

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

JONE DOS SANTOS SODRÉ CORREIA

**UM PROCESSO PARA A AQUISIÇÃO DE RELAÇÕES
TAXONÔMICAS DE UMA ONTOLOGIA**

São Luís
2011

JONE DOS SANTOS SODRÉ CORREIA

**UM PROCESSO PARA A AQUISIÇÃO DE RELAÇÕES
TAXONÔMICAS DE UMA ONTOLOGIA**

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
2011

Correia, Jone dos Santos Sodré

Um Processo para a Aquisição de Relações Taxonômicas de uma Ontologia

/ Jone dos Santos Sodré Correia. – 2011.
96 f.

Impresso por computador (Fotocópia).

Orientadora: Maria del Rosario Girardi Gutiérrez.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Curso de Pós-graduação em Engenharia de Eletricidade, 2011.

1. Ontologias (Computação) – relações. 2. Ontologias (Computação) - conceitos. 3. Processamento de Linguagem Natural. I. Gutiérrez, María del Rosario Girardi, orientadora. I Título.

CDU 004.775

UM PROCESSO PARA A AQUISIÇÃO DE RELAÇÕES TAXONÔMICAS DE UMA ONTOLOGIA

Jone dos Santos Sodré Correia

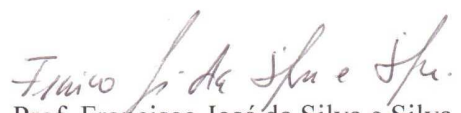
Dissertação aprovada em 06 de maio de 2011.



Profa. María del Rosario Girardi Gutiérrez, Ph.D.
(Orientadora)



Prof. Omar Andres Carmona Cortés, Dr.
(Membro da Banca Examinadora)



Prof. Francisco José da Silva e Silva, Dr.
(Membro da Banca Examinadora)

A todas as pessoas gostam de mim
e, principalmente, as que me amam

AGRADECIMENTOS

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

Aos meus pais Jonas Correia e Luzia Correia, pela educação e apoio que me deram durante toda a minha vida. Por serem pessoas excelentes, sempre preocupadas seus filhos. Por serem para os seus filhos pais amados. Por serem eles.

Ao meu irmão Jean, pelo apoio, pela paciência, por sempre entender o meu lado nos momentos de conflito e por ser tão amigo.

A minha namorada Mara que sempre desejou o meu bem, por ser uma pessoa tão importante para mim e por sempre me apoiar e confortar.

A todos os meus familiares pela sua importância na vida de qualquer ser humano.

À Prof^a Rosario Girardi, por sempre ter cobrado o máximo de mim, por aceitar me ceder a sua orientação e pelas palavras que de alguma forma me enriqueceram na vida.

A todos os meus colegas do GESEC, em especial a Carla, Ivo e Raimundo, que contribuíram complementando minha orientação neste trabalho. Pelas palavras de apoio e conforto quando eu precisei.

A todos os meus atuais e ex-colegas do NTI que tanto me deram apoio e tanto me ensinaram.

A todos os meus atuais e ex-colegas da FGS, em especial o André, a Carmem, o Tiago e o Rayan. Por serem ótimas pessoas.

A todos os meus amigos. Pois todos são muito importantes para mim.

"Se você consegue aprender através dos duros golpes, você também consegue aprender pelos suaves toques."
(Carolyn Kenmore)

RESUMO

Ontologias são uma forma de representação de conhecimento capaz de expressar um conjunto de entidades e suas relações, restrições, axiomas e vocabulário de um determinado domínio. A construção manual de ontologias por especialistas de domínio e engenheiros de conhecimento é uma tarefa cara e demorada e a automatização/semi-automatização desta tarefa é uma necessidade. O aprendizado de ontologias visa automatizar ou semi-automatizar a identificação de elementos de uma ontologia como classes, relações taxonômicas e não-taxonômicas, propriedades e axiomas de fontes textuais. Este trabalho propõe um processo de aprendizagem automática de ontologias a partir de fontes textuais enfocando a aplicação de técnicas de processamento de linguagem natural para adquirir relações taxonômicas. Alguns experimentos utilizando um corpus jurídico foram realizados para a avaliação da abordagem proposta. Os resultados iniciais são promissores.

Palavras-chave: Ontologias, Aprendizagem de ontologias. Processamento da Linguagem Natural. Relacionamentos Taxonômicos

ABSTRACT

Ontologies are an approach for knowledge representation capable of expressing a set of entities and their relationships, constraints, axioms and vocabulary of a given domain. Manual construction of ontologies by domain experts and knowledge engineers is an expensive and time consuming task so, automatic and/or semi-automatic approaches are needed. Ontology Learning looks for automatically or semi-automatically identifying ontology elements like classes, taxonomic and non-taxonomic relationships, properties and axioms from textual resources. This work proposes a process for automatic learning of ontologies from text focusing on the application of natural language processing techniques to acquire taxonomic relationships. Some experiments using a legal corpus were conducted in order to evaluate it. Initial results are promising.

Keywords: Ontology, Ontology Learning, Natural Language Processing, Taxonomic relationships

LISTA DE FIGURAS

Figura 1: Exemplo de parte de uma ontologia que descreve o domínio do direito de família.....	19
Figura 2: Taxonomias em redes semânticas e em lógica de primeira ordem	21
Figura 3: Hierarquia do aprendizado	22
Figura 4: Pipeline do Processamento da Linguagem Natural	24
Figura 5: Exemplo de Synset do WordNet	25
Figura 6: Exemplo de resultado das anotações realizado pelo GATE	28
Figura 7: Processo geral para extração de hierarquias.....	31
Figura 9: Evidências para identificação de hierarquias técnica PREHE.....	33
Figura 10: Formulas determinação pesos na técnica PREHE.....	33
Figura 11: Etapas da técnica ONTRETTEX	34
Figura 12: Arquivo .arff usado pela ONTRETTEX.....	35
Figura 13: Determinação da hierarquia de conceitos da ONTRETTEX.....	35
Figura 14: Cálculo da similaridade entre termos	36
Figura 15: Cálculo da similaridade entre dois grupos de termos.....	36
Figura 16: Cálculo da similaridade ponderada	37
Figura 17: Um processo para a aquisição de relações taxonômicas de uma ontologia	43
Figura 18: Atividades da fase “Marcação”	43
Figura 19: Exemplo de um texto marcado.....	44
Figura 20: Exemplo de aplicação da “Marcação”	45
Figura 21: Classes candidatas	46
Figura 22: Vetores de Sinônimos para “father”, “mother”, “child”, “doll” e “toy”	47
Figura 23: Vetores de Hiperônimos para “parent”, “father”, “mother”, “child”, “doll”, “mask” e “toy”	47
Figura 24: Padrões Heurísticos de Hearst	48
Figura 25: Outros padrões heurísticos para identificação de hierarquias.....	49
Figura 26: Identificação de relações taxonômicas.....	49
Figura 27: Taxonomy-NLPDumper: a extensão do projeto NPLDumper.....	52

Figura 28: Exemplos de resultados do Taxonomy-NLPDumper com o JGraph	53
Figura 29: Exemplo dos resultados do Taxonomy-NLPDumper em OWL	53
Figura 30: resultados do Taxonomy-NLPDumper em um editor de ontologias	54
Figura 31: Informações do processamento geradas pelo Taxonomy-NLPDumper ...	55
Figura 32: Apresentação parcial dos resultados em OWL em um editor de ontologias	56

LISTA DE TABELAS

Tabela 1: Exemplos de relações de hiperonímia.....	20
Tabela 2: Classificação das técnicas de aprendizado de máquina	22
Tabela 3: Relação entre tarefas de PLN e Plugins do GATE	28
Tabela 4: tabela comparativa das principais técnicas de aprendizado de relações taxômicas	40
Tabela 5: Resumo dos resultados parciais obtidos no desenvolvimento do estudo de caso.....	56

LISTA DE ABREVIATURAS E SIGLAS

FCA	– Formal Concept Analysis
GATE	– General Architecture for Text Engineering
POS	– Part of Speech
PRECE	– Probabilistic Relational Concept Extraction
PREHE	– Probabilistic Relational Hierarchy Extraction
RLM	– Rede Lógica de Markov
PLN	– Processamento de Linguagem Natural
RI	– Recuperação de informação
EI	– Extração de informação
HTML	– HyperText Markup Language
XML	– Extensible Markup Language
RTF	– Rich Text Format
DAML	– DARPA Agent Markup Language
ONTRETEX	– Ontology Taxonomical Relationships Extraction From Text
TF-IDF	– Term Frequency–Inverse Document Frequency
OWL	– Web Ontology Language

Sumário

1	Introdução	14
1.1	Objetivos	15
1.2	Estrutura do Trabalho.....	16
2	Aprendizagem de Ontologias	17
2.1	Ontologias	17
2.2	Aprendizagem de relações taxonômicas.....	20
2.2.1	Taxonomias	20
2.2.2	Aprendizado de Máquina.....	21
2.2.3	Processamento da Linguagem Natural.....	23
2.2.4	Wordnet	25
2.2.5	GATE.....	26
2.2.6	OWL	29
2.2.7	Abordagens para o aprendizado de relações taxonômicas	29
2.3	Considerações Finais.....	41
3	Um processo para a aquisição de relações taxonômicas de uma ontologia	42
3.1	O processo.....	42
3.1.1	Marcação.....	43
3.1.2	Extração de Classes Candidatas.....	46
3.1.3	Identificação de Hiperônimos e Sinônimos.....	46
3.1.4	Identificação e Representação de Relações Taxonômicas	48
3.2	Considerações Finais.....	50
4	Avaliação do processo proposto	51
4.1	Uma ferramenta para aquisição de relações taxonômicas de uma ontologia 51	
4.2	Estudo de Caso.....	55
4.2.1	O corpus utilizado	55
4.2.2	Resultados.....	56
4.3	Considerações Finais.....	57
5	Conclusões	59
5.1	Limitações	60
5.2	Trabalhos Futuros	61
	REFERÊNCIAS	62

1 Introdução

O aprendizado humano se dá, em sua maioria, com a análise do que é observado, ou seja, armazenamos informação e tomamos decisões a partir de fatos históricos. Muitos dos Sistemas de conhecimento também têm essa propriedade. Essa característica desses sistemas provavelmente tem demonstrado contribuições com o meio empresarial através do apoio a decisão, com buscadores de informações com resultados mais inteligentes e mais próximos das reais necessidades dos usuários, e em muitas áreas de conhecimento como, por exemplo, a medicina.

Muitos trabalhos científicos apresentam formas de representar o conhecimento. Por outro lado, pontos importantes relacionados à representação conhecimento como, por exemplo, a utilização de ontologias - que são uma forma muito precisa de representação de conhecimento de um domínio (Girardi, 2010) – apresenta o problema da escassez de ferramentas para o aprendizado automático, dado que não encontramos facilmente muitas ferramentas para esta tarefa. Além disso, mesmo com tanta necessidade do elemento conhecimento, muitos trabalhos atuais, dentre eles (Drumond, 2009) e (Ting Wang, 2006), citam dificuldades na aquisição de conhecimento, problema também chamado de gargalo da aquisição de conhecimento.

Este trabalho utiliza ontologias como representação de conhecimento, mais especificamente a formalização de ontologias adotada no GESEC (GESEC, 2010). A escolha de ontologias, para tanto, se dá pelas suas características que perfazem a capacidade de expressar um conjunto de entidades e suas relações, restrições, axiomas e vocabulário de um domínio, que, neste caso, são características desejáveis.

Para se construir e manter ontologias manualmente é necessário um profissional caro (um especialista de domínio, por exemplo), sendo um trabalho difícil e tedioso.

A redução de custos e a construção de mecanismos que possibilitem a rápida e automática aquisição de conhecimento não só resolve o problema do gargalo já citado, mas, também, toca no âmago dos sistemas de conhecimento. Logo, uma forma possível de se amenizar esses custos com construção de

ontologias manualmente é a utilização de ferramentas que permitam a automatização ou semi-automatização deste trabalho.

Esta dissertação propõe um processo que visa à identificação de relações taxonômicas de uma ontologia e mostra é possível fazer isso em uma proposta utilizando o Processamento de Linguagem Natural (PLN) (Allen, 1995), o Wordnet (Wordnet, 2010) e Padrões Heurísticos. Esse trabalho também apresenta uma ferramenta para uma avaliação preliminar do processo proposto em um estudo de caso dirigido a área do direito da família, que permite a partir de um texto de entrada a extração de relacionamentos taxonômicos e armazenamento em um arquivo em formato OWL (OWL, 2010).

1.1 Objetivos

O objetivo geral deste trabalho é contribuir com uma alternativa para contornar ou amenizar o problema do gargalo da aquisição de conhecimento com o desenvolvimento de uma ferramenta para a aprendizagem de relações taxonômicas de uma ontologia. Pretende-se, para isso, alcançar os seguintes objetivos específicos:

- Analisar e compreender os conceitos de ontologias adotados no GESEC.
- Analisar as técnicas para a aprendizagem de relações taxonômicas;
- Analisar experimentalmente e selecionar as técnicas adequadas para a aprendizagem de ontologias;
- A partir das técnicas de aprendizagem selecionadas, projetar uma ferramenta de aprendizado para a aquisição de relacionamentos taxonômicos de uma ontologia;
- Avaliar a efetividade da técnica proposta através do desenvolvimento de um estudo de caso.

1.2 Estrutura do Trabalho

Este trabalho, incluindo a introdução, está estruturado em cinco capítulos. No Capítulo 2 é apresentada a definição formal da ontologia adotada. Em seguida é feita uma revisão do estado da arte da aprendizagem de ontologias a partir de fontes textuais com ênfase na aprendizagem de relações taxonômicas.

No Capítulo 3, é apresentado o processo para aprendizagem de relações taxonômicas proposto neste trabalho. Ele utiliza técnicas de Processamento de Linguagem Natural e padrões heurísticos para trabalhar com textos na língua inglesa.

No capítulo 4 é apresentada a ferramenta Taxonomy-NLPDumper desenvolvida para automatizar a aplicação do processo proposto. É ainda descrito um estudo de caso no domínio jurídico do direito de família para avaliar os resultados da aplicação da técnica e o suporte da ferramenta.

O capítulo 5 apresenta as conclusões do trabalho, incluindo as contribuições da pesquisa e sugerindo trabalhos que podem ainda ser realizados nessa área.

2 Aprendizagem de Ontologias

Esse capítulo apresenta uma descrição sobre uma recente área de pesquisa, a aprendizagem de ontologias, com ênfase na aprendizagem de um de seus componentes, as taxonomias. Algumas das principais abordagens para a aquisição de relações taxonômicas a partir de fontes textuais são descritas e avaliadas comparativamente, dando assim, uma visão do estado da arte nessa área. O capítulo está organizado da seguinte forma: a Seção 2.1 apresenta a definição formal de ontologia que é utilizada neste trabalho. A Seção 2.2 define o problema da aprendizagem de ontologias a partir de fontes textuais e discute algumas das abordagens propostas na literatura. Por fim, a Seção 2.3 apresenta as considerações finais.

2.1 Ontologias

Para contextualizar o aprendizado de relações taxonômicas de uma ontologia é necessário primeiro definir o que é uma ontologia. Algumas definições para ontologia podem ser encontradas em (BIEMANN, 2005). Para este trabalho é assumido um conceito de ontologia específico que é apresentado nesta Seção.

Uma ontologia é uma especificação formal e explícita de uma conceituação compartilhada de um domínio de interesse (Gruber, 1995). Conceituação refere-se a um modelo abstrato de algum fenômeno do mundo. A especificação é explícita porque os conceitos utilizados e as limitações do seu uso são explicitamente definidos; formal porque é processável computacionalmente; compartilhada porque captura o conhecimento consensual, isto é, não é privada de algum indivíduo, mas aceita por um grupo.

Formalmente uma ontologia pode ser definida como a 6-tupla:

$$O = (C, H, I, R, P, A)$$

onde,

$C = C_C \cup C_I$ é o conjunto de entidades do domínio sendo modelado. O conjunto C_C é formado por classes, ou seja, conceitos que representam entidades que descrevem um conjunto de objetos (por exemplo, "Mãe" $\in C_C$) enquanto que o

conjunto C_I é formado por instâncias, ou seja, entidades únicas no domínio (por exemplo, “Anne Smith” $\in C_I$).

$H = \{\text{tipo_de}(c_1, c_2) \mid c_1 \in C_C \wedge c_2 \in C_C\}$ é o conjunto de relações taxonômicas que definem a hierarquia de classes da ontologia e são denotadas por “tipo_de(c_1, c_2)” indicando que c_1 é uma subclasse de c_2 . Um exemplo desse relacionamento é “tipo_de(Mãe, Pessoa)”.

$I = \{\text{é_um}(c_1, c_2) \mid c_1 \in C_I \wedge c_2 \in C_C\} \cup \{\text{prop}_K(c_i, \text{valor}) \mid c_i \in C_I\} \cup \{\text{rel}_K(c_1, c_2, \dots, c_n) \mid \forall i, c_i \in C_I\}$ é o conjunto de relacionamentos entre classes e instâncias (relacionamento “é um”), por exemplo “é_um(Anne, Mãe)”, prop_K define os valores específicos dos tipos de dados das instâncias, por exemplo, “data_de_nascimento(Anne Smith, 12/02/1980)” e as instâncias de relacionamentos não taxonômicos, por exemplo, “mãe_de(Anne Smith, Clara Smith)”.

$R = \{\text{rel}_K(c_1, c_2, \dots, c_n) \mid \forall i, c_i \in C_C\}$ é o conjunto de relacionamentos não taxonômicos de uma ontologia. Por exemplo, “mãe_de(Mãe, Filha)”.

$P = \{\text{prop}_K(c_i, \text{tipo}) \mid c_i \in C_C\}$ é o conjunto de propriedades das classes de uma ontologia e seu tipo de dados básico. Por exemplo, “data_de_nascimento(Mãe, dd/mm/aaaa)”.

$A = \{\text{condition}_x \Rightarrow \text{conclusion}_y(c_1, c_2, \dots, c_n) \mid \forall j, c_j \in C_C\}$ 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: $\text{condition}_x = \{(cond_1, cond_2, \dots, cond_n) \mid \forall z, cond_z \in H \cup I \cup R\}$. Por exemplo, “Mãe, Filha1, Filha2, mãe_de(Mãe, Filha1), mãe_de(Mãe, Filha2) \Rightarrow irmã_de(Filha1, Filha2)” é uma regra que indica que se duas filhas têm a mesma mãe, então as filhas são irmãs.

Para ilustrar, uma ontologia simples que descreve parte do domínio jurídico do direito da família é apresentada na Figura 1.

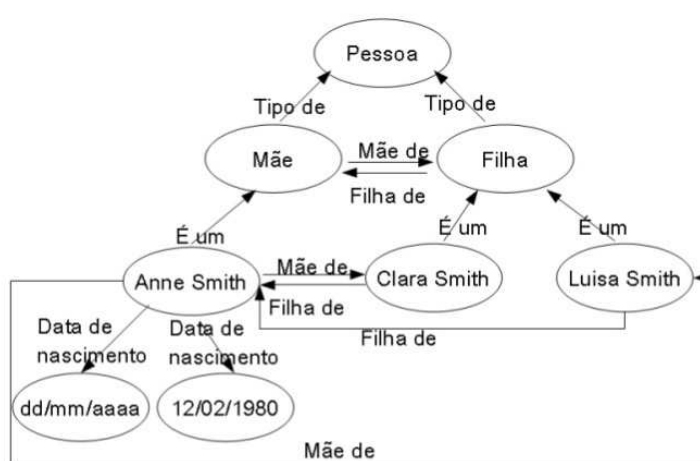


Figura 1: Exemplo de parte de uma ontologia que descreve o domínio do direito de família

Considerando a definição e a ontologia da Figura 1, os seguintes conjuntos podem ser identificados:

$$C_C = \{\text{mãe, filha, pessoa}\}$$

$$C_I = \{\text{Anne Smith, Luisa Smith, Clara Smith}\}$$

$$H = \{\text{tipo_de(mãe, pessoa), tipo_de(filha, pessoa)}\}$$

$$I = \{\text{é_um(Anne Smith, mãe), é_um(Clara Smith, filha), é_um(Luisa Smith, filha), mãe_de(Anne Smith, Clara Smith), mãe_de(Anne Smith, Luisa Smith), filha_de(Clara Smith, Anne Smith), filha_de(Luisa Smith, Anne Smith), data_de_nascimento (Anne, 12/02/1980)}\}$$

$$R = \{\text{mãe_de(mãe, filha), filha_de(filha, mãe)}\}$$

$$P = \{\text{data_de_nascimento (mãe, dd/mm/aaaa)}\}$$

$$A = \{\forall \text{ Mãe, Filha1, Filha2, mãe_de(Mãe, Filha1), mãe_de(Mãe, Filha2)} \Rightarrow \text{irmã_de(Filha1, Filha2)}\}$$

Na próxima seção é discutido o conjunto H da definição de ontologia o qual corresponde à hierarquia de conceitos. Além disso, serão discutidas diferentes abordagens para sua extração a partir de fontes textuais, o que é também conhecido como aprendizagem de relações taxonômicas.

2.2 Aprendizagem de relações taxonômicas

Esta seção apresenta alguns conceitos importantes que contribuem com o entendimento dos trabalhos relacionados apresentados na Seção 2.2.7. São apresentados os conceitos de taxonomias e uma revisão de Aprendizado de Máquina, da base lexical Wordnet, do *framework* GATE, da Linguagem OWL e do Processamento da Linguagem Natural, uma vez que a maioria dos processos para extração de relações taxonômicas utiliza combinações destes recursos.

2.2.1 Taxonomias

Segundo (TAN et al., 2009), uma hierarquia (ou taxonomia) é uma organização multinível das entidades ou conceitos definidos em um determinado domínio. Neste trabalho, classes ou conceitos devem ser assumidos como os de uma ontologia, isto é, como entidades que nomeiam grupos de objetos, ou conjunto C da definição da seção 2.1. As instâncias são as concretizações das classes; são entidades únicas no domínio (conjunto I da seção 2.1).

O conceito “mãe” é uma especialização ou caso particular do conceito “pessoa”, dessa forma diz-se que “mãe” é uma hiponímia de “pessoa” e que “pessoa” é uma hiperonímia de “mãe” (Tabela 1).

Tabela 1: Exemplos de relações de hiperonímia

HIPERONÍMIA	HIPONÍMIA
Pessoa	Mãe
	Pai
	Filho

Segundo uma gramática disponível na web (Mundo Educação, 2010), “**Hiperônimo** é toda palavra que possui sentido amplo, geral; enquanto **Hipônimo** é justamente o contrário, é a palavra na sua especificidade.”

As taxonomias podem ser representadas de diferentes formas, dentre elas redes semânticas e predicados da lógica de primeira ordem.

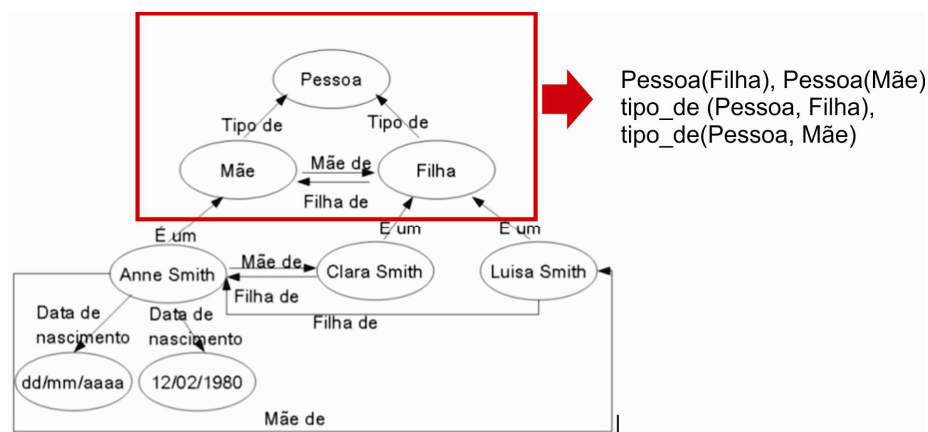


Figura 2: Taxonomias em redes semânticas e em lógica de primeira ordem

A Figura 2 ilustra uma ontologia apresentada na forma de redes semânticas, na qual um pequeno fragmento em que é mostrada uma taxonomia Pessoa/Mãe/Filha representa estas em lógica de primeira ordem.

2.2.2 Aprendizado de Máquina

É uma área de estudo que trata da elaboração de técnicas que possibilitem a construção de sistemas capazes de incrementar suas *expertises* de forma automática.

Tendo o acréscimo de conhecimento como propriedade, esses sistemas devem ter a capacidade de tomar novas decisões baseadas em experiências armazenadas (Monard; Baranauskas, 2005) para propor soluções mais “inteligentes”.

A aquisição do conhecimento pode se dar por meio de três mecanismos: o dedutivo, o indutivo e o analógico. No mecanismo dedutivo o conhecimento é adquirido a partir do que já existe, mas que estava implícito, enquanto que na indução um novo conhecimento é adquirido pela análise de exemplos, partindo-se da parte para o todo (generalização). O mecanismo analógico caracteriza-se pela apropriação de novo conhecimento a partir de uma analogia entre dois problemas e suas respectivas soluções.

Técnicas de aprendizado de máquina utilizam o mecanismo indutivo para atualização de sua base de conhecimento. Monard e Baranauskas (2005) definem o processo de indução como sendo uma “forma de inferência lógica que permite obter

conclusões genéricas sobre um conjunto particular de exemplos”. Com base no exposto, é válido afirmar que na aprendizagem de máquina um novo conceito é aprendido por meio de inferências indutivas sobre uma coleção de exemplos.

O aprendizado indutivo é classificado em supervisionado e não supervisionado. Há, ainda, na literatura uma terceira categoria: o aprendizado por reforço, que é classificado como uma técnica intermediária entre o aprendizado supervisionado e o não supervisionado. A Tabela 2 resume as principais características de cada um desses paradigmas.

Tabela 2: Classificação das técnicas de aprendizado de máquina

Tipo de Aprendizado	Caracterização	Aplicação
Supervisionado	Envolve a aprendizagem de uma função a partir de um conjunto de exemplos de suas entradas e saídas.	É tipicamente utilizado para prever categorias apropriadas de um exemplo a partir de um conjunto de categorias representadas.
Não-supervisionado	Envolve a aprendizagem de padrões de entrada, visto que não são fornecidos ao algoritmo de aprendizagem os valores de saída específicos.	É utilizado para encontrar aglomerados de conjuntos de dados semelhantes entre si (<i>clusters</i>).
por Reforço	A aprendizagem se dá a partir da observação das consequências das ações no ambiente.	Utilizado em tarefas de controle e robótica.

A Figura 3 ilustra a hierarquia do aprendizado segundo Monard e Baranauskas (2005).

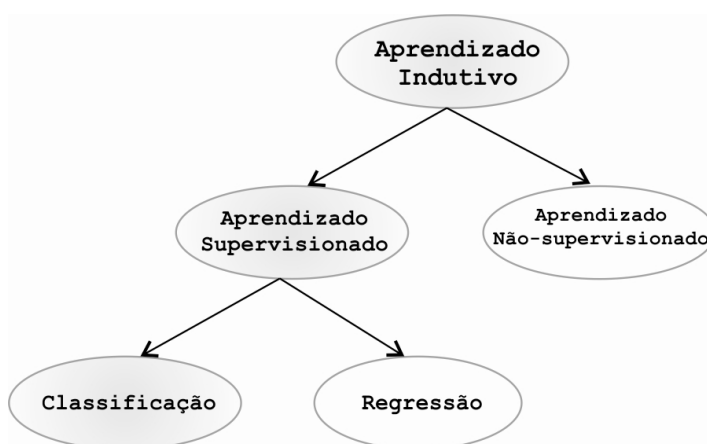


Figura 3: Hierarquia do aprendizado (Monard; Baranauskas, 2005)

Muitas técnicas para a aprendizagem de ontologias a partir de fontes textuais têm sido propostas com o uso de tarefas da aprendizagem de máquina. Na Seção 2.2.12 é apresentada uma tabela comparativa que ilustra técnicas que utilizam aprendizado de máquina para esse fim.

2.2.3 Processamento da Linguagem Natural

O Processamento da Linguagem Natural (PLN) é um subcampo da Inteligência Artificial que estuda os problemas de geração automatizada e compreensão de linguagens naturais humanas (Jielin, 2007).

A seguir são apresentados alguns tipos de aplicações, citados por Drake (2003), que usam PLN:

- Recuperação de informação (RI) – Sistemas de RI fornecem uma lista de documentos potencialmente relevantes em resposta a uma consulta do usuário.
- Extração de informação (EI) – É a mais recente área de aplicação, focalizando o reconhecimento, rotulação e extração de elementos de informação (pessoas, companhias, localizações, organizações, etc.) a partir de grandes coleções de textos.
- Pergunta-resposta – Sistemas de perguntas e respostas fornecem ao usuário apenas o texto da própria resposta ou passagens de respostas disponíveis.
- Resumo – Os altos níveis de PLN, particularmente o nível de discurso, podem ser utilizados na implementação de aplicações que produzem a partir de um texto, outro menor constituído de uma abreviada representação narrativa do documento original.
- Máquina de tradução – Talvez a mais antiga de todas as aplicações que usam PLN. Os vários níveis da PLN têm sido usados em tradutores, que vão desde as abordagens baseadas em palavras (nível léxico) até aplicações que incluem altos níveis de análise.
- Sistema de Diálogo – Talvez esse tipo de sistema seja a aplicação do futuro. Normalmente se concentram em tarefas estritamente definidas (por exemplo, sistemas de reconhecimento de voz usados em portais

de voz e agendas de telefones celulares) e usam predominantemente os níveis fonéticos e lexicais de processamento de linguagem.

O processamento da linguagem natural consiste de uma aplicação seqüencial de diferentes componentes de análise em uma arquitetura pipeline.

A definição de etapas ou fases de processamento da linguagem natural se baseia nos conhecimentos lingüísticos necessários a compreensão da linguagem natural tais com o fonético, léxico, morfológico, etc. Em geral o PLN é composto por etapas de tokenização, análise morfo-léxica, análise sintática, análise semântica e análise contextual.

Dependendo do autor, algumas dessas etapas são subdivididas, omitidas ou agrupadas em uma única fase. Além disso, podem ser incorporadas novas técnicas (etapas) de acordo com a saída que se pretende obter no final do processamento. Dale, Moisl e Somers (2000) definem cinco estágios do PLN: tokenização, análise léxica, análise sintática, análise semântica e análise pragmática.

Já para Cimiano (2006), o processamento da linguagem natural consiste nas fases de: pré-processamento, análise sintática, análise semântica e interpretação contextual. A fase de pré-processamento ainda tem seu próprio pipeline como mostra a Figura 4. O pipeline de pré-processamento é composto pela tokenização e normalização, POS tagging, análise morfológica, reconhecimento de entidades nomeadas e resolução de co-referências.

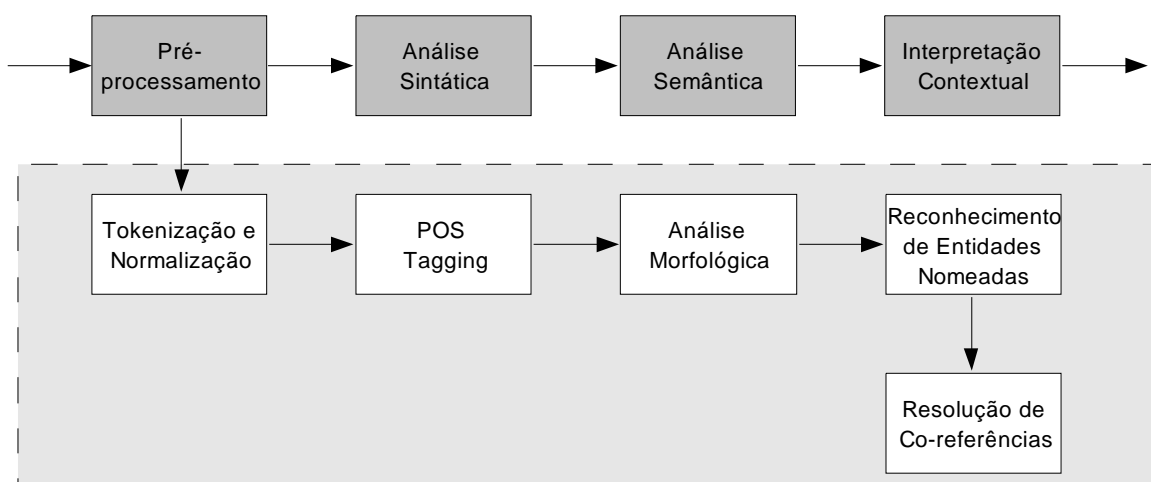


Figura 4: Pipeline do Processamento da Linguagem Natural (CIMIANO, 2006)

Uma grande parte das técnicas de aprendizagem de ontologias e em particular as de extração de taxonomias fazem uso de técnicas de PLN com o objetivo de estruturar o corpus (conjunto de documentos) para algum processamento subsequente. Veremos mais adiante que este trabalho não foge a esta tendência e que o PLN é substancialmente importante para o processo proposto no próximo capítulo.

2.2.4 Wordnet

O WordNet (FELLBAUM, 1998) é uma base de dados lexical organizada por significado. Ela foi desenvolvida na Universidade de Princeton por George A. Miller e categoriza conceitos e expressões em língua inglesa em quatro grupos: substantivo, verbo, adjetivo e advérbio. Os itens lexicais são apresentados através de suas diversas definições e suas relações com outros itens lexicais.

A sinonímia é o relacionamento semântico básico do WordNet, ele é expresso por synsets (conjunto de sinônimos). Através de synsets relacionados é formada uma hierarquia lexical entre eles pela hiponímia – relação entre um hiperônimo, mais genérico, e um hipônimo, mais específico.

```
{robin, redbreast } → {bird} → {animal, animate_being} → {organism, life_form,
living_thing}
({tordo, pisco-de-peito-ruivo} → {pássaro} → {animal, ser_animado} → {organismo,
forma-de-vida, ser-vivo})
```

Figura 5: Exemplo de Synset do WordNet

No exemplo da Figura 5, os elementos que estão entre chaves são os sinônimos que formam um synset e as setas apontam para os hiperônimos de cada synset.

O WordNet também representa relações parte/todo, denominadas relações de meronímia/holonímia. Por exemplo, bird(pássaro) é holônimo (o todo) de plumage(plumagem) e, conseqüentemente, plumage é merônimo (parte) de bird.

Sendo uma base lexical, o WordNet pode ser usado para apoiar etapas do processo de aprendizagem de ontologias. Por exemplo, a estrutura de synsets ajuda

na descoberta de sinônimos e as relações de hiponímia auxiliam na determinação de hierarquias de uma ontologia.

Segundo Lima et. al. (2007) a partir da WordNet de Princeton “têm sido geradas outras, para várias línguas, incluindo a do Português brasileiro, em elaboração e ainda indisponível”.

2.2.5 GATE

O GATE (General Architecture for Text Engineering – Arquitetura Genérica para Engenharia de Texto) é uma infra-estrutura para desenvolvimento e implantação de componentes de software que processam a linguagem natural (CUNNINGHAM, 2002). GATE foi desenvolvido em linguagem Java (Java, 2010) pela Universidade de Sheffield na Inglaterra.

Essa ferramenta guia os desenvolvedores de três formas:

- Especificando uma arquitetura genérica para o processamento da linguagem;
- Fornecendo um framework (conjunto de classes em Java) que implementa a arquitetura;
- Fornecendo um ambiente gráfico de desenvolvimento (Figura 10).

Arquitetura do GATE é baseada em Componentes conhecidos como Resources (recursos). Os recursos no GATE são diferenciados em três categorias:

- Recursos de Linguagem (*LanguageResources* – *LR*): são entidades lingüísticas, tais como documentos e corpora.
- Recursos de Processamento (*ProcessingResources* – *PR*): são entidades algorítmicas, efetuam alguma forma de processamento. Exemplo: analisadores (*parsers*), geradores, etc.
- Recursos Visuais (*VisualResources* – *VR*): são componentes da GUI que permitem a visualização ou edição de outros componentes.

O conjunto de recursos integrados ao GATE é chamado CREOLE (*Collection of Reusable Objects for Language Engineering* – *Coleção de Objetos de Engenharia da Linguagem Reusáveis*).

Todos os recursos são construídos baseados no modelo de Java Beans – “modelo de componente reutilizável para a linguagem Java e que pode ser manipulado por ferramentas gráficas de forma similar como é feito por outras ferramentas como Delphi e Visual Basic.” (MOSQUEIRA et. al., 2006)

Recursos de processamento podem ser combinados em aplicações. Aplicações controlam a execução dos PR através de pipeline. Há duas formas de pipeline:

- Pipeline Simples: consiste de um grupo de PR que serão executados seqüencialmente em um único documento.
- Pipeline em corpus: os PR são aplicados a cada um dos documentos do Corpus seqüencialmente.

Existem ainda versões condicionais desses pipelines, onde um PR pode ou não ser executado dependendo de alguma característica do documento.

O GATE suporta documentos nos formatos Plain Text, HTML, XML e RTF. Durante o processamento o GATE realiza anotações nos textos. Essas anotações constituem o resultado do processamento lingüístico realizado pelo framework.

Outro ponto importante é que o GATE possui um único modelo de informação que descreve documentos, coleção de documentos e anotações em documentos, baseado em pares de atributos/valores denominados de *features*. Nomes de atributos são *Strings* e valores são quaisquer objetos Java.

Todos os Recursos de Linguagem (LR) do GATE possuem um conjunto de features que caracteriza o recurso. Com base nisso, Cunningham (2009) apresenta os seguintes conceitos:

- Corpora: grupo de documentos + features
- Documento: texto + features + anotações
- Anotações: consistem de um campo de identificação, um que especifica o seu tipo, um que determina a posição de início do token, outro que determina a posição final e um conjunto de características (features) referentes a essa anotação.

Na Figura 6, as anotações mostradas são realizadas sobre a frase *Cyndi savored the soup (Cyndi saboreou a sopa)*. Por exemplo, a anotação com id 1 é do

tipo *token*, começa na posição 0 e termina na 5 e possui como *feature* o atributo POS (Part of Speech) com o valor *NP* (nome próprio).

Text				
Cyndi savored the soup.				
^0...^5...^10..^15..^20				
Annotations				
Id	Type	SpanStart	Span End	Features
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=DT
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	name_type=person
7	sentence	0	23	constituents=[1],[2],[3].[4],[5]

Figura 6: Exemplo de resultado das anotações realizado pelo GATE

A relação entre os recursos de processamento do GATE e as tarefas de PLN é mostrada na Tabela 3.

Tabela 3: Relação entre tarefas de PLN e Plugins do GATE

Tarefas de PLN	Recurso GATE
<i>Tokenização</i>	<i>ANNIE English Tokeniser</i>
<i>Normalização</i>	<i>ANNIE Gazetter</i>
<i>Divisão em Sentenças</i>	<i>Sentence Splitter</i>
<i>POS Tagging</i>	<i>ANNIE POS Tagger</i>
<i>Lematização</i>	<i>GATE Morphological Analyser</i>
<i>Stemming</i>	<i>Stemmer</i>
<i>Reconhecimento de Entidades Nomeadas</i>	<i>NE Transducer</i>
<i>Co-referência entre Entidades Nomeadas</i>	<i>ANNIE OrthoMatcher</i>
<i>Co-referencia pronominal</i>	<i>Pronominal Coreference</i>
<i>Chunking</i>	<i>Verb Group Chunker e Noun Phrase Chunker</i>
Parsing	<i>StanfordParser</i>

É ainda importante observar que o GATE possui toda uma biblioteca de classes escritas em Java com todos os recursos de processamento presentes na interface gráfica, o que facilita o desenvolvimento de novas aplicações de PLN independentes do ambiente gráfico. O GATE foi utilizado na elaboração da

ferramenta de software que dá suporte ao processo proposto para a aprendizagem de relações taxonômicas. O Capítulo 4 descreve esta ferramenta.

2.2.6 OWL

OWL (Web Ontology Language) é uma linguagem para definir e instanciar ontologias. Uma ontologia OWL pode incluir descrições de classes e suas respectivas propriedades e seus relacionamentos. Essa linguagem foi projetada para para uso por aplicações que precisam processar o conteúdo da informação ao invés de apenas apresentá-la aos humanos. Ela facilita a interpretação por máquinas do conteúdo da Web em relação a XML, RDF e RDF Schema por fornecer vocabulário adicional com uma semântica formal.

OWL é baseada nas linguagens OIL e DAML + OIL, sendo hoje uma recomendação da W3C. Ela vem ocupando um papel importante em um número cada vez maior de aplicações e vem sendo foco de pesquisa para ferramentas, técnicas de inferência e extensões de linguagens.

O processo para extração de taxonomias, objeto desse trabalho, tem como produto o conjunto H (da sêxtupla $O=(C,H,I,R,P,A)$) em formato OWL conforme descrito na seção 3.1.4. O Apêndice B mostra a ontologia da aplicação de um corpus do Direito de Família representada em formato OWL. Esta mesma ontologia pode ser vista em formato de redes semânticas no Apêndice A.

2.2.7 Abordagens para o aprendizado de relações taxonômicas

O aprendizado de relações taxonômicas consiste na utilização de técnicas como o aprendizado supervisionado, o aprendizado não-supervisionado, técnicas estatísticas, reconhecimento de padrões, etc. que permitem ao computador reconhecer hierarquias semi-automaticamente ou automaticamente. Essas abordagens permitem encontrar hierarquias que podem ser aplicadas a um domínio específico ou serem armazenadas em uma forma de representação de conhecimento, como por exemplo, uma ontologia.

Muitas abordagens podem ser encontradas em trabalhos científicos para a aquisição de relações taxonômicas. Esta seção visa a apresentação de trabalhos

relacionados com o tema desta dissertação. A maioria desses trabalhos pode ser encontrada em documentos científicos demonstrando como é possível extrair relacionamentos taxonômicos, ou simplesmente hierarquias, de fontes textuais.

As técnicas aqui descritas foram organizadas de acordo com características importantes tais como: utilização de Análise Formal de Conceitos, Métodos Estatísticos, Agrupamento (clustering) e Casamento de Padrões.

2.2.8 Aprendizagem de taxonomias usando Análise Formal de Conceitos

Cimiano (et. al., 2004) propôs uma abordagem baseada em Análise Formal de Conceitos (AFC), que é um método usado para análise de dados. Segundo o autor, esta técnica se baseia na identificação direta de conceitos utilizando uma hipótese distribucional com base no contexto de um determinado termo. O contexto é determinado por um vetor de dependências sintáticas adquiridas do texto com um *parser* linguístico. Segundo Mitchel (1980), um *parser* é capaz de analisar um texto de entrada e, com base em alguma gramática, determinar sua estrutura gramatical. Assim, a hipótese de distribuição utilizada por Cimiano assume que os termos são semelhantes à medida que eles se apresentam em contextos lingüísticos similares.

De modo geral, o Cimiano utiliza Análise Formal de Conceitos (AFC) para analisar o corpus com o intuito de descobrir relacionamento implícitos entre objetos (representados por um conjunto de atributos) e seus atributos. O autor extrai verbos e seus relacionamentos (por exemplo, o verbo e o sujeito de uma oração) e aplica a lematização a esse verbo (etapa Lematizer) e a um termo principal do relacionamento (chamado termo cabeça). É realizada a medida da frequência, sendo dado um peso para cada par (verbo/cabeça) com base estatística. Somente pares que tem o peso em um limite estabelecido pelo autor são selecionados em um contexto formal para a aplicação da AFC. O processo para extração de hierarquias de Cimiano é apresentado na Figura 7.

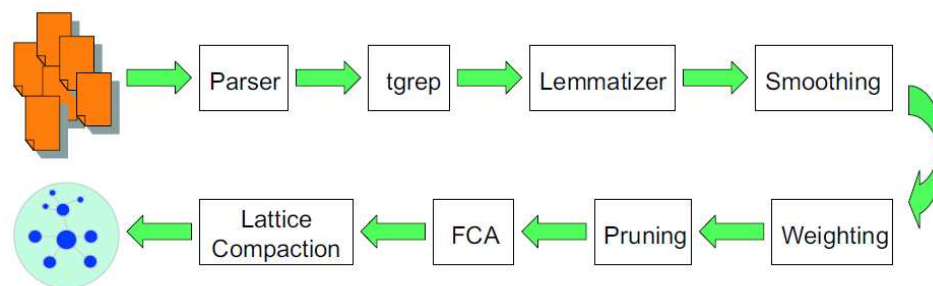


Figura 7: Processo geral para extração de hierarquias (Cimiano et al, 2005)

A primeira parte do processo é a identificação de classes gramaticais (Part-of-Speech) que é realizado com uma ferramenta chamada TreeTagger (TreeTagger, 2010). Então é feito o *parsing* (análise para identificação da estrutura gramatical) utilizando a ferramenta Lopar (Lopar, 2010). O resultado da primeira etapa é uma árvore de análise sintática. A fase subsequente é a extração de dependências do tipo verbo/sujeito, verbo/objeto e verbo/frases preposicionais das árvores criadas no passo anterior. O autor realiza a aplicação de uma técnica particular de nivelamento (*Smoothing*) para lidar com escassez de dados. A aplicação dessa técnica estima a frequência de alguns pares que não aparecem no texto estimando-os com base em outros pares para que sejam aplicados pesos (*Weighting*). Essa ponderação permite que somente pares que estejam em certo limiar sejam transformados em um conceito formal. Segundo o autor, essa “filtragem” permite alcançar um nível de hierarquia de conceitos.

A partir da etapa de atribuição de pesos (que realiza a seleção dos termos que estatisticamente são os melhores candidatos) os pares são sucessivamente refinados. Uma etapa de poda (Pruning) é realizada antes da AFC com o intuito de remover pares que não estejam em um limiar de pesos aceitáveis. Após a AFC (aplicação do aprendizado), os resultados são mais uma vez refinados com remoção de conceitos abstratos e então o resultado é a taxonomia de conceitos.

Cimiano (et al, 2004) cita que foram alcançadas as efetividades de 29,33% para o domínio do turismo e 33,11% para finanças.

2.2.9 Aprendizagem de taxonomias usando Métodos Estatísticos

Drumond (Drumond, 2009) apresentou um processo que utiliza redes lógicas de markov para identificação de classes e hierarquias de uma ontologia. Seu trabalho é aplicado a língua inglesa e utiliza uma técnica para extração de conceitos chamada de PRECE (Probabilistic Relational Concept Extraction) e uma para extração de hierarquias chamada de PREHE (Probabilistic Relational Hierarchy Extraction).

Sua técnica para extração de hierarquias (PREHE) visa extrair o conjunto H. O autor chamou esta tarefa como predição de links, ou seja, a própria identificação das relações entre entidades do tipo classe que forem extraídas de uma fonte textual.

Basicamente a técnica tenta determinar as mais prováveis relações entre dois conceitos a partir de evidências. Essas evidências são relacionamentos semânticos e relacionamentos sintáticos, ambos extraídos do corpus onde cada conceito foi encontrado.

Finalmente, o autor utiliza uma Rede Lógica de Markov (RLM) para identificar os relacionamentos mais prováveis do tipo *kind_of*. As evidências (Figura 8) que são extraídas do corpus são utilizadas em fórmulas (Figura 9) para determinação dos pesos para uma rede lógica de markov.

No arquivo de evidências, as dependências sintáticas entre os termos são representadas pelo predicado `Depends(term1,term2,dependency)`, enquanto que o predicado `Hypernym(term1,term2)` reflete a relação de hiperonímia entre os termos, já o predicado `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 *t1* e *t2* possuem o mesmo significado então é criado um conceito *c* e os átomos `F(c, t1)` e `F(c, t2)`. Por exemplo, "country", "nation" e "brazil" referem-se ao mesmo conceito.

A Figura 8 mostra as evidências usadas pela RLM da PREHE. As fórmulas desta mesma RLM (Figura 9) foram definidas manualmente com base em idéias extraídas de técnicas do estado da arte e algumas intuições sobre o problema e representadas em formato de lógica de primeira ordem.

<i>F(1,"country")</i>	<i>F(3,"brasilia")</i>
<i>F(1,"nation")</i>	<i>Depends("city","capital",NSUBJ)</i>
<i>F(1,"brazil")</i>	<i>Depends("capital","country",POSS)</i>
<i>F(2,"city")</i>	<i>Depends("capital","brazil",POSS)</i>
<i>F(2,"capital")</i>	<i>Depends("capital","brasilia",APPOS)</i>
<i>F(2,"brasilia")</i>	<i>Hypernym("city","capital")</i>
<i>F(3,"capital")</i>	

Figura 8: Evidências para identificação de hierarquias técnica PREHE (Drumond, 2009)

Por exemplo, 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. Esse é o caso de um conceito formado pelos termos “cidade” e “centro urbano”, que 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 Figura 10.

<i>Identificador</i>	<i>Fórmula</i>
1	$\neg Kind_of(c_1, c_2)$
2	$Kind_of(c_1, c_2) \Rightarrow \neg Kind_of(c_2, c_1)$
3	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Depends(t_3, t_1, +dep) \wedge Depends(t_3, t_2, +dep) \Rightarrow Kind_of(c_1, c_2)$
4	$F(c_1, t_1) \wedge F(c_2, t_2) \wedge Hypernym(t_1, t_2) \Rightarrow Kind_of(c_1, c_2)$
5	$F(c_1, t_1) \wedge F(c_2, t_1) \wedge F(c_1, t_2) \wedge \neg F(c_2, t_2) \Rightarrow Kind_of(c_1, c_2)$

Figura 9: Formulas determinação pesos na técnica PREHE (Drumond, 2009)

Já, a fórmula 3, expressa o fato de que certas dependências sintáticas entre realizações léxicas de dois conceitos pode indicar a existência de uma relação taxonômica entre eles. A notação +dep indica que um peso diferente será calculado para cada tipo de dependência sintática.

O produto da técnica PREHE é o conjunto de formas básicas do predicado *Kind_Of* consideradas verdadeiras no processo de inferência, o que corresponde ao conjunto H da definição de ontologia.

2.2.10 Aprendizagem de taxonomias usando Agrupamento

Essa seção descreve a ONTRETTEX (Ontology Taxonomical Relationships Extraction From Text) (Sarmiento, 2009), uma técnica para extração de relações taxonômicas a partir de fontes textuais baseada em algoritmo de agrupamento hierárquico. Sua aplicação foi dividida em três etapas, a saber: Estruturação dos Dados, Aplicação do Agrupamento e Identificação das Relações Taxonômicas (conforme apresentado na Figura 10).

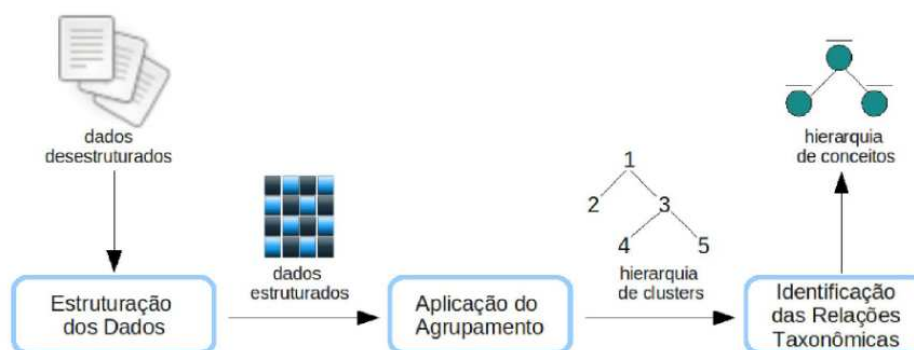


Figura 10: Etapas da técnica ONTRETTEX (Sarmiento, 2009)

A etapa “Estruturação dos Dados” é responsável por receber um documento em linguagem natural e convertê-lo em um documento estruturado. Para isso é utilizada uma ferramenta de software desenvolvida no Grupo de Pesquisa em Engenharia de Software e Engenharia de Conhecimento (GESEC, 2010) chamada PLSA Dumper, que é utilizada para transformar um arquivo semi-estruturado de texto em outros formatos, dentre eles o formato .arff, utilizado por uma ferramenta chamada WEKA (WEKA, 2009).

O autor utiliza TF-IDF (Frequência de Termos e a Frequência Inversa de Documentos) que é uma medida estatística que indica o quão relevante uma palavra é para um documento em um corpus e é utilizada para qualificar os atributos dos dados em sua técnica de agrupamento.

Após a estruturação, um arquivo em um formato específico é gerado para que possa ser utilizado na ferramenta WEKA (WEKA, 2009). Esse arquivo é composto por um cabeçalho que define o nome do documento (@RELATION seguido do nome), juntamente com identificação de cada atributo que representa um elemento no documento (@ATTRIBUTE seguido do nome do atributo e seu tipo) e,

finalmente, o corpo do arquivo que nada mais é do que uma lista de elementos (@DATA seguida por linhas que representam o nome do elemento e seus respectivos valores para os atributos do cabeçalho), como ilustrado na Figura 11. Por exemplo, o termo “*datum*” da Figura 11 é seguido por números que representam pesos separados por vírgulas para os documentos indicados no cabeçalho (DOC1, DOC2, DOC3,...).

```

@RELATION corpus_model

@ATTRIBUTE token {datum,engineering,chemistry,nature,material,webmaster,technology}
@ATTRIBUTE DOC1 REAL
@ATTRIBUTE DOC2 REAL
@ATTRIBUTE DOC3 REAL
@ATTRIBUTE DOC4 REAL
@ATTRIBUTE DOC5 REAL
@ATTRIBUTE type {term}

@DATA
datum,0.0,0.0,0.0,2.80,0.0,term
engineering,0.0,1.58,0.0,0.0,0.0,term
chemistry,0.0,0.0,0.0,0.0,2.32,term
nature,1.04,0.0,0.0,0.0,0.0,term
material,0.0,1.0,0.0,1.0,0.0,term
webmaster,1.0,0.0,0.0,0.0,2.03,term
technology,0.0,1.02,1.83,0.0,0.0,term

```

Figura 11: Arquivo .arff usado pela ONTRETEx (Sarmento, 2009)

O arquivo da Figura 11 é a entrada utilizada pelo autor para um algoritmo chamado Cobweb, que é uma implementação de aprendizado de máquina usando agrupamento e está disposto na API do WEKA. Esse algoritmo realiza a hierarquização de termos encontrados no corpus (organizados no arquivo .arff). Um refinamento subsequente dessas hierarquias encontradas é feita com a API do Wordnet (Wordnet, 2010) (de acordo com a Figura 12).

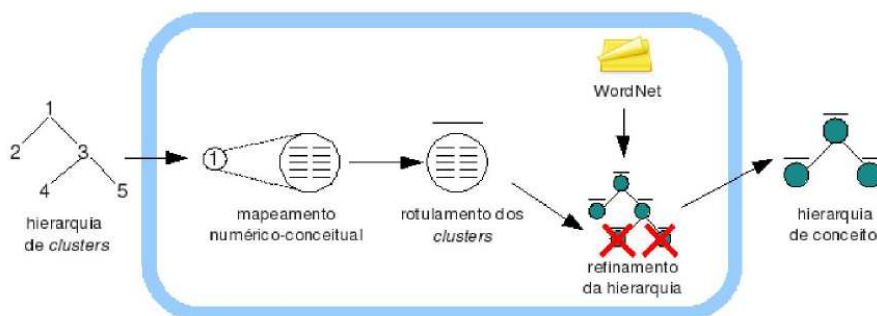


Figura 12: Determinação da hierarquia de conceitos da ONTRETEx (Sarmento, 2009)

O Cobweb utiliza algumas funções que são responsáveis por organizar os termos em grupos e eleger um mais representativo para cada grupo. Uma vez eleito o mais representativo, os outros termos são analisados e os que não apresentam uma relação de subconceito com o mais representativo são descartados. Para a seleção dos conceitos aos grupos, o algoritmo implementa uma função de utilidade.

Outra aplicação de agrupamento foi desenvolvida por Caraballo (1999) que descreveu uma forma de aprendizado baseada em agrupamento de nomes. Seus experimentos se basearam primeiramente no agrupamento de nomes de um corpus de dados do Jornal de Wall Street (Wall Street Journal). Uma vez realizado o agrupamento de nomes, é feita a identificação daquilo que o autor chamou de termos “*appositives*”, que nada mais são do que duas ou mais palavras juntas em texto que se referem a uma mesma entidade (exemplo: Roberto Justus, um grande empresário). Também são extraídos fragmentos do texto como, por exemplo, “vice-presidente e tesoureiro” ou “equipamentos científicos, aparatos e dispositivos”, que tem como característica a idéia de enumeração (normalmente a presença vírgulas e conjunções), considerando que estes estão semanticamente relacionados e que também podem ser chamados de “frases nominais”.

De cada frase nominal é selecionado um termo principal. Esses termos juntamente com a contagem de quantas vezes o mesmo ocorre nos fragmentos selecionados, são organizados em vetores. Caraballo assumiu a similaridade entre dois vetores de termos pela fórmula presente na Figura 13 e a similaridade entre dois grupos de termos conforme a Figura 14.

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| |\mathbf{w}|}$$

Figura 13: Cálculo da similaridade entre termos (Caraballo, 1999)

$$sim(A, B) = \frac{\sum_{\mathbf{v}, \mathbf{w}} \cos(\mathbf{v}, \mathbf{w})}{size(A) size(B)}$$

Figura 14: Cálculo da similaridade entre dois grupos de termos (Caraballo, 1999)

A similaridade entre grupos de termos é razão do cosseno entre cada par de nomes sobre cada nome. As variáveis \mathbf{v} e \mathbf{w} representam respectivamente todos os vetores de um grupo A e B.

O objetivo é fazer uma árvore com todos os termos utilizando técnicas *bottom-up clustering* (*clustering* aglomerativo) colocando cada termo em um nó da árvore.

A similaridade é calculada entre cada par de nós usando a fórmula do cosseno (Figura 13) e os nós mais similares são colocados sob o mesmo nó pai. Quando os dois ou mais nós são colocados sob um mesmo pai, a nova similaridade entre esse (*i*) e um nó pai *C* é calculada com uma fórmula de similaridade ponderada (Figura 15) e esse procedimento é repetido até que todos os termos sejam associados a um ancestral.

$$sim(C, i) = \frac{sim(A, i)size(A) + sim(B, i)size(B)}{size(A) + size(B)}$$

Figura 15: Cálculo da similaridade ponderada (Caraballo, 1999)

Na prática e na época, essa técnica exigiu um poder de processamento muito alto e uma grande quantidade de memória. O próprio autor considerou que para uma análise de 50.000 termos, uma matriz era gerada na memória com dimensão 50.000 x 50.0000, dado que cada elemento é comparado com todos os outros. A sugestão dada para contornar o problema foi o processamento em lotes para diminuir a quantidade de recursos de uma só vez para o processamento.

2.2.11 Aprendizagem de taxonomias usando Casamento de Padrões

A utilização de padrões léxico-sintáticos é datada de um bom tempo, pelo menos, com a introdução dos padrões de Hearst (1992). Hearst se propõe a realizar a expansão de dicionários e bases lexicais com a obtenção de relações de hiponímia/hiperonímia baseada em padrões, que foram encontrados com a simples observação de fragmentos textuais.

Martir Hearst foi a primeira a utilizar esses padrões (Freitas, 2007), que servem de base e de comparação, ainda, para muitos dos trabalhos que realizam identificação de relações taxonômicas.

Os padrões de Hearst são apresentados a seguir e enumerados por identificadores romanos i, ii, iii, iv, v e vi:

- (i) NP0 such as NP1 {, NP2 ... , (and | or) NPi}
- (ii) such NP0 as {NP ,}* {(and | or)} NP
- (iii) NP {, NP}* {,} or other NP0
- (iv) NP {, NP}* {,} and other NP0
- (v) NP0 {,} including { NP ,}* {or | and} NP
- (vi) NP0 {,} especially { NP ,}* {or | and} NP

Onde os termos representados por NP0 são considerados hipônimos dos demais termos (representados por NP).

Segundo Hearst, os padrões de (i) até o (iii) foram encontrados com base na observação, porém, um procedimento padrão pode ser utilizado para a descoberta de novos padrões (Freitas, 2007):

- Decidir qual a relação lexical de interesse;
- Derivar, por meio da WordNet (Wordnet, 2010), uma lista de pares de palavra na qual a relação esteja expressa: por exemplo, para a relação de meronímia, o par carro-volante;
- Extrair sentenças do corpus em que ambas as palavras (carro e volante) apareçam, registrando o contexto lexical e sintático em que foram encontradas;
- Encontrar semelhanças entre esses contextos e tentar generalizar: contextos comuns levam a padrões que indicam a relação de interesse.

Um exemplo do casamento dos padrões de Hearst pode ser visto no, exemplo (iv) como:

Temples, treasuries, and other important civic buildings.

A abordagem possui limitações (por exemplo, nem sempre é possível achar quantidade significativa de padrões em textos pequenos), mas tem bons resultados. Segundo (Freitas, 2007):

“Consciente do alto grau de subjetividade deste tipo de avaliação, e assumindo uma abordagem “cautelosa” na avaliação, 63% das relações extraídas foram consideradas corretas, isto é, passíveis de serem inseridas na Wordnet”

(Freitas, 2007, p. 60):

Outras abordagens exemplificam o uso de padrões para a identificação de relações presentes em um texto como (Sundblad, 2002) e (Ting Wang et al, 2006). Este último apresenta uma abordagem para aquisição de relações taxonômicas que se baseia na utilização de Máquinas de Vetores de Suporte (SVMs, do Inglês Support Vector Machines) que constituem uma técnica de aprendizado de máquina. O artigo de Ting Wang apresenta o programa ACE (*Automatic Content Extraction*) em sua versão 2004.

As SVMs são utilizadas com o intuito de realizar a identificação do tipo e do subtipo das relações entre cada par de entidades encontradas em uma mesma sentença de um conjunto de tipos predefinidos. Por exemplo, em “*Texas has many cars*” o autor indica que “Texas” e “many cars” possuem, segundo sua tabela de tipos, uma relação (Agente/artefato) do subtipo de (Usuário/proprietário).

O PLN é utilizado por Ting Wang (2006) com o intuito de realizar a identificação de evidências para o aprendizado de máquina. É citado pelo autor que para maioria das técnicas de aprendizado de máquina é muito importante a escolha de características, e que esta etapa é realizada cuidadosamente para um bom resultado (a exemplo de árvores de decisão e aprendizado de regras), mas para SVMs isso não é tão importante porque, características irrelevantes para um problema binário normalmente são distribuídos uniformemente no treinamento positivo e negativo e, portanto, teria pouca contribuição no mecanismo de aprendizagem. Esta deve ter sido uma das vantagens que motivaram o uso desta técnica pelo autor.

2.2.12 Resumo das técnicas

As técnicas de extração de hierarquias apresentadas anteriormente são comparadas na Tabela 4 com um pequeno resumo de suas propriedades. Isto pode ajudar a entender as diferenças entre elas, bem como suas limitações.

Tabela 4: tabela comparativa das principais técnicas de aprendizado de relações taxômicas

Abordagem/Ano	Principais Tecnologias	Principais Ferramentas	Efetividade (<i>Precision</i>) / Domínio	Nível de Automação
Cimiano et al., 2004	AFC PLN	TreeTagger Lopar	29,33% / Turismo 33,11% / Finanças	Automático
Drumond, 2009	PLN MLN Dicionário PREHE	GATE Wordnet	17% / Turismo	Semi-Automático
Caraballo, 1999	AM-Clustering	-	39% / Wall Street Journal	Automático
Sarmento, 2009	AM - Agrupamento NLP Dicionário	WEKA GATE Wordnet	n/a / Turismo	Semi-Automático
Hearst, 1992	Padrões Léxico Sintáticos	-	63%	Automático
Ting Wang et al. 2006	Padrões Léxico Sintáticos SVM PLN Dicionário	Wordnet GATE	73,87% /-	Automático
Correia, 2011	PLN Padrões Léxico Sintáticos Dicionário	GATE Wordnet	81,57% / Direito de Família	Automático

As principais técnicas usadas (segunda coluna da Tabela 4) são baseadas em padrões léxicos sintáticos (Ting Wang et al, 2006) (Hearst, 1992), métodos estatísticos com Redes Lógicas de Markov (RLM) (Drumond, 2009) e Técnicas de Aprendizado de Máquina (AM) (Cimiano; Hotho; Staab, 2004) (Caraballo, 1999) (Sarmento, 2009). Os trabalhos mais recentes têm em comum a aplicação de técnicas de Processamento da Linguagem Natural (PLN) em associação com outras técnicas.

As abordagens descritas em (Ting Wang et al, 2006), (Sarmento, 2009), (Drumond, 2009) e o presente trabalho usam o Wordnet como uma base lexical para a extração de hipônimos e o GATE (GATE, 2010) para realizar o PLN. A ferramenta Weka (Weka, 2010) foi utilizada como um framework para a implementação de um algoritmo de *clustering* usado por (Sarmento, 2009), enquanto as ferramentas TreeTagger e LoPar foram usadas por (Cimiano; Hotho; Staab, 2004), respectivamente, para realizar a identificação de classes gramaticais e como *parser*.

A efetividade é exibida em termos da medida *precision* (Dellschaft e Staab, 2006) obtida por cada abordagem em seus experimentos. Contudo, estes números não devem ser considerados para fins comparativos, pois, cada avaliação foi realizada com diferentes entradas e/ou comparadas com diferentes ontologias.

As abordagens (Cimiano; Hotho; Staab, 2004), (Caraballo, 1999) e (Hearst, 1992) apresentam soluções automáticas para a aquisição de relações taxômicas, enquanto as abordagens (Drumond, 2009) e (Sarmiento, 2009) propõem soluções semi-automáticas.

As abordagens são bem distintas e mesmo as que obtêm relações taxonômicas, apresentam representações de relações taxonômicas diferentes. No proposto neste trabalho, são consideradas relações taxonômicas as mesmas relações encontradas nos trabalhos de Sarmiento (2009), Drumond (2009) e Macedo (2010) que são as relações formalizadas no conceito de ontologias da Seção 2.1. Contudo, Ting Wang representa relações taxonômicas como relações *é_um*, que neste contexto são relações de instanciação (conjunto I). Outra consideração importante é que as medidas de efetividade são aferidas de formas diferentes e são exibidas nas tabelas apenas para fins informativos, não se tratando de uma forma de comparação fiel para experimentos de mesmo aspecto e condições.

2.3 Considerações Finais

Este capítulo apresentou o conceito de relações taxonômicas e uma revisão dos conceitos de aprendizagem máquina e Processamento da Linguagem Natural. Foi apresentado também o estado da arte da aquisição de relações taxonômicas com algumas abordagens atuais. A seção “Abordagens para o aprendizado de relações taxonômicas” apresentou várias técnicas que permitem a aquisição de relações taxonômicas apresentada de forma resumida em uma tabela comparativa.

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) (Ting Wang, 1992) e estatística (Drumond, 2009) fazendo uso de tarefas da aprendizagem de máquina como o agrupamento (Cimiano; Hotho; Staab, 2004) (Caraballo, 1999) (Sarmiento, 2009).

3 Um processo para a aquisição de relações taxonômicas de uma ontologia

Este capítulo propõe um processo para a aquisição automática de relações taxonômicas de uma ontologia baseado na aplicação de técnicas de Processamento da Linguagem Natural e Padrões Heurísticos, mostrando é possível unir a utilização de Padrões Léxico-Sintáticos (como os descritos por Hearst em 1992) com o Processamento da Linguagem Natural em um processo automático para o aprendizado de relações taxonômicas de uma ontologia. Está organizado como segue: A Seção 3.1 descreve o processo proposto e detalha as suas fases e a Seção 3.2 apresenta as considerações finais do capítulo.

3.1 O processo

O processo proposto é apresentado na Figura 16 e consiste de quatro fases: “Marcação”, “Extração de Classes Candidatos”, “Identificação de Hiperônimos e Sinônimos” e “Identificação e Representação de Relações Taxonômicas”.

A entrada para o processo de aquisição de relações taxonômicas é um corpus contendo um conjunto de documentos textuais em um determinado domínio de aplicação. A fase “Marcação” tem por objetivo a identificação dos tokens, sentenças, classes gramaticais e lemas com a aplicação de técnicas de PLN. Em seguida, a atividade “Extração de Classes Candidatas” é responsável por separar os tokens que provavelmente serão as classes na hierarquia de uma ontologia. A fase “Identificação de Sinônimos e Hiperônimos” identifica os sinônimos e hiperônimos dos termos que representam as classes obtidas da fase anterior. A última fase do processo é a “Identificação e Representação de Relações Taxonômicas” com a utilização de heurísticas para identificação de padrões lingüísticos. Como produto obtém-se a representação das relações taxonômicas encontradas em uma linguagem de especificação de ontologias.

Nas próximas subseções são apresentadas detalhadamente as transformações realizadas em cada fase do processo.

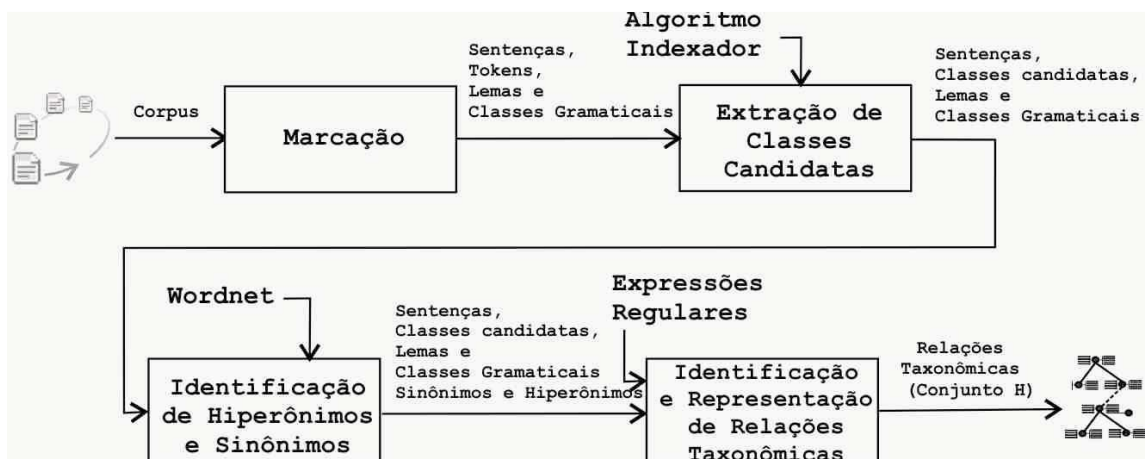


Figura 16: Um processo para a aquisição de relações taxonômicas de uma ontologia

3.1.1 Marcação

A fase “Marcação” tem como entrada um corpus e visa a sua transformação em uma representação, um modelo, capaz de ser processado computacionalmente. Isso permite que uma fonte desestruturada escrita em linguagem natural possa ser interpretada adequadamente. Esta fase consiste das seguintes atividades: “Tokenização”, “Divisão em Sentenças”, “Lematização” e “Análise Lexical” (como ilustrado na Figura 17).

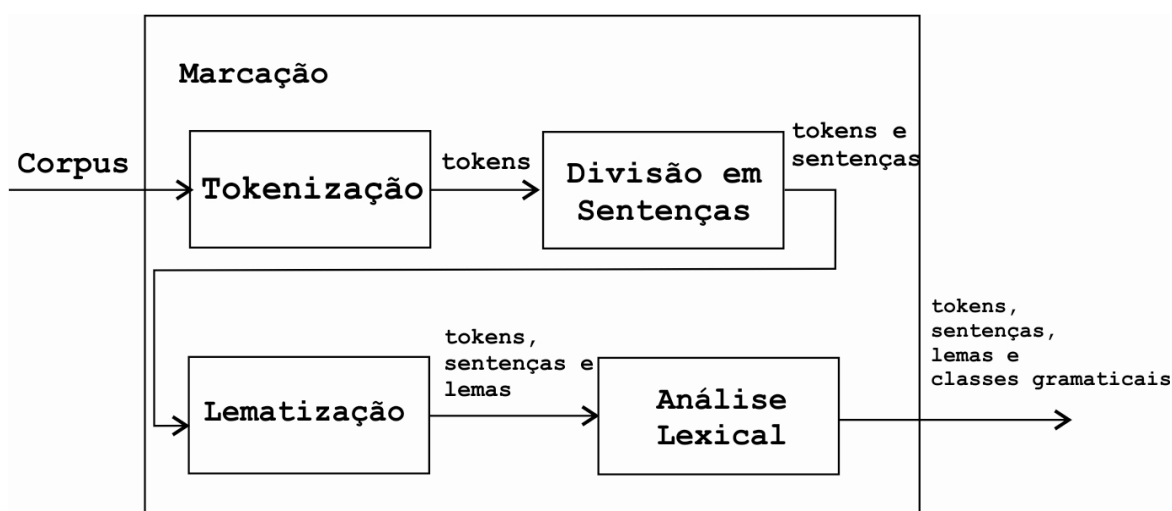


Figura 17: Atividades da fase “Marcação”

A atividade “Tokenização” busca identificar os tokens existentes no documento. Os tokens podem, por exemplo, ser palavras, pontuações, espaços em branco, etc., mas somente palavras são selecionadas. A atividade “Divisão em

Sentenças” organiza os tokens identificados agrupando-os de acordo com a sentença onde estão localizados. A atividade “Lematização” realiza a redução de cada token a sua forma básica, por exemplo, a sentença “*Eu conheci os filhos do meu amigo*” se tornaria, após o processo de lematização, “*Eu conhecer o filho do meu amigo*”. Essa forma base de cada palavra generaliza as várias formas flexionadas de um token tornando possível o agrupamento dos tokens em um único lema.

A atividade “Análise lexical”, visa a identificação das classes gramaticais de cada um dos tokens selecionados na atividade “Tokenização”. A identificação de classes gramaticais através da análise lexical também é conhecida como Part-of-Speech (POS) Tagging. Uma lista das classes gramaticais passíveis de serem encontradas através da análise lexical é apresentada no Anexo A (Marcus; Santorini; Marcinkiewicz, 1993).

A aplicação das atividades da fase de marcação resulta em texto marcado. Por exemplo, para o fragmento de texto “*Parents like my father or my mother are amazing. All children love dolls, masks and other toys*” após ser submetido à marcação terá a identificação de partes do texto conforme apresentado na Figura 18.

	term_id [PK] serial	term character var	num_doc integer	sentence_id integer	pos character var	t_start integer	t_end integer	gate_id integer
1	125752	parent	1	8507	NNS	0	7	1
2	125753	like	1	8507	IN	8	12	3
3	125754	my	1	8507	PRP\$	13	15	5
4	125755	father	1	8507	NN	16	22	7
5	125756	or	1	8507	CC	23	25	9
6	125757	my	1	8507	PRP\$	26	28	11
7	125758	mother	1	8507	NN	29	35	13
8	125759	be	1	8507	VBP	36	39	15
9	125760	amazing	1	8507	JJ	40	47	17
10	125761	.	1	8507	.	47	48	18
11	125762	all	1	8508	DT	49	52	20
12	125763	child	1	8508	NNS	53	61	22
13	125764	love	1	8508	VBP	62	66	24
14	125765	,	1	8508	,	72	73	27
15	125766	doll	1	8508	NNS	67	72	26
16	125767	mask	1	8508	NNS	74	79	29
17	125768	and	1	8508	CC	80	83	31
18	125769	toy	1	8508	NNS	90	94	35
19	125770	other	1	8508	JJ	84	89	33

	sentence_id [PK] integer	sentence text	doc_num integer	t_start integer	t_end integer	gate_id integer
1	8507	Parents like my father or my mother are amazing.	1	0	48	38
2	8508	All children love dolls, masks and other toys	1	49	94	39

Figura 18: Exemplo de um texto marcado

A Figura 19 ilustra o fragmento de texto usado na Figura 18 passando pelas atividades da fase “Marcação” (desprezando a pontuação). A esquerda é apresentada cada atividade e a direita da figura o resultado da aplicação da atividade. O primeiro quadro mostra os termos que foram marcados como tokens: “parents”, “like”, “my”, “father”, “or”, “my”, “mother”, “are”, “amazing”, “all”, “children”, “love”, “dolls”, “masks”, “and”, “other” e “toys”. Pontos e espaços também são tokens, porém, para fins ilustrativos, somente os termos foram mostrados. No segundo quadro é mostrado o resultado da divisão em sentenças: “Parents like my father or my mother are amazing.” e “All children love dolls, masks and other toys”. Em seguida é mostrado o resultado da “Lematização” realizada sobre cada termo. Por exemplo, o termo “children” possui o lema “child”. O último quadro que corresponde a Análise Lexical, mostra cada token e sua respectiva classe gramatical. Por exemplo, a marcação NNS (Marcus, Santorini, Marcinkiewicz, 1993) para o termo “parents” indica que este é um substantivo no plural.

Tokenização	Parents like my father or my mother are amazing. All children love dolls, masks and other toys
Divisão em Sentenças	Parents like my father or my mother are amazing. All children love dolls, masks and other toys <small>↳ sentence</small> <small>↳ sentence</small>
Lematização	parent like my father or my mother are amazing. All children love dolls, masks and other toys <small>↳ sentence</small> <small>↳ sentence</small>
Análise Lexical	NNS IN PRP-NN CC PRP-NN VBP-JJ DT NNS VBP NNS NNS CC JJ NNS parent like my father or my mother are amazing. All children love dolls, masks and other toys <small>↳ sentence</small> <small>↳ sentence</small>

Figura 19: Exemplo de aplicação da “Marcação”

O produto desta fase é o corpus marcado com os tokens, as sentenças, os lemas e as classes gramaticais, com mostra, por exemplo, a Figura 18.

3.1.2 Extração de Classes Candidatas

A “Extração de Classes Candidatas” tem como objetivo a seleção de substantivos concretos e abstratos que serão as classes candidatas pertencentes ao conjunto C_c (conjunto de classes da seção 2.1) de uma ontologia.

A hipótese assumida neste trabalho é que somente substantivos podem ser considerados classes. Desta forma, conceitualmente falando, podemos encontrar substantivos de dois tipos: substantivos concretos e substantivos abstratos. Os substantivos concretos podem ser próprios ou comuns (GrammarNet, 2011). Os substantivos concretos próprios servem para representar nomes de pessoas (antropônimos) ou nomes de lugares (topônimos), o que essencialmente os caracteriza como a concretização de classes (instâncias). Portanto, na extração de tokens candidatos a classes são desconsiderados os substantivos próprios e selecionados apenas os substantivos concretos comuns e os substantivos abstratos.

O produto desta fase são as classes candidatas, as sentenças onde as mesmas se encontram, os lemas para essas classes e suas respectivas classes gramaticais.

Por exemplo, o resultado da execução desta fase aplicada a saída da fase anterior (Figura 19) consiste das classes candidatas “parent”, “father”, “mother”, “child”, “doll”, “mask” e “toy”, conforme podemos ver na Figura 20.

term_id [PK] serial	term character vai	num_doc integer	sentence_id integer	pos character vai	t_start integer	t_end integer	gate_id integer
125771	parent	1	8509	NNS	0	7	1
125772	father	1	8509	NN	16	22	7
125773	mother	1	8509	NN	29	35	13
125774	child	1	8510	NNS	53	61	22
125775	doll	1	8510	NNS	67	72	26
125776	mask	1	8510	NNS	74	79	29
125777	toy	1	8510	NNS	90	94	35

Figura 20: Classes candidatas

3.1.3 Identificação de Hiperônimos e Sinônimos

A “Identificação de Hiperônimos e Sinônimos” tem como objetivo a identificação de sinônimos e hiperônimos no Wordnet (Wordnet, 2010), uma base de

dados que contêm termos em linguagem natural, suas respectivas definições e seus relacionamentos semânticos como a sinonímia, hiperonímia e hiponímia. Cada classe selecionada na fase anterior é representada por um termo (por exemplo, a classe “Mother”). Esses termos são localizados no Wordnet e quando é encontrado um sinônimo ou hiperônimo presente no próprio texto essa ocorrência é armazenada. Ou seja, cada classe será associada a dois vetores: um de sinônimos e um de hiperônimos presentes no texto. Por exemplo, para as classes candidatas “parent”, “father”, “mother”, “child”, “doll”, “mask” e “toy”, através da execução desta fase, os seguintes vetores de sinônimos (Figura 21) e de hiperônimos (Figura 22) são identificados:

term character varying	synonym character varying
father	male parent
father	begetter
mother	female parent
child	kid
child	youngster
child	minor
child	shaver
child	nipper
child	small fry
child	tiddler
child	tike
child	tyke
child	fry
child	nestling
doll	dolly
toy	plaything

Figura 21: Vetores de Sinônimos para “father”, “mother”, “child”, “doll” e “toy”

term character varying	hypernym character varying
parent	genitor
father	parent
mother	parent
child	juvenile person
child	juvenile
doll	toy
doll	plaything
mask	disguise
toy	artefact
toy	artifact

Figura 22: Vetores de Hiperônimos para “parent”, “father”, “mother”, “child”, “doll”, “mask” e “toy”

O produto desta fase é o mesmo produto da fase anterior acrescido das relações de sinonímia e hiperonímia obtidas do Wordnet.

3.1.4 Identificação e Representação de Relações Taxonômicas

A fase “Identificação e Representação de Relações Taxonômicas” tem como objetivo descobrir relações taxonômicas através da aplicação de padrões heurísticos e representá-las em uma linguagem de especificação de ontologias. Cada sentença é analisada com o objetivo de identificar e realizar o casamento de padrões.

Para testar a existência de relações de hiperonímia são aplicadas expressões regulares sobre as sentenças com padrões heurísticos como, por exemplo, os padrões de Hearst (Hearst, 1992) (Figura 23 e Figura 24). As expressões regulares são consideradas um método formal para especificar um padrão de texto (Jargas, 2008). Por exemplo, aplicando a expressão regular padrão $([A-Za-z]+,.)+[A-Za-z]+ \text{ and other } [A-Za-z]+$ ao texto “*All children love dolls, masks and other toys*” resultando em “*dolls, masks and other toys*”. As classes “Doll” e “Mask” estão relacionadas com a classe “Toy” por uma relação de hiponímia. Para o caso apresentado, isto significa que “Toy” é uma classe superior as classes “Doll” and “Mask” na hierarquia de uma ontologia. Esse casamento representa o padrão (iv) da tabela de padrões heurísticos de Hearst (Figura 24), onde NP0 representa um termo hierarquicamente superior a os demais termos (NP1, NP2,...):

```
(i) NP0 such as NP1 {, NP2 ... , (and | or) NPi}
(ii) such NP0 as {NP ,}* {(and | or)} NP
(iii) NP {, NP}* {,} or other NP0
(iv) NP {, NP}* {,} and other NP0
(v) NP0 {,} including { NP ,}* {or | and} NP
(vi) NP0 {,} especially { NP ,}* {or | and} NP
```

Figura 23: Padrões Heurísticos de Hearst (Hearst, 1992)

- (i) NP is (a | an) NP0
- (ii) NP was (a | an) NP0
- (iii) NP0 was thought of NP
- (iv) which defined NP0 as NP
- (v) what defines NP0 as NP
- (vi) NP have been recognized as NP0
- (vii) NP was recognized as NP0

Figura 24: Outros padrões heurísticos para identificação de hierarquias

A identificação de sinônimos permite tratar termos que tenham o mesmo significado com as mesmas relações de herança, enquanto as relações hiperonímia, encontrados na fase anterior, são organizadas juntamente com as relações encontradas com o casamento de padrões, resultando em um conjunto de relações de heranças que são representadas em uma linguagem de especificação de ontologias.

O produto desta fase são as relações taxonômicas, conjunto H da definição de ontologia, representadas em uma linguagem de especificação de ontologia. Por exemplo, para as classes “parent”, “father”, “mother”, “child”, “doll”, “mask” e “toy” as relações de herança seriam encontradas conforme a Figura 25.

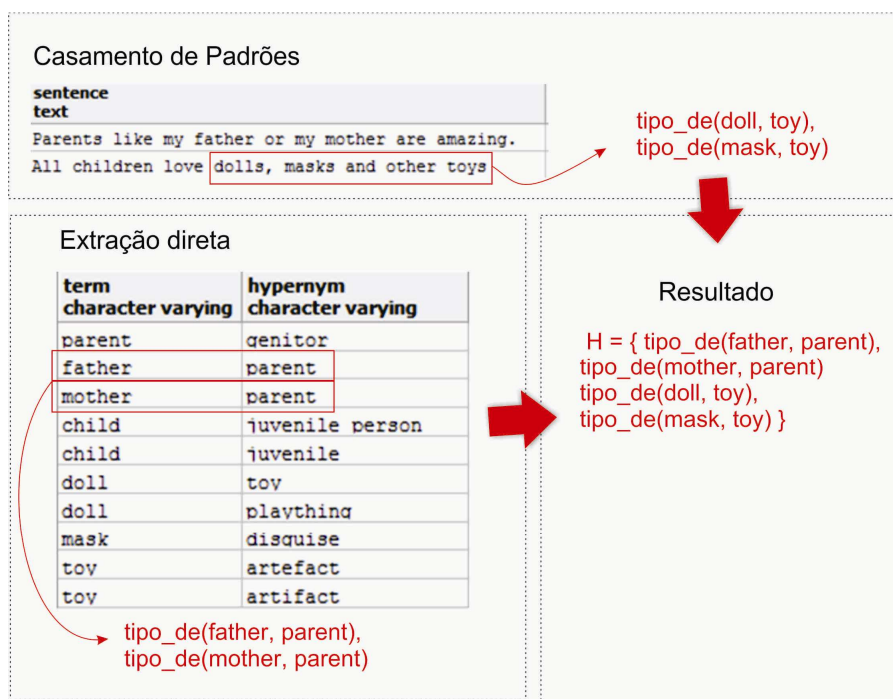


Figura 25: Identificação de relações taxonômicas

3.2 Considerações Finais

Este capítulo apresentou um processo para a aquisição de relações taxonômicas de uma ontologia. A primeira fase do processo apresentado, fase “Marcação”, utiliza o processamento de linguagem natural para transformar o texto original em uma representação viável para o processamento em computador. As etapas seguintes utilizam as informações da primeira fase para selecionar os termos (tokens) que melhor se candidatam a classes (nós das hierarquias) e, identificar os sinônimos e hipônimos desses termos com a ajuda da base lexical (Wordnet). São utilizados padrões heurísticos, escritos na forma de expressões regulares, para identificar relações taxonômicas presentes no texto. Essas hierarquias encontradas com padrões são somadas às que são encontradas com as extrações diretas da base lexical (Wordnet), perfazendo o conjunto H da definição de ontologia. Finalmente, o conjunto H é representado em uma linguagem de especificação de ontologias.

Este processo é passível de automatização quando considerado que os padrões escritos em expressões regulares são uma entrada fixada antes da execução do processo. Uma vez que esses padrões são genéricos e relacionam classes que sempre são oriundas da fonte de entrada, É possível assumir que o processo é independente de domínio, desde que a fonte de entrada (corpus) seja escrita em língua inglesa.

4 Avaliação do processo proposto

Este capítulo apresenta uma avaliação preliminar do processo proposto. Para tanto, foi desenvolvida uma ferramenta para a aquisição de relações taxonômicas de uma ontologia baseada na aplicação do processo proposto no Capítulo 3. O capítulo está organizado da seguinte maneira. A Seção 4.1 faz uma breve apresentação da ferramenta e a descreve, detalhando a plataforma e o seu funcionamento. A Seção 4.2 apresenta o estudo de caso conduzido para fins de avaliação do processo e, finalmente, a Seção 4.3 apresenta as considerações finais do capítulo.

4.1 Uma ferramenta para aquisição de relações taxonômicas de uma ontologia

Para automatizar a execução das fases descritas no capítulo anterior, que explana como podemos fazer a aquisição automática de relações taxonômicas, foi desenvolvida uma ferramenta que recebe um corpus escrito em linguagem natural em formato texto (.txt) e que gera a hierarquia de uma ontologia em um arquivo de extensão .owl, o qual armazena as classes ligadas hierarquicamente por relações de herança (relações “tipo_de”, ou conjunto H da definição de ontologia). O projeto foi chamado de Taxonomy-NLPDumper por se tratar de uma extensão do NLPDumper (Macedo, 2010).

A ferramenta foi implementada na linguagem de programação Java (Java, 2010) e configurada para realizar conexões de banco de dados para fins de indexação e persistência de dados. O projeto apresenta a estrutura disposta na Figura 26.

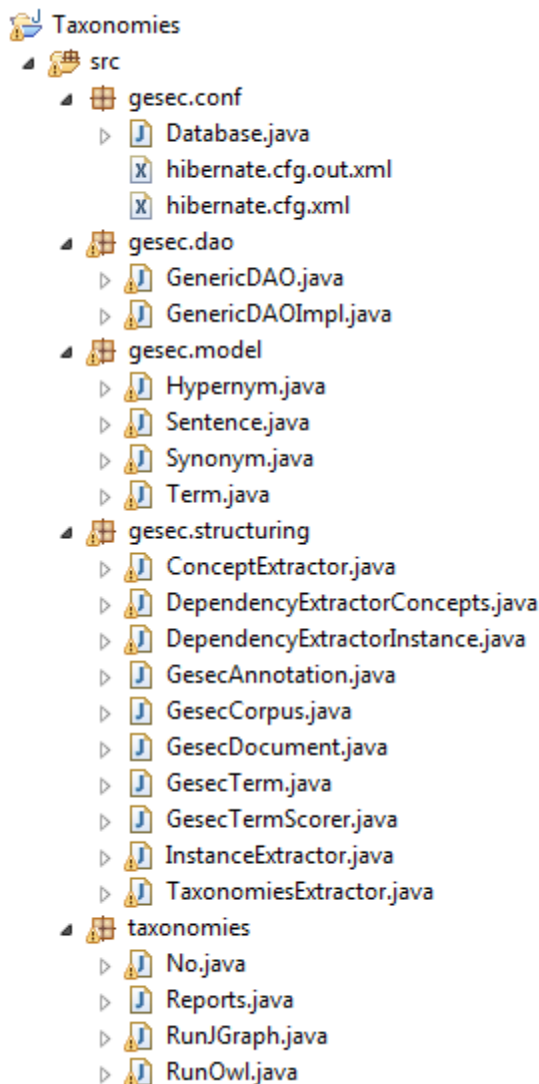


Figura 26: Taxonomy-NLPDumper: a extensão do projeto NPLDumper

O pacotes apresentados são o *gesec.conf*, que é o que possui as configurações da ferramenta; o *gesec.dao*, que possui as classes de manipulação dos dados seguindo o padrão de projetos DAO (Data Access Object) e faz a conexão do banco de dados (Date, 2000); o *gesec.model*, que implementa as classes de modelo representam os elementos manipulados pela ferramenta; o *gesec.structuring*, que possui as classes que realizam o processamento principal do projeto; e, finalmente, o pacote *taxonomies* que faz a instanciação das demais classes, gera e apresenta os resultados.

A partir de uma fonte textual, a ferramenta encontra automaticamente as hierarquias e as apresenta de duas formas: uma ilustração gráfica com os relacionamentos utilizando a API gráfica Jgraph (Jgraph, 2010), que é uma biblioteca que permite a exibição gráfica de árvores, grafos, fluxogramas, etc.; e a

representação em formato OWL (Web Ontology Language) (OWL, 2010), que é uma linguagem de especificação de ontologias. Por exemplo, utilizando o fragmento de texto “*Mary is a good parent. Mary is the mother of Bryan, your child. Steve, the father of Lenn, is Scientist*” como entrada, os resultados obtidos devem ser respectivamente a Figura 27 para o JGraph e a Figura 28 para o formato OWL.

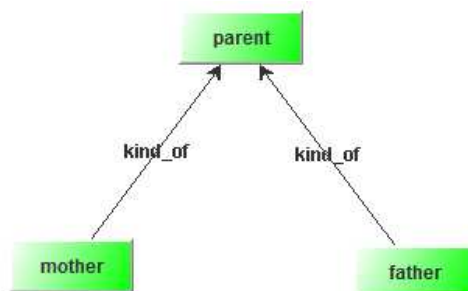


Figura 27: Exemplos de resultados do Taxonomy-NLPDumper com o JGraph

```

<?xml version="1.0"?>
<rdf:RDF xmlns="TaxonomyOntology#"
  xml:base="TaxonomyOntology"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:TaxonomyOntology="TaxonomyOntology#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<owl:Ontology rdf:about="TaxonomyOntology"/>
<!--
// Classes
-->
<!-- TaxonomyOntology#father -->
<owl:Class rdf:about="TaxonomyOntology#father">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#parent"/>
</owl:Class>
<!-- TaxonomyOntology#mother -->
<owl:Class rdf:about="TaxonomyOntology#mother">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#parent"/>
</owl:Class>
<!-- TaxonomyOntology#parent -->
<owl:Class rdf:about="TaxonomyOntology#parent"/>
</rdf:RDF>
  
```

Figura 28: Exemplo dos resultados do Taxonomy-NLPDumper em OWL (OWL, 2010)

Para gerar a ontologia com as relações taxonômicas em formato OWL foi utilizada a biblioteca OWL API (Horridge, Bechhofer, 2009) (Horridge, Bechhofer, Noppens, 2007) (Bechhofer, Lord, Volz, 2003). As taxonomias armazenadas em uma ontologia em formato OWL permitem a fácil utilização dos resultados da ferramenta em um sistema de conhecimento, bem como, tem a vantagem de ser

facilmente editado em uma ferramenta de edição de ontologias que suporte este formato. Os resultados apresentados na Figura 28 quando abertos em um editor de ontologias, podem ser também graficamente visualizados (Figura 29).

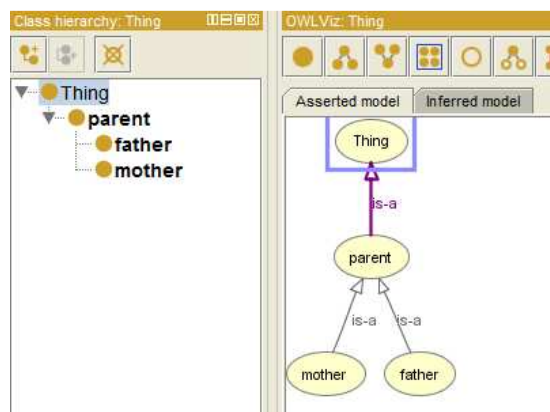


Figura 29: resultados do Taxonomy-NLPDumper em um editor de ontologias (Protégé, 2010)

Além da apresentação gráfica disponível na própria ferramenta e do arquivo OWL que contém as hierarquias, a ferramenta gera um arquivo em formato .txt (Figura 30) que traz informações do corpus processado tais como: total de *tokens*, total de *tokens* candidatos, total de sentenças, número relações taxonômicas, número de *tokens* combinados pelo lema, número de termos da ontologia, número de termos excluídos, número de termos ontologia final, quantas classes têm herança direta da superclasse '*thing*' e total de relações taxonômicas.

A Figura 30 mostra informações adquiridas pelo processamento do mesmo fragmento de texto que deu origem as Figura 27, Figura 28 e Figura 29 e indica que a ferramenta, usando o fragmento "*Mary is a good parent. Mary is the mother of Bryan, your child. Steve, the father of Lenn, is Scietist*", encontrou 25 tokens (incluindo tokens de pontuação), e desses, apenas 10 foram considerados candidatos (conforme especificado na fase "Extração de Classes Candidatas" da Seção 3.1.2) em 3 sentenças. Desses 10 tokens, três foram relacionados, resultando em duas relações taxonômicas. Contando com a herança da superclasse "thing", que é a classe da qual todas as outras herdam, três relações são mostradas na ontologia final como podemos ver na Figura 30.

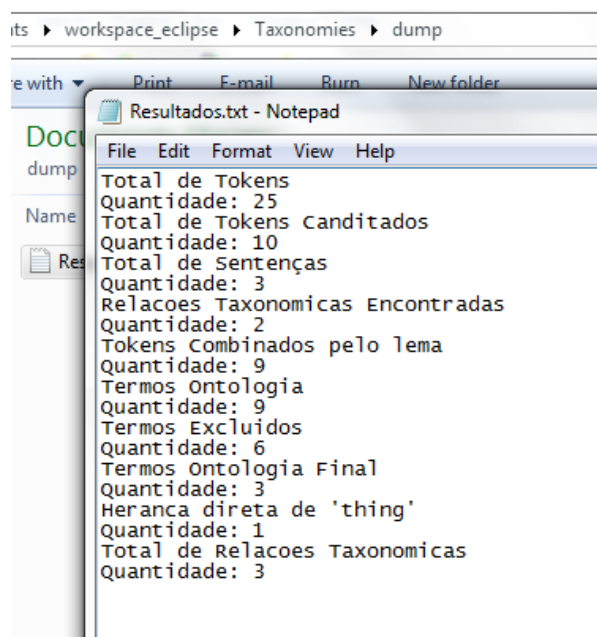


Figura 30: Informações do processamento geradas pelo Taxonomy-NLPDumper

As tarefas que realizam o processamento da linguagem natural foram implementadas utilizando o ambiente GATE (CUNNINGHAM, 2002). Através dele é possível realizar tarefas da fase de marcação (seção 3.1.1) do capítulo 3. A análise sintática foi realizada adicionando o plugin Stanford Parser ao GATE.

4.2 Estudo de Caso

Um estudo de caso na área do direito de família foi conduzido para uma avaliação preliminar da efetividade do processo proposto, utilizando a ferramenta descrita na seção anterior.

4.2.1 O corpus utilizado

Os documentos em linguagem natural utilizados no experimento são oriundos de um corpus da doutrina do direito de família (Family Law, 2010) (ANEXO C) composto por várias sentenças das quais as cem primeiras foram escolhidas e analisadas manualmente por um especialista de domínio. Essa análise manual foi realizada a partir da observação por um profissional do domínio do direito, gerando os resultados disponíveis no Anexo B. Esses resultados foram contabilizados para o

teste da medida *precision* da aplicação do corpus Family Law ao Taxonomy-NLPDumper.

4.2.2 Resultados

Os resultados da aplicação do Taxonomy-NLPDumper foram analisados por um especialista de domínio que concluiu que as relações extraídas, algumas das quais são apresentadas na Figura 31, foram satisfatórias, ou seja, em maioria representam relações taxonômicas acertadas.

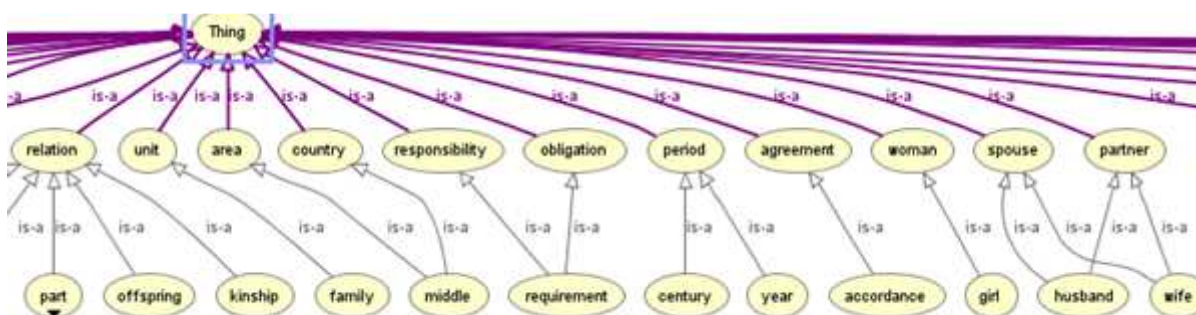


Figura 31: Apresentação parcial dos resultados em OWL em um editor de ontologias

O conjunto completo de relações taxonômicas extraídas (conjunto H da definição de ontologia) é apresentado na forma de rede semântica no Apêndice A e em formato OWL no Apêndice B. Um resumo dos resultados parciais gerados pela ferramenta é apresentado na Tabela 5, indicando que foram encontradas 77 relações taxonômicas

Tabela 5: Resumo dos resultados parciais obtidos no desenvolvimento do estudo de caso

Métrica	Resultado
Tokens	2935
Tokens Candidatos	737
Total de Sentenças	107
Tokens agrupados pelo lema	322
Número de Relações Taxonômicas	77

Termos não relacionados	218
Termos da ontologia final	104
Herdeiros diretos da superclasse "Thing"	37
Total de Relações Taxonômicas	114

A efetividade dos resultados foi avaliada com uma adaptação da medida de *precision* da área de Recuperação de Informação (Dellschaft e Staab, 2006), considerando o número de relações taxonômicas extraídas corretamente.

Precision é a razão entre o número de relações taxonômicas extraídas corretamente (NREC) e o número de relações taxonômicas extraídas (NRE) conforme a fórmula abaixo:

$$P = \text{NREC} / \text{NRE} \quad (2)$$

Logo, para o cálculo desta medida temos que considerar os resultados analisados pelo especialista de domínio como a variável NREC (93 relações como corretas, conforme o Anexo B) e dados da saída do Taxonomy-NLPDumper como a variável NRE (77 relações diretas e 37 da superclasse) resultando em um valor de *precision* de 81,57% (NREC = 93 e NRE = 114).

4.3 Considerações Finais

Este capítulo apresentou a ferramenta Taxonomy-NLPDumper para a extração automática de relações taxonômicas de uma ontologia. Foram apresentados os pacotes que fazem parte do código fonte da ferramenta e suas características. Foi apresentado um estudo de caso dirigido à área do direito de família desenvolvido com o intuito de realizar uma avaliação preliminar do processo proposto no Capítulo 3, bem como foi detalhada, também, a descrição do corpus que foi submetido à experimentação.

As hierarquias resultantes do processamento do corpus foram apresentadas tanto para o fragmento aplicado ao estudo de caso como para o corpus completo e estão disponíveis respectivamente nos Apêndices A e C.

Finalmente foi realizado o cálculo de uma adaptação da medida *precision* da área de Recuperação de Informação para o experimento realizado resultando em

uma alta medida (*precision* de cerca de 80%). Podemos concluir que para o experimento aplicado foi obtido uma alta precisão como ponto forte.

A utilização das ferramentas GATE (GATE, 2010), OWL API (Horridge, Bechhofer, 2009) (Horridge, Bechhofer, Noppens, 2007) (Bechhofer, Lord, Volz, 2007) e JGraph (JGraph, 2010), reduziram generosamente o esforço de implementação da técnica proposta e se mostraram boas alternativas em suas respectivas aplicações.

5 Conclusões

Os sistemas de conhecimento se mostram cada vez mais importantes no mundo de hoje além de serem mais efetivos em diversas tarefas. Com isso, a importante necessidade de contribuições que automatizem tarefas voltadas à aquisição de conhecimento se mostra um ponto crucial para viabilizar a produção destes sistemas.

Ontologias se mostram hoje uma promissora forma de armazenar conhecimento (Girardi, 2010) e o aprendizado destas constitui uma promissora abordagem, que tem sua contribuição na amenização do gargalo da aquisição de conhecimento. Em especial, como tratado neste trabalho, o aprendizado de relações taxonômicas de uma ontologia fornece informações ricas sobre um domínio, bem como tem sua contribuição no alívio do gargalo da aquisição de conhecimento citado no capítulo introdutório, uma vez que estes relacionamentos também são conhecimento.

Muitas das abordagens propostas com o intuito de realizar o aprendizado de relações taxonômicas fazem uso de técnicas de aprendizagem de máquina, como técnicas de agrupamento (Caraballo, 1999) (Sarmiento, 2010) e outras variações de aprendizagem não supervisionada (Cimiano; Hotho; Staab, 2004), processamento da linguagem natural (Hearst, 1992), Máquina de Suporte Vetorial (Ting Wang et al, 2006) e estatística com Redes Lógica de Markov (Drumond, 2009).

Este trabalho apresentou um processo dividido em quatro fases para a aprendizagem de relações taxonômicas de uma ontologia a partir de fontes textuais, aplicando técnicas de Processamento da Linguagem Natural e Padrões Heurísticos. O Wordnet, usado na terceira fase do processo, de um lado contribui com a identificação de parte das hierarquias enquanto, de outro, os padrões heurísticos escritos com Expressões Regulares determinam novas hierarquias em uma quarta e última fase, também responsável por representá-las em uma linguagem de especificação de ontologias.

Podemos considerar o processo proposto no Capítulo 3 como uma proposta automática independente de domínio, desde que a fonte de entrada (corpus) seja escrita em língua inglesa.

Para a construção do Taxonomy-NLPDumper, a ferramenta construída com o intuito de realizar uma avaliação preliminar da efetividade do processo, as

ferramentas GATE (GATE, 2010), OWL API (Horridge, Bechhofer, 2009) (Horridge, Bechhofer, Noppens, 2007) (Bechhofer, Lord, Volz, 2003) e JGraph (JGraph, 2010) reduziram generosamente o esforço de implementação e se mostraram boas alternativas em suas respectivas aplicações. Na avaliação realizada, a medida de *precision* foi de cerca de 80%, o que podemos considerar uma alta precisão.

5.1 Limitações

Dentre as limitações do processo, se mostram a dependência de uma base lexical e a necessidade de escrita de padrões heurísticos que podem não atender as necessidades de domínios muito específicos para um bom funcionamento.

A necessidade de se obter como entrada textos somente em linguagem natural e em formato *.txt*, limita a ferramenta ao uso em documentos que devem ser “tratados” antes da execução da ferramenta, deixando de lado entradas interessantes como documentos HTML, XML, planilhas ou arquivos de editores de texto proprietários.

Como os documentos de entrada podem conter irregularidades de escrita, muitos termos podem ser interpretados erroneamente como, por exemplo, a falta de letras maiúsculas em nomes próprios ou mesmo termos separados por hífen podem ser ignorados ou interpretados erroneamente.

Como (Cimiano et al, 2005) citou em seu trabalho, existe uma escassez de apresentações comparativas entre métodos que visam o aprendizado. Uma boa proposta de trabalho é a comparação dos resultados encontrados com este processo e ferramenta com outros trabalhos. Porém, experimentos desta natureza nem sempre são simples e, possivelmente, necessitariam de mais recursos e a colaboração de outros grupos para realizar experimentos mais fieis a suas respectivas propostas e sob as mesmas condições.

Para o experimento realizado, foi utilizado somente corpus do direito da família, sendo que para o processamento de corpora de grande volume é esperada uma razoável demora nos resultados, contudo, a contagem manual de relações taxonômicas na maioria dos casos é também inviável.

A implementação da ferramenta e, provavelmente, as futuras extensões também, requerem o conhecimento de muitas tecnologias e APIs tais como: PLN, Bancos de Dados, Linguagem de Programação, Ontologias, Wordnet, OWL, etc.;

5.2 Trabalhos Futuros

Algumas das limitações da ferramenta podem ser corrigidas em trabalhos futuros, como por exemplo:

- A utilização de outras ferramentas de processamento de linguagem natural pode melhorar a qualidade dos termos recuperados/desprezados;
- A escrita de padrões de hierarquias mais complexos que possam aumentar a taxa de identificação de relações taxonômicas;
- Anexar à ferramenta conversores de formato pode apresentar aperfeiçoamentos de grande contribuição;
- Corretores da formatação da escrita (retirada de hífens, correção de letras maiúsculas, etc.) ajudam em uma aplicação melhor sucedida do PLN;
- A identificação de entidades nomeadas pode dar mais riqueza as classes e hierarquias recuperadas.
- Melhorias no desempenho da ferramenta.

As relações de hiperonímia que estão presentes na base lexical, mas que não existem nos documentos analisados pela ferramenta Taxonomy-NLPDumper, podem contribuir com informações importantes que podem ajudar a identificar o domínio analisado. Isto pode ser considerado um conhecimento implícito útil para alguns tipos de aplicação e também que pode contribuir para a solução de problemas de identificação do domínio como o apresentado em Ting Wang (2006).

Outras extensões da ferramenta Taxonomy-NLPDumper podem ser realizadas como, por exemplo, a utilização dos padrões propostos no processo para determinação de hierarquias ou localizar outros conjuntos pertencentes à definição de ontologia apresentada aqui como, por exemplo, o conjunto R que representa as demais relações (relações não taxonômicas) que fazem parte de uma ontologia.

REFERÊNCIAS

- Allen J.: Natural Language Understanding. Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc (1995).
- Bechhofer, Sean; Lord ,Phillip; Volz, Raphael. Cooking the Semantic Web with the OWL API. 2nd International Semantic Web Conference, ISWC, Sanibel Island, Florida, October 2003.
- Biemann, C. Ontology learning from text: a survey of methods. LDV Forum, [S. I.], v. 20, n. 2, p. 75 – 93, 2005.
- Caraballo, S.A. Automatic construction of a hypernym-labeled noun hierarchy from text. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics Morristown, NJ, USA, p. 120–126, 1999.
- Cimiano, P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.
- Cimiano, P.; Hotho, A.; Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. 2004.
- Cimiano, P.; Staab, S. Learning concept hierarchies from text with a guided hierarchical clustering algorithm. In: ICML workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods. 2005.
- Cunningham, H. Developing Language Processing Components with GATE - Version 5 (a User Guide). University of Sheffield, 2009.
- Cunningham, H. Gate, a general architecture for text engineering.Computers And The Humanities, [S. I.], v. 36, n. 2, p. 223 -2 54, 2002.
- Dale, R.; Moisl, H.; Somers, H. L. Handbook of natural language processing. CRC Press, 2000.
- Drake, Mirian A. Encyclopedia of Library and Information Science: Lib-Pub. CRC Press, 2003
- Dellschaft K., Staab S. (2006) “On how to perform a gold standard based evaluation of ontology learning”, In: Proceedings of the 5th International Semantic Web Conference, p. 228 – 241, Athens. Springer.
- Drumond, L. Aquisição Automatizada de Hierarquias de Conceitos de Ontologias Utilizando Aprendizagem Estatística Relacional.Dissertação de Mestrado, Universidade Federal do Maranhão, 2009.
- Fellbaum, C. WordNet: An Electronic Lexical Database.Cambridge: MIT Press, 1998. 23-24 p.

- Freitas, M. Elaboração automática de ontologias de domínio: discussão e resultados. Tese de Doutorado, Pontifícia Universidade Católica do Rio de Janeiro. 2007
- GATE. General Architecture for Text Engineering. Disponível em <<http://gate.ac.uk/>> Último acesso em 10 de agosto de 2010.
- GESEC. Grupo de pesquisa em Engenharia de Software e Engenharia do Conhecimento. Disponível em <<http://maae.deinf.ufma.br>> Último acesso em agosto de 2010.
- Girardi, R. Guiding Ontology Learning and Population by Knowledge System Goals. International Conference on Knowledge Engineering and Ontology Development. Valencia. 2010
- GrammarNet. Gramática da Língua Inglesa. Disponível em <<http://www.grammarnet.com/ghtml/ntypes.htm>> Último acesso em fevereiro de 2011.
- Gruber, T. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, p. 907-928, 1995.
- 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.
- Hearst, M. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics. 1992.
- Horridge, Matthew; Bechhofer, Sean ; Noppens, Olaf. Igniting the OWL 1.1 Touch Paper: The OWL API. OWLED 2007, 3rd OWL Experienced and Directions Workshop, Innsbruck, Austria, June 2007.
- Horridge, Matthew; Bechhofer, Sean. The OWL API: A Java API for Working with OWL 2 Ontologies. OWLED 2009, 6th OWL Experienced and Directions Workshop, Chantilly, Virginia, October 2009.
- Jargas, Aurélio Marinho. Expressões Regulares. Editora Novatec. 2ed. 2008.
- Java. Oracle. Disponível em: <<http://www.oracle.com/technetwork/java/index.html/>>. Último acesso em dezembro de 2010.
- JGraph. JGraph:Visualize Everything. Disponível em <<http://jgraph.com/>> Último acesso em 01 de outubro de 2010.
- Jielin, D. Network Dictionary. Javvin Technologies, 2007. Disponível em: <http://books.google.com.br/books?id=On_Hh23IXDUC&printsec=frontcover&dq=Network+Dictionary#v=onepage&q=&f=false>. Último acesso em setembro 2010.

- Lima, V.L.S; Nunes, M.G.V; Vieira, R. Desafios do Processamento de Línguas Naturais. In: SEMISH - XXXIV Seminário Integrado de Software e Hardware. Anais do XXVII Congresso da SBC, Rio de Janeiro, 2007.
- Lopar. An implementation of a parser for head-lexicalised probabilistic context-free grammars. Disponível em <<http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LoPar.html>> Último acesso em setembro de 2010.
- Macedo, M. Processamento da linguagem natural para identificação de classes e instâncias de uma ontologia. Monografia de Graduação, Universidade Federal do Maranhão, 2010.
- 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.
- Mitchell P. Marcus. Theory of Syntactic Recognition for Natural Languages. Cambridge: MIT Press, 1980. 335 pg.
- Mitchell, T. Machine Learning, McGraw Hill (1997).
- Monard, Maria Carolina; Baranauskas, José Augusto. Indução de Regras e Árvores de Decisão. In: REZENDE, Solange Oliveira (coord.). Sistemas Inteligentes. Barueri: Manole, 2005.
- Mosqueira, A. et al. JavaBeans. Faculdade de Computação e Informática. Universidade Presbiteriana Mackenzie, 2006.
- Mundo Educação. Hipônimos e hiperônimos, holónimos e merónimos. Disponível em: <<http://www.mundoeducacao.com.br/gramatica/hiponimos-hiperonimos-holonimos-meronimos.htm>>. Último acesso em outubro de 2010.
- OWL. OWL Web Ontology Language Overview. Disponível em <<http://www.w3.org/TR/owl-features/>> Último acesso em 10 de dezembro de 2010.
- Sarmiento, G. Aplicação de Técnicas de Agrupamento Hierárquico para a Extração das Relações Taxonômicas de uma Ontologia. Monografia de Graduação, Universidade Federal do Maranhão, 2009.
- Sundblad, H. (2002). Automatic Acquisition of Hyponyms and Meronyms from Question Corpora. In Proceedings from the Workshop on Natural Language Processing and Machine Learning for Ontology Engineering at ECAI2002. Lyon, France
- Protégé - Ontology Editor and Knowledge Acquisition System. Disponível em <http://protege.stanford.edu>. Último acesso em março de 2010.
- Ting Wang et al. Automatic Extraction of Hierarchical Relations from Text. 2006

Treetagger. A language independent part-of-speech tagger. Disponível em <<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>> Último acesso em setembro de 2010.

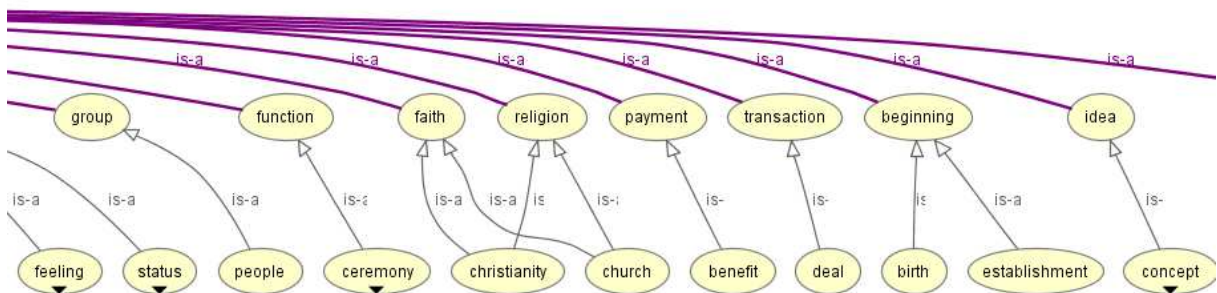
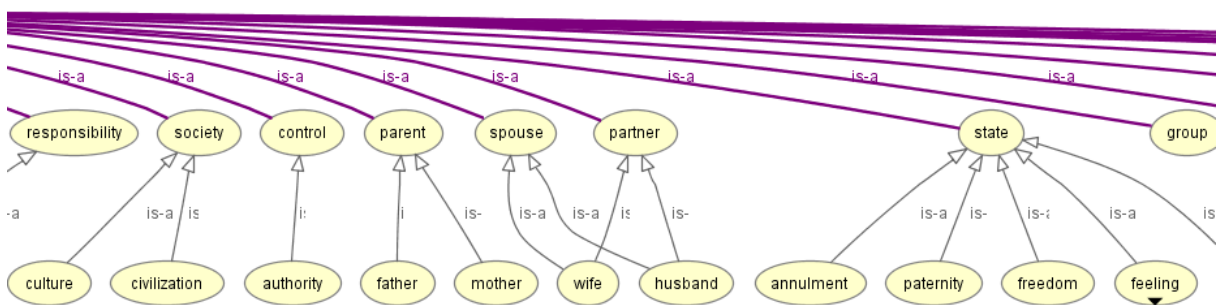
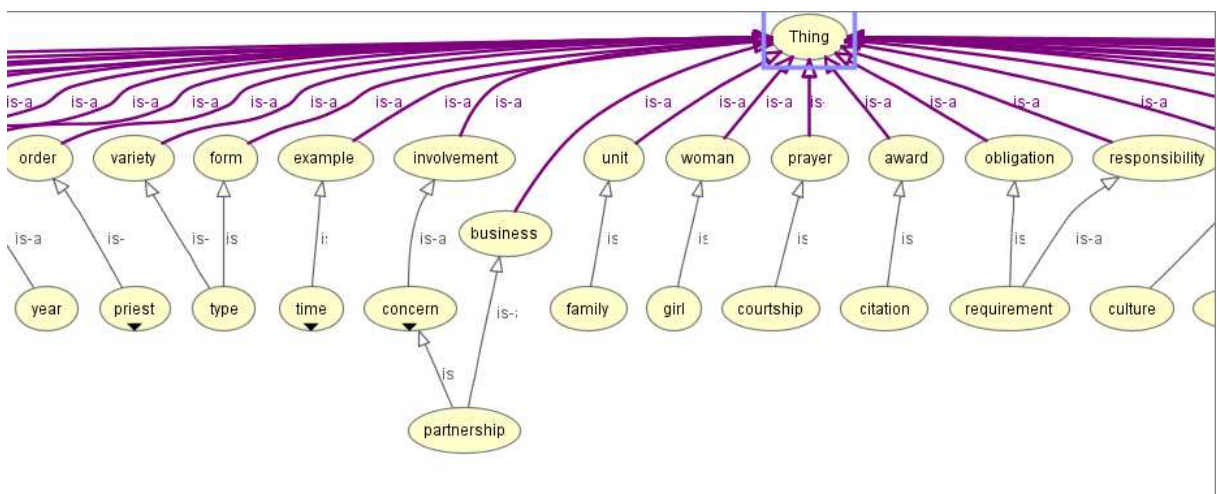
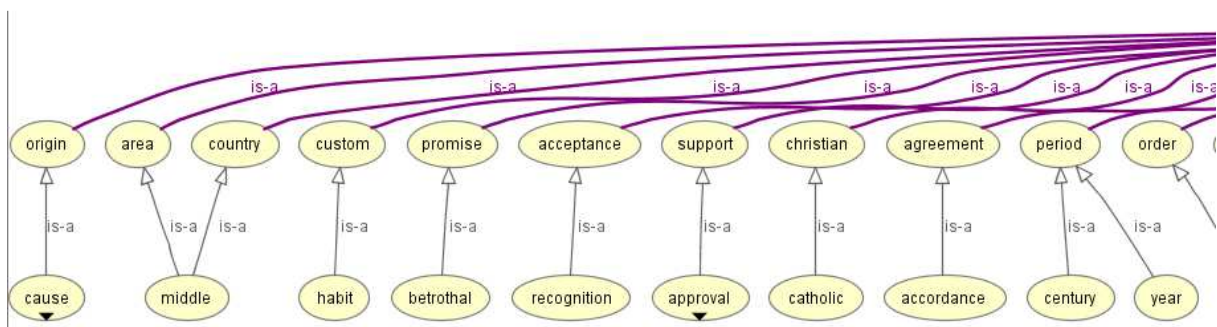
Weka project. Disponível em: <<http://www.cs.waikato.ac.nz/ml/weka/>>. Acessado em: 13 de novembro de 2009.

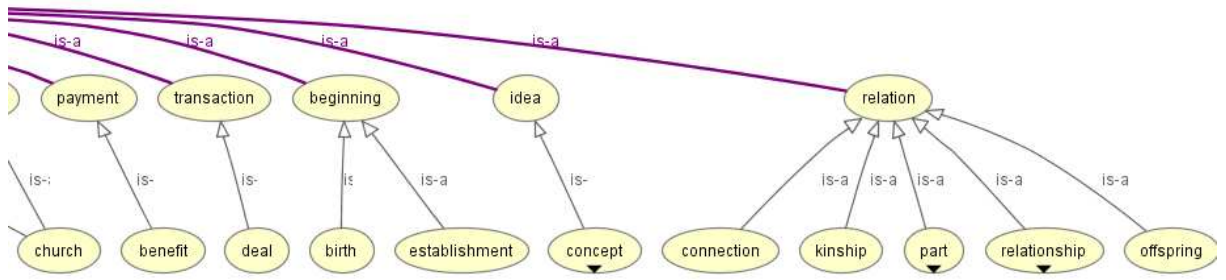
Wordnet. WordNet: A lexical database for English. Cognitive Science Laboratory, Princeton University, Disponível em: <<http://wordnet.princeton.edu/>>. Último acesso em setembro de 2010.

Fisher, D. Knowledge Acquisition Via Incremental Conceptual Clustering, Machine Learning, 2, 139--172. Reprinted in J. Shavlik & T. Dietterich (eds.), Readings in Machine Learning, 267--283, Morgan Kaufmann, 1990.

APÊNDICES

APÊNDICE A – Conjunto H de relações taxonômicas encontradas pela ferramenta Taxonomy-NLPDumper nas cem primeiras sentenças avaliadas do corpus “Family Law”





APÊNDICE B – Conjunto H de relações taxonômicas geradas no estudo de caso em formato OWL

```

<?xml version="1.0"?>
<rdf:RDF xmlns="TaxonomyOntology#"
  xml:base="TaxonomyOntology"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:TaxonomyOntology="TaxonomyOntology#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<owl:Ontology rdf:about="TaxonomyOntology"/>
<!--
// Classes
//
// Classes
//
-->
<!-- TaxonomyOntology#acceptance -->
<owl:Class rdf:about="TaxonomyOntology#acceptance"/>
<!-- TaxonomyOntology#accordance -->
<owl:Class rdf:about="TaxonomyOntology#accordance">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#agreement"/>
</owl:Class>
<!-- TaxonomyOntology#adoption -->
<owl:Class rdf:about="TaxonomyOntology#adoption">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#approval"/>
</owl:Class>
<!-- TaxonomyOntology#advocate -->
<owl:Class rdf:about="TaxonomyOntology#advocate">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#individual"/>
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#person"/>
</owl:Class>
<!-- TaxonomyOntology#affection -->
<owl:Class rdf:about="TaxonomyOntology#affection">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#feeling"/>
</owl:Class>
<!-- TaxonomyOntology#affiliation -->
<owl:Class rdf:about="TaxonomyOntology#affiliation">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#relationship"/>
</owl:Class>
<!-- TaxonomyOntology#agreement -->
<owl:Class rdf:about="TaxonomyOntology#agreement"/>
<!-- TaxonomyOntology#allowance -->
<owl:Class rdf:about="TaxonomyOntology#allowance">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#part"/>
</owl:Class>
<!-- TaxonomyOntology#annulment -->
<owl:Class rdf:about="TaxonomyOntology#annulment">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#state"/>
</owl:Class>
<!-- TaxonomyOntology#approval -->
<owl:Class rdf:about="TaxonomyOntology#approval">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#support"/>
</owl:Class>
  <!-- TaxonomyOntology#area -->
<owl:Class rdf:about="TaxonomyOntology#area"/>
  <!-- TaxonomyOntology#authority -->
<owl:Class rdf:about="TaxonomyOntology#authority">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#control"/>
</owl:Class>
  <!-- TaxonomyOntology#award -->
<owl:Class rdf:about="TaxonomyOntology#award"/>
  <!-- TaxonomyOntology#beginning -->
<owl:Class rdf:about="TaxonomyOntology#beginning"/>
  <!-- TaxonomyOntology#benefit -->
<owl:Class rdf:about="TaxonomyOntology#benefit">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#payment"/>
</owl:Class>
  <!-- TaxonomyOntology#betrothal -->
<owl:Class rdf:about="TaxonomyOntology#betrothal">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#promise"/>
</owl:Class>
  <!-- TaxonomyOntology#birth -->
<owl:Class rdf:about="TaxonomyOntology#birth">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#beginning"/>
</owl:Class>
  <!-- TaxonomyOntology#bishop -->
<owl:Class rdf:about="TaxonomyOntology#bishop">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#priest"/>
</owl:Class>
  <!-- TaxonomyOntology#business -->
<owl:Class rdf:about="TaxonomyOntology#business">
  <!-- TaxonomyOntology#catholic -->
<owl:Class rdf:about="TaxonomyOntology#catholic">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#christian"/>
</owl:Class>
  <!-- TaxonomyOntology#cause -->
<owl:Class rdf:about="TaxonomyOntology#cause">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#origin"/>
</owl:Class>
  <!-- TaxonomyOntology#century -->
<owl:Class rdf:about="TaxonomyOntology#century">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#period"/>
</owl:Class>
  <!-- TaxonomyOntology#ceremony -->
<owl:Class rdf:about="TaxonomyOntology#ceremony">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#function"/>
</owl:Class>
  <!-- TaxonomyOntology#christian -->
<owl:Class rdf:about="TaxonomyOntology#christian"/>
  <!-- TaxonomyOntology#christianity -->
<owl:Class rdf:about="TaxonomyOntology#christianity">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#faith"/>
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#religion"/>
</owl:Class>
  <!-- TaxonomyOntology#church -->
<owl:Class rdf:about="TaxonomyOntology#church">

```

```

    <rdfs:subClassOf rdf:resource="TaxonomyOntology#faith"/>
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#religion"/>
  </owl:Class>
  <!-- TaxonomyOntology#circumstance -->
  <owl:Class rdf:about="TaxonomyOntology#circumstance">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#status"/>
  </owl:Class>
  <!-- TaxonomyOntology#citation -->
  <owl:Class rdf:about="TaxonomyOntology#citation">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#award"/>
  </owl:Class>
  <!-- TaxonomyOntology#civilization -->
  <owl:Class rdf:about="TaxonomyOntology#civilization">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#society"/>
  </owl:Class>
  <!-- TaxonomyOntology#concept -->
  <owl:Class rdf:about="TaxonomyOntology#concept">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#idea"/>
  </owl:Class>
  <!-- TaxonomyOntology#concern -->
  <owl:Class rdf:about="TaxonomyOntology#concern">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#involvement"/>
  </owl:Class>
  <!-- TaxonomyOntology#connection -->
  <owl:Class rdf:about="TaxonomyOntology#connection">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#relation"/>
  </owl:Class>
  <!-- TaxonomyOntology#control -->
  <owl:Class rdf:about="TaxonomyOntology#control"/>
  <!-- TaxonomyOntology#country -->
  <owl:Class rdf:about="TaxonomyOntology#country"/>
  <!-- TaxonomyOntology#courtship -->
  <owl:Class rdf:about="TaxonomyOntology#courtship">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#prayer"/>
  </owl:Class>
  <!-- TaxonomyOntology#culture -->
  <owl:Class rdf:about="TaxonomyOntology#culture">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#society"/>
  </owl:Class>
  <!-- TaxonomyOntology#custom -->
  <owl:Class rdf:about="TaxonomyOntology#custom"/>
  <!-- TaxonomyOntology#deal -->
  <owl:Class rdf:about="TaxonomyOntology#deal">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#transaction"/>
  </owl:Class>
  <!-- TaxonomyOntology#element -->
  <owl:Class rdf:about="TaxonomyOntology#element">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#part"/>
  </owl:Class>
  <!-- TaxonomyOntology#establishment -->
  <owl:Class rdf:about="TaxonomyOntology#establishment">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#beginning"/>
  </owl:Class>
  <!-- TaxonomyOntology#example -->
  <owl:Class rdf:about="TaxonomyOntology#example"/>
  <!-- TaxonomyOntology#factor -->
  <owl:Class rdf:about="TaxonomyOntology#factor">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#cause"/>

```

```

  </owl:Class>
  <!-- TaxonomyOntology#faith -->
  <owl:Class rdf:about="TaxonomyOntology#faith"/>
  <!-- TaxonomyOntology#family -->
  <owl:Class rdf:about="TaxonomyOntology#family">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#unit"/>
  </owl:Class>
  <!-- TaxonomyOntology#father -->
  <owl:Class rdf:about="TaxonomyOntology#father">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#parent"/>
  </owl:Class>
  <!-- TaxonomyOntology#feeling -->
  <owl:Class rdf:about="TaxonomyOntology#feeling">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#state"/>
  </owl:Class>
  <!-- TaxonomyOntology#form -->
  <owl:Class rdf:about="TaxonomyOntology#form"/>
  <!-- TaxonomyOntology#freedom -->
  <owl:Class rdf:about="TaxonomyOntology#freedom">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#state"/>
  </owl:Class>
  <!-- TaxonomyOntology#function -->
  <owl:Class rdf:about="TaxonomyOntology#function"/>
  <!-- TaxonomyOntology#future -->
  <owl:Class rdf:about="TaxonomyOntology#future">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#time"/>
  </owl:Class>
  <!-- TaxonomyOntology#girl -->
  <owl:Class rdf:about="TaxonomyOntology#girl">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#woman"/>
  </owl:Class>
  <!-- TaxonomyOntology#group -->
  <owl:Class rdf:about="TaxonomyOntology#group"/>
  <!-- TaxonomyOntology#habit -->
  <owl:Class rdf:about="TaxonomyOntology#habit">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#custom"/>
  </owl:Class>
  <!-- TaxonomyOntology#heterosexual -->
  <owl:Class rdf:about="TaxonomyOntology#heterosexual">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#individual"/>
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#person"/>
  </owl:Class>
  <!-- TaxonomyOntology#husband -->
  <owl:Class rdf:about="TaxonomyOntology#husband">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#partner"/>
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#spouse"/>
  </owl:Class>
  <!-- TaxonomyOntology#idea -->
  <owl:Class rdf:about="TaxonomyOntology#idea"/>
  <!-- TaxonomyOntology#illegitimacy -->
  <owl:Class rdf:about="TaxonomyOntology#illegitimacy">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#status"/>
  </owl:Class>
  <!-- TaxonomyOntology#individual -->
  <owl:Class rdf:about="TaxonomyOntology#individual">
    <rdfs:subClassOf rdf:resource="TaxonomyOntology#cause"/>
  </owl:Class>
  <!-- TaxonomyOntology#involvement -->

```

```

<owl:Class rdf:about="TaxonomyOntology#involvement"/>
  <!-- TaxonomyOntology#kinship -->
<owl:Class rdf:about="TaxonomyOntology#kinship">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#relation"/>
</owl:Class>
  <!-- TaxonomyOntology#matter -->
<owl:Class rdf:about="TaxonomyOntology#matter">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#concern"/>
</owl:Class>
  <!-- TaxonomyOntology#middle -->
<owl:Class rdf:about="TaxonomyOntology#middle">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#area"/>
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#country"/>
</owl:Class>
  <!-- TaxonomyOntology#mother -->
<owl:Class rdf:about="TaxonomyOntology#mother">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#parent"/>
</owl:Class>
  <!-- TaxonomyOntology#necessity -->
<owl:Class rdf:about="TaxonomyOntology#necessity">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#need"/>
</owl:Class>
  <!-- TaxonomyOntology#need -->
<owl:Class rdf:about="TaxonomyOntology#need">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#status"/>
</owl:Class>
  <!-- TaxonomyOntology#obligation -->
<owl:Class rdf:about="TaxonomyOntology#obligation"/>
  <!-- TaxonomyOntology#offspring -->
<owl:Class rdf:about="TaxonomyOntology#offspring">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#relation"/>
</owl:Class>
  <!-- TaxonomyOntology#order -->
<owl:Class rdf:about="TaxonomyOntology#order"/>
  <!-- TaxonomyOntology#ordinance -->
<owl:Class rdf:about="TaxonomyOntology#ordinance">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#rule"/>
</owl:Class>
  <!-- TaxonomyOntology#origin -->
<owl:Class rdf:about="TaxonomyOntology#origin"/>
  <!-- TaxonomyOntology#parent -->
<owl:Class rdf:about="TaxonomyOntology#parent"/>
  <!-- TaxonomyOntology#part -->
<owl:Class rdf:about="TaxonomyOntology#part">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#relation"/>
</owl:Class>
  <!-- TaxonomyOntology#partner -->
<owl:Class rdf:about="TaxonomyOntology#partner"/>
  <!-- TaxonomyOntology#partnership -->
<owl:Class rdf:about="TaxonomyOntology#partnership">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#business"/>
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#concern"/>
</owl:Class>
  <!-- TaxonomyOntology#paternity -->
<owl:Class rdf:about="TaxonomyOntology#paternity">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#state"/>
</owl:Class>
  <!-- TaxonomyOntology#payment -->

```

```

<owl:Class rdf:about="TaxonomyOntology#payment"/>
  <!-- TaxonomyOntology#people -->
<owl:Class rdf:about="TaxonomyOntology#people">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#group"/>
</owl:Class>
  <!-- TaxonomyOntology#period -->
<owl:Class rdf:about="TaxonomyOntology#period"/>
  <!-- TaxonomyOntology#person -->
<owl:Class rdf:about="TaxonomyOntology#person">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#cause"/>
</owl:Class>
  <!-- TaxonomyOntology#prayer -->
<owl:Class rdf:about="TaxonomyOntology#prayer"/>
  <!-- TaxonomyOntology#priest -->
<owl:Class rdf:about="TaxonomyOntology#priest">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#order"/>
</owl:Class>
  <!-- TaxonomyOntology#promise -->
<owl:Class rdf:about="TaxonomyOntology#promise">
  <!-- TaxonomyOntology#recognition -->
<owl:Class rdf:about="TaxonomyOntology#recognition">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#acceptance"/>
</owl:Class>
  <!-- TaxonomyOntology#relation -->
<owl:Class rdf:about="TaxonomyOntology#relation"/>
  <!-- TaxonomyOntology#relationship -->
<owl:Class rdf:about="TaxonomyOntology#relationship">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#relation"/>
</owl:Class>
  <!-- TaxonomyOntology#religion -->
<owl:Class rdf:about="TaxonomyOntology#religion"/>
  <!-- TaxonomyOntology#requirement -->
<owl:Class rdf:about="TaxonomyOntology#requirement">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#obligation"/>
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#responsibility"/>
</owl:Class>
  <!-- TaxonomyOntology#responsibility -->
<owl:Class rdf:about="TaxonomyOntology#responsibility"/>
  <!-- TaxonomyOntology#rule -->
<owl:Class rdf:about="TaxonomyOntology#rule">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#concept"/>
</owl:Class>
  <!-- TaxonomyOntology#society -->
<owl:Class rdf:about="TaxonomyOntology#society"/>
  <!-- TaxonomyOntology#spouse -->
<owl:Class rdf:about="TaxonomyOntology#spouse"/>
  <!-- TaxonomyOntology#state -->
<owl:Class rdf:about="TaxonomyOntology#state"/>
  <!-- TaxonomyOntology#status -->
<owl:Class rdf:about="TaxonomyOntology#status">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#state"/>
</owl:Class>
  <!-- TaxonomyOntology#support -->
<owl:Class rdf:about="TaxonomyOntology#support"/>
  <!-- TaxonomyOntology#time -->
<owl:Class rdf:about="TaxonomyOntology#time">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#example"/>
</owl:Class>

```



```

<!-- TaxonomyOntology#transaction -->
<owl:Class rdf:about="TaxonomyOntology#transaction"/>
<!-- TaxonomyOntology#type -->
<owl:Class rdf:about="TaxonomyOntology#type">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#form"/>
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#variety"/>
</owl:Class>
<!-- TaxonomyOntology#unit -->
<owl:Class rdf:about="TaxonomyOntology#unit"/>
<!-- TaxonomyOntology#variety -->
<owl:Class rdf:about="TaxonomyOntology#variety"/>
<!-- TaxonomyOntology#virginity -->
<owl:Class rdf:about="TaxonomyOntology#virginity">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#status"/>
</owl:Class>
<!-- TaxonomyOntology#wedding -->
<owl:Class rdf:about="TaxonomyOntology#wedding">

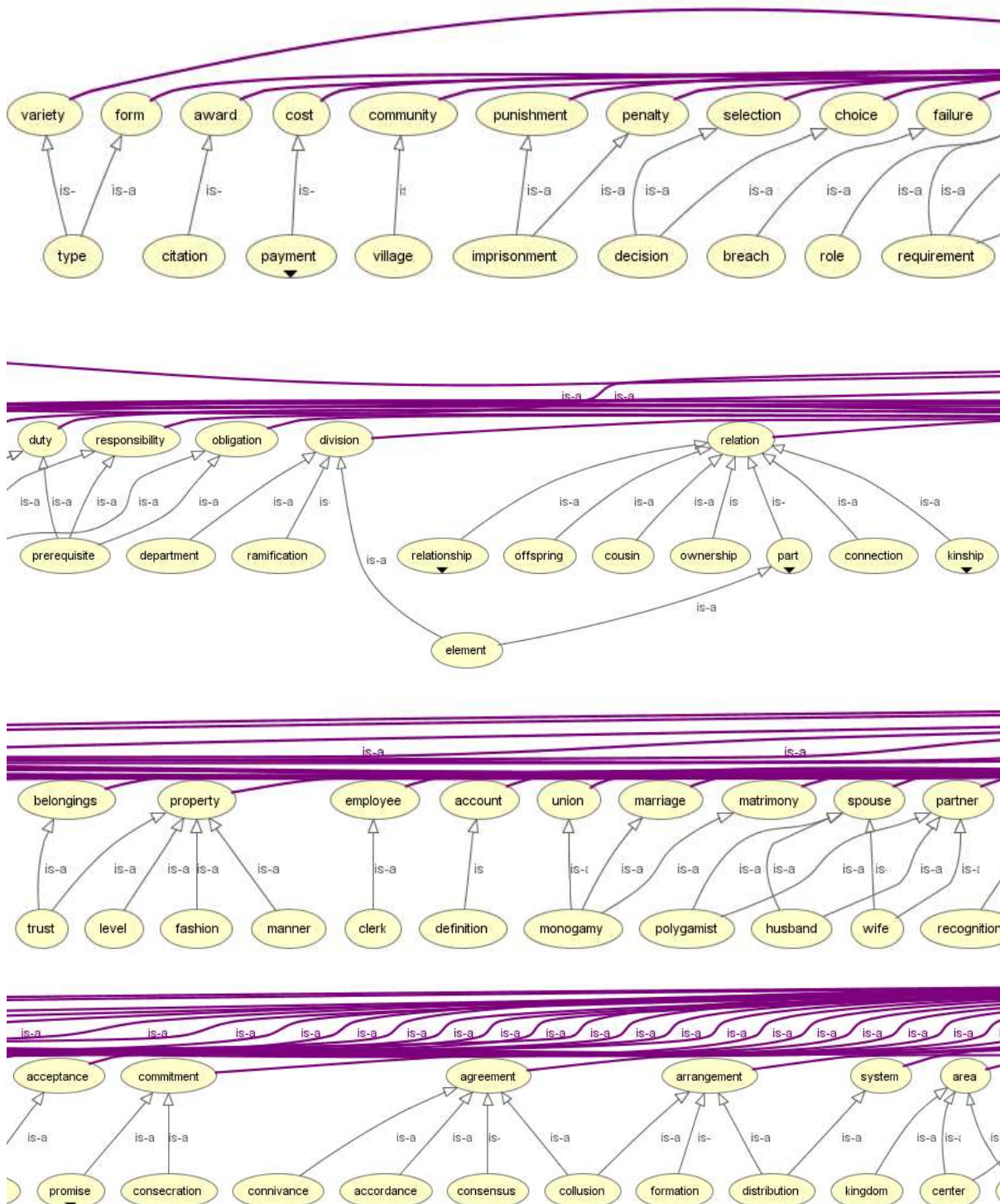
```

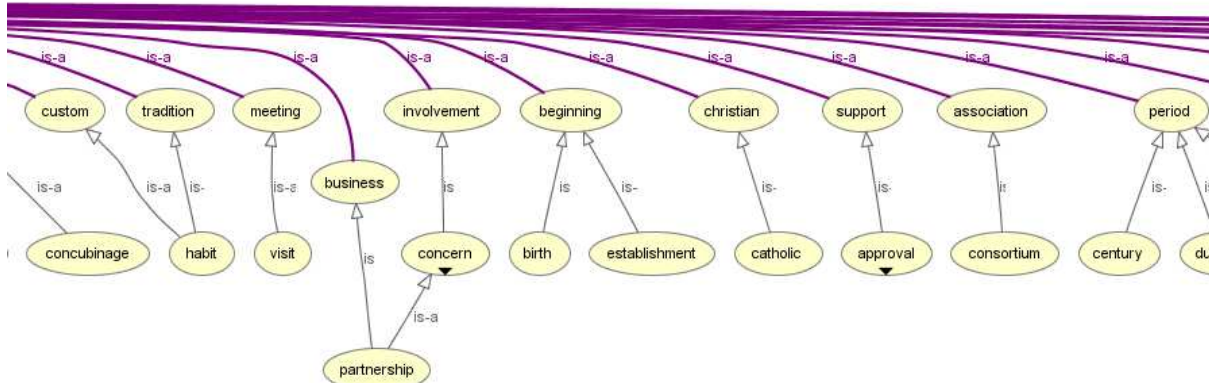
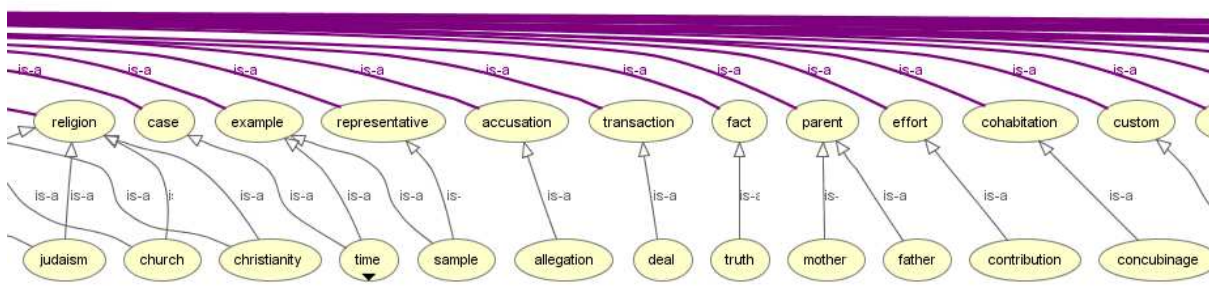
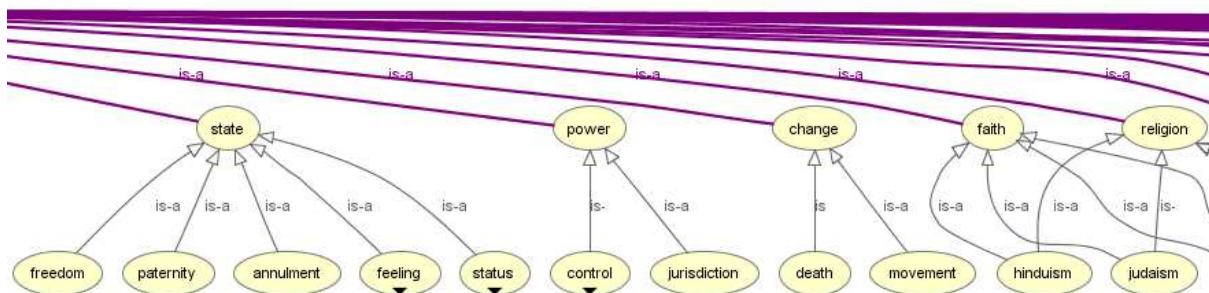
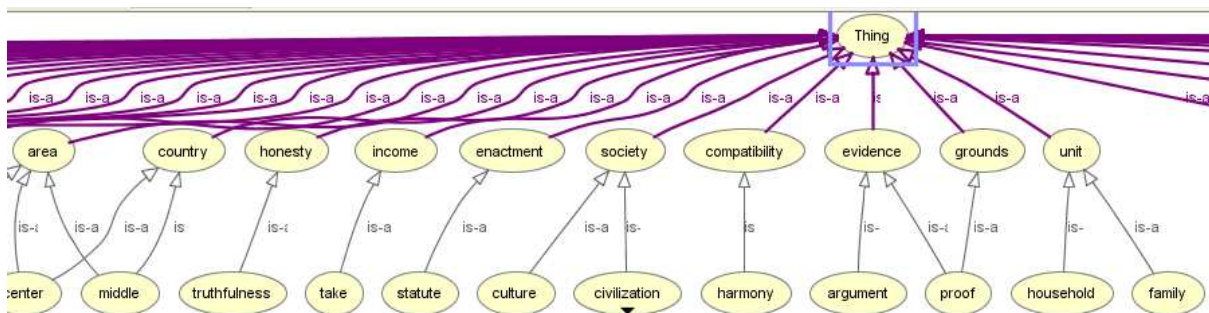
```

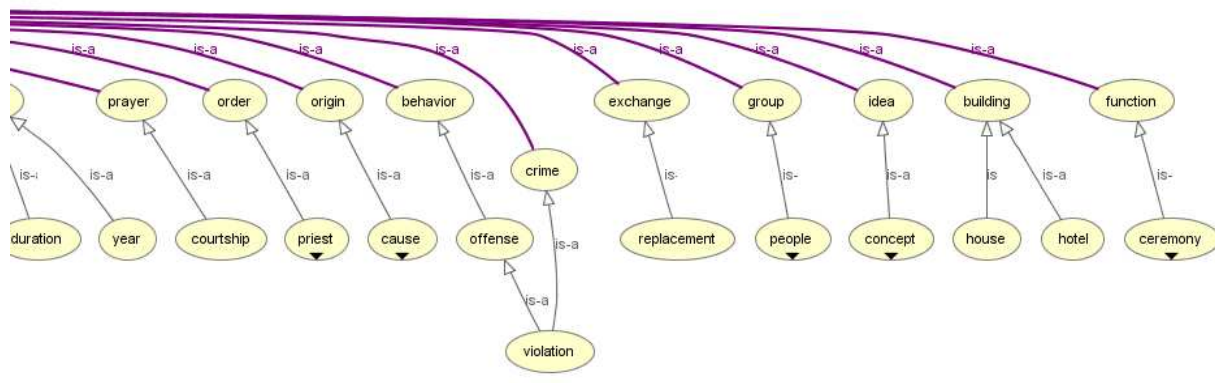
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#ceremony"/>
</owl:Class>
<!-- TaxonomyOntology#wife -->
<owl:Class rdf:about="TaxonomyOntology#wife">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#partner"/>
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#spouse"/>
</owl:Class>
<!-- TaxonomyOntology#woman -->
<owl:Class rdf:about="TaxonomyOntology#woman"/>
<!-- TaxonomyOntology#year -->
<owl:Class rdf:about="TaxonomyOntology#year">
  <rdfs:subClassOf rdf:resource="TaxonomyOntology#period"/>
</owl:Class>
</rdf:RDF>
<!-- Generated by the OWL API (version 3.1.0.1592)
http://owlapi.sourceforge.net -->

```

APÊNDICE C – Conjunto H de relações taxonômicas encontradas pela ferramenta Taxonomy-NLPDumper no corpus Family Law do Anexo C







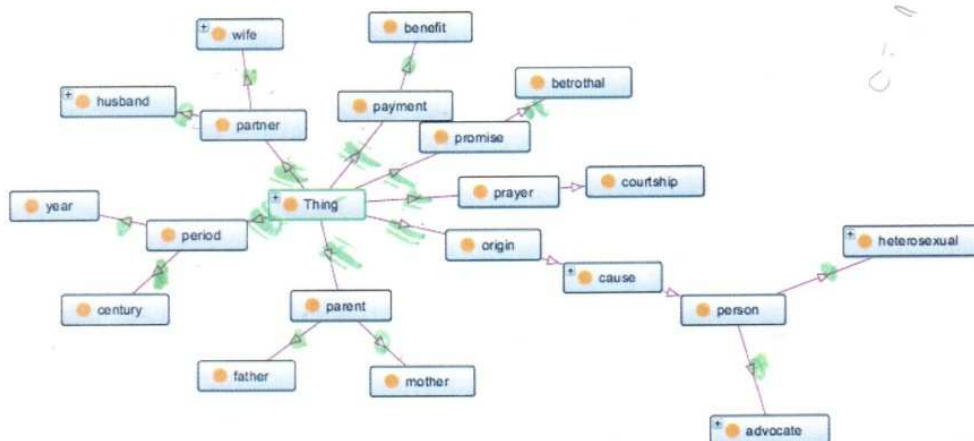
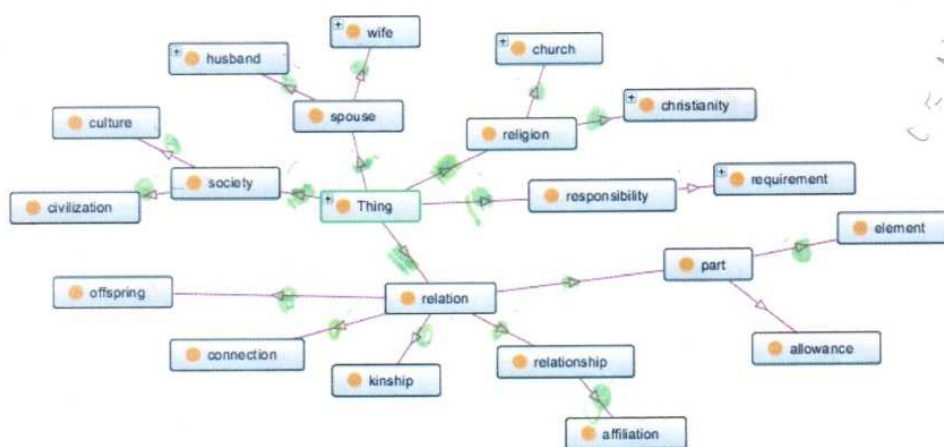
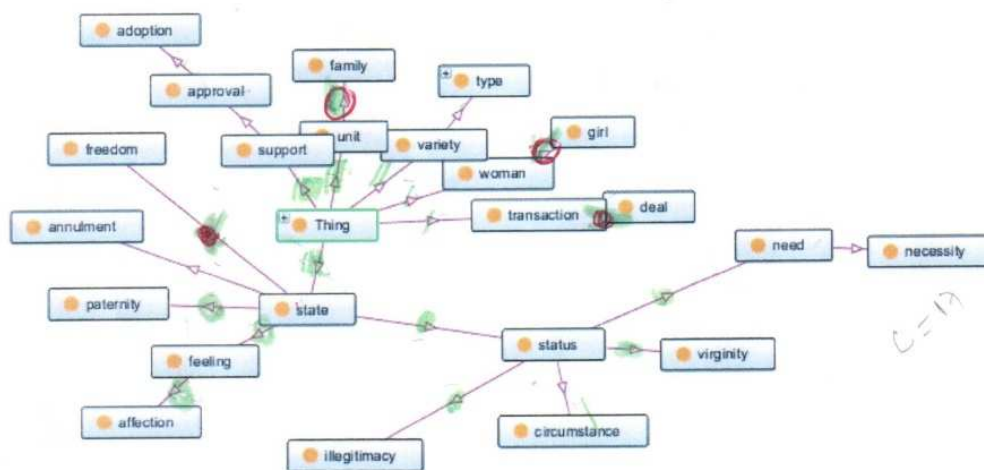
ANEXOS

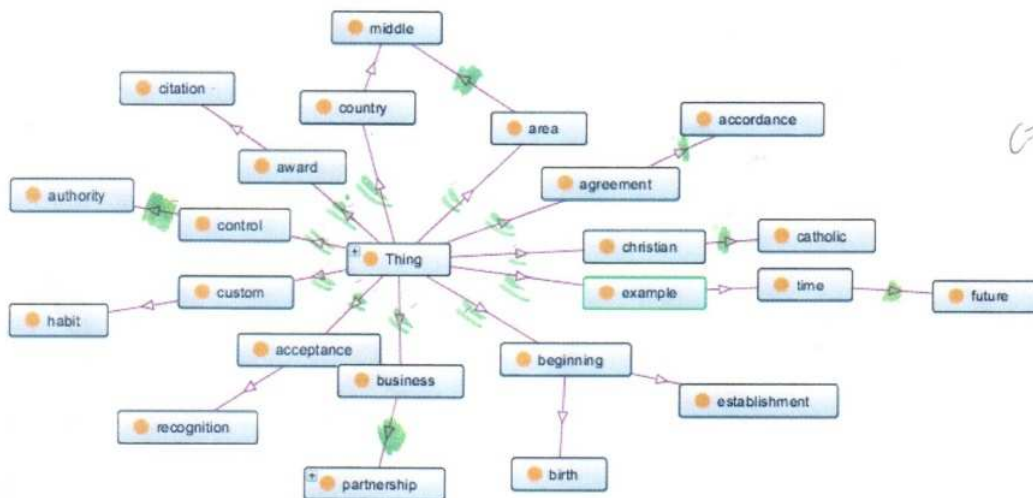
ANEXO A - Conjunto de Tags Penn Treebank

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

<i>Tag</i>	<i>Significado</i>
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

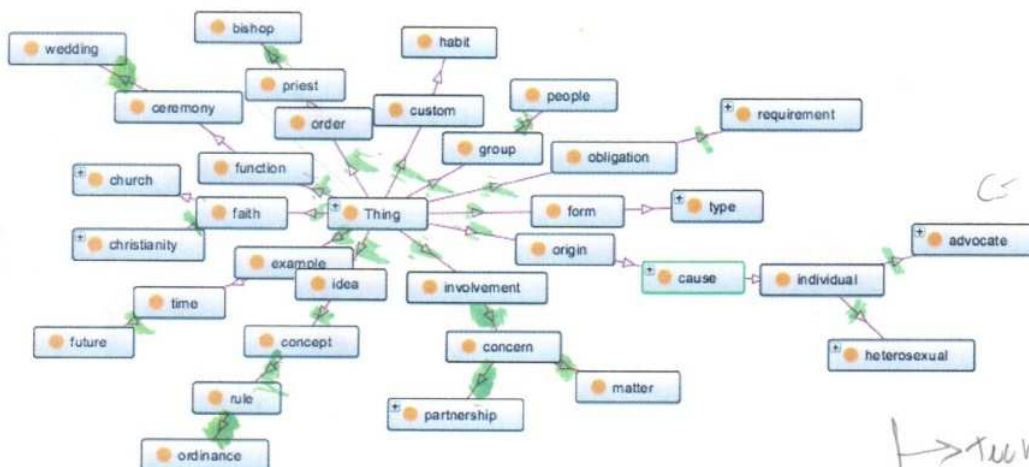
ANEXO B – Contagem manual das relações taxonômicas por especialista de domínio.





C-176

36



C-25

→ técnica
 ↳ Avaliações

119 10

Total = 119

CORRETAS = 93

ANEXO C – Corpus Family Law

Family law is an area of the law that deals with family-related issues and domestic relations including:

the nature of marriage, civil unions, and domestic partnerships;
 issues arising during marriage, including spousal abuse, legitimacy, adoption, surrogacy, child abuse, and child abduction
 the termination of the relationship and ancillary matters including divorce, annulment, property settlements, alimony, and parental responsibility orders (in the United States, child custody and visitation, child support and alimony awards).

This list is by no means dispositive of the potential issues that come through the family court system. In many jurisdictions in the United States, the family courts see the most crowded dockets. Litigants representative of all social and economic classes are parties within the system.

For the conflict of laws elements dealing with transnational and interstate issues, see marriage (conflict), divorce (conflict) and nullity (conflict).

Family Law may also refer to the marriage contract in Islamic faith, which includes the allowance of men to marry up to four wives.[1]

Members of the fathers' rights movement criticize the "win or lose" nature of family law in determining issues of divorce and child custody in many Western countries. Cross-national parties dealing with legal systems in different countries simultaneously grapple with substantive and procedural issues regarding child concerns.[2]

Advocates of Alimony reform also critique the Family Law system. They argue that current system pits divorcing couples against each other over child support and alimony, creating a hostile environment for the family and requiring large payments to divorce attorneys.[3][4]

Marriage is a social union or legal contract between individuals that creates kinship. It is an institution in which interpersonal relationships, usually intimate and sexual, are acknowledged in a variety of ways, depending on the culture or subculture in which it is found. Such a union

may also be called matrimony, while the ceremony that marks its beginning is usually called a wedding.

People marry for many reasons, most often including one or more of the following: legal, social, emotional, economical, spiritual, and religious. These might include arranged marriages, family obligations, the legal establishment of a nuclear family unit, the legal protection of children and public declaration of love.[1][2]

Marriage practices are very diverse across cultures, may take many forms, and are often formalized by a wedding.[3] The act of marriage usually creates normative or legal obligations between the individuals involved. In some societies these obligations also extend to certain family members of the married persons. Almost all cultures that recognize marriage also recognize adultery as a violation of the terms of marriage.[4]

External recognition can manifest in a variety of ways. Some examples include the state, a religious authority, or both. Marriage is often viewed as a contract. Civil marriage is the legal concept of marriage as a governmental institution irrespective of religious affiliation, in accordance with marriage laws of the jurisdiction. If recognized by the state, by the religion(s) to which the parties belong or by society in general, the act of marriage changes the personal and social status of the individuals who enter into it.

Anthropologists have proposed several competing definitions of marriage so as to encompass the wide variety of marital practices observed across cultures.[6] In his book *The History of Human Marriage* (1921), Edvard Westermarck defined marriage as "a more or less durable connection between male and female lasting beyond the mere act of propagation till after the birth of the offspring." [7] In *The Future of Marriage in Western Civilization* (1936), he rejected his earlier definition, instead provisionally defining marriage as "a relation of one or more men to one or more women that is recognised by custom or law". [8]

The anthropological handbook *Notes and Queries* (1951) defined marriage as "a union between a man and a woman such that children born to the woman are the recognized legitimate offspring of both partners." [9] In recognition of a practice by the Nuer of Sudan allowing women to act as a husband in certain circumstances, Kathleen Gough suggested modifying this to "a woman and one or more other persons." [10]

Edmund Leach criticized Gough's definition for being too restrictive in terms of recognized legitimate offspring and suggested that marriage be viewed in terms of the different types of rights it serves to establish. Leach expanded the definition and proposed that "Marriage is a relationship established between a woman and one or more other persons, which provides that a child born to the woman under circumstances not prohibited by the rules of the relationship, is accorded full birth-status rights common to normal members of his society or social stratum"[11] Leach argued that no one definition of marriage applied to all cultures. He offered a list of ten rights associated with marriage, including sexual monopoly and rights with respect to children, with specific rights differing across cultures.[12]

Duran Bell also criticized the legitimacy-based definition on the basis that some societies do not require marriage for legitimacy, arguing that in societies where illegitimacy means only that the mother is unmarried and has no other legal implications, a legitimacy-based definition of marriage is circular. He proposed defining marriage in terms of sexual access rights.[6]

Various definitions of marriage often include the concept of conjugal rights which may include mutual rights of companionship and support (consortium), co-habitation, joint property rights, and sexual relations. The concept of conjugal rights may be applied to "traditional" or non-traditional definitions of marriage. In June 2007, the California Department of Corrections announced it would allow same-sex conjugal visits. The policy was enacted to comply with a 2005 state law requiring state agencies to give the same rights to domestic partners that heterosexual couples receive.[13]

In the 21st century United States, it has become popular to describe marriages as either traditional or non-traditional. Traditional marriage has been conservatively described as being a life-long legal union between one man and one woman for the support of family and presumably offspring. Following the Anglican Book of Common Prayer, the marriage vows of many Christian denominations included "Till death do us part" as part of the marriage liturgy. With the advent of No-fault divorce in 1970, and the widespread acceptance of Birth control, the definition of traditional marriage has become less narrowly defined as a legal bond "between a man and a woman". The Defense of Marriage Act, DOMA, of 1996 uses this

definition. Non-traditional marriage assumes a conjugal relationship other than that assumed in the traditional definition, and most commonly refers to Same-sex marriage.

Although the institution of marriage pre-dates reliable recorded history, many cultures have legends concerning the origins of marriage. The way in which a marriage is conducted and its rules and ramifications has changed over time, as has the institution itself, depending on the culture or demographic of the time.[15]

One of the oldest known and recorded marriage laws is discerned from Hammurabi's Code, enacted in ancient Mesopotamia (widely considered as the cradle of civilization). Various cultures have had their own theories on the origin of marriage. One example may lie in a man's need for assurance as to paternity of his children. He might therefore be willing to pay a bride price or provide for a woman in exchange for exclusive sexual access.[16] Legitimacy is the consequence of this transaction rather than its motivation. In Comanche society, married women work harder, lose sexual freedom, and do not seem to obtain any benefit from marriage.[17] But nubile women are a source of jealousy and strife in the tribe, so they are given little choice other than to get married. "In almost all societies, access to women is institutionalized in some way so as to moderate the intensity of this competition." [18]

In English common law, a marriage was a voluntary contract by a man and a woman, in which by agreement they choose to become husband and wife.[19] Edvard Westermarck proposed that "the institution of marriage has probably developed out of a primeval habit".[20]

Forms of group marriage which involve more than one member of each sex, and therefore are not either polygyny or polyandry, have existed in history. However, these forms of marriage are extremely rare. Of the 250 societies reported by the American anthropologist George P. Murdock in 1949, only the Caingang of Brazil had any group marriages at all.[21]

For most of European history, marriage was more or less a business agreement between two families who arranged the marriages of their children. Romantic love, and even simple affection, were not considered essential.[22] Historically, the perceived necessity of marriage has been stressed.[23]

In Ancient Greece, no specific civil ceremony was required for the creation of a marriage - only mutual agreement and the fact that the couple must regard each other as husband and wife accordingly.[24] Men usually married when they were in their 20s or 30s [25] and expected their wives to be in their early teens. It has been suggested that these ages made sense for the Greek because men were generally done with military service by age 30, and marrying a young girl ensured her virginity.[26] Married Greek women had few rights in ancient Greek society and were expected to take care of the house and children.[27] Time was an important factor in Greek marriage. For example, there were superstitions that being married during a full moon was good luck and, according to Robert Flacelière, Greeks married in the winter.[26] Inheritance was more important than feelings: A woman whose father dies without male heirs can be forced to marry her nearest male relative—even if she has to divorce her husband first.[28]

There were several types of marriages in ancient Roman society. The traditional ("conventional") form called *conventio in manum* required a ceremony with witnesses and was also dissolved with a ceremony.[23] In this type of marriage, a woman lost her family rights of inheritance of her old family and gained them with her new one. She now was subject to the authority of her husband.[29] There was the free marriage known as *sine manu*. In this arrangement, the wife remained a member of her original family; she stayed under the authority of her father, kept her family rights of inheritance with her old family and did not gain any with the new family.[29] The minimum age of marriage for girls was 12.[30]

From the early Christian era (30 to 325 CE), marriage was thought of as primarily a private matter,[citation needed] with no uniform religious or other ceremony being required. However, bishop Ignatius of Antioch writing around 110 to bishop Polycarp of Smyrna exhorts, "[I]t becomes both men and women who marry, to form their union with the approval of the bishop, that their marriage may be according to God, and not after their own lust." [31]

In the 12th century women were obligated to take the name of their husbands and starting in the second half of the 16th century parental consent along with the churches consent was required for marriage .[32]

With few local exceptions, until 1545, Christian marriages in Europe were by mutual consent, declaration of intention to marry and upon the subsequent physical union of the parties.[33][34] The couple would promise verbally to each other that they would be married to each other; the presence of a priest or witnesses was not required.[35] This promise was

known as the "verbum." If freely given and made in the present tense (e.g., "I marry you"), it was unquestionably binding;[33] if made in the future tense ("I will marry you"), it would constitute a betrothal. One of the functions of churches from the Middle Ages was to register marriages, which was not obligatory. There was no state involvement in marriage and personal status, with these issues being adjudicated in ecclesiastical courts. During the Middle Ages marriages were arranged, sometimes as early as birth, and these early pledges to marry were often used to ensure treaties between different royal families, nobles, and heirs of fiefdoms. The church resisted these imposed unions, and increased the number of causes for nullification of these arrangements.[32] As Christianity spread during the roman period and the Middle Ages, the idea of free choice in selecting marriage partners increased and spread with it.[32]

The average age of marriage in the late 1200s into the 1500s was around 25 years of age.[36]

As part of the Counter-Reformation, in 1563 the Council of Trent decreed that a Roman Catholic marriage would be recognized only if the marriage ceremony was officiated by a priest with two witnesses. The Council also authorized a Catechism, issued in 1566, which defined marriage as, "The conjugal union of man and woman, contracted between two qualified persons, which obliges them to live together throughout life." [37]

In England, under the Anglican Church, marriage by consent and cohabitation was valid until the passage of Lord Hardwicke's Act in 1753. This act instituted certain requirements for marriage, including the performance of a religious ceremony observed by witnesses.[38]

As part of the Reformation, the role of recording marriages and setting the rules for marriage passed to the state. By the 1600s many of the Protestant European countries had a state involvement in marriage. As of 2000, the average marriage age range was 25-44 years for men and 22-39 years for women.

In the early modern period, John Calvin and his Protestant colleagues reformulated Christian marriage by enacting the Marriage Ordinance of Geneva, which imposed "The dual requirements of state registration and church consecration to constitute marriage"[37] for recognition.

In England and Wales, Lord Hardwicke's Marriage Act 1753 required a formal ceremony of marriage, thereby curtailing the practice of Fleet Marriage.[39] These were clandestine or irregular marriages performed at Fleet Prison, and at hundreds of other places. From the 1690s until the Marriage Act of 1753 as many as 300,000 clandestine marriages were performed at Fleet Prison alone.[40] The Act required a marriage ceremony to be officiated by an Anglican priest in the Anglican Church with two witnesses and registration. The Act did not apply to Jewish marriages or those of Quakers, whose marriages continued to be governed by their own customs.

In England and Wales, since 1837, civil marriages have been recognized as a legal alternative to church marriages under the Marriage Act of 1836. In Germany, civil marriages were recognized in 1875. This law permitted a declaration of the marriage before an official clerk of the civil administration, when both spouses affirm their will to marry, to constitute a legally recognized valid and effective marriage, and allowed an optional private clerical marriage ceremony.

The selection of a marriage partner may involve either the couple going through a selection process of courtship or the marriage may be arranged by the couple's parents or an outside party, a matchmaker.

A pragmatic (or 'arranged') marriage is made easier by formal procedures of family or group politics. A responsible authority sets up or encourages the marriage; they may, indeed, engage a professional matchmaker to find a suitable spouse for an unmarried person. The authority figure could be parents, family, a religious official, or a group consensus.

In some cases, the authority figure may choose a match for purposes other than marital harmony.

In rural Indian villages, child marriage is also practiced, with parents at times arranging the wedding, sometimes even before the child is born. This practice is now illegal under the Child Marriage Restraint Act.

In some societies ranging from Central Asia to the Caucasus to Africa, the custom of bride kidnapping still exists, in which a woman is captured by a man and his friends. Sometimes this covers an elopement, but sometimes it depends on sexual violence. In previous times, raptio was a larger-scale version of this, with groups of women captured by groups of men, sometimes

in war; the most famous example is The Rape of the Sabine Women, which provided the first citizens of Rome with their wives.

Other marriage partners are more or less imposed on an individual. For example, widow inheritance provides a widow with another man from her late husband's brothers.

A marriage is usually formalized at a wedding or marriage ceremony. The ceremony may be officiated either by a religious official, by a government official or by a state approved celebrant. In many European and some Latin American countries, any religious ceremony must be held separately from the required civil ceremony. Some countries - such as Belgium, Bulgaria, France, the Netherlands, Romania and Turkey[46] - require that a civil ceremony take place before any religious one. In some countries - notably the United States, Canada, the United Kingdom, the Republic of Ireland, Norway and Spain - both ceremonies can be held together; the officiant at the religious and civil ceremony also serving as agent of the state to perform the civil ceremony. To avoid any implication that the state is "recognizing" a religious marriage (which is prohibited in some countries) - the "civil" ceremony is said to be taking place at the same time as the religious ceremony. Often this involves simply signing a register during the religious ceremony. If the civil element of the religious ceremony is omitted, the marriage is not recognized by government under the law.

While some countries, such as Australia, permit marriages to be held in private and at any location, others, including England and Wales, require that the civil ceremony be conducted in a place open to the public and specially sanctioned by law. In England, the place of marriage need no longer be a church or register office, but could also be a hotel, historic building or other venue that has obtained the necessary license. An exception can be made in the case of marriage by special emergency license, which is normally granted only when one of the parties is terminally ill. Rules about where and when persons can marry vary from place to place. Some regulations require that one of the parties reside in the locality of the registry office.

Within the parameters set by the law of the jurisdiction in which a marriage or wedding takes place, each religious authority has rules for the manner in which weddings are to be conducted by their officials and members.

Marriage is an institution which can join together people's lives in a variety of emotional and economic ways. In many Western cultures, marriage

usually leads to the formation of a new household comprising the married couple, with the married couple living together in the same home, often sharing the same bed, but in some other cultures this is not the tradition.[47] Among the Minangkabau of West Sumatra, residency after marriage is matrilocal, with the husband moving into the household of his wife's mother.[48] Residency after marriage can also be patrilocal or avunculocal. Such marriages have also been increasingly common in modern Beijing. Guo Jianmei, director of the center for women's studies at Beijing University, told a Newsday correspondent, "Walking marriages reflect sweeping changes in Chinese society." [49] A similar arrangement in Saudi Arabia, called misyar marriage, also involves the husband and wife living separately but meeting regularly.[50]

Conversely, marriage is not a prerequisite for cohabitation. In some cases couples living together do not wish to be recognized as married, such as when pension or alimony rights are adversely affected, or because of taxation consideration, or because of immigration issues, and for many other reasons. In modern western societies some couples cohabit before marriage to test whether such an arrangement might work in the long term.

In some cases cohabitation may constitute a common-law marriage, and in some countries the laws recognize cohabitation in preference to the formality of marriage for taxation and social security benefits. This is the case, for example, in Australia.[51]

Adultery, also called philandery or infidelity, is a form of extramarital sex. It originally referred more specifically to sex between a woman who was married and a person other than her spouse.[1] The term "adultery" for many people carries a moral or religious association, while the term "extramarital sex" is morally or judgmentally neutral.

Adultery is illegal in some countries. The interaction between laws on adultery with those on rape has and does pose particular problems in societies that are especially sensitive to sexual relations by a married woman and men, such as some Muslim countries.[2] The difference between the offenses is that adultery is voluntary, while rape is not.

The term adultery has a Judeo-Christian origin, though the concept predates Judaism and is found in many other societies. Though the definition and consequences vary between religions, cultures, and legal jurisdictions, the concept is similar in Judaism, Christianity, and Islam. Hinduism has a similar concept.[3]

Historically, adultery has been considered to be a serious offense by many cultures. In some countries, adultery is a crime. However, even in jurisdictions where adultery is not itself a criminal offense, it may still have legal consequences, particularly in divorce cases. For example, where there is fault-based family law, it almost always constitutes grounds for divorce, it may be a factor to consider in a property settlement, it may affect the status of children, the custody of children, etc. Moreover, adultery can result in social ostracism in some parts of the world.

Three recent studies in the United States, using nationally representative samples, have found that about 10-15% of women and 20-25% of men admitted to having engaged in extramarital sex.[4][5][6]

In some East Asian countries, including Korea, and Taiwan, adultery continues to be a crime. In the Philippines, adultery (defined as consensual sexual intercourse between a married woman and a man who is not her husband) and a related act of concubinage (a man cohabiting with a woman who is not his wife), are considered crimes under the Revised Penal Code of the Philippines. Adultery is not a crime in China, but constitutes grounds for divorce.

In Pakistan, adultery is a crime under the Hudood Ordinance. The Ordinance sets a maximum penalty of death, although only imprisonment and corporal punishment have ever actually been imposed.[citation needed] The Ordinance has been particularly controversial because it requires a woman making an accusation of rape to provide extremely strong evidence to avoid being charged with adultery herself. A conviction of a man for rape is only possible with evidence from no less than four witnesses. In recent years high-profile rape cases in Pakistan have given the Ordinance more exposure than similar laws in other countries.[41] Similar laws exist in some other Muslim countries, such as Saudi Arabia.

In Indian law, adultery is defined as sex between a man and a woman without the consent of the woman's husband. The man is prosecutable and can be sentenced for up to five years (even if he himself was unmarried) whereas the married woman can not be jailed[42]. Men have called the law gender discrimination in that women cannot be prosecuted for adultery[43] and the National Commission of Women has criticized the British era law of being anti-feminist as it treats women as the property of their husbands and has consequentially recommended deletion of the law or reducing it to a civil offense. The Government is yet to act[44]. Extramarital sex without the

consent of one's partner can be a valid grounds for monetary penalty on government employees, as ruled by the Central Administrative Tribunal[45].

In Southwest Asia, adultery has attracted severe sanctions, including death penalty. In some places, such as Iran[46] and Saudi Arabia,[47] the method of punishment for adultery is stoning to death. It has been suggested that Iranian officials are avoiding imposing the penalty because of social objections. Proving adultery under Muslim law can be a very difficult task as it requires the accuser to produce four eye witnesses to the act of sexual intercourse, each of whom should have a good reputation for truthfulness and honesty. The criminal standards do not apply in the application of social and family consequences of adultery, where the standards of proof are not as exacting.

Most European countries have decriminalized adultery. Adultery is not a crime in most countries of the European Union, including Austria, the Netherlands, Belgium, Finland or Sweden. In some Southern-European countries, adultery can lead to the so called vendetta, which is illegal (with penalties up to life sentence), but rarely prosecuted.

Annulment is a legal procedure for declaring a marriage null and void. Unlike divorce, it is retroactive: an annulled marriage is considered to be invalid from the beginning almost as if it had never taken place. In strict legal terminology, (annulment) refers only to making a voidable marriage null; if the marriage is void ab initio, then it is automatically null, although a legal declaration of nullity is required to establish this. The process of obtaining such a declaration is similar to the annulment process. Generally speaking, annulment, despite its retrospective nature, still results in any children born being considered legitimate in the United States and many other countries.

A reason for annulment is called a diriment impediment to the marriage. Prohibitory impediments (which no longer exist in the Latin Code, CIC83) make entering a marriage wrong but do not invalidate the marriage, such as being betrothed to another person at the time of the wedding; diriment impediments, such as being brother and sister, or being married to another person at the time of the wedding, prevent such a marriage from being contracted at all. Such unions are called putative marriages.

Diriment impediments include:

Consanguinity

Insanity precluding ability to consent

Not intending, when marrying, to remain faithful to the spouse (simulation of consent)

One partner had been deceived by the other in order to obtain consent, and if the partner had been aware of the truth, would not have consented to marry

Abduction of a person, with the intent to compel them to marry (known as raptus), constitutes an impediment as long as they remain in the kidnapper's power.

Failure to adhere to requirements of canon law for marriages, such as clandestinity

the couple killed the spouse of one of them in order to be free to marry

the couple committed adultery, and one of the couple killed the spouse of one of them, in order to be free to marry

Some impediments can be dispensed, in which the Church exempts a couple, prior to the marriage, to the obligation to conform to the canon law. While some relationships can not have the impediment of consanguinity dispensed, a marriage can be sanctioned between cousins. This renders the marriage valid, and so non-annulable. Again, if an invalid marriage has been contracted, and the diriment impediment can be removed, a convalidation or sanatio in radice can be performed to make the marriage valid.

Divorce or dissolution of marriage is the final termination of marriage, canceling the legal duties and responsibilities of marriage and dissolving the bonds of matrimony between married persons. In most countries divorce requires the sanction of a judge or other authority in a legal process.

In western countries, a divorce does not declare a marriage null and void, as in an annulment, but divorce cancels the marital status of the parties. Where monogamy is law, this allows each partner to marry another. Where polygyny is legal, divorce allows the woman to marry another.

Divorce laws vary considerably around the world. Divorce is not permitted in some countries, such as in Malta and in the Philippines, though an annulment is permitted.

The legal process for divorce may also involve issues of spousal support, child custody, child support, distribution of property and division of debt, though these matters are usually only ancillary or consequential to the dissolution of the marriage.

Types of divorce

Though divorce laws vary between jurisdiction, there are two basic approaches to divorce: fault based and no-fault based. However, even in

some jurisdictions that do not require a party to claim fault of their partner, a court may still take into account the behavior of the parties when dividing property, debts, evaluating custody, and support.

Laws vary as to the waiting period before a divorce is effective. Also, residency requirements vary. However, issues of division of property are typically determined by the law of the jurisdiction in which the property is located.

[edit] No-fault divorce

Under a no-fault divorce system, the dissolution of a marriage does not require an allegation or proof of fault of either party.

The application can be made by either party or by both parties jointly.[where?] The same holds true for Canada[citation needed] and Germany.[citation needed]

Forty-nine states of the United States have adopted no-fault divorce laws, with grounds for divorce including incompatibility, irreconcilable differences, and irremediable breakdown of the marriage.[citation needed]

[edit] At-fault divorce

Prior to 1975, countries which permitted divorces also required proof by one party that the other party had committed an act incompatible to the marriage. This was termed "grounds" for divorce (popularly called "fault") and was the only way to terminate a marriage. Most jurisdictions around the world still require such proof of fault. In the United States, only New York state still requires fault for a divorce.

Fault-based divorces can be contested and may involve allegations of collusion of the parties, or condonation, connivance, or provocation by the other party. Contested fault divorces can be expensive, and not usually practical as eventually most divorces are granted. Comparative rectitude is a doctrine used to determine which spouse is more at fault when both spouses are guilty of breaches.

Cohabitation is an arrangement whereby two people decide to live together on a longterm or permanent basis in an emotionally and/or sexually intimate relationship. The term is most frequently applied to couples who are not married.

Today, cohabitation is a common pattern among people in the Western world. People may live together for a number of reasons. These may include wanting

to test compatibility or to establish financial security before marrying. It may also be because they are unable to legally marry, because for example same-sex, some interracial or interreligious marriages are not legal or permitted. Other reasons include living with someone before marriage in an effort to avoid divorce, a way for polygamists or polyamorists to avoid breaking the law, a way to avoid the higher income taxes paid by some two-income married couples (in the United States), negative effects on pension payments (among older people), and philosophical opposition to the institution of marriage and seeing little difference between the commitment to live together and the commitment to marriage. Some individuals also may choose cohabitation because they see their relationships as being private and personal matters, and not to be controlled by political, religious or patriarchal institutions.

Some couples prefer cohabitation because it does not legally commit them for an extended period, and because it is easier to establish and dissolve without the legal costs often associated with a divorce. In some jurisdictions cohabitation can be viewed legally as common-law marriages, either after the duration of a specified period, or the birth of the couple's child, or if the couple consider and behave accordingly as husband and wife. (This helps provide the surviving partner a legal basis for inheriting the deceased's belongings in the event of the death of their cohabiting partner.)

In the Western world, a man and a woman who lived together without being married were once socially shunned and persecuted and potentially prosecuted by law. In some jurisdictions, cohabitation was illegal until relatively recently. Other jurisdictions have created a Common-law marriage status when two people of the opposite sex live together for a prescribed period of time. Most jurisdictions no longer prosecute this choice.

Opposition to cohabitation comes mainly from religious groups, but also some factions of feminists as well.[citation needed] Opponents of cohabitation usually argue that living together in this fashion is less stable and hence harmful. According to one argument, the total and unconditional commitment of marriage strengthens a couple's bond and makes the partners feel more secure, more relaxed, and happier than those that have chosen cohabitation.[1] Opponents of cohabitation commonly cite statistics that indicate that couples who have lived together before marriage are more likely to divorce, and that unhappiness, ill health, poverty, and domestic violence are more common in unmarried couples than in married ones.[2] Cohabitation advocates, in turn, cite limited research

that either disproves these claims or indicates that the statistical differences are due to other factors than the fact of cohabitation itself.[3]

A civil union is a legally recognized union similar to marriage. Beginning with Denmark in 1989, civil unions under one name or another have been established by law in many developed countries in order to provide same-sex couples with rights, benefits, and responsibilities similar (in some countries, identical) to opposite-sex civil marriage. In some jurisdictions, such as Quebec, New Zealand, and Uruguay, civil unions are also open to opposite-sex couples.

Most civil-union countries recognize foreign unions if those are essentially equivalent to their own; for example, the United Kingdom lists equivalent unions in Civil Partnership Act Schedule 20.

Supporters of civil unions contend that civil unions grant same-sex couples equal rights to married couples.[1] Some commentators, such as Ian Ayres, are critical of civil unions because they say they represent a separate status unequal to marriage.[2][3] Others, such as Robert Knight, are critical because they say civil unions endow the same rights and privileges of heterosexual marriages – alleging that they allow same-sex marriage by using a different name.[4]

The terms used to designate recognized same-sex unions are not standardized, and vary widely from country to country. Government-sanctioned relationships that may be similar or equivalent to civil unions include civil partnerships, registered partnerships, domestic partnerships, significant relationships, reciprocal beneficiary relationships, common-law marriage, adult interdependent relationships, life partnerships, stable unions, civil solidarity pacts, and so on. The exact level of rights, benefits, obligations, and responsibilities also varies, depending on the laws of a particular country. Some jurisdictions allow same-sex couples to adopt, while others forbid them to do so, or allow adoption only in specified circumstances.

As used in the United States, beginning with the state of Vermont in 2000, the term civil union has connoted a status equivalent to marriage for same-sex couples; domestic partnership, offered by some states, counties, cities, and employers since as early as 1985,[5] has generally connoted a lesser status with fewer benefits.[6] However, the legislatures of the West Coast states of California, Oregon and Washington have preferred the term

domestic partnership for enactments similar or equivalent to civil union laws in East Coast states.

Civil unions are not seen as a replacement for marriage by many in the gay community. "Marriage in the United States is a civil union; but a civil union, as it has come to be called, is not marriage," said Evan Wolfson of Freedom to Marry.[7] "It is a proposed hypothetical legal mechanism, since it doesn't exist in most places, to give some of the protections but also withhold something precious from gay people. There's no good reason to do that." However, some supporters of traditional marriage view the matter differently; Randy Thomasson, Executive Director of the Campaign for California Families, calls civil unions "homosexual marriage by another name" and contends that civil unions provide same-sex couples "all the rights of marriage available under state law." [4] The California Supreme Court, in the *In Re Marriage Cases* decision, noted nine differences [8] in state law.

A domestic partnership is a legal or personal relationship between two individuals who live together and share a common domestic life but are neither joined by marriage nor a civil union. However, in some jurisdictions, such as Australia, New Zealand, the United States of Oregon, Washington, Nevada, and California, a domestic partnership is almost equivalent to marriage, or to other legally recognized same-sex or different-sex unions. The terminology for such unions is still evolving, and the exact level of rights and responsibilities conferred by a domestic partnership varies widely from place to place.

Some legislatures have voluntarily established domestic partnership relations by statute instead of being ordered to do so by a court. Although some jurisdictions have instituted domestic partnerships as a way to recognize same-sex unions, domestic partnerships may involve either different-sex or same-sex couples.

In some legal jurisdictions, domestic partners (or 'domparts') who live together for an extended period of time but are not legally entitled to common-law marriage may be entitled to legal protection in the form of a domestic partnership. Some domestic partners may enter into domestic partnership agreements in order to agree contractually to issues involving property ownership, support obligations, and similar issues common to marriage. (See effects of marriage and palimony.)

One of the purposes of domestic partnership relation is to recognize the contribution of one partner to the property of the other. In the common law, devices such as the constructive trust are available to protect spouses in legal or common-law marriages. In civil law jurisdictions, such trusts are generally not available, prompting courts to find alternative ways to protect the partner who contributes to the other's property.