PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

FACULDADE DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil

TÚLIO LIMA BASÉGIO

Dissertação de Mestrado

PORTO ALEGRE
2007

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL

FACULDADE DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TÚLIO LIMA BASÉGIO

Uma abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre, pelo Programa de Pós-Graduação em Ciência da Computação da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profª. Drª. Vera Lúcia Strube de Lima

Porto Alegre

2007

Dados Internacionais de Catalogação na Publicação (CIP)

B299a Baségio, Túlio Lima

Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil / Túlio Lima Baségio. – Porto Alegre, 2008.

124 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.

Ficha Catalográfica elaborada pelo

Setor de Tratamento da Informação da BC-PUCRS

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

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada ***"Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil"***, apresentada por Túlio Lima Baségio, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 05/01/2007 pela Comissão Examinadora:

Profa. Dra. Vera Lúcia Strube de Lima – PPGCC/PUCRS
Orientador (a)

Prof. Dr. Marcelo Blois Ribeiro – PPGCC/PUCRS

Profa. Dra. Renata Vieira – UNISINOS

Homologada em 14/04/2009, conforme Ata No. 006 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 16 – sala 106 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@inf.pucrs.br

AGRADECIMENTOS

À minha orientadora, professora Vera Lúcia Strube de Lima, por toda sua dedicação, ensinamentos e principalmente compreensão.

Aos meus pais, Lauri e Marlene, pelo apoio, paciência, cobranças e compreensão que tiveram neste período. Tudo o que eu consegui até hoje devo a vocês.

À minha namorada, Ana Paula, pelo amor, carinho e apoio em todos os momentos. Eu sei que muitos deles não foram fáceis. Tua dedicação e companherismo foram fundamentais.

À meu irmão, Felipe, por sempre acreditar e incentivar meu trabalho.

Aos amigos que fiz durante o curso e que compartilharam bons e maus momentos.

Às professoras Simone Sarmento e Marutschka Moesch pelo apoio na avaliação deste trabalho.

Ao convênio Dell-PUCRS por viabilizarem a bolsa de estudos durante quase todo o mestrado.

Ao Programa de Pós-Graduação em Ciência da Computação e a todos os professores dos quais pude conviver durante o curso.

"É melhor tentar e falhar que preocupar-se e ver a vida passar.
É melhor tentar, ainda que em vão, que sentar-se fazendo nada até o final.
Eu prefiro na chuva caminhar que em casa me esconder
Prefiro ser feliz embora louco, que em conformidade viver…"

(Martin Luther King)

RESUMO

Para várias áreas de aplicação, a construção semi-automática ou automática de ontologias seria extremamente útil. Abordagens semi-automáticas para a extração de ontologias a partir de textos têm sido propostas na literatura, as quais sugerem a extração de conhecimento encontrado nos textos de um domínio, com o apoio de técnicas de processamento da língua natural. Este trabalho propõe uma abordagem para suportar algumas fases do processo de aquisição de estruturas ontológicas, mais especificamente as fases de extração de conceitos e relações taxonômicas, de modo a semi-automatizar os passos da construção de ontologias a partir de textos na língua portuguesa do Brasil. O resultado obtido serve como ponto de partida ao engenheiro de ontologia. Para avaliação da abordagem proposta, foi desenvolvido um protótipo que incorpora mecanismos de importação de corpus, identificação de termos relevantes, identificação de relações taxonômicas entre esses termos e geração da estrutura ontológica em OWL. Este protótipo foi utilizado num estudo de caso sobre o domínio do Turismo, possibilitando a avaliação com relação a diferentes aspectos do processo de aquisição de conceitos e relações.

Palavras-Chave: Construção de ontologias a partir de textos, Processamento da língua Portuguesa.

ABSTRACT

Automatic or semi-automatic ontology building would be extremely useful for several application areas. Semi-automatic approaches for ontology extraction from texts have been proposed in the literature, which suggest knowledge extraction from texts of a certain domain supported by natural language processing techniques. This work proposes an approach to support some phases of the acquisition of ontological structures, more specifically the phases of concept extraction and taxonomic relations extraction, in order to semi-automatize the steps to build ontologies from Brazilian Portuguese texts. The results from these phases represent an initial structure to help the ontology engineer in the ontology building process. The evaluation of this approach was done through a prototype developed with functionalities such as corpus uploading, identification of relevant terms and taxonomic relations among these terms, additionally providing ontological structure generation in OWL. This prototype was used in a case study on the Tourism domain, enabling the evaluation of different aspects of the concepts and relations acquisition process.

Key-words: Ontology extraction from texts, Portuguese language processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS

ADS	Ambiente de Desenvolvimento de Software
CoPS	*Comunity of Practices*
HTML	*HyperText Markup Language*
IA	Inteligência artificial
MO	Memória organizacional
NILC	Núcleo Interinstitucional de Lingüística Computacional
ODE	*Ontology-based software Development*
OWL	*Web Ontology Language*
PDF	*Portable Document Format*
PLN	Processamento de linguagem natural
SMES	*Saarbrücken Message Extraction System*
TF	*Term frequency*
TFIDF	*Term frequency x inverted document frequency*
XML	*eXtensible Markup Language*
XSL	*Extensible Stylesheet Language*
XTM	*XML Topic Maps*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto geral

Há alguns anos o termo “ontologia” remetia apenas a um campo da Filosofia. Porém, a partir do início dos anos 90, ontologias se tornaram um tópico de investigação e aplicação em comunidades de pesquisa de Inteligência Artificial (IA), incluindo engenharia do conhecimento, representação do conhecimento e processamento de linguagem natural. Hoje, encontramos pesquisas em torno de ontologias sendo desenvolvidas em diversas áreas, tais como sistemas de informação cooperativos, recuperação de informação, comércio eletrônico e gestão do conhecimento, entre outras [BAS04a].

Segundo Fensel [FEN03], a principal razão para esse interesse é que ontologias “prometem” entendimento comum e compartilhado de um mesmo domínio, que pode ser comunicado entre pessoas e sistemas de aplicação. Noy e McGuinness, em [NOY01], também citam algumas razões para se construir uma ontologia:

- Compartilhar entendimento comum da estrutura de informação entre pessoas e agentes de software;
- Permitir reuso do conhecimento de um domínio;
- Criar compreensão do domínio explicitamente;
- Separar o conhecimento do domínio do conhecimento operacional;
- Permitir a análise do conhecimento do domínio.

Mas, apesar do número de pesquisas sobre o tema ontologia, sua construção ainda é uma questão crítica. Segundo Blázquez em [BLA98], não existe uma metodologia generalizada e suficientemente testada para construção de ontologias e, dessa forma, os engenheiros de ontologias acabam criando seu próprio processo. A afirmação de Blázquez ainda pode ser considerada válida hoje, visto que várias propostas de metodologias têm sido apresentadas na literatura nos últimos anos como, por exemplo, as encontradas em [USC95], [FER97], [FAL98], [STA01], [HOL02] e [KIS04].

De acordo com Noy e McGuinness em [NOY01], não existe uma metodologia “mais correta” a ser usada na construção de ontologias e não há modo “mais correto” para modelar um domínio, pois sempre existem alternativas viáveis. A escolha da solução vai depender da aplicação que se tem em mente e do escopo desejado para a ontologia.

Metodologias como as encontradas em [USC95], [FER97], [FAL98], [STA01], [HOL02] e [KIS04] incluem a construção manual de ontologias suportada por algumas ferramentas computacionais. Porém, conforme [BRE03] e [MAE02], a construção manual de ontologias é um processo complexo, tedioso e de alto custo, e por ser um processo extremamente artesanal é também propenso a erro. Além disso, a manutenção de uma ontologia, inclusão ou alteração de conceitos existentes, também pode ser onerosa.

Sendo assim, para várias áreas de aplicação, a construção semi-automática ou automática de ontologias seria extremamente útil e este fato tem levado pesquisadores a propor, num primeiro momento, abordagens para construção semi-automática de ontologias.

Entretanto, para se construir uma ontologia de modo automático ou semi-automático, é necessária uma fonte a partir da qual possa ser extraído conhecimento relevante. Nesse contexto, textos eletrônicos são fonte de informação bastante interessante, pois a cada dia que passa, mais textos estão disponíveis ao acesso das pessoas, representando o conhecimento dos mais diversos domínios.

Porém, de acordo com Davies *et al*. em [DAV02], existem alguns problemas, não triviais, relacionados à construção de ontologias a partir de textos como, por exemplo: o que deve ser representado? Um outro problema, apontado em [BRE03], é a identificação de fontes corretas e apropriadas para a extração do conhecimento.

Apesar destes problemas, abordagens e métodos utilizados para a construção (automática ou semi-automática) de ontologias a partir de textos têm sido propostos na literatura como, por exemplo, os encontrados em [VEL01], [MAE02], [LAM03], [BUI04], [CEL04] e [DEG04]. Estas frentes propõem a construção semi-automática a partir do conhecimento encontrado nos textos de um dado domínio, com o apoio de técnicas de processamento de linguagem natural (PLN) [BAS04b].

Face ao exposto, constituem a motivação deste trabalho: o significado e a importância de ontologias na solução de problemas de Inteligência Artificial; os problemas relacionados à construção manual de ontologias; e o surgimento de abordagens semi-automáticas para a construção de ontologias a partir de textos. Outra importante motivação para este trabalho é que, durante a pesquisa realizada, não foram encontradas abordagens para a construção automática ou semi-automática de ontologias a partir de textos da língua portuguesa do Brasil. O trabalho mais próximo, em se tratando de textos escritos em nosso idioma, é o apresentado

em [PER04], onde a autora propõe um processo semi-automático para a extração de conhecimento a partir de textos da língua portuguesa do Brasil, voltado à construção de mapas conceituais.

Nesse sentido, a proposta de uma abordagem para semi-automatizar os passos típicos do processo de construção de ontologias (identificação de conceitos, de relações taxonômicas e não-taxonômicas e identificação de instâncias) a partir textos na língua portuguesa do Brasil é o que nos vem em mente. Porém, devido à complexidade e diversidade de técnicas envolvidas em cada um destes passos, fica inviável neste trabalho uma solução completa que integre todos os passos mencionados. No contexto desta dissertação, "conceito" e "termo" são usados sem distinção, isto é, tem o mesmo significado.

Visando evoluirmos em direção a uma abordagem para semi-automatizar os passos da construção de ontologias a partir textos na língua portuguesa do Brasil, este trabalho apresenta uma proposta para suportar algumas fases deste processo, mais especificamente a extração de conceitos e de relações taxonômicas (ou relações hierárquicas entre conceitos). O resultado da execução destas fases será o que chamamos de estrutura ontológica, ou seja, uma estrutura inicial que servirá como ponto de partida ao engenheiro da ontologia.

Apesar da variedade de relações entre palavras que podem ser encontradas quando analisando um corpus, a escolha por trabalhar com relações taxonômicas se deve ao fato de que, segundo [RUI05], na maioria dos casos, ontologias são estruturadas como hierarquias de conceitos (taxonomias), ou seja, arranjam conceitos dos mais gerais aos mais específicos.

Neste sentido, emergiu a questão de pesquisa deste trabalho:

"É possível avançar na proposta de uma abordagem para aquisição de estruturas ontológicas a partir de textos da língua portuguesa do Brasil?".

Partindo-se desta questão de pesquisa, tínhamos a seguinte hipótese:

- Existem algoritmos e técnicas que podem ser adaptados para o processo de construção de estruturas ontológicas a partir de textos da língua portuguesa do Brasil, apoiando este processo através da geração de resultados parciais.

1.2 Objetivo do trabalho

Visando evoluirmos em direção à solução do problema colocado, o principal objetivo deste trabalho foi desenvolver uma abordagem para suportar fases do processo de aquisição

de estruturas ontológicas a partir de textos na língua portuguesa do Brasil, mais especificamente as fases de extração de conceitos e relações taxonômicas.

1.2.1 Objetivos específicos

- Propor/adaptar mecanismos que facilitem a extração de conceitos relevantes ao domínio ao qual os textos pertencem;
- Propor/adaptar mecanismos que facilitem a extração de relações taxonômicas entre os conceitos extraídos previamente;
- Propor/adaptar mecanismos que auxiliem a criação de uma estrutura ontológica a partir dos dados extraídos previamente;
- Definir um ambiente de apoio aos mecanismos anteriores.

1.3 Método de pesquisa

A Figura 1.1 representa as principais etapas desenvolvidas na condução deste trabalho. Inicialmente foram estudadas as principais fontes sobre o processo de construção de ontologias, donde se obteve um entendimento geral sobre o que são ontologias e suas aplicações. Posteriormente, o estudo restringiu-se às abordagens para a construção de ontologias a partir de textos, onde o problema motivador para esta pesquisa, a falta de uma abordagem para a construção de ontologias a partir de textos da língua portuguesa do Brasil, foi identificado.

Estas abordagens forneceram subsídios para a definição das funcionalidades de apoio necessárias para a consecução da proposta aqui apresentada. O passo seguinte consistiu no desenvolvimento de um ambiente de apoio através de um protótipo, o qual teve seus resultados analisados através de dois estudos de caso.

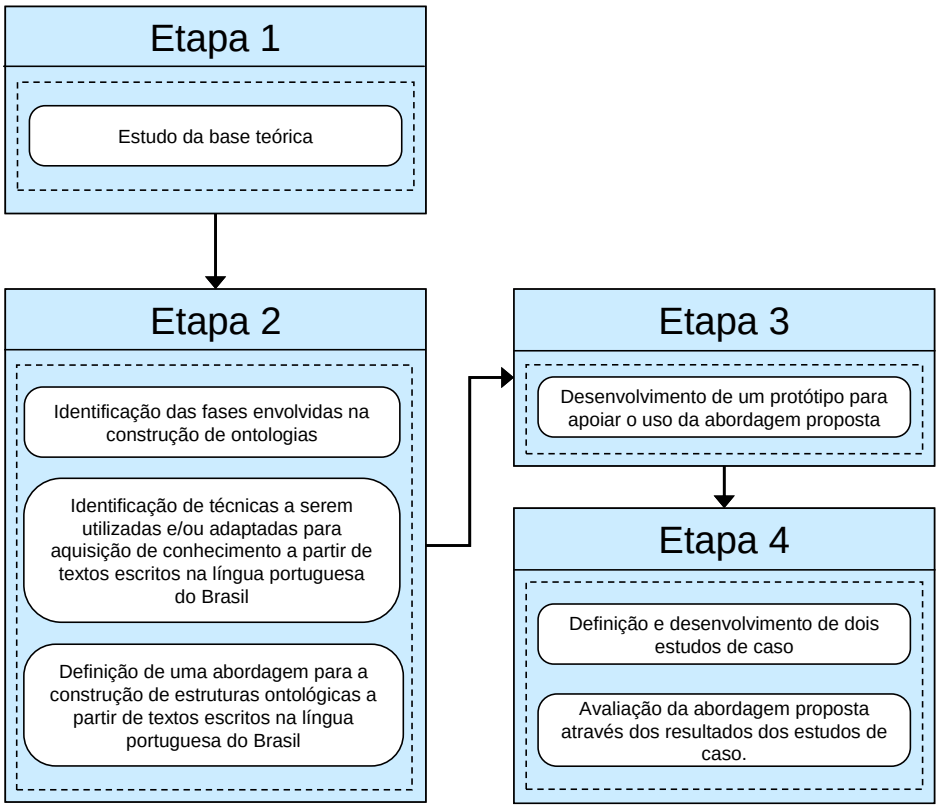

Figura 1.1: Etapas desenvolvidas na condução da pesquisa

1.4 Organização do texto

Este trabalho está organizado em 7 capítulos. O Capítulo 2 apresenta conceituações e definições para o termo ontologia e a temática da sua classificação, fornecendo ainda uma visão geral sobre diferentes áreas de aplicação de ontologias. O Capítulo 3 aborda uma questão crítica em relação a ontologias: sua construção. Nele são apresentados aspectos e abordagens propostos na literatura relacionados à construção de ontologias a partir de textos, sendo aqui reportados os principais trabalhos relacionados à proposta deste trabalho.

O Capítulo 4 descreve os mecanismos propostos para auxiliar na construção de estruturas ontológicas a partir de textos da língua portuguesa do Brasil e o Capítulo 5 descreve o ambiente de apoio à abordagem proposta. O Capítulo 6 apresenta os estudos de caso realizados a partir desta proposta, fornecendo uma análise dos resultados obtidos. O Capítulo 7 discorre sobre conclusões e trabalhos futuros. Por último, encontram-se as Referências Bibliográficas e os anexos.

2 ONTOLOGIAS

Este capítulo apresenta definições e classificações de ontologias, relacionando também algumas áreas de aplicação.

2.1 Conceituação

O termo "ontologia" originou-se na Filosofia, através de Aristóteles, que o utilizava no intuito de especificar o que existe ou o que podemos dizer sobre o mundo. Segundo Alexander Maedche em [MAE02], "Ontologia é um ramo da Filosofia que lida com a natureza e a organização do ser, e foi introduzido por Aristóteles na Metafísica". Aristóteles estava interessado nas definições das coisas. Sua noção de definição não era simplesmente o significado de uma palavra. Uma definição tinha a intenção de explicar claramente "o que uma coisa é" por ser um relato da essência da entidade. Ele acreditava que, para dizer "o que uma coisa é", precisava-se sempre dizer "por que alguma coisa é" [MAE02].

Na Ciência da Computação, o uso do termo ontologia teve origem na comunidade de IA. Segundo Thomas Gruber em [GRU93a], para sistemas de IA o que existe é aquilo que pode ser representado e, quando o conhecimento de um domínio é representado através de um formalismo declarativo, o conjunto de objetos que pode ser representado é chamado de universo do discurso. Mas nem sempre esse conceito precisa ser tão abrangente: Russel e Norvig, em [RUS03], definem ontologia como uma teoria da natureza do ser ou da existência, expressa por meio de um vocabulário. Tais autores consideram uma ontologia apenas como um vocabulário, ou seja, uma lista informal dos conceitos em um domínio.

O termo ontologia pode assumir diferentes significados para a Filosofia e para a comunidade de Ciência da Computação. No sentido filosófico, pode-se referir a uma ontologia como um sistema particular de categorias a considerar para uma certa visão do mundo e, no seu uso mais comum na Ciência da Computação, uma ontologia refere-se a um artefato de engenharia, constituído por uma espécie de vocabulário usado para descrever uma certa realidade [MAE02].

Dentre as várias definições existentes para o termo ontologia na literatura, a mais citada é a oferecida por Gruber [GRU93a]: "Uma ontologia é uma especificação explícita e

formal de uma conceituação compartilhada". Segundo Fensel [FEN97], que remonta esse assunto:

- Conceituação: refere-se a um modelo abstrato de algum fenômeno do mundo, por terem sido identificados os conceitos relevantes para aquele fenômeno.
- Explícito: significa que o conjunto de conceitos utilizados e as restrições aplicadas são previamente e explicitamente definidos.
- Formal: refere-se ao fato de que se espera que uma ontologia seja processável por computador, o que exclui definições diretamente em linguagem natural, por exemplo.
- Compartilhada: descreve um conhecimento consensual, que é utilizado por mais de um indivíduo e aceito por um grupo.

Várias outras definições para o termo ontologia têm sido adotadas conforme o contexto de sua utilização. Para Neches e co-autores [NEC91], "uma ontologia define termos e relações compreendendo o vocabulário de um tópico de uma área assim como as regras para combinar termos e relações para definir extensões para o vocabulário". Já para Chandrasekaran *et al*. [CHA99]: "Ontologias são teorias de conteúdo sobre tipos de objetos, propriedade dos objetos, e os relacionamentos entre os objetos que são possíveis em um domínio específico do conhecimento".

Uschold e co-autores, em [USC98], afirmam que uma ontologia é uma reunião explícita de conhecimento compartilhado em uma área específica e que, conseqüentemente, pode resolver problemas de comunicação entre pessoas, organizações e aplicativos. Studer [STU98] comenta que uma das finalidades das ontologias é permitir a interoperabilidade de fontes heterogêneas de informação e, assim, realizar o compartilhamento de conhecimento.

Uma ontologia, segundo Noy e McGuinness em [NOY01], é formada basicamente pelos seguintes componentes: classes (organizadas em uma taxonomia), relações (representando os tipos de interação entre os conceitos de um domínio), axiomas (usados para modelar sentenças sempre verdadeiras) e instâncias (utilizadas para representar elementos específicos, ou seja, os próprios dados).

Maedche, em [MAE02], oferece uma definição mais formal, na qual uma ontologia é constituída por cinco elementos, $O = \{C, R, H^C, rel, A^O\}$, onde:

- *C* refere-se a um conjunto de conceitos, muitas vezes chamados de classes;
- *R* refere-se a um conjunto de relações;
- H^C representa a hierarquia dos conceitos, também chamada de taxonomia. Por exemplo, H^C(*C1*, *C2*) significa que *C1* é um sub-conceito de *C2*.
- *rel* é uma função referente a um relacionamento não-taxonômico entre conceitos. Por exemplo, (*R*)=(*C1*, *C2*) significa que *C1* e *C2* estão relacionados de forma não-taxonômica através de (R).
- A^O é um conjunto de axiomas da ontologia, expressos em linguagem lógica apropriada (por exemplo, lógica de primeira ordem).

2.2 Classificação de ontologias

As ontologias apresentam propriedades distintas mas, mesmo assim, é possível identificar tipos bem definidos de ontologias. Existem diferentes classificações para ontologias na literatura, sendo que algumas propostas definem tipos de ontologias relacionando a capacidade de uma ontologia modelar um domínio ou, por exemplo, tipos de ontologia quanto a sua aplicação. A seguir apresentamos algumas destas classificações.

Segundo Vergara em [VER03], ontologias podem ser classificadas de acordo com a forma como são capazes de modelar as informações de um determinado domínio. Sendo assim, ontologias que dificultam as inferências, por não possuírem axiomas e restrições, são classificadas como superficiais (*lightweight*), enquanto que ontologias que incluem todos os elementos que permitem inferências sobre o conhecimento que representam, são classificadas como profundas (*heavyweigth*).

Em [HEI97], Heijst classifica ontologias nas quatro categorias apresentadas a seguir:

- **Ontologias de representação do conhecimento:** capturam a representação primitiva usada para formalizar conhecimento em paradigmas de representação de conhecimento. Explicam as conceituações que fundamentam os formalismos de representação de conhecimento.
- **Meta-ontologias:** definem conceitos tais como estado, evento, processo, ação, etc., com o intuito de serem especializadas na definição de conceitos em uma ontologia de domínio. Esse tipo de ontologia é também chamado de ontologia genérica.

- **Ontologias de domínio:** fornecem um vocabulário dos conceitos (dentro de um domínio) e seus relacionamentos, das atividades que acontecem naquele domínio e das teorias e princípios elementares que governam aquele domínio.
- **Ontologias de aplicação:** contêm o conhecimento necessário para modelar uma aplicação em particular.

Guarino, em [GUA98], apresenta uma classificação de ontologias com base no conteúdo, que se organiza em:

- **Ontologias de alto nível:** provêem noções gerais às quais todos os termos nas ontologias estão relacionados. Equivalem basicamente as meta-ontologias genéricas de Heijst.
- **Ontologias de domínio:** têm a mesma característica encontrada na definição de Heijst. Descrevem o vocabulário relacionado a um domínio genérico pela especialização dos conceitos introduzidos na ontologia de alto nível.
- **Ontologias de tarefas:** descrevem o vocabulário relacionado a uma tarefa ou atividade genérica pela especialização das ontologias de alto nível. Seu objetivo é facilitar a integração dos conhecimentos da tarefa e do domínio em uma abordagem mais uniforme e consistente, tendo por base o uso de ontologias.
- **Ontologias de aplicação:** são as ontologias mais específicas. Os conceitos de uma ontologia de aplicação devem ser especializações dos termos das ontologias de domínio e de tarefa correspondentes.

Guarino em [GUA 98] fornece também outra distinção em relação a ontologias. Para ele, uma ontologia pode ser “refinada” ou “não refinada”. Ontologias não refinadas têm um número mínimo de axiomas e o objetivo de serem compartilhadas por usuários que concordem sobre uma determinada visão do mundo. Uma ontologia refinada precisa de uma linguagem de alta expressividade e tem um grande número de axiomas. Ontologias não refinadas têm mais chance de serem compartilhadas e deveriam ser usadas *on-line* para dar suporte à funcionalidade de sistemas de informação. Já as ontologias refinadas deveriam ser usadas *off-line* e somente para referência.

2.3 Aplicações de ontologias

Como dito anteriormente, muitas pesquisas vêm sendo desenvolvidas em torno de ontologias em diversas áreas, tais como integração de informação inteligente, sistemas de informação cooperativa, recuperação de informação, comércio eletrônico e gerenciamento do conhecimento. Para Fensel [FEN03], a razão pela qual ontologias se tornaram populares é, em grande parte, o que elas prometem: entendimento comum e compartilhado de um mesmo domínio que pode ser comunicado entre pessoas e sistemas de aplicação. Segundo Noy e McGuinnes em [NOY01], uma ontologia define um vocabulário comum para pesquisadores que precisam compartilhar informação em um domínio.

Seguem algumas visões de outros autores sobre a utilidade de ontologias.

- Ontologias habilitam o compartilhamento de conhecimento e a análise ontológica clarifica a estrutura do conhecimento [CHA99].
- Ontologias são essenciais para o desenvolvimento e uso de sistemas inteligentes bem como para interoperabilidade de sistemas heterogêneos. Ontologias são úteis de muitas maneiras para o entendimento e interação humana [FAR97].
- Uma razão típica para a construção de ontologias é oferecer uma linguagem comum para compartilhamento e reuso de informações sobre um fenômeno de interesse no mundo [HOL02].

Maedche [MAE02] aponta a *Web* semântica, a compreensão da linguagem natural, a gestão do conhecimento e o comércio eletrônico como as principais áreas de aplicação de ontologias. Cita ainda outras áreas de aplicação, como modelagem de processos de negócios, bibliotecas digitais e agentes inteligentes, entre outras. A seguir são tecidos comentários a respeito de áreas de aplicação mencionadas.

2.3.1 Web Semântica

A *Web* trouxe novas possibilidades de acesso à informação e às aplicações em geral. Porém essa facilidade tem levado a um crescimento exponencial do material disponibilizado, gerando alguns problemas para os usuários. Um problema encontrado hoje está na busca de informação a qual, na maioria das vezes, ocorre através do uso de palavras-chave, ou seja, uma busca por comparação lexical.

Devido à grande quantidade de informação disponível e à falta de significado desta, as buscas na *Web* retornam hoje uma grande quantidade de informações, em diversos contextos irrelevantes para os usuários. O uso de semântica na *Web* tem sido visto, então, como um fator fundamental para encontrar uma saída para os problemas causados por essa expansão.

> *"o desenvolvimento da World Wide Web está a ponto de amadurecer de uma plataforma técnica que permite o transporte de informação de fontes da Web para humanos (embora em muitos formatos) para uma plataforma que permite a comunicação do conhecimento de fontes Web para máquinas" (Berners-Lee et al., 2001) apud [MAE02].*

Segundo Grüninger [GRU02], o criador da *Web*, Tim Berners-Lee, considera o uso de ontologias como uma parte crítica da *Web* semântica (WS), que necessita ontologias formais para definir o significado dos dados. Segundo Maedche [MAE02], já existem vários exemplos de protótipos de aplicações fazendo uso de ontologias nesta área.

2.3.2 Comércio eletrônico

A área de comércio eletrônico vem crescendo muito ultimamente. Porém, segundo Fensel [FEN03], de forma geral, a automatização de transações nesta área não cumpriu as expectativas anunciadas. Essa automatização requer descrições formais de produtos, além de uma sintaxe para formatos de troca. Assim, um entendimento comum dos termos e suas interpretações deve ser capturado na forma de ontologias, permitindo interoperabilidade e meios para integração de informação inteligente [MAE02].

Outro problema no comércio eletrônico é o fato de muitas informações sobre produtos estarem disponibilizadas somente em linguagem natural e, portanto, serem entendidas somente por humanos. Nesse caso, ontologias poderiam ser usadas para descrever os produtos, facilitando a navegação e recuperação automática de informações.

2.3.3 Gestão do conhecimento

A área de gestão do conhecimento trata da aquisição, manutenção e acesso ao conhecimento de uma organização [MAE02]. No que se refere à gestão do conhecimento, a tecnologia da *Web* semântica, que inclui as ontologias, permitirá definições semântica e estrutural de documentos fornecendo possibilidades completamente novas [FEN03]:

- Busca inteligente ao invés de busca por palavra-chave;
- Resposta a consultas ao invés de recuperar documentos;

- Troca de documentos via mapeamento de ontologia;
- Definição de visões personalizadas de documentos.

Como exemplo de aplicação do uso de ontologias na gestão do conhecimento, pode-se citar o projeto On-To-Knowledge[1], que constrói um ambiente para gerenciamento do conhecimento em grandes *intranets* e *web sites* e usa ontologia para modelar as semânticas de fontes de informação de maneira processável por máquina.

2.3.4 Processamento de linguagem natural

Compreensão de linguagem natural requer uma integração de muitas fontes de conhecimento. Sendo assim, o conhecimento do domínio, na forma de ontologias, é essencial para entender profundamente os textos [MAE02]. A extração de informação também é uma aplicação da área de processamento da linguagem natural que pode usar a noção de ontologia para preencher modelos com instâncias, e assim retornar as informações solicitadas.

2.3.5 Outras aplicações

Ontologias também têm sido aplicadas na área de Engenharia de Software. Falbo *et al*. em [FAL00], apresentam uma ontologia de qualidade de software visando apoiar o entendimento sobre o domínio em questão. Esta ontologia foi formalizada em lógica de primeira ordem e durante sua construção ocorreu integração com uma ontologia de processo de software.

Já em [FAL02a], Falbo *et al*. apresentam ODE (*Ontology-based Software Development Environment*), um Ambiente de Desenvolvimento de Software (ADS) baseado em ontologias. Segundo tais autores, por ser baseado em ontologias, este ambiente possui algumas vantagens como a criação de um repositório de conhecimento que proporciona ao ambiente uma uniformidade de conceitos, primordial na integração de ferramentas.

Em relação a reuso de software, Falbo *et al*. [FAL02b] também sugerem o uso de ontologias. Porém, a falta de abordagens para inserir ontologias em um processo de desenvolvimento de software mais convencional é uma das principais desvantagens, quando

[1] http://www.ontoknowledge.org

se busca um uso mais amplo de ontologias na Engenharia de Software [FAL02b]. Os autores apresentam o ambiente ODE como uma abordagem ontológica para o domínio da engenharia de software com o objetivo de unir ontologias com a tecnologia de orientação a objetos. Essa abordagem está baseada em duas fases: construção de ontologias, seguida da derivação de *frameworks* de objetos a partir dessas ontologias.

Já Edward Hovy em [HOV03] apresenta o uso de ontologias com o objetivo de simplificar acesso a dados em bases do governo dos EUA. Segundo o autor, devido à divisão em muitas esferas (federal, estadual e local; executivo, judicial e legislativo; etc.), os dados do governo acabam sendo coletados por diferentes pessoas, em tempos e locais diferentes. Muitos dos dados coletados envolvem outros dados ou então são complementares, e a heterogeneidade resultante desta coleta acaba criando incompatibilidade no compartilhamento dos dados. O que se buscava então era um mecanismo para padronizar os tipos de dados de forma a permitir compartilhamento. Para tal, foi definida uma ontologia simplesmente como uma taxonomia de termos, abrangendo desde termos mais gerais, no topo, até termos mais específicos, na base.

Outro trabalho interessante é o que John Everett *et al*. descrevem em [EVE02]. Os autores apresentam um sistema de compartilhamento de conhecimento da Xerox, denominado Eureka, que contém aproximadamente 40.000 documentos textuais. Seu objetivo é construir um sistema que possa identificar documentos conceitualmente similares, de forma a permitir fazer manutenção em várias coleções de documentos. Para isso, os autores estão utilizando ontologias, que devem suportar a normalização de diferentes representações de conteúdo similar, para permitir a detecção de similaridades. No projeto das ontologias, desenvolveram critérios que suportam comparações de textos em linguagem natural. Segundo os autores, foi construído um protótipo do sistema, mas que está muito longe, ainda, de estar completo.

Já em [KAL02], Kalfoglou *et al*. apresentam uma forma de aplicar ontologias em Memória Organizacional (MO). O objetivo de uma MO é prover acesso fácil e recuperação de informações relevantes aos seus usuários. Um dos problemas encontrados quando uma MO está sendo desenvolvida diz respeito a sua população inicial. Outro problema encontrado é a identificação de Comunidades de Prática (CoPs) dentro de uma organização. Essas comunidades criam e compartilham conhecimento, e armazená-lo é muito importante para a memória da organização. Para resolver esses problemas, os autores propuseram o uso de ontologias, aplicando um método denominado Análise de Rede de Ontologia.

Assim, os autores criaram uma instanciação particular desse método, denominada Ontocopi, para tentar identificar CoPs dentro das organizações. Ontocopi é implementada como um *plugin* para a ferramenta Protégé[2] e também está disponível via *Web*. Os autores aplicaram um algoritmo de ativação de propagação, para identificar quais objetos são mais importantes em uma ontologia de modo a usá-los para povoar inicialmente a MO. Para operacionalizar a Ontocopi, assumiram que ontologias populadas existiam na organização.

2.4 Considerações

Este capítulo apresentou um embasamento teórico sobre o que são ontologias, suas diferentes classificações e sua aplicação em diferentes áreas, entre as quais engenharia do conhecimento, recuperação de informação e comércio eletrônico.

A partir da pesquisa realizada constatou-se que o tema ‘ontologia’ tem sido abordado em várias pesquisas, em diversas áreas, confirmando sua importância no contexto tecnológico atual. O principal objetivo nestas pesquisas tem sido fornecer um entendimento comum e compartilhado sobre um determinado domínio.

Porém, apesar da importância e da grande quantidade de pesquisas realizadas sobre o tema, a construção de ontologias não conta com metodologias e modelos maduros e bem aceitos para este processo. No próximo capítulo apresentamos algumas abordagens encontradas na literatura para a construção de ontologias a partir de textos, que serviram de base para a abordagem que está sendo proposta neste trabalho.

[2] http://protégé.stanford.edu

3 TRABALHOS RELACIONADOS

Este capítulo descreve os principais trabalhos relacionados à nossa pesquisa, principalmente no que se refere a construção de ontologias a partir de textos.

3.1 Ontologias e texto

Para várias áreas de aplicação, a construção automática de ontologias a partir de um determinado conjunto de textos seria extremamente útil. Porém, segundo Davies *et al.* [DAV02] existem alguns problemas, não triviais, relacionados à geração de ontologias a partir de textos, alguns deles já conhecidos do campo da representação do conhecimento. Dentre estes problemas está a identificação de qual conhecimento deve ser representado, neste caso, em nível da ontologia.

Um problema adicional está em determinar uma linha divisória entre o que é expresso explicitamente em um texto e o que é assumido implicitamente. Por exemplo, quando um texto está sendo escrito, seu autor assume várias questões como sabidas, isto é, ele assume que o leitor compartilha do mesmo, ou quase o mesmo, conhecimento, deixando assim de explicitar conhecimento. Conseqüentemente, é muito difícil construir um processo computacional que capturará o que em essência não está presente no texto.

Segundo Brewster *et al.* em [BRE03], alguns aspectos nos textos podem afetar uma ontologia de domínio. Estes aspectos são:

- Um texto pode reforçar as suposições e estruturas de conhecimento de uma ontologia através da aproximação de conceitos;
- Um texto pode alterar as ligações, associações e instanciações de conceitos existentes. Este tipo de atividade pode ser visto como uma tentativa de re-estruturar a ontologia de domínio.
- A maneira mais óbvia como um texto afeta uma ontologia de domínio é adicionando conceitos novos.

Ainda em [BRE03], segundo os autores, a experiência tem mostrado que um certo número de contextos textuais é necessário para que o conhecimento ontológico esteja

explicitamente disponível. Assim, haverá sempre uma quantidade de termos de baixa freqüência para os quais é difícil achar contextos suficientes ou apropriados dentro do corpus. Nesses casos, poderiam ser procuradas fontes textuais externas para superar a ausência de comunicação explícita de conhecimento no corpus específico do domínio.

O problema encontrado nesse caso é identificar a fonte externa correta e apropriada para um determinado conjunto de textos. Existem várias fontes potenciais de conhecimento ontológico [BRE03]:

- Enciclopédias: parecem fontes ideais para o conhecimento ontológico, pois incluem definição e textos explicativos que poderiam ser explorados. Seu principal problema é o fato de que elas provavelmente não sejam muito atualizadas.
- Livros didáticos e manuais: associados ao domínio, têm utilidade potencial. Aqui o problema principal é identificar os textos relevantes e obtê-los eletronicamente.
- Internet: esta é a fonte mais óbvia:
  - Vantagens: devido ao seu tamanho, a informação desejada provavelmente será encontrada; devido ao seu dinamismo, é provável que conceitos novos estejam disponíveis prontamente; ela é facilmente acessada.
  - Desvantagens: o conceito de determinado termo pode aparecer em diversos textos de domínios diferentes; a Internet tende a repetir a mesma informação em muitos lugares porque as pessoas freqüentemente copiam informações umas das outras; é difícil determinar critérios para decidir se um *web site* será confiável ou não.

A seguir apresentamos abordagens encontradas na literatura que estão relacionadas a nossa proposta de construção de estruturas ontológicas a partir de textos.

3.2 A abordagem de Buitelaar *et al.*

Em [BUI04], é apresentada uma abordagem para extração ou extensão (também denominada enriquecimento) de ontologias a partir de documentos textuais. Segundo Buitelaar *et al.*, esta abordagem segue os passos típicos de aprendizado de ontologia, porém objetivando integrar mais diretamente a engenharia de ontologia com análise lingüística, através da definição de regras de mapeamento, que relacionam entidades lingüísticas, em

coleções de texto anotadas, a conceitos e atributos. A seguir provemos uma descrição em alto nível dos passos desta abordagem, que é implementada como um *plugin*, denominado OntoLT[3], para a ferramenta de desenvolvimento de ontologia Protégé.

3.2.1 Anotação lingüística do corpus

A primeira parte do processo de extração consiste em realizar anotação lingüística nos textos do corpus. Este passo é realizado por um sistema baseado em regras para análise do alemão e do inglês denominado Schug[4] que, segundo os autores [BUI04], provê as seguintes informações: *part-of-speech* (categoria gramatical), informação morfológica **(**flexão, derivação ou composição de uma palavra), estrutura sintática da frase e da sentença.

3.2.2 Regras de mapeamento

A próxima etapa desta abordagem consiste em definir regras de mapeamento entre a estrutura lingüística e o conhecimento ontológico. Previamente são providas algumas regras de mapeamento, mas o usuário tem a liberdade para criar novas regras se achar necessário. A seguir apresentamos dois exemplos de regras de mapeamento previamente definidos:

- HeadNounToClass_ModToSubClass: mapeia o substantivo principal para uma classe (conceito) e, em combinação com seus modificadores, para uma ou mais sub-classes.

- SubjToClass_PredToSlot: mapeia um sujeito para uma classe (conceito), e seu predicado para um atributo (*slot*) dessa classe.

A idéia é executar as regras de mapeamento coletivamente e, à medida que as pré-condições sejam satisfeitas, gerar os conceitos e atributos para uma nova ontologia ou integrá-los em uma ontologia existente, sempre de forma automática. Deve-se observar que as regras de mapeamento somente serão executadas para aquela informação considerada relevante no pré-processamento estatístico (vide próxima sub-seção). Por fim, os conceitos e atributos extraídos são validados pelo usuário.

---

[3] http://olp.dfki.de/OntoLT/OntoLT.htm

[4] Maiores informações sobre o sistema Schug podem ser encontradas em [DEC02].

3.2.3 Pré-processamento estatístico

Este passo serve para filtrar, a partir da informação lingüística extraída, aquela relevante ao domínio. Para realizar essa tarefa a abordagem baseia-se em uma função denominada qui-quadrado[5], a partir da qual é determinada a relevância da informação ao domínio através da comparação da sua freqüência no corpus do domínio com sua freqüência em um corpus de referência.

3.2.4 Geração semi-automática de regras de mapeamento

A partir do passo anterior é possível a geração semi-automática de regras de mapeamento, as quais simplesmente poderiam ser geradas para todos os possíveis elementos da anotação lingüística, porém limitados às palavras que foram selecionadas pela medida qui-quadrado.

3.2.5 Considerações

O uso de mapeamentos entre a estrutura lingüística e o conhecimento ontológico, similar ao proposto por Velardi e co-autores em [VEL01], é uma abordagem que poderia ser utilizada, principalmente, no que diz respeito à construção de ontologias de domínio, onde algumas informações importantes precisam de regras específicas para ser extraídas. Porém, é nosso objetivo focar em regras mais genéricas, que possibilitem a geração de estruturas ontológicas para diferentes domínios.

Quanto ao processamento estatístico, nosso objetivo inicial era trabalhar somente com o corpus do domínio para o qual a estrutura ontológica deveria ser gerada. Sendo assim, o processamento estatístico utilizado em [BUI04] não poderia ser utilizado em nossa proposta, pois a medida qui-quadrado precisa de um corpus de referência para determinar qual informação é relevante ao domínio. Assim, utilizaríamos a medida TFIDF (subseção 3.5.2.2) para determinar os termos relevantes do domínio. Porém, a medida TFIDF retorna apenas uma classificação dos termos, não ocorrendo nenhuma poda e nenhum valor ou regra para poda são sugeridos. Isso acaba resultando em uma quantidade muito grande de termos não relevantes sendo apresentados ao engenheiro de ontologia.

---

[5] Maiores informações sobre a função qui-quadrado podem ser encontradas em [AGI01].

Nesse caso, a medida qui-quadrado poderia ser utilizada para podar termos através da sua comparação em um corpus de referência. Uma medida alternativa a medida qui-quadrado é a medida *Log-Likelihood*. De acordo com Rayson *et al.* [RAY04], a medida *Log-Likelihood* é muito semelhante à medida qui-quadrado, com uma ligeira melhora nos resultados para algumas situações.

3.2.5.1 *A medida Log-Likelihood*

A medida *Log-Likelihood* calcula a relevância de um termo do corpus do domínio com base na sua freqüência no corpus do domínio e no corpus de referência. É possível calcular a medida *Log-Likelihood* (G2) com a seguinte fórmula:

$$G2 = 2*((a*\ln (a/E1)) + (b*\ln (b/E2)))$$

Onde,

- E1 = c*(a+ b) / (c+ d);
- E2 = d*(a+ b) / (c+ d);
- a: é a freqüência da palavra observada no corpus de referência;
- b: é a freqüência da palavra observada no corpus do domínio;
- c: corresponde ao número de palavras no corpus de referência;
- d: corresponde ao número de palavras no corpus do domínio.

Quanto mais alto o valor de G2, mais significante é a diferença entre duas freqüências. Vale observar que o valor de G2 será sempre um número positivo. Para definir se o termo no corpus de referência é mais significante que no corpus do domínio, ou vice-versa, calcula-se como segue:

- Se a*(ln(a/E1) > = 0 então o termo é mais relevante no corpus de referência;
- Se a*(ln(a/E1) < 0 então o termo é mais relevante no corpus do domínio.

A medida *Log-Likelihood* será utilizada em nossa abordagem para poda de termos na fase de identificação de termos relevantes.

3.3 A abordagem de Degeratu e Hatzivassiloglou

Segundo os autores [DEG04], algumas pesquisas desenvolvidas no intuito de semi-automatizar a construção de ontologias a partir de textos ou, até mesmo, automatizar completamente algumas tarefas, estão apoiadas no uso de ontologias já existentes, *thesauri* e dicionários *on-line*, entre outros. Diferentemente, os autores propõem um método para

construir ontologias a partir de textos escritos em linguagem natural, não anotados. As fases deste método, que são descritas a seguir, foram implementadas em uma ferramenta denominada Ontostruct.

3.3.1 Pré-processamento

A primeira parte do pré-processamento é transformar diferentes formatos de texto, por exemplo, PDF ou HTML, em um formato único. Porém, segundo os autores, nesta transformação existe perda significativa de informação de formatação. Os marcadores de lista, por exemplo, assim como parágrafos aninhados, provêem valiosa informação sobre a divisão e estrutura do texto e, a partir deles, pode-se descobrir se dois termos compartilham uma mesma estrutura de lista, e isso pode servir como um indicador de relação entre esses termos. Para resolver este problema, os autores usam uma ferramenta denominada MXTERMINATOR[6] e um algoritmo que procura por seqüências crescentes de contadores (ex.: 1, 2, 3, ... ou a, b, c, ...) e aprende padrões que extraem estes tipos de seqüências.

Depois, são usados um etiquetador de categorias gramaticais e um *chunk parser*[7] para etiquetar o conteúdo, e identificar sintagmas nominais e sintagmas verbais nas orações.

3.3.2 Identificação de termos

A partir dos sintagmas nominais, identificados no pré-processamento, são extraídos os candidatos a termos. Com o objetivo de obter somente a forma singular dos substantivos e seus modificadores adjetivais (variantes), são aplicados filtros sintáticos e morfológicos a cada sintagma nominal.

Os autores optam por considerar essas variantes como termos candidatos, não as eliminando quando encontradas. Por exemplo, em um texto, o termo "*profit*" (lucro) pode vir acompanhado de adjetivos, como em "*domestic profit*" (lucro nacional) ou "*foreign profit*" (lucro estrangeiro). Estas variantes para o termo "*profit*" são então aceitas como termos

[6] Maiores informações podem ser encontradas em [REY97].

[7] Em resumo, um *chunk parser* é um tipo de *parser* que identifica grupos lingüísticos (como sintagmas nominais) em texto irrestrito, tipicamente uma sentença de cada vez.

candidatos desde que possuam um valor de informação mútua[8] positivo entre o adjetivo e o substantivo principal.

Por fim, dentre os termos candidatos que ficam, são selecionados todos aqueles que aparecem pelo menos duas vezes em um dos documentos da coleção, ou aparecem em múltiplos documentos.

3.3.3 Extração de relações

Nesta fase, para derivar relações conceituais entre os termos extraídos do corpus, primeiramente são analisados vários padrões sintáticos que acontecem no texto. Segundo os autores, esses padrões são usados para recuperar relações *is-a*, equivalências, atributos gerais de termos (inclusive relações *has-a* (tem um) ou *part-of* (parte de)), e outras relações gerais entre termos. A seguir brevemente descrevemos os tipos de relações extraídas no processo.

Relações hierárquicas: são extraídas do texto usando padrões léxico-sintáticos pré-definidos por Hearst em [HEA92]. Exemplo de um padrão: <*TERM*> *is a* <*HEADTERM*> (<Termo1> é um <Termo principal>). Como exemplo, considere o seguinte trecho de texto: "*... the dog is a animal...*". A partir do padrão em questão seria extraída uma relação hierárquica definindo que *dog* (cão) é um tipo de *animal* (animal).

Relações de equivalência: pelo uso de padrões parentéticos[9] podem ser extraídas relações de equivalência entre dois termos, ou um termo e seu acrônimo como, por exemplo, "*standard industrial code* (SIC)". Relações de equivalência ainda podem ser extraídas pela seguinte regra: <*TERM-1*> *(also|formerly) called* <*TERM-2*> (<Termo-1> (também|anteriormente) chamado <Termo-2>). A partir de tal regra pode-se, por exemplo, identificar a equivalência entre os termos "beija-flor" e "colibri" no trecho de texto "O beija-flor também chamado de colibri...".

---

[8] De acordo com Church em [CHU89], informação mútua é uma medida que compara a probabilidade de se observar dois pontos (palavras), x e y, juntos, com a probabilidade de se observar x e y independentemente.

[9] Do inglês *parenthetical*. Expresso entre parênteses.

Atributos de termos e propriedades: podem ser extraídos usando informação contextual. A regra <*ATTRIBUTE*> *of* <*TERM*> (<Atributo> de <termo>) permite identificar atributos como em "*name of person*" (nome de pessoa), onde se identifica que "nome" é um atributo de "pessoa".

3.3.4 Agrupamento de termos

O objetivo desta fase é colocar termos de significado equivalente, ou quase equivalente, em um mesmo grupo. Isso é feito com base na similaridade entre as relações das quais os termos participam.

Primeiramente são produzidas três listas para um dado termo: *attribute of (x)*, correspondendo aos atributos associados ao termo; *verb with subject(x)*, correspondendo aos verbos onde este termo aparece como sujeito e *verb with object(x)*, correspondendo aos verbos onde este termo aparece como objeto. Então, para dois termos, é medida a desigualdade entre cada par de listas correspondentes através do coeficiente de Lance e Williams (Lance e Williams, 1967 *apud* [DEG04]).

O resultado final, calculado entre as três listas, corresponde ao valor de desigualdade lexical média entre os dois termos**.** Tal valor sofre ainda interferência pela ocorrência ou não destes dois termos juntos em um documento. No final, os termos com valor de desigualdade mínimo são então agrupados.

3.3.5 Criando a hierarquia

As classes de termos semanticamente agrupados e as relações de equivalência e relações hierárquicas aprendidas são parte da entrada para o algoritmo de construção de hierarquia, que gerará como resultado um conjunto de hierarquias formando um grafo acíclico [DEG04].

No experimento relatado em [DEG04], foram usados documentos como formulários e suas correspondentes instruções de preenchimento, desenvolvidos sem nenhuma estrutura e formatação, com grande variedade no seu conteúdo. A partir de informações do domínio e definições para os diferentes tipos de relações fornecidas, avaliadores humanos julgaram a precisão. Como resultado, a ferramenta alcançou a precisão de 71-83% na extração de relações e obteve agrupamentos corretos de termos em 54% das vezes [DEG04].

3.3.6 Considerações

Computar uma forma padrão para os termos (como o lema da palavra) é um aspecto muito importante na fase de identificação de termos, pois em um corpus um conceito normalmente aparece com diferentes formas (com variações de gênero, número e grau). Em um texto de esportes, por exemplo, o conceito "goleiro" pode aparecer no singular (goleiro) ou no plural (goleiros). Computando-se o lema das palavras obtemos apenas um conceito (goleiro), o que faz mais sentido, pois este é o conceito envolvido nas duas situações. Dada a importância desta abordagem, ela será utilizada em nossa proposta para identificação de termos. Além disso, por se tratar da construção de uma ontologia de domínio, é interessante, como em [DEG04], não eliminar as variantes de um termo (substantivo + modificadores adjetivais), como "*profit*", "*domestic profit*" e "*foreign profit*". Tais variantes podem representar conceitos mais específicos dentro de um domínio e, portanto não serão eliminadas. A diferença em nossa proposta é que adotaremos todas as variantes dos termos como possíveis candidatas, deixando que sua permanência seja decidida pelo engenheiro de ontologia, enquanto que em [DEG04] é utilizada a medida de informação mútua para determinar se a variante deve ser considerada ou não. Em relação à freqüência dos termos, em [DEG04] são selecionados todos aqueles termos que aparecem ao menos duas vezes em um dos documentos da coleção, ou que aparecem em múltiplos documentos. Nosso entendimento é que assim podem ser selecionados muitos termos, sendo que grande quantidade não relevante ao domínio. Nossa estratégia, como descrito anteriormente, será a de utilizar uma medida para filtrar parte dos candidatos a termos e deixar ao engenheiro de ontologia a tarefa de definir, conforme sua percepção, quais termos são ou não relevantes.

3.4 A abordagem de Lame

O método apresentado por Lame em [LAM03] tem por objetivo a identificação de conceitos e relações semânticas entre esses conceitos. Segundo o autor, as técnicas utilizadas são automáticas e contam com ferramentas tais como analisadores sintáticos e estatísticos, porém não substituem os projetistas de ontologias, apenas os ajudam no processo.

No contexto apresentado em [LAM03], o método foi utilizado para a construção de uma ontologia referente à lei francesa, a partir de textos de Códigos (leis), e foi aplicada para a construção de uma ontologia dedicada a recuperação de informação. Segundo Lame, o método também pode ser usado para construir ontologias dedicadas a outras tarefas.

De modo geral, a idéia aqui consiste em identificar conceitos, através de termos (palavras ou grupos de palavras) encontrados nos textos, e identificar relações semânticas entre esses conceitos, através da busca de relações sintáticas entre os termos. Para atingir esses objetivos, o método conta com as etapas descritas a seguir.

3.4.1 Análise sintática

Consiste na análise de textos com o objetivo de identificar a categoria gramatical das palavras (substantivos, verbos, etc.). Para realizar esta etapa o autor utiliza uma ferramenta denominada Syntex[10], obtendo como resultado uma lista de termos etiquetados com as respectivas categorias gramaticais. A ferramenta inclui um conjunto de regras sintáticas e uma lista de "*stopwords*[11]" e estabelece também dependências sintáticas entre termos [LAM03].

No contexto onde foi aplicado este método, o autor optou por trabalhar somente com substantivos pois, segundo ele, a maioria dos conceitos está rotulada por substantivos. Além disso, são excluídos os termos que contêm caracteres não-alfabéticos, como números e símbolos. Dessa forma, ao final da análise estatística obtém-se como resultado uma lista de termos relevantes (substantivos) que são considerados nas próximas etapas.

3.4.2 Análise das relações de coordenação

A análise das relações de coordenação tem como objetivo identificar termos separados por conjunções "*e"* ou "*ou"* como, por exemplo, "nome e sobrenome" ou "carro ou ônibus". A partir da lista de termos, é realizada a análise dos documentos identificando ocorrências desses termos separados com "*e"* ou "*ou"*, pois isso pode indicar uma relação entre eles. O resultado desta etapa, uma lista de possíveis relacionamentos, deve ser manualmente conferido para validar quais relações são semanticamente relevantes e quais não são.

---

[10] Maiores informações sobre a ferramenta Syntex podem ser encontradas em [BOU02].

[11] *Stopwords* são palavras comuns que não agregam valor por possuírem significado semântico limitado; palavras cuja finalidade é auxiliar a estruturação da linguagem. Ex.: preposições, artigos, alguns pronomes e conjunções. Estas palavras pdem ser eliminadas, numa aplicação tal como a estudada.

3.4.3 Análise estatística

Tem como objetivo identificar relações entre os termos relevantes identificados na análise sintática. De acordo com [LAM03], dois termos semanticamente relacionados ocorrem com freqüência em contextos similares. Esses contextos estão relacionados com as palavras que cercam um termo. Por exemplo, o termo “nacionalidade” sempre aparece no mesmo contexto que “registro”, “data de nascimento”, etc. Cada palavra do contexto recebe uma medida de informação mútua que quantifica sua dependência em relação a um determinado termo. Conforme Lame, deve-se definir um limiar mínimo com o objetivo de validar as relações obtidas. Uma validação manual desses resultados também poderia ser realizada.

3.4.4 *Pattern-matching*

Muitas vezes um termo está dentro de outro termo como, por exemplo, o termo “contrato” está em “contrato de depósito” ou em “contrato de locação”. Sendo assim, esta etapa consiste em relacionar tais termos (no exemplo, “contrato”), com os termos dos quais fazem parte (neste caso, “contrato de depósito” e “contrato de locação”). Segundo Lame, esse método é grosseiro mas, se aplicado a uma lista de termos bem identificados, pode dar bons resultados, especialmente no contexto de uma ontologia dedicada à recuperação de informação. O resultado desta etapa será um conjunto de relações hierárquicas entre os termos identificados na análise sintática.

3.4.5 Considerações

Lame, em suas pesquisas, optou por trabalhar somente com substantivos que, segundo ele, representam a maioria dos conceitos. Como dito anteriormente, iremos trabalhar com os substantivos e também com seus modificadores. Lame utiliza uma lista de *stopwords* para filtrar palavras que não representam conceitos do domínio e também exclui palavras que contêm caracteres não-alfabéticos como números e outros símbolos. A utilização dessas estratégias nos permitirá excluir palavras que não acrescentam valor para a construção da estrutura ontológica.

Ao observar o termo “contrato” como parte do termo “contrato de depósito”, identificamos que “contrato de depósito” é um tipo de “contrato” e, portanto, os termos estão relacionados hierarquicamente (relação taxonômica). A identificação de termos que fazem

parte de outros, como proposto por Lame, auxilia na identificação de relações taxonômicas e também será utilizada em nossa proposta.

3.5 A abordagem de Maedche

Em [MAE02], Maedche e co-autores apresentam um método semi-automático para obter uma ontologia de domínio baseada em fontes de texto de uma *intranet*. As etapas desse método, que são suportadas por uma ferramenta denominada Text-To-Onto [MAE02], são descritas a seguir. De acordo com os autores, a ontologia de domínio resultante deste processo pode, se necessário, ser refinada e melhorada através da repetição do processo de aquisição [MAE02]. Para extração de informação, é usado um processador superficial de texto para o idioma alemão denominado SMES[12], o qual é composto por um *tokenizador*, um analisador lexical e um *chunk parser*. O resultado deste processamento são textos do domínio lingüisticamente anotados.

3.5.1 Escolha de uma ontologia base

Esta etapa consiste em escolher uma ontologia genérica (ou ainda redes léxico-semânticas ou uma ontologia relacionada ao domínio) para ser usada como uma estrutura base para a ontologia específica que se deseja construir. De acordo com os autores, os algoritmos trabalham sem a necessidade de qualquer estrutura de conhecimento; entretanto, se algum tipo de conhecimento estiver disponível, ele será utilizado como base para o restante do processo.

3.5.2 Aquisição de conceitos

O objetivo desta etapa é extrair conceitos específicos do domínio. Os autores trabalham com duas abordagens nesta etapa. Em uma delas utilizam um dicionário contendo termos importantes do domínio e a outra é baseada em medida de freqüência.

[12] Maiores informações sobre o processador de textos SMES podem ser encontradas em [MAE02].

3.5.2.1 *Aquisição de conceitos via dicionário de domínio*

O primeiro passo consiste em converter em conceitos as *headwords*[13] do dicionário. Quando duas entradas possuírem uma mesma definição, elas são unidas e as *headwords* são consideradas sinônimos e, quando uma entrada contiver referência à outra entrada, então esta ligação é convertida em uma relação conceitual [MAE02].

A partir dos conceitos criados são então extraídos *stems*[14]. Se um *stem* já existe na ontologia, é necessário verificar se a entrada extraída do dicionário descreve de forma diferente o mesmo conceito contido na ontologia e, caso isso aconteça, tal fato é classificado como conflito [MAE02]. Para tentar resolver esse tipo de conflito de forma automática, os autores propõem o uso de algumas heurísticas. Caso o conflito não seja resolvido de forma automática é então solicitada intervenção do usuário.

3.5.2.2 *Aquisição de conceitos baseada em freqüência*

Consiste em obter a freqüência das entradas lexicais em um conjunto de documentos do domínio com o objetivo de indicar possíveis conceitos deste domínio. A técnica está baseada na idéia de que uma entrada lexical freqüente pode sugerir um conceito. Porém, segundo [MAE02], existem meios mais efetivos para se obter a classificação dessas entradas além da simples contagem de freqüência.

Dessa forma, os autores usam então uma combinação de medidas denominada *tfidf* (*term frequency inverse document frequency*), a qual pesa a freqüência de uma entrada lexical em um documento.

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

onde:

- $lef_{l,d:}$ refere-se à freqüência da entrada lexical $l \in L$ em um documento $d \in D$.
- df_l: é o número de documentos do corpus D em que l ocorre.

---

[13] Uma *headword*, neste caso, é uma palavra-chave do dicionário, isto é, uma palavra que representa uma entrada específica do dicionário.

[14] *Stem*, no contexto desta dissertação, é o mesmo que radical de uma palavra.

Depois de calculada a freqüência de cada entrada lexical em cada documento, através da medida mencionada anteriormente, tem-se uma lista de todas as entradas lexicais. As *stopwords* são retiradas desta lista.

De acordo com os autores em [MAE02], para calcular o valor *tfidf* para uma determinada entrada lexical no corpus, basta fazer um somatório das freqüências desta entrada em cada documento do corpus como segue:

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

Para auxiliar na extração de entradas lexicais relevantes ao domínio, um limiar mínimo pode ser definido e para a entrada ser considerada, o valor $tfidf_l$ deve alcançar este limiar.

A determinação da freqüência também pode ser feita com a utilização de um segundo corpus contendo documentos genéricos. Assim, as freqüências dos conceitos em ambos os corpora são comparadas. Aqueles conceitos que tiverem maior freqüência no corpus específico do domínio permanecem como relevantes, enquanto os demais são removidos [MAE02]. Segundo os autores, é dada liberdade ao usuário para intervir no processo de modo, por exemplo, a excluir conceitos que permaneceram ou incluir conceitos não selecionados.

3.5.3 Aquisição de taxonomia

Os autores utilizam duas técnicas para a extração de taxonomias (ou hierarquia de conceitos), sendo uma a partir de textos lingüisticamente anotados, usando agrupamento hierárquico, e outra a partir de definições de dicionários pré-processados lingüisticamente, usando padrões.

3.5.3.1 *Técnica de agrupamento hierárquico*

Fazer um agrupamento significa formar grupos de objetos semelhantes em algum ponto. Existem duas formas de se criar um agrupamento hierárquico: *bottom-up* e *top-down*.

O algoritmo *bottom-up* adotado pelos autores inicia com cada objeto formando um grupo e, a cada iteração, os grupos mais similares são determinados e agrupados em um novo grupo, terminando quando somente um grande grupo existir [MAE02].

Já o algoritmo *top-down* adotado pelos autores inicia com um grupo contendo todos os objetos, a partir do qual, a cada iteração, são selecionados e divididos em grupos menos coerentes, ou seja, aqueles com objetos menos similares [MAE02].

Os autores trabalham com três estratégias em relação à medida de similaridade para agrupamentos hierárquicos:

- *single linkage*: a similaridade entre dois grupos é dada pela similaridade entre os dois objetos mais similares entre os grupos.
- *complete linkage*: a similaridade entre dois grupos é baseada na similaridade dos dois objetos menos similares.
- *group-average*: a similaridade entre dois grupos é baseada na média de similaridade entre os objetos desses grupos.

3.5.3.2 *Exploração de dicionário baseada em padrões*

O objetivo é usar a informação estruturada contida em dicionários específicos de domínio como entrada para a extração de relações taxonômicas. A idéia é definir uma expressão regular que captura trechos recorrentes e mapear os resultados para uma estrutura semântica como $H^C(C_1,C_2)$. Em seu *framework*, os autores definem várias expressões regulares para adquirir relações taxonômicas. Um problema, porém, é que domínios específicos requerem padrões específicos. Mesmo assim, segundo os autores, várias heurísticas pode ser reusadas. A desvantagem do uso dessa abordagem está na necessidade de definição desses padrões, pois esta é uma atividade que consome tempo [MAE02].

3.5.4 Aquisição de relações conceituais

Para descobrir relações conceituais, os autores utilizam uma abordagem estatística a partir de um algoritmo baseado em regras de associação. A idéia consiste em descobrir quais uniões entre conceitos aparecem freqüentemente nas orações, pois estas podem representar relações relevantes entre conceitos [MAE02].

Por exemplo, o processamento lingüístico poderia indicar que a entrada lexical "*costs*" freqüentemente aparece junto com palavras como "*hotel*", "*guest house*", e "*youth hostel*". Os dados estatísticos indicam então que pode existir uma relação relevante entre "*costs*" e cada um desses conceitos, e esta relação pode ser proposta para inclusão na ontologia. Esses dados

estatísticos são denotados por duas medidas, *support* e *confidence*[15], para as quais um valor mínimo aceitável deve ser especificado. Segundo os autores, o resultado deste passo são sugestões de relações, devendo o usuário manualmente selecionar e nomear as relações desejadas [MAE02].

3.5.5 Considerações

Assim como Lame [LAM03], Maedche [MAE02] também utiliza uma lista de *stopwords* para filtrar palavras que não representam conceitos do domínio e também exclui palavras que incluem caracteres não-alfabéticos como números e símbolos. Em relação à extração de termos relevantes do domínio, Maedche propõe duas abordagens. Uma delas é baseada em um dicionário contendo termos importantes do domínio o que, dada nossa proposta de não utilizar outras fontes além dos textos, não é viável. A segunda abordagem é baseada na combinação de medidas *tfidf* (*term frequency* x *inverted document frequency*), a qual tem por objetivo pesar a freqüência de uma entrada lexical em um conjunto de documentos. Essa medida é adotada em nossa proposta para determinar a ordem de relevância dos termos do domínio. Para ficar somente com os termos relevantes do domínio, um limiar mínimo para esta medida ainda poderia ser definido. Porém, como nossa idéia é possibilitar que nossa abordagem seja utilizada para diferentes domínios e, portanto, muito provavelmente para diferentes tamanhos de corpus, um limiar fixo não será muito adequado. Dessa forma, como dito anteriormente, nossa proposta é deixar a definição de tal limiar para o engenheiro de ontologia.

Para identificar as relações taxonômicas, Maedche também utiliza duas abordagens. Uma abordagem, denominada agrupamento hierárquico, é o processo de organizar objetos em grupos nos quais os membros são similares de alguma forma. A outra abordagem usada é denominada *pattern-matching*, onde a partir da informação sintática, várias heurísticas (padrões) para extrair relações taxonômicas são aplicadas. Maedche usa nesta abordagem os padrões léxico-sintáticos identificados por Hearst [HEA92]. A mesma abordagem e padrões são utilizados também por Degeratu e Hatzivassiloglou em [DEG04]. Os padrões de Hearst inicialmente propostos para textos escritos na língua inglesa, como veremos adiante, foram

[15] Maiores informações sobre as medidas de *support* e *confidence* podem ser encontradas em [MAE00].

adaptados para uso em textos em língua portuguesa e também serão utilizados em nossa abordagem.

3.6 A abordagem de Velardi *et al.*

Em [VEL01], Velardi *et al*. apresentam técnicas de mineração de texto com o intuito de melhorar a produtividade humana durante o processo de construção de ontologia. Tais técnicas foram implementadas em uma ferramenta, denominada OntoLearn[16], com o objetivo de extrair conceitos importantes do domínio e descobrir as relações semânticas entre eles. Para a mineração dos textos, os autores usam um processador de corpus chamado Ariosto[17], o qual teve seu desempenho melhorado com a adição de um reconhecedor de entidades mencionadas[18] e um *chunk parser* chamado Chaos[19] [VEL01].

Segundo os autores, conceitos são denotados por termos que são classificados em três classes:

- Entidades mencionadas do domínio – nesta classe encontram-se, por exemplo, nomes próprios, complexos, como *gulf of Mexico, Texas Country*, etc.
- Termos multi-palavra específicos do domínio – ou seja, termos compostos por mais de uma palavra como *travel agent, preservation área*, etc.
- Palavras singulares específicas do domínio - por exemplo, *hotel, reservation*, etc.

3.6.1 Identificação de instâncias de conceitos

Segundo os autores, identificar instâncias de conceitos significa identificar termos da classe "entidades mencionadas do domínio" como, por exemplo, nomes próprios, os quais representam instâncias de conceitos do domínio. Entidades mencionadas são muito comuns em textos e, de acordo com Velardi *et al.* em [VEL01], na maioria dos domínios, representam mais de 20% do total de palavras de um texto.

---

[16] Maiores informações sobre a ferramenta OntoLearn podem ser encontradas em [VEL01].

[17] Maiores informações sobre o processador de corpus Ariosto podem ser encontradas em [BAS96].

[18] Do inglês *named entities*.

[19] Maiores informações sobre o *chunk parser* Chaos podem ser encontradas em [BAS00].

Através de Ariosto, os termos desta classe são identificados e etiquetados semanticamente conforme uma das três categorias conceituais: lugares, organizações e pessoas. Essa identificação é baseada em regras contextuais anotadas manualmente ou aprendidas por máquina [VEL01].

3.6.2 Identificação de conceitos

O método proposto pelos autores explora propriedades lingüísticas e estatísticas para construir um glossário terminológico específico do domínio. Assim, expressões terminológicas candidatas são capturadas por meio de técnicas executadas em quatro passos através do *chunk parser* Chaos:

- Etiquetagem de categorias gramaticais;
- *Chunking*;
- Comparação de estruturas de argumentos de verbos;
- Análise gramatical superficial[20].

Segundo os autores, o problema de usar abordagens puramente sintáticas é que os termos candidatos são extraídos em uma quantidade muito maior do que as verdadeiras entradas terminológicas. Então, existe a necessidade de filtrar candidatos não-terminológicos e, para tanto, duas novas métricas foram definidas. Essa filtragem é executada pela combinação de ambas:

Relevância de Domínio: um termo candidato é medido pela análise comparativa entre diferentes domínios. Uma definição quantitativa da relevância de domínio pode ser dada de acordo com a quantia de informação capturada dentro do corpus designado, em relação à coleção inteira [VEL01].

Consenso de domínio: mede o uso distribuído de um termo em um domínio, isto é, sua distribuição ao longo de todos os documentos pertencentes ao domínio.

3.6.3 Detecção de hierarquias

O passo anterior retorna uma lista de termos sem nenhuma relação hierárquica. Esta lista então é processada no intuito de identificar relações verticais entre os termos (em outras

[20] Maiores detalhes sobre esta análise gramatical superficial podem ser encontradas em [MAE02].

palavras, as relações hierárquicas). Essas relações hierárquicas formam sub-árvores que facilitam sua união a um nodo apropriado da ontologia.

O método extrai relações taxonômicas a partir da do núcleo do sintagma de termos multi-palavras. Segundo Velardi *et al.*, a estruturação hierárquica dos termos reduz significativamente o trabalho manual, pois somente núcleos de termos devem ser ligados diretamente à ontologia.

Considere, por exemplo, que uma sub-árvore para o termo *card* foi encontrada com vários termos associados (*credit card*, *golden card*, *business card*, etc.). Nesse caso, só *card* deverá ser ligado diretamente a um nodo da ontologia. Essa ligação é feita de modo manual.

3.6.4 Descoberta de relações entre conceitos

De acordo com os autores, relações tais como hiponímia e hiperonímia são difíceis de extrair a partir de um corpus. Sendo assim, os autores sugerem que relações entre Sujeito, Verbo e Objeto sejam mais facilmente detectadas a partir de técnicas de mineração de texto.

3.6.5 Considerações

Um ponto considerado importante em [VEL01], e que também adotaremos, é a extração dos termos multi-palavra (ou termos compostos) como, por exemplo, “pano de prato”. As variantes de termos, identificadas em [DEG04], também são classificadas como termos compostos.

Para identificar relações taxonômicas, Velardi e co-autores usam uma heurística especificamente para termos multi-palavra, baseada no núcleo do sintagma desses termos. Por exemplo: a partir de dois termos, t1 e t2, se t1 é igual a t2, e t1 é modificado por um adjetivo, então é possível derivar a relação taxonômica IS-A (t1, t2). Considerando os termos *credit card* como t1 e *card* como t2, a partir da heurística temos a relação IS-A(*credit card*, *card*). Visto que estaremos identificando termos compostos em nossa abordagem, é interessante utilizarmos também essa heurística para identificar relações taxonômicas. Segundo os autores, mesmo sendo um método grosseiro, pode dar bons resultados se for aplicado sobre uma lista de termos bem definidos.

3.7 Construção de mapas conceituais a partir de textos – o trabalho de Cláudia Pérez

Um trabalho correlato a esta proposta foi desenvolvido por Pérez [23], onde a autora apresenta um processo semi-automático para a extração de conhecimento a partir de textos da língua portuguesa do Brasil, objetivando a construção de Mapas Conceituais. Estes mapas têm o objetivo de representar o conhecimento como armazenado na estrutura cognitiva, ou seja, como um conjunto de conceitos, organizados de forma hierárquica, representando o conhecimento e as experiências adquiridas por uma pessoa. O relacionamento entre os conceitos é formado por palavras de ligação, que são usadas para formar proposições simples (união entre dois ou mais termos conceituais) formando uma unidade semântica [PER04]. A seguir são brevemente descritas as etapas utilizadas pela autora na geração dos Mapas Conceituais. O corpus selecionado para a realização dos experimentos é composto por artigos jornalísticos da Folha de S. Paulo do ano de 1994 e textos didáticos da Editora Scipione do ano de 1983 [PER04].

3.7.1 Análise sintática

De acordo com [PER04], para a análise sintática foi utilizado o parser PALAVRAS [BIC00], que realiza as etapas de *tokenização*, processamento léxico-morfológico, e a análise sintática propriamente dita.

3.7.2 Seleção das estruturas

A identificação das estruturas para a aquisição de conhecimento a partir dos textos está baseada no formato de mapas conceituais, ou seja, são representadas por triplas (conceito – relação – conceito), que em textos da língua natural tendem a aparecer como Sujeito – Verbo – Objeto [PER04]. O verbo tem a função de estabelecer o relacionamento entre dois conceitos. Para a representação das triplas no formato de mapas conceituais foi utilizada a ferramenta Cmap Tools, desenvolvida pelo IHMC- University of West Florida[21].

3.7.3 Cadeias de co-referência

Co-referência é o termo usado para especificar que duas ou mais expressões de um texto se referem a uma mesma entidade do discurso. Depois de os textos estarem anotados,

[21] http://www.ihmc.us/

são usadas folhas de estilo XSL para extrair as cadeias de co-referência, que podem ser usadas para identificar os termos relevantes dos textos, a partir dos quais é possível observar aqueles que são mais referenciados ao longo do texto e também podem ser úteis na identificação de termos equivalentes presentes nos mapas [PER04].

3.7.4 Codificação XTM

A partir do arquivo no formato de triplas Relação(Argumento1, Argumento2), aplicam-se scripts na linguagem *shell* (Linux) para a gerar a codificação XTM.

3.7.5 Resultados

O primeiro experimento da autora foi baseado apenas no núcleo dos sintagmas nominais que exercem a função de sujeito e objeto direto, e o núcleo dos sintagmas verbais, formando a tripla <sujeito – verbo – objeto direto>, representando as relações entre os conceitos [PER04]. Após a análise qualitativa das triplas formadas por essas estruturas, a autora constatou que os mapas conceituais gerados automaticamente não apresentaram relações conceituais tão relevantes como nos mapas construídos manualmente. A autora então observou que a extração poderia melhorar ao levar em consideração pronomes que exercem o papel de sujeito ou objeto, verbos com complementos proposicionais, nomes compostos, e ambigüidade [PER04].

Assim, para o segundo experimento foi utilizado um corpus, com estruturas de seleção mais complexas e foi realizada uma avaliação quantitativa com base em mapas gerados manualmente. Assim, as estruturas que passaram a ser de interesse foram: verbos, sujeito, objeto direto ou indireto, agente da passiva, predicativo do sujeito, adjuntos adverbial e adnominal. A relação das estruturas consideradas para a formação de triplas foi: Argumento1 - Relação – Argumento2.

- **Conceitos**
  a. Argumento 1 (Sujeito): sintagma nominal que exerça a função de sujeito, pronome relativo “que” exercendo a função de sujeito, e verbo no particípio que exerça a função de sujeito.
  b. Argumento 2 (Complemento): núcleo do sintagma nominal que exerce a função de objeto direto e o adjetivo se existir, sintagma preposicional que exerce a função de objeto indireto, predicativo do sujeito, etc.

- **Relação:** núcleo do sintagma verbal e advérbio que antecede um verbo, pois a retirada do mesmo pode produzir um sentido oposto à tripla.

De acordo com a autora, a consideração das novas estruturas possibilitou a formação de mais triplas completas, refletindo no aumento do número de triplas que constituem os mapas. As triplas extraídas automaticamente de cada texto foram então avaliadas utilizando as medidas de abrangência e precisão. Segundo Pérez, questões semânticas como a ambigüidade, a resolução anafórica de pronomes da terceira pessoa (ele, eles, etc) e complementos proposicionais ficaram fora do escopo do trabalho.

3.7.6 Considerações

Dois pontos diferenciam o trabalho realizado por Pérez em relação ao que está sendo aqui proposto. O primeiro diz respeito ao fato de a autora não mencionar o uso de medida para selecionar os conceitos. Um segundo ponto refere-se ao tipo de relação que é identificado em [PER99]. A autora trabalha relações predicativas, ou seja, onde os conceitos estão ligados a verbos ao quais servem de sujeito ou objeto. Diferentemente, nossa proposta é a de identificar as relações taxonômicas existentes entre os conceitos.

Além disso, segundo [PER04], apesar de serem construções similares e com muitas características em comum, mapas conceituais são mais flexíveis que ontologias e possuem proposições simples, enquanto ontologias possuem classes, sub-classes e outras relações bem definidas como, por exemplo, meronímia e hiperonímia. Assim, apesar de os trabalhos serem similares e igualmente voltados à língua portuguesa, nossa abordagem não está baseada nesta pesquisa.

3.8 Padrões para identificação das relações de hiponímia e hiperonímia

Apesar da variedade de relações que podem existir entre conceitos, segundo [RUI05] na maioria dos casos ontologias são estruturadas como hierarquias de conceitos (taxonomias). Esse tipo de relação é identificado nos textos por meio da relação semântica hiponímia e seu inverso, hiperonímia, que arranjam conceitos dos mais gerais aos mais específicos.

- *Hiponímia*: segundo [JUR00] é a relação em que um conceito denota uma subclasse do outro. O conceito mais específico é um hipônimo do mais genérico.

Por exemplo, a relação entre “turista” e “pessoa” é uma hiponímia, em que “turista” é um hipônimo de “pessoa”.

- *Hiperonímia*: é uma relação semântica inversa à hiponímia, ou seja, é a relação em que um conceito denota uma generalização do outro. O conceito mais genérico é um hiperônimo do mais específico. Neste caso, “pessoa” é um hiperônimo de “turista”.

Apesar de não estarem diretamente relacionadas à construção de ontologias, os padrões léxico-sintáticos identificados por Hearst [HEA92] e por Morin e Jacquemin [MOR03] serão apresentados a seguir, pois tais padrões serão utilizados como parte do processo para recuperação de relações taxonômicas em nossa abordagem.

3.8.1 Padrões de Hearst

Em [HEA92], Hearst apresenta uma lista com seis padrões léxico-sintáticos que indicam a relação hiponímia a partir de textos escritos na língua inglesa. Esses padrões são também utilizados em [MAE02] e [DEG04]. A Tabela 3.1 apresenta tais padrões, com sua tradução/adaptação para uso em textos escritos na língua portuguesa do Brasil.

Tabela 3.1: Padrões léxico-sintáticos de Hearst e suas adaptações para o português

	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 {,} especially {NP,}*{or|and} NP**	SUB {,} especialmente {SUB,}*{ou|e} SUB
		SUB {,} principalmente {SUB,}*{ou|e} SUB
		SUB {,} particularmente {SUB,}*{ou|e} SUB
		SUB {,} em especial { SUB,}*{ou|e} SUB
		SUB {,} em particular { SUB,}*{ou|e} SUB
		SUB {,} de maneira especial { SUB,}*{ou|e} SUB
		SUB {,} sobretudo { SUB,}*{ou|e} SUB

Na Tabela 3.1,

SUB: Substantivo

NP: Sintagma Nominal

Seguem, como exemplo, dois trechos de texto provenientes de [HEA92], a partir dos quais pretendemos demonstrar como esses padrões podem ser empregados em nosso trabalho:

Trecho 1: "*The bow lute such as the Bambara ndang, is ....*".

Trecho 2: "*...most European countries, especially France, England and Spain*."

Podemos observar que o primeiro trecho se enquadra no primeiro padrão da Tabela 3.1 (*NP such as NP*), tendo "*bow lute*" e "*Bambara ndang*" como NP, gerando, neste caso, a relação:

- hiponímia ("*Bambara ndang*", "*bow lute*").

Em relação ao segundo trecho, podemos observar que se trata do sexto padrão (*NP , especially {NP,}* and NP*) apresentado na Tabela 3.1. Desta forma, obteríamos as seguintes relações:

- hiponímia ("*France*" , "*European country*")
- hiponímia ("*England*", "*European country*")
- hiponímia ("*Spain*", "*European country*")

3.8.2 Padrões de Morin e Jacquemin

Em [MOR03], Morin e Jacquemin apresentam padrões para a aquisição de relações de hiperonímia para um corpus de textos escritos na língua francesa. A Tabela 3.2 apresenta tais padrões, onde as linhas subseqüentes a cada padrão referem-se a nossa adaptação para uso em textos escritos na língua portuguesa do Brasil.

Tabela 3.2: Padrões léxico-sintáticos de Morin e Jacquemin adaptados ao português

	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_SUB1
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

Na Tabela 3.2,

SUB1, SUB2: Substantivos
NP1, NP2: Sintagmas Nominais
LIST_SUB = refere-se a uma lista de substantivos
*: significa que o texto entre chaves anterior ao símbolo não é obrigatório na identificação do padrão

A partir dos exemplos que seguem, provenientes de [MOR03], demonstraremos como tais padrões funcionam.

Exemplo 1: {deux|trois...|2|3|4...} NP1: LIST2

Trecho: "*...et place dans la succession de trois arbres guyanais: Trema micrantha, Goupia glabra et Eperua grandiora.*"

A partir do padrão em questão, podemos identificar no trecho de texto, as seguintes relações:

- hiperônimo ("*Trema micrantha*", "*arbre guyanais*")
- hiperônimo ("*Goupia glabra*", "*arbre guyanais*")
- hiperônimo ("*Eperua grandiora*", "*arbre guyanais*")

Exemplo 2: {deux|trois...|2|3|4...} NP1 (LIST2)

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

A partir do padrão em questão, podemos identificar as seguintes relações no trecho de texto:

- hiperônimo ("*chêne*", "*espèce ligneux*")
- hiperônimo ("*frêne*", "*espèce ligneux*")
- hiperônimo ("*lierre*", "*espèce ligneux*")
- hiperônimo ("*cornouiller*", "*espèce ligneux*")

3.8.3 Considerações

Podemos observar, pelo disposto na Tabela 3.3, que alguns padrões de Morin e Jacquemin [MOR03] se equivalem a padrões identificados por Hearst [HEA92]. Desta forma, somente alguns padrões de Morin e Jacquemin (1, 2, 3, 4, 10) serão acrescentados ao nosso conjunto de regras. Como pode ser visto na Tabela 3.4, os padrões 1, 2, 3, e 4 foram

generalizados, pois entendemos que esta medida nos permitirá recuperar relações que não seriam identificadas caso mantidos os padrões no formato original.

Tabela 3.3: Equivalência entre os padrões de Morin e Jacquemin e padrões de Hearst

Morin e Jacquemin	Hearst
5 - {de autre} NP1 tel que LIST2	**1 –** NP *such as* {(NP,)*(*or*|*and*)} NP
6 - NP1, particulierement NP2	**6 –** NP {,} especially {NP,}*{or|and} NP
7 - {de autre} NP1 comme LIST2	**1 –** NP *such as* {(NP,)*(*or*|*and*)} NP
8 - NP1 tel LIST2	**1 -** NP *such as* {(NP,)*(*or*|*and*)} NP
9 - NP2 {et|ou} de autre NP1	**3 -** NP {, NP}* {,} or other NP
	4 - NP {, NP}* {,} and other NP

Assim, passamos a considerar os padrões generalizados na Tabela 3.4:

Tabela 3.4: Padrões de Morin e Jacquemin (adaptados)

Padrões de Morin e Jacquemin adaptados
1- SUB1 (LIST_SUB2)
2- SUB1: LIST_SUB1

Visto que estes padrões podem auxiliar na identificação de relações taxonômicas de interesse para uma ontologia de domínio, os mesmos serão utilizados em nossa proposta com as adaptações realizadas para seu uso em textos na língua portuguesa do Brasil.

3.9 Visão integrada das abordagens estudadas

Nossa primeira etapa, neste trabalho, foi fornecer uma visão geral sobre abordagens existentes para construção de ontologias a partir de textos. Na Tabela 3.5 apresentamos uma visão integrada dessas abordagens. Podemos ver que, de modo geral, a construção de ontologias passa pelos seguintes objetivos: extrair conceitos, extrair relações taxonômicas e não-taxonômicas e popular a ontologia com instâncias.

Dessa forma podemos classificar abordagens como “mais completas” ou “menos completas” conforme a quantidade de objetivos que elas se propõem a alcançar. Por esse ponto de vista, e conforme a Tabela 3.5, a abordagem de Buitelaar seria a menos abrangente (menos completa) entre as estudadas, tratando apenas da identificação de conceitos. Já a abordagem de Maedche, devido ao número de objetivos que visa alcançar, estaria entre as mais completas.

Tabela 3.5: Visão integrada das características gerais das abordagens estudadas

Autor*	Objetiva identificar	Principais etapas	Principais técnicas usadas	Nível de automatização	Intervenção do usuário
Buitelaar [BUI04]	• Conceitos e atributos	• Anotação lingüística • Pré-processamento estatístico • Definição de regras de mapeamento • Geração semi-automática de regras de mapeamento • Validação manual de conceitos e atributos pré-selecionados • Integração dos conceitos e atributos validados em uma ontologia (nova ou existente)	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• O usuário define regras de mapeamento e valida conceitos e atributos extraídos
Degeratu [DEG04]	• Termos relevantes • Relações taxonômicas • Relações não-taxonômicas	• Pré-processamento • Identificação de termos • Extração de relacionamentos (taxonômicos e não-taxonômicos) • Agrupamento de termos • Criação de hierarquia	• Análise e anotação lingüística • Abordagem estatística • Uso de padrões léxico-sintáticos	Automático	• A referência utilizada descreve que o usuário somente "interage" avaliando a precisão da ontologia resultante, ou seja, após o processo de construção automática da ontologia
Lame [LAM03]	• Conceitos • Relações não-taxonômicas	• Análise sintática • Análise de relações de coordenação • Análise estatística • *Pattern matching*	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• O usuário valida algumas saídas e determina alguns limiares
Maedche [MAE02]	• Conceitos • Relações taxonômicas • Relações não-taxonômicas	• Escolha da ontologia base (a ser ampliada) • Extração de informação • Aquisição de conceitos • Aquisição de taxonomia • Aquisição de relações conceituais	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• O usuário seleciona e nomeia relações • Caso necessário, na resolução de conflitos
Velardi [VEL01]	• Conceitos • Relações taxonômicas • Relações não-taxonômicas	• Identificar conceitos • Identificar instâncias de conceitos • Organizar os conceitos em hierarquias • Descobrir relações entre conceitos	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• Resultado é avaliado por especialistas • O usuário interage na integração de sub-árvores a nodos apropriados na ontologia e na definição de regras
Hearst [HEA94]	• Relações taxonômicas	• Identificar relações	• Identificação baseada em padrões	Semi-automático	• O usuário valida as relações extraídas
Morin [MOR]	• Relações taxonômicas	• Identificar relações	• Identificação baseada em padrões	Semi-automático	• O usuário valida as relações extraídas

* As abordagens estão identificadas pelo primeiro autor.

Outro ponto importante em relação a construção de ontologias é o nível de automatização proposto pela abordagem. Quanto a esse aspecto, a abordagem de Degeratu e Hatzivassiloglou propõe um processo totalmente automatizado. Isso, porém, nos deixa receosos quanto à qualidade da ontologia resultante, pois podemos verificar que as demais abordagens estudadas propõem a construção semi-automática de ontologias a partir de textos, requisitando intervenções do usuário em alguma parte importante do processo como, por exemplo, validação de conceitos ou de relações extraídas.

Na Tabela 3.6, podemos ver outras características mais específicas das abordagens estudadas. O reuso ou não de ontologias existentes, bem como o uso de fontes de conhecimento adicionais (por exemplo, dicionários ou corpora de texto mais genéricos) para auxiliar na extração de conceitos e relacionamentos, são aspectos muito importantes a serem considerados.

Tabela 3.6: Visão integrada de características mais específicas das abordagens estudadas

Autor*	Reuso de outras ontologias	Fontes de conhecimento utilizadas	Ferramentas associadas	Domínio onde foi aplicada
Buitelaar [BUI04]	• Não consta	• O corpus MuchMore[22] foi usado como corpus de referência contrastante, representando o domínio médico em geral	• *OntoLT - plugin* da ferramenta *Protégé*	• Neurologia
Degeratu [DEG04]	• Não utiliza	• Não utiliza	• *OntoStruct* • *MxTerminator*	• Comércio eletrônico
Lame [LAM03]	• Não consta	• Não consta	• *Syntex*	• Legislação francesa
Maedche [MAE02]	• Permite utilizar uma ontologia para servir de estrutura central (usou GermaNet)	• Podem ser usados dicionários, ontologias de domínio e genéricas como WordNet e GermaNet	• *Text-To-Onto* • *Shug*	• Seguros
Velardi [VEL01]	• Permite usar uma ontologia de domínio para ligar as sub-árvores geradas	• Não consta	• *OntoLearn* • *Chaos* • *Ariosto*	• Turismo

* Os métodos estão identificados pelo primeiro autor.

[22] http://muchmore.dfki.de/resources1.htm

A Tabela 3.7 refere-se a avaliação da ontologia resultante. Através dela podemos ver que a avaliação manual é a mais utilizada dentre as abordagens estudadas, tanto para a validação da ontologia resultante quanto para a validação das saídas em cada etapa como, por exemplo, na abordagem de Lame[LAM03], que apresenta apenas avaliação dos resultados de cada etapa, sem uma avaliação final. Segundo Maedche em [MAE02], não existe medida padrão para avaliação de ontologias extraídas de texto, e então o autor propõe uma abordagem de avaliação baseada nas medidas de precisão e *recall*.

Tabela 3.7: Avaliação da ontologia resultante

Autor*	Avaliação
Buitelaar [BUI04]	Uma plataforma para avaliação de ontologias extraídas é um dos seus trabalhos futuros.
Degeratu [DEG04]	Avalia somente precisão. A avaliação foi realizada por dois especialistas.
Lame [LAM03]	Cita somente validação manual das saídas de cada etapa e não de avaliação da ontologia resultante.
Maedche [MAE02]	Através das medidas de precisão e *recall* e com validação humana.
Velardi [VEL01]	Avaliação realizada por um especialista.

* As abordagens estão identificadas pelo primeiro autor

O próximo capítulo descreve a abordagem proposta nesta pesquisa, a qual foca nas atividades de identificação de termos relevantes do domínio e relações taxonômicas entre esses termos, bem como a fase relacionada à geração da estrutura ontológica. Descreve ainda as etapas de cada uma das fases.

4 ABORDAGEM PARA A CONSTRUÇÃO DE ESTRUTURAS ONTOLÓGICAS A PARTIR DE TEXTOS NA LÍNGUA PORTUGUESA DO BRASIL

Este capítulo apresenta nossa proposta para construção de estruturas ontológicas, fornecendo um detalhamento sobre cada etapa do processo.

Neste capítulo apresentamos nossa proposta para construção de estruturas ontológicas a partir de textos na língua portuguesa do Brasil. A abordagem, que tem como foco as atividades de identificação de termos relevantes do domínio e identificação de relações taxonômicas entre esses termos, surgiu da combinação de abordagens e técnicas apresentadas no capítulo anterior. A Figura 4.1 representa, de forma simplificada, as fases da abordagem que está sendo proposta. Nas próximas seções apresentamos uma descrição de cada uma dessas fases.

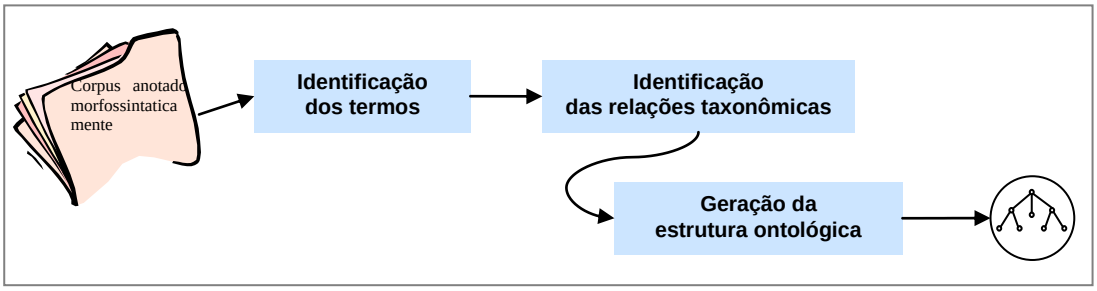

Figura 4.1: Visão simplificada da abordagem proposta

4.1 Entrada

O primeiro ponto importante a ser considerado sobre nossa proposta é a entrada utilizada pela mesma. As abordagens estudadas no Capítulo 3 incluem uma fase inicial para realizar anotação lingüística do texto (com auxílio de alguma ferramenta), o que envolve *tokenização*, processamento léxico-morfológico e análise sintática do corpus, entre outras ações. Sabemos que a eficácia de nossa proposta, assim como é o caso das abordagens estudadas, depende da correta identificação das etiquetas gramaticais. Embora existam ferramentas de pré-etiquetagem gramatical do corpus, desenvolvidas dentro do Grupo de Pesquisa em Processamento da Linguagem Natural da PUCRS, como descrito em [OLI02], as mesmas ainda não se encontram com níveis de confiabilidade compátiveis com o desejado ao

escopo deste trabalho. Por não dispormos de uma ferramenta com alta confiabilidade na etiquetagem e nem de tempo hábil para construí-la, optamos por não realizar a etapa de anotação lingüística. Assim, assumimos como entrada um corpus com textos já anotados lingüisticamente, 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.).

O padrão de etiquetas considerado neste trabalho é o utilizado pelo NILC[23]. A Tabela 4.1 apresenta tais categorias.

Tabela 4.1: Categorias gramaticais consideradas neste trabalho

Categoria	Etiqueta
Artigo definido	_AD
Artigo indefinido	_AI
Adjetivo	_AJ
Verbo no particípio	_AP
Advérbio	_AV
Conjunção coordenativa	_CC
Conjunção subordinativa	_CS
Interjeição	_IN
Locução adverbial	_LA
Locução conjuntiva	_LC
Locução interjetiva	_LI
Locução prepositiva	_LP
Numeral cardinal	_NC
Numeral ordinal	_NO
Pronome demonstrativo	_PD
Pronome indefinido	_PI
Pronome relativo	_PL
Pontuação (exceto a vírgula)	_PN
Pronome pessoal	_PP
Preposição	_PR
Pronome possessivo	_PS
Substantivo	_SU
Verbo auxiliar	_VA
Verbo (exceto particípio)	_VB
Vírgula	_VG

[23] Núcleo Interinstitucional de Lingüística Computacional - NILC (www.nilc.icmc.usp.br/nilc/)

4.2 Identificação de termos

A primeira fase de nossa abordagem é a identificação de termos relevantes do domínio, que é constituída pelas cinco etapas apresentadas na Figura 4.2. Esta fase utiliza como entrada um corpus com as características apresentadas na seção anterior. A execução das etapas 1 a 4 resulta em uma lista de termos relevantes simples (termos mono palavra). A execução da quinta etapa resulta em uma lista de termos compostos. A seguir descrevemos cada etapa desta primeira fase.

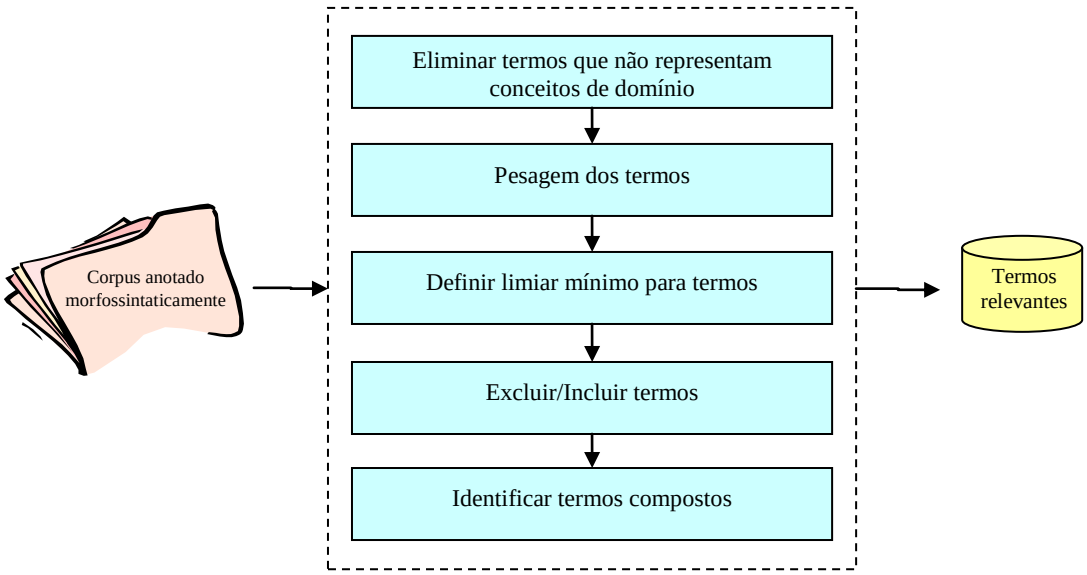

Figura 4.2: Etapas da fase de identificação de termos

1. **Eliminar termos que não representam conceitos de domínio:** através de uma lista de *stopwords* são excluídas do corpus palavras comuns que possuem significado semântico limitado e, portanto, não são relevantes o domínio. Atualmente essa lista conta com aproximadamente 550 palavras (entre artigos, preposições, advérbios, etc.). A lista de *stopwords* encontra-se no Anexo A. Nesta etapa, além das *stopwords*, são também removidos do corpus todos os termos contendo caracteres não-alfabéticos como números e símbolos, os quais estão rotulados pelas etiquetas apresentadas na Tabela 4.2. Tais termos são excluídos para serem identificados como possíveis termos relevantes. Porém vale salientar que os termos excluídos nesta etapa ainda podem ser utilizados em regras para identificação de termos compostos e relações taxonômicas entre os termos.

Tabela 4.2: Etiquetas de termos que não representam conceitos de domínio

Etiqueta	Nome etiqueta	Exemplo
_NC	Numeral cardinal	“61”, “70”, “2.800”
_NO	Numeral ordinal	“16º”, “segunda”
_PN	Pontuação (exceto a vírgula)	“!”, “?”, “.”
_VG	Vírgula	“,”

Outra característica da proposta é o fato de não trabalharmos com nomes próprios como termos (por exemplo, ‘Robert’, ‘Kyoto’, ‘Espanha’, ‘Sergipe’ e ‘Brasil’), visto que os mesmos representam instâncias de um domínio e, por conseqüência, seriam também instâncias em uma ontologia. Por exemplo, ‘Brasil’ e ‘Espanha’ seriam provavelmente instâncias do conceito (classe) ‘país’. Como a extração de instâncias não faz parte do escopo desta pesquisa, nesta etapa, palavras que representam nomes próprios são também desconsideradas. A identificação de nomes próprios no texto é realizada a partir das heurísticas apresentadas na Tabela 4.3.

Tabela 4.3: Regras para identificação de nomes próprios

Regra (SE)	Ação (ENTÃO)
1. A primeira e a segunda letra da palavra são maiúsculas.	Desconsidera o termo nas próximas etapas.
2. A palavra inicia com letra maiúscula e caractere anterior não é um ponto (“.”).	Desconsidera o termo nas próximas etapas.

Além disso, também são identificadas, a partir de heurísticas, palavras abreviadas como, por exemplo, “jr.”, “tel.”, “av.”, “sr.” e “pág.”. Assim como acontece com os nomes próprios, estas abreviaturas são igualmente desconsideradas nas etapas seguintes. A heurística usada para identificar uma palavra abreviada consiste em verificar se a palavra termina com um ponto (“.”). Em caso positivo (é abreviatura), a mesma é desconsiderada nas próximas etapas.

Resumindo, esta etapa consiste em identificar e desconsiderar nas próximas etapas, palavras que estejam na lista de *stopwords*, palavras contendo caracteres não-alfabéticos e palavras que representem nomes próprios e abreviaturas, pois estas palavras tendem a não constituir termos de um domínio.

2. Pesagem dos termos: A segunda etapa consiste em pesar as palavras candidatas a termos relevantes do domínio. Para isso utilizamos duas medidas: TFIDF (*term frequency* x *inverted document frequency*) e *Log-Likelihood*. Inicialmente apenas a medida TFIDF foi utilizada para pesar os termos e apresentá-los em ordem de relevância ao engenheiro de ontologia. Porém, tal medida retornava apenas a classificação de todos termos, resultando em uma quantidade muito grande de termos não relevantes sendo apresentados ao engenheiro de ontologia. Foi então adicionada a medida *Log-Likelihood* para comparar a freqüência dos termos no corpus do domínio face a sua freqüência em um corpus de referência, promovendo assim a exclusão automática de termos não relevantes ao domínio (ou seja, termos que aparecem em maior proporção no corpus de referência).

Nesta etapa, utilizamos o lema da palavra (disponível no corpus etiquetado) para realizar a pesagem. A utilização do lema na pesagem se deve ao fato de que os termos normalmente aparecem no corpus com diferentes propriedades (gênero, número e grau). Assim, considerando-se o lema da palavra, evita-se que um mesmo termo, representado com diferentes propriedades, receba distintos pesos como se fossem diferentes termos. Por exemplo, podem aparecer nos textos os termos “praia” e “praias”. Se não fosse utilizado o lema da palavra para computar os pesos, teríamos dois termos diferentes, cada um com seu peso associado. Utilizando-se o lema, estes termos passam a ser pesados como um único termo (“praia”).

3. Definição de limiar mínimo para termos: A terceira etapa consiste em definir uma freqüência (tfidf) mínima aceitável para um termo no corpus ser considerado relevante ao domínio. Com a definição desse limiar, os termos com freqüência abaixo do mesmo são excluídos. A definição desse limiar é responsabilidade do engenheiro de ontologia. Entretanto, a poda conforme a freqüência deve ser aplicada com cautela, visto que termos que aparecem poucas vezes ou apenas uma vez em um texto podem ser mais relevantes para o domínio do que termos mais freqüentes.

4. Excluir e Incluir termos: A idéia aqui é sugerir ao engenheiro de ontologia, como termos relevantes, a lista de termos resultantes das etapas anteriores. O engenheiro de ontologia pode então excluir os termos que julgar desnecessários ou por ventura incorretos. Nesta etapa é

permitido ao engenheiro de ontologia incluir termos relevantes não selecionados previamente. Todos os termos resultantes desta etapa serão considerados nas etapas subseqüentes.

5. Identificar termos compostos: A quinta etapa corresponde à identificação de termos compostos (ou termos multi-palavra). A partir da lista de termos relevantes, resultantes da execução das etapas 1 a 4, são selecionados termos compostos que contenham ao menos um termo relevante em sua composição. A identificação dos termos compostos é realizada com base em regras expressas por seqüências de etiquetas que, quando encontradas no texto, podem representar termos compostos. A Tabela 4.4 apresenta as seqüências de etiquetas utilizadas na identificação de termos compostos. Os termos compostos resultantes desta etapa são também considerados termos relevantes do domínio. A validação dos termos compostos extraídos deve ser realizada pelo engenheiro de ontologia.

Tabela 4.4: Regras para identificação 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

Na Tabela 4.4,

_SU: etiqueta para substantivos
_AJ: etiqueta para adjetivos
_PR: etiqueta para preposições
_AD: etiqueta para advérbios

A Tabela 4.5 apresenta as entradas e saídas de cada uma das etapas envolvidas na fase de identificação de termos e ainda observações quanto a sua automatização.

Tabela 4.5: Entradas e saídas em cada etapa

Fase	Entrada	Saída	Modo de Execução
1	Corpus original etiquetado	Corpus filtrado	Automática
2	Corpus filtrado	Lista de termos com peso (*tfidf*) associado	Automática
3	Lista de termos com peso (*tfidf*) associado	Lista de termos relevantes	Definir limiar: Manual Excluir com base no limiar: Automática
4	Lista de termos relevantes	Lista de termos relevantes revisada	Manual
5	Lista de termos relevantes revisada e corpus original etiquetado	Lista de termos compostos revisada	Identificar termos compostos: Automática Validação: Manual

4.3 Extração de relações taxonômicas

A segunda fase da abordagem refere-se à identificação de relações taxonômicas entre os termos relevantes derivados da fase anterior. Cada etapa desta fase procura extrair um conjunto de relações taxonômicas a partir de uma determinada abordagem. A Figura 4.3 apresenta as etapas desta fase.

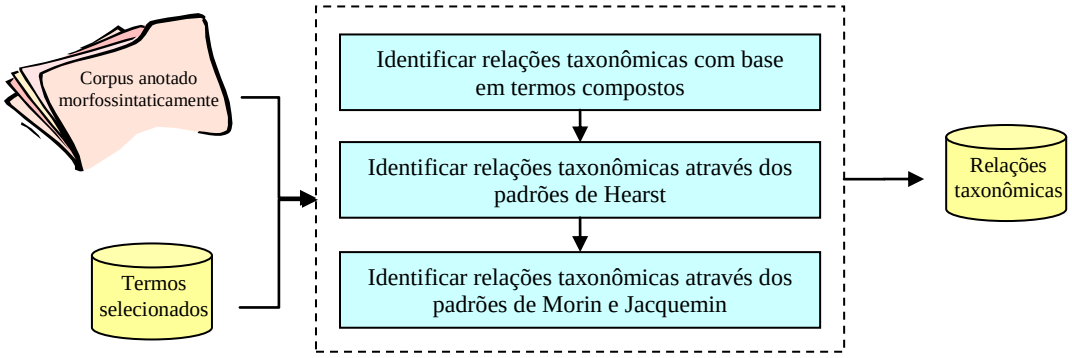

Figura 4.3: Etapas da fase de identificação de relações taxonômicas

1 Identificar relações taxonômicas com base em termos compostos: O primeiro passo na busca por relações taxonômicas é a identificação de relações a partir do núcleo do sintagma de termos compostos. A idéia aqui é relacionar cada termo composto ao termo relevante que faz parte da sua composição. Por exemplo, se foram identificados o termo relevante "contrato" e o termo composto "contrato de venda", a idéia é relacionar taxonomicamente esses dois

termos, de forma a identificar que “contrato de venda” é um tipo de “contrato”. A validação dessas relações é realizada pelo engenheiro de ontologia.

2 Identificar relações taxonômicas através dos padrões de Hearst: este segundo passo tem por objetivo a identificação de relações taxonômicas nos textos através dos padrões léxico-sintáticos propostos por Hearst em [HEA92]. A idéia aqui é encontrar no corpus os padrões de Hearst onde exista ao menos um termo relevante envolvido. A validação dessas relações também é realizada pelo engenheiro de ontologia.

Os padrões de Hearst foram criados inicialmente para a língua inglesa. Para sua utilização em textos da língua portuguesa do Brasil eles precisaram ser adaptados. A Tabela 4.6 apresenta os padrões originais e as traduções/adaptações realizadas. Primeiramente vale salientar que Hearst [HEA92] trabalha com sintagma nominal (*noun phrase* – NP) em seus padrões e que não dispomos desta de informação nos arquivos utilizados como entrada para a abordagem. Nesse sentido, nossa adaptação foi substituir a informação de NP diretamente por um substantivo (SU).

Tabela 4.6: Padrões léxico-sintáticos de Hearst e suas adaptações para o português

	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 {,} especially {NP,}*{or|and} NP**	SUB {,} especialmente {SUB,}*{ou|e} SUB
		SUB {,} principalmente {SUB,}*{ou|e} SUB
		SUB {,} particularmente {SUB,}*{ou|e} SUB
		SUB {,} em especial { SUB,}*{ou|e} SUB
		SUB {,} em particular { SUB,}*{ou|e} SUB
		SUB {,} de maneira especial { SUB,}*{ou|e} SUB
		SUB {,} sobretudo { SUB,}*{ou|e} SUB

3 Identificar relações taxonômicas através dos padrões de Morin e Jacquemin: Esta etapa objetiva identificar relações taxonômicas através dos padrões léxico-sintáticos propostos por Morin e Jacquemin em [MOR03]. A idéia, aqui, também é encontrar no corpus padrões

onde exista ao menos um termo relevante envolvido. A validação dessas relações é realizada pelo engenheiro de ontologia.

Os padrões de Morin e Jacquemin foram desenvolvidos para a língua francesa, e por isso foi necessário traduzi-los/adaptá-los para a língua portuguesa do Brasil. Os padrões com suas adaptações são apresentados na Tabela 4.7. Da mesma forma que Hearst [HEA92], Morin e Jacquemin trabalham em seus padrões com informação em nível de sintagma nominal e, da mesma forma, substituímos essa informação diretamente por um substantivo (SU).

Tabela 4.7: Padrões léxico-sintáticos de Morin e Jacquemin adaptados ao português

	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_SUB1
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

A numeração que acompanha SU, NP ou LIST_SU refere-se a sua posição na relação: (1) hiperonímia e (2) hiponímia.

A Tabela 4.8 apresenta as entradas e saídas de cada uma das etapas envolvidas na fase de identificação de relações taxonômicas.

Tabela 4.8: Entradas e saídas em cada etapa

Fase	Entrada	Saída	Modo de Execução
1	Lista de termos relevantes e lista de termos compostos	Lista de relações taxonômicas entre termos relevantes e termos compostos	Identificar relações: Automática Validação: Manual
2	Lista de termos relevantes e corpus original etiquetado	Lista de relações taxonômicas extraídas através dos padrões de Hearst [HEA92]	Identificar relações: Automática Validação: Manual
3	Lista de termos relevantes e corpus original etiquetado	Lista de relações taxonômicas extraída através dos padrões de Morin [MOR03]	Identificar relações: Automática Validação: Manual

4.4 Geração do código da estrutura ontológica

Nesta fase, o objetivo é utilizar o conhecimento adquirido, termos relevantes e relações taxonômicas, para gerar uma estrutura ontológica em uma linguagem de representação ontológica. A linguagem escolhida foi a OWL [W3C05], suportada pelo *framework* Jena[JEN05], o qual é um projeto *open-source* desenvolvido pelo HP Labs Semantic Web Programme. Em OWL os termos são representados por classes e as relações são representadas por propriedades, mapeadas através de relações hierárquicas (*rdfs:subClassOf*). A linguagem OWL permite que a ontologia seja editada ou estendida em uma ferramenta para edição de ontologias.

Nesta etapa o engenheiro de ontologia poderá gerar a estrutura ontológica 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.

A geração da estrutura ontológica em OWL está disponível no protótipo desenvolvido e apresentado no próximo capítulo.

4.5 Avaliação

Como pôde ser visto durante a descrição da proposta, existe a necessidade de avaliação dos resultados em algumas etapas. A forma avaliação proposta segue a forma mais utilizada dentre as abordagens estudadas: a validação manual por um especialista. A Tabela 4.9 refere-se à forma de avaliação da ontologia resultante proposta por diferentes autores, em comparação à proposta de avaliação desta dissertação.

Tabela 4.9: Avaliação da ontologia resultante

Autor*	Avaliação
Baségio [proposta]	Avaliação manual realizada por um especialista nas saídas de cada etapa e também da ontologia resultante.
Buitelaar [BUI04]	Uma plataforma para avaliação de ontologias extraídas é um dos seus trabalhos futuros.
Degeratu [DEG04]	Avalia somente precisão. A avaliação foi realizada por dois especialistas.
Lame [LAM03]	Cita somente validação manual das saídas de cada etapa e não de avaliação da ontologia resultante.
Maedche [MAE02]	Através das medidas de precisão e *recall* e com validação humana.
Velardi [VEL01]	Avaliação realizada por um especialista.

* As abordagens estão identificadas pelo primeiro autor

Em nossa proposta, a avaliação dos resultados se torna necessária a partir da quarta etapa da primeira fase, logo após a pesagem dos termos e definição do limiar. Nesta etapa são apresentados ao engenheiro de ontologia (especialista) os termos identificados e considerados relevantes ao domínio. Cabe ao engenheiro de ontologia avaliar os termos e então excluir os aqueles julgados incorretos. Caso algum termo considerado relevante pelo engenheiro não tenha sido selecionado, o mesmo pode incluí-lo nesta etapa. A quinta etapa desta mesma fase corresponde à identificação de termos compostos. A validação dos termos compostos extraídos também deve ser realizada pelo engenheiro de ontologia. Uma avaliação correta neste ponto (termos simples e termos compostos) é muito importante, pois todos os termos resultantes desta etapa servirão de base para a segunda fase.

Na segunda fase da abordagem, relacionada à identificação de relações taxonômicas, a avaliação dos resultados pelo engenheiro de ontologia deve ocorrer em todas as três fases: Identificar relações taxonômicas com base em termos compostos; Identificar relações taxonômicas através dos padrões de Hearst; Identificar relações taxonômicas através dos

padrões de Morin e Jacquemin. A avaliação dos resultados das etapas citadas nesta sessão implicará na qualidade da ontologia resultante.

4.6 Considerações

A abordagem aqui proposta, de modo geral, está baseada nas abordagens e medidas estudadas durante o desenvolvimento deste trabalho. A Tabela 4.10 mostra em destaque as principais contribuições de cada autor para a definição desta proposta.

Tabela 4.10: Principais contribuições de cada autor para definir esta proposta

Autor	Medidas	Termos	Relações	Automático	Intervenções do usuário
Buitelaar [BUI04]	Qui-quadrado	**Termos simples**	Baseado em regras de mapeamento	Semi-automático	Define regras de mapeamento e valida conceitos e atributos extraídos.
Degeratu [DEG04]	Informação mútua para variantes. Seleciona termos que ocorram 2 ou mais vezes em um texto ou múltiplos textos.	**Variantes** e termos simples	Padrões de Hearst	Automático	Avalia a precisão da ontologia resultante.
Lame [LAM03]	Não consta	**Simples e compostos**	**Relações baseadas em termos compostos**	**Semi-automático**	**Validação manual das saídas de cada etapa.**
Maedche [MAE02]	**TFIDF**	**Termos simples**	Padrões de Hearst e agrupamento hierárquico por similaridade.	Semi-automático	Valida relações. Se necessário, auxilia na resolução de conflitos.
Velardi [VEL01]	Relevância de domínio e consenso de domínio	**Simples e compostos**	**Relações baseadas em termos compostos**	Semi-automático	Validação da ontologia resultante e sua integração a uma ontologia base.
Hearst [HEA94]	N/A	N/A	**Padrões léxico-sintáticos**	Semi-automático	O usuário valida as relações extraídas.
Morin [MOR03]	N/A	N/A	**Padrões léxico-sintáticos**	Semi-automático	O usuário valida as relações extraídas.
Rayson [RAY04]	**Log-likelihood**	-	-	-	-

As tabelas 4.11 e 4.12 posicionam a abordagem aqui proposta face a características gerais e características mais específicas das abordagens estudadas.

Tabela 4.11: Abordagem proposta face à características gerais das abordagens estudadas

Autor*	Objetiva identificar	Principais etapas	Principais técnicas	Aautomatização	Intervenção do usuário
Buitelaar [BUI04]	• Conceitos e atributos	• Anotação lingüística • Pré-processamento estatístico • Definição de regras de mapeamento • Geração semi-automática de regras de mapeamento • Validação manual de conceitos e atributos pré-selecionados • Integração dos conceitos e atributos validados em uma ontologia (nova ou existente)	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• O usuário define regras de mapeamento e valida conceitos e atributos extraídos
Degeratu [DEG04]	• Termos relevantes • Relações taxonômicas • Relações não-taxonômicas	• Pré-processamento • Identificação de termos • Extração de relacionamentos (taxonômicos e não-taxonômicos) • Agrupamento de termos • Criação de hierarquia	• Análise e anotação lingüística • Abordagem estatística • Uso de padrões léxico-sintáticos	Automático	• A referência utilizada descreve que o usuário somente "interage" avaliando a precisão da ontologia resultante, ou seja, após o processo de construção automática da ontologia
Lame [LAM03]	• Conceitos • Relações não-taxonômicas	• Análise sintática • Análise de relações de coordenação • Análise estatística • *Pattern matching*	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• O usuário valida algumas saídas e determina alguns limiares
Maedche [MAE02]	• Conceitos • Relações taxonômicas • Relações não-taxonômicas	• Escolha da ontologia base (a ser ampliada) • Extração de informação • Aquisição de conceitos • Aquisição de taxonomia • Aquisição de relações conceituais	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• O usuário seleciona e nomeia relações • Caso necessário, na resolução de conflitos
Velardi [VEL01]	• Conceitos • Relações taxonômicas • Relações não-taxonômicas	• Identificar conceitos • Identificar instâncias de conceitos • Organizar os conceitos em hierarquias • Descobrir relações entre conceitos	• Análise e anotação lingüística • Abordagem estatística	Semi-automático	• Resultado é avaliado por especialistas • O usuário interage na integração de sub-árvores a nodos apropriados na ontologia e na definição de regras
Hearst [HEA94]	• Relações taxonômicas	• Identificar relações	• Identificação baseada em padrões	Semi-automático	• O usuário valida as relações extraídas
Morin [MOR]	• Relações taxonômicas	• Identificar relações	• Identificação baseada em padrões	Semi-automático	• O usuário valida as relações extraídas
Baségio [proposta]	• Termos relevantes • Relações taxonômicas	• Identificar termos relevantes simples e compostos • Identificar relações taxônomicas	• Abordagem estatística • Identificação baseada em padrões	Semi-automático	• O usuário valida saídas e determina alguns limiares se desejado

* As abordagens estão identificadas pelo primeiro autor.

Tabela 4.12: Abordagem proposta face à características específicas das abordagens estudadas

Autor*	Reuso de outras ontologias	Fontes de conhecimento utilizadas	Ferramentas associadas	Domínio onde foi aplicada
Buitelaar [BUI04]	• Não consta	• O corpus MuchMore foi usado como corpus de referência contrastante, representando o domínio médico em geral	• *OntoLT - plugin* da ferramenta *Protégé*	• Neurologia
Degeratu [DEG04]	• Não utiliza	• Não utiliza	• *OntoStruct* • *MxTerminator*	• Comércio eletrônico
Lame [LAM03]	• Não consta	• Não consta	• *Syntex*	• Legislação francesa
Maedche [MAE02]	• Permite utilizar uma ontologia para servir de estrutura central (usou GermaNet)	• Podem ser usados dicionários, ontologias de domínio e genéricas como WordNet e GermaNet	• *Text-To-Onto* • *Shug*	• Seguros
Velardi [VEL01]	• Permite usar uma ontologia de domínio para ligar as sub-árvores geradas	• Não consta	• *OntoLearn* • *Chaos* • *Ariosto*	• Turismo
Baségio [proposta]	• Não utiliza	• Corpus de referência(geral) disponibilizado pelo NILC	• *Protótipo desenvolvido no escopo desta dissertação*	• Turismo

* Os métodos estão identificados pelo primeiro autor.

5 AMBIENTE PARA A CONSTRUÇÃO DE ONTOLOGIAS A PARTIR DE TEXTOS/ PROTÓTIPO E IMPLEMENTAÇÃO

Este capítulo descreve o protótipo desenvolvido no intuito de auxiliar na validação e avaliação da abordagem para identificação de estruturas ontológicas proposta nesta dissertação.

Neste capítulo é descrito o protótipo de software desenvolvido no contexto deste trabalho. O protótipo foi desenvolvido com o objetivo principal de verificar a aplicabilidade da abordagem para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil proposta nesta dissertação. É importante ressaltar que nosso objetivo, durante a construção do protótipo, foi exclusivamente desenvolver a abordagem proposta, sendo que questões relacionadas à modelagem do protótipo e otimização dos algoritmos foram tratadas com menor prioridade.

Assim, neste capítulo apresentamos, de modo geral, o funcionamento dos módulos do protótipo sem entrar em maiores detalhes sobre as características específicas da implementação.

5.1 Características gerais da implementação

O protótipo de software foi implementado na linguagem Java, versão 1.5.0, através do IDE Net Beans versão 4.1. Durante seu desenvolvimento, ele foi testado em um ambiente com plataforma Windows XP utilizando um computador Pentium 3, com 1,2GHz e 256MB de memória.

5.2 Funcionalidades

As funcionalidades do protótipo estão descritas na Tabela 5.1. O ator corresponde ao engenheiro de ontologias que fará uso do ambiente de apoio para gerar uma estrutura ontológica. As funcionalidades em negrito são aquelas onde existe especificidade quanto à língua portuguesa.

Tabela 5.1: Funcionalidades do protótipo

Funcionalidade	Descrição
Importar Corpus	O engenheiro de ontologia deve ser capaz de importar um arquivo texto contendo o corpus de um determinado domínio.
Visualizar Textos	O engenheiro de ontologia deve ser capaz de visualizar os textos do corpus importado.
Excluir *Stopwords* e Palavras contendo Caracteres Não-alfabéticos	O engenheiro de ontologia deve ser capaz de excluir do processo de pesagem, as stopwords e palavras contendo caracteres não-alfabéticos.
Pesar termos	O engenheiro de ontologia deve ser capaz de pesar os termos através da medida TFIDF.
Visualizar termos	O engenheiro de ontologia deve ser capaz de visualizar os termos com seus pesos associados.
Filtrar termos	O engenheiro de ontologia deve ser capaz de filtrar os termos através da definição de um limiar para seu peso.
Excluir termos	O engenheiro de ontologia deve ser capaz de excluir termos considerados irrelevantes.
Incluir termos	O engenheiro de ontologia deve ser capaz de incluir termos não selecionados.
Identificar termos compostos	O engenheiro de ontologia deve ser capaz de identificar termos compostos.
Excluir termos compostos	O engenheiro de ontologia deve ser capaz de excluir termos compostos identificados.
Identificar relações taxonômicas a partir de termos compostos	O engenheiro de ontologia deve ser capaz de identificar relações taxonômicas a partir de termos compostos.
Identificar relações taxonômicas a partir dos padrões de Hearst	O engenheiro de ontologia deve ser capaz de identificar relações taxonômicas a partir dos padrões de Hearst[HEA92].
Identificar relações taxonômicas a partir dos padrões de Morin e Jacquemin	O engenheiro de ontologia deve ser capaz de identificar relações taxonômicas a partir dos padrões de Morin e Jacquemin [MOR03].
Excluir relações	O engenheiro de ontologia deve ser capaz de excluir relações identificadas.
Identificar relações duplicadas	O engenheiro de ontologia deve ser capaz de identificar relações duplicadas.
Gerar estrutura ontológica em OWL	O engenheiro de ontologia deve ser capaz de gerar a estrutura ontológica em OWL.
Visualizar a estrutura gerada	O engenheiro de ontologia deve ser capaz de visualizar a estrutura ontológica gerada.
Salvar estrutura ontológica em OWL	O engenheiro de ontologia deve ser capaz de salvar a estrutura ontológica em um arquivo OWL.

5.3 Arquitetura do protótipo

A Figura 5.1 fornece uma visão geral do processo de identificação de estruturas ontológicas do protótipo, caracterizando as entradas e saídas geradas em cada etapa.

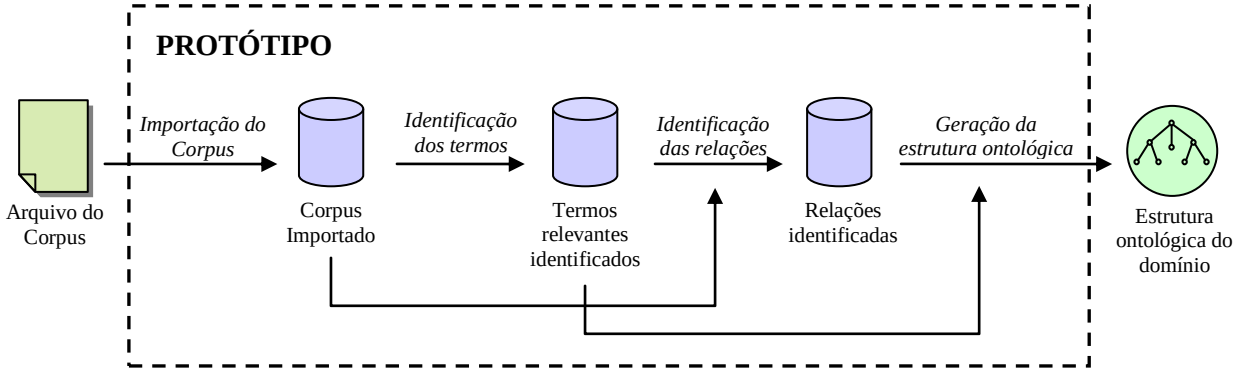

Figura 5.1: Processo de identificação de estruturas ontológicas

Na Figura 5.2 apresentamos a arquitetura do protótipo e o modo como seus módulos estão estruturados para atender às funcionalidades propostas.

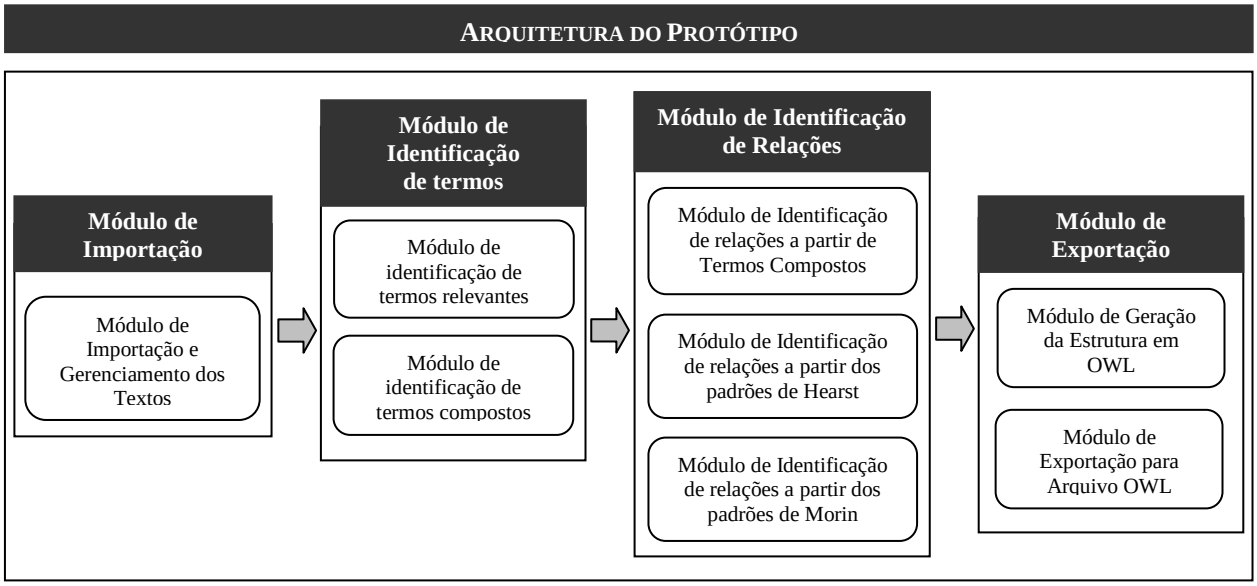

Figura 5.2: Arquitetura do protótipo

5.3.1 Módulo de importação

5.3.1.1 *Módulo de Importação e Gerenciamento de Textos*

Este módulo é o primeiro a ser utilizado no protótipo, pois possibilita ao usuário selecionar um arquivo texto contendo os textos de um determinado domínio; obter informações sobre o tamanho do corpus em questão e visualizar os textos importados, que são utilizados nos demais módulos.

a) Arquivo de entrada

Assume-se como entrada um arquivo texto contendo artigos de um determinado

domínio, sendo cada artigo formado por um título e um texto. Um título é identificado pela informação <tit id>, onde id é o número que identifica o artigo no corpus. O texto do artigo inicia pela informação <art id> onde id tem a mesma identificação do título do artigo. Cada artigo termina quando um novo título é identificado. Além disso, cada linha do arquivo, exceto aquelas que definem o início do título ou do texto do artigo, contém as seguintes informações, separadas por um espaço em branco: a palavra no formato original, o lema e a etiqueta gramatical da palavra. A Figura 5.3 representa um arquivo texto contendo as informações citadas. Em nossa abordagem, apenas a palavra no formato original, o lema e a etiqueta gramatical da palavra são utilizados, sendo as demais informações desconsideradas.

```
<tit 340>
Temperatura temperatura _SU
surpreende surpreender _VB
turista turista _SU
. . _PN
<art 340>
Para para _PR
os o _AD
turistas turista _SU
que que _PL
esperavam esperar _VB
o o _AD
frio frio _SU
de de _PR
montanha montanha _SU
, , _VG
```

Figura 5.3: Exemplo de formato de um artigo do corpus

b) Seleção do Arquivo do Corpus

A Figura 5.4 representa a interface do protótipo responsável pela importação do arquivo do corpus. Primeiramente, o protótipo permite selecionar um arquivo texto contendo o corpus. Uma vez selecionado o arquivo, é permitido ao usuário visualizar cada texto importado em uma pequena tela.

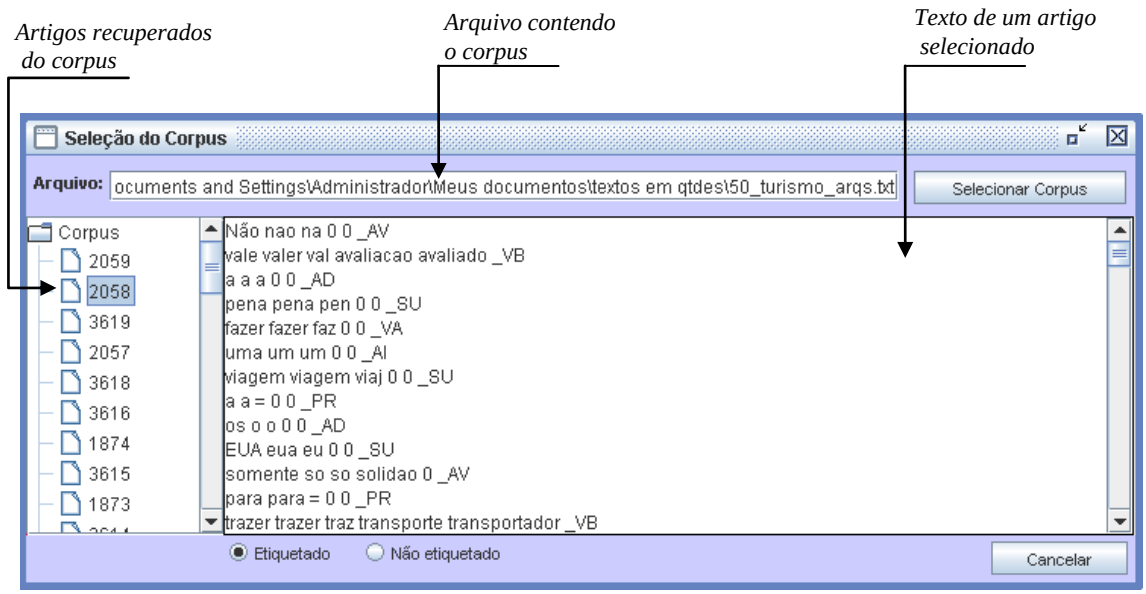

Figura 5.4: Interface para seleção e visualização dos textos do corpus

5.3.2 Módulo de identificação de termos

Este módulo possibilita ao usuário identificar os termos relevantes contidos no corpus do domínio selecionado.

5.3.2.1 *Módulo de Identificação de Termos Relevantes*

A Figura 5.5 apresenta a interface para a identificação dos termos relevantes simples, juntamente com os aspectos considerados durante o processo.

1. *Remover Stopwords*: a seleção deste item implica que, durante a pesagem dos termos, serão desconsideradas todas as palavras identificadas como *stopwords* e palavras contendo caracteres não-alfabéticos. Apesar da proposta estar baseada na exclusão de *Stopwords*, o protótipo possibilita, para fins de testes, a não seleção deste item, resultando na não execução deste passo.
2. *Pesar termos*: este botão, quando acionado, ativa o processo responsável pela pesagem dos termos. A pesagem é realizada através das medidas *Log-Likelihood e tfidf*. Como resultado, os termos com seus respectivos pesos (*tf*, *Log-Likelihood*, *tfidf*) são apresentados em uma tabela.

3. *Filtrar*: este botão, quando acionado, exclui os termos relevantes da tabela que possuírem um peso abaixo do limiar definido para a medida *tfidf*. Porém, também é possível utilizar no filtro, a simples freqüência do termo (*tf*) ou o peso *Log-Likelihood* associado ao termo.
4. *Excluir selecionados*: este botão, quando acionado, exclui todos os termos selecionados pelo usuário. A idéia é possibilitar a exclusão de termos julgados desnecessários ou irrelevantes para o domínio.
5. *Incluir termo*: tem por objetivo possibilitar ao usuário incluir termos relevantes para o domínio, que não tenham sido selecionados pela ferramenta. Ao clicar no botão o sistema abre uma nova janela permitindo a entrada de uma palavra e sua adição.

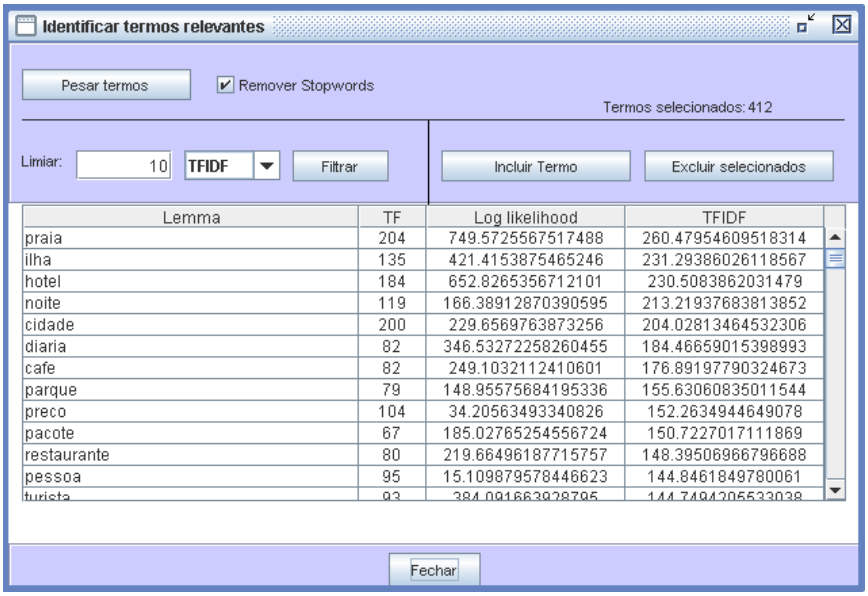

Figura 5.5: Interface para identificação de termos relevantes com pesos associados

5.3.2.2 *Módulo de Identificação de Termos Compostos*

A Figura 5.6 apresenta a interface para a identificação de termos compostos, juntamente com os aspectos considerados durante o processo.

1. *Identificar*: este botão, quando acionado, identifica termos compostos que contenham ao menos um termo relevante em sua composição. A extração é baseada em regras explícitas de seqüências de etiquetas nos textos. Como resultado são apresentados os termos compostos juntamente com sua freqüência no texto.
2. *Excluir selecionados*: este botão, quando acionado, exclui os termos selecionados pelo usuário que foram julgados desnecessários ou irrelevantes.

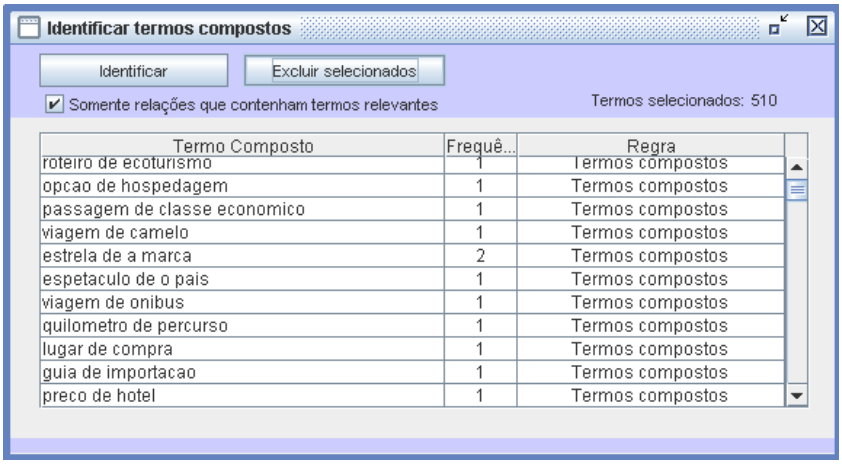

Figura 5.6: Interface para seleção de termos compostos

5.3.3 Módulo de identificação de relações

As interfaces disponibilizadas na ferramenta para a identificação de relações taxonômicas apresentam as funcionalidades de identificação e exclusão de relações. Em ambas, são apresentadas como resultado, as relações identificadas, suas freqüências no texto e a regra pela qual foram extraídas. Como exemplo, a Figura 5.7 apresenta a interface para identificação de relações taxonômicas a partir dos padrões de Morin e Jacquemin.

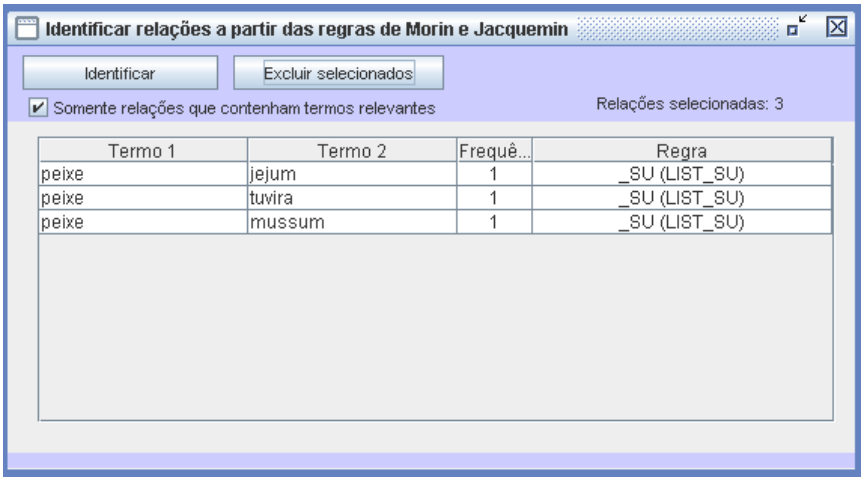

Figura 5.7: Interface para seleção de relações a partir das regras de Morin e Jacquemin

5.3.4 Módulo de exportação

Este módulo tem como objetivo fornecer funcionalidades que possibilitem ao usuário a geração da estrutura ontológica em uma linguagem de representação de ontologias.

5.3.4.1 *Módulo de Geração da Estrutura em OWL*

O protótipo utiliza o *framework* Jena[JEN05] para criar, em OWL[W3C05], uma estrutura inicial da ontologia do domínio. A Figura 5.8 mostra a interface para geração do código OWL, tendo-se como base os termos e relações selecionados.

1. *Termos*: possibilita ao usuário selecionar quais termos (Relevantes e Compostos) serão incluídos na estrutura ontológica a ser gerada.
2. *Relações*: possibilita ao usuário selecionar quais relações (Compostas, Hearst e/ou Morin e Jacquemin) serão incluídas na estrutura ontológica a ser gerada.
3. *Gerar*: ativa a geração do código OWL, de acordo com os termos e relações selecionados. Cada termo é representado em OWL por uma classe e as relações entre os termos (propriedades em OWL) são mapeadas através de relações hierárquicas (*rdfs:subClassOf*).

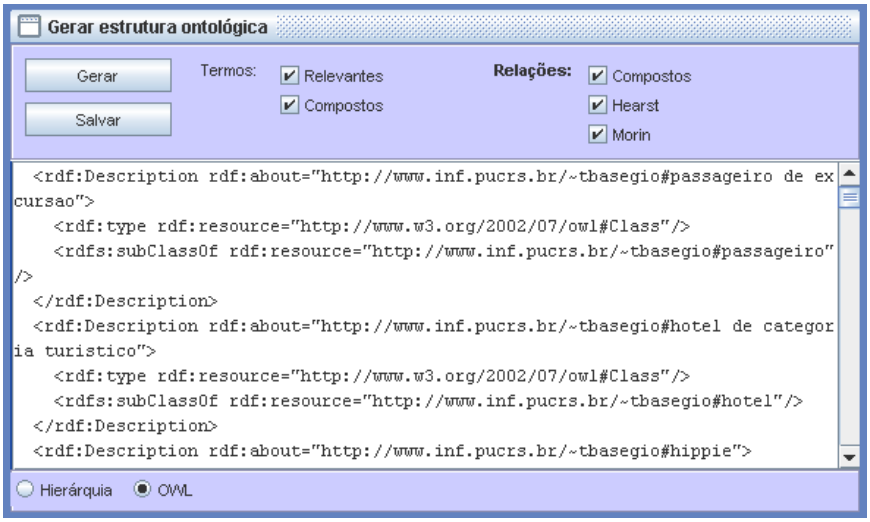

Figura 5.8: Interface para geração do código OWL

5.3.4.2 *Módulo de Exportação para Arquivo OWL*

O módulo de exportação possibilita, através do botão “Salvar” apresentado na interface mostrada na Figura 5.8, salvar a estrutura ontológica gerada em um arquivo OWL. A partir do arquivo OWL gerado é possível estender a ontologia manualmente ou com uso de ferramentas que apóiam a edição de ontologias. Como exemplo, na Figura 5.9 apresentamos parte de um código OWL gerado sendo visualizado na ferramenta Protégé [PRO05].

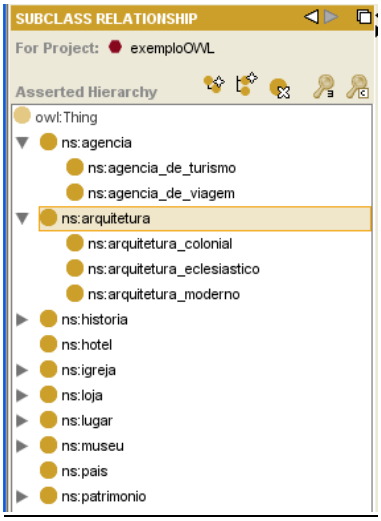

Figura 5.9: Estrutura taxonômica visualizada na ferramenta Protégé

5.4 Considerações

Este capítulo destacou o protótipo de software desenvolvido no contexto desta pesquisa. Foram apresentados os módulos do protótipo, com suas características, funcionalidades e interfaces, bem como o formato utilizado no arquivo de entrada.

O próximo capítulo apresenta o estudo de caso realizado com o objetivo de validar a abordagem proposta com a ferramenta desenvolvida. Serão ainda descritos o processo de condução da avaliação e seus resultados.

6 ESTUDO DE CASO NO DOMÍNIO DO TURISMO

O presente capítulo descreve os estudos de caso realizados sobre o domínio do Turismo com o objetivo de validar a abordagem para identificação de estruturas ontológicas proposta nesta pesquisa.

Para validar a abordagem proposta nesta pesquisa, foi definido um estudo de caso consistindo na identificação de termos e relações taxonômicas a partir de um corpus do domínio do Turismo, fazendo uso do protótipo apresentado no Capítulo 5. Seguindo a abordagem proposta, em cada etapa ocorreu a validação dos resultados por um especialista na área do Turismo. Os dados selecionados em uma etapa serviram de entrada e direcionaram os resultados das etapas subseqüentes, ou seja, termos excluídos em uma etapa foram desconsiderados nas demais. O estudo de caso teve como objetivo analisar o modo como a abordagem proposta auxilia na identificação de estruturas ontológicas.

Um segundo estudo de caso foi realizado utilizando o mesmo corpus do domínio do primeiro estudo de caso. A diferença deste para o primeiro está na forma como os dados foram apresentados ao especialista. Após a seleção do corpus, a abordagem foi executada sem a intervenção do especialista, isto é, sem que termos ou relações fossem excluídos entre as etapas, fazendo com que um número maior de termos e relações fossem extraídos. Ao final do processo, todos os termos e relações extraídos foram apresentados ao especialista para exclusão daqueles não relevantes ao domínio. Este segundo estudo de caso teve como objetivo verificar a viabilidade de a abordagem proposta ser executada sem intervenção humana, possibilitando maior automatização, com validação do resultado pelo especialista apenas no final do processo.

O especialista que participou dos estudos de caso aqui apresentados, sendo responsável pela validação tanto das saídas de cada etapa quanto da ontologia final, é um profissional altamente especializado[24] na área do Turismo.

Nas seções seguintes, discorremos sobre o corpus utilizado, bem como os resultados obtidos em cada estudo de caso.

---

[24] Doutor na área do Turismo, coordenador do curso de pós-graduação em Turismo da PUCRS, tendo grande atuação profissional na área em questão.

6.1 Corpus de referência e corpus do domínio

O corpus de referência (ou corpus geral), utilizado no estudo de caso é formado a partir de uma coleção de documentos do Jornal Folha de São Paulo do ano de 1994[25], já no formato apresentado na seção 4.1 do Capítulo 4. Os textos utilizados pertencem a diferentes seções do jornal e compreendem um total de 3.862 documentos com 974.685 palavras.

O corpus do domínio utilizado para o estudo de caso é constituído por um total de 294 documentos com 88.601 palavras, documentos estes extraídos do Caderno de Turismo da Folha (retirados do corpus de referência citado). Um ponto importante a ser considerado em relação ao corpus utilizado é o fato de que textos de jornal, de modo geral, mesmo que de uma seção específica, não podem ser considerados textos especializados de um domínio. Textos com essa característica são considerados semi-especializados. Assim, por não ser um corpus especializado do domínio do Turismo, os textos utilizados no estudo de caso não contêm apenas termos relacionados ao domínio do Turismo. A utilização desse corpus se deu pelo fato de não dispormos de um corpus específico de domínio etiquetado conforme nossa necessidade.

6.2 Estudo de caso 1

A execução deste estudo de caso iniciou com a seleção do corpus do Turismo descrito na seção anterior. A primeira etapa realizada foi a identificação de termos, sendo o especialista do domínio responsável por selecionar, dentre os termos extraídos pela ferramenta (simples e compostos), aqueles considerados relevantes ao domínio do Turismo. Depois o especialista selecionou as relações taxonômicas relevantes, identificadas pela ferramenta com base nos termos selecionados previamente. Por fim gerou-se a estrutura ontológica em OWL. Nas seções seguintes detalhamos os resultados obtidos no desenvolvimento do estudo de caso.

---

[25] Este corpus foi gentilmente disponibilizado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC), ao grupo de pesquisa em PLN da PUCRS. Maiores informações sobre o NILC podem ser encontradas em www.nilc.icmc.usp.br/nilc/ .

6.2.1 Identificação de termos

Nas subseções seguintes são apresentados os resultados referentes a cada passo da etapa de identificação de termos relevantes do domínio.

6.2.1.1 *Eliminar termos que não representam conceitos de domínio*

Neste passo foram excluídas as palavras que não têm significado para o domínio, caracterizadas como *stopwords*, palavras contendo caracteres não-alfabéticos (números e símbolos), nomes próprios e abreviaturas.

A Tabela 6.1 apresenta a quantidade de palavras identificadas como não representando conceitos do domínio, explicitando a categoria pela qual foram desconsideradas. A coluna "Diferentes" apresenta quantas palavras distintas foram excluídas, ou seja, considera cada palavra apenas uma vez, ignorando as repetições. A coluna "Total" apresenta a quantidade total de palavras excluídas, contabilizando todas as vezes que a palavra aparece. Por exemplo, se a palavra "aquele" foi excluída 5 vezes, ela será contabilizada 5 vezes na coluna "Total" e apenas 1 vez na coluna "Diferentes". A coluna "% do corpus" refere-se ao percentual de palavras excluídas do corpus do domínio utilizando como referência a coluna "Total" em relação ao total de palavras do corpus.

Tabela 6.1: Quantidade de palavras não consideradas relevantes ao domínio

Categoria	Diferentes	Total	% do corpus
Stopwords	256	47292	53,37
Abreviaturas	19	120	0,14
Nomes próprios	2946	7762	8,76
Não-alfabéticas	1600	6200	7,00
Total	**4821**	**61374**	**69,71**

Como podemos verificar na Tabela 6.1, grande parte do corpus do domínio não foi considerada como candidata a termo relevante. No total foram excluídas 61.374 palavras (69,72%) do corpus. O maior percentual de palavras excluídas foi classificado como *stopwords*, contabilizando mais de 50% do corpus utilizado. Também chama à atenção, o grande número de nomes próprios encontrados no texto, representando possíveis instâncias do domínio. O Gráfico 6.1 apresenta a distribuição das palavras do corpus, excluídas de acordo com sua categoria, bem como o total de palavras restantes após a execução deste passo.

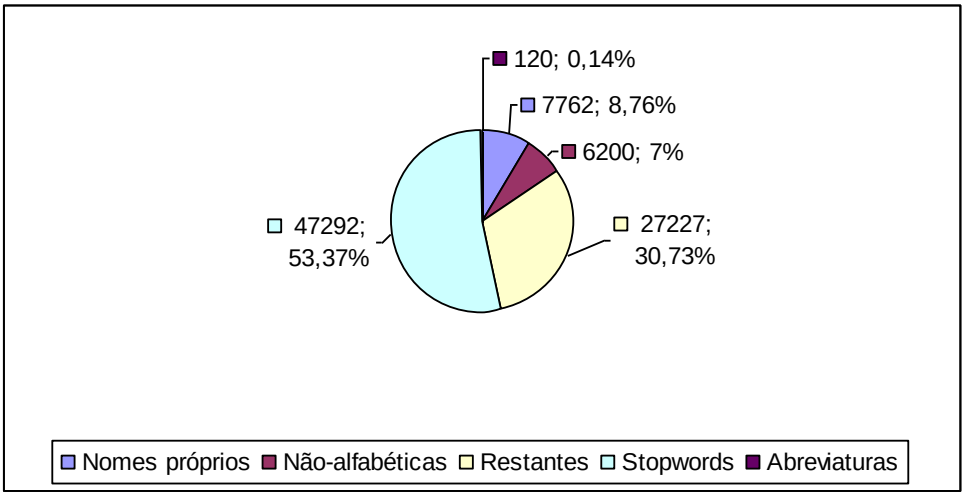

Gráfico 6.1: Distribuição das palavras excluídas por categoria e palavras restantes

6.2.1.2 *Pesagem dos termos*

Com a eliminação das palavras não relevantes ao domínio, restaram 27.227 palavras (30,73% do corpus), dentre as quais 14.485 são substantivos (4.047 substantivos diferentes), correspondendo a 16,35% do corpus.

Os 4.047 substantivos restantes, que são os verdadeiros candidatos a termos relevantes, tiveram sua freqüência no corpus do domínio comparada a sua freqüência no corpus de referência com a utilização da medida *Log-Likelihood*. Este passo nos possibilitou excluir automaticamente 3.635 diferentes substantivos (7.288 no total) não específicos ao domínio do Turismo, ou seja, substantivos que aparecem em maior proporção no corpus de referência. Assim, restaram apenas 412 substantivos candidatos a termos relevantes do domínio. O Gráfico 6.2 apresenta a proporção entre os termos filtrados pela medida *Log-Likelihood* e os termos que permaneceram após a filtragem.

Para apresentar os termos ao especialista em ordem de relevância, foi utilizada a medida TFIDF, associando-se um peso a cada uma das palavras restantes. A utilização da medida TFIDF se deve ao fato de que essa medida mostrou um melhor desempenho em testes realizados do que a simples freqüência do termo no corpus.

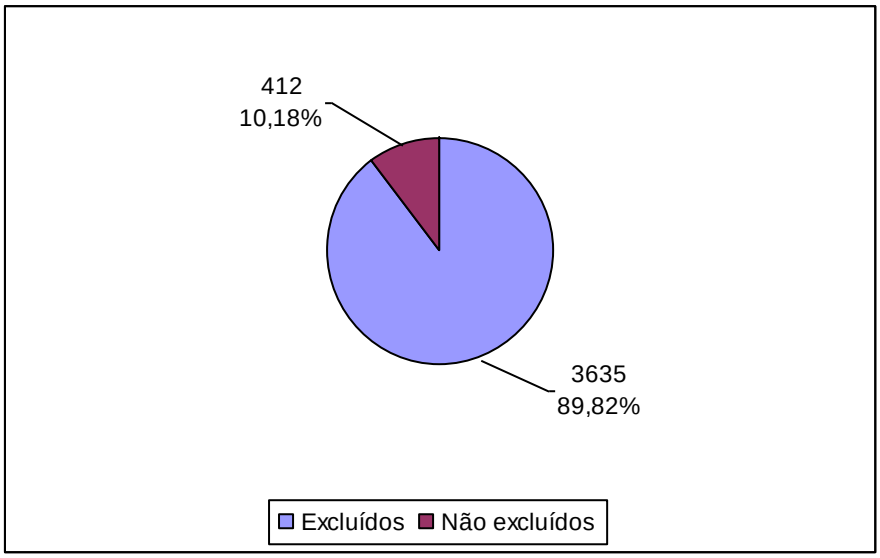

Gráfico 6.2: Distribuição dos substantivos candidatos a termos, resultantes da pesagem

6.2.1.3 *Definição de limiar mínimo para termos*

Nossa intenção neste passo foi possibilitar a poda de termos a partir da definição de uma freqüência mínima para o termo ser considerado relevante. Na ferramenta foi disponibilizada a poda pela medida TFIDF, pela simples freqüência (TF) ou ainda pela medida *Log-Likelihood*. Como mencionado anteriormente, a poda através da freqüência deve ser seguida com cautela, visto que termos que aparecem poucas vezes ou apenas uma vez em um texto podem ser mais relevantes, para o domínio, do que termos mais freqüentes. Foi possível verificar tal afirmação através da utilização da medida *Log-Likelihood*, onde termos com menor freqüência no corpus do domínio foram considerados mais relevantes do que termos com maior freqüência, quando comparados com o corpus de referência. O especialista teve a mesma percepção e optou por não realizar uma poda pela definição de um limiar.

6.2.1.4 *Excluir/Incluir termos:*

Esse passo foi muito importante na execução deste estudo de caso. Neste ponto o especialista pôde excluir termos extraídos que não considerou relevantes para o domínio do Turismo. Dos 412 termos resultantes apresentados ao especialista, 362 foram excluídos, ou seja, 50 termos foram considerados relevantes, correspondendo a 12,14% dos termos extraídos. Vale ressaltar que não houve interesse do especialista na inclusão de termos além

dos identificados pelo protótipo como relevantes. A Tabela B.1 do Anexo B apresenta os termos selecionados pelo especialista.

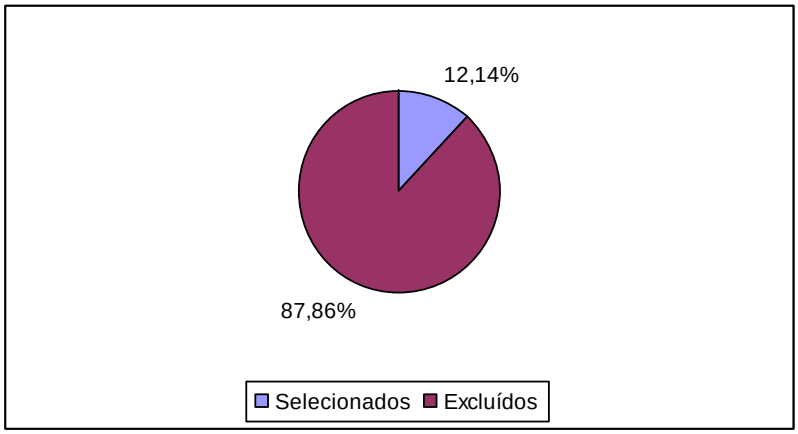

Gráfico 6.3: Distribuição dos candidatos a termos excluídos e selecionados

Os 412 candidatos a termos do domínio foram apresentados ao especialista na ordem de relevância identificada pela medida TFIDF. A Tabela 6.2 apresenta o percentual de termos selecionados pelo especialista em comparação à quantidade de termos extraídos, considerando-se a ordem de relevância.

Tabela 6.2: Percentual de termos extraídos por faixa de relevância

N primeiros termos	Termos selecionados
50	30%
100	52%
150	70%
200	80%
250	96%
300	100%

A Tabela 6.2 nos mostra que mais de 50% dos termos selecionados pelo especialista estavam entre os 100 termos mais relevantes, de acordo com a medida TFIDF, e 80% dos termos selecionados estavam entre os 200 termos mais relevantes, ou seja, na primeira metade dos termos apresentados ao especialista. Considerando o total de termos selecionados pelo especialista, todos estão entre os 300 termos mais relevantes.

6.2.1.5 *Identificar termos compostos a partir da lista de termos relevantes*

Este passo se propôs a identificar termos compostos a partir da lista de termos relevantes selecionados previamente pelo especialista. Foram extraídos 284 termos

compostos, dentre os quais 154 foram selecionados pelo especialista, ou seja, 54,23% dos termos extraídos. A Tabela B.2 do Anexo B apresenta os termos compostos selecionados pelo especialista. Na Tabela 6.3 podemos verificar os resultados obtidos por cada regra utilizada na identificação de termos compostos. E os resultados obtidos diferem bastante entre as regras, tanto na quantidade de termos extraídos quanto na quantidade de termos selecionados. A regras 5, por exemplo, não extraiu nenhum termo composto. Já as regras 1 e 2, por exemplo, foram responsáveis pela extração de poucos termos e as regras 8 e 9 extraíram um número maior de termos. Em relação aos termos selecionados pelo especialista, as regras também tiveram diferentes resultados. Algumas regras não tiveram termo algum selecionado e outras tiveram um grande número de termos selecionados como, por exemplo, a regra 8, que também obteve o melhor percentual de termos selecionados entre as regras utilizadas. Já a regra 9 foi responsável pelo maior número de termos extraídos (55,28% do total extraído) e também pelo maior número de termos selecionados (57,14% do total de termos selecionados), ou seja, uma quantidade maior que a soma dos resultados obtidos nas demais regras.

Tabela 6.3: Termos compostos extraídos *versus* selecionados

Nro	Regra	Extraídos	Selecionados	% selecionados
1	_SU _AJ _PR _AD _SU _AJ	4	2	50,00
2	_SU _AJ _PR _AD _SU	4	0	0,00
3	_SU _PR _AD _SU _AJ	4	0	0,00
4	_SU _PR _AD _SU	40	14	35,00
5	_SU _AJ _PR _SU _AJ	0	0	0,00
6	_SU _AJ _PR _SU	4	2	50,00
7	_SU _PR _SU _AJ	12	7	58,33
8	_SU _PR _SU	59	41	69,49
9	_SU _AJ	157	88	56,05
Total		**284**	**154**	**54,23**

O Gráfico 6.4 apresenta a distribuição dos termos compostos selecionados pelo especialista de acordo com a regra pelos quais foram extraídos.

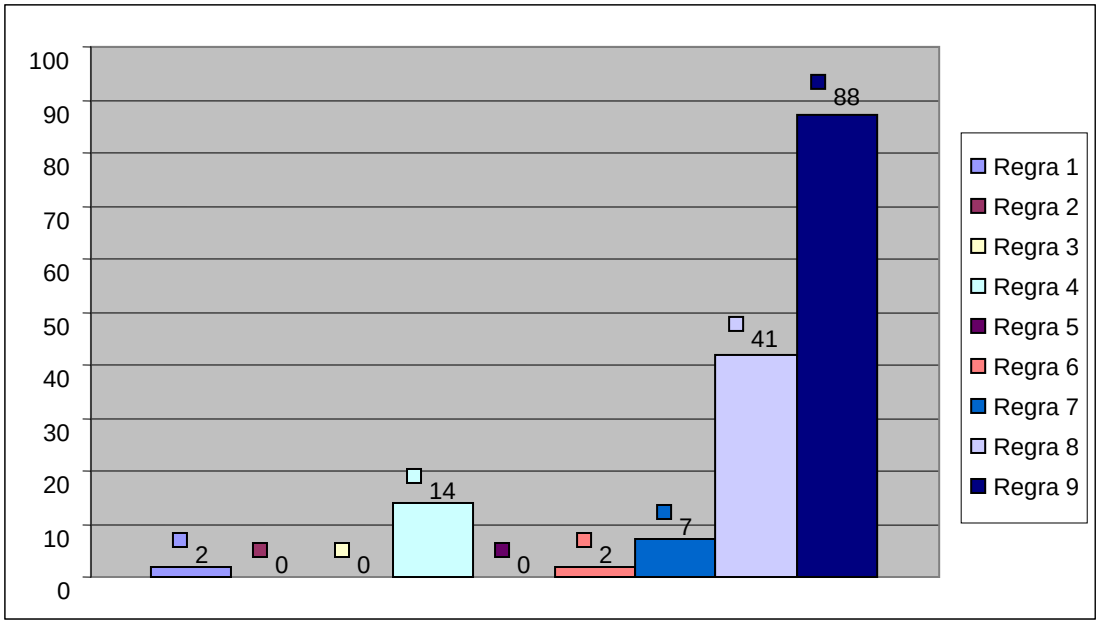

Gráfico 6.4: Distribuição dos termos compostos selecionados pelo especialista

Com exceção da regra 9, as demais regras têm como base uma preposição (_PR), sendo que a mesma pode assumir 3 diferentes valores ("de", "da" e "do"). Assim, analisamos como se comportaram as regras para cada uma dessas preposições. A Tabela 6.4 apresenta a relação entre termos extraídos e termos selecionados, de acordo com a preposição pela qual foram extraídos.

Tabela 6.4: Termos compostos extraídos e selecionados de acordo com a preposição

	Extraídos				**Selecionados**				
Regra	**da**	**do**	**de**	**Total**	**da**	**do**	**de**	**Total**	**%**
_SU _AJ _PR _AD _SU _AJ	3	1	-	**4**	2	0	-	**2**	**50,00**
_SU _AJ _PR _AD _SU	1	3	-	**4**	0	0	-	**0**	**0,00**
_SU _PR _AD _SU _AJ	1	3	-	**4**	0	0	-	**0**	**0,00**
_SU _PR _AD _SU	18	22	-	**40**	7	7	-	**14**	**35,00**
_SU _AJ _PR _SU _AJ	-	-	0	**0**	-	-	0	**0**	**0,00**
_SU _AJ _PR _SU	-	-	4	**4**	-	-	2	**2**	**50,00**
_SU _PR _SU _AJ	-	-	12	**12**	-	-	7	**7**	**58,33**
_SU _PR _SU	-	-	59	**59**	-	-	41	**44**	**69,49**
	23	**29**	**75**		**9**	**7**	**50**		**51,97**
	127				**66**				

Podemos verificar na Tabela 6.4 que o número de termos compostos extraídos utilizando a preposição "de" é bem maior que o número de termos compostos extraídos para

as preposições “da” e “do”, chegando a 59,05% do total de termos extraídos. O mesmo pode ser dito em relação aos termos selecionados pelo especialista, onde o número de termos selecionados com a preposição “de” é maior que a soma do número de termos selecionados para as preposições “da” e “do”, chegando 75,75% dos termos selecionados entre as regras que utilizam preposição como base. O Gráfico 6.5 apresenta a distribuição dos termos selecionados, com relação à preposição pela qual foram extraídos.

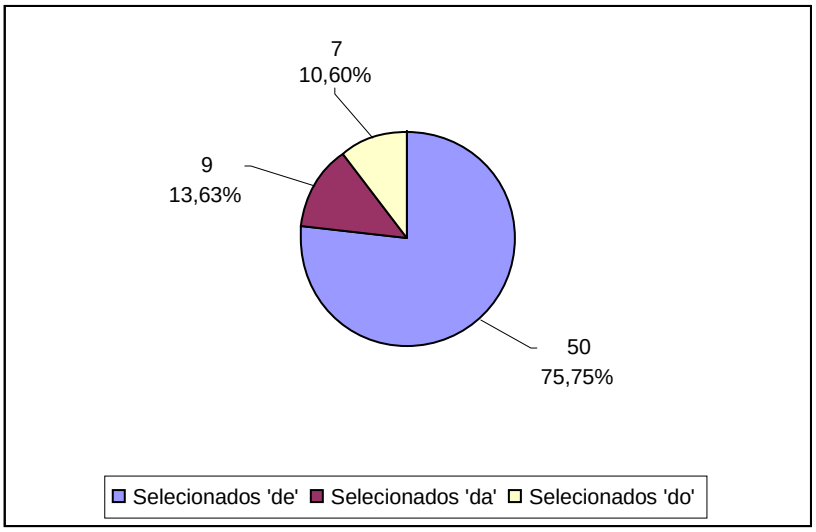

Gráfico 6.5: Distribuição dos termos compostos selecionados de acordo com a preposição

6.2.2 Extração de relações taxonômicas

A segunda fase da abordagem refere-se à identificação de relações taxonômicas entre os termos relevantes (simples e compostos) derivados da fase anterior. Nas subseções seguintes são apresentados os resultados referentes a cada passo desta etapa. A Tabela B.3 do Anexo B apresenta os termos compostos selecionados pelo especialista.

6.2.2.1 *Identificar relações taxonômicas com base em termos compostos*

Esse passo buscou identificar relações taxonômicas a partir do núcleo do sintagma de termos compostos, relacionando cada termo composto ao termo relevante que faz parte da sua composição. Nesta etapa foram extraídas 284 relações taxonômicas. Das relações extraídas, o especialista selecionou 152, o que representa 53,52% do total extraído. Apesar de ser um método grosseiro, esta regra para identificação de relações taxonômicas foi a que obteve o

melhor resultado dentre as regras utilizadas. O Gráfico 6.6 apresenta a proporção entre relações selecionadas e relações excluídas.

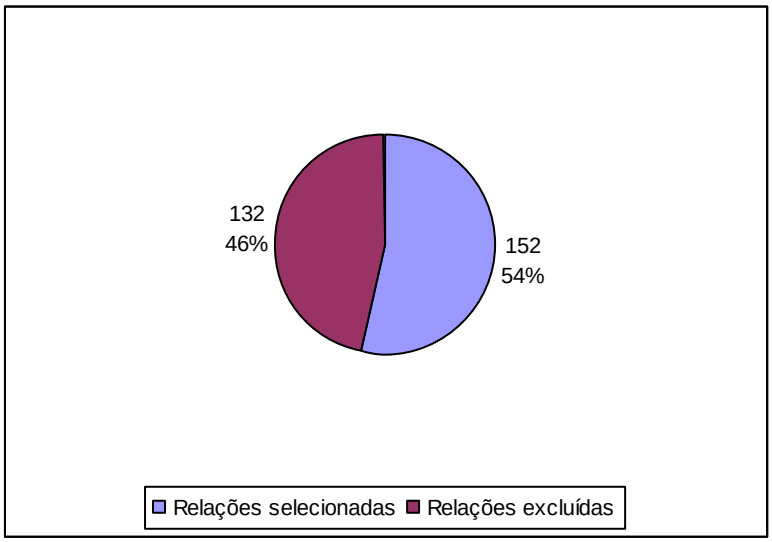

Gráfico 6.6: Proporção entre relações taxonômicas selecionadas e excluídas

6.2.2.2 *Identificar relações taxonômicas através dos padrões de Hearst*

Apesar do grande número de padrões utilizados nesta etapa, para identificação de relações taxonômicas, foram extraídas pela ferramenta apenas oito relações, sendo que apenas uma delas foi selecionada pelo especialista. A Tabela 6.5 apresenta os padrões de Hearst (adaptados), a partir dos quais as oito relações foram extraídas.

Tabela 6.5: Relações taxonômicas pelos padrões de Hearst (adaptados)

	Padrões de Hearst adaptados	Extraídas	Selecionadas
1	SUB como {(SUB,)*(ou|e)} SUB	7	1
2	SUB {, SUB}* {,} ou outro(s) SUB	1	0

6.2.2.3 *Identificar relações taxonômicas através dos padrões de Morin e Jacquemin*

Neste passo foram extraídas pela ferramenta apenas quatro relações taxonômicas e nenhuma delas foi selecionada pelo especialista como relevante. A Tabela 6.6 apresenta os padrões de Morin e Jacquemin (adaptados), a partir dos quais as relações foram extraídas.

Tabela 6.6: Relações taxonômicas pelos padrões de Morin e Jacquemin (adaptados)

	Padrões de Morin e Jacquemin adaptados	Extraídas	Selecionadas
1	SUB1 (LIST_SUB2)	4	0

6.2.3 Geração da estrutura ontológica

Nesta fase o objetivo foi utilizar os termos e relações taxonômicas selecionados pelo especialista para gerar uma estrutura ontológica na linguagem de representação ontológica denominada OWL. O resultado deste passo foi a criação de um arquivo OWL, que pode ser utilizado em editores de ontologias como Protégé, permitindo ao especialista do domínio continuar o desenvolvimento da ontologia.

6.3 Estudo de caso 2

A execução do segundo estudo de caso seguiu basicamente as mesmas etapas do primeiro, porém sem a intervenção do usuário no que se refere à exclusão de termos e relações no decorrer do processo. Como foi utilizado o mesmo corpus nos dois estudos de caso, os resultados dos primeiros passos são exatamente os mesmos e, portanto, não serão repetidos neste segundo estudo de caso. Aqui serão apresentados os resultados a partir do passo relativo à primeira intervenção do especialista no primeiro estudo de caso, ou seja, a exclusão/inclusão de termos (item 6.4.4 do estudo de caso 1).

Ao final do processo, coube ao especialista selecionar os termos e as relações taxonômicas relevantes extraídas pela ferramenta. A diferença, neste caso, é que os dados foram apresentados ao especialista uma única vez e em maior quantidade, pois não ocorreu exclusão de termos no decorrer do processo. Por fim gerou-se a estrutura ontológica em OWL.

Nas seções seguintes detalhamos os resultados obtidos na execução deste estudo de caso. Apesar de os termos e relações terem sido validados somente no final do processo, para melhor compreensão e visualização os resultados das validações pelo especialista serão apresentados juntamente com os resultados referentes à extração dos termos e relações.

6.3.1 Identificação de termos

6.3.1.1 *Excluir/Incluir termos*

Ao contrário do primeiro estudo de caso, no qual termos não relevantes foram excluídos neste passo, neste estudo de caso os 412 termos resultantes dos passos anteriores foram utilizados nas etapas subseqüentes como termos relevantes ao domínio. Somente no final do processo o especialista selecionou os 50 termos tidos como relevantes ao domínio do Turismo.

6.3.1.2 *Identificar termos compostos a partir da lista de termos relevantes*

Este passo se propôs a identificar termos compostos a partir dos 412 termos relevantes resultantes do passo anterior. Foram extraídos 1.245 termos compostos, dentre os quais 203 foram selecionados pelo especialista ao final do processo, ou seja, 16,31% dos termos extraídos. Na Tabela 6.7 podemos verificar que existem regras com poucos ou nenhum termo composto extraído (regras 1 e 5) e regras a partir das quais foi extraído um número maior de termos (regras 8 e 9).

Na regra 1, por exemplo, temos poucos termos extraídos, porém com maior precisão. Por outro lado temos a regra 3 a partir da qual foram extraídos 23 termos e nenhum foi selecionado pelo especialista. Já a regra 9 foi responsável por um grande número de termos extraídos (52,93% do total extraído) e também por um grande número de termos selecionados (56,65% do total de termos selecionados).

Tabela 6.7: Termos compostos extraídos *versus* selecionados

Nro	Regra	Extraídos	Selecionados	% selecionados
1	_SU _AJ _PR _AD _SU _AJ	5	2	40,00
2	_SU _AJ _PR _AD _SU	24	1	4,17
3	_SU _PR _AD _SU _AJ	23	0	0,00
4	_SU _PR _AD _SU	220	22	10,00
5	_SU _AJ _PR _SU _AJ	0	0	0,00
6	_SU _AJ _PR _SU	16	2	12,50
7	_SU _PR _SU _AJ	39	8	20,51
8	_SU _PR _SU	259	53	20,46
9	_SU _AJ	659	115	17,45
Total		**1245**	**203**	**16,31**

O Gráfico 6.7 apresenta a distribuição dos termos compostos selecionados pelo especialista de acordo com a regra pelos quais foram extraídos.

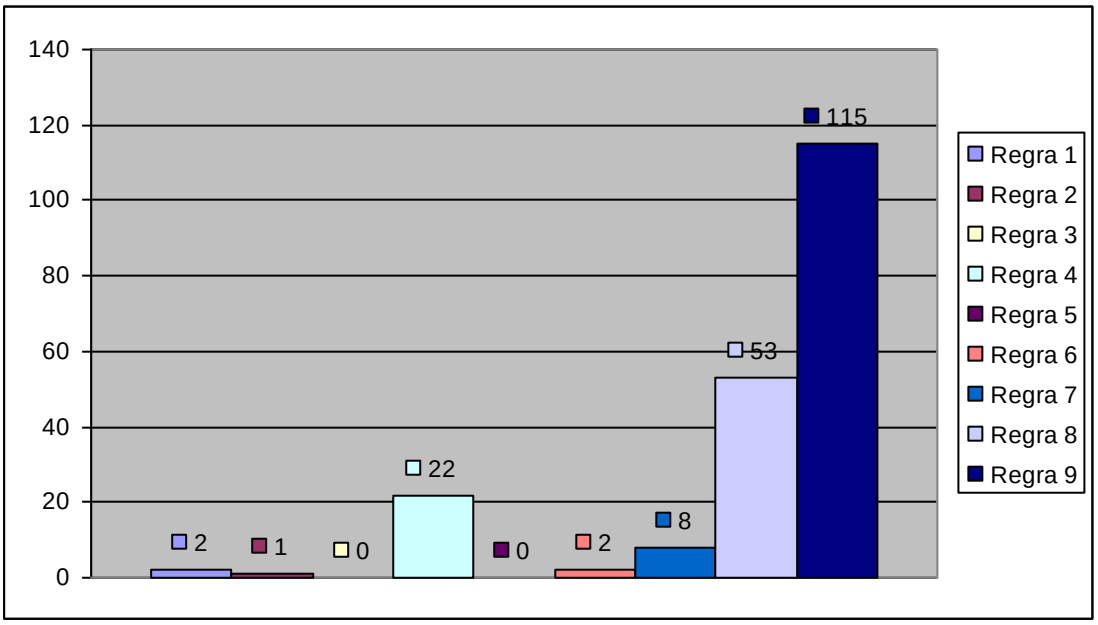

Gráfico 6.7: Distribuição dos termos compostos selecionados pelo especialista

Assim como no primeiro estudo de caso, com exceção da regra 9, as demais regras têm como base uma preposição (_PR), que pode assumir 3 diferentes valores (“de”, “da” e “do”). A Tabela 6.8 apresenta a relação entre termos extraídos e termos selecionados de acordo com a preposição pela qual foram extraídos.

Tabela 6.8: Termos compostos extraídos *versus* selecionados de acordo com a preposição

	Extraídos				**Selecionados**			
Regra	**da**	**do**	**de**	**Total**	**da**	**do**	**de**	**Total**
_SU _AJ _PR _AD _SU _AJ	4	1	-	**5**	2	0	-	**2**
_SU _AJ _PR _AD _SU	14	10	-	**24**	0	1	-	**1**
_SU _PR _AD _SU _AJ	12	11	-	**23**	0	0	-	**0**
_SU _PR _AD _SU	102	118	-	**220**	11	11	-	**22**
_SU _AJ _PR _SU _AJ	-	-	0	**0**	-	-	0	**0**
_SU _AJ _PR _SU	-	-	16	**16**	-	-	2	**2**
_SU _PR _SU _AJ	-	-	39	**39**	-	-	8	**8**
_SU _PR _SU	-	-	259	**259**	-	-	53	**53**
	132	**140**	**314**	**586**	**13**	**12**	**63**	**88**
					9,85%	**8,57%**	**20,06%**	

Podemos verificar na Tabela 6.8 que o número de termos compostos extraídos utilizando a preposição “de” é bem maior do que aquele para as preposições “da” e “do”,

chegando a 53,58% do total de termos compostos extraídos. O mesmo pode ser dito em relação aos termos selecionados pelo especialista, onde o número de termos com preposição "de" é maior que a soma do número de termos selecionados para as preposições "da" e "do", chegando a 71,59% dos termos selecionados entre as regras que utilizam preposição como base. O Gráfico 6.8 apresenta a distribuição dos termos selecionados, de acordo a preposição pela qual foram extraídos.

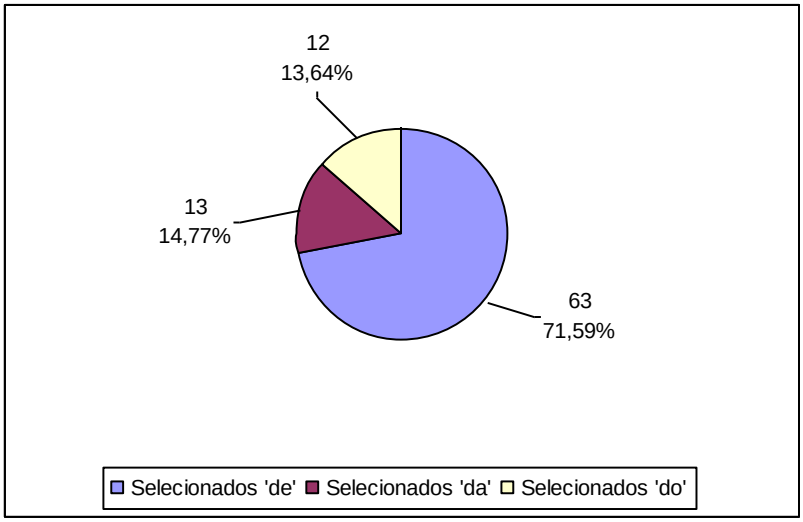

Gráfico 6.8: Distribuição dos termos compostos selecionados de acordo com a preposição

6.3.2 Extração de relações taxonômicas

A segunda fase da abordagem refere-se à identificação de relações taxonômicas entre os termos relevantes derivados da fase anterior. Nas subseções seguintes são apresentados os resultados referentes a cada passo desta etapa.

6.3.2.1 *Identificar relações taxonômicas com base em termos compostos*

Este passo buscou identificar relações taxonômicas a partir do núcleo do sintagma de termos compostos, relacionando cada termo composto ao termo relevante que faz parte da sua composição. Nesta etapa foram extraídas 1.104 relações taxonômicas tendo como base os termos compostos da Tabela 6.7. Das relações extraídas, o especialista selecionou 161, o que representa 14,58%.

Tabela 6.9: Relações taxonômicas baseada nos termos compostos

	Regra	Extraídas	Selecionadas
1	Relações taxonômicas baseada em termos compostos	1.104	161

6.3.2.2 *Identificar relações taxonômicas através dos padrões de Hearst*

Neste passo foram extraídas pela ferramenta apenas 17 relações taxonômicas e apenas uma delas foi selecionada pelo especialista, representando 5,89%. A Tabela 6.10 apresenta os padrões de Hearst (adaptados) a partir das quais as relações foram extraídas.

Tabela 6.10: Relações taxonômicas pelos padrões de Hearst (adaptados)

	Tradução/Adaptação	Extraídas	Selecionadas
1	SUB como {(SUB,)*(ou|e)} SUB	12	1
2	SUB {, SUB}* {,} ou outro(s) SUB	5	0

6.3.2.3 *Identificar relações taxonômicas através dos padrões de Morin e Jacquemin*

Neste passo foram extraídas pela ferramenta apenas 13 relações taxonômicas e nenhum foi selecionada pelo especialista como relevante. A Tabela 6.11 apresenta os padrões de Morin e Jacquemin (adaptados), a partir dos quais as regras foram extraídas.

Tabela 6.11: Relações taxonômicas pelos padrões de Morin e Jacquemin (adaptados)

	Padrões de Morin e Jacquemin adaptados	Extraídas	Selecionadas
1	SUB1 (LIST_SUB2)	13	0

6.3.3 Geração da estrutura ontológica

Como mencionado no primeiro estudo de caso, este passo serviu para gerar uma estrutura ontológica em linguagem de representação ontológica com base nos termos e relações taxonômicas selecionados pelo especialista. O resultado deste passo foi a criação de um arquivo OWL, que pode ser utilizado em editores de ontologias como Protégé, permitindo ao especialista do domínio continuar o desenvolvimento da ontologia.

6.4 Considerações quanto à Análise dos resultados

Esta seção apresenta uma análise sobre o desempenho da abordagem proposta, discutindo sobre resultados obtidos, identificando problemas e levantando possíveis causas. A idéia é analisar a eficiência da abordagem no que diz respeito a sua capacidade em identificar termos e relações taxonômicas corretamente. Nesse contexto, foi objetivo dos estudos de caso apresentados neste capítulo possibilitar uma avaliação do modo como a abordagem proposta auxilia na identificação de estruturas ontológicas.

Um ponto importante a ser considerado, como descrito no inicio do capítulo, é o fato de que textos de jornal, devido a sua característica, não são considerados específicos de um domínio, mas semi-especializados para o domínio. Assim, por não ser um corpus especializado do domínio do Turismo, os textos utilizados no estudo de caso não contêm apenas termos relacionados ao domínio do Turismo. Dessa forma, a utilização do corpus nos estudos de caso pode ter gerado um resultado não tão preciso quanto se fosse utilizado um corpus específico do domínio.

Durante a execução dos primeiros passos, que foram iguais para os dois estudos de caso, alguns dados nos chamaram a atenção. O primeiro diz respeito ao grande número de palavras que não representam conceitos no domínio, chegando a quase 70% do corpus utilizado. Outro dado que merece atenção foi o grande número de nomes próprios encontrados no texto, pois nomes próprios podem representar instâncias (indivíduos) de um domínio. A identificação dessas instâncias e das classes de uma ontologia às quais estão relacionadas pode gerar outro trabalho interessante.

A utilização da medida *Log-Likelihood* foi muito importante para os resultados obtidos nos estudos de caso. Com o uso da medida foram excluídos muitos termos sem significado para o domínio do Turismo, pelo menos quando comparados ao corpus de referência. Para se ter uma idéia, em uma execução como a do segundo estudo de caso se, ao invés da medida *Log-Likelihood*, utilizássemos apenas a medida TFIDF, obteríamos um total de 3.308 candidatos a termo relevante ao invés dos 412 retornados com o uso da medida *Log-Likelihood*. E isso traria conseqüências para as etapas seguintes. A Tabela 6.12 apresenta os dados resultantes de uma execução como a do segundo estudo de caso, porém apenas com a utilização da medida TFIDF. Como podemos ver na tabela, a diferença de termos e relações extraídas é relativamente grande quando comparadas aos resultados obtidos com o uso da medida *Log-Likelihood*.

Tabela 6.12: Resultados com uso da medida Log-Likelihood e uso da medida TFIDF

	Log-Likelihood	TFIDF
Simples	412	3308
Compostos	1245	3606
Relações por Compostos	1245	3306
Relações por Hearst	14	28
Relações por Morin	11	35

Um ponto importante a ser considerado é o fato de que o tamanho do corpus utilizado é pequeno, e que um corpus maior irá conter possivelmente muito mais termos e relações do que os apresentados na Tabela 6.12.

A utilização da medida TFIDF para apresentar os candidatos a termo relevante 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 exclusão de termos através da definição de um limiar, como proposto na abordagem, não foi executada nos estudos de caso apresentados neste capítulo, devido à possibilidade de termos com grande relevância serem excluídos por possuírem um peso menor. Talvez o tamanho do corpus utilizado nos estudos de caso não tenha gerado a necessidade de uma exclusão por limiar, mas em um corpus maior este passo pode vir a ser importante, mesmo que alguns termos relevantes possam vir a ser perdidos com sua utilização. Este passo também poderia ser utilizado em uma possível execução automática da abordagem, semelhante ao que ocorreu no segundo estudo de caso. Porém seriam necessários vários testes até encontrar um bom ponto de corte.

O baixo número de termos selecionados pelo especialista (12,14% dos termos extraídos) nos leva a entender que uma solução totalmente automatizada e com alto grau de precisão não seja viável quando se utilizando apenas de técnicas estatísticas para identificação de termos. Esse baixo resultado pode ser conseqüência do uso de um corpus semi-especializado como o utilizado nos estudos de caso.

A identificação de termos compostos obteve um bom resultado no primeiro estudo de caso, visto que o especialista selecionou 57% dos termos extraídos. Dentre as regras utilizadas, tem grande destaque a regra 9 (_SU _AJ) que 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.

Outro ponto importante na identificação de termos compostos está relacionado às regras que utilizam preposição como base. Como visto anteriormente, a preposição aqui modelada pode assumir três diferentes valores (“de”, “da” e “do”). Os resultados do primeiro estudo de caso nos mostraram que a preposição “de” foi responsável por 59% dos termos

extraídos e 77% dos termos selecionados. Analisando os termos extraídos com uso das preposições “da” e “do”, é possível observar que grande parte deles poderia ser identificada como atributos de outros termos extraídos.

Já no segundo estudo de caso o número de termos compostos extraídos foi bem maior e o número de termos selecionados pelo especialista também aumentou (41 termos a mais). Da mesma forma que o estudo de caso anterior, algumas regras continuaram sem termos extraídos ou selecionados neste estudo de caso, e a regra 9 continuou sendo a responsável pela maioria dos termos extraídos e selecionados. A mesma observação quanto aos termos compostos identificados por regras que utilizam preposição como base foi confirmada no segundo estudo de caso.

Quanto às relações taxonômicas identificadas no primeiro estudo de caso, foram observados resultados bons e ruins. A identificação de relações taxonômicas com base em termos compostos, apesar de ser um método relativamente grosseiro, teve um grau de acerto de 53%. Já os padrões adaptados de Hearst e também de Morin e Jacquemin tiveram um resultado muito abaixo do esperado, tanto na extração de termos quanto na seleção dos termos pelo especialista. Isso pode ser devido à característica de escrita dos textos.

No segundo estudo de caso, a identificação de relações taxonômicas por meio dos termos compostos resultou basicamente na mesma quantidade de termos obtidos no primeiro estudo de caso, ou seja, um maior número de termos compostos gerados não resultou em um ganho significativo. Da mesma forma, as relações identificadas através dos padrões adaptados de Hearst e também de Morin e Jacquemin tiveram um resultado muito abaixo do esperado.

Como pôde ser visto no decorrer do capítulo, apenas a medida Precisão (*Precision*) foi utilizada para análise dos resultados. Outras medidas como *Recall* e *Accuracy* poderiam ser interessantes nesta análise, mas para sua utilização seriam necessários dados sobre acerto e erro dentre os termos não selecionados pela ferramenta. Estes dados porém não foram produzidos pelo especialista. Apenas os termos selecionados foram validados.

7 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as considerações finais quanto ao trabalho realizado, descrevendo suas principais contribuições e limitações. Ainda, destaca rumos para futuras pesquisas na área.

Para o desenvolvimento do trabalho, inicialmente realizou-se um embasamento teórico sobre o que são ontologias, suas classificações e sua aplicação em diferentes áreas. Este embasamento nos propiciou constatar que o tema ‘ontologia’ tem sido abordado em várias pesquisas, em diversas áreas, confirmando sua importância no contexto tecnológico atual.

Posteriormente, o estudo restringiu-se às alternativas para a construção de ontologias a partir de textos, onde um dos problemas motivadores para esta pesquisa, a falta de uma abordagem automática ou semi-automática para a construção de ontologias a partir de textos da língua portuguesa do Brasil, foi identificado.

Como descrito no trabalho, apesar da variedade de relações entre palavras que podem ser encontradas em um corpus, nossa escolha em lidar com relações taxonômicas se deve ao fato de que, de acordo com [RUI05], na maioria dos casos ontologias são estruturadas como hierarquias de conceitos (taxonomias).

Frente ao exposto, nosso objetivo neste trabalho foi propor uma abordagem para semi-automatizar passos do processo de aquisição de estruturas ontológicas a partir de textos escritos na língua portuguesa do Brasil, mais especificamente as fases de extração de conceitos e relações taxonômicas.

Um protótipo de software foi desenvolvido com o objetivo principal de validar a abordagem proposta para identificação de estruturas ontológicas. Com auxílio do protótipo foi realizado um estudo de caso sobre o domínio do Turismo. Os resultados obtidos a partir da aplicação da abordagem sobre o corpus do Turismo foram considerados relevantes à pesquisa, não se tendo noticia de outro trabalho que tivesse realizado essa tarefa. No entanto, algumas considerações precisam ser feitas.

Um ponto importante a ser considerado é o fato de que textos de jornal, como os utilizados no estudo de caso, não são considerados textos especializados de um domínio, mas

sim textos semi-especializados. Assim, a utilização desse corpus pode ter gerado alguma distorção nos resultados, tanto na identificação de termos quanto na identificação de relações. Para obtermos resultados mais precisos, será necessário um estudo de caso com corpus específico de um domínio.

Outra consideração importante diz respeito ao tamanho do corpus utilizado para o estudo de caso. Um corpus contendo em torno de 100 mil palavras é considerado pequeno, mas relevante, para um estudo de caso como o executado neste trabalho. Porém, a quantidade de termos e relações resultantes da aplicação da abordagem proposta em um corpus maior (e específico) pode vir a ser considerável, e um novo estudo de caso nesse sentido poderá ser desenvolvido, dando continuidade à pesquisa.

Durante a realização deste trabalho, como já descrito, foram realizados dois estudos de caso, cujo resultados estão sintetizados na Tabela 7.1. O objetivo destes estudos de caso foi analisar a eficiência da abordagem no que diz respeito a sua capacidade em identificar corretamente termos e relações taxonômicas.

No primeiro estudo de caso os resultados de cada etapa foram avaliados pelo especialista. Isto significa que apenas os dados selecionados como corretos em uma etapa, serviram de entrada e direcionaram os resultados das etapas seguintes. Já no segundo estudo de caso, os dados foram apresentados ao especialista somente ao final do processo, sem que termos e relações fossem excluídos entre as etapas. Conseqüentemente, como pode ser visto na Tabela 7.1, mais termos e relações foram extraídos. A exclusão dos dados irrelevantes pelo especialista ocorreu apenas ao final do processo.

Como pôde ser visto nos resultados dos estudos de caso, a maior parte do corpus do domínio, quase 70% do total de palavras no que se refere à possibilidade de serem termos relevantes, foi desconsiderado já na primeira etapa da abordagem proposta. Consideradas apenas as *stopwords,* estas já representam mais de 50% do corpus. Logo após a eliminação das palavras não relevantes ao domínio, as palavras restantes tiveram sua freqüência no corpus do domínio comparada a sua freqüência no corpus de referência, com a utilização da medida Log-Likelihood. Esse processo resultou na exclusão, de forma automática, de 3.635 diferentes substantivos não específicos ao domínio do Turismo, ou seja, substantivos que aparecem em maior proporção no corpus de referência. Desta forma restaram apenas 412 candidatos a termos relevantes do domínio. Como pode ser visto na Tabela 7.1, o especialista

selecionou, entre estes, apenas 50 termos, correspondendo a 12,14% dos termos extraídos (ou ainda a 0,06% do total do corpus).

Tabela 7.1 Comparação entre os estudos de caso

	Estudo de Caso 1			**Estudo de Caso 2**		
Fase	**Extraídos**	**Selecionados**		**Extraídos**	**Selecionados**	
Termos relevantes simples	412	50	12,14%	412	50	12,14%
Termos compostos	284	154	54,23%	1245	203	16,31%
Relações baseadas nos termos compostos	154	152	98,70%	1245	161	12,93%
Relações pelos padrões de Hearst	8	1	12,5%	17	1	5,88%
Relações pelos padrões de Morin	4	0	0%	13	0	0%

No primeiro estudo de caso, fase de identificação de termos compostos, foram extraídos 284 termos compostos a partir dos 50 termos simples selecionados pelo especialista. Destes termos compostos identificados, 154 foram selecionados pelo especialista, ou seja, 54,23% dos termos extraídos. Dentre as regras utilizadas para essa identificação, se destaca a regra _SU _AJ (o termo é formado por um substantivo seguido de um adjetivo), sendo esta responsável por mais de 50% dos termos extraídos e também dos termos selecionados. No segundo estudo de caso, como não houve exclusão de termos pelo especialista na primeira etapa, o número de termos compostos selecionados foi muito superior (1245 termos compostos).

No que diz respeito às relações taxonômicas identificadas no primeiro estudo de caso, foram observados resultados bons e também resultados ruins. Durante a identificação de relações taxonômicas a partir dos termos compostos, foram extraídas 284 relações taxonômicas, sendo que 152 destas foram selecionadas pelo especialista, representando 53,52% do total extraído. Apesar de ser este um método não muito sofisticado, a identificação de relações taxonômicas baseada em termos compostos foi a que obteve o melhor resultado dentre as regras utilizadas. Já no segundo estudo de caso pode ser visto que o número de termos e relações extraídas foi bem maior, pois não ocorreram cortes durante o processo.

Já a identificação de relações taxonômicas, seja através dos padrões de Hearst seja pelos padrões de Morin e Jacquemin, não trouxe resultados significativos nos dois estudos de

caso. Houve poucas relações identificadas, e apenas uma selecionada pelo especialista, em cada estudo de caso. Este resultado pode ser consequência da característica da linguagem empregada na escrita dos textos.

Apesar de serem relevantes, os resultados dos estudos de caso nos levam a acreditar que uma solução totalmente automatizada, não levaria a um alto grau de precisão, quando se utilizando apenas técnicas estatísticas para identificação de termos e relações taxonômicas.

7.1 Contribuições

A abordagem proposta, juntamente com o protótipo desenvolvido, constitui a principal contribuição deste trabalho, auxiliando na identificação de termos e relações taxonômicas, oferecendo apoio ao engenheiro de ontologia em fases importantes do processo de construção. Em uma visão mais detalhada, as contribuições deste trabalho são:

- Levantamento e análise de abordagens e técnicas para identificação de termos e relações taxonômicas e sua aplicação ao português;
- Mecanismos que auxiliam a identificação de conceitos relevantes a partir de um corpus de domínio específico;
- Mecanismos que auxiliam a identificação de relações taxonômicas entre os conceitos extraídos previamente;
- Mecanismos que auxiliam a criação de uma estrutura ontológica a partir dos conceitos e relações identificadas previamente e;
- O protótipo desenvolvido como uma contribuição prática que possibilita validar a abordagem proposta.

7.2 Limitações e trabalhos futuros

As principais limitações da pesquisa são mencionadas a seguir:

- Utilização de um corpus que, devido a suas características, é considerado semi-especializado;
- Dependência de um corpus anotado como entrada para a abordagem;

- Utilização de um corpus de tamanho limitado por falta de um *parser* que permitisse realizar a etiquetagem de textos do domínio;
- Validação dos resultados dos estudos de caso realizada por apenas um especialista do domínio;
- Realização de estudos de caso para apenas um domínio.

A partir do trabalho apresentado e das limitações colocadas acima é possível identificar novos trabalhos como continuidade da pesquisa:

- Ampliar o processo de construção de modo a torná-lo recorrente, permitindo a ampliação do número de níveis da ontologia;
- Enriquecer ontologias: utilizar uma ontologia existente para, a partir de um corpus, adicionar novos conceitos e relações;
- Utilizar um dicionário de sinônimos evitando que termos com mesmo significado sejam identificados como conceitos diferentes;
- Identificar outros tipos de relações, como relações “parte-de”, entre outras citadas no decorrer do trabalho;
- Descobrir automaticamente novos padrões que indiquem relacionamento entre termos;
- Incorporar um etiquetador e um lematizador à ferramenta, possibilitando a utilização de corpora que não estejam marcados;
- Identificar restrições e propriedades para conceitos de uma ontologia;
- Identificar instâncias e relacioná-las a conceitos da ontologia;
- Desenvolver um estudo de caso com corpus específico de um domínio;
- Executar automaticamente um número maior de tarefas.

REFERÊNCIAS

[AGI01] AGIRRE, E.; ANSA, O.; MARTINEZ, D.; HOVY, E. **Enriching WordNet concepts with topic signatures**. In: Proceedings of the SIGLEX workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations, in conjunction with NAACL, Pittsburg, vol. 0109031, 2001. Capturado em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.2151>. Último acesso: 23 de Julho de 2006.

[BAS00] BASILI, R.; PAZIENZA, M.T.; ZANZOTTO, F. **Customizable Modular Lexicalized Parsing.** In: Proceedings of the Int. Workshop on Parsing Technology, Povo (Trento), 2000. Capturado em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.3.1330>. Último acesso: 23 de Julho de 2006.

[BAS04a] BASEGIO, T. L. **Um estudo sobre ontologias e seu uso na engenharia de software.** Trabalho Individual I, Mestrado em Ciência da Computação, PUCRS, Porto Alegre, 46 páginas, 2004.

[BAS04b] BASEGIO, T. L. **Aquisição de conhecimento a partir de textos para a construção de ontologias.** Trabalho Individual II, Mestrado em Ciência da Computação, PUCRS, Porto Alegre, 48 páginas, 2004.

[BAS96] BASILI, R.; PAZIENZA, M.T.; VELARDI, P. **An Empyrical Symbolic Approach to Natural Language Processing**. In: Artificial Intelligence, vol. 85, p. 59-99, 1996.

[BIC00] BICK, E. **The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.** PH. D. thesis, Arhus University Press, 505 páginas, 2000.

[BLA98] BLÁZQUEZ, M.; FERNÁNDEZ, M.; GARCÍA-PINAR, J. M.; PEREZ, A. G. **Building Ontologies at the Knowledge Level** Using the Ontology Design Environment**.** In: Proceedings of the Knowledge Acquisition Workshop, KAW98, Banff, Canada, p. 29-30, 1998.

[BOU02] BOURIGAULT, D. **Analyse distributionnelle étendue.** In: Proceedings of Traitement Automatique des Langues, Nancy, France, p. 75-84, 2002.

[BRE03] BREWSTER, C.; CIRAVEGNA, F.; WILKS, Y. **Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance.** In: Proceedings of the Semantic Web Workshop, SIGIR, Toronto, Canada, p. 150-158, 2003.

[BUI04] BUITELAAR, P.; OLEJNIK, D.; SINTEK, M. **A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis**. In: Proceedings of the European Semantic Web Symposium (ESWS), Heraklion, Greece, vol. 3053, p. 31-44, 2004.

[CEL04] ČELJUSKA, D. **Semi-automatic Construction of Ontologies from Text**. Master's Thesis, Department of Artificial Intelligence and Cybernetics, Technical University Košice, 2004.

[CHA99] CHANDRASEKARAN, B.; JOSEPHSON, J.R.; and BENJAMINS, V.R. **What are ontologies, and why do we need them?** IEEE Intelligent Systems, vol. 14, n.1, p. 20-26, 1999.

[CHU89] CHURCH, K. W.; HANKS, P. **Word association norms, mutual information, and lexicography.** In: Proceedings of the 27th conference on Association for Computational Linguistics , Vancouver, British Columbia, Canada, p. 76-83, 1989.

[DAV02] DAVIES, J.; FENSEL, D.; VAN HARMELEN, F. **Towards the semantic web: ontology-driven knowledge management.** John Wiley & Sons Ltd., England, 310 páginas, 2002.

[DEC02] DECLERCK, T. **A set of tools for integrating linguistic and non-linguistic information**. In: Proceedings of the SAAKM workshop at ECAI, Lyon, volume 100, 2002. Capturado em: < ftp://ftp.informatik.rwth-aachen.de/pub/publications/CEUR-WS/>. Último acesso: 12 de Novembro de 2006.

[DEG04] DEGERATU, M.; HATZIVASSILOGLOU, V. **An Atomatic Method for Constructing Domain-Specific Ontology Resources**. In: Proceedings of the Language Resources and Evaluation Conference (LREC2004), Lisbon, Portugal, p. 2001-2004, 2004.

[EVE02] EVERETT, J. O.; *et al*. **Making Ontologies Work for Resolving Redundancies Across Documents.** Communications of the ACM, vol. 45, n. 2, p. 55-60, 2002.

[FAL98] FALBO, R. A.; MENEZES, C. S.; ROCHA, A. R. C. **A Systematic approach for Building Ontologies**. In: Proceedings of the IBERAMIA 98, Lisboa, Portugal, vol. 1484, p. 349-360, 1998.

[FAL00] FALBO, R.A.; DUARTE, K.C. **Uma Ontologia de Qualidade de Software**. Anais do VII Workshop de Qualidade de Software, XIV Simpósio Brasileiro de Engenharia de Software, João Pessoa, Paraíba, Brasil, p. 275-285, 2000.

[FAL02a] FALBO, R.A.; BERTOLLO, G.; RUY, F. B.; MIAN, P. G.; PEZZIN, J.; SHWAMBACH, M.M.; NATALI, A.C.C. **ODE - Um Ambiente de Desenvolvimento de Software Baseado em Ontologias.** Anais do XVI Simpósio Brasileiro de Engenharia de Software - SBES'2002. Caderno de Ferramentas, Gramado - RS, Brasil, p. 438-443, 2002.

[FAL02b] FALBO, R. A.; GUIZZARDI, G.; DUARTE, K. C. **An ontological approach to domain engineering.** In**:** 14th International Conference on Software Engineering and Knowledge Engineering, Ischia, Italy, vol. 27, p. 351-358, 2002.

[FAR97] FARQUHAR, A.; FIKES, R.; RICE, J. **The Ontolingua Server: a Tool for Collaborative Ontology Construction**. International Journal of Human-Computer Studies, vol. 46, n. 6, p. 707-727, 1997.

[FEN97] FENSEL, D. *et al*. **Using ontologies for defining tasks, problem-solving methods and their mappings**. In: Proceedings of the 10th European Workshp in Knowledge Acquisition, Modeling and Management, EKAW, Saint Felin de Guixols, Catalonia, vol. 1319, p.113-128, 1997.

[FEN03] FENSEL, D.; HENDLER, J.; LIEBERMAN, H.; WAHLSTER, W. **Spinning the Semantic Web**: bringing the world wide web to its full potential**.** Cambridge: The Mit Press, 479 páginas, 2003.

[FER97] FERNÁNDEZ, M.; GÓMEZ-PÉREZ, A.; JURISTO, N. **Methontology: From Ontological Art towards Ontological Engineering**. In: Proceedings of Workshop on Ontological Engineering: AAAI97 Spring Symposium Series, Stanford, CA, p. 33-40, 1997.

[GRU02] GRUNINGER, M.; LEE, J. **Ontology: Applications and Design.** Communications of the ACM, vol. 45, n. 2, p. 39-41, 2002.

[GRU93a] GRUBER, T. R. **Towards Principles for the Design of Ontologies Used for Knowledge Sharing.** International Journal of Human and Computer Studies, vol. 43, n. 5/6, p. 907-928, 1993.

[GUA98] GUARINO, N. **Formal ontology and Information Systems.** In Proceedings of 1st International Conference on Formal Ontology in Information Systems (FOIS'98), IOS Press, Trento, Itália, p. 3-15, 1998.

[HEA92] HEARST, M. A. **Automatic acquisition of hyponyms from large text corpora**. In: Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, p. 539-545, 1992.

[HEI97] HEIJST, G. van; SCHREIBER, A. T.; WIELINGA, B. J. **Using explicit ontologies in KBS development.** International Journal of Human-Computer Studies, vol. 46, n. 2/3, p. 183-292, 1997.

[HOL02] HOLSAPPLE, C. W.; JOSHI, K. D. **A Collaborative Approach to Ontology Design.** Communications of the ACM, vol. 45, n. 2, p. 42-47, 2002.

[HOV03] HOVY, E. **Using an Ontology to Simplify Data Access.** Communications of the ACM, vol. 46, n. 1, p. 47-49, 2003.

[JEN05] **JENA - A Semantic Web Framework for Java**. Capturado em:

<http://jena.sourceforge.net/>. Último acesso: 23 de Outubro de 2005.

[JUR00] JURAFSKY, D. S. **Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition**. Upper Saddle River, New Jersey, Prentice Hall, 934 páginas, 2000.

[KAL02] KALFOGLOU, Y.; ALANI, H.; O'HARA, K.; SHADBOLT, N. **Initiating Organizational Memories Using Ontology Network Analysis**. In: Proceedings of ECAI 2002, Workshop on Knowledge Management and Organizational Memories (W-5, KMOM), Lyon, France, p. 79-89, 2002.

[KIS04] KISHORE, R.; ZHANG, H.; RAMESH, R. **A Helix-Spindle Model for Ontological Engineering.** Communications of the ACM, vol. 47, n. 2, p. 69-75, 2004.

[LAM03] LAME, G. **Using text analysis techniques to identify legal ontologies' components**. In: Proceedings of the Workshop on Legal Ontologies of the International Conference on Artificial Intelligence and Law, Edinburgh, UK, Junho, 24-28, 2003.

[MAE00] MAEDCHE, A.; STAAB, S. **Discovering conceptual relations from text**. In: Proceedings of the 14th European Conference on Artificial Intelligence: ECAI-2000. IOS Press, Amsterdam, p. 321-325, 2000.

[MAE02] MAEDCHE, A. **Ontology Learning for the Semantic Web.** Massachusetts: Kluwer Academic Publishers, 272 páginas, 2002.

[MOR03] MORIN, E.; JACQUEMIN, C. **Automatic acquisition and expansion of hypernym links**. Computer and the humanities, Kluwer Academic Press, vol. 38, n.4, p. 33, 2003

[NEC91] NECHES R, FIKES, R.; FININ, T.; GRUBER T.R.; SENATOR, T.; SWARTOUT, W.R. **Enabling technology for knowledge sharing**. AI Magazine vol. 12, n.3, p. 36–56, 1991.

[NOY01] NOY, N. F.; McGUINNESS, D. L. **Ontology Development 101: A Guide to Creating Your First Ontology**. Technical Report KSL-01-05 and SMI-2001-0880, Stanford Knowledge Systems Laboratory and Stanford Medical Informatics, March 2001.

[OLI02] OLIVEIRA, M.C.; PICADA, R.C.; GONZALEZ, M.A.I. **Pré-etiquetador de Categorias Gramaticais Orientado a Morfologia.** Trabalho de Conclusão (Graduação em Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, 94páginas, 2002.

[PER04] PÉREZ, C. **Aquisição de conhecimento a partir de textos para a construção de mapas conceituais**. Dissertação (Mestrado em Computação Aplicada)–Universidade do Vale do Rio dos Sinos, São Leopoldo, 85p,2004.

[PER99] PÉREZ, A. G.; BENJAMINS, V. R. **Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods.** Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends (IJCAI99), Estocolmo, vol. 18, p. 1.1–1.15, 1999.

[PRO05] **PROTÉGÉ - Ontology Editor and Knowledge Acquisition System** (2005). Capturado em: <http://protege.stanford.edu/>. Úlimo acesso em Julho de 2005.

[RAY04] RAYSON P.; BERRIDGE D.; FRANCIS B. **Extending the Cochran rule for the comparison of word frequencies between corpora.** In Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium, vol. II, p. 926-936, 2004.

[REY97] REYNAR, J. C.; RATNAPARKHI A. **A Maximum Entropy Approach to Identifying Sentence Boundaries.** In: Proceedings of the Fifth Conference on Applied Natural Language Processing , Washington, D.C, p. 16-19, 1997.

[RUI05] RUIZ-CASADO, M.; ALFONSECA, E.; CASTELLS, P. **Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia.** In: Proceedings of NLDB-2005: Natural Language Processing and Information Systems, v. 3513, p. 67-79, Alicante, Spain, 2005

[RUS03] RUSSEL, S.; NORVIG, P. **Artificial Intelligence – A Modern Approach.** Prentice Hall, 2003.

[STA01] STAAB, S.; STUDER, R.; SCHNURR, H.; SURE, Y. **Knowledge Process and Ontologies.** IEEE Inteligent Systems, p. 26-34, 2001.

[STU98] STUDER, R.; BENJAMINS, V. R.; FENSEL, D. **Knowedge Engineering: Principles e Methods**. Data & Knowedge Engineering, vol. 25, p.161-197, 1998.

[USC95] USCHOLD, M.; KING, M. **Towards a Methodology for Building Ontologies.** In: IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Candá, 1995. Capturado em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.5357>. Último acesso: 05 de Novembro de 2006.

[USC98] USCHOLD, M.; CLARK, P.; HEALY, M.; WILLIAMSON, K.; WOODS, S. **An Experiment in Ontology Reuse**. In: Proceedings of the Knowledge Acquisition Workshop, KAW98, Banff, Canada, 1998. Capturado em: < http://ksi.cpsc.ucalgary.ca/KAW/KAW98/uschold/ >. Úlimo acesso em Julho de 2005.

[VEL01] VELARDI, P.; PAOLO, F.; MISSIKOFF M. **Using text processing techniques to automatically enrich a domain ontology.** In: Proceedings of the International Conference on Formal Ontology in Information Systems – FOIS. Ogunquit, Maine, USA, v. 2001, p. 270-284, 2001.

[VER03] VERGARA, J. E. L. *et al*. **Ontologies Giving Semantics to Network Management.** IEEE Network, special issue on Networks Management, vol. 17, n. 3, p. 15-21, 2003.

[W3C05] "OWL - Web Ontology Language".

Capturado em: <http://www.w3.org/2004/OWL/>. Último acesso: Outubro de 2005.

APÊNDICE A - Lista de Stopwords

Artigos: o, a, os, as, um, uma, uns, umas.

Pronomes pessoais: eu, tu, ele, ela, nós, vós, eles, elas, me, mim, comigo, te, ti, contigo, se, si, lhe, o, a, nos, conosco, vos, lhes, os, as.

Pronomes possessivos: meu, minha, meus, minhas, teu, tua, teus, tuas, seu, sua, seus, suas, nosso, nossa, nossos, nossas, vosso, vossa, vossos, vossas.

Pronomes demonstrativos: este, estes, esta, estas, esse, esses, essa, essas, aquele, aqueles, aquela, aquelas, mesmo, mesmos, mesma, mesmas, próprio, próprios, própria, próprias, tal, tais, isto, isso, aquilo, o, a, os, as.

Pronomes indefinidos: algo, alguém, nada, ninguém, outrem, quem, tudo, cada, certo, certos, certa, certas, algum, alguns, alguma, algumas, bastante, demais, mais, menos, muito, muitos, muita, muitas, nenhum, nenhuns, nenhuma, nenhumas, outro, outros, outra, outras, pouco, poucos, pouca, poucas, qualquer, quaisquer, qual, que, quanto, quantos, quanta, quantas, tal, tais, tanto, tantos, tanta, tantas, todo, todos, toda, todas, um, uns, uma, umas, vários, várias.

Pronomes relativos: cujo, cujos, cuja, cujas, quanto, quantos, quanta, quantas, quem, que, onde.

Pronomes interrogativos: que, quê, quem, qual, quantos, quantas.

Advérbios: sim, deveras, talvez, quiçá, acaso, porventura, decerto, muito, pouco, assaz, bastante, mais, menos, tão, demasiado, meio, todo, demais, que, quão, quanto, quase, como, abaixo, acima, acolá, cá, lá, aqui, ali, aí, além, aquém, algures, alhures, nenhures, atrás, fora, afora, dentro, perto, longe, adiante, diante, onde, avante, através, defronte, aonde, donde, bem, mal, assim, depressa, devagar, debalde, alerta, melhor, pior, não, tampouco, agora, hoje, amanhã, depois, ontem, anteontem, já, sempre, amiúde, nunca, jamais, ainda, logo, antes, cedo, tarde, ora, outrora, então, absolutamente, breve, calmamente, certamente, corretamente, efetivamente, fielmente, levemente, possivelmente, primeiramente, provavelmente, quiçá, realmente, tanto, tarde, ultimamente.

Palavras e locuções denotativas: eis, exclusive, menos, exceto, fora, salvo, senão, sequer, inclusive, também, mesmo, ainda, até, ademais, além disso, de mais a mais, só, apenas, é que, sobretudo, embora, aliás, ou, melhor, isto é, ou antes, a saber, por exemplo, ou seja, afinal, agora, então, mas.

Preposição: a, ante, após, até, com, contra, de, desde, em, entre, para, per, perante, por, sem, sob, sobre, trás, conforme, segundo, durante, mediante, visto, como.

Combinações e contrações com preposições: ao, aos, aonde, à, às, àquele, àquela, àquilo, de, dele, deles, dela, delas, deste, deste, desta, destas, disto, daqui, dentre, nesse, nesses, nessa, nessas, no, nos, na, nas, num, nuns, naquele, naquela, naqueles, naquelas, pelo, pela, pelos, pelas.

Conjunções: e, nem, mas também, mas ainda, senão, também, como também, bem como, mas, porém, todavia, contudo, entretanto, senão, ao passo que, antes, no entanto, não obstante,

apesar disso, em todo caso, ou, logo, portanto, por conseguinte, pois, por isso, que, porque, porquanto.

Conjunções subordinativas: porque, que, pois, como, porquanto, visto que, visto como, já que, uma vez que, desde que, como, qual, tal qual, tal e qual, assim como, como, tal como, tão, como, tanto como, que, mais que, mais do que, menos que, menos do que, quanto, tanto quanto, que nem, feito, o mesmo que, embora, conquanto, que, ainda que, mesmo que, posto que, por mais que, por menos que, se bem que, nem que, dado que, se caso, contanto que, desde que, salvo se, sem que, a não ser que, a menos que, como, conforme, segundo, consoante, de sorte, que, de modo que, de forma que, de maneira que, sem que, para que, a fim de que, à proporção que, à medida que, ao passo que, quando, enquanto, logo que, mal, sempre que, assim que, desde que, antes que, depois que, até agora, agora que, que, se.

APÊNDICE B – Resultados do Estudo de Caso 1

Apêndice B.1: Termos simples selecionados pelo especialista

Nro	Termo	TFIDF
1	praia	260.479546
2	cidade	204.028135
3	diaria	184.46659
4	parque	155.630608
5	pacote	150.722702
6	restaurante	148.39507
7	turista	144.749421
8	viagem	143.007709
9	passeio	133.873062
10	voo	132.725961
11	visitante	101.353533
12	turismo	98.9808879
13	roteiro	92.7604274
14	passagem	88.7293419
15	agencia	86.6349087
16	museu	81.0030323
17	reveillon	76.7681548
18	compra	75.1657781
19	guia	72.6234082
20	paisagem	72.5910448
21	pousada	71.4127096
22	resort	69.8302002
23	litoral	68.6154906
24	passageiro	65.4912386
25	viajante	65.4616505

Nro	Termo	TFIDF
26	onibus	60.8904488
27	aeroporto	58.2198209
28	temporada	57.8459263
29	excursao	57.5761161
30	traslado	55.781683
31	categoria	54.0620734
32	embarque	52.5709519
33	esqui	52.3273747
34	lazer	52.043675
35	cruzeiro	47.2701395
36	aventura	43.6608257
37	navio	43.2496587
38	trem	39.4282139
39	tarifa	37.3766962
40	rota	36.6672767
41	cassino	35.0263827
42	culinaria	33.8099467
43	complexo	32.5931348
44	passaporte	32.5931348
45	cambio	32.437244
46	tour	32.437244
47	artesanato	31.3771967
48	destino	28.8331058
49	hospede	27.2427421
50	polo	24.4448511

Apêndice B.2: Termos compostos selecionados pelo especialista

Nro.	Termo composto	Regra
1	roteiro ecologico	_SU _AJ
2	hotel africano	_SU _AJ
3	turismo local	_SU _AJ
4	voo eletrico	_SU _AJ
5	passeio pitoresco	_SU _AJ
6	viagem internacional	_SU _AJ
7	viagem interessante	_SU _AJ
8	turista novo	_SU _AJ
9	resort grande	_SU _AJ
10	voo inaugural	_SU _AJ
11	viajante europeu	_SU _AJ
12	turista ocidental	_SU _AJ
13	roteiro cultural	_SU _AJ
14	hotel unico	_SU _AJ
15	turista preciso	_SU _AJ
16	cambio oficial	_SU _AJ
17	parque religioso	_SU _AJ
18	turismo interno	_SU _AJ
19	turista estrangeiro	_SU _AJ
20	culinaria local	_SU _AJ
21	turismo ecologico	_SU _AJ
22	passeio inusitado	_SU _AJ
23	culinaria regional	_SU _AJ
24	hotel excelente	_SU _AJ
25	parque marinho	_SU _AJ
26	hotel central	_SU _AJ
27	viagem gratuito	_SU _AJ
28	paisagem belo	_SU _AJ
29	paisagem virgem	_SU _AJ
30	cambio competitive	_SU _AJ
31	categoria turistico	_SU _AJ
32	turismo passageiro	_SU _AJ
33	viagem poeirento	_SU _AJ
34	voo curto	_SU _AJ
35	turismo academic	_SU _AJ
36	tour europeu	_SU _AJ
37	voo livre	_SU _AJ
38	voo direto	_SU _AJ
39	aeroporto local	_SU _AJ
40	agencia brasileiro	_SU _AJ
41	viagem aereo	_SU _AJ
42	passeio verde	_SU _AJ
43	roteiro gastronomic	_SU _AJ
44	turismo regional	_SU _AJ
45	esqui alpino	_SU _AJ
46	resort familiar	_SU _AJ
47	turista norte-americano	_SU _AJ

48	hotel novo	_SU _AJ
49	hotel caro	_SU _AJ
50	voo noturno	_SU _AJ
51	voo seminal	_SU _AJ
52	turista alternative	_SU _AJ
53	voo regular	_SU _AJ
54	hotel official	_SU _AJ
55	voo extra	_SU _AJ
56	passeio noturno	_SU _AJ
57	hotel maior	_SU _AJ
58	tarifa promocional	_SU _AJ
59	lazer principal	_SU _AJ
60	viagem seguro	_SU _AJ
61	paisagem local	_SU _AJ
62	culinaria peninsular	_SU _AJ
63	passeio circular	_SU _AJ
64	destino principal	_SU _AJ
65	polo charmoso	_SU _AJ
66	roteiro complete	_SU _AJ
67	trem turistico	_SU _AJ
68	turismo conveniente	_SU _AJ
69	voo regional	_SU _AJ
70	tour asiatico	_SU _AJ
71	voo ultimo	_SU _AJ
72	viagem barato	_SU _AJ
73	tarifa inaugural	_SU _AJ
74	tarifa normal	_SU _AJ
75	viagem gratis	_SU _AJ
76	resort luxuoso	_SU _AJ
77	tour insolito	_SU _AJ
78	viagem delicioso	_SU _AJ
79	destino novo	_SU _AJ
80	viagem semelhante	_SU _AJ
81	destino tradicional	_SU _AJ
82	viagem dificil	_SU _AJ
83	hotel concorrente	_SU _AJ
84	roteiro cronologico	_SU _AJ
85	polo turistico	_SU _AJ
86	roteiro exclusive	_SU _AJ
87	polo importante	_SU _AJ
88	turismo aereo	_SU _AJ
89	agencia de ecoturismo	_SU _PR _SU
90	agencia de viagens	_SU _PR _SU
91	categoria de preco	_SU _PR _SU
92	compra de produto	_SU _PR _SU
93	guia de bolso	_SU _PR _SU
94	hotel de lazer	_SU _PR _SU
95	hotel de selva	_SU _PR _SU
96	hotel de sonho	_SU _PR _SU
97	navio de cruzeiro	_SU _PR _SU
98	pacote de ecoturismo	_SU _PR _SU

99	pacote de feria	_SU _PR _SU
100	pacote de reveillon	_SU _PR _SU
101	passageiro de excursao	_SU _PR _SU
102	passeio de barco	_SU _PR _SU
103	passeio de catamaran	_SU _PR _SU
104	passeio de escuna	_SU _PR _SU
105	passeio de lancha	_SU _PR _SU
106	passeio de teste	_SU _PR _SU
107	passeio de treno	_SU _PR _SU
108	polo de esqui	_SU _PR _SU
109	polo de turismo	_SU _PR _SU
110	roteiro de carro	_SU _PR _SU
111	roteiro de ecoturismo	_SU _PR _SU
112	roteiro de esqui	_SU _PR _SU
113	roteiro de lazer	_SU _PR _SU
114	roteiro de viagem	_SU _PR _SU
115	tour de carro	_SU _PR _SU
116	tour de treno	_SU _PR _SU
117	traslado de ida	_SU _PR _SU
118	turismo de alemao	_SU _PR _SU
119	turismo de aventura	_SU _PR _SU
120	turismo de pesca	_SU _PR _SU
121	viagem de barco	_SU _PR _SU
122	viagem de camelo	_SU _PR _SU
123	viagem de carro	_SU _PR _SU
124	viagem de feria	_SU _PR _SU
125	viagem de onibus	_SU _PR _SU
126	viagem de turismo	_SU _PR _SU
127	viagem de verao	_SU _PR _SU
128	voo de garca	_SU _PR _SU
129	voo de para-quedas	_SU _PR _SU
130	agencia de turismo local	_SU _PR _SU _AJ
131	compra de passagem aereo	_SU _PR _SU _AJ
132	hotel de categoria turistico	_SU _PR _SU _AJ
133	paisagem de agua cristalino	_SU _PR _SU _AJ
134	passagem de classe economic	_SU _PR _SU _AJ
135	polo de atracao turistico	_SU _PR _SU _AJ
136	turismo de pesca tradicional	_SU _PR _SU _AJ
137	paisagem pedregoso de montanha	_SU _AJ _PR _SU
138	temporada oficial de viagem	_SU _AJ _PR _SU
139	aeroporto de a capital	_SU _PR _AD _SU
140	compra de a passage	_SU _PR _AD _SU
141	diaria de o hotel	_SU _PR _AD _SU
142	hospede de o hotel	_SU _PR _AD _SU
143	hotel de a rede	_SU _PR _AD _SU
144	hotel de a regiao	_SU _PR _AD _SU
145	hotel de o interior	_SU _PR _AD _SU
146	lazer de a cidade	_SU _PR _AD _SU
147	lazer de o hotel	_SU _PR _AD _SU
148	lazer de o turista	_SU _PR _AD _SU
149	pacote de a agencia	_SU _PR _AD _SU

150	paisagem de a ilha	_SU _PR _AD _SU
151	roteiro de o pacote	_SU _PR _AD _SU
152	viagem de o aficionado	_SU _PR _AD _SU
153	roteiro gastronomico de a capital cearense	_SU _AJ _PR _AD _SU _AJ
154	voo regular de a companhia aereo	_SU _AJ _PR _AD _SU _AJ

Apêndice B.3: Relações taxonômicas selecionadas pelo especialista

Nro.	Termo 1	Termo 2	Regra
1	aeroporto	aeroporto local	Baseada em Termos compostos
2	aeroporto	aeroporto de a capital	Baseada em Termos compostos
3	agencia	agencia de ecoturismo	Baseada em Termos compostos
4	agencia	agencia de turismo local	Baseada em Termos compostos
5	agencia	agencia brasileiro	Baseada em Termos compostos
6	agencia	agencia de viagens	Baseada em Termos compostos
7	cambio	cambio competitivo	Baseada em Termos compostos
8	cambio	cambio oficial	Baseada em Termos compostos
9	categoria	categoria de preco	Baseada em Termos compostos
10	categoria	categoria turistico	Baseada em Termos compostos
11	compra	compra de passagem aereo	Baseada em Termos compostos
12	compra	compra de a passagem	Baseada em Termos compostos
13	compra	compra de produto	Baseada em Termos compostos
14	culinaria	culinaria local	Baseada em Termos compostos
15	culinaria	culinaria regional	Baseada em Termos compostos
16	destino	destino principal	Baseada em Termos compostos
17	destino	destino novo	Baseada em Termos compostos
18	destino	destino tradicional	Baseada em Termos compostos
19	diaria	diaria de hotel	Baseada em Termos compostos
20	esqui	esqui alpino	Baseada em Termos compostos
21	guia	guia de bolso	Baseada em Termos compostos
22	hospede	hospede de o hotel	Baseada em Termos compostos
23	hotel	hotel maior	Baseada em Termos compostos
24	hotel	hotel oficial	Baseada em Termos compostos
25	hotel	hotel novo	Baseada em Termos compostos
26	hotel	hotel de sonho	Baseada em Termos compostos
27	hotel	hotel central	Baseada em Termos compostos
28	hotel	hotel excelente	Baseada em Termos compostos
29	hotel	hotel caro	Baseada em Termos compostos
30	hotel	hotel de a rede	Baseada em Termos compostos
31	hotel	hotel unico	Baseada em Termos compostos
32	hotel	hotel de a regiao	Baseada em Termos compostos
33	hotel	hotel de lazer	Baseada em Termos compostos
34	hotel	hotel de categoria turistico	Baseada em Termos compostos
35	hotel	hotel concorrente	Baseada em Termos compostos
36	hotel	hotel africano	Baseada em Termos compostos
37	hotel	hotel de selva	Baseada em Termos compostos
38	hotel	hotel de o interior	Baseada em Termos compostos
39	lazer	lazer de a cidade	Baseada em Termos compostos
40	lazer	lazer de o hotel	Baseada em Termos compostos
41	lazer	lazer principal	Baseada em Termos compostos
42	lazer	lazer de o turista	Baseada em Termos compostos
43	navio	navio de cruzeiro	Baseada em Termos compostos
44	pacote	pacote de a agencia	Baseada em Termos compostos
45	pacote	pacote de ecoturismo	Baseada em Termos compostos
46	pacote	pacote de feria	Baseada em Termos compostos
47	pacote	pacote de reveillon	Baseada em Termos compostos

48	paisagem	paisagem pedregoso de montanha	Baseada em Termos compostos
49	paisagem	paisagem de a ilha	Baseada em Termos compostos
50	paisagem	paisagem belo	Baseada em Termos compostos
51	paisagem	paisagem virgem	Baseada em Termos compostos
52	paisagem	paisagem local	Baseada em Termos compostos
53	paisagem	paisagem de agua cristalino	Baseada em Termos compostos
54	parque	parque religioso	Baseada em Termos compostos
55	passageiro	passageiro de excursao	Baseada em Termos compostos
56	passagem	passagem de classe economico	Baseada em Termos compostos
57	passeio	passeio inusitado	Baseada em Termos compostos
58	passeio	passeio de treno	Baseada em Termos compostos
59	passeio	passeio de escuna	Baseada em Termos compostos
60	passeio	passeio pitoresco	Baseada em Termos compostos
61	passeio	passeio de catamara	Baseada em Termos compostos
62	passeio	passeio de barco	Baseada em Termos compostos
63	passeio	passeio de teste	Baseada em Termos compostos
64	passeio	passeio de lancha	Baseada em Termos compostos
65	passeio	passeio verde	Baseada em Termos compostos
66	passeio	passeio circular	Baseada em Termos compostos
67	passeio	passeio noturno	Baseada em Termos compostos
68	polo	polo de esqui	Baseada em Termos compostos
69	polo	polo de atracao turistico	Baseada em Termos compostos
70	polo	polo charmoso	Baseada em Termos compostos
71	polo	polo importante	Baseada em Termos compostos
72	polo	polo de turismo	Baseada em Termos compostos
73	polo	polo turistico	Baseada em Termos compostos
74	resort	resort familiar	Baseada em Termos compostos
75	resort	resort luxuoso	Baseada em Termos compostos
76	resort	resort grande	Baseada em Termos compostos
77	roteiro	roteiro gastronomico	Baseada em Termos compostos
78	roteiro	roteiro exclusivo	Baseada em Termos compostos
79	roteiro	roteiro gastronomico de a capital cearense	Baseada em Termos compostos
80	roteiro	roteiro de lazer	Baseada em Termos compostos
81	roteiro	roteiro de esqui	Baseada em Termos compostos
82	roteiro	roteiro de o pacote	Baseada em Termos compostos
83	roteiro	roteiro cronologico	Baseada em Termos compostos
84	roteiro	roteiro de carro	Baseada em Termos compostos
85	roteiro	roteiro completo	Baseada em Termos compostos
86	roteiro	roteiro cultural	Baseada em Termos compostos
87	roteiro	roteiro de viagem	Baseada em Termos compostos
88	roteiro	roteiro de ecoturismo	Baseada em Termos compostos
89	roteiro	roteiro ecologico	Baseada em Termos compostos
90	tarifa	tarifa inaugural	Baseada em Termos compostos
91	tarifa	tarifa normal	Baseada em Termos compostos
92	tarifa	tarifa promocional	Baseada em Termos compostos
93	temporada	temporada oficial de viagem	Baseada em Termos compostos
94	tour	tour de treno	Baseada em Termos compostos
95	tour	tour asiatico	Baseada em Termos compostos
96	tour	tour de carro	Baseada em Termos compostos
97	tour	tour insolito	Baseada em Termos compostos
98	tour	tour europeu	Baseada em Termos compostos

99	traslado	traslado de ida	Baseada em Termos compostos
100	trem	trem turistico	Baseada em Termos compostos
101	turismo	turismo de alemao	Baseada em Termos compostos
102	turismo	turismo interno	Baseada em Termos compostos
103	turismo	turismo regional	Baseada em Termos compostos
104	turismo	turismo academico	Baseada em Termos compostos
105	turismo	turismo conveniente	Baseada em Termos compostos
106	turismo	turismo de aventura	Baseada em Termos compostos
107	turismo	turismo ecologico	Baseada em Termos compostos
108	turismo	turismo passageiro	Baseada em Termos compostos
109	turismo	turismo aereo	Baseada em Termos compostos
110	turismo	turismo de pesca tradicional	Baseada em Termos compostos
111	turismo	turismo local	Baseada em Termos compostos
112	turismo	turismo de pesca	Baseada em Termos compostos
113	turista	turista norte-americano	Baseada em Termos compostos
114	turista	turista preciso	Baseada em Termos compostos
115	turista	turista alternativo	Baseada em Termos compostos
116	turista	turista novo	Baseada em Termos compostos
117	turista	turista ocidental	Baseada em Termos compostos
118	turista	turista estrangeiro	Baseada em Termos compostos
119	viagem	viagem barato	Baseada em Termos compostos
120	viagem	viagem de verao	Baseada em Termos compostos
121	viagem	viagem dificil	Baseada em Termos compostos
122	viagem	viagem gratis	Baseada em Termos compostos
123	viagem	viagem semelhante	Baseada em Termos compostos
124	viagem	viagem delicioso	Baseada em Termos compostos
125	viagem	viagem seguro	Baseada em Termos compostos
126	viagem	viagem de barco	Baseada em Termos compostos
127	viagem	viagem aereo	Baseada em Termos compostos
128	viagem	viagem poeirento	Baseada em Termos compostos
129	viagem	viagem de carro	Baseada em Termos compostos
130	viagem	viagem de o aficionado	Baseada em Termos compostos
131	viagem	viagem de camelo	Baseada em Termos compostos
132	viagem	viagem gratuito	Baseada em Termos compostos
133	viagem	viagem interessante	Baseada em Termos compostos
134	viagem	viagem internacional	Baseada em Termos compostos
135	viagem	viagem de turismo	Baseada em Termos compostos
136	viagem	viagem de feria	Baseada em Termos compostos
137	viagem	viagem de onibus	Baseada em Termos compostos
138	viajante	viajante europeu	Baseada em Termos compostos
139	voo	voo de garca	Baseada em Termos compostos
140	voo	voo regular de a companhia aereo	Baseada em Termos compostos
141	voo	voo eletrico	Baseada em Termos compostos
142	voo	voo de para-quedas	Baseada em Termos compostos
143	voo	voo curto	Baseada em Termos compostos
144	voo	voo livre	Baseada em Termos compostos
145	voo	voo direto	Baseada em Termos compostos
146	voo	voo noturno	Baseada em Termos compostos
147	voo	voo semanal	Baseada em Termos compostos
148	voo	voo regular	Baseada em Termos compostos
149	voo	voo extra	Baseada em Termos compostos

150	voo	voo regional	Baseada em Termos compostos
151	voo	voo ultimo	Baseada em Termos compostos
152	voo	voo inaugural	Baseada em Termos compostos
153	turismo	fonte	Baseado nos padrões de Hearst