito. É utilizada aqui a mesma expressão regular criada para extração de termos, a única diferença é que se a palavra for uma preposição ou um artigo, então ela não é verificada pela expressão. Esta exceção é considerada porque muitas destas palavras possuem tamanho menor que três, porém, não devem ser excluídas, pois são relevantes para construção de um conceito.

Arquivos de Saída

Os resultados finais do processo de extração de termos e conceitos são disponibilizadas em arquivos de texto, que contém o conceito gerado seguido de sua frequência relativa e TFIDF. Os arquivos são gerados de acordo com o número de palavra que compõem o conceito. Para termos (unigramas) é gerado um único arquivo chamado "termos.txt". E para conceitos, por ser possível gerar conceitos de até seis palavras através das regras de Baségio, são gerados arquivos chamados "N_gramas.txt", para N variando de 2 a 6. A Figura 5.11 mostra um exemplo de arquivo de saída gerado para conceitos compostos por quatro palavras.

```
aumentos pontuais de torque 0.066 13.999
ameaças de prisão constantes 0.044 10.144
avanço com torque elevado 0.044 10.144
perda parcial de circulação 0.044 10.144
aplicação de compressão máxima 0.022 5.765
aumento abrupto de pressão 0.022 5.765
aumento gradativo da pressão 0.022 5.765
bbl de tampão fino 0.022 5.765
bbl de tampão lubrificante 0.022 5.765
bbl de tampão pesado 0.022 5.765
bbl de tampão tensoativo 0.022 5.765
centralizador superior do perfil 0.022 5.765
coluna por torque excessivo 0.022 5.765
conclusão da manobra curta 0.022 5.765
```

Figura 5.11: Exemplo de arquivo de saída da aplicação.

6 ANÁLISE DE RESULTADOS

Para análise de resultados foram processados 2.088 textos referentes à relatórios técnicos de anormalidades na perfuração e produção de petróleo. O número de termos e conceitos geradas pela ferramenta LPhDic, sem a intervenção do usuário, são apresentadas na Tabela 6.1.

Tabela 6.1: Número de termos e conceitos extraídos pelo LPhDic.

Lista	Nro. de Extraídos
unigramas - termos	728
bigramas	301
trigramas	1.245
quadrigramas	249
pentigramas	7
hexigramas	0

O desempenho da ferramenta LPhDic foi avaliado através de comparações entre as listas geradas de unigramas e bigramas com as respectivas listas de referência extraídas manualmente pelo especialista do domínio. Além disto, também foi feita uma comparação do desempenho com as ferramentas PhDic e ExATOlp . Para fazer as comparações foram usadas três medidas estatísticas: Precisão, Abrangência e *F-mesuare,* definidas na sessão 2.3.2.

Para geração de listas da ferramenta ExATOlp não houve nenhuma intervenção do usuário. Já para as extraídas pelo PhDic, foi definido um limiar mínimo de corte, sendo extraídos somente termos com frequência maior que 1. Os resultados são apresentados na Tabela 6.2.

Para a extração de termos (unigramas), a precisão da LPhDic foi menor do que a precisão das outras ferramentas e a abrangência foi maior do que a abrangência da ExATOlp, sendo a média harmônica entre essas duas medidas menor do que a média das outras duas ferramentas. Já para bigramas, a precisão da LPhDic foi maior do que a precisão das outras ferramentas e a abrangência foi maior do que a abrangência do PhDic, sendo a média harmônica entre essas duas medidas praticamente igual a da ExATOlp, que obteve melhor média.

Tabela 6.2: Comparação de Resultados.

LPhDic						
Tipo de Lista	**Termos Extraídos**	**Termos Corretos**	**Lista de Referência**	**Precisão (%)**	**Abrangência (%)**	***F-measure* (%)**
unigramas	728	204	654	28,02	31,19	29,52
bigramas	301	235	528	**78,07**	44,5	56,69
PhDic						
unigramas	947	322	654	34	**49,23**	**40,22**
bigramas	2383	198	528	8,3	37,5	13,59
ExATOlp						
unigramas	464	175	654	**37,71**	26,76	31,3
bigramas	463	281	528	60,69	**53,22**	**56,71**

7 CONCLUSÃO

Para o desenvolvimento deste trabalho, inicialmente foi feito um estudo detalhado das metodologias existentes para aquisição de ontologias a partir de textos em Língua Portuguesa. Em seguida, foi desenvolvida uma nova metodologia para aprimoramento dos métodos de anotação linguística e extração de conceitos da ferramenta PhDic - utilizada na construção do conhecimento e ontologias a partir de relatórios técnicos de anormalidades na perfuração e produção de petróleo.

A metodologia criada teve como base informações linguísticas e foi implementada na aplicação LPhDic (*Linguistic* PhDic), desenvolvida utilizando uma metodologia orienta a objeto.

A partir da comparação feita entre a ferramenta LPhDic e a PhDic, por meio da análise de resultados, foi possível observar que, apesar do PhDic ter obtido um resultado satisfatório para extração de termos, com uma abrangência de aproximadamente 50%, a ferramenta LPhDic obteve melhores resultados para extração de conceitos, com precisão de aproximadamente 78% para extração de bigramas, contra 8,3% do PhDic.

Além disto, através da adoção do anotador sintático PALAVRAS, foi possível melhorar a automatização do processo de extração de conceitos, haja vista que no processo usado pelo PhDic os termos extraídos são anotados manualmente com uma sintaxe pré definida pelo usuário. Porém, é importante ressaltar que a dependência de um analisador sintático pago pode ser considerada como uma limitação da ferramenta LPhDic.

No que diz respeito aos benefícios deste trabalho, cito o conhecimento adquirido em metodologias para extração de conceitos na aquisição de ontologias a partir de textos, em tecnologias e ferramentas utilizadas. Também o desenvolvimento integral do trabalho, desde a elaboração do projeto para submissão de bolsa de iniciação científica até sua implementação, trouxe enorme experiência acadêmica.

7.1 TRABALHOS FUTUROS

A partir deste trabalho, é possível identificar os seguintes trabalhos como continuidade da pesquisa:

- A partir das listas obtidas na extração de termos e conceitos, desenvolver a próxima etapa de um processo de aquisição de ontologias a partir de textos, hierarquia de conceitos;
- Possibilitar o uso de formatos de entrada para a ferramenta que tenham sido processados por analisadores sintáticos gratuitos e analisar as alterações de desempenho;
- Estudar os métodos de construção de anotadores sintáticos, para que a partir dos conceitos extraídos, possam ser criados anotadores de textos específicos para o domínio estudado.

REFERÊNCIAS

ALMEIDA, G. M. B.; VALE, O. A. **Do Texto ao Termo: Interação entre Terminologia, Morfologia e Linguística de Corpus na Extração Semi-automática de Termos**. In: ISQUERDO, A. N.; FINATTO, M. J. B. (Orgs.). As ciências do Léxico: Lexicologia, Lexicografia e Terminologia. 1ª ed. Campo Grande: Editora da UFMS, 2008, v. IV, p. 483-499.

BASÉGIO, T. L. **Uma Abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil**. Dissertação (Mestrado). Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, 2006.

BERNERS-LEE, T.; HANDLER, J.; LASSILA, O. **The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities**. Scientific American, Maio de 2001. Disponível em http://www.sciam.com/article.cfm?id=the-semantic-web&page=1>. Acesso em: 31 Jul. 2009.

BICK, E. The parsing System "Palavras": **Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework**. PhD thesis, Arhus University, 2000.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. **Ontology Learning from Text: An Overview.** IOS Press, 2003.

GUARINO, N.; GIARETTA, P. **Ontologies and Knowledge Bases: Towards a Terminological Clarification. Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing**, p. 25–32, 1995. Disponível em: http://www.csee.umbc.edu/771/papers/KBKS95.pdf.Z. Acesso em: 5 de Ago. 2009.

GUILHERME, I. R.; SERAPIAO, A. B. S.; RABELO, C.; MENDES, J. R. P.; **An Ontology Based for Drilling Report Classification**; 2006; Lecture Notes in Computer Science, v. 1, p.inicial 1037, p.final 1046, ISSN: 0302-9743.

KONIG, E.; LEZIUS, W.;VOORMANN, H. **TIGERSearch 2.1 – User's Manual**, IMS, University of Stuttgart, 2003. Disponível em: http://www.ims.unistuttgart.de/projekte/TIGER/TIGERSearch/doc/html/. Acesso em: 27 de Set. 2009.

LARMAN, C. **Utilizando UML e Padrões: Uma Introdução à Análise e ao Projeto Orientado a Objetos e ao Desenvolvimento Iterativo**. Porto Alegre: Bookman, 2007.

LOPES, L.; FERNANDES, P. ; VIEIRA, R. ; FEDRIZZI, G. . **ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora.** In: LTC'09 - 4th Language and Technology Conference, 2009, Poznan, 2009, Poznan. Proceedings of the Fourth Language and Technology Conference. Poznan : Adam Mickiewicz University, 2009. p. 427-431.

LOPES, L. ; OLIVEIRA, L. H. M ; VIEIRA, R. . **Portuguese Term Extraction Methods: Comparing Linguistic and Statistical Approaches**. In: International Conference on Computational Processing of Portuguese Language - PROPOR, 2010, Porto Alegre. Proceedings of PROPOR 2010,2010. p. 1-6.

RIBEIRO JUNIOR, L. C. **OntoLP: Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa**. 2008. 68 f. Dissertação (Mestrado) - Universidade Do Vale Do Rio Dos Sinos, São Leopoldo, 2008. Disponível em: <http://www.inf.pucrs.br/~ontolp/downloads-ontolpplugin.php>. Acesso em: 31 Jul. 2009.

SANTANA, F. A. P. **Métodos para Construção Semi-Automática de Ontologias a Partir de Textos**. Relatório Técnico, Universidade Estadual Paulista, Rio Claro, 2009.

SILVA, M. C. P. d. S.; KOCH, I. G. V. **Lingüística Aplicada ao Português: Sintaxe.** São Paulo: Cortez, 2001. ISBN 85-249-0204-3. Disponível em: <citeseer.ist.psu.edu/635422.html>.

TELINE, M. F.; MANFRIN, A. M. P.; ALUÍSIO, S. M. **Extração Automática de Termos de Textos em Português: Aplicação e Avaliação de Medidas Estatísticas de Associação de Palavras.** Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC-ICMC-USP, São Carlos, 2003.

TELINE, M.F.; ALMEIDA, G.M.B.; ALUÍSIO, S.M. **Extração Manual e Automática de Terminologia: Comparando Abordagens e Critérios** . In: 1o. Workshop em Tecnologia da Informação e da Linguagem Humana, 2003, São Carlos. *Anais do TIL'2003*, 2003.