UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE
ÁREA DE CIÊNCIA DA COMPUTAÇÃO

LUCAS RÊGO DRUMOND

AQUISIÇÃO AUTOMATIZADA DE HIERARQUIAS DE CONCEITOS DE ONTOLOGIAS UTILIZANDO APRENDIZAGEM ESTATÍSTICA RELACIONAL

São Luís
2009

LUCAS RÊGO DRUMOND

AQUISIÇÃO AUTOMATIZADA DE HIERARQUIAS DE CONCEITOS DE ONTOLOGIAS UTILIZANDO APRENDIZAGEM ESTATÍSTICA RELACIONAL

Dissertação de Mestrado apresentada ao curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Eletricidade, na área de Ciência da Computação.

Orientadora: Profª Drª Rosario Girardi

São Luís

2009

Drumond, Lucas Rêgo.
Aquisição Automatizada de Hierarquias de Conceitos de Ontologias Utilizando Aprendizagem Estatística Relacional / Lucas Rêgo Drumond. – São Luís, 2009.

105 f.
Impresso por computador (fotocópia).
Orientadora: Rosario Girardi.
Dissertção (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia da Eletricidade, 2009.
1. Web – Aquisição de conhecimento. 2. Ontologias – Aprendizagem. 3. Aprendizagem estatística relacional. 4. Redes lógicas de Markov. I. Título.

CDU: 004.775

AQUISIÇÃO AUTOMATIZADA DE HIERARQUIAS DE CONCEITOS DE ONTOLOGIAS UTILIZANDO APRENDIZAGEM ESTATÍSTICA RELACIONAL

Lucas Rêgo Drumond

Dissertação aprovada em 23 de outubro de 2009.

Profa. Maria del Rosário Girardi, Ph. D.
(Orientadora)

Profa. Maria Carolina Monard, Dra.
(Membro da Banca Examinadora)

Prof. Zair Abdelouahab, Ph. D.
(Membro da Banca Examinadora)

Ao meu irmão Rafael.

AGRADECIMENTOS

Em primeiro lugar a Deus, por tudo que me tem concedido nestes anos.

Aos meus pais Alberto e Celi, pela educação, exemplo e suporte que me deram, não só durante a elaboração deste trabalho, mas durante toda minha vida e por serem os melhores pais que alguém pode ter.

Ao meu irmão e amigo Rafael, pela paciência, apoio e bom humor, principalmente nos momentos mais difíceis.

A todos os meus familiares, que sempre foram muito importantes para mim e sempre estiveram ao meu lado.

À Profª Rosario Girardi, por ter aceitado minha participação no grupo de pesquisa, por toda a dedicação com a qual orientou meus trabalhos, por todo o esforço empreendido na minha formação e por ter sempre cobrado o máximo de mim.

Ao meu ex-colega do GESEC e amigo, Leandro cujas ponderações, sugestões e questionamentos muito ajudaram a enriquecer este trabalho.

A todos os meus atuais e ex-colegas do grupo GESEC cujos questionamentos e ponderações foram muito importantes para o desenvolvimento deste trabalho.

A todos os meus valorosos amigos por nunca terem deixado a chama do *rock* se apagar.

À “Baixinha”, pelo café e pelo bom humor que sempre levou ao laboratório do grupo GESEC.

Aqueles que foram, de algum modo, importantes para a elaboração deste estudo e que têm consciência disso.

O ser humano é quase único dentre os animais por ter a capacidade de aprender com a experiência dos outros, e notável também pela sua aparente relutância em fazê-lo.

Douglas Adams

RESUMO

Os formalismos de representação do conhecimento como as ontologias têm se mostrado uma poderosa ferramenta para melhorar a efetividade de sistemas de processamento da linguagem natural, recuperação e filtragem de informação e muitas outras tarefas. Além disso, as ontologias são essenciais para a Web Semântica, uma nova geração da Web que visa estruturar o conteúdo da mesma de modo que este possa ser processado de forma mais efetiva pelas máquinas. Entretanto, os sistemas de conhecimento sofrem do problema conhecido como o gargalo da aquisição do conhecimento, que nada mais é do que a dificuldade de construção das bases de conhecimento. Uma abordagem para este problema é o suporte automático ou semi-automático à construção de ontologias. Este campo de pesquisa é conhecido como aprendizagem de ontologias. Este trabalho discute o estado da arte das técnicas de aprendizagem de ontologias e propõe uma abordagem para o suporte ao processo de construção de ontologias através da automatização da extração de hierarquias de conceitos a partir de fontes textuais. O processo proposto é composto por duas técnicas, a PRECE (*Probabilistic Relational Concept Extraction*) para a extração de conceitos e a PREHE (*Probabilistic Relational Hierarchy Extraction*) para a descoberta de relacionamentos taxonômicos entre os conceitos extraídos pela PRECE. As duas técnicas fazem uso das Redes Lógicas de Markov, uma abordagem da aprendizagem probabilística relacional que combina a lógica de primeira ordem com as redes de Markov. As técnicas PRECE e PREHE foram avaliadas no domínio turístico comparando os seus resultados com uma ontologia desenvolvida manualmente por especialistas neste domínio.

Palavras-chave: Aprendizagem de ontologias. Aquisição de conhecimento. Aprendizagem estatística relacional. Redes lógicas de Markov.

ABSTRACT

Knowledge representation formalisms, such as ontologies, have proven to be a powerful tool for enhancing the effectiveness of natural language processing, information filtering and retrieval and so on. Besides these tasks, ontologies are also crucial for the Semantic Web, a new generation of the Web that aims at structuring its content in such a way that it can be more effectively processed by machines. However, knowledge systems suffer from the so called knowledge acquisition bottleneck, i.e. the difficulty in constructing knowledge bases. An approach for this problem is to provide automatic or semi-automatic support for ontology construction, a field of research known as ontology learning. This work discusses the state of the art of ontology learning techniques and proposes and approach for supporting the ontology construction process through the automatization of the concept hierarchy extraction from textual sources. The proposed process is composed by two techniques, namely PRECE (Probabilistic Relational Concept Extraction) and PREHE (Probabilistic Relational Hierarchy Extraction). The PRECE technique extracts ontology concepts from textual sources while the PREHE technique extracts taxonomic relationships between the concepts extracted by PRECE. Both techniques use Markov logic networks, an approach for statistical relational learning that combines first order logic with Markov networks. The PRECE and PREHE techniques were evaluated in the touristic domain and their results were compared with an ontology manually developed by a domain expert.

Keywords: Ontology learning. Knowledge acquisition. Statistical relational learning. Markov logic networks.

LISTA DE FIGURAS

Figura 1 - Exemplo de ontologia 19
Figura 2 - Modelo do Processo de aprendizagem de ontologias baseado na aprendizagem de máquina 22
Figura 3 - Camadas do Processo de Desenvolvimento de Ontologias adaptadas de Buitelaar, Cimiano e Magnini (2005) 23
Figura 4 - Técnicas de agrupamento hierárquico para a aprendizagem de ontologias 27
Figura 5 - Exemplo de hierarquia gerada a partir da FCA (CIMIANO; HOTHO; STAAB, 2004b) 28
Figura 6 - Exemplo de uma rede de Markov gerada a partir da RLM da Tabela 2 ..39
Figura 7 - Processo de aprendizagem de ontologias proposto 46
Figura 8 - Detalhamento do pré-processamento dos corpora 47
Figura 9 - Exemplo de árvore sintática 50
Figura 10 - Passos da Técnica PRECE 51
Figura 11 - Exemplo de parte de uma hierarquia de conceitos 59
Figura 12 - Etapas da fase de aprendizagem de pesos 60
Figura 13 - Exemplo de ontologia resultante do pré-processamento 64
Figura 14 - Resultados dos experimentos com o Corpus *LonelyPlanet* 73
Figura 15 - Parte da hierarquia $HPREHE$ gerada a partir do corpus *LonelyPlanet* 75
Figura 16 - Parte da hierarquia $HFCA$ gerada a partir do corpus *LonelyPlanet* 75
Figura 17 - Parte da hierarquia $HHearst$ gerada a partir do corpus *LonelyPlanet* 76
Figura 18 - Resultado dos Experimentos com o corpus *LonelyPlanet* 78

LISTA DE TABELAS

Tabela 1 - Exemplo de contexto formal no domínio turístico 28
Tabela 2 - Exemplo de uma Rede Lógica de Markov 38
Tabela 3 - Documentos utilizados como exemplo no decorrer da seção 48
Tabela 4 - Sentença após tokenização e anotação dos tokens com rótulos de parte do discurso e lemas 49
Tabela 5 - Exemplos de dependências sintáticas 49
Tabela 6 - Exemplo de Corpus Pré-processado 50
Tabela 7 - Predicados da RLM utilizada pela técnica PRECE 52
Tabela 8 - Estrutura da RLM utilizada pela técnica PRECE 53
Tabela 9 - Exemplo de resultado do processo de inferência 55
Tabela 10 - Arquivo contendo a saída da técnica PRECE 55
Tabela 11 - Exemplo de arquivo contendo instanciações de predicados de evidência da técnica PREHE 57
Tabela 12 - Estrutura da RLM utilizada pela técnica PREHE 58
Tabela 13 - Mapeamento entre as entradas e saídas do processo de criação do conjunto de treinamento para a técnica PRECE 61
Tabela 14 - Exemplo de conjunto de treinamento utilizado para aprender os pesos da RLM da técnica PRECE 62
Tabela 15 - Mapeamento entre as entradas e saídas do processo de criação do conjunto de treinamento para a técnica PREHE 62
Tabela 16 - Exemplo de conjunto de treinamento utilizado para aprender os pesos da RLM da técnica PREHE 63
Tabela 17 - Exemplo de resultados combinados das técnicas PRECE e PREHE 64
Tabela 18 - Corpora utilizados nos experimentos 67
Tabela 19 - Ontologias utilizadas nos experimentos 68
Tabela 20 - Comparativo das técnicas em relação ao número de conceitos extraídos 72
Tabela 21 - Comparativo das técnicas em relação ao número de relacionamentos taxonômicos extraídos 79

LISTA DE SIGLAS

FCA	– Formal Concept Analysis
GATE	– General Architecture for Text Engineering
MAP	– Maximum a posteriori
MCMC	– Markov chain Monte Carlo
PLI	– Programação em Lógica Indutiva
POS	– Part of Speech
PRECE	– Probabilistic Relational Concept Extraction
PREHE	– Probabilistic Relational Hierarchy Extraction
RLM	– Rede Lógica de Markov
TF-IDF	– Term Frequency - Inverse Document Frequency

SUMÁRIO

1 INTRODUÇÃO 13
1.1 Objetivos 15
1.2 Estrutura da dissertação 15
2 APRENDIZAGEM DE ONTOLOGIAS 17
2.1 Ontologias 17
2.2 Aprendizagem de ontologias a partir de fontes textuais 20
2.2.1 O Processo de aprendizagem de ontologias 21
2.2.2 Camadas do processo de aprendizagem de ontologias 22
2.2.2.1 Termos e sinônimos 23
2.2.2.2 Conceitos 25
2.2.2.3 Hierarquias de conceitos 26
2.2.2.4 Relacionamentos 29
2.2.2.5 Axiomas 29
2.3 Avaliação de ontologias 30
2.4 Considerações finais 32
3 APRENDIZAGEM ESTATÍSTICA RELACIONAL 33
3.1 Lógica de primeira ordem 35
3.2 Redes de Markov 36
3.3 Redes Lógicas de Markov 37
3.3.1 Inferência 39
3.3.2 Aprendizagem de pesos 42
3.4 Considerações finais 44
4 UM PROCESSO PARA A EXTRAÇÃO DE HIERARQUIAS DE CONCEITOS A PARTIR DE FONTES TEXTUAIS 45
4.1 Pré-processamento do corpus 46
4.2 Extração de conceitos através da técnica PRECE 50
4.2.1 Identificação de conceitos 51
4.2.2 Nomeação de conceitos 55

4.3 Extração de hierarquia através da técnica PREHE 57
4.4 Aprendizagem de pesos 60
4.5 Pós-processamento 63
4.6 Considerações finais 64
5 AVALIAÇÃO 66
5.1 Dados 67
5.1.1 Corpora de texto 67
5.1.2 Ontologias utilizadas 68
5.2 Ferramentas utilizadas 69
5.3 Descrição do experimento 70
5.4 Avaliação da técnica PRECE 70
5.4.1 Métricas 71
5.4.2 Resultados 72
5.5 Avaliação da Técnica PREHE 74
5.5.1 Hierarquia extraída com a técnica PREHE 74
5.5.2 Hierarquia extraída com a análise formal de conceitos 75
5.5.3 Hierarquia extraída a partir dos padrões de Hearst 76
5.5.4 Métricas 76
5.5.5 Resultados 78
5.6 Considerações finais 79
6 CONCLUSÃO 81
6.1 Resultados 82
6.2 Trabalhos futuros 83
REFERÊNCIAS 85
APÊNDICES 95
ANEXOS 102

1 INTRODUÇÃO

O conhecimento se tornou um fator importante de sucesso para as organizações. Assim como o advento dos bancos de dados relacionais permitiu aos sistemas de computação lidar com a informação de forma mais efetiva, surgindo assim os chamados Sistemas de Informação, a utilização de formalismos de representação de conhecimento para a construção de bases de conhecimento, possibilitou o surgimento dos Sistemas de Conhecimento.

Uma abordagem para a representação de conhecimento é através das ontologias. Elas são capazes de expressar um conjunto de entidades e seus relacionamentos, restrições e axiomas de um dado domínio. Uma vez que elas são formais, o conhecimento que elas representam pode ser facilmente processado pelas máquinas. Elas possibilitam que o conhecimento seja reusado e compartilhado por aplicações heterogêneas.

As ontologias possuem uma grande importância para os modernos sistemas de conhecimento, uma vez que já mostraram ser uma poderosa ferramenta para o processamento da linguagem natural (GIRARDI; IBRAHIM, 1995) (GIRARDI, 1995), filtragem de informação (MIDDLETON; SHADBOLT; ROURE, 2004) (DRUMOND; GIRARDI; SILVA, 2008), recuperação de informação (SILVA; GIRARDI; DRUMOND, 2009) (VAILET; FERN ANDEZ; CASTELLS, 2005) além de constituírem a espinha dorsal da Web Semântica (SHADBOLT; HALL; BERNERS-LEE, 2006) (DRUMOND; GIRARDI, 2006).

A Web Semântica, uma nova geração da Web na qual a semântica dos documentos, na maioria dos casos expressada apenas em linguagem natural, seria expressada utilizando ontologias, é uma das maiores aplicações dessa modalidade de representação de conhecimento. Assim, a Web Semântica é uma abordagem para melhorar a efetividade do acesso à informação na Web.

Tradicionalmente, o desenvolvimento de bases de conhecimento (e de ontologias) tem sido executado manualmente por especialistas de domínio e/ou engenheiros do conhecimento. Contudo, esta é uma tarefa cara e sujeita a erros. Esta dificuldade em tornar explícito o conhecimento implicitamente contido nos textos, bases de dados ou mentes dos especialistas de domínio é conhecida como o

gargalo da aquisição do conhecimento. Superar este problema é crucial para o sucesso das aplicações de conhecimento.

O desenvolvimento rápido e barato de ontologias é crucial para o sucesso dos sistemas de conhecimento, em especial, para a Web Semântica. Uma abordagem para este problema é o suporte automático ou semi-automático à construção de ontologias. Este campo de pesquisa é conhecido como aprendizagem de ontologias (BUITELAAR; CIMIANO; MAGNINI, 2005) (CIMIANO; VÖLKER; STUDER, 2006) (CIMIANO, 2006) (DRUMOND; GIRARDI, 2008).

Muitas técnicas para a aprendizagem de ontologias a partir de fontes textuais têm sido propostas. Grande parte delas se baseiam em análise lingüística (HEARST, 1992) (SNOW; JURAFSKY; NG, 2005) e estatística (HARRIS, 1968) (MAEDCHE. STAAB, 2004) (ZAVITSANOS et al., 2007) e fazem uso de tarefas da aprendizagem de máquina como o agrupamento (CIMIANO; HOTHO; STAAB, 2004b) (ZHAO; KARYPIS, 2002).

As técnicas de aprendizagem estatística apresentam a vantagem de lidar muito bem com o ruído presente em textos em linguagem natural. Contudo, elas assumem que os termos não apresentam relacionamento entre si. Geralmente elas consideram apenas atributos das palavras, como a parte do discurso ou a freqüência. Os relacionamentos entre as palavras tais como a hiperonímia, hiponímia e as dependências sintáticas, podem ser explorados por técnicas de aprendizagem relacional, que, por sua vez, não são capazes de lidar com o ruído advindo da ambigüidade da linguagem natural.

A aprendizagem estatística relacional (RAEDT; KERSTING, 2003) (NEVILLE; RATTIGAN; JENSEN, 2003) (POPESCUL et al., 2003) figura como uma abordagem viável a este problema, uma vez que ela combina o poder expressivo dos formalismos de representação de conhecimento utilizados pela aprendizagem relacional com as abordagens de aprendizagem probabilística.

Este trabalho propõe uma abordagem para o suporte ao processo de construção de ontologias através da automatização da extração de hierarquias de conceitos a partir de fontes textuais. O processo proposto é composto por duas técnicas, a PRECE (*Probabilistic Relational Concept Extraction*), para a extração de conceitos e a PREHE (*Probabilistic Relational Hierarchy Extraction*), para a descoberta de relacionamentos taxonômicos entre os conceitos extraídos pela PRECE. As duas técnicas fazem uso das Redes Lógicas de Markov (RICHARDSON;

DOMINGOS, 2006), uma abordagem da aprendizagem probabilística relacional que combina a lógica de primeira ordem com as redes de Markov. As técnicas PRECE e PREHE foram avaliadas no domínio turístico comparando os seus resultados com uma ontologia desenvolvida manualmente por especialistas neste domínio.

1.1 Objetivos

O objetivo geral deste trabalho é contribuir para a resolução do problema do gargalo da aquisição de conhecimento através de uma técnica para o suporte automatizado ao desenvolvimento de ontologias a partir de fontes textuais.

Para alcançar esse objetivo geral, planeja-se atingir os seguintes objetivos específicos:

a) Contribuir com a implantação da Web Semântica, através da automatização parcial da conversão do conhecimento presente nas páginas atuais para uma representação ontológica;
b) Revisar e avançar na compreensão do estado da arte das técnicas de aprendizagem de ontologias;
c) Compreender e avançar nas técnicas de aprendizagem estatística relacional, compreendendo como tais técnicas podem ser utilizadas na melhoria da efetividade da aprendizagem de ontologias a partir de fontes textuais;
d) Desenvolver e avaliar uma técnica para a extração de conceitos e outra para a extração de hierarquias a partir de fontes textuais, ambas baseadas na aprendizagem estatística relacional.

1.2 Estrutura da dissertação

Este trabalho, incluindo esta introdução, está estruturado em seis capítulos. No Capítulo 2 é apresentada uma revisão do estado da arte da aprendizagem de ontologias a partir de fontes textuais. Em primeiro lugar é

apresentada a definição para o termo “ontologia” utilizada durante todo este trabalho. Em seguida a tarefa de aprendizagem de ontologias é definida e sua divisão em camadas é apresentada. Algumas das principais abordagens utilizadas em cada uma das camadas da aprendizagem de ontologias são discutidas. Por fim, este capítulo traz uma discussão sobre os métodos para avaliação de ontologias.

No Capítulo 3, a aprendizagem estatística relacional é definida. Neste capítulo, são apresentadas as redes lógicas de Markov, a abordagem de aprendizagem estatística relacional utilizada neste trabalho. São discutidos os conceitos básicos acerca da lógica de primeira ordem e das redes de Markov, essenciais para a compreensão das lógicas de Markov. Por fim as redes lógicas de Markov são apresentadas e algoritmos para inferência e aprendizagem destes modelos são discutidos.

O Capítulo 4, parte central deste trabalho, apresenta as técnicas PRECE (*Probabilistic RElational Concept Extraction*) e PREHE (*Probabilistic RElational Hierarchy Extraction*) para a extração de conceitos e hierarquia de conceitos respectivamente. As duas técnicas são integradas em um processo que também envolve uma fase de pré-processamento do corpus de texto, uma fase de pós-processamento dos resultados obtidos e uma fase de aprendizagem. Cada uma dessas fases é discutida em detalhe neste capítulo.

A avaliação das técnicas PRECE e PREHE é apresentada no Capítulo 5. Neste capítulo são descritos os dados utilizados na avaliação, a metodologia de avaliação e as medidas consideradas. Por fim são apresentados os resultados obtidos, acompanhados de uma discussão sobre os mesmos.

No último capítulo, as conclusões do trabalho são apresentadas, incluindo as contribuições das técnicas propostas, as limitações que as mesmas ainda possuem e que precisam ser superadas e as perspectivas de trabalhos relacionados que estão sendo ou que serão desenvolvidos no futuro.

2 APRENDIZAGEM DE ONTOLOGIAS

Este capítulo apresenta uma revisão do estado da arte da aprendizagem de ontologias, o problema tratado neste trabalho, e está organizado como segue. A Seção 2.1 apresenta a definição de ontologia utilizada aqui. A Seção 2.2 define o problema da aprendizagem de ontologias a partir de fontes textuais e discute algumas das abordagens propostas na literatura. O problema da avaliação de ontologias é discutido na Seção 2.3. Por fim, a Seção 2.4 apresenta as considerações finais do capítulo.

2.1 Ontologias

Para definir a tarefa de aprendizagem de ontologias é necessário primeiro definir o que é uma ontologia. Existem várias definições para este termo e algumas delas podem ser encontradas em (BIEMANN, 2005). Para os objetivos deste trabalho, o termo ontologia será utilizado segundo a definição proposta nesta seção.

De acordo com Russel e Norvig (2003), o termo “ontologia” diz respeito a uma teoria sobre a existência. A Inteligência Artificial considera as ontologias como especificações formais de conceitos de um domínio de interesse, onde seus relacionamentos, restrições e axiomas são expressados definindo assim um vocabulário comum para compartilhar o conhecimento (GRUBER, 1995).

Uma ontologia é composta por um lado de conceitos e relacionamentos taxonômicos (que definem uma hierarquia de conceitos) e não taxonômicos entre eles e por outro por instâncias dos conceitos e asserções sobre as mesmas. Ela deve ser formal e, portanto, compreensível para os agentes e outras entidades computacionais. Desta forma, as ontologias podem fornecer um vocabulário comum entre várias aplicações e por isso também devem ser compartilhadas. Esta estrutura de representação de conhecimento, normalmente, consiste de um conjunto de classes organizadas hierarquicamente descrevendo um domínio. Mais formalmente, uma ontologia pode ser definida de acordo com a equação 1.

$$\mathcal{O} = (C, H, I, R, P, A) \quad (1)$$

Onde:

a) $C = C^C \cup C^I$ é o conjunto de entidades da ontologia, i.e., ele representa as entidades do domínio sendo modelado. São designados por um ou mais termos em linguagem natural. O conjunto C^C é formado por classes, ou seja, conceitos que representam entidades genéricas que descrevem um conjunto de objetos (por exemplo, "Pessoa" $\in C^C$) enquanto o conjunto C^I é formado por instâncias, ou seja, entidades únicas no domínio (por exemplo "umberto eco" $\in C^I$);

b) $H = \{\text{tipo_de}(c_1, c_2) | \, c_1 \in C^C \wedge c_2 \in C^C \}$ é o conjunto das relações taxonômicas entre os conceitos. Tais relações definem a hierarquia de conceitos e são denotadas por $\text{tipo_de}(c_1, c_2)$ significando que c_1 é um tipo de c_2. Um exemplo desse relacionamento é $\text{tipo_de}(Autor, Pessoa)$;

c) $I = \{\text{é_um}(c_1, c_2) | c_1 \in C^I \wedge c_2 \in C^C\}$ é o conjunto de relacionamentos entre classes e instâncias (relacionamento "é um") de uma ontologia, por exemplo $\text{é_um}(UmbertoEco, Autor)$;

d) $R = \{ rel_k \, (c_1, c_2, \ldots, c_n) \, | \forall i, \; c_i \in C\}$ é o conjunto de relacionamentos que não são nem taxonômicos nem de instanciação entre classes e instâncias de uma ontologia. Alguns exemplos são $escreve(Autor, Livro)$ e $escreve(UmbertoEco, oNomedaRosa)$;

e) $P = \{ prop_k \, (c_i, tipo/valor) \mid c_i \in C\}$ é o conjunto de propriedades das entidades de uma ontologia. Tais propriedades relacionam conceitos a um tipo básico de dados, como *inteiro*, *real* ou *string*, ou podem relacionar instâncias a valores específicos dos tipos de dados. Alguns exemplos são $idade(Pessoa, Inteiro)$ e $idade(joao, 30)$;

f) $A = \{Condition_x \implies conclusion_y \, (c_1, c_2, \ldots c_n) | \forall_j, \; c_j \in C^C\}$ é um conjunto de axiomas, regras que permitem checar a consistência da ontologia e deduzir novos conhecimentos através de algum mecanismo de inferência. O termo $Condition_x$ é dado por: $Condition_x := \{(\text{cond}_1, \text{cond}_2, \ldots, \text{cond}_\text{n}) | \, \forall_z, \text{cond}_\text{z} \in H \cup I \cup R\}$. Por exemplo: $\{comprou(Cliente, Livro1), autor(Autor, Livro1\,),$ $autor(Autor, Livro2\,)\} \rightarrow \text{provável-comprador}\,(Cliente, Livro2)$.

Para esclarecer esta definição, tomemos como exemplo uma ontologia de uma livraria. A ontologia aqui mostrada é bem restrita, apenas para ilustrar a definição apresentada. Uma livraria vende livros e lida com pessoas, que podem ser autores ou clientes. Os autores escrevem livros, enquanto os clientes compram os livros e têm interesse por seus autores. A partir desta descrição identifica-se a ontologia a seguir, também mostrada na Figura 1.

a) $C^C = \{pessoa, autor, cliente, livro\}$

b) $C^I = \{UmbertoEco, ONomeDaRosa, Joao\}$

c) $H = \{\text{tipo_de}(raiz, pessoa), \text{tipo_de}(raiz, livro), (pessoa, cliente), (pessoa, autor)\}$

d) $I = \{\text{é_um}(UmbertoEco, Autor), \text{é_um}(ONomeDaRosa, Livro), \text{é_um}(Joao, Cliente)\}$

e) $R = \{\text{tem_interesse}(cliente, autor), \text{compra}(cliente, livro), \text{escreve}(autor, livro), \text{escreve}(UmbertoEco, ONomeDaRosa)\}$

f) $P = \{preco(Livro, Real), preco(oNomedaRosa, 15.82)\}$

g) $A = \{\text{compra}(cliente_1, livro_1), \text{escreve}(autor_1, livro_1)\} \Rightarrow tem_interesse(cliente_1, autor_1)$

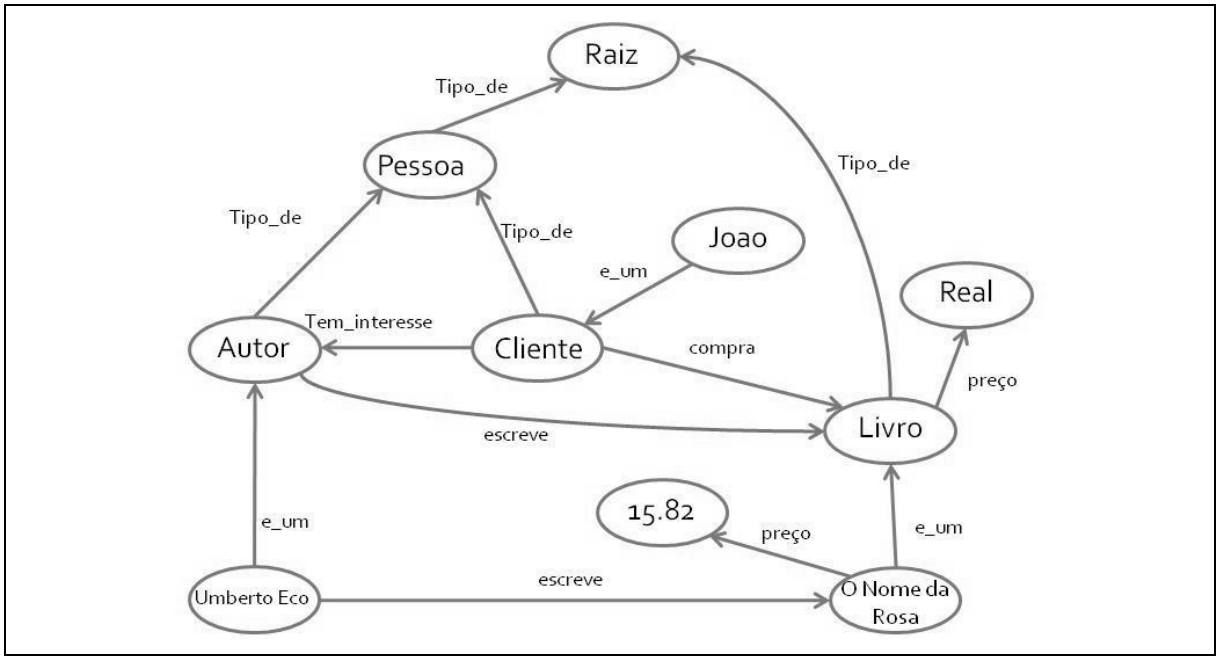


Figura 1 - Exemplo de ontologia

Cada conceito, instância e relação de uma ontologia possui um identificador único. Além disso, os conceitos, instâncias e relações em um domínio

são referenciados por um ou mais termos em linguagem natural. Por isso, algumas abordagens para aprendizagem de ontologias utilizam um lexicon, uma estrutura auxiliar utilizada para mapear termos em linguagem natural a conceitos, relações e instâncias em uma ontologia. Um lexicon é definido de acordo com a equação 2.

$$L = (L^C, L^I, L^R, F, G, K) \quad (2)$$

Um lexicon L é formado pelos conjuntos de entradas léxicas para conceitos L^C, instâncias L^I e relações L^R bem como dos relacionamentos $F \subseteq L^C \times C^C$ associando uma entrada léxica a um determinado conceito, $K \subseteq L^I \times C^I$, associando uma entrada léxica a uma determinada instância e $G \subseteq L^R \times R$ associando uma entrada léxica a uma relação.

2.2 Aprendizagem de ontologias a partir de fontes textuais

A tarefa da aquisição de conhecimento é representar através de um formalismo o conhecimento implícito presente em uma ou mais fontes. O termo aprendizagem de ontologias refere-se ao suporte automático ou semi-automático à construção de uma ontologia, enquanto o suporte automático ou semi-automático à instanciação de uma dada ontologia é chamado de povoamento de ontologia (BUITELAAR; CIMIANO; MAGNINI, 2005). Juntos, a aprendizagem e o povoamento de ontologias constituem uma abordagem para automatizar a aquisição de conhecimento através da descoberta de conhecimento em diferentes fontes de dados e representando-o através de ontologias.

De acordo com Benz (2007) há dois aspectos fundamentais da aprendizagem de ontologias. O primeiro é a disponibilidade de conhecimento prévio, que pode ser na forma de uma ontologia a ser estendida ou pode ser transformado na primeira versão da ontologia. Tal versão é então estendida automaticamente por procedimentos de aprendizagem ou manualmente pelo engenheiro de conhecimento (DELLSCHAFT, 2005).

O outro aspecto é o formato das fontes de dados a partir das quais se deseja extrair conhecimento. Existem três diferentes tipos de fontes de dados (BENZ, 2007):

a) fontes desestruturadas: documentos em linguagem natural, como documentos PDF, Word e como a maioria das páginas da Web tradicional;
b) fontes semi-estruturadas: dicionários e folksonomias;
c) fontes estruturadas: esquemas de bancos de dados.

Algumas abordagens para a aprendizagem de ontologias a partir de fontes estruturadas (LEHMANN; HITZLER, 2007) e semi-estruturadas (BENZ, 2007) (MARINHO; BUZA; SCHMIDT-THIEME, 2008) (WU; WELD, 2008) foram propostas e apresentaram bons resultados. Contudo, apesar de tais abordagens proverem um determinado suporte ao desenvolvimento de ontologias, a maior parte do conhecimento disponível, especialmente na Web, está na forma de textos em linguagem natural (MAEDCHE; STAAB, 2001). Por isso, a aprendizagem de ontologias a partir de fontes textuais (BIEMANN, 2005)(CIMIANO, 2006) é um ponto central para algumas áreas como o estabelecimento da Web Semântica e constitui o foco deste trabalho.

2.2.1 O Processo de aprendizagem de ontologias

Antes de apresentar algumas abordagens existentes para a aprendizagem de ontologias, convém discutir um modelo genérico para a aprendizagem de ontologias baseada na aprendizagem de máquina. Tal modelo é mostrado na Figura 2.

A utilização da aprendizagem de máquina na aprendizagem de ontologias consiste em aplicar algoritmos de mineração de dados aos textos para detectar padrões recorrentes que possam ser mapeados para uma ontologia. Entretanto, os algoritmos de aprendizagem de máquina são normalmente aplicados a dados estruturados, uma vez que a mineração de dados visa à descoberta de conhecimento em bases de dados estruturadas. Por isso o corpus de onde se deseja extrair a ontologia é primeiramente pré-processado.

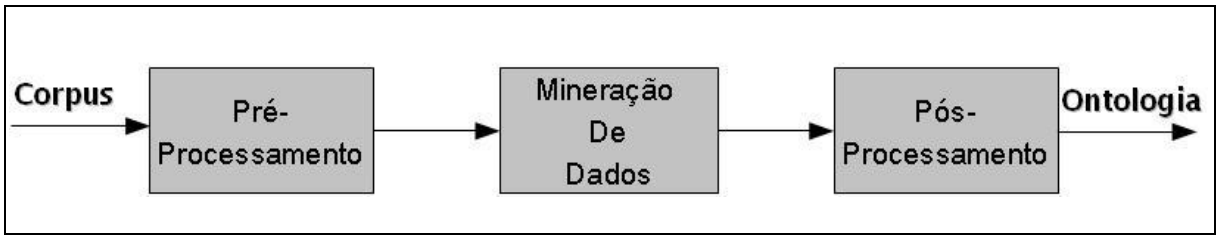


Figura 2 - Modelo do Processo de aprendizagem de ontologias baseado na aprendizagem de máquina

O pré-processamento visa extrair dos textos em linguagem natural, dados que possam ser utilizados como entrada para o processo de mineração de dados. Para isso são utilizadas técnicas de Processamento da Linguagem Natural e análise estatística. Através do processamento da linguagem natural é possível extrair informações léxicas e sintáticas sobre os termos. A análise estatística permite extrair dados baseados na freqüência com que os termos aparecem no corpus ou em determinados documentos, bem como dados de co-ocorrência de termos (termos que ocorrem freqüentemente juntos).

A seguir os dados extraídos são utilizados como entrada para um processo de mineração de dados para descobrir e reconhecer padrões que possam ser utilizados, por exemplo, no agrupamento de termos com significados semelhantes ou na classificação de um termo como referente a uma classe ou uma instância de uma ontologia.

Por fim, o resultado da mineração de dados é organizado e escrito em uma linguagem de especificação de ontologias na fase de pós-processamento.

2.2.2 Camadas do processo de aprendizagem de ontologias

Apesar de não haver um consenso quanto às etapas do processo de desenvolvimento de ontologias, Buitelaar, Cimiano e Magnini (2005) organizam os aspectos e tarefas deste processo de desenvolvimento em um conjunto de camadas, como mostra a Figura 3.

De acordo com a equação 1, uma ontologia consiste basicamente de conceitos, relacionamentos entre eles e axiomas. Para identificar os conceitos de um domínio, é necessário primeiramente identificar os termos em linguagem natural que

se referem a eles. A identificação de sinônimos contribui para evitar conceitos redundantes uma vez que dois ou mais termos em linguagem natural podem representar o mesmo conceito. Os termos são a fonte para identificar os conceitos que farão parte da ontologia, i.e. o conjunto C^C da equação 1.

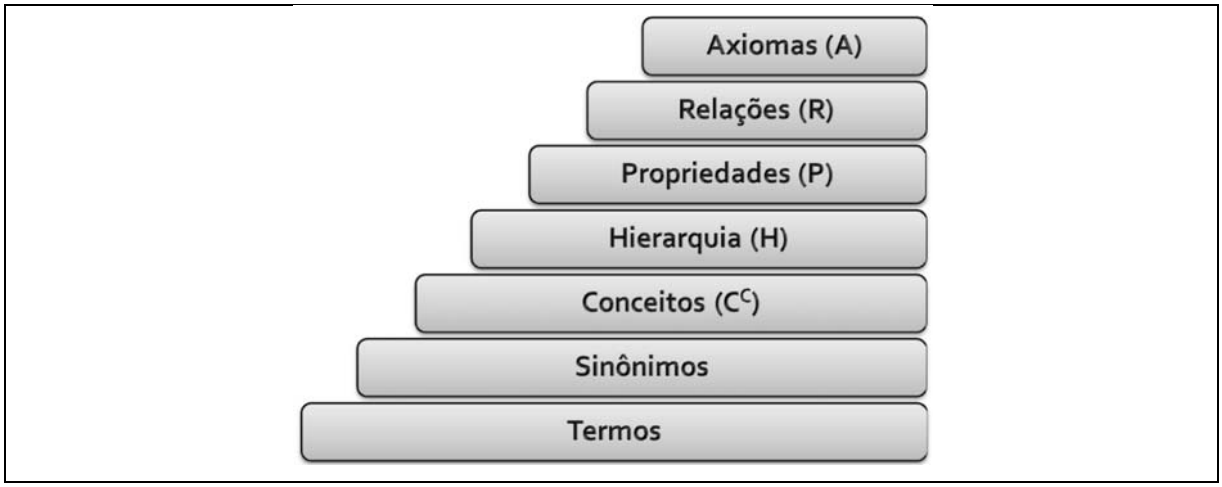


Figura 3 - Camadas do Processo de Desenvolvimento de Ontologias adaptadas de Buitelaar, Cimiano e Magnini (2005)

O próximo passo é identificar as relações taxonômicas (generalização e especialização) entre os conceitos. O produto desta tarefa é o conjunto H. Também é necessário identificar as relações não taxonômicas entre os conceitos, determinando assim o conjunto R. Alguns autores consideram também a aquisição dos axiomas que irão constituir o conjunto A.

Cada uma das camadas da Figura 3 representa um subproblema da aprendizagem de ontologias. A seguir cada subproblema será definido bem como as respectivas soluções propostas na literatura.

2.2.2.1 Termos e sinônimos

Um documento em linguagem natural pode ser visto como um conjunto ordenado de termos. Os termos constituem a base do processo de aprendizagem de ontologias uma vez que cada termo possui um significado que pode ser um conceito ou um relacionamento relevante para o domínio em questão. O objetivo desta primeira etapa do processo de aprendizagem de ontologias é identificar quais destes

termos possuem um significado relevante para as etapas subseqüentes, ou seja, aprender os conjuntos L^C, L^R, e L^I, da equação 2.

Os métodos para realizar esta tarefa se utilizam de análise estatística e lingüística. Em geral, ambas as abordagens são utilizadas de forma combinada realizando-se um pré-processamento lingüístico e aplicando-se técnicas estatísticas apenas sobre o resultado deste processamento.

Os métodos estatísticos para a extração de termos vêm em geral das técnicas de recuperação de informação (SALTON; BUCKLEY, 1987). Tais métodos são baseados na idéia de que um termo que apareça freqüentemente na coleção de documentos indica um conceito relevante do domínio em questão. Entretanto, a pesquisa em recuperação de informação mostra que há métodos mais eficazes para a ponderação de termos que a simples contagem da freqüência dos termos. Uma medida simples e muito utilizada é a TF-IDF (SALTON; BUCKLEY, 1987) que considera a freqüência direta dos termos e a freqüência inversa.

Sejam $freq_{t,i}$ o número de vezes que o termo *t* aparece no documento d_i e df_t o número de documentos da coleção *D* onde o termo *t* ocorre, a medida TF-IDF é definida pela equação 3.

$$TF - IDF_{t,i} = freq_{t,i} * \log\left(\frac{|D|}{df_t}\right) \quad (3)$$

A equação 4 computa a relevância de um termo para um dado documento. A relevância de um termo para o corpus todo geralmente é considerada como a soma da relevância do termo para cada documento no corpus (MAEDCHE; STAAB, 2004) assim como na equação 4.

$$Relevancia_t = \sum_{i \in D} TF - IDF_{t,i} \quad (4)$$

A descoberta de sinônimos é útil para evitar a formação de conceitos redundantes. Uma abordagem comum para esta tarefa é a utilização de técnicas de agrupamento. Para tanto, a similaridade entre os termos pode ser calculada através de métodos estatísticos, como os baseados na co-ocorrência, na hipótese distribucional de Harris (HARRIS, 1968), que diz que palavras similares tendem a

aparecer em contextos similares ou utilizando medidas de similaridade semântica na estrutura do WordNet (FELLBAUM, 1998) (RESNICK, 1999).

2.2.2.2 Conceitos

O objetivo desta fase é aprender o conjunto C^C . Entretanto, a extração de conceitos a partir de texto é uma tarefa controversa pois não está claro o que é um conceito. Segundo Buitelaar, Cimiano e Magnini (2005) um conceito pode ser definido pela sua intensão, ou seja, uma descrição formal do mesmo, pela sua extensão que é o conjunto de instâncias do conceito ou pelas suas realizações léxicas, conjunto de termos em linguagem natural que o designam.

A extração da intensão do conceito a partir de texto é normalmente abordada através da Programação em Lógica Indutiva (LEHMANN; HITZLER, 2007) e através da Análise Formal de Conceitos (CIMIANO; HOTHO; STAAB, 2004b) (FCA do inglês *Formal Concept Analysis*).

A Análise Formal de Conceitos é um método utilizado para a análise de dados. Os dados são estruturados em unidades que são abstrações dos conceitos presentes no pensamento humano. Cada unidade dessas é descrita como um conjunto de objetos e seus respectivos atributos. Mais formalmente, a FCA analisa a correlação entre os objetos *G* e suas características *M*. As características são associadas aos objetos através de uma relação binária $I \subseteq G \times M$. Um conceito formal é uma relação bijetiva *(A,B)* tal que $A \subseteq G$ e $B \subseteq M$.

Na aplicação da FCA para a extração de conceitos, os objetos *G* são os termos extraídos e as características *M*, os verbos com os quais eles se relacionam. Termos que aparecem freqüentemente no corpus como sujeitos dos mesmos verbos são agrupados em um mesmo conjunto. Como será visto mais adiante a FCA também pode ser utilizada para extrair a hierarquia dos conceitos.

A aprendizagem de conceitos através de suas realizações lingüísticas normalmente é realizada através de técnicas de agrupamento. Uma abordagem muito utilizada é o agrupamento hierárquico (ZHAO; KARYPIS, 2002) utilizada também para a extração de hierarquia de conceitos.

Outro grupo de abordagens para a extração de conceitos através de suas realizações léxicas é composto pelas abordagens baseadas na análise semântica latente probabilística (HOFMANN, 2001), um modelo probabilístico que descreve os tópicos presentes nos documentos de um corpus como distribuições de probabilidade sobre os termos. Em Zavitsanos et al. (2007), cada tópico extraído é mapeado para um conceito da ontologia.

2.2.2.3 Hierarquias de conceitos

Esta etapa do processo de aprendizagem de ontologias é de longe a mais pesquisada e mais avançada. Aqui o objetivo é extrair o conjunto H a partir do corpus e dos conceitos extraídos na etapa anterior. Novamente, entre as abordagens propostas se destacam as lingüísticas e as estatísticas, quase sempre com o apoio de alguma técnica de aprendizagem de máquina.

Uma das abordagens lingüísticas é o uso de padrões léxico-sintáticos para identificar relações de hiperonímia no texto. Os relacionamentos de hiperonímia indicam uma generalização entre os termos, como acontece entre os termos *automóvel* e *carro*, uma vez que todo carro é um *tipo de* automóvel. Dentre estes padrões destacam-se os padrões de Hearst (HEARST, 1992). Hearst definiu alguns padrões como “Ws tais como Xs, Ys e Zs”, “Xs, Ys e outros Zs”. Por exemplo a frase “carros, motocicletas e outros automóveis” indica a existência de relacionamentos taxonômicos entre os conceitos indicados por *carros*, *motocicletas* e *automóveis*, i.e, *tipo_de(carro, automóvel)* e *tipo_de(motocicleta,automóvel)*. Algumas extensões foram propostas a esta abordagem (IWANSKA; MATA; KRUGER, 2000) (SNOW; JURAFSKY; NG, 2005). O grande problema destas abordagens é que, como tais padrões aparecem com pouca freqüência nos documentos, elas, em geral, extraem uma baixa porcentagem dos relacionamentos presentes no corpus (CIMIANO; VöLKER; STUDER, 2006).

Outra classe de abordagens para a extração da hierarquia de conceitos é composta por aquelas baseadas em agrupamento (MAEDCHE; STAAB, 2004). Dentre estas abordagens destacam-se aquelas que utilizam agrupamento hierárquico (CARABALLO, 1999) (FAURE; NEDELLEC, 1999). Tais métodos podem

ser divididos em aglomerativos e particionais. As técnicas de agrupamento hierárquico são ilustradas na Figura 4.

As abordagens aglomerativas constroem a hierarquia ”de baixo para cima”. Elas partem do princípio de que cada termo é por si só um grupo. Então grupos maiores são formados iterativamente fundindo grupos similares até que um critério de parada seja alcançado. Já nas abordagens particionais, o conjunto de todos os termos é considerado um único grupo que é iterativamente particionado em grupos menores de termos mais similares, sendo essa uma abordagem para a construção da hierarquia ”de cima para baixo”. Os grupos são então mapeados para conceitos e a fusão de grupos em um grupo maior indica um relacionamento taxonômico, extraindo-se assim a hierarquia de conceitos. Estas técnicas em geral são muito ruidosas, i.e. recuperam muitas relações taxonômicas que não existem, por isso têm-se adicionado a elas algum tipo de supervisão no processo de agrupamento (FAURE; NEDELLEC, 1999) (CIMIANO; STAAB, 2005).

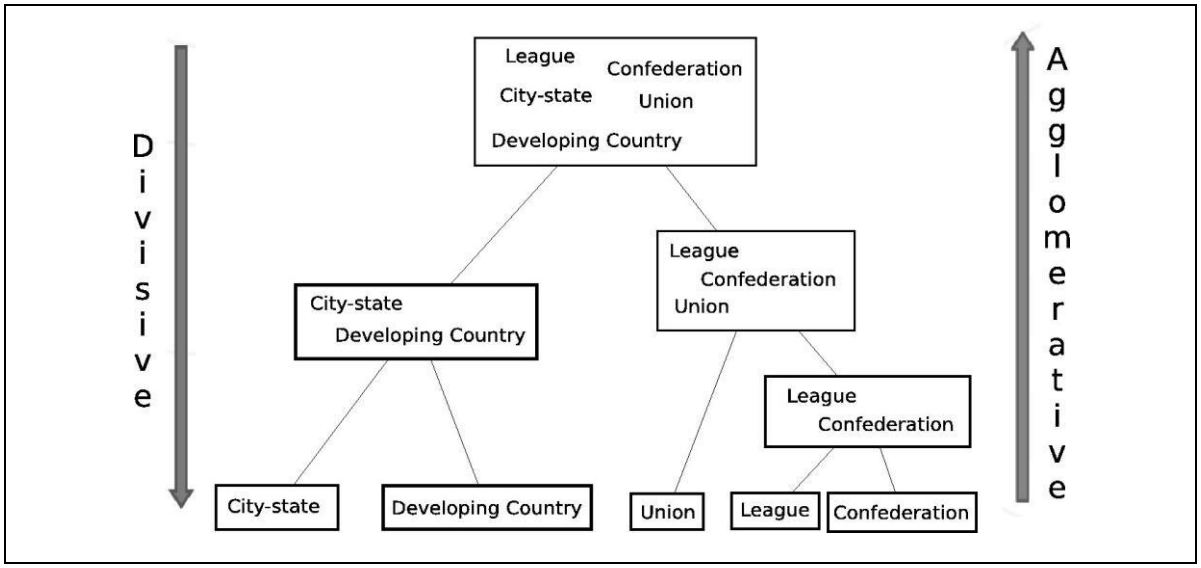


Figura 4 - Técnicas de agrupamento hierárquico para a aprendizagem de ontologias

Há também outro tipo de agrupamento, o agrupamento conceitual. Trabalhos neste sentido têm-se baseado na Análise Formal de Conceitos (CIMIANO; HOTHO; STAAB, 2004b). A hierarquia de conceitos é extraída a partir de uma ordenação nos conceitos formais. Esta ordenação é definida na equação 5.

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \quad (5)$$

Uma vez que o corpus esteja representado através deste formalismo, os conceitos formais são mapeados para a os conceitos da ontologia e as relações taxonômicas extraídas a partir da ordenação dos mesmos. Por exemplo $(A_1, B_1) \leq (A_2, B_2)$ significa que o conceito definido por (A_2, B_2) é uma generalização do conceito definido por (A_1, B_1). Deste modo, dado o contexto formal da Tabela 1, exemplo dado em (CIMIANO; HOTHO; STAAB, 2004b), extrai-se a hierarquia da Figura 4 construída segundo a equação 5. Na Tabela 1 as linhas são grupos de objetos (e.g. hotel se refere a todos os hotéis) e as colunas representam características. Um *X* indica que um grupo de objetos possui uma característica. Por exemplo, um apartamento, um carro e uma bicicleta podem ser alugados (*rentable*).

Tabela 1 - Exemplo de contexto formal no domínio turístico

	bookable	*rentable*	*driveable*	*rideable*	*joinable*
Hotel	*X*				
Apartment	*X*	*X*			
Car	*X*	*X*	*X*		
Bike	*X*	*X*	*X*	*X*	
Excursion	*X*				*X*
Trip	*X*				*X*

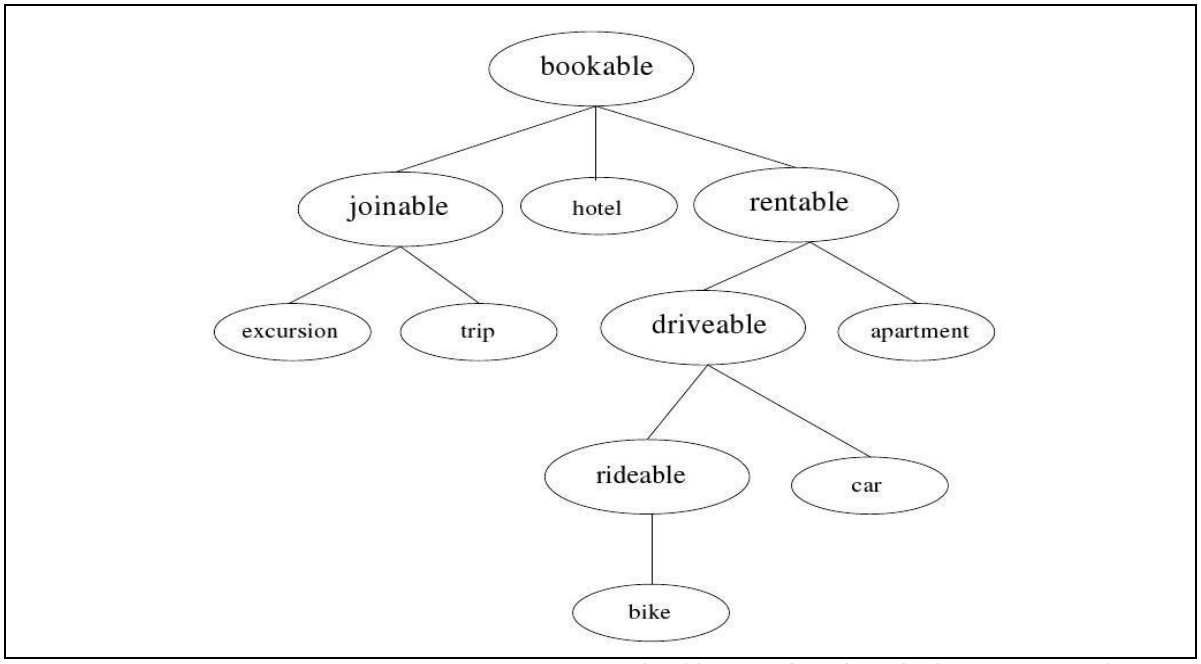


Figura 5 - Exemplo de hierarquia gerada a partir da FCA (CIMIANO; HOTHO; STAAB, 2004b)

2.2.2.4 Relacionamentos

Além das relações taxonômicas e das relações de instanciação existem outros tipos de relações como, por exemplo, as relações “parte de” e outras específicas de cada domínio. Estas relações constituem o conjunto *R*. A aquisição do conjunto *P* das propriedades, é abordada junto com o conjunto *R* na literatura.

O trabalho na extração de relações a partir de texto tem sido influenciado pela mineração de dados. Uma técnica da mineração de dados muito utilizada nesse contexto são as regras de associações (AGRAWAL; IMIELINSKI; SWAMI, 1993). Essas técnicas visam descobrir associações entre grandes itens de dados. No contexto da aprendizagem de ontologias a aprendizagem de regras de associação é aplicada sobre as estruturas sintáticas e/ou sobre as co-ocorrências estatísticas. As associações descobertas entre os itens de dados são mapeadas para as relações não taxonômicas do conjunto *R*. Esta abordagem é descrita em (MAEDCHE; STAAB, 2000).

Novamente as abordagens lingüísticas mostram seu valor para a aprendizagem de ontologias. Alguns trabalhos consideram que os verbos que aparecem nos textos são indicadores de relacionamentos entre seus argumentos (sujeito e objeto) como em (CIARAMITA et al., 2005). Por exemplo, a sentença “um professor ensina uma disciplina” significa que há um relacionamento, indicado pelo verbo ensina entre o sujeito professor e o objeto disciplina.

2.2.2.5 Axiomas

A extração de axiomas a partir de texto ainda é um problema ainda pouco estudado e, uma vez que as técnicas propostas ainda não apresentam resultados satisfatórios, este é considerado ainda um problema em aberto. Ele pode ser classificado em dois problemas distintos. Por um lado esta atividade pode ser vista como a instanciação de esquemas de axiomas previamente definidos. Tais axiomas são aqueles comumente utilizados na engenharia de ontologias (CIMIANO, 2006) como, por exemplo, a disjunção e a equivalência para os conceitos e a transitividade

e simetria para as relações. Um trabalho nesse sentido é apresentado em Haase e Völker (2005). Esta abordagem é baseada em padrões lingüísticos. Ela assume que termos que aparecem em expressões como ”homens e mulheres“ têm alta probabilidade de serem disjuntos.

Por outro lado a extração de axiomas pode ser vista como a aprendizagem dos mesmos e não apenas a sua instanciação. Existem poucas abordagens propostas. Uma delas é a utilizada no sistema HASTI (SHAMSFARD; BARFOROUSH, 2002), onde axiomas em sentenças condicionais e quantificadas em linguagem natural são traduzidos para axiomas em uma linguagem lógica formal. Para isto usa-se uma abordagem semelhante aos padrões de Hearst para a busca de padrões no texto que indiquem a presença de sentenças condicionais e quantificadas.

2.3 Avaliação de ontologias

Comparar técnicas para a aprendizagem de ontologias não é uma tarefa trivial. Para um dado domínio não há uma possibilidade única de conceitualização (SMITH, 2003) e cada uma das possibilidades existentes podem ser mais ou menos úteis para determinadas tarefas e ainda assim serem justificáveis (BIEMANN, 2005). Além disso, não existe uma definição clara do conhecimento que deve ser adquirido (BREWSTER et al., 2004).

Apesar da avaliação das técnicas de aprendizagem de ontologias ainda ser um problema em aberto, já existem trabalhos nessa direção. Em (SHAMSFARD; BARFOROUSH, 2003) duas abordagens básicas para a avaliação dessas técnicas são discutidas: a avaliação dos métodos de aprendizagem utilizados e a avaliação da ontologia aprendida. Entretanto, por causa da dificuldade de medir o grau de correção dos procedimentos de aprendizagem, a primeira abordagem é menos utilizada (SHAMSFARD; BARFOROUSH, 2003).

De acordo com Dellschaft e Staab (2006), as ontologias resultantes podem ser comparadas avaliando-as em uma aplicação executável, por avaliação a posteriori por especialistas ou pela comparação dos resultados aprendidos com uma ontologia de referência pré-definida.

As ontologias aprendidas automaticamente são úteis na medida em que melhoram a efetividade dos sistemas nos quais elas são empregadas. Assim, a comparação de ontologias em uma aplicação executável visa medir a efetividade de um sistema que utiliza as ontologias sendo avaliadas. A vantagem desta abordagem é que é possível utilizar aplicações com métodos de avaliações simples e diretos. Por exemplo, uma comparação de hierarquias de conceitos no contexto da tarefa de desambiguar o sentido de palavras é mostrado por Agirre et al. (2000) e no contexto do agrupamento de texto por Bloehdorn, Cimiano e Hotho (2006). Entretanto os resultados de tal avaliação mostram se uma ontologia é adequada a uma tarefa mas é difícil generalizar esta conclusão para outras tarefas. Outra desvantagem dessa abordagem é que outros fatores podem ter impacto na saída do sistema e, algumas vezes a ontologia é, de fato, uma parte pequena do sistema e que pouco impactua nos seus resultados (BRANK; GROBELNIK; MLADENIC, 2005).

A avaliação manual tem suas vantagens, uma vez que espera-se que os especialistas conheçam os conceitos de relacionamentos dos seus domínios de atuação e, portanto, eles são supostamente capazes de dizer se uma dada ontologia representa bem o domínio ou não.

Apesar da avaliação de ontologias a posteriori por especialistas e sua avaliação em uma aplicação terem suas vantagens, elas também apresentam problemas. Por exemplo, a avaliação manual é subjetiva e demorada. Além disso, estes dois métodos não são viáveis para avaliações em larga escala (DELLSCHAFT; STAAB, 2006). Assim, a comparação com uma ontologia de referência (CIMIANO et al., 2005) é uma alternativa plausível. Um trabalho primário na comparação de duas ontologias é mostrado em Maedche e Staab (2002). Dellschaft e Staab (2006) apresentam um framework para a avaliação de aprendizagem de ontologias baseado em uma ontologia de referência.

Entretanto, como se pode afirmar que uma ontologia é boa o suficiente para ser uma ontologia de referência? A ontologia de referência é uma ontologia feita a mão, desenvolvida pelos mesmos processos custosos e propensos a erros que a aprendizagem de ontologias tenta evitar. Se a ontologia de referência apresentar problemas de modelagem, o método de avaliação recompensa ontologias com problemas similares e penaliza ontologias com conceitos ou relacionamentos que não aparecem na ontologia de referência.

Ao invés de comparar uma ontologia com outra ontologia feita à mão, as abordagens dirigidas aos dados (BREWSTER et al., 2004) (PATEL et al., 2003) comparam as ontologias aprendidas com os dados existentes. Por exemplo, em Brewster et al. (2004) um conjunto de termos foi extraído de um corpus de documentos utilizando análise semântica latente de modo que a adequação da ontologia ao corpus pudesse ser medida pela interseção entre os termos extraídos e os termos que aparecem na ontologia.

2.4 Considerações finais

Este capítulo apresentou uma revisão do estado da arte das técnicas para a aprendizagem automática de ontologias a partir de fontes textuais. Foi também apresentada a definição de ontologia considerada no decorrer deste trabalho. A aprendizagem de ontologias foi apresentada não como uma única tarefa mas dividida em sub-tarefas organizadas em camadas como na Figura 3. Este trabalho se concentra nas camadas de conceitos e de hierarquias.

A partir da discussão apresentada aqui é possível perceber que as técnicas para aprendizagem de ontologias se baseiam na aprendizagem de máquina e no processamento da linguagem natural. Uma possibilidade para aumentar a efetividade dos métodos propostos baseia-se no fato de que o processamento da linguagem natural revela relacionamentos entre os termos como as relações de hiperonímia descobertas pelos padrões de Hearst e os relacionamentos sintáticos revelados pela análise sintática. Abordagens de aprendizagem relacional representam de forma compacta e elegante esses relacionamentos, mas não conseguem lidar de forma efetiva com o ruído presente nos textos em linguagem natural. Abordagens de aprendizagem estatística, por sua vez, lidam bem com o ruído presente nos textos, mas negligenciam os relacionamentos existentes entre os termos.

Deste modo, a aprendizagem estatística relacional parece ser uma abordagem promissora para aumentar a efetividade da aprendizagem de ontologias em relação às técnicas apresentadas neste capítulo. A aprendizagem estatística relacional é tratada no próximo capítulo.

3 APRENDIZAGEM ESTATÍSTICA RELACIONAL

A aprendizagem de máquina (BISHOP, 2006) (MITCHELL, 1997) é uma disciplina que trata do desenvolvimento de sistemas capazes de melhorar o seu desempenho através da própria experiência (MITCHELL, 2006). Tais técnicas são aplicadas em diversos campos como, por exemplo, na robótica, para a construção de robôs que aprendam a navegar por um terreno explorando o mesmo; na medicina, para a construção de sistemas que aprendam a partir do histórico de vários pacientes a realizar um diagnóstico mais confiável de novos pacientes; e nos sistemas de informação, pois permitem desenvolver sistemas que se adaptem às preferências dos usuários.

Sistemas de aprendizagem de máquina devem operar em conjuntos de dados do mundo real, que normalmente são ruidosos e/ou incompletos. Estes sistemas precisam lidar com a incerteza uma vez que o conhecimento disponível sobre o ambiente nem sempre é completo, dada a complexidade do mundo real (RUSSELL; NORVIG, 2003). A aprendizagem probabilística ou estatística utiliza a teoria da probabilidade para lidar com essa incerteza.

Em geral, as técnicas de aprendizagem estatística assumem que elas trabalham com apenas um tipo de objeto e que as instâncias de objetos não possuem nenhuma relação entre elas. Contudo, este pressuposto nem sempre é verdadeiro uma vez que, em muitos casos, existem relacionamentos entre as instâncias de dados. Por exemplo, páginas Web relacionam-se umas com as outras através dos *hyperlinks* e filmes podem estar relacionados por terem o mesmo diretor e/ou ator principal.

A aprendizagem relacional utiliza formalismos como lógica de primeira ordem para lidar com conjuntos de dados relacionais. Uma abordagem popular da aprendizagem relacional é a PLI (Programação em Lógica Indutiva) (MUGGLETON; RAEDT, 1994). A PLI visa aprender um programa lógico a partir de um conjunto de exemplos positivos e negativos e, possivelmente a partir de uma base de conhecimento.

Lehmann e Hitzler (2007) adaptaram a PLI para aprender definições em lógica de descrições, aplicando esta técnica na aprendizagem de ontologias a partir de dados estruturados (exemplos fornecidos pelos usuários).

Métodos de aprendizagem relacional não eram capazes de lidar com a incerteza tão bem quanto a aprendizagem estatística. Uma vez que a lógica de primeira ordem nos permite representar uma vasta variedade de conhecimentos acerca de um domínio (entidades e seus relacionamentos no domínio) e a probabilidade é um arcabouço comum para lidar com a incerteza, o interesse em técnicas de aprendizagem estatística relacional (KERSTING, 2006) (RAEDT; KERSTING, 2003) (NEVILLE; RATTIGAN; JENSEN, 2003) (POPESCUL et al., 2003) tem crescido recentemente. A aprendizagem estatística relacional combina o poder expressivo de formalismos de representação de conhecimento tais como a lógica de primeira ordem com abordagens probabilísticas de aprendizagem.

Dados os bons resultados alcançados por (LEHMANN; HITZLER, 2007) utilizando a PLI na aprendizagem de ontologias e os resultados promissores da PLI probabilística em relação aos métodos tradicionais, esta última configura-se como uma abordagem promissora para a aprendizagem de ontologias a partir de fontes textuais.

Para ilustrar a relevância das técnicas de aprendizagem relacional, suponha o exemplo a seguir. Suponha que se deseja determinar o tópico de uma página Web. Técnicas tradicionais de aprendizagem estatística o fariam com base nos seus atributos (i.e., suas palavras chave). Entretanto uma informação importante para este problema é omitida: os *links*. Páginas da Web tendem a referenciar páginas semelhantes. Essa informação pode ser utilizada por abordagens de aprendizagem relacional, porém, a observação de que existe um *link* entre duas páginas apenas aumenta a probabilidade de que elas possuam o mesmo tópico. A aprendizagem estatística relacional é capaz de representar de forma compacta os atributos e os relacionamentos entre os objetos e ainda capturar informação estatística sobre os relacionamentos.

Muitas abordagens de aprendizagem estatística relacional foram propostas na literatura. A maioria delas combina modelos gráficos probabilísticos, como as redes Bayesianas e as redes de Markov, com formalismos de representação do conhecimento como a lógica de primeira ordem (KERSTING; RAEDT, 2001), sistemas baseados em *frames* (FRIEDMAN et al., 1999) ou linguagens de consultas a bancos de dados (POPESCUL; UNGAR, 2003). Exemplos de tais abordagens são as Redes Lógicas de Markov (RICHARDSON; DOMINGOS, 2006), os Programas Lógicos Bayesianos (KERSTING; RAEDT, 2001), os Modelos

Probabilísticos Relacionais (FRIEDMAN; GETOOR; KOLLER; PFEFFER, 1999), a Regressão Logística Estrutural (POPESCUL; UNGAR, 2003) e os Programas Lógicos Estocásticos (MUGGLETON, 1996).

Dentre estas abordagens, as Redes Lógicas de Markov encontram-se em um estágio maduro de desenvolvimento, já dispondo inclusive de um software eficiente e bem documentado implementando diversos algoritmos de aprendizagem e inferência, o Alchemy (KOK et al., 2009). Além disso, em alguns estudos realizados, as Redes Lógicas de Markov mostraram resultados promissores para a automatização da aquisição de conhecimento (KOK; DOMINGOS, 2008) (WU; WELD, 2008). Assim essa foi a abordagem escolhida para este trabalho.

Este capítulo apresenta as Redes Lógicas de Markov (RLM) e está organizado como a seguir. A Seção 3.1 apresenta alguns conceitos básicos sobre a lógica de primeira ordem, o formalismo de representação de conhecimento utilizado pelas RLMs. A Seção 3.2 introduz brevemente as redes de Markov, o modelo probabilístico utilizado pelas RLMs. Por fim, a Seção 3.3 apresenta as RLMs discutindo o formalismo e algumas abordagens para aprendizagem e inferência em Lógica de Markov.

3.1 Lógica de primeira ordem

O objetivo desta seção não é realizar uma revisão exaustiva dos conceitos da Lógica de Primeira ordem, e sim esclarecer aqueles conceitos importantes para a compreensão da Lógica de Markov.

A lógica de primeira ordem é um formalismo de representação de conhecimento capaz de representar os objetos de um domínio e seus relacionamentos. Uma base de conhecimento de primeira ordem é um conjunto de sentenças (ou fórmulas) em lógica de primeira ordem. Cada sentença representa uma restrição rígida sobre o conjunto de mundos possíveis, ou seja, um mundo que viole uma restrição é impossível.

Uma sentença pode ser escrita utilizando predicados, funções, constantes e variáveis. As constantes representam objetos em um dado domínio (por exemplo, países como *Brasil, Uruguai*, *Alemanha* e *Estados Unidos* ou cidades como *São*

Luís, *Brasília* e *Nova Iorque*) enquanto variáveis são definidas a partir desses objetos. Os predicados representam relacionamentos entre esses objetos (e.g. *Vizinhos*) ou atributos dos objetos (e.g. *PaísEmDesenvolvimento*). Por fim, as funções mapeiam objetos ou tuplas de objetos a outros objetos (e.g. *Capital(Brasil)* = *Brasília*).

O bloco básico de construção de uma sentença é a sentença atômica ou átomo. Uma sentença atômica é um símbolo de predicado aplicado a uma tupla de termos. Um termo por sua vez é uma expressão que representa um objeto em um domínio, podendo ser uma constante, uma variável ou uma função aplicada a uma tupla de termos. Exemplos de sentenças atômicas são *Vizinhos(X,Y), PaísEmDesenvolvimento(Brasil)*, *Vizinhos(Capital(Brasil),Taguatinga)*. A forma básica de uma sentença, átomo ou predicado é obtida substituindo todas as variáveis por constantes.

As sentenças são construídas a partir de sentenças atômicas utilizando quantificadores e conectores. Os conectores utilizados são a negação (¬), a conjunção (∧), a disjunção (∨), a implicação (⟹) e a equivalência (⟺). Os quantificadores utilizados são o quantificador universal (∀) e o quantificador existencial (∃).

Por fim, dados um conjunto de predicados e funções e um conjunto de constantes, um mundo é uma atribuição de um valor verdade (*verdadeiro* ou *falso*) para cada um dos possíveis átomos básicos.

Uma discussão mais detalhada sobre a lógica de primeira ordem pode ser encontrada em (RUSSELL; NORVIG, 2003).

3.2 Redes de Markov

Uma rede de Markov (KINDERMANN; SNELL, 1980) representa uma distribuição de probabilidade conjunta de um conjunto de variáveis $X = (X_1, X_2, \dots, X_n)$. Ela é composta por um grafo não dirigido onde cada nó representa uma variável e as arestas representam relações entre elas. Além do grafo há um conjunto de funções potenciais, uma para cada clique no grafo. Um

clique é um subconjunto dos nós tal que para cada par de nós no clique exista uma aresta conectando eles. Em outras palavras, um clique de um grafo *G* é um subgrafo totalmente conectado formado apenas por nós presentes em *G*. Uma função potencial é uma função do estado do clique definida arbitrariamente determinando a probabilidade de cada estado. Assim sendo, a distribuição de probabilidade representada por uma rede de Markov é dada pela equação 6.

$$P(X = x) = \frac{1}{Z}\prod_{c}\phi_c\left(x_{\{c\}}\right) \tag{6}$$

Na equação 6, $x_{\{c\}}$ é o estado do clique *c* e *Z*, dado por $Z = \sum_{x\in X}\prod_c \phi_c\left(x_{\{c\}}\right)$, é uma constante de normalização. As redes de Markov são freqüentemente representadas como modelos logarítmicos lineares, o que é especialmente útil para a compreensão do seu papel na Lógica de Markov. Nesta representação, cada potencial de um clique é representado como uma soma ponderada de características do estado. Uma característica pode ser qualquer função real do estado de um clique. Assim a equação 6 pode ser reescrita como a equação 7.

$$P(X = x) = \frac{1}{Z}\exp\sum_{j} w_j f_j(x) \tag{7}$$

onde *j* é um clique no grafo, w_j é um peso associado ao clique *j* e f_j é uma característica (função potencial) do clique *j*. O valor $f_j(x)$ denota o valor da função potencial do clique *j* para o seu estado onde $X = x$.

3.3 Redes Lógicas de Markov

As Redes Lógicas de Markov (RICHARDSON; DOMINGOS, 2006) ou RLM combinam uma classe de modelos probabilísticos gráficos, as redes de

Markov, para lidar com a incerteza, com a expressividade da lógica de primeira ordem para representar de forma compacta uma grande variedade de conhecimento.

Uma base de conhecimento em lógica de primeira ordem é um conjunto de restrições rígidas sobre o conjunto de mundos possíveis, ou seja, mundos que violem alguma fórmula são impossíveis. Uma RLM, no entanto, é um conjunto de restrições relaxadas, o que significa que mundos que violem uma fórmula são menos prováveis, mas não impossíveis. Isto é representado associando pesos a cada uma das fórmulas. Quanto mais alto o peso da fórmula violada, menos provável o mundo é. Assim, uma restrição rígida pode ser representada por uma fórmula com peso infinito e uma base de conhecimento em lógica de primeira ordem pode ser vista como um caso específico de uma RLM onde todas as fórmulas possuem pesos infinitos.

Mais formalmente uma RLM pode ser definida com um conjunto de pares (*F,w*) onde:

a) *F* é uma fórmula em lógica de primeira ordem;

b) *w* é um número real representando o peso da fórmula *F*.

A Tabela 2 mostra um exemplo de uma pequena RLM para prever a relevância de páginas Web baseando-se nos seus links. A primeira fórmula indica que se existe um link entre duas páginas, então elas possuem o mesmo tópico e a segunda indica que se duas páginas possuem o mesmo tópico, então dizer que uma delas é relevante é equivalente a dizer que a outra também é relevante.

Tabela 2 - Exemplo de uma Rede Lógica de Markov

Fórmula	Peso
$\forall_{x,y} Link(x,y) \Rightarrow MesmoTopico(x,y)$	1.1
$\forall_{x,y} MesmoTopico(x,y) \Rightarrow (Relevante(x) \Leftrightarrow Relevante(y))$	1.6

Uma RLM pode ser vista como um *template* para a construção de redes de Markov, possibilitando assim inferência probabilística. Uma rede de Markov é construída a partir de uma RLM e um conjunto de constantes criando-se um nó para cada átomo básico. Então é adicionada uma aresta entre cada par de nós presentes na mesma fórmula. Deste modo, cada fórmula gera um clique no grafo.

A Figura 6 mostra a rede de Markov gerada a partir da RLM da Tabela 2 e de um conjunto simples de constantes formados pelas páginas *A* e *B*.

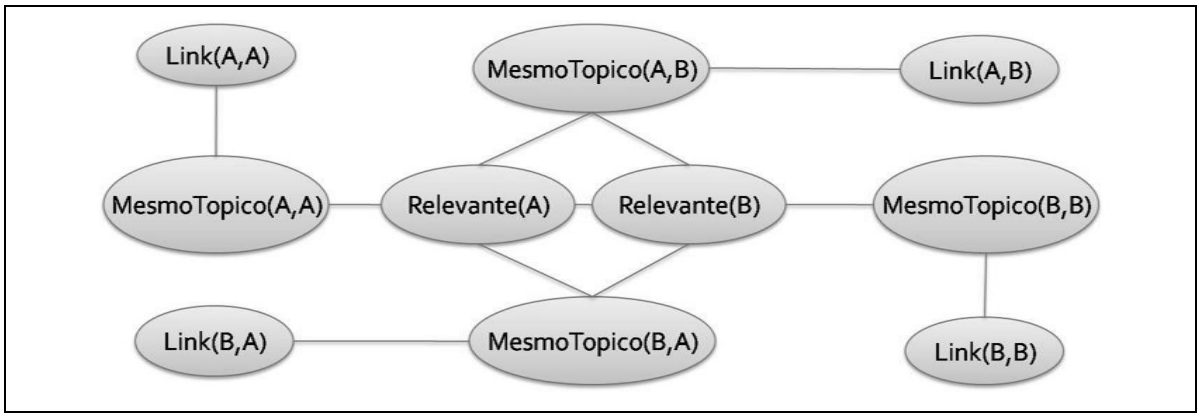


Figura 6 - Exemplo de uma rede de Markov gerada a partir da RLM da Tabela 2

De acordo com a equação 7, é possível definir a probabilidade de um dado mundo (i.e. uma atribuição de um valor verdade para cada átomo básico) como na equação 8.

$$P(x) = \frac{1}{Z} \exp \sum_j w_j n_j(x) \tag{8}$$

Na equação 8, a característica $n_j(x)$ é o número de formas básicas da fórmula *j* que são satisfeitas pelo mundo x e w_j é o peso da fórmula *j.*

A lógica de Markov suporta tanto a hipótese do mundo fechado quanto a hipótese do mundo aberto. Na hipótese do mundo fechado, os valores verdade de todos os átomos básicos são conhecidos. Já na hipótese do mundo aberto, os valores verdade de instanciações de átomos que não são fornecidos previamente são considerados desconhecidos.

3.3.1 Inferência

Os predicados de uma RLM podem ser separados em dois grupos: os predicados observados e os não observados. Os predicados observados (ou de evidência) são predicados cujos valores verdade dos respectivos predicados básicos são conhecidos previamente quando a inferência é realizada. O segundo grupo é formado pelos predicados não observados, ou seja, predicados cujos valores

verdade dos predicados básicos são desconhecidos. O objetivo do processo de inferência é inferir os valores verdade das formas básicas de um subconjunto de predicados não observados. Estes predicados são chamados de predicados de consulta.

Duas tarefas comuns de inferência em Lógica de Markov são as inferências MAP (maximum a posteriori) e probabilística. A inferência probabilística visa determinar a probabilidade de uma fórmula dado um conjunto de constantes e, talvez, outras fórmulas como evidência. A probabilidade de uma fórmula é a soma das probabilidades dos mundos onde ela é satisfeita. Calcular tais probabilidades pode ser muito caro computacionalmente, assim métodos aproximativos como a inferência MCMC (do inglês *Markov chain Monte Carlo*) (GILKS; RICHARDSON; SPIEGELHALTER, 1996) apresentam-se como uma alternativa viável e são geralmente utilizadas em combinação com o algoritmo MC-SAT (POON; DOMINGOS, 2006). Uma vez que a inferência MAP é de maior interesse para este trabalho, a inferência probabilística não será discutida em maiores detalhes aqui.

A inferência MAP visa encontrar o estado do mundo mais provável dada alguma evidência ou, mais formalmente, dado um conjunto de evidências *x* encontrar o estado do mundo *y* dado pela equação 9.

$$\max_{y} P(y|x) \tag{9}$$

A evidência em lógica de Markov é o conjunto de predicados básicos de evidência. Assim, dados uma RLM e um conjunto de predicados básicos de evidência (e seus respectivos valores verdade), a inferência MAP encontra o mundo mais provável. A partir das equações 8 e 9, a inferência MAP em lógica de Markov pode ser definida pela equação 10.

$$\max_{y} \frac{1}{Z} \exp \sum_{j} w_j n_j(x, y) \tag{10}$$

Uma vez que maximizar a expressão $\frac{1}{Z} \exp \sum_j w_j n_j(x, y)$ é o mesmo que maximizar o expoente da mesma, a equação 10 pode ser reescrita como na equação 11.

$$\max_{y} \sum_{j} w_j n_j(x, y) \tag{11}$$

A partir da equação 11 percebe-se que esta tarefa consiste em encontrar as atribuições de valores verdade que maximizem a soma dos pesos das fórmulas satisfeitas. Esta é exatamente a tarefa de um resolvedor SAT ponderado. Por exemplo variantes do algoritmo WalkSat (KAUTZ; SELMAN; JIANG, 1997) como o MaxWalkSat (KAUTZ; SELMAN; JIANG, 1997) têm sido utilizadas para esta tarefa (RICHARDSON; DOMINGOS, 2006).

O pseudocódigo do MaxWalkSat é mostrado no Algoritmo 1. Ele busca pela melhor solução escolhendo aleatoriamente uma sentença não satisfeita e invertendo o valor verdade de um dos átomos da mesma. O átomo a ser invertido é escolhido aleatoriamente ou então é escolhido o átomo que maximize o somatório da equação 11 quando invertido. A probabilidade de que o átomo a ser invertido seja escolhido aleatoriamente é dada por *p*. A função *Prob(solução,RLM)* computa o somatório $\sum_j w_j n_j(solução)$ para as cláusulas da RLM dada e a função *Uniforme(0,1)* retorna um valor do intervalo [0,1] escolhido aleatoriamente com distribuição uniforme.

Algoritmo 1 - MaxWalkSat

MaxWalkSat(RLM*, max-tentativas, max-flips, limite, p*)

```
para i ← 1 até max-tentativas faça
   solução = atribuição aleatória de valores verdade
   para j ← 1 até max-flips faça
      se prob(solução,RLM) > limite então
         retorne solução
      fim se
      c ← sentença não satisfeita escolhida aleatoriamente
      se  Uniforme(0,1) < p então
         inverte uma variável em c escolhida aleatoriamente
      senão
         inverte a variável em c que maximiza
            prob(solução,RLM)
      fim se
   fim para
fim para
retorne falha, melhor solução encontrada até aqui
```

3.3.2 Aprendizagem de pesos

Tanto a estrutura (fórmulas) quanto os pesos de uma RLM podem ser automaticamente aprendidas a partir dos dados. Uma vez que apenas os pesos são automaticamente aprendidos neste trabalho, a discussão aqui se concentrará na aprendizagem de pesos. Os dados de treinamento são um conjunto de constantes e predicados básicos.

O problema da aprendizagem de pesos em uma RLM pode ser visto como um problema de otimização numérica onde o objetivo é encontrar o conjunto de pesos para as fórmulas dadas que melhor explicam os dados no conjunto de treinamento. Há duas abordagens para aprender os pesos dado um conjunto de fórmulas e um conjunto de treinamento: a abordagem gerativa e a discriminativa.

A aprendizagem gerativa visa maximizar a verossimilhança conjunta de todos os predicados. As técnicas do estado da arte para esta tarefa utilizam o gradiente para buscar a solução ótima. O gradiente do logaritmo da função de probabilidade de um mundo dada pela equação 8 em relação aos pesos das fórmulas é dado pela equação 12.

$$\frac{\partial}{\partial w_i} \log P_w(x) = n_i(x) - E_w[n_i(x)] \tag{12}$$

Na equação 12 w_i denota o peso da fórmula *i* na RLM, $P_w(x)$ denota a probabilidade de um mundo x dado um conjunto de pesos *w*, $n_i(x)$ é o número de vezes que a fórmula *i* é satisfeita no conjunto de treinamento e $E_w[n_i(x)]$, a esperança do número de vezes que a fórmula *i* é satisfeita dada a distribuição de probabilidade P_w, ou seja, quantas vezes o modelo composto pelos pesos *w* prevê que a fórmula *i* será satisfeita.

Simplificando, o gradiente dado pela equação 12 é a diferença entre o número de vezes que a fórmula *i* é realmente satisfeita no conjunto de treinamento e o número de vezes que o modelo prevê que a fórmula será satisfeita. Isso nos diz que se a grandeza $\frac{\partial}{\partial w_i} \log P_w(x)$ for maior que 0 (o modelo prevê que a fórmula *i* será

satisfeita menos freqüentemente do que ela realmente é) o peso w_i deve aumentar, se for menor que zero w_i deve diminuir.

A desvantagem desta abordagem é que computar a grandeza $E_w[n_i(x)]$ é computacionalmente intratável, uma vez que exige um processo de inferência para cada um dos mundos possíveis. Uma saída para este problema é utilizar a função de pseudo-verossimilhança, uma aproximação da equação 12, cuja computação consome menos tempo de processamento. Entretanto, como mostrado por Domingos et al. (2008), esta abordagem pode levar, em alguns casos, a resultados ruins.

Uma alternativa à pseudo-verossimilhança é a aprendizagem discriminativa. Esta abordagem requer que se saiba, no momento da aprendizagem, quais predicados serão utilizados apenas como evidência e quais serão os predicados de consulta. Como este é o caso neste trabalho (como será mostrado mais adiante), a abordagem utilizada aqui é a aprendizagem discriminativa

A aprendizagem discriminativa visa maximizar a verossimilhança condicional dos predicados de consulta, dados os predicados de evidência. Ao particionar o conjunto de treinamento em um conjunto de átomos de consulta *Y* e um conjunto de átomos de evidência *X*, a equação 12 torna-se a equação 13.

$$\frac{\partial}{\partial w_i} \log P_w(y|x) = n_i(x,y) - E_w[n_i(x,y)] \tag{13}$$

Na equação 13 $P_w(y|x)$ é a probabilidade de um mundo y dado que o mundo x é verdadeiro e um conjunto de pesos *w*; $n_i(x,y)$ por sua vez é o número instanciações verdadeiras da fórmula *i* envolvendo algum átomo de consulta. A computação de $E_w[n_i(x,y)]$ também é computacionalmente intratável, assim como na abordagem gerativa. Entretanto, foi demonstrado que uma boa aproximação para $E_w[n_i(x,y)]$ é a contagem $n_i(x,y_w^*)$ onde y_w^* é o estado MAP $y_w^*(x)$, o estado mais provável de y dado x e pode ser computado através da inferência MAP, utilizando o Algoritmo 1. A grande vantagem dessa abordagem é que é possível aproximar $E_w[n_i(x,y)]$ apenas com uma inferência MAP (não necessitando que seja realizado um processo de inferência para cada mundo possível) e ainda assim obter resultados satisfatórios. Em Lowd e Domingos (2007) é possível encontrar uma discussão mais detalhadas e resultados sobre o tópico.

3.4 Considerações finais

Este capítulo apresentou uma breve discussão acerca das técnicas de aprendizagem estatística relacional. Esta abordagem da aprendizagem de máquina combina técnicas estatísticas com formalismos de representação do conhecimento, como a lógica de primeira ordem. Dentre os diversos modelos propostos para a aprendizagem estatística relacional, as redes lógicas de Markov foram discutidas em mais detalhes, uma vez que elas estão em uma fase madura de desenvolvimento e apresentam técnicas consolidadas para a sua aprendizagem e inferência, estando inclusive implementadas em um pacote de software livre, o Alchemy.

Como já discutido anteriormente, a aprendizagem estatística relacional apresenta-se como uma abordagem promissora para a aprendizagem de ontologias. O próximo capítulo apresenta um processo que utiliza redes lógicas de Markov para extrair a hierarquia de conceitos de uma ontologia a partir de um corpus em linguagem natural.

4 UM PROCESSO PARA A EXTRAÇÃO DE HIERARQUIAS DE CONCEITOS A PARTIR DE FONTES TEXTUAIS

As redes lógicas de Markov constituem uma abordagem promissora para a automatização do processo de aquisição de conhecimento. Neste capítulo é apresentada uma abordagem para a automatização parcial da construção de ontologias. Esta abordagem é constituída por um processo suportado por uma metodologia composta por duas técnicas, uma para a extração de conceitos e outra para a extração de hierarquias a partir de fontes textuais.

As fontes textuais utilizadas neste trabalho são escritas na língua inglesa. O motivo para se trabalhar com a língua inglesa é a existência de um vasto conjunto de ferramentas para suportar o processamento da linguagem natural em inglês, em especial o WordNet (FELLBAUM, 1998), cuja versão em português ainda não encontra-se disponível. O WordNet é uma base de dados que contém termos em linguagem natural, suas respectivas definições e seus relacionamentos semânticos como a sinonímia, hiperonímia e hiponímia.

O processo proposto para a aprendizagem de ontologias é mostrado na Figura 7. Uma vez que as RLMs trabalham com dados relacionais, os corpora em linguagem natural devem ser pré-processados para que sejam extraídos dados relacionais. Esta fase de pré-processamento é descrita na subseção 4.1.

Uma vez que o corpus esteja pré-processado, ele é utilizado como entrada para a etapa de *Extração de Conceitos* suportada pela técnica PRECE (*Probabilistic Relational Concept Extraction*) (DRUMOND; GIRARDI, 2009). A subseção 4.2 descreve a técnica PRECE e como ela extrai o conjunto C^C da equação 1 através das realizações léxicas das classes (representadas pelo conjunto *F*) a partir do corpus pré-processado. Em seguida os conceitos extraídos são organizados hierarquicamente na fase de *Extração de Hierarquia* produzindo assim o conjunto *H* da equação 1. Esta fase é suportada pela técnica PREHE (*Probabilistic Relational Hierarchy Extraction*) (DRUMOND; GIRARDI, 2010), descrita na subseção 4.3. Tanto a técnica PRECE quanto a técnica PREHE são baseadas em inferência em Lógica de Markov. Cada uma delas utiliza uma Rede Lógica de Markov (RLM) cuja estrutura foi definida manualmente e os parâmetros aprendidos automaticamente através da aprendizagem discriminativa durante a fase de

Aprendizagem de Pesos, discutida na subseção 4.4. As estruturas das RLMs utilizadas pelas técnicas PRECE e PREHE foram definidas manualmente devido à dificuldade em aprendê-las automaticamente a partir dos dados. A desvantagem dessa abordagem é que não há como garantir que a estrutura definida manualmente seja a melhor possível. Entretanto, o alto custo computacional da aprendizagem automática da estrutura de modelos como as RLMs e os maus resultados obtidos com estruturas aprendidas automaticamente justificam a definição manual das fórmulas utilizadas neste trabalho.

Por fim, o conjunto C^C extraído na fase de *Extração de Conceitos* e os relacionamentos taxonômicos (conjunto *H*) extraídos na fase de *Extração de Hierarquia* são escritos em uma linguagem de especificação de ontologias, o OWL, na fase de *Pós-processamento*.

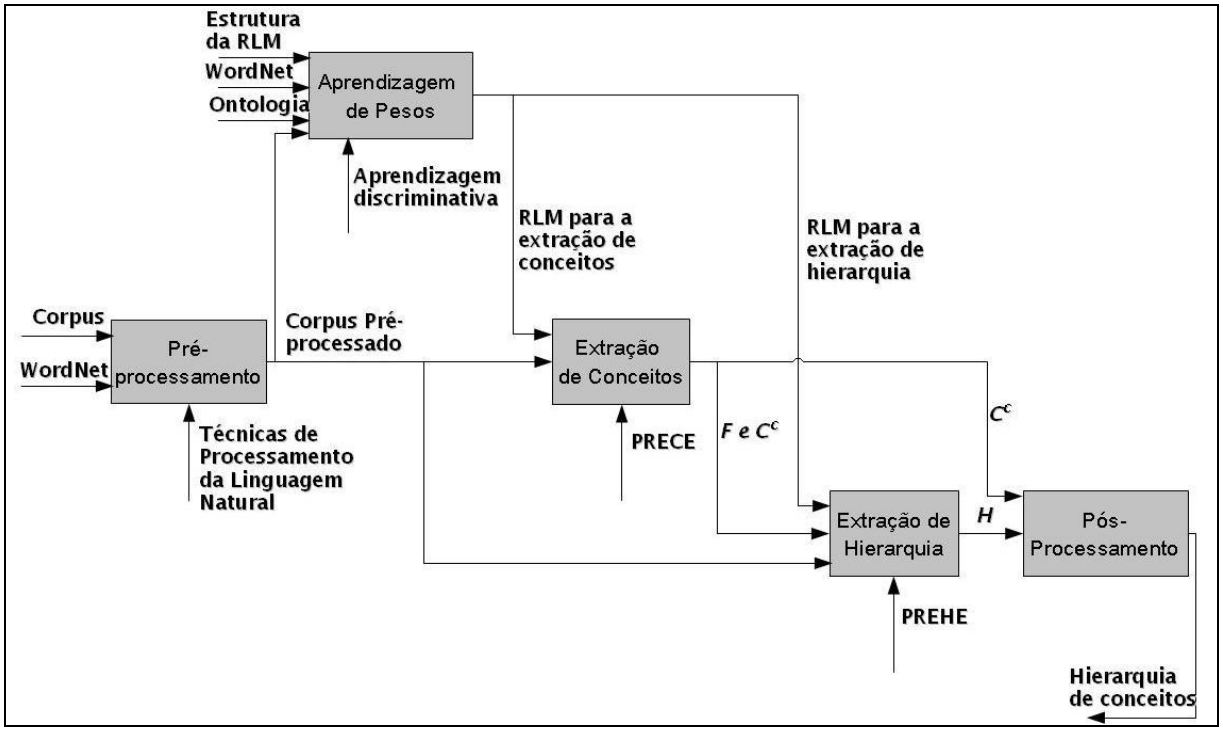


Figura 7 - Processo de aprendizagem de ontologias proposto

4.1 Pré-processamento do corpus

As técnicas de aprendizagem estatística relacional aprendem modelos para realizar predições baseadas nos atributos dos objetos e nos relacionamentos

entre os mesmos. Neste trabalho, os objetos são termos em linguagem natural e seus atributos são os documentos nos quais eles aparecem enquanto seus relacionamentos são representados pelas dependências sintáticas e pelo relacionamento semântico de hiperonímia. As dependências sintáticas são relacionamentos a nível sintático que as palavras possuem dentro das sentenças. Elas indicam, por exemplo, quem são o sujeito e o objeto de um determinado verbo ou qual substantivo é modificado por um dado adjetivo. O relacionamento de hiperonímia, por sua vez, é semelhante ao relacionamento *Tipo_de* entre conceitos de ontologias. Ele indica uma relação de generalização entre os termos. Por exemplo, o termo *automóvel* é uma hiperonímia do termo *carro*, uma vez que todo carro é um *tipo de* automóvel. A existência de tais relacionamentos no texto justifica a utilização de técnicas de aprendizagem estatística relacional ao invés de técnicas puramente estatísticas.

Os corpora utilizados no processo de aprendizagem de ontologias aqui proposto são pré-processados conforme ilustra a Figura 8. O corpus passa primeiramente por um processo de *tokenização*. O objetivo deste processo é dividir os textos em tokens, que podem ser palavras, pontuações ou sentenças, para prepará-los para as etapas subseqüentes do pré-processamento.

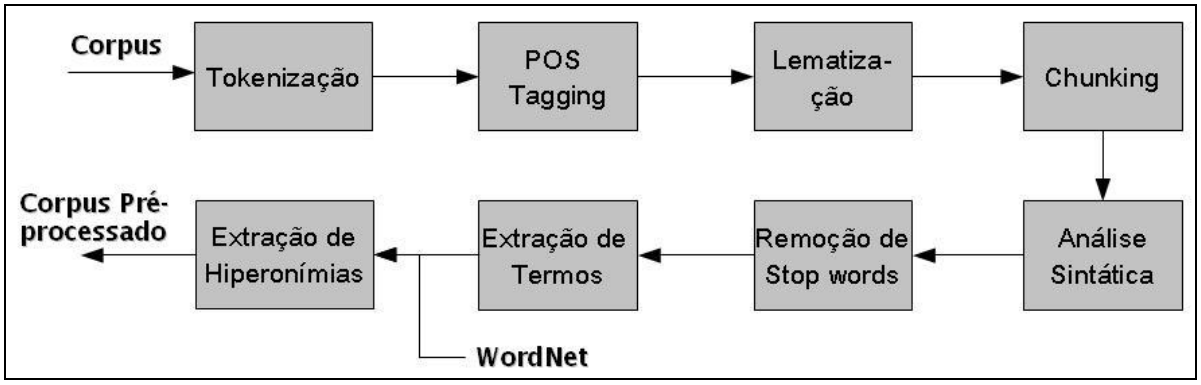


Figura 8 - Detalhamento do pré-processamento dos corpora

A próxima etapa do pré-processamento é a atribuição de *tags* de parte do discurso. Esta segunda etapa consiste em atribuir a cada palavra (identificada no processo de tokenização) um rótulo com a sua categoria sintática como, por exemplo, substantivo, verbo, adjetivo, etc. As *tags* utilizadas neste trabalho são aquelas do conjunto de *tags Penn Treebank* (MARCUS; SANTORINI; MARCINKIEWICZ, 1993) e são mostradas no Anexo A. Note que tais *tags* referem-

se à língua inglesa. Em seguida, os tokens passam por um processo de lematização, que consiste em encontrar a forma básica das palavras, chamada de lema, como por exemplo a forma no singular dos substantivos e o infinitivo dos verbos. Por exemplo a sentença "*Eu vi os carros dos vizinhos*" se tornaria, após o processo de lematização, "*Eu ver o carro do vizinho*". Depois da lematização, é realizado o *chunking*. O objetivo desta fase é descobrir os conjuntos de palavras que, juntas, formam uma unidade sintática. Esta tarefa é importante porque muitos conceitos são expressados em linguagem natural por mais de uma palavra, como ocorre com as frases nominais. Por exemplo, "museu de arte moderna" é uma especialização do conceito denotado por "museu".

O próximo passo é a realização da análise sintática para extrair as árvores sintáticas de cada sentença no corpus, identificando assim as dependências sintáticas entre as palavras. As dependências sintáticas consideradas neste trabalho são representadas de acordo com as dependências de Stanford (MARNEFFE; MANNING, 2008), mostradas no Anexo B.

Em seguida, é realizada a extração de termos. Em primeiro lugar, uma vez que o interesse aqui é extrair conceitos e conceitos são expressos por substantivos e frases nominais, palavras como preposições, advérbios, pronomes e verbos são removidos. Isto é feito observando as *tags* da parte do discurso. Os termos selecionados são ponderados e apenas os termos cujo peso está acima de um determinado valor são selecionados. Os termos são ponderados utilizando a medida de relevância descrita pela equação 4. As freqüências dos termos são computadas baseadas nos lemas dos termos. Por exemplo, se um documento possui uma ocorrência para o termo "célula" e uma para o termo "células", a freqüência para o lema "célula" é 2. Por último, são extraídos, a partir do WordNet, os relacionamentos de hiperonímia entre os termos selecionados.

Para ajudar na compreensão do processo, será utilizado como exemplo um corpus contendo dois documentos: *DOC1* e *DOC2*. Por razões de simplicidade cada um destes documentos contém apenas uma sentença, conforme mostrado na Tabela 3.

Tabela 3 - Documentos utilizados como exemplo no decorrer da seção

DOC1	*DOC2*
A country's capital is the most important city of the nation.	*Brazil's capital, Brasilia, is a modern city.*

A Tabela 4 mostra o documento *DOC1* após passar pela tokenização, atribuição de rótulos de parte do discurso e a lematização do pré-processamento da Figura 8. A tabela foi gerada utilizando o sistema *TreeTagger*[1].

Tabela 4 - Sentença após tokenização e anotação dos tokens com rótulos de parte do discurso e lemas

Token	*Parte do Discurso*	*Lema*
A	DT	a
Country	NN	country
's	POS	's
Capital	NN	capital
Is	VBZ	be
The	DT	the
Most	RBS	most
Important	JJ	important
City	NN	city
Of	IN	of
The	DT	the
Nation	NN	nation

A análise sintática, por sua vez, gera como resultados uma árvore sintática de onde são extraídas as dependências sintáticas. A árvore sintática gerada a partir do conteúdo do documento *DOC1* é mostrada na Figura 9.

As dependências sintáticas são representadas através do predicado *Depends(t_1,t_2,dep).* Este predicado significa que o termo t_1 governa o termo t_2 através da dependência do tipo *dep*. A lista dos tipos de dependências sintáticas utilizados aqui encontra-se no Anexo B. As dependências sintáticas extraídas a partir da árvore da Figura 9 e da árvore extraída a partir do documento *DOC2* são mostradas na Tabela 5, ambas geradas utilizando o analisador sintático *Stanford Parser*[2].

Tabela 5 - Exemplos de dependências sintáticas

Depends("city","capital",NSUBJ)
Depends("capital","country",POSS)
Depends("capital","brazil",POSS)
Depends("capital","brasilia",APPOS)

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[2] http://nlp.stanford.edu/software/lex-parser.shtml

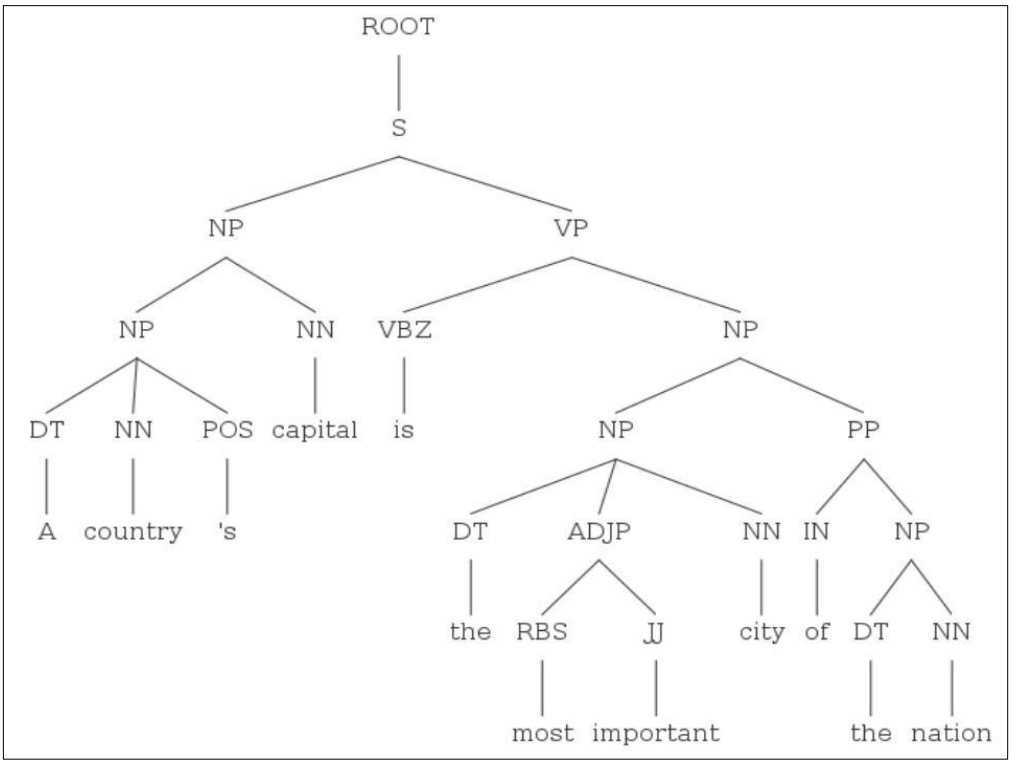


Figura 9 - Exemplo de árvore sintática

O resultado final do pré-processamento do corpus da Tabela 3 é mostrado na Tabela 6.

Tabela 6 - Exemplo de Corpus Pré-processado

HasTerm(DOC1,"country")	*HasTerm(DOC2,"city")*
HasTerm(DOC1,"capital")	*Depends("city","capital",NSUBJ)*
HasTerm(DOC1,"city")	*Depends("capital","country",POSS)*
HasTerm(DOC1,"nation")	*Depends("capital","brazil",POSS)*
HasTerm(DOC2,"brazil")	*Depends("capital","brasilia",APPOS)*
HasTerm(DOC2,"capital")	*Hypernym("city","capital")*
HasTerm(DOC2,"brasilia")	

4.2 Extração de conceitos através da técnica PRECE

Esta seção descreve a PRECE (*Probabilistic RElational Concept Extraction*), uma técnica para extrair conceitos de ontologias a partir de corpora em linguagem natural, que usa aprendizagem probabilística relacional.

O problema aqui é aprender o conjunto de conceitos C^C de uma ontologia (Definição 1) a partir de um dado corpus de texto *D*. A abordagem proposta aqui

aprende conceitos através das suas realizações lingüísticas. De acordo com o que já foi discutido anteriormente, as realizações lingüísticas de um conceito são termos em linguagem natural cujo significado é o conceito em questão. Uma vez que o conjunto de realizações lingüísticas de um conceito mapeia termos em linguagem natural ao conceito, o objetivo desta tarefa de aprendizagem é aprender o conjunto C^C da definição 1 através da aprendizagem do conjunto *F* da definição 2. Esta tarefa é realizada por dois passos: Identificação de Conceitos e Nomeação dos Conceitos, como mostrado na Figura 10.

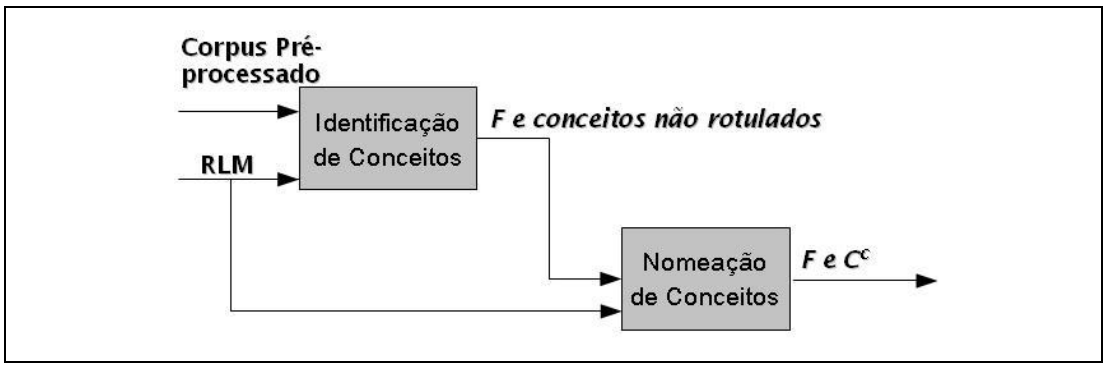


Figura 10 - Passos da Técnica PRECE

4.2.1 Identificação de conceitos

Uma vez que a técnica PRECE extrai conceitos a partir de suas realizações léxicas, nesta etapa do trabalho um conceito é considerado como um conjunto de termos em linguagem natural que possuem o mesmo significado. Para extrair tais conjuntos, a técnica PRECE utiliza aprendizagem estatística relacional para realizar uma tarefa de reconhecimento de entidade. Esta tarefa, também conhecida como identificação de objetos, consiste em descobrir quais registros representam um mesmo objeto do mundo real, para assim, remover redundâncias, compactar bancos de dados ou auxiliar o processo de tomada de decisão. Em lógica de Markov, esse problema pode ser formulado definindo-se um predicado de igualdade e definindo fórmulas que relacionem objetos com as mesmas características através do predicado de igualdade. A aprendizagem estatística

relacional tem apresentado bons resultados nesta tarefa como mostrado por Singla e Domingos (2004).

No contexto da técnica PRECE, o objetivo do reconhecimento de entidades é descobrir quais termos referem-se ao mesmo conceito do mundo real. O fato de dois termos representarem o mesmo conceito (o predicado de igualdade) é representado pelo predicado *SameMeaning(term$_1$,term$_2$)*.

A identificação de conceitos é realizada através de inferência em Lógica de Markov. O objetivo do processo de inferência é inferir os valores verdade para as possíveis instanciações do predicado *SameMeaning* baseado na evidência. A evidência utilizada é composta pelos atributos e relacionamentos entre os termos extraídos durante a fase de pré-processamento. Uma vez que a lógica de Markov provê um formalismo que possibilita realizar inferência estatística e representar relacionamentos entre os termos, a evidência é composta por um conjunto de predicados. A ocorrência de um termo em um documento é representada pelo predicado *HasTerm(document,term)*. As dependências sintáticas entre os termos são representadas pelo predicado *Depends(term$_1$,term$_2$,dependency)* enquanto que o predicado *Hypernym(term$_1$,term$_2$)* reflete a relação de hiperonímia entre os termos. A Tabela 7 mostra um resumo dos predicados utilizados pela técnica PRECE e suas interpretações.

O arquivo de evidências, gerado a partir do corpus da Tabela 3, com as instanciações dos predicados *HasTerm, Depends* e *Hypernym* é mostrado na Tabela 6. É importante frisar que o valor deste exemplo é meramente didático, uma vez que a técnica PRECE requer uma grande quantidade de texto para operar corretamente.

Tabela 7 - Predicados da RLM utilizada pela técnica PRECE

Predicado	Interpretação
HasTerm(document,term)	Captura a distribuição de probabilidade de termos e documentos
Depends(t_1,t_2,dependency)	Indica que o termo t_1 governa o termo t_2 através da dependência do tipo *dependency*
Hypernym(t_1,t_2)	Indica que o termo t_1 é uma hiperonímia do termo t_2
SameMeaning(term,term)	Indica que dois termos possuem o mesmo significado, i.e. referem-se ao mesmo conceito

Conforme o que foi apresentado no Capítulo 3, uma RLM é composta por um conjunto de fórmulas em lógica de primeira ordem (a sua estrutura) e seus

respectivos pesos (os parâmetros). A estrutura da RLM utilizada pela técnica PRECE é mostrada na Tabela 8. As fórmulas desta RLM foram definidas manualmente, observando-se algumas regularidades nos corpora utilizados e codificando idéias de técnicas do estado da arte através da lógica de primeira ordem.

Tabela 8 - Estrutura da RLM utilizada pela técnica PRECE

Identificador	*Fórmula*
1	$HasTerm(d, t_1) \wedge HasTerm(d, t_2) \Rightarrow SameMeaning(t_1, t_2)$
2	$Hypernym(t_3, t_1) \wedge Hypernym(t_3, t_2) \Rightarrow SameMeaning(t_1, t_2)$
3	$Hypernym(t_3, t_1) \wedge Hypernym(t_4, t_2) \wedge \neg SameMeaning(t_3, t_4) \Rightarrow \neg SameMeaning(t_1, t_2)$
4	$Depends(t_3, t_1, +dep) \wedge Depends(t_3, t_2, +dep) \Rightarrow SameMeaning(t_1, t_2)$
5	$\neg SameMeaning(t_1, t_2)$

É importante frisar que, apesar de escritas utilizando a sintaxe da lógica de primeira ordem, as fórmulas da Tabela 8 não são regras em lógica de primeira ordem. Elas são sentenças com uma determinada probabilidade de serem verdadeiras. Esta probabilidade é dada em função dos pesos das mesmas, automaticamente aprendidos durante a fase de Aprendizagem de Pesos da Figura 7. A versão completa da RLM (sua estrutura e os pesos) encontra-se no Apêndice A.

A fórmula 1 da Tabela 8 expressa que dois termos que aparecem em um mesmo documento possuem uma probabilidade de estarem relacionados ao mesmo conceito. Isso ocorre porque os autores tendem a evitar repetir excessivamente o mesmo termo em um mesmo texto, utilizando sinônimos para se referenciar a conceitos recorrentes em um texto. Isto pode ser observado no documento *DOC1* do corpus da Tabela 3. Os termos *country* e *nation* são utilizados para referenciar o mesmo conceito sem utilizar o mesmo termo duas vezes na mesma frase.

Da mesma forma, é possível perceber intuitivamente que dois termos distintos que possuem uma relação de hiperonímia com o mesmo termo, têm uma maior probabilidade de estarem relacionados ao mesmo conceito. Este fato ocorre com os termos *country* e *nation* do exemplo dado. Ambos possuem a mesma hiperonímia, *political entity* e possuem o mesmo significado. Isto é representado na fórmula 2 da Tabela 8.

Assim como a nossa intuição nos diz que termos que possuem a mesma hiperonímia são mais prováveis de estarem relacionados ao mesmo conceito, o fato de dois temos possuírem hiperonímias diferentes é uma evidência de que ambos não possuem o mesmo significado, como expressa a fórmula 3 da Tabela 8.

As dependências sintáticas são representadas pelo predicado *Depends.* De acordo com a hipótese de Harris (HARRIS, 1968), duas palavras com significado similar tendem a aparecer em contextos similares. Uma maneira de expressar isto é afirmar que duas palavras que possuem as mesmas dependências sintáticas com as mesmas palavras estão mais inclinadas a denotarem o mesmo conceito. Isto pode ser observado no exemplo dado nos termos *Brazil* e *country*. Na Tabela 6 é possível perceber que ambos estão relacionados ao termo *capital* através da dependência *POSS*. Isto é expresso pela fórmula 4 da Tabela 8.

Algumas dependências sintáticas são mais significativas do que outras para determinar se dois termos possuem significados similares. Por exemplo, não espera-se que uma dependência entre um substantivo e uma conjunção seja muito informativa. Por isso a fórmula 4 da Tabela 8 apresenta a notação +*dep*. Neste caso, o operador + indica que um peso diferente será calculado para cada substituição do termo +*dep*, por cada tipo de dependência sintática do Anexo B.

Para evitar que uma quantidade excessiva de termos seja designada a um conceito, a fórmula 5 da Tabela 8 é adicionada à RLM. Essa sentença, juntamente com o seu peso, captura a probabilidade de que dois termos escolhidos aleatoriamente possuam o mesmo significado. O sinal de negação indica que quanto maior o peso atribuído a essa fórmula, menor o número de formas básicas do predicado *SameMeaning* serão consideradas verdadeiras. Vale ressaltar que o peso desta fórmula é estimado automaticamente a partir de um conjunto de treinamento como mostrado na Seção 4.4

Os conceitos são extraídos realizando inferência probabilística na RLM apresentada. Os predicados de evidência *HasTerm, Depends* e *Hypernym* seguem a hipótese de mundo fechado. As instanciações destes predicados são extraídas automaticamente do corpus durante as fases de pré-processamento e extração de termos. O predicado de consulta é o predicado *SameMeaning*.

O objetivo do processo de inferência é determinar os valores verdade mais prováveis para as instanciações do predicado *SameMeaning*. Isto é feito através da inferência MAP, descrita na Subseção 3.3.1. A PRECE utiliza o Algoritmo

1 para realizar a inferência MAP. Nos exemplos aqui discutidos, dado o arquivo de evidências da Tabela 6 e a RLM composta pelas fórmulas apresentadas aqui e seus respectivos pesos, o processo de inferência retorna o arquivo da Tabela 9, onde as instanciações que nele aparecem são consideradas verdadeiras e todas as outras, falsas.

Tabela 9 - Exemplo de resultado do processo de inferência

SameMeaning("country","country")	*SameMeaning("capital","capital")*
SameMeaning("country","nation")	*SameMeaning("capital","brasilia")*
SameMeaning("country","brazil")	*SameMeaning("nation","country")*
SameMeaning("city","city")	*SameMeaning("nation","nation")*
SameMeaning("city","capital")	*SameMeaning("nation","brazil")*
SameMeaning("city","brasilia")	

O arquivo da Tabela 9 indica quais termos possuem o mesmo significado, mas não indica explicitamente os conceitos como grupos de termos. Essa indicação é feita pelo predicado *F(conceito,termo)*. Um átomo *F(c,t)* indica que o termo *t* é uma realização léxica do conceito *c*, onde *c* é um identificador único do conceito. Assim, se dois termos t_1 e t_2 possuem o mesmo significado, i.e. o átomo $SameMeaning(t_1,t_2)$ foi considerado verdadeiro, então é criado um conceito *c* e os átomos $F(c, t_1)$ e $F(c, t_2)$. Cada conceito *c* criado faz parte do conjunto de conceitos não rotulados, indicados como parte da saída da fase de Identificação de Conceitos na Figura 10. A outra parte da saída, é o conjunto de instanciações do predicado *F*. O arquivo contendo estas instanciações, gerado a partir do arquivo da Tabela 9 é mostrado na Tabela 10.

Tabela 10 - Arquivo contendo a saída da técnica PRECE

F(1,"country")	*F(2,"capital")*
F(1,"nation")	*F(2,"brasilia")*
F(1,"brazil")	*F(3,"capital")*
F(2,"city")	*F(3,"brasilia")*

4.2.2 Nomeação de conceitos

Uma vez que os conceitos aprendidos utilizando a abordagem apresentada aqui são representados apenas por conjuntos de palavras, nenhum

rótulo é designado a eles. Uma forma de rotular os conceitos é extrair tais rótulos a partir de um corpus de texto com conhecimento básico no domínio desejado. Isto é feito extraindo termos deste corpus e classificando-os de acordo com os conceitos extraídos.

Entretanto tais corpora não são tão fáceis de obter. Uma maneira alternativa de realizar isto é considerar os pesos dos termos. Existem várias abordagens para ponderar os pesos. Este trabalho utiliza uma medida baseada na entropia de uma variável definida por Shannon (1948). Kao et al. (2004) adaptaram a entropia para ser utilizada como uma medida de peso de termos em linguagem natural. Dado um termo *t* a sua entropia *H(t)* é definida pela equação 14.

$$H(t_i) = -\sum_{j=1}^{|D|} \mathrm{w_{ij}} * \log_{|D|} w_{ij} \qquad (14)$$

Na equação 14, w_{ij} é o peso normalizado do termo t_i em relação ao documento *j* e |*D*| é a quantidade de documentos no corpus. O peso w_{ij} utilizado para os termos é simplesmente a freqüência do termo t_i no documento j, normalizada para que $0 \leq w_{ij} \leq 1$. Assim, o rótulo de um conceito *c* é o termo *t* que possua a menor entropia *H(t)* dentre os termos para os quais *F(c,t)* é verdadeiro.

Devido ao seu tamanho, o corpus da Tabela 3 não fornece um bom exemplo da aplicação da equação 14 na nomeação de conceitos. Portanto A aplicação da entropia para a nomeação de conceitos é exemplificada aqui utilizando dados extraídos do corpus *LonelyPlanet* (KAVALEC; SVATEC, 2005). Ao aplicar a técnica PRECE neste corpus, os termos *city* e *town* foram agrupados no mesmo conceito. A entropia calculada para o termo *city* foi de 0.7272 e para o termo *town* foi de 0.799621. Por isso o rótulo escolhido para este conceito foi o termo *city*, uma vez que possui a menor entropia.

O conjunto de conceitos extraídos na fase de identificação de conceitos, cada um com seu rótulo atribuído na Nomeação de Conceitos, constituem o conjunto C^C e, junto com as instanciações do predicado *F*, constituem as saída da técnica PRECE indicada na Figura 7.

4.3 Extração de hierarquia através da técnica PREHE

Uma vez extraídos os conceitos, o problema é organizá-los hierarquicamente, ou seja, aprender o conjunto *H* da equação 1. Esta tarefa pode ser definida como uma atividade de predição de links. A predição de links consiste em descobrir se uma relação existe entre dois objetos ou, em outras palavras, dada um relacionamento *r*, descobrir os objetos *x* e *y* que se relacionam através de *r*, i.e., para os quais *r(x,y)* é verdadeira. Em Popescul e Ungar (2003) a aprendizagem estatística relacional foi utilizada para realizar a tarefa de predição de links em ambientes com dados ruidosos e incompletos. No contexto da extração de hierarquias, o relacionamento *r* é representado pelo predicado *Kind_Of* e os objetos são as classes em C^C extraídas pela técnica PRECE.

A técnica PREHE (*Probabilistic RElational Hierarchy Extraction*) visa encontrar as atribuições de valores de verdade mais prováveis para as formas básicas do predicado de consulta *Kind_Of(concept,concept)*, dada a evidência. Esta é claramente uma tarefa de inferência MAP. A evidência é composta pelas formas básicas dos predicados *Hypernym* e *Depends*, também utilizados pela PRECE bem como o predicado *F*, resultado da inferência na fase de extração de conceitos. Um exemplo de arquivo de entrada gerado a partir dos resultados do exemplo dado na seção anterior é mostrado na Tabela 11. A partir desse arquivo, a técnica PREHE extrairia apenas o relacionamento taxonômico *Kind_Of(2,3)*.

Tabela 11 - Exemplo de arquivo contendo instanciações de predicados de evidência da técnica PREHE

F(1,“country”)	*F(3,“brasilia”)*
F(1,“nation”)	*Depends(“city”,“capital”,NSUBJ)*
F(1,“brazil”)	*Depends(“capital”,“country”,POSS)*
F(2,“city”)	*Depends(“capital”,“brazil”,POSS)*
F(2,“capital”)	*Depends(“capital”,“brasilia”,APPOS)*
F(2,“brasilia”)	*Hypernym(“city”,“capital”)*
F(3,“capital”)	

A estrutura da RLM utilizada pela técnica PREHE é mostrada na Tabela 12. Assim como na técnica PRECE, as fórmulas desta RLM foram definidas manualmente, codificando idéias de técnicas do estado da arte e algumas intuições

sobre o problema através da lógica de primeira ordem. A versão completa da RLM (sua estrutura e os pesos) encontra-se no Apêndice B.

A fórmula 1 da Tabela 12 é análoga à fórmula 5 da Tabela 8. Juntamente com o seu peso, estimado automaticamente a partir de um conjunto de treinamento (Seção 4.4), ela captura a probabilidade a priori de que exista um relacionamento taxonômico entre dois conceitos escolhidos aleatoriamente. O peso desta fórmula irá controlar a quantidade de relacionamentos taxonômicos extraídos.

Tabela 12 - Estrutura da RLM utilizada pela técnica PREHE

Identificador	*Fórmula*
1	$\neg Kind_of(c_1, c_2)$
2	$Kind_Of(c_1, c_2) \Longrightarrow \neg Kind_of(c_2, c_1)$
3	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, +dep) \wedge Depends(t_3, t_2, +dep) \Longrightarrow Kind_of\ (c_1, c_2)$
4	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Hypernym(t_1, t_2) \Longrightarrow Kind_of\ (c_1, c_2)$
5	$F(c_1, t_1) \wedge F(c_2, t_1) \wedge F(c_1, t_2) \wedge \neg F(c_2, t_2) \Longrightarrow Kind_Of(c_1, c_2)$

Em seguida, uma propriedade importante do predicado *Kind_Of* deve ser declarada. Ela indica que este relacionamento não é simétrico, ou seja, se é verdadeiro que *Kind_Of(c_1,c_2)* então é falso que *Kind_Of(c_2,c_1)*. Esta propriedade é expressa pela fórmula 2 da Tabela 12.

Assim como termos que possuem certas dependências sintáticas com os mesmos termos, podem apresentar uma relação de sinonímia, a presença de certas dependências sintáticas entre realizações léxicas de dois conceitos pode indicar a existência de uma relação taxonômica entre os conceitos. Esse fato é capturado pela fórmula 3 da Tabela 12. A notação *+dep* indica que um peso diferente será calculado para cada tipo de dependência sintática, da mesma maneira que ocorre com a fórmula 4 da Tabela 8.

Intuitivamente, pode-se perceber que a existência de uma relação de hiperonímia entre dois termos pertencentes às realizações léxicas de dois conceitos diferentes é uma evidência de que pode haver uma relação taxonômica entre estes dois conceitos. Por exemplo, um conceito formado pelos termos “cidade” e “centro urbano” é provavelmente um superconceito de um conceito formado pelo termo “capital” uma vez que o termo “cidade” é uma hiperonímia do termo “capital”. Este fato é expresso pela fórmula 4 da Tabela 12.

As abordagens para a extração de hierarquias de conceitos que utilizam algoritmos de agrupamento hierárquico, assumem que o conjunto das realizações léxicas de um conceito está contido no conjunto das realizações léxicas do seu superconceito (CARABALLO, 1999) (FAURE; NEDELLEC, 1999). Isto significa que se há uma interseção entre os conjuntos das realizações léxicas de dois conceitos, então é provável que exista uma relação taxonômica entre eles. Outro fato a ser observado é que o superconceito, por ser mais geral, deve possuir mais realizações léxicas que não estão presentes no subconceito. Para ilustrar esse fato a Figura 11 mostra parte de uma hierarquia de conceitos. Nesta figura o conceito formado pelos termos “animal”, “réptil”, “mamífero”, “gato”, “macaco” e “cobra” é uma generalização dos conceitos formados apenas pelos termos “mamífero”, “gato” e “macaco” e pelos termos “réptil” e “cobra” uma vez que contém este termo e outros. Este fato é expresso pela fórmula 5 da Tabela 12.

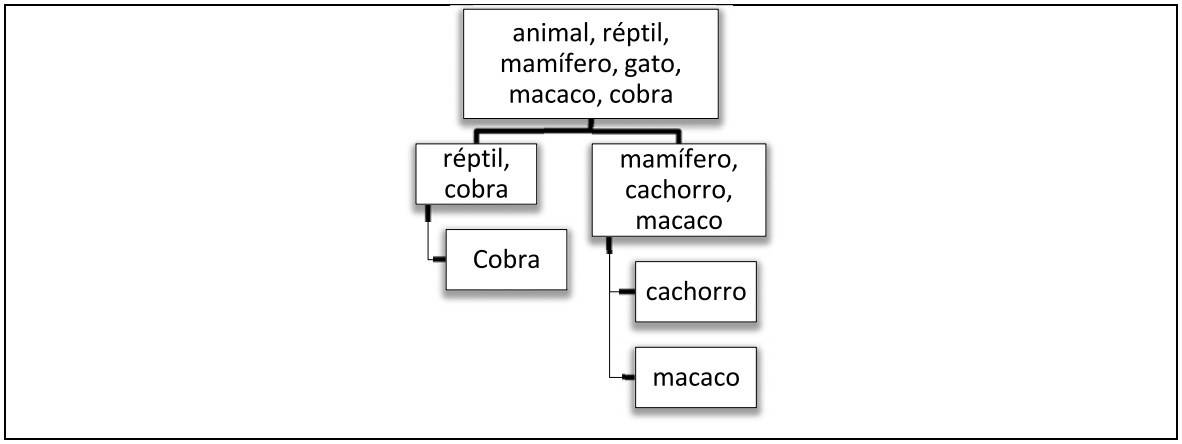


Figura 11 - Exemplo de parte de uma hierarquia de conceitos

A técnica PREHE utiliza inferência MAP (*maximum a posteriori*), descrita na Subseção 3.3.1 para descobrir os relacionamentos taxonômicos. A inferência MAP é realizada através do Algoritmo 1. O resultado da inferência MAP é um conjunto de atribuições de valores de verdade para todas as possíveis formas básicas do predicado de consulta, *Kind_Of*.

A saída da técnica PREHE é o conjunto de formas básicas do predicado *Kind_Of* consideradas verdadeiras no processo de inferência. Este conjunto é o conjunto *H* indicado na Figura 7.

4.4 Aprendizagem de pesos

As subseções anteriores apresentaram as fórmulas que constituem a estrutura das RLMs utilizadas pelas técnicas PRECE e PREHE. Como dito anteriormente, uma RLM é composta por fórmulas e seus respectivos pesos. O objetivo desta etapa é estimar os pesos para as estruturas das RLMs mostradas na Tabela 8 e na Tabela 12. Tais pesos são aprendidos automaticamente em um processo *off-line*. A fase de aprendizagem de pesos da Figura 7 é dividida em duas subfases conforme mostra a Figura 12.

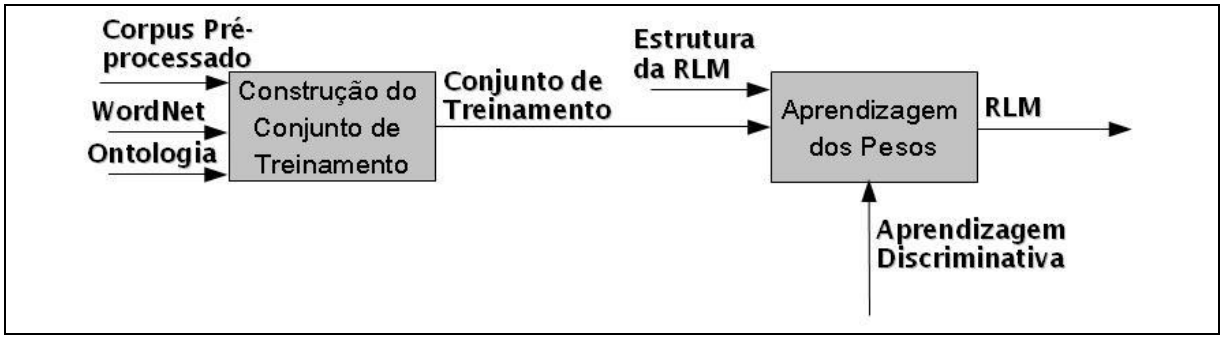


Figura 12 - Etapas da fase de aprendizagem de pesos

A aprendizagem de pesos requer um conjunto de dados de treinamento, que é produto da fase de *Construção do Conjunto de Treinamento* da Figura 12. Uma vez que a aprendizagem discriminativa é um processo de aprendizagem supervisionada, são necessários dados de entrada e as respectivas saídas desejadas. Em uma RLM isto significa que o conjunto de treinamento deve ser composto por um mundo, que atribui valores verdade tanto para as formas básicas dos predicados de evidência quanto dos predicados de consulta.

Como o objetivo das técnicas PRECE e PREHE é extrair uma hierarquia de conceitos de uma ontologia a partir de um corpus, os dados de treinamento das RLMs são obtidos a partir de um corpus (dados de entrada) e de uma hierarquia de conceitos de uma ontologia no mesmo domínio do corpus (dados de saída). Os conjuntos de treinamento utilizados são arquivos com as instanciações verdadeiras dos predicados. Como é utilizada a hipótese do mundo fechado, as instanciações que não aparecem no conjunto de treinamento são consideradas falsas.

Na RLM utilizada pela técnica PRECE, os predicados de evidência são os predicados *HasTerm*, *Depends* e *Hypernym*. Assim, os dados de entrada são obtidos através da etapa de Pré-processamento, e são o próprio corpus pré-processado. O predicado de consulta da técnica PRECE é o predicado *SameMeaning*. Os valores verdade para as suas instanciações são obtidos a partir do conjunto C^C da ontologia. Para cada classe pertencente ao conjunto C^C da ontologia, é adicionado ao conjunto de treinamento o átomo $SameMeaning(r_c,r_c)$, onde r_c é o rótulo da classe em questão. Esses mapeamentos são mostrados na Tabela 13.

Tabela 13 - Mapeamento entre as entradas e saídas do processo de criação do conjunto de treinamento para a técnica PRECE

Predicados	***Origem***
HasTerm, Depends, Hypernym	*Pré-processamento*
SameMeaning(c,c)	*Ontologia*
SameMeaning(c1,c2)	*Synsets do WordNet*

O WordNet é então utilizado para enriquecer o conjunto de treinamento. Isso é feito da seguinte forma. Para cada instanciação $SameMeaning(r_c,r_c)$ considerada verdadeira, é feita uma consulta no WordNet pelos *synsets* que possuem o termo r_c. Dentre os *synsets* retornados como resultado, aquele com o maior número de termos que aparecem no corpus é escolhido. Então, para cada termo *w* pertencente ao mesmo *synset* de r_c e que também aparece no corpus, o átomo $SameMeaning(r_c,w)$ é adicionado ao conjunto de treinamento. Depois, procura-se nas hiponímias do synset escolhido por termos que também aparecem no corpus. Cada termo *w'* encontrado também dá lugar a novos átomos $SameMeaning(r_c,w')$.

A Tabela 14 mostra um exemplo de conjunto de treinamento gerado de acordo com o descrito acima.

A RLM utilizada pela técnica PREHE, por sua vez, utiliza como evidência os predicados *Depends, Hypernym* e *F*. As instanciações dos predicados *Depends e Hypernym* são extraídas a partir do pré-processamento e as instanciações do predicado *F* são extraídas de forma semelhante ao que é feito com o predicado *SameMeaning*. Para cada classe *c* pertencente ao conjunto C^C da ontologia, é adicionado ao conjunto de treinamento um átomo $F(c,r_c)$, onde r_c é o rótulo da classe

c. Então é feita uma consulta no WordNet pelos *synsets* que possuem o termo r_c. Para cada termo *w* pertencente ao mesmo *synset* de r_c e que também aparece no corpus, é adicionado ao conjunto de treinamento da PREHE, um átomo do tipo *F(c,w)*.

Tabela 14 - Exemplo de conjunto de treinamento utilizado para aprender os pesos da RLM da técnica PRECE

HasTerm(DOC1,"country")	*Hypernym("city","capital")*
HasTerm(DOC1,"capital")	*SameMeaning("country","country")*
HasTerm(DOC1,"city")	*SameMeaning("country","nation")*
HasTerm(DOC1,"nation")	*SameMeaning("country","brazil")*
HasTerm(DOC2,"brazil")	*SameMeaning("city","city")*
HasTerm(DOC2,"capital")	*SameMeaning("city","capital")*
HasTerm(DOC2,"brasilia")	*SameMeaning("city","brasilia")*
HasTerm(DOC2,"city")	*SameMeaning("capital","capital")*
Depends("city","capital",NSUBJ)	*SameMeaning("capital","brasilia")*
Depends("capital","country",POSS)	*SameMeaning("nation","country")*
Depends("capital","brazil",POSS)	*SameMeaning("nation","nation")*
Depends("capital","brasilia",APPOS)	*SameMeaning("nation","brazil")*

O predicado de consulta da técnica PREHE é o predicado *Kind_Of*. As instanciações desse predicado são obtidas a partir do conjunto *H* da ontologia utilizada no treinamento. Para cada relacionamento taxonômico em *H*, uma instanciação equivalente do predicado *Kind_Of* é adicionada ao conjunto de treinamento. Os mapeamentos realizados para a construção do conjunto de treinamento da técnica PREHE são mostrados na Tabela 15.

Tabela 15 - Mapeamento entre as entradas e saídas do processo de criação do conjunto de treinamento para a técnica PREHE

Predicados	***Origem***
Depends, Hypernym	*Pré-processamento*
F, Kind_Of	*Ontologia*
F	*Synsets do WordNet*

A Tabela 16 mostra um exemplo de conjunto de treinamento utilizado para aprender os pesos da estrutura de RLM mostrada na Tabela 12.

Uma vez construído o conjunto de treinamento e dada a estrutura de uma RLM, é possível realizar efetivamente a aprendizagem dos pesos, como mostrado na Figura 12. Para tanto, utiliza-se a aprendizagem de pesos discriminativa (LOWD;

DOMINGOS, 2007), descrita na Seção 3.3.2. A aprendizagem discriminativa visa encontrar a distribuição de pesos que maximize a função de verossimilhança condicional dos predicados de consulta, dados os predicados de evidência.

Tabela 16 - Exemplo de conjunto de treinamento utilizado para aprender os pesos da RLM da técnica PREHE

F(1,"country")	*F(3,"brasilia")*
F(1,"nation")	*Depends("city","capital",NSUBJ)*
F(1,"brazil")	*Depends("capital","country",POSS)*
F(2,"city")	*Depends("capital","brazil",POSS)*
F(2,"capital")	*Depends("capital","brasilia",APPOS)*
F(2,"brasilia")	*Hypernym("city","capital")*
F(3,"capital")	*Kind_Of(2,3)*

4.5 Pós-processamento

A fase de pós-processamento é a fase mais simples do processo de aprendizagem de ontologias. Esta fase consiste em gerar a hierarquia de conceitos de uma ontologia a partir dos resultados das inferências realizadas nas etapas de Extração de Conceitos e Hierarquias. No processo proposto esta fase consiste em converter os resultados das técnicas PRECE e PREHE em uma ontologia OWL (WORLD WIDE WEB CONSORTIUM, 2009). Para tanto, o framework JENA (CARROLL et al., 2004) é utilizado.

Como se pode perceber na Figura 7, esta fase gera uma hierarquia de conceitos a partir do conjunto C^C, extraído pela técnica PRECE, e do conjunto *H*, extraído pela técnica PREHE. Assim, cada classe do conjunto C^C é mapeada para uma classe OWL. Da mesma forma, cada instanciação do predicado *Kind_Of* presente no conjunto *H* é mapeada para uma relação taxonômica na ontologia. Por fim, todas as classes que não tiverem uma generalização são consideradas especializações diretas da classe *Thing* do OWL.

Para exemplificar o pós-processamento, considere os resultados combinados das técnicas PRECE e PREHE mostrados na Tabela 17. A ontologia gerada a partir do processo de pós-processamento e do arquivo da Tabela 17 é mostrada na Figura 13.

Tabela 17 - Exemplo de resultados combinados das técnicas PRECE e PREHE

F(1,“country”)	*F(3,“brasilia”)*
F(1,“nation”)	*Depends(“city”,“capital”,NSUBJ)*
F(1,“brazil”)	*Depends(“capital”,“country”,POSS)*
F(2,“city”)	*Depends(“capital”,“brazil”,POSS)*
F(2,“capital”)	*Depends(“capital”,“brasilia”,APPOS)*
F(2,“brasilia”)	*Hypernym(“city”,“capital”)*
F(3,“capital”)	*Kind_Of(2,3)*

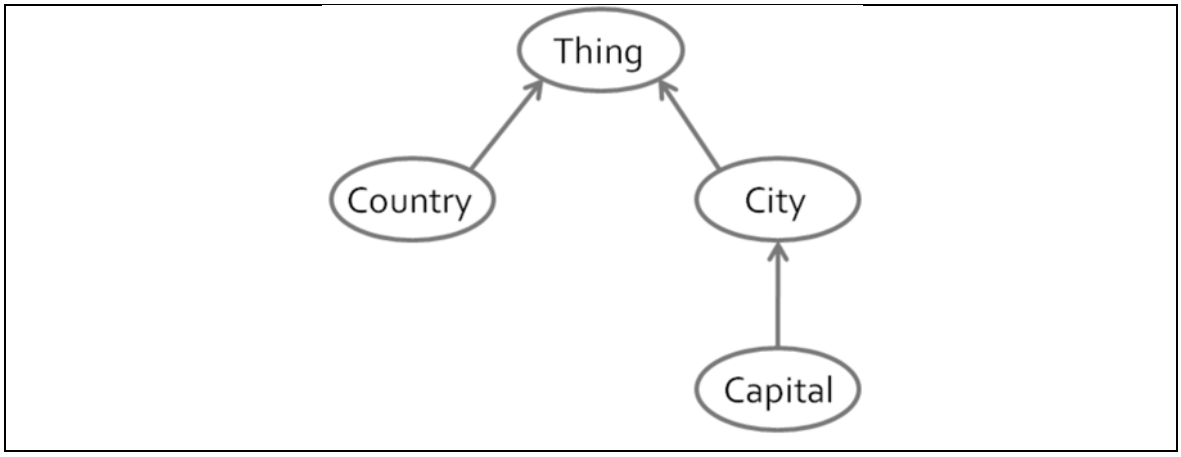


Figura 13 - Exemplo de ontologia resultante do pré-processamento

4.6 Considerações finais

Este capítulo apresentou um processo para extrair hierarquias de conceitos de ontologias a partir de fontes textuais. Este processo é composto basicamente pelas técnicas PRECE e PREHE.

Para que as técnicas PRECE e PREHE possam ser utilizadas, o corpus deve ser pré-processado, utilizando processamento da linguagem natural e análise estatística.

A técnica PRECE extrai os conceitos a partir de suas realizações léxicas, i.e. termos em linguagem natural que referenciam esse conceito. Os termos são agrupados baseados nas suas relações de hiperonímias e nas dependências sintáticas entre eles. A técnica PREHE organiza os conceitos extraídos em uma hierarquia. Ambas as técnicas utilizam a lógica de Markov para representar idéias de vários métodos do estado da arte da aprendizagem de ontologias. As redes lógicas de Markov permitem representar os relacionamentos entre os termos bem como

capturar informações estatísticas dos mesmos. Os resultados das técnicas PRECE e PREHE são escritos na linguagem OWL, constituindo a saída do processo proposto.

É importante ressaltar que este capítulo se concentrou apenas nas camadas de conceitos e de hierarquia da Figura 3. A extensão da abordagem proposta para contemplar as outras camadas é proposta como trabalho futuro.

O próximo capítulo trata da avaliação das técnicas PRECE e PREHE, comparando-as com algumas abordagens do estado da arte.

5 AVALIAÇÃO

Os experimentos realizados têm o intuito de responder à seguinte questão: o quanto as técnicas de aprendizagem estatística relacional podem melhorar os resultados da aprendizagem de ontologias em relação às técnicas existentes? Para isso, as técnicas propostas deverão ser avaliadas em domínios diferentes e comparadas com as técnicas do estado da arte.

A partir da discussão sobre métodos de avaliação de técnicas de aprendizagem de ontologias da Seção 2.3, propõe-se a realização da avaliação em relação a uma ontologia de referência. A comparação com uma ontologia de referência pode ser feita através de medidas adequadas à camada de conceitos (DELLSCHAFT; STAAB, 2006) e à camada de hierarquia (MAEDCHE; STAAB, 2002). Para os fins deste trabalho será utilizada a abordagem de comparação com uma ontologia de referência.

Para isso serão utilizados corpora de texto nos domínios turístico e da biologia. O conjunto de dados do domínio turístico a ser utilizado é o *LonelyPlanet*. Este corpus consiste de 1801 documentos em HTML, utilizados pela primeira vez por Kavalec e Svatek (2005) na avaliação de técnicas de extração de relações não taxonômicas. No domínio da biologia será utilizado o corpus GENIA (OHTA; TATEISI; KIM, 2002), desenvolvido junto com a ontologia GENIA que é uma coleção de documentos no domínio da biologia desenvolvida pelo laboratório Tsuji.

Este capítulo está organizado da seguinte forma. A Seção 5.1 detalha os conjuntos de dados a serem utilizados nos experimentos. As ferramentas utilizadas neste experimento são discutidas na Seção 5.2. A Seção 5.3 descreve os passos que foram realizados nesta avaliação. A Seção 5.4 apresenta a avaliação da extração de conceitos, explicitando as técnicas comparadas, as métricas de avaliação e discutindo os resultados obtidos. Por fim, a seção 5.5 apresenta a avaliação da extração de hierarquia, também apresentando as técnicas avaliadas, medidas utilizadas e discutindo os resultados.

5.1 Dados

Os dados utilizados nos experimentos aqui descritos são constituídos por dois corpora de texto, um no domínio da biologia e outro no turismo, descritos na Subseção 5.1.1 e duas ontologias também nos domínios da biologia e do turismo, descritas na subseção 5.1.2.

5.1.1 Corpora de texto

Os corpora utilizados nos experimentos aqui descritos são o Genia e o *LonelyPlanet*. O corpus Genia[3], é uma coleção de documentos sobre biologia molecular desenvolvido pelos laboratórios Tsuji. Tais documentos são resumos de artigos extraídos do banco de dados MEDLINE[4]. O objetivo do projeto Genia é criar um recurso para apoiar o desenvolvimento de aplicações de processamento de linguagem natural no domínio da biologia molecular. A versão do corpus utilizada neste trabalho é composta por 2000 documentos, 18.546 sentenças e 400.000 tokens.

O corpus *LonelyPlanet* consiste de 1801 documentos HTML, originalmente descarregados de http://www.lonelyplanet.com/destinations em 2003, descrevendo destinos turísticos (países, cidades, etc) de todos os continentes. Este corpus foi utilizado como entrada para as técnicas avaliadas neste experimento.

A Tabela 18 apresenta um resumo dos corpora aqui descritos.

Tabela 18 - Corpora utilizados nos experimentos

Corpus	***Domínio***	***Número de Documentos***
Genia	*Biologia*	*2000*
LonelyPlanet	*Turismo*	*1801*

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/genia/topics/Corpus/

[4] http://www.ncbi.nlm.nih.gov

5.1.2 Ontologias utilizadas

Assim como os corpora, duas ontologias foram utilizadas nos experimentos deste trabalho, uma no domínio da biologia, a ontologia Genia e uma no domínio do turismo, a ontologia GETESS.

O corpus Genia é anotado semanticamente de acordo com os conceitos da ontologia Genia, também desenvolvida pelo laboratório Tsuji, com o intuito específico de anotar o corpus Genia. Esta ontologia representa os conceitos acerca das reações sinalizadoras entre células. A versão 3.02 da ontologia Genia, utilizada neste trabalho, é composta por 54 conceitos e 45 relações taxonômicas e será, daqui por diante, chamada de O_{Genia}. Os termos relevantes do corpus Genia são anotados de acordo com os conceitos em O_{Genia}. Para ilustrar como esse processo é feito, tomemos como exemplo a seguinte anotação semântica do termo *monocytes*:

```
<conslex=”monocyte”sem=”G#cell_type”>monocytes</cons>
```

Os termos relevantes são anotados com a tag *cons* que possui dois atributos: *lex* e *sem*. O atributo *lex* indica o lema do termo anotado enquanto que o *sem*, o rótulo do conceito da ontologia com o qual o termo está relacionado.

Os conceitos automaticamente extraídos do corpus *LonelyPlanet* foram comparados aos conceitos da ontologia para o domínio turístico desenvolvida no contexto do projeto GETESS, a partir de agora chamada de $O_{Tourism}$. A ontologia $O_{Tourism}$ foi desenvolvida manualmente por um engenheiro de ontologias e consiste de 969 conceitos.

A Tabela 19 apresenta um resumo dos corpora aqui descritos.

Tabela 19 - Ontologias utilizadas nos experimentos

Ontologia	***Domínio***	***Número de Conceitos***	***Número de Relações Taxonômicas***
O_{Genia}	*Biologia*	*54*	*45*
$O_{Tourism}$	*Turismo*	*969*	*961*

5.2 Ferramentas utilizadas

As tarefas do experimento relativas ao processamento da linguagem natural foram realizadas utilizando o ambiente GATE (CUNNINGHAM, 2002). O GATE (*General Architecture for Text Engineering*) provê um framework e um ambiente de desenvolvimento para aplicações de processamento da linguagem natural. Através dele é possível realizar tarefas como tokenização, atribuição de *tags* de parte do discurso, lematização e *chunking* de textos em inglês. A análise sintática foi realizada acoplando-se o Stanford Parser ao ambiente de desenvolvimento do GATE.

A aprendizagem de pesos e a inferência em lógica de Markov foram realizadas utilizando o pacote de software Alchemy (KOK et al., 2009). O Alchemy é um pacote de software de código aberto que provê implementações de uma série de algoritmos para aprendizagem estatística relacional e inferência lógica probabilística baseados em uma representação em lógica de Markov. Através dele é possível realizar inferência do tipo MAP e probabilística, além de aprendizagem gerativa e discriminativa de pesos.

A utilização destas duas ferramentas reduziu o esforço de implementação das técnicas PRECE e PREHE à conversão dos resultados do GATE ao formato utilizado pelo Alchemy, que provê uma linguagem simples para a especificação de RLMs.

Para efeito de comparação com as técnicas PRECE e PREHE, utilizou-se as implementações das abordagens baseadas nos padrões de Hearst e na Análise Formal de Conceitos providas pelo ambiente TextToOnto (CIMIANO; VÖLKER, 2005), um ambiente para a aprendizagem de ontologias a partir de fontes textuais. A implementação da Análise Semântica Latente Probabilística utilizada foi a provida no software GibbsLDA++ (PHAN; NGUYEN; HORIGUCHI, 2008).

5.3 Descrição do experimento

Os experimentos realizados consistiram na comparação com uma ontologia de referência assim, ela se encaixa no grupo dos métodos de avaliação de técnicas de aprendizagem de ontologias através da avaliação das ontologias aprendidas.

O experimento consiste em extrair automaticamente uma ontologia a partir de um corpus de texto para então compará-la com uma ontologia de referência. Neste trabalho o corpus Genia e a ontologia O_{Genia} foram utilizados na fase de treinamento das técnicas PRECE e PREHE. A RLM resultante do treinamento com a técnica para a técnica PRECE é mostrada no Apêndice A e a da técnica PREHE, no Apêndice B. O corpus *LonelyPlanet* e a ontologia $O_{Tourism}$, por sua vez, foram utilizados para avaliar as técnicas propostas. Assim, para cada técnica, as seguintes tarefas foram executadas:

a) extrair uma hierarquia de conceitos a partir do corpus *LonelyPlanet*;
b) comparar os resultados obtidos com a ontologia $O_{Tourism}$ utilizando as métricas descritas adiante.

Deste modo foram realizados dois experimentos, um para realizar uma avaliação no nível da camada de extração de conceitos e outro no nível a extração da hierarquia de conceitos.

5.4 Avaliação da técnica PRECE

Quatro técnicas para a extração de conceitos foram utilizadas para extrair os conceitos do corpus *LonelyPlanet*. Os conjuntos de conceitos extraídos bem como as respectivas técnicas utilizadas para extraí-los são:

a) $C_{Trivial}$ - conjunto de conceitos extraídos a partir dos 500 termos mais freqüentes no corpus. Cada termo é considerado uma classe;
b) C_{PRECE} – conjunto de conceitos extraídos utilizando a técnica PRECE. Nos experimentos aqui descritos, a aprendizagem e inferência em redes lógicas de Markov foram executadas através do software

Alchemy (KOK et al., 2009). As tarefas relacionadas ao pré-processamento do corpus foram realizadas utilizando o GATE (CUNNINGHAM, 2002);

c) C_{PLSA} - conjunto de conceitos extraídos utilizando Análise Semântica Latente Probabilística (PLSA). Uma vez que a PLSA aprende conceitos como distribuições de probabilidade sobre os termos, os conceitos foram rotulados com o termo com maior probabilidade dado o respectivo tópico. Para a extração dos conceitos foi utilizado o software GibbsLDA++ (PHAN; NGUYEN; HORIGUCHI, 2008);

d) C_{FCA} - conjunto de conceitos extraídos utilizando a técnica baseada na Análise Formal de Conceitos (FCA) proposta por Cimiano, Hotho e Staab (2004b).

5.4.1 Métricas

Para a comparação dos conjuntos de conceitos das ontologias automaticamente aprendidas com a ontologia de referência são utilizadas as medidas definidas por Dellschaft e Staab (2006), o recall, a precisão e a medida-f, adaptadas das versões já conhecidas da recuperação de informação. Tais medidas consideram o número de conceitos corretamente extraídos.

Sejam C_R o conjunto de conceitos na ontologia de referência (o conjunto de conceitos que deveriam ser extraídos do corpus) e C_L o conjunto de conceitos na ontologia aprendida (o conjunto de conceitos realmente extraídos pelo procedimento de aprendizagem) então, o recall e a precisão são definidos nas equações 15 e 16.

$$Recall = \frac{|C_R \cap C_L|}{|C_R|} \tag{15}$$

$$Precisão = \frac{|C_R \cap C_L|}{|C_L|} \tag{16}$$

É desejável que as técnicas avaliadas não apresentem bons valores apenas para o recall ou para a precisão. É importante que a técnica apresente uma boa combinação dos dois. Uma medida constantemente utilizada para refletir essa combinação em um único valor é a medida-F, uma média harmônica de ambos. Esta medida é descrita pela equação 17.

$$medida - F = \frac{2 * Recall * Precisão}{Recall + Precisão} \quad (17)$$

Para computar o valor $|C_R \cap C_L|$ presente nas equações do recall e da precisão é preciso determinar um método para a comparação de conceitos em ontologias diferentes. Um método simples e direto é a simples comparação entre os rótulos de cada conceito.

5.4.2 Resultados

A Tabela 20 mostra um comparativo entre as técnicas avaliadas em relação ao total de conceitos extraídos e à quantidade de conceitos presentes tanto no conjunto extraído automaticamente quanto na ontologia $O_{Tourism}$. A partir dessa tabela é possível perceber que os conjuntos de conceitos $C_{Trivial}$ e C_{FCA} possuem o maior número de conceitos corretamente extraídos (95 e 86 respectivamente). Entretanto, esses conjuntos também possuem o maior número total de conceitos extraídos (500 e 612 respectivamente) o que indica baixa precisão, principalmente em relação ao conjunto C^C_{PRECE}. Esta observação é confirmada pelos valores das medidas recall, precisão e *F* apresentadas na Figura 14.

Tabela 20 - Comparativo das técnicas em relação ao número de conceitos extraídos

Conjunto de Conceitos	***Número de Conceitos***	***Número de conceitos corretamente extraídos (em relação a $O_{Tourism}$)***
$C_{Trivial}$	500	95
C_{FCA}	612	86
C_{PRECE}	263	85
C_{PLSA}	405	47

A Figura 14 mostra os resultados da extração de conceitos medidos segundo a precisão, o recall e a medida *F*, apresentados na seção anterior. Em relação a Análise Formal de Conceitos e a abordagem trivial, puramente estatística, a figura evidencia que a técnica PRECE apresenta um recall ligeiramente menor sendo, entretanto, a mais precisa de todas. Este fato é uma evidência de que a combinação de técnicas de aprendizagem relacional com abordagens estatísticas é uma abordagem viável para aumentar a precisão da extração automática de classes de uma ontologia.

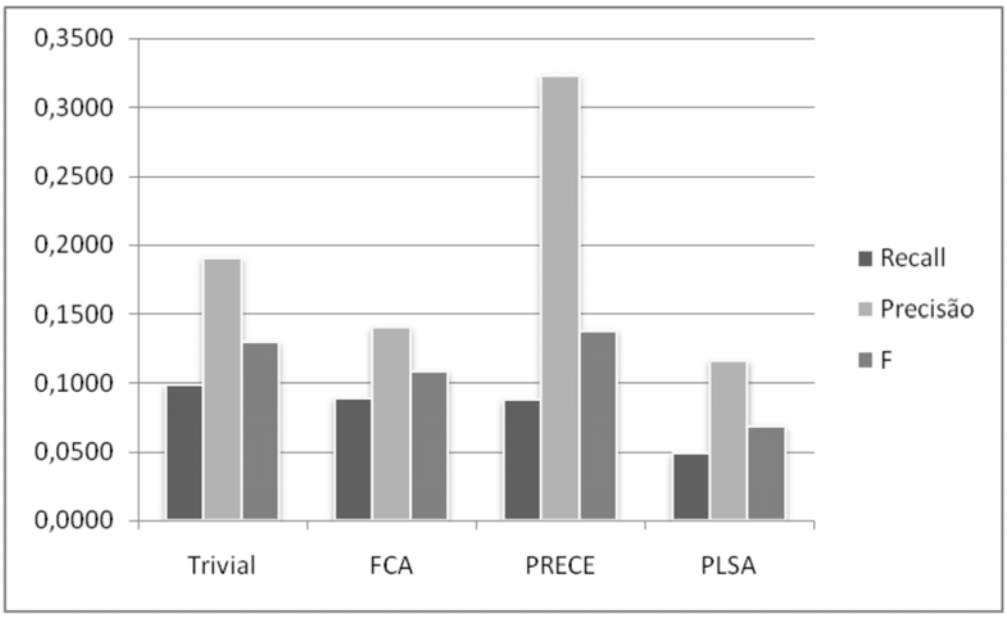


Figura 14 - Resultados dos experimentos com o Corpus *LonelyPlanet*

Uma vez que os termos não são i.i.d. (eles são correlacionados pelas suas dependências sintáticas e por relacionamentos semânticos como a hiperonímia) os relacionamentos entre eles fornecem evidências úteis para determinar sua proximidade semântica, o que explica o fato de a técnica PRECE ter obtido resultados melhores que a análise semântica latente. Como visto no Capítulo 1, a Análise Formal de Conceitos também considera relacionamentos sintáticos, mas apenas aqueles entre sujeito e verbo. A técnica PRECE, por sua vez, estende essa idéia fazendo uso de todos os relacionamentos sintáticos no Anexo B e do relacionamento de hiperonímia, baseando-se assim em um conjunto maior de evidências, o que pode explicar seu melhor resultado do ponto de vista da precisão.

Por outro lado, a técnica PRECE também possui suas desvantagens. Uma vez que ela é uma abordagem supervisionada, ela é muito sensível a qualidade

do conjunto de treinamento. Além disso, construir um conjunto de treinamento como o necessário a esta tarefa é uma tarefa árdua. Este problema pode ser amenizado através da investigação e desenvolvimento de técnicas para a aprendizagem não supervisionada de Redes Lógicas de Markov.

5.5 Avaliação da Técnica PREHE

Três técnicas para a extração de hierarquias de conceitos foram utilizadas para extrair os conceitos do corpus *LonelyPlanet*. A hierarquia extraída com a técnica PREHE é mostrada na subseção 5.4.1, enquanto que a extraída com a análise formal de conceitos é mostrada na subseção 5.4.2 e a extraída através dos padrões de Hearst, na subeção 5.4.3.

5.5.1 Hierarquia extraída com a técnica PREHE

A hierarquia de conceitos extraída utilizando a técnica PREHE será denotada por H_{PREHE}. Nos experimentos aqui descritos, a aprendizagem e inferência em RLM foram executadas através do software Alchemy (KOK et al., 2009). As tarefas relacionadas ao pré-processamento do corpus foram realizadas utilizando o GATE (CUNNINGHAM, 2002). Parte desta hierarquia é mostrada na Figura 15. A sua versão completa encontra-se no Apêndice C.

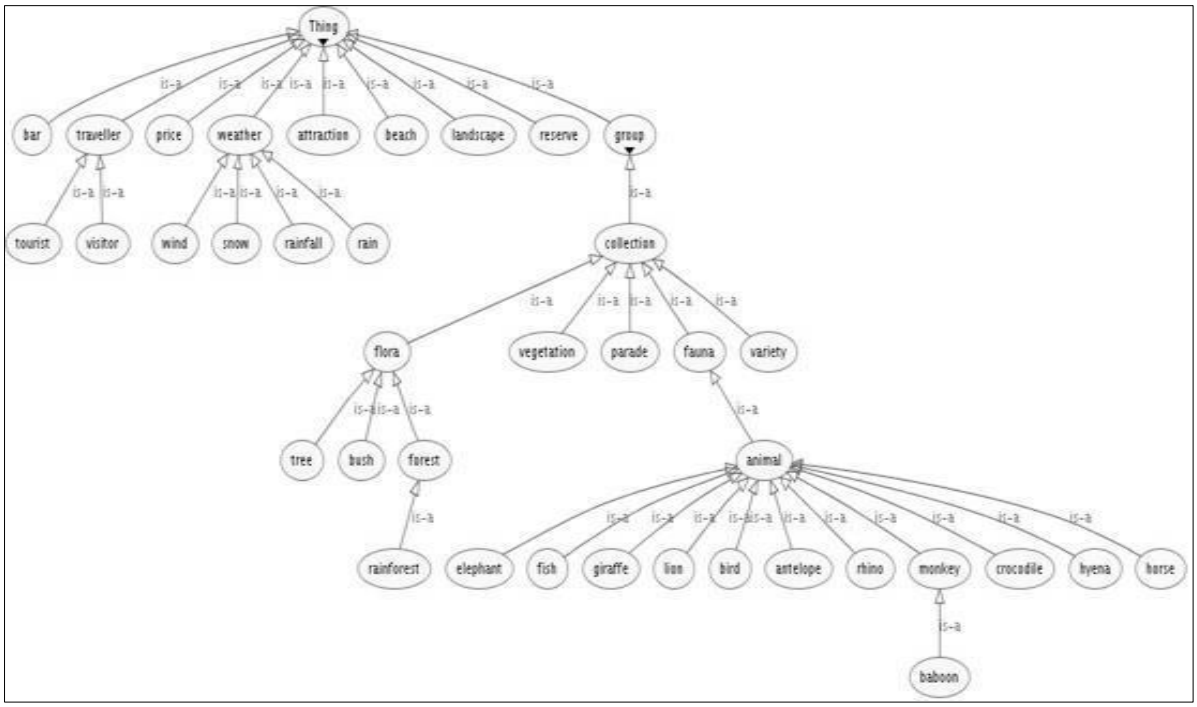


Figura 15 - Parte da hierarquia H_{PREHE} gerada a partir do corpus *LonelyPlanet*

5.5.2 Hierarquia extraída com a análise formal de conceitos

Esta hierarquia de conceitos foi extraída utilizando a técnica baseada na análise formal de conceitos proposta por Cimiano, Hotho e Staab (2004b). Ela será denotada por H_{FCA}. A Figura 16 mostra uma parte da hierarquia H_{FCA}.

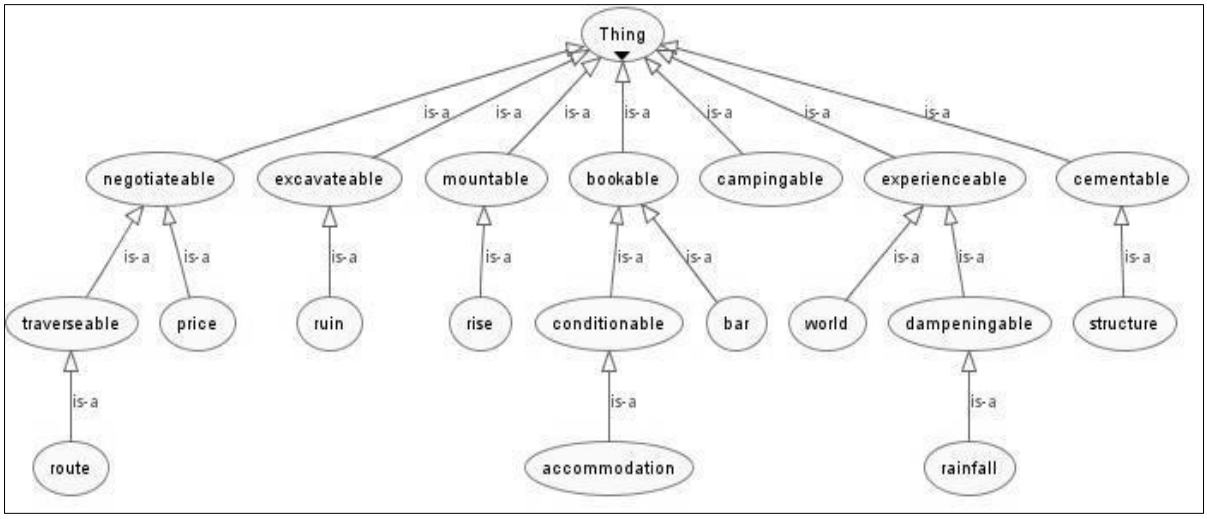


Figura 16 - Parte da hierarquia H_{FCA} gerada a partir do corpus *LonelyPlanet*

5.5.3 Hierarquia extraída a partir dos padrões de Hearst

A hierarquia H_{Hearst} foi extraída a partir dos conceitos presentes no conjunto $C_{Trivial}$, utilizando padrões de Hearst (HEARST, 1992) para identificar relacionamentos taxonômicos entre eles.

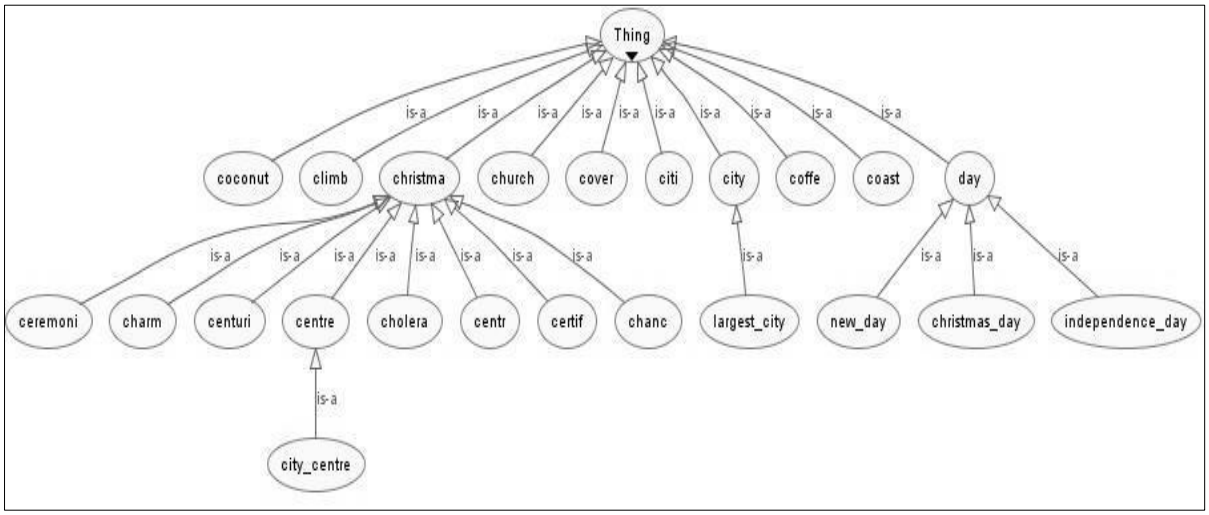


Figura 17 - Parte da hierarquia H_{Hearst} gerada a partir do corpus *LonelyPlanet*

5.5.4 Métricas

Antes de definir as medidas utilizadas nesta avaliação, é necessária uma breve explicação sobre a notação utilizada nesta seção. Seja $\mathcal{O}_i$ uma ontologia, o conjunto de classes desta ontologia é denotado por C_i^C e o conjunto de relações taxonômicas, por H_i, sendo uma relação taxonômica deste conjunto denotada por $Kind_Of_i(c_k, c_j)$.

Para comparar duas taxonomias Dellschaft e Staab (2006) propuseram o recall taxonômico (RT) e a precisão taxonômica (PT). Ambas as medidas fazem uso do conceito de cotopia semântica (MAEDCHE; STAAB, 2002). A cotopia semântica de um conceito pode ser definida como o conjunto de seus superconceitos e subconceitos. Mais formalmente, dada uma ontologia $\mathcal{O}_1$, a cotopia semântica (*CS*) de um conceito é definida pela equação 18.

$$CS(c_i, \mathcal{O}_1) \coloneqq \{c_j \in C_1^C \,|\, Kind_Of_{i_1}(c_i, c_j) \vee Kind_Of_{i_1}(c_j, c_i)\} \quad (18)$$

Entretanto, como demonstrado por Maedche e Staab (2002), comparar apenas a cotopia semântica dos conceitos não reflete a similaridade das hierarquias de conceitos como desejado. Por tanto é utilizada a cotopia semântica comum que considera apenas conceitos presentes nas duas ontologias sendo comparadas e é definida na equação 19.

$$CS(c_i, \mathcal{O}_1, \mathcal{O}_2) \coloneqq \left\{c_j \in C_1^C \cap C_2^C \,\middle|\, \left(Kind_Of_{i_1}(c_i, c_j) \vee Kind_Of_{i_1}(c_j, c_i)\right) \wedge c_i \neq c_j\right\} \quad (19)$$

A comparação de duas hierarquias geralmente é definida por medidas locais e globais (DELLSCHAFT; STAAB, 2006). As medidas locais comparam a posição de dois conceitos nas duas hierarquias, enquanto que as medidas globais comparam as hierarquias como um todo. A precisão taxonômica também é definida localmente e globalmente. A precisão taxonômica local pt_{local} compara as cotopias semânticas comuns de dois conceitos, um de cada hierarquia. Seja $\mathcal{O}_L$ a ontologia sendo avaliada e $\mathcal{O}_R$ a ontologia de referência, e c_1 e c_2 duas classes tais que $c_1 \in C_L^C$ e $c_2 \in C_R^C$ a medida pt_{local} é definida na equação 20.

$$pt_{local}(c_1, c_2, \mathcal{O}_L, \mathcal{O}_R) \coloneqq \frac{|CS(c_1, \mathcal{O}_L, \mathcal{O}_R) \cap CS(c_2, \mathcal{O}_L, \mathcal{O}_R)|}{|CS(c_1, \mathcal{O}_L, \mathcal{O}_R)|} \quad (20)$$

A taxonomia semântica global (*PT*) de uma ontologia $\mathcal{O}_L$ em relação a uma ontologia de referência $\mathcal{O}_R$ é definida pela equação 21.

$$PT(\mathcal{O}_L, \mathcal{O}_R) \coloneqq \frac{1}{|C_L^C \cap C_R^C|} \sum_{c \in C_L^C \cap C_R^C} pt_{local}(c, c, \mathcal{O}_L, \mathcal{O}_R) \quad (21)$$

O recall taxonômico (*RT*) é definido por sua vez em função a precisão taxonômica como na equação 22.

$$RT(\mathcal{O}_L, \mathcal{O}_R) \coloneqq PT(\mathcal{O}_R, \mathcal{O}_L) \quad (22)$$

Uma medida combinada do recall e precisão taxonômica é a medida *F* assim como para o recall e precisão utilizados para avaliar os resultados da extração de classes. A medida *F* para taxonomias, representada por *FT* é dada pela equação 23.

$$FT(\mathcal{O}_L, \mathcal{O}_R) = \frac{2 * PT(\mathcal{O}_L, \mathcal{O}_R) * RT(\mathcal{O}_L, \mathcal{O}_R)}{PT(\mathcal{O}_L, \mathcal{O}_R) + RT(\mathcal{O}_L, \mathcal{O}_R)} \quad (23)$$

5.5.5 Resultados

A Figura 18 mostra os resultados do experimento realizado para a extração de hierarquias. Nesta figura é possível perceber que, ao contrário do que acontece com a técnica PRECE, a técnica PREHE parece ter uma precisão mais baixa do que os métodos do estado da arte como a Análise Formal de Conceitos e a abordagem baseada nos padrões de Hearst. Entretanto a técnica PREHE apresentou um recall maior, o que lhe proporcionou os melhores resultados em termos da medida *FT*.

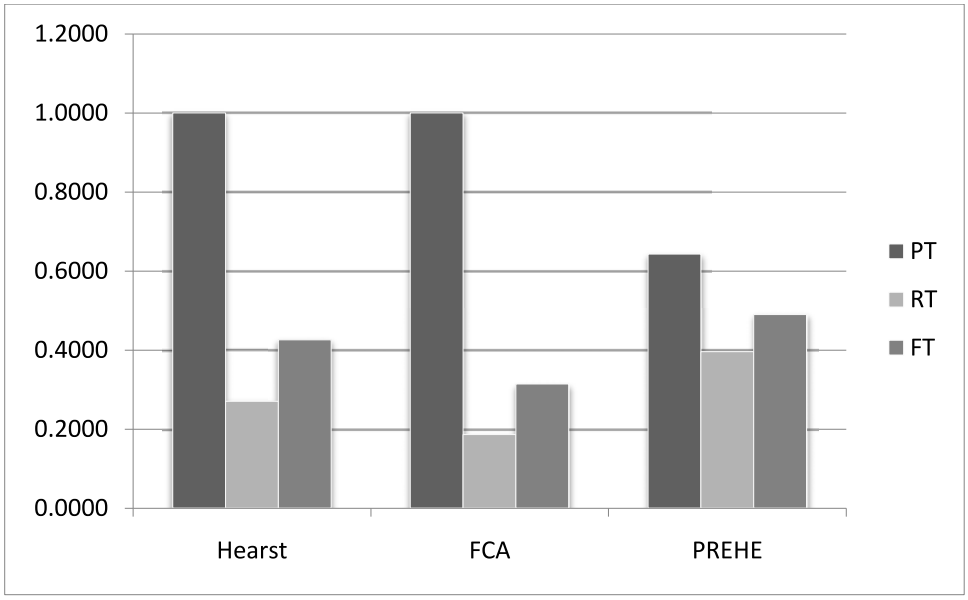


Figura 18 - Resultado dos Experimentos com o corpus *LonelyPlanet*

A Tabela 21 ajuda a interpretar os dados da Figura 18. A Tabela 21 mostra o número de relacionamentos taxonômicos extraídos por cada técnica. É

importante frisar que os relacionamentos taxonômicos entre os conceitos e a raiz da hierarquia não são considerados na Tabela 21. Nela pode-se perceber que a alta precisão obtida pela abordagem baseada nos padrões de Hearst deve-se ao baixo número de relacionamentos taxonômicos extraídos. Isso significa que os padrões de Hearst constituem uma forte evidência para a descoberta de relacionamentos taxonômicos, porém eles não aparecem com freqüência nos textos, o que explica o baixo número de relacionamentos taxonômicos extraídos e, conseqüentemente o baixo recall taxonômico.

A alta precisão taxonômica da abordagem baseada na FCA e o alto número de relacionamentos taxonômicos extraídos apontam para a boa efetividade desta abordagem. Entretanto ela apresenta um baixo recall taxonômico, medida na qual a técnica PREHE apresenta vantagem sobre as outras duas.

O maior recall da técnica PREHE deve-se ao fato de a mesma utilizar tanto a hiperonímia quanto uma gama maior de dependências sintáticas para representar relacionamentos entre os termos em relação à abordagem baseada na FCA. Entretanto o ganho no recall obtido pela técnica PREHE custou uma redução na precisão o que pode indicar que algumas das dependências sintáticas consideradas não apresentam uma evidência sólida da presença de um relacionamento taxonômico.

Tabela 21 - Comparativo das técnicas em relação ao número de relacionamentos taxonômicos extraídos

Hierarquias	***Número de relacionamentos taxonômicos***
H_{Hearst}	*17*
H_{FCA}	*364*
H_{PREHE}	*170*
$H_{Tourism}$	*961*

5.6 Considerações finais

Este capítulo apresentou uma avaliação das técnicas PRECE e PREHE propostas neste trabalho. A avaliação da efetividade destas técnicas se deu através

da comparação dos seus resultados com uma ontologia de referência, utilizando as medidas propostas por Dellschaft e Staab (2006).

Os resultados mostraram que tanto a técnica PRECE quanto a PREHE apresentaram bons resultados em comparação com as técnicas do estado da arte utilizadas na avaliação, fornecendo assim uma evidência para suportar a afirmação de que as técnicas de aprendizagem estatística relacional são adequadas para a aprendizagem de ontologias a partir de fontes textuais. Entretanto os resultados mostraram também que ainda há espaço para melhorar a efetividade das técnicas propostas.

Um ponto que é necessário observar é que a comparação com uma ontologia de referência indica o quanto as técnicas para a extração automatizada de ontologias se aproximam do julgamento dos especialistas de domínio que desenvolveram a ontologia de referência utilizada. Uma vez que não é possível determinar com exatidão a qualidade da ontologia de referência, o método de avaliação aqui utilizado recompensa ontologias com problemas similares e penaliza aquelas que não os têm. Por isso, é necessário também realizar a avaliação da PRECE e da PREHE através de outras abordagens, como as discutidas na Seção 2.3.

6 CONCLUSÃO

Com o advento da Web Semântica e os sistemas de conhecimento se mostrando cada vez mais efetivos em diversas tarefas, é crescente a necessidade de diminuir o custo do desenvolvimento de bases de conhecimento sem sacrificar a efetividade deste processo.

A aprendizagem de ontologias é uma abordagem promissora para aliviar o problema do gargalo da aquisição de conhecimento. As abordagens propostas fazem uso de técnicas de aprendizagem de máquina, como técnicas de agrupamento e outras variações de aprendizagem não supervisionada (CIMIANO; HOTHO; STAAB, 2004b) (MAEDCHE; STAAB, 2004), processamento da linguagem natural (HEARST, 1992) (IWANSKA; MATA; KRUGER, 2000) (SNOW; JURAFSKY; NG, 2005) e análise estatística como co-ocorrência e ponderação de termos baseada na freqüência (HARRIS, 1968) (MAEDCHE; STAAB, 2004). Apesar de tais métodos estarem bem consolidados, avaliações de abordagens propostas (CIMIANO; HOTHO; STAAB, 2004a) evidenciam que, apesar das técnicas propostas diminuírem o esforço do especialista humano ainda há a necessidade de sua interferência no processo, refinando, estendendo e corrigindo as ontologias aprendidas.

Este trabalho abordou o problema da aprendizagem de ontologias a partir de fontes textuais a partir da aprendizagem estatística relacional, utilizando a mesma para agrupar termos com significados semelhantes e descobrir relacionamentos taxonômicos a partir das dependências sintáticas e relacionamentos de hiperonímia entre os termos. Os resultados da avaliação realizada mostram que esta é uma abordagem promissora para melhorar a efetividade da aprendizagem de ontologias. Entretanto este ainda é um problema em aberto, uma vez que os resultados mostram que ainda há espaço para melhorar a efetividade da abordagem proposta. Além disso, apenas as camadas de extração de conceitos e de hierarquias foram contempladas, não tendo sido abordadas as camadas de relacionamentos não taxonômicos e de axiomas.

6.1 Resultados

Este trabalho apresentou uma abordagem para a extração de hierarquias de conceitos a partir de fontes textuais. O processo proposto é composto por duas técnicas: a PRECE para a extração de conceitos e a PREHE para organizar hierarquicamente os conceitos extraídos. Estas técnicas combinam idéias de vários métodos do estado da arte com as redes lógicas de Markov, um modelo de aprendizagem estatística relacional. Através das redes lógicas de Markov é possível utilizar o poder expressivo da lógica de primeira ordem para representar os relacionamentos entre os termos bem como utilizar a teoria da probabilidade, através das redes de Markov, para lidar com o ruído presente em textos em linguagem natural.

As redes lógicas de Markov não trabalham diretamente com textos em linguagem natural. Por isso, a fase de pré-processamento do processo proposto extrai dados do corpus, como os termos relevantes e as dependências sintáticas entre eles, e os representa no formato com o qual as RLMs podem trabalhar. No processo proposto também há uma etapa de pós-processamento que visa escrever em uma linguagem de especificação de ontologias, o OWL, o resultado produzido pela aplicação das técnicas PRECE e PREHE.

As abordagens estatísticas tradicionais assumem que os dados são independentes e identicamente distribuídos (i.i.d.), ou seja, não consideram os relacionamentos entre os objetos do domínio. Foi dito na parte introdutória deste trabalho que uma vez que os termos apresentam relacionamentos sintáticos (as dependências sintáticas) e semânticos (hiperonímia), assumir que os dados são i.i.d. no problema da aprendizagem de ontologias poderia implicar em uma perda na efetividade. Os experimentos conduzidos neste trabalho mostraram que considerando as dependências sintáticas e os relacionamentos de hiperonímia entre os termos, as técnicas PRECE e PREHE conseguiram bons resultados se comparados aos atingidos por técnicas do estado da arte. Isto é uma evidência de que a aprendizagem estatística relacional é, de fato, uma abordagem promissora para a descoberta de conhecimento em fontes textuais.

Uma das limitações da aprendizagem proposta é que ela é supervisionada, requerendo portanto uma ontologia e um corpus para o treinamento.

A efetividade das técnicas PRECE e PREHE é portanto limitada pela qualidade da ontologia utilizada na aprendizagem de pesos. Uma maneira de superar este problema é através da aprendizagem não supervisionada de RLMs. Poon e Domingos (2008), por exemplo, realizam resolução de correferência de forma não supervisionada utilizando RLMs.

Outra desvantagem da abordagem proposta foi a baixa precisão taxonômica obtida pela técnica PREHE nos experimentos realizados. Isto é um indício de que algumas das dependências sintáticas consideradas não constituem uma evidência sólida da presença de um relacionamento taxonômico. Este problema pode ser tratado considerando apenas as dependências sintáticas mais relevantes. Exemplos de medidas que podem determinar a relevância de uma dependência sintática, baseando-se na freqüência com que elas aparecem no corpus é a informação mútua (HINDLE, 1990) e a medida apresentada por Resnick (1997).

Entretanto, é necessário observar que a comparação com uma ontologia de referência, realizada neste trabalho, indica o quanto as técnicas para a extração automatizada de ontologias se aproximam do julgamento dos especialistas de domínio que desenvolveram a ontologia de referência utilizada. Através dessa avaliação não é possível determinar o quanto uma ontologia é adequada a uma determinada tarefa. Além disso, se a ontologia de referência apresentar problemas de modelagem, o método de avaliação aqui utilizado recompensa ontologias com problemas similares e penaliza aquelas que não os têm.

6.2 Trabalhos futuros

Dadas as limitações discutidas da comparação com uma ontologia de referência para a avaliação de métodos de aprendizagem de ontologia, é proposta como trabalho futuro a avaliação das técnicas PRECE e PREHE utilizando uma aplicação (PORZEL; MALAKA, 2004). Para esta tarefa propõe-se a avaliação do modelo de recuperação de informação baseado em ontologias proposto por Silva, Girardi e Drumond (2009). Esta avaliação será feita testando o mesmo com ontologias construídas utilizando diferentes abordagens e medindo a sua efetividade

com cada ontologia. Assim também será possível testar a abordagem proposta em domínios diferentes daqueles utilizados neste trabalho.

Outro aspecto das técnicas propostas a ser avaliado é a complexidade das mesmas. É necessário avaliar se a complexidade introduzida pela aprendizagem estatística relacional é compensada pelos resultados. Deste modo, a avaliação da complexidade das técnicas PRECE e PREHE e das técnicas do estado da arte aqui analisadas é proposta. Os resultados dessa análise de complexidade devem ser comparados com a efetividade de cada técnica avaliada.

Juntas as técnicas PRECE e PREHE se mostraram capazes de extrair uma hierarquia de conceitos a partir de um corpus de texto. Entretanto, de acordo com a equação 1, uma ontologia é mais do que uma hierarquia de conceitos. O próximo passo neste trabalho é aprender os relacionamentos não taxonômicos, os axiomas e as instâncias da ontologia.

A abordagem com redes lógicas de Markov pode ser estendida também para as camadas subseqüentes da aprendizagem de ontologias. Por exemplo, a extração de relacionamentos não taxonômicos pode ser abordada através da invenção estatística de predicados (KOK, DOMINGOS, 2007). A invenção estatística de predicados consiste em descobrir novas propriedades e relacionamentos em dados estruturados. Os dados estruturados podem ser obtidos através da fase de pré-processamento do processo descrito no Capítulo 4 e os relacionamentos descobertos podem ser mapeados para relacionamentos não taxonômicos. A extração de axiomas, por sua vez, pode ser realizada através da aprendizagem de estrutura de redes lógicas de Markov (KOK, DOMINGOS, 2009), uma vez que as fórmulas de uma RLM podem ser mapeadas para axiomas de uma ontologia.

Outro tópico promissor para pesquisa é combinar abordagens que exploram diferentes tipos de fontes de dados. Para se tornar mais efetiva em ambientes como a Web, a abordagem aqui proposta poderia ser estendida para tirar proveito da estrutura de links da Web tradicional, por exemplo, ou ser combinada com abordagens para aprendizagem de ontologias a partir de folksonomias (BENZ, 2007) (MARINHO; BUZA; SCHMIDT-THIEME, 2008), que se tornaram populares com o advento da Web 2.0.

REFERÊNCIAS

AGIRRE, E. et al. Enriching very large ontologies using the WWW. In: ECAI 2000 WORKSHOP ON ONTOLOGY LEARNING (OL'2000), Berlin, **Proceedings…** IOS Press, 2000.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, New York, **Proceedings…** ACM, 1993. p. 207 – 216.

BENZ, D.. **Collaborative Ontology Learning. Dissertação.** 2007. 95 f. Dissertação (Mestrado) - Universität Freiburg, Freiburg, 2007.

BIEMANN, C. Ontology learning from text: a survey of methods. **LDV Forum**, [S. l.], v. 20, n. 2, p. 75 – 93, 2005.

BISHOP, C.. **Pattern recognition and machine learning.** Springer, 2006. 740 p.

BLOEHDORN, Stephan; CIMIANO, Philipp; HOTHO, Andreas. Learning ontologies to improve text clustering and classification. In: Annual Conference of the German Classification Society (GfKl 2005), 29., Magdeburg, 2005. **Proceedings…**, [S.l.] Springer, 2006, p. 334 – 34.

BRANK, J; GROBELNIK, M; MLADENIC, D. A survey of ontology evaluation techniques. In: CONFERENCE ON DATA MINING AND DATA WAREHOUSES, 8., 2005, Ljubljana. **Proceedings...** Ljubljana, 2005. p. 166 - 169.

BREWSTER, C et al. Data driven ontology evaluation. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4., 2004, Lisbon. **Proceedings...** .Lisbon, 2004.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: an overview. **Ontology learning from text**: methods, evaluation and applications. Amsterdam: IOS Press, 2005. (Frontiers in Artificial Intelligence and Applications Series, v. 123).

CARABALLO, S. Automatic construction of a hypernym-labeled noun hierarchy from text. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, 37., 1999, Maryland.

Proceedings... . Morristown: Association For Computational Linguistics, 1999. p. 120 - 126.

CARROLL, J et al. Jena: Implementing the semantic web recommendations. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 13., 2004, New York. **Proceedings...** . New York: ACM, 2004. p. 74 - 83.

CIARAMITA, M. et al. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 19., 2005, Edinburgh. **Proceedings...** . Professional Book Center, 2005. p. 659 - 664.

CIMIANO, P. et al. Learning taxonomic relations from heterogeneous sources of evidence. In: BUITELLAR, P.; CIMIANO, P.; MAGNINI, B. (Comp.).**Ontology learning from text:** Methods, applications, evaluation. Amsterdam: IOS Press, 2005. p. 59-73.

CIMIANO, P. **Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.** New York: Springer-verlag, 2006

CIMIANO, P.; HOTHO, A.; STAAB, S. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE (ECAI), Valencia, **Proceedings…** IOS Press, 2004a. p. 435 – 443.

CIMIANO, P.; HOTHO, A.; STAAB, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. **Journal Of Artificial Intelligence Research**, [S. l.], v. 24, p.305 - 339, 2004b.

CIMIANO, P.; STAAB, S. Learning concept hierarchies from text with a guided hierarchical clustering algorithm. In: WORKSHOP ON LEARNING AND EXTENDING LEXICAL ONTOLOGIES WITH MACHINE LEARNING METHODS, 2005, Bonn. **Proceedings...** . 2005.

CIMIANO, P.; VÖLKER, J.; STUDER, R. Ontologies on demand? – a description of the state-of-the-art, applications, challenges and trends for ontology learning from text. **Information, Wissenschaft Und Praxis**, [S. l.], v. 57, n. 6-7, p. 315 - 320, 2006.

CIMIANO, P.; VÖLKER, J.. Text2onto: a framework for ontology learning and data-driven change discovery. In: INTERNATIONAL CONFERENCE ON APPLICATIONS

OF NATURAL LANGUAGE TO INFORMATION SYSTEMS, 10., 2005, Alicante. **Proceedings...** . Alicante: Springer, 2005. p. 227 - 238.

CUNNINGHAM, H. GATE, a general architecture for text engineering.**Computers And The Humanities**, [S. l.], v. 36, n. 2, p. 223 -2 54, 2002.

D'AMATO, C.; FANIZZI, N.; ESPOSITO, F.. A semantic similarity measure for expressive description logics. In: CONVEGNO ITALIANO DI LOGICA COMPUTAZIONALE, 2005, Rome. **Proceedings...** . Rome, 2005.

DELLSCHAFT, K.; STAAB, S. On how to perform a gold standard based evaluation of ontology learning. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 5., 2006, Athens. **Proceedings...** . [S.l.]: Springer, 2006. p. 228 - 241.

DELLSCHAFT, K. **Measuring the Similarity of Concept Hierarchies and its Influence on the Evaluation of Learning Procedures.** 2005. 83 f. Dissertação (Mestrado) - Universität Koblenz Landau, Koblenz, 2005.

DOMINGOS, P. et al. Markov Logic. In: RAEDT, L. De et al. (Comp.).**Probabilistic Inductive Logic Programming.** New York: Springer, 2008. p. 92 - 117.

DRUMOND, L.; GIRARDI, R. Uma análise das técnicas e ferramentas para o desenvolvimento de aplicações para a Web Semântica. **REIC - Revista Eletrônica de Iniciação Científica**, v. 6, n.1, mar. 2006.

DRUMOND, L.; GIRARDI, R. A survey of ontology learning procedures. In: WORKSHOP ON ONTOLOGIES AND THEIR APPLICATIONS (WONTO 2008), Salvador, **Proceedings**…, CEURS, 2008.

DRUMOND, L.; GIRARDI, R. An Experiment Using Markov Logic Network to Extract Ontology Concepts from Text. In: INTERNATIONAL WORKSHOP ON WEB AND TEXT INTELLIGENCE (WTI 2009), 2., 2009, São Carlos. **Proceedings...** . São Carlos, 2009. p. 1 - 10. CD-ROM.

DRUMOND, L.; GIRARDI, R. Extracting Ontology Concept Hierarchies from Text using Markov Logic. In: INTERNATIONAL SYMPOSIUM ON APPLIED COMPUTING, 25., 2010, Sierre. **Proceedings...** . Sierre: ACM, 2010. (a ser publicado)

DRUMOND, L.; GIRARDI, R.; SILVA, F. A similarity analysis model for semantic web information filtering applications. In: INTERNATIONAL CONFERENCE ON

SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, Redwood City, **Proceedings**… Redwood City: Knowledge Engineering Institute, 2008. p. 558 - 563.

FAURE, D.; NEDELLEC, C. Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM.**Lecture Notes In Computer Science**, Berlin, v. 1621, p. 329 - 334, 1999. Springer.

FELLBAUM, C. **Wordnet:** An Electronic Lexical Database. Cambridge: MIT Press, 1998.

FRIEDMAN, N. et al. Learning probabilistic relational models. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 16. Stockholm, **Proceedings...**, Morgan Kaufmann, 1999. v. 16, p. 1300 - 1309.

GILKS, W.; RICHARDSON, S.; SPIEGELHALTER, D. **Markov chain Monte Carlo in practice.** [S. l.]: Chapman & Hall/CRC, 1996.

GIRARDI, R. **Classification and Retrieval of Software through their Descriptions in Natural Language.**1995. 212 f. Tese (Doutorado) - Université de Geneve, Genebra, 1995.

GIRARDI, R.; IBRAHIM, B. Using english to retrieve software. **Journal Of Systems Software**: Special Issue on Software Reusability, New York, v. 30, n. 3, p. 249 - 270, 1995. Elsevier.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. **International Journal Of Human-computer Studies**, Duluth, Usa, v. 43, n. 5-6, p. 907-928, 1995. Academic Press.

HAASE, P.; VOLKER, J. Ontology Learning and Reasoning - Dealing with Uncertainty and Inconsistency. In: WORKSHOP ON UNCERTAINTY REASONING FOR THE SEMANTIC WEB, 1., 2005, Galway. **Proceedings...** .2005. p. 45 - 55.

HARRIS, Z. **Mathematical Structures of Language.** New York: Wiley, 1968.

HEARST, M. Automatic acquisition of hyponyms from large text corpora. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14., 1992, Nantes. **Proceedings...** . Morristown: Association For Computational Linguistics, 1992. p. 539 - 545.

HINDLE, D. Noun classification from predicate-argument structures. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 28., 1990, Pittsburgh. **Proceedings...** . Morristown: Association For Computational Linguistics, 1990. p. 268 - 275.

HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. **Machine Learning**, [S. l.], v. 42, n. 1, p. 177 - 196, 2001.

IWANSKA, L.; MATA, N.; KRUGER, K. Fully automatic acquisition of taxonomic knowledge from large corpora of texts: limited-syntax knowledge representation system based on natural language. In: IWANSKA, L.; SHAPIRO, S. (Ed.). **Natural language processing and knowledge representation:** language for knowledge and knowledge for language. Cambridge, USA: MIT Press, 2000. p. 335-345.

KAO, H. et al. Mining Web Informative Structures and Contents Based on Entropy Analysis. **IEEE Transactions On Knowledge And Data Engineering**, Piscataway, v. 16, n. 1, p. 41-55, 2004.

KAUTZ, H.; SELMAN, B.; JIANG, Y. A general stochastic approach to solving problems with hard and soft constraints. **The Satisfiability Problem: Theory And Applications**, [S. l.], p. 573-586, 1997.

KAVALEC, M.; SVATEK, V. A study on automated relation labelling in ontology learning. In: BUITELLAR, P.; CIMIANO, P.; MAGNINI, B. (Ed.). Ontology learning from text: Methods, applications, evaluation. Amsterdam: IOS Press, 2005. p. 44 - 58.

KERSTING, K. **An Inductive Logic Programming Approach to Statistical Relational Learning.** Amsterdam: IOS Press, 2006.

K. KERSTING, L. DE RAEDT. **Bayesian Logic Programs**. Technical Report No. 151, Institute for Computer Science, University of Freiburg, Germany, April 2001.

KINDERMANN, R.; SNELL, J. **Markov Random Fields and Their Applications.** [S. l.]: American Mathematical Society, 1980.

KOK, S.; DOMINGOS, P. Statistical Predicate Invention. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 24., 2007, Corvallis. **Proceedings...** . New York: ACM Press, 2007. p. 433 - 440.

KOK, S.; DOMINGOS, P.. Extracting Semantic Networks from Text via Relational Clustering. In: EUROPEAN CONFERENCE ON MACHINE LEARNING, 19., 2008, Antwerp. **Proceedings...** . Berlin: Springer-Verlag, 2008. p. 624 - 639.

KOK, S.; DOMINGOS, P.. Learning Markov logic network structure via hypergraph lifting. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 26., 2009, Montreal. **Proceedings...** . New York: ACM Press, 2009. p. 505 - 512.

KOK, S. et al. **The alchemy system for statistical relational AI.** Seattle: Department of Computer Science And Engineering, University Of Washington, 2009.

LEHMANN, J.; HITZLER, P. A refinement operator based learning algorithm for the alc description logic. In: INTERNATIONAL CONFERENCE ON INDUCTIVE LOGIC PROGRAMMING, 17., 2007, Corvallis. **Proceedings...** . Berlin: Springer-verlag, 2007. p. 147 - 160.

LOWD, D.; DOMINGOS, P.. Efficient weight learning for Markov logic networks. In: EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES, 11., 2007, Warsaw. **Proceedings...** . Berlin: Springer-verlag, 2007. p. 200 - 211.

MAEDCHE, A.; STAAB, S.. Discovering Conceptual Relations from Text. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., 2000, Berlin. **Proceedings...** . Amsterdam: IOS Press, 2000. p. 321 - 325.

MAEDCHE, A.; STAAB, S. Measuring similarity between ontologies. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT, 13., 2002, London. **Proceedings...** . Berlin: Springer-verlag, 2002. p. 251 - 263.

MAEDCHE, A.; STAAB, S. Ontology learning for the Semantic Web. **IEEE Intelligent Systems**, Piscataway, v. 16, n. 2, p. 72-79, 2001.

MAEDCHE, A.; STAAB, S. Ontology learning. In: In: STAAB, S.; STUDER, R. (Ed.). **Handbook on Ontologies.** Berlin: Springer, 2004. p. 173-190.

MARCUS, M.; SANTORINI, B.; MARCINKIEWICZ, M. Building a Large Annotated Corpus of English: The Penn Treebank. **Computational Linguistics**: Special Issue on Using Large Corpora, [S. l.], v. 19, n. 2, p. 313-330, 1993.

MARINHO, L.; BUZA, K.; SCHMIDT-THIEME, L.. Folksonomy-based Collabulary Learning. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 7., 2008, Karlsruhe. **Proceedings... .** Berlin: Springer-verlag, 2008. p. 261 - 276.

MARNEFFE, M.; MANNING, C.. The Stanford typed dependencies representation. In: WORKSHOP ON CROSS-FRAMEWORK AND CROSS-DOMAIN PARSER EVALUATION, 2008, Manchester. **Proceedings... .** 2008. p. 1 - 8.

MIDDLETON, S.; SHADBOLT, N.; ROURE, D. Ontological user profiling in recommender systems. **ACM Transactions On Information Systems**, New York, v. 22, n. 1, p. 54-88, 2004.

MITCHELL, T. **Machine Learning.** Singapore: Mcgraw Hill, 1997.

MITCHELL, T et al. **The discipline of machine learning.** Pittsburgh: School Of Computer Science, Carnegie Mellon University, 2006. 9 p.

MUGGLETON, S.. Stochastic logic programs. **Advances In Inductive Logic Programming**, [S. l.], v. 32, p. 254-264, 1996.

MUGGLETON, S.; RAEDT, L.. Inductive Logic Programming: Theory and Methods. **Journal Of Logic Programming**, [S. l.], v. 19, n. 20, p. 629-679, 1994.

NEVILLE, J.; RATTIGAN, M.; JENSEN, D.. Statistical relational learning: Four claims and a survey. In: WORKSHOP ON LEARNING STATISTICAL MODELS FROM RELATIONAL DATA, 2003, Acapulco. **Proceedings... .** 2003.

OHTA, T.; TATEISI, Y.; KIM, J. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In: INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2., 2002, San Diego. **Proceedings... .** San Francisco: Morgan Kaufmann Publishers, 2002. p. 82 - 86.

PATEL, C et al. A semantic web portal for ontology searching, ranking and classification. In: WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT, 1., 2003, New Orleans. **Proceedings...** . New York: ACM Press, 2003. p. 58 - 61.

PHAN, X.; NGUYEN, L.; HORIGUCHI, S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In: INTERNATIONAL

CONFERENCE ON WORLD WIDE WEB, 17., 2008, Beijing. **Proceedings... .** New York: ACM Press, 2008. p. 91 - 100.

POON, H.; DOMINGOS, P. Sound and efficient inference with probabilistic and deterministic dependencies. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 23., 2006, Boston. **Proceedings... .** Boston: MIT Press, 2006. p. 458 - 463.

POON, H.; DOMINGOS, P.. Joint Unsupervised Coreference Resolution with Markov Logic. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2008, Honolulu. **Proceedings... .** Honolulu: Association For Computational Linguistics, 2008. p. 649 - 658.

POPESCUL, A. et al. Statistical relational learning for document mining. In: INTERNATIONAL CONFERENCE ON DATA MINING, 3., 2003, Melbourne. **Proceedings... .** Washington: IEEE Computer Society, 2003. p. 275 - 282.
PORZEL, R.; MALAKA, R. A task-based approach for ontology evaluation. In: ECAI 2004 WORKSHOP ON ONTOLOGY LEARNING AND POPULATION, 2004, Valencia. **Proceedings... .** 2004.

RAEDT, L.; KERSTING, K. Probabilistic logic learning. **ACM SIGKDD Explorations Newsletter**, New York, v. 5, n. 1, p. 31-48, 2003. ACM Press.

RESNIK, P. Selectional preference and sense disambiguation. In: ANLP WORKSHOP "TAGGING TEXT WITH LEXICAL SEMANTICS: WHY WHAT AND HOW?", 1., 1997, Washington. **Proceedings... .** Washington: Association For Computational Linguistics, 1997. p. 52 - 57.

RESNICK, P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. **Journal of Artificial Intelligence Research,** [S. l.], v. 11, p. 95–130, 1999.

RICHARDSON, M.; DOMINGOS, P. Markov logic networks**. Machine Learning**, Berlin, v. 62, n. 1, p. 107-136, 2006. Springer.

RUSSEL, S.; NORVIG, P. **Artificial Intelligence:** A Modern Approach. 2. ed. [S. l.]: Prentice Hall, 2003.

SALTON, G.; BUCKLEY, C.. **Term Weighting Approaches in Automatic Text Retrieval.** [S. l.]: Cornell University, 1987.

SHADBOLT, N.; HALL, W.; BERNERS-LEE, T. The semantic web revisited. **IEEE Intelligent Systems**, Piscataway, v. 21, n. 3, p. 96–101, 2006.

SHAMSFARD, M.; BARFOROUSH, A. An introduction to HASTI: An ontology learning system. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING, 6., 2002, Banff. **Proceedings...** . ACTA Press, 2002.

SHAMSFARD, M.; BARFOROUSH, A. The state of the art in ontology learning: a framework for comparison. **Knowledge Engineering Review**, New York, v. 18, n. 4, p. 293–316, 2003.

SHANNON, C., A mathematical theory of communication. **Bell System Technical Journal**, [S. l.], v. 27, p. 379-423 and 623-656, 1948.

SILVA, F.; GIRARDI, R.; DRUMOND, L. A knowledge-based retrieval model. In: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, 21., 2009, Boston. **Proceedings... .** Boston: Knowledge Systems Institute Graduate School, 2009. p. 558 - 563.

SINGLA, P.; DOMINGOS, P. Multi-relational record linkage. In: KDD WORKSHOP ON MULTI-RELATIONAL DATA MINING, 1., 2004, Seattle. **Proceedings... .** New York: ACM Press, 2004. p. 31 - 48.

SMITH, B. Ontology. In: FLORIDI, L. (Ed.). **The Blackwell guide to philosophy of computing and information.** Malden: Blackwell, 2003. p. 155-166.

SNOW, R.; JURAFSKY, D.; NG, A.Y. Learning syntactic patterns for automatic hypernym discovery. **Advances in Neural Information Processing Systems**, [S. l.], v. 17, p. 1297-1304, 2005.

VAILET, D.; FERNANDEZ, M.; CASTELLS, P. An Ontology-Based Information Retrieval Model. In: DAVIES, J. et al. (Ed.). **The Semantic Web: Research and Applications.** Berlin: Springer, 2005. p. 455-470.

WORLD WIDE WEB CONSORTIUM. OWL: Web Ontology Language Guide. 10 fev. 2004. Disponível em:<http://www.w3.org/TR/owl-guide/>. Acesso em: 29 ago. 2009.

WU, F.; WELD, D. Automatically refining the wikipedia infobox ontology. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 17., 2008, Beijing. **Proceedings... .** New York: ACM Press, 2008. p. 635 - 644.

ZAVITSANOS, E et al. Discovering subsumption hierarchies of ontology concepts from text corpora. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, 6., 2007, Silicon Valley. **Proceedings... .** Washington: IEEE Computer Society, 2007. p. 402 - 408.

ZHAO, Y.; KARYPIS, G. Discovering subsumption hierarchies of ontology concepts from text corpora. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 11., 2002, Mclean. **Proceedings...** . New York: ACM Press, 2002. p. 515 - 524.

APÊNDICES

APÊNDICE A - Rede lógica de Markov utilizada pela técnica PRECE

Os pesos foram aprendidos a partir do corpus Genia (OHTA; TATEISI; KIM, 2002) no experimento descrito na Seção 5.4.

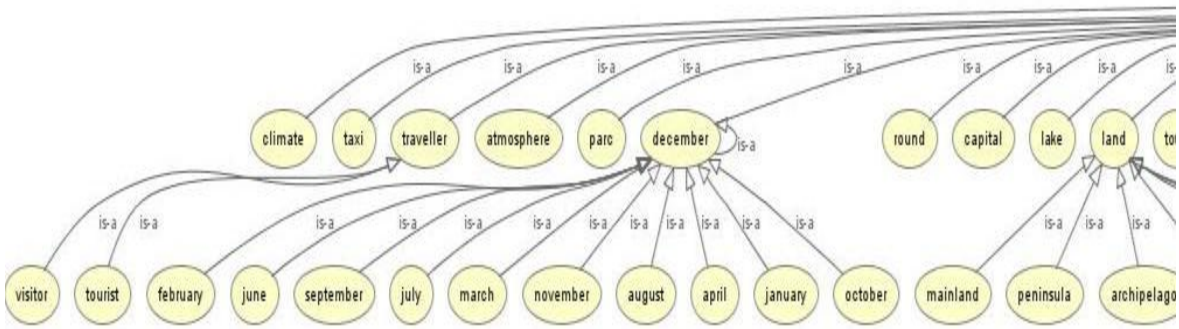
Peso	*Fórmula*
5.88725	$\neg SameMeaning(t_1, \mathrm{t}_2)$
0.0345846	$HasTerm(d, t_1) \wedge HasTerm(d, t_2) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
3.3345	$Hypernym(t_3, t_1) \wedge Hypernym(t_3, t_2) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
0.790565	$Hypernym(t_3, t_1) \wedge Hypernym(t_4, t_2) \wedge \neg SameMeaning(t_3, t_4) \Rightarrow \neg SameMeaning(t_1, t_2)$
0.778915	$Depends(t_3, t_1, \mathrm{Dep}) \wedge Depends(t_3, t_2, \mathrm{Dep}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
2.2124	$Depends(t_3, t_1, \mathrm{NSUBJ}) \wedge Depends(t_3, t_2, \mathrm{NSUBJ}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
1.45431	$Depends(t_3, t_1, \mathrm{CONJ}) \wedge Depends(t_3, t_2, \mathrm{CONJ}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
2.05647	$Depends(t_3, t_1, \mathrm{DOBJ}) \wedge Depends(t_3, t_2, \mathrm{DOBJ}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
2.44838	$Depends(t_3, t_1, \mathrm{CCOMP}) \wedge Depends(t_3, t_2, \mathrm{CCOMP}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
1.55356	$Depends(t_3, t_1, \mathrm{NN}) \wedge Depends(t_3, t_2, \mathrm{NN}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
1.50767	$Depends(t_3, t_1, \mathrm{NSUBJPASS}) \wedge Depends(t_3, t_2, \mathrm{NSUBJPASS}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
1.53968	$Depends(t_3, t_1, \mathrm{APPOS}) \wedge Depends(t_3, t_2, \mathrm{APPOS}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
1.82079	$Depends(t_3, t_1, \mathrm{RCMOD}) \wedge Depends(t_3, t_2, \mathrm{RCMOD}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
0.821218	$Depends(t_3, t_1, \mathrm{TMOD}) \wedge Depends(t_3, t_2, \mathrm{TMOD}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
0.233324	$Depends(t_3, t_1, \mathrm{AMOD}) \wedge Depends(t_3, t_2, \mathrm{AMOD}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$
1.06635	$Depends(t_3, t_1, \mathrm{POSS}) \wedge Depends(t_3, t_2, \mathrm{POSS}) \Rightarrow SameMeaning(t_1, \mathrm{t}_2)$

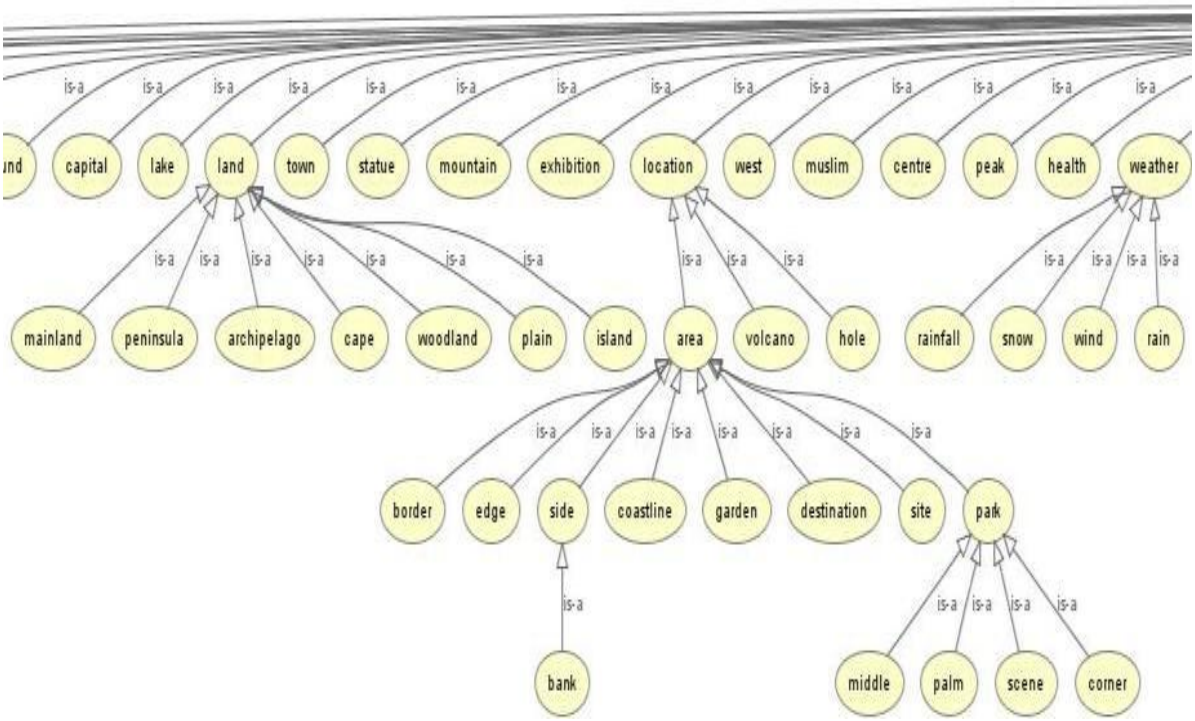
APÊNDICE B - Rede lógica de Markov utilizada pela técnica PREHE

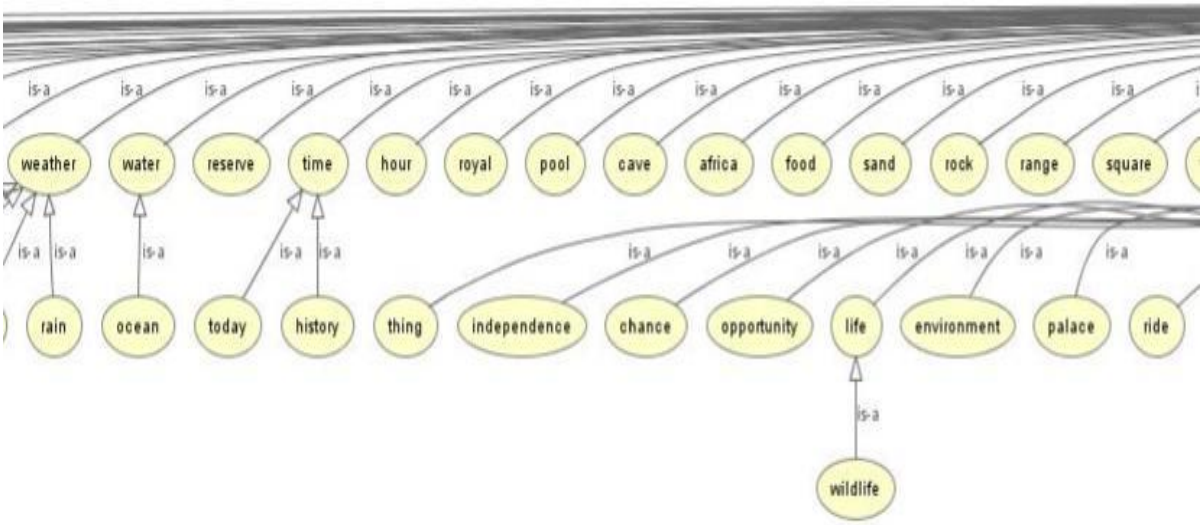
Os pesos foram aprendidos a partir do corpus Genia (OHTA; TATEISI; KIM, 2002) no experimento descrito na Seção 5.5.

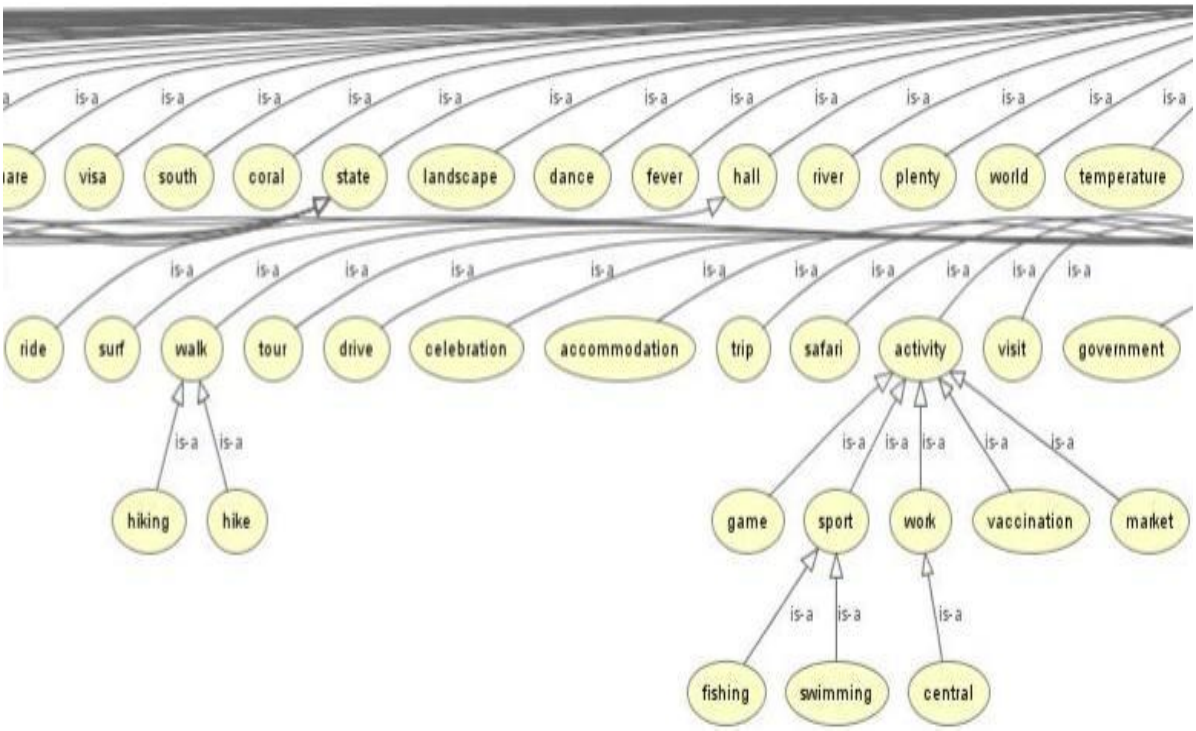
Peso	*Fórmula*
3.48646	$\neg Kind_of(c_1, c_2)$
14.3513	$Kind_Of(c_1, c_2) \Longrightarrow \neg Kind_of(c_2, c_1)$
1.54267	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Hypernym(t_1, t_2) \Longrightarrow Kind_of\ (c_1, c_2)$
1.33383	$F(c_1, t_1) \wedge F(c_2, t_1) \wedge F(c_1, t_2) \wedge \neg F(c_2, t_2) \Longrightarrow Kind_Of(c_1, c_2)$
0.693368	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{DEP}) \wedge Depends(t_3, t_2, \text{DEP}) \Longrightarrow Kind_of\ (c_1, c_2)$
4.29078	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{NSUBJ}) \wedge Depends(t_3, t_2, \text{NSUBJ}) \Longrightarrow Kind_of\ (c_1, c_2)$
0.91315	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{CONJ}) \wedge Depends(t_3, t_2, \text{CONJ}) \Rightarrow Kind_of\ (c_1, c_2)$
1.61126	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{DOBJ}) \wedge Depends(t_3, t_2, \text{DOBJ}) \Rightarrow Kind_of\ (c_1, c_2)$
0.199214	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{CCOMP}) \wedge Depends(t_3, t_2, \text{CCOMP}) \Rightarrow Kind_of\ (c_1, c_2)$
4.22751	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{NN}) \wedge Depends(t_3, t_2, \text{NN}) \Rightarrow Kind_of\ (c_1, c_2)$
4.18069	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{NSUBJPASS}) \wedge Depends(t_3, t_2, \text{NSUBJPASS}) \Rightarrow Kind_of\ (c_1, c_2)$
3.21001	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{APPOS}) \wedge Depends(t_3, t_2, \text{APPOS}) \Rightarrow Kind_of\ (c_1, c_2)$
1.52378	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{RCMOD}) \wedge Depends(t_3, t_2, \text{RCMOD}) \Rightarrow SameMeaning(t_1, t_2)$
0.171287	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{TMOD}) \wedge Depends(t_3, t_2, \text{TMOD}) \Rightarrow Kind_of\ (c_1, c_2)$
0.2355	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, \text{AMOD}) \wedge Depends(t_3, t_2, \text{AMOD}) \Rightarrow Kind_of\ (c_1, c_2)$

APÊNDICE C - Ontologia gerada com as técnicas PRECE e PREHE a partir do corpus LonelyPlanet

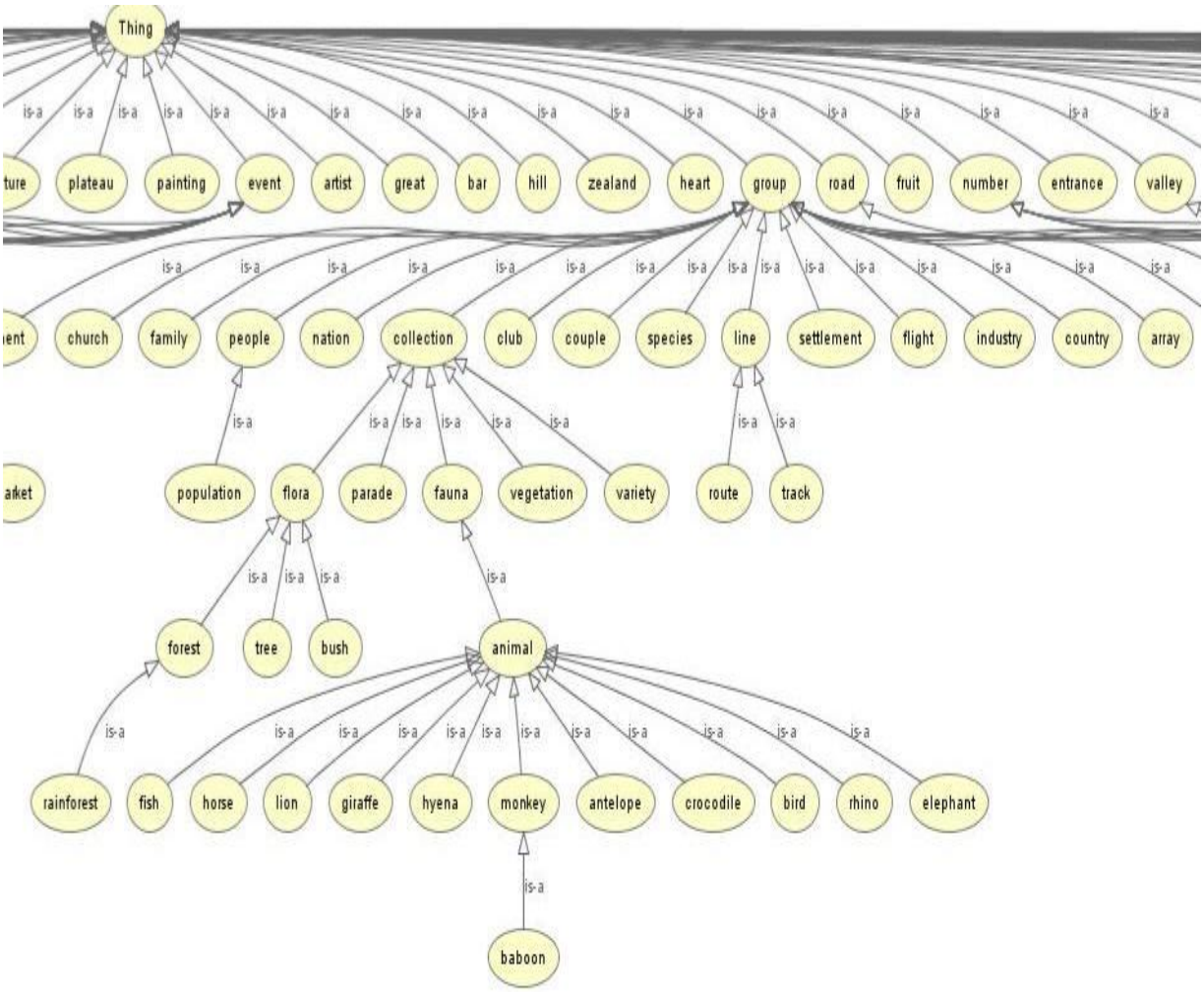


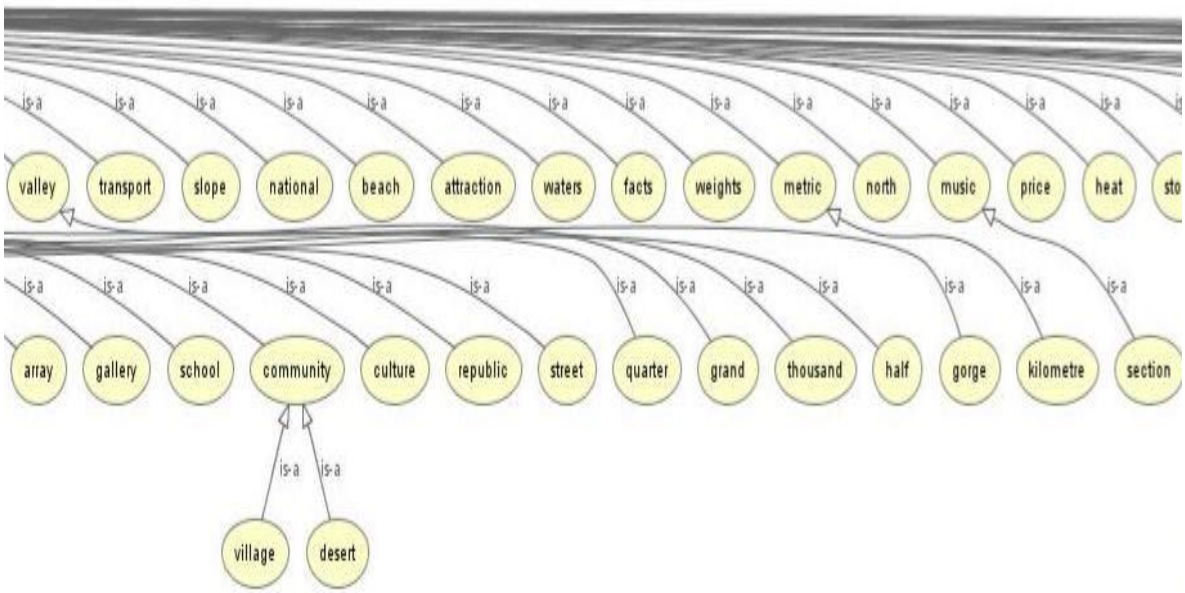


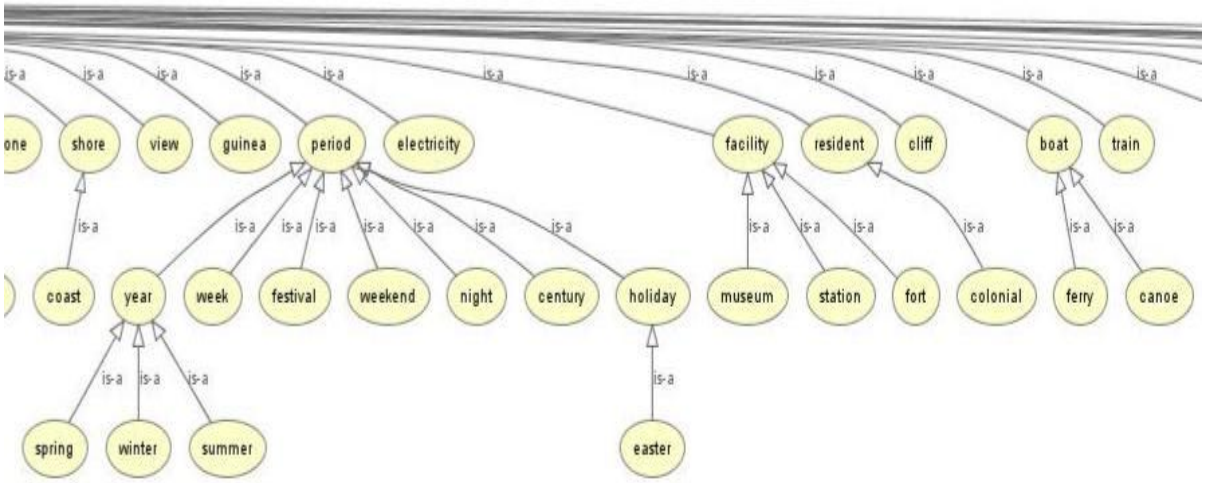

is-a
weather
water
reserve
time
hour
royal
pool
cave
africa
food
sand
rock
range
square
rain
ocean
today
history
thing
independence
chance
opportunity
life
environment
palace
ride
wildlife

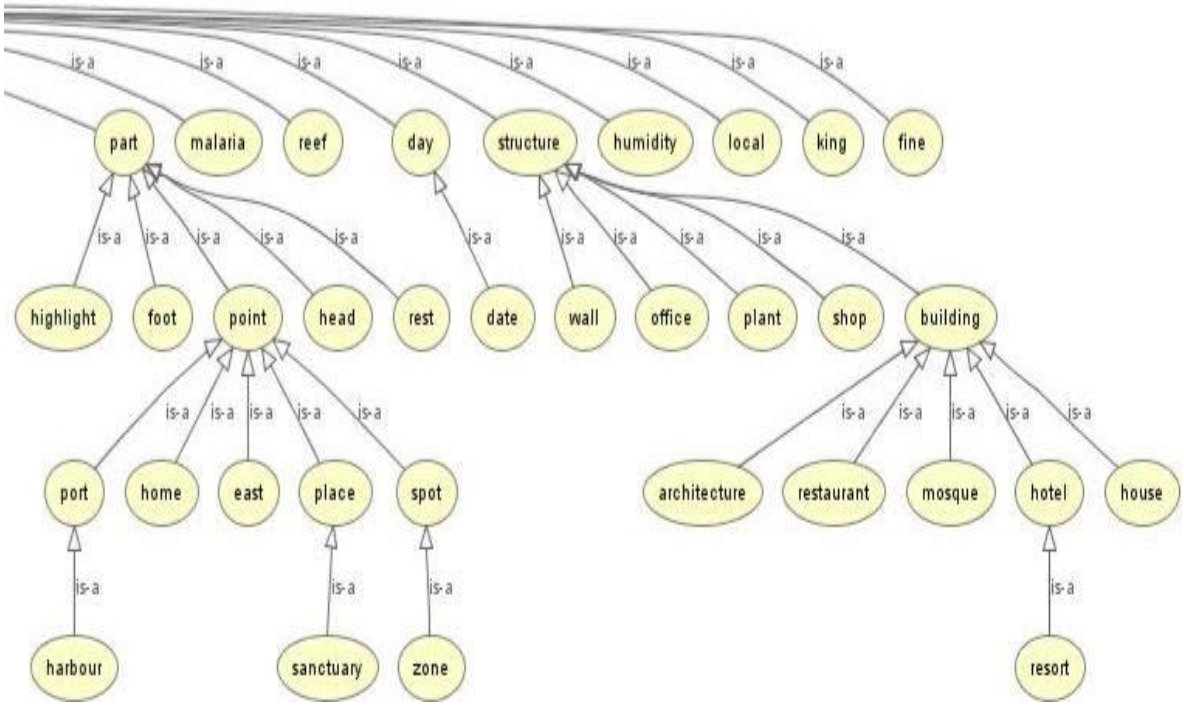

is-a
visa
south
coral
state
landscape
dance
fever
hall
river
plenty
world
temperature
ride
surf
walk
tour
drive
celebration
accommodation
trip
safari
activity
visit
government
hiking
hike
game
sport
work
vaccination
market
fishing
swimming
central

,


Thing
is-a
ture
plateau
painting
event
artist
great
bar
hill
zealand
heart
group
road
fruit
number
entrance
valley
ient
church
family
people
nation
collection
club
couple
species
line
settlement
flight
industry
country
array
arket
population
flora
parade
fauna
vegetation
variety
route
track
forest
tree
bush
animal
rainforest
fish
horse
lion
giraffe
hyena
monkey
antelope
crocodile
bird
rhino
elephant
baboon


is-a
valley
transport
slope
national
beach
attraction
waters
facts
weights
metric
north
music
price
heat
array
gallery
school
community
culture
republic
street
quarter
grand
thousand
half
gorge
kilometre
section
village
desert


is-a
shore
view
guinea
period
electricity
facility
resident
cliff
boat
train
coast
year
week
festival
weekend
night
century
holiday
museum
station
fort
colonial
ferry
canoe
spring
winter
summer
easter


is-a
part
malaria
reef
day
structure
humidity
local
king
fine
highlight
foot
point
head
rest
date
wall
office
plant
shop
building
port
home
east
place
spot
architecture
restaurant
mosque
hotel
house
harbour
sanctuary
zone
resort

ANEXOS

ANEXO A - Conjunto de Tags Penn Treebank

Conjunto de tags descritas por Marcus, Santorini e Marcinkiewicz (1993).

Tag	*Significado*
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	To
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

ANEXO B - As dependências de Stanford

Dependências descritas e detalhadas por Marneffe e Manning (2008).

Tipo da Dependência	***Significado***
dep	Dependent
aux	Auxiliary
auxpass	passive auxiliary
cop	Copula
arg	Argument
agent	Agent
comp	Complement
acomp	adjectival complement
attr	Attributive
ccomp	clausal complement with internal subject
xcomp	clausal complement with external subject
compl	Complementizer
obj	Object
dobj	direct object
iobj	indirect object
pobj	object of preposition
mark	marker (word introducing an advcl)
rel	relative (word introducing a rcmod)
subj	Subject
nsubj	nominal subject
nsubjpass	passive nominal subject
csubj	clausal subject
csubjpass	passive clausal subject
cc	Coordination
conj	Conjunct
expl	expletive (expletive \there")
mod	modifier
abbrev	abbreviation modifier
amod	adjectival modifier
appos	appositional modifier
advcl	adverbial clause modifier
purpcl	purpose clause modifier
det	Determiner
predet	Predeterminer
preconj	Preconjunct
infmod	infinitival modifier
partmod	participial modifier
advmod	adverbial modifier
neg	negation modifier
rcmod	relative clause modifier

quantmod	quantifier modifier
tmod	temporal modifier
measure	Measure
nn	noun compound modifier
num	numeric modifier
number	element of compound number
prep	prepositional modifier
poss	possession modifier
possessive	possessive modifier ('s)
prt	phrasal verb particle
parataxis	Parataxis
punct	Punctuation
ref	Referent
sdep	semantic dependent
xsubj	controlling subject