UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2007.2

Aprendizagem de Ontologias a partir de Texto

TRABALHO DE GRADUAÇÃO

Aluno: Zinaldo Araujo Barros Jr (zabj@cin.ufpe.br)

Orientador: Frederico Luiz Gonçalves de Freitas (fred@cin.ufpe.br)

29 de Janeiro de 2008

Agradecimentos

Agradeço primeiramente a Deus pelo suporte espiritual e por manter minha sanidade mental. Agradeço a meus pais e familiares pelo apoio financeiro e emocional em todos os instantes da minha árdua caminhada. Não posso esquecer os laços de amizade criados na graduação e estreitados fora dela. Gabriel, meu irmão, amigo de felizes momentos de descontração fora do CIn, também devo a ele os poucos momentos em que estive presente as aulas, uma vez que ele foi o grande responsável pelo meu transporte esses anos todos. Também não posso me esquecer do meu amigo Roberto, e de sua vodka batizada. Não posso deixar de citar Firma, pela sua paciência, ou falta dela, não sei ao certo, no desenvolvimento dos projetos das disciplinas, o que importa é que a metodologia dele é eficaz.

Agradeço aos meus chefes, Jairson e Alessandro, pela compreensão nas situações em que a vida acadêmica atrapalhava as atividades profissionais. Agradeço também ao meu orientador Frederico Freitas, pela oportunidade em desenvolver este trabalho. Infelizmente não posso citar todos, mas fica aqui minha gratidão àqueles que de alguma forma contribuíram para conclusão do curso.

Sumário

Lista de Figuras

Lista de Quadros

1. Introdução

A necessidade da existência de uma representação de conhecimento que permita a automatização do raciocínio intensificou o desenvolvimento de tecnologias associadas à manipulação e construção de ontologias nos últimos anos. Ferramentas capazes de dar suporte a construção de ontologias, ou sistemas de aprendizagem automática de ontologia são essenciais, uma vez que o processo de construção manual de ontologias é uma tarefa extremamente custosa, que demanda bastante tempo, tediosa e passível de erros. Ontologias construídas de forma manual, ainda correm o risco de serem parciais, influenciadas de alguma forma pelo engenheiro que as cria.

Este trabalho define um modelo de aprendizagem de ontologias a partir de texto no domínio da documentação presente em arquivos de código fonte JAVA aprendendo conceitos de duas hierarquias, para isso, se utiliza de uma abordagem híbrida de aprendizagem baseada em processamento lingüístico e casamento de padrões. O modelo também é instanciado no GATE [Bontcheva & Cunningham 2002], uma ferramenta de suporte a processamento de linguagem natural, que dá suporte a todas as técnicas empregadas no modelo proposto e as abordagens de avaliação eleitas.

O restante deste capítulo apresenta duas definições de ontologia: uma de origem filosófica e outra mais adequada ao seu emprego na computação. Trata também de motivar o leitor por meio das possibilidades em termo das aplicações de ontologias em sistemas computacionais e ainda aborda de forma sucinta o problema a ser resolvido por esse trabalho.

1.1. Motivação

A motivação para o emprego de esforços na direção de encontrar uma solução definitiva para o problema de aprendizagem de ontologias a partir de texto pode ser encontrada nas várias aplicações que podem se beneficiar com o surgimento de tal solução.

1.1.1. Web Semântica

A Web Semântica promete um suporte nunca visto a colaboração automática. Colaboração tem a ver com cooperação de indivíduos com a finalidade atingir objetivos complexos. Uma configuração básica dessa colaboração automática prometida pela Web Semântica pode ser vista em agentes de software servindo como representantes de indivíduos do mundo real interagindo de maneira cooperativa para atingir os objetivos de seus donos como já exemplificado por [Berners-Lee et al 2001]. Esses agentes devem ser capazes de acessar e trocar informação com anotações semânticas pela Web, utilizando, por exemplo, Web Services como um meio facilitador. A resolução automática de tarefas por parte desses agentes, representantes de entidades do mundo real, está condicionada a disponibilidade de ontologias capazes de prover tais anotações semânticas a informação a ser processada e trocada por eles.

1.1.2. Recuperação de informação

É possível e recuperar informação baseada em seu significado. Sistemas que tiram proveito de aspectos semânticos podem, por exemplo, se valer de equivalência de conceitos, considere o cenário, onde, por exemplo, um usuário de um engenho de busca dispara uma consulta por "George W. Bush", mas também espera documentos que não façam referências explicitas ao termo pesquisado, mas que tratem de alguma forma sobre o presidente dos Estados Unidos da América. A semântica permite enriquecer a forma como a informação é apresentada, no nosso cenário, o mesmo engenho de busca pode exibir os documentos clusterizados de acordo com seu significado, ao invés, por exemplo, de listá-los em qualquer ordem. Ainda, com o poder semântico provido pela utilização de ontologias não seria impossível imaginar que o engenho fosse capaz de consolidar informação de todos os documentos relevantes excluindo possíveis redundâncias e resumindo o que fosse apropriado. Tudo isso só é permitido por meio de eficiente processamento computacional, que consiste, basicamente, em inferir a partir de conhecimento existente, novos conhecimentos.

1.1.3. Integração de dados

Em uma organização, por exemplo, o uso de ontologias também pode exercer papel importante na integração de informação oriunda de fontes heterogêneas, seja dentro da própria organização ou externamente. Tipicamente, diferentes esquemas são usados para descrever e classificar informação e essa diferença se estende também para a terminologia contida na própria informação, por exemplo, para um determinado sistema, a noção de "computador", existe, para um outro, a mesma entidade do mundo real é modelada como "PC". Com a criação de mapeamentos entre esses esquemas pode-se criar uma visão unificada e alcançar interoperabilidade entre os processos que usam tal informação. O que implica também que ontologias favorecem o reuso de informação.

1.1.4. Web Services

A semântica tem sido explorada fortemente na internet. Descrições semânticas têm sido aplicadas a Web Services. Uma vez que um determinado serviço é anotado semanticamente, ele pode ser descoberto com mais facilidade e precisão, pois uma vez munido de um mecanismo de busca semântica, o usuário, ou até mesmo um computador, não está limitado a resultados simplesmente filtrados por ocorrências de palavras chaves, mas por conceitos e relações entre esses conceitos. Ainda, com tais descrições semânticas disponíveis é possível criar Web Services completamente novos de mera composição de diferentes Web Services existentes [Mahmoud 2005].

1.2. Ontologia

Em seu sentido filosófico, trata-se de um termo relativamente novo, introduzido com o objetivo de distinguir o estudo do ser como tal. O Dicionário Oxford de Filosofia define ontologia como "[...] o termo derivado da palavra grega que significa 'ser', mas usado desde o século XVII para denominar o ramo da metafísica que diz respeito àquilo que existe" [Blackburn 1996].

De forma simples, para elaborar ontologias, definem-se categorias para as coisas que existem em um mesmo domínio. Ontologia é um "catálogo de tipos de coisas" em que se supõe existir um domínio, na perspectiva de uma pessoa que usa uma determinada linguagem [Sowa 1999]. Trata-se de "uma teoria que diz respeito a tipos de entidades e, especificamente, a tipos de entidades abstratas que são aceitas em um sistema com uma linguagem" [Gove 2002].

1.3. Aprendizagem de Ontologias

Aprendizagem de ontologias está relacionada à aquisição de conhecimento, muito do trabalho desenvolvido nesse campo, portanto, está na direção de áreas como processamento de linguagem natural, inteligência artificial, e aprendizagem de máquina. Dessa forma, surge a legítima pergunta se a roda não está sendo reinventada. Seria aprendizagem de ontologias meramente rotular idéias e técnicas existentes com um novo nome? A resposta a essa questão deve ser: não.

Embora o objetivo de aquisição de conhecimento e a aprendizagem de ontologias a partir de texto são os mesmos, em essência, a aquisição de conhecimento implicitamente contido em dados textuais, existe, no entanto, alguns aspectos inovadores a aprendizagem de ontologias que a diferencia de muito trabalho desenvolvido na aquisição de conhecimento [Cimiano et al 2005].

Aprendizagem de ontologias está relacionada a extrair elementos lógicos (conhecimento conceitual) de dados de entrada e construir uma ontologia a partir dela. Construção manual de ontologias é uma tarefa custosa e que demanda bastante tempo, tediosa e passível de erros. Ontologias construídas de forma manual são caras, influenciadas pelo seu desenvolvedor, não flexíveis a mudanças e específicas ao propósito para o qual foram construídas. Automação da construção de ontologias não somente elimina os custos, mas também resultam em um melhor casamento da ontologia com as aplicações que a utilizarão.

1.4. Resumo do documento

Além deste capítulo inicial que apresenta uma breve introdução do trabalho, definição de ontologia, aplicação de ontologias em sistemas computacionais como motivação para a área e uma descrição do problema a ser tratado pelo trabalho este documento ainda é composto de mais cinco capítulos:

O capítulo 2 introduz com um pouco mais de profundidade a tarefa aprendizagem de ontologias a partir de texto, dá uma definição mais rigorosa de ontologias, e apresenta as ferramentas de suporte e sistemas de aprendizagem bem como suas classificações em várias categorias.

O capítulo 3 aborda as ferramentas do método de aprendizagem adotado, ou seja, técnicas de pré-processamento utilizadas e padrões de casamento, o capítulo também apresenta as ferramentas para a construção e avaliação da ontologia utilizadas para instanciar o método proposto.

O capítulo 4 apresenta o domínio explorado além do método que foi implementado para aprendizagem de ontologias a partir de texto que vai da formação do corpus a fase de pruning da ontologia extraída.

O capítulo 5 começa descrevendo as diferentes formas de avaliação: performance de extração; avaliação de especialista; adequação a tarefa, além de apresentar os resultados do estudo de caso para duas dessas formas de avaliação.

Por fim, o capítulo 6 apresenta as conclusões e uma proposta de trabalhos futuros.

2. Aprendizagem de ontologias a partir de texto

Aprendizagem de ontologias é uma área essencialmente multidisciplinar devido a sua forte conexão com a Web Semântica, que tem atraído pesquisadores de uma larga variedade de disciplinas: representação de conhecimento, filosofia, banco de dados, aprendizagem de máquina, processamento de linguagem natural, processamento de imagem, etc. Em conseqüência, aprendizagem de ontologias tem se beneficiado de uma grande troca de idéias e técnicas que de certa forma construíram uma visão diferente da de aquisição de conhecimento.

Aprendizagem de ontologias, no contexto da Web Semântica, é antes de mais nada relacionada com aquisição de conhecimento de e para o conteúdo Web, dessa forma se distancia de pequenas e homogêneas coleções de dados para se importar com a enorme heterogeneidade dos dados da Web.

Uma vez que, muito do trabalho desenvolvido em aprendizagem de ontologias é oriundo do aprendizado de máquina, métodos rigorosos que são centrais ao aprendizado de máquina estão sendo rapidamente adaptados para a área. Dessa forma, a aprendizagem de ontologias será impactada por esforços para sistematicamente avaliar e comparar abordagens em tarefas e medidas de avaliação bem definidas, fazendo da aprendizagem de ontologias uma área altamente desafiadora onde somente abordagens demonstráveis e competitivas irão sobreviver.

Neste capítulo daremos uma definição formal de ontologia e de suas classificações, também serão apresentadas as categorias existentes de ferramentas de suporte a construção de ontologias e sistemas de aprendizagem bem como uma seleção desses sistemas e ferramentas. Finalmente, são feitas algumas considerações a respeito do que foi apresentado no capítulo.

2.1. Ontologia

Uma ontologia é uma especificação explícita e formal de uma conceitualização compartilhada de um domínio de interesse [Gruber 1993], onde formal implica que a ontologia deve ser compreensível por um computador e compartilhada significa que é consensual, aceita por um grupo ou comunidade e não apenas por um indivíduo. Além disto, a ontologia deve estar restrita a um dado domínio de interesse, dessa forma, modelar conceitos e relacionamentos que são relevantes a uma tarefa particular ou a um domínio de aplicação.

O termo conceitualização corresponde a uma coleção de objetos, conceitos e outras entidades que se assume existirem em um domínio e os relacionamentos entre eles [Genesereth & Nilsson 1987].

Uma conceitualização é uma visão abstrata e simplificada do mundo que se deseja representar.

Assim como em um modelo relacional as ontologias formalizam os aspectos intencionais de um domínio, que o estruturam, enquanto que a extensão é provida pela base de conhecimento que contem declarações sobre instancias de conceitos e relações definidas pela própria ontologia.

Staab [Staab & Studer 2004] apresenta uma definição mais formal para ontologia: “Ontologias consistem de conceitos (classes), relações (propriedades), instâncias e axiomas, daí uma definição mais sucinta de uma ontologia é uma 4-tupla (C, R, I, A), onde C é um conjunto de conceitos, R, um conjunto de relações, I, um conjunto de instancias, e A, um conjunto de axiomas.”

A seguir as ontologias são classificadas em três tipos básicos: Ontologias Formais; Ontologias Baseadas em Protótipos e Ontologias Terminológicas.

2.1.1. Tipos de Ontologias

Sowa [Sowa 2003] propõe a classificação das ontologias em três tipos:

2.1.1.1. Ontologias Formais

São declaradas em lógica ou alguma linguagem computacional que pode ser automaticamente traduzida para a lógica, os axiomas e definições dão suporte a inferências mais complexas. A figura 1 [Biemann 2005] ilustra um exemplo de Ontologia Formal.

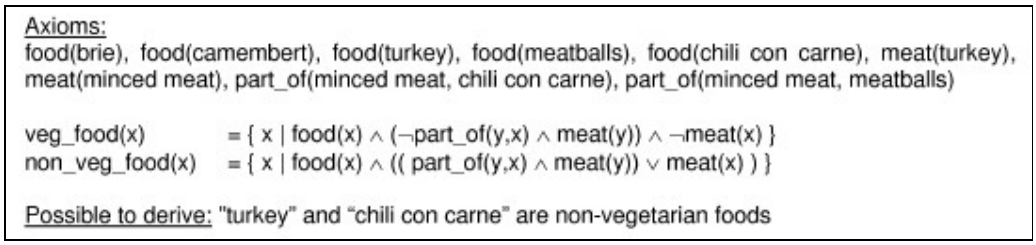

Figura 1: Exemplo de Ontologia Formal

2.1.1.2. Ontologias Baseadas em Protótipos

São distinguidas por instâncias típicas ou protótipos, ao invés dos axiomas e definições em lógica das ontologias formais. Categorias são formadas de forma extensional, ou seja, por coleções de instancias ao invés de forma intencional, pela descrição de todas as instâncias possíveis. Para seleção dessas instâncias, alguma métrica de similaridade de instâncias deve ser definida. A figura 2 [Biemann 2005] ilustra um exemplo de Ontologia Baseada em Protótipos.

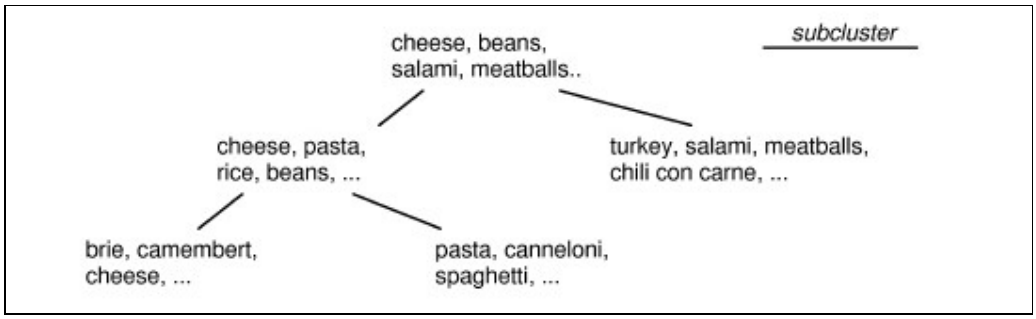

Figura 2: Exemplo de Ontologia Baseada em Protótipo

2.1.1.3. Ontologias Terminológicas

É o caso de ontologias cujas categorias não precisam ser totalmente especificadas por axiomas e definições como nas ontologias formais, mas são parcialmente especificadas por relações tais como subtipo/supertipo ou parte/todo, que determinam a posição relativa dos conceitos em relação aos outros, mas que não os definem completamente. Ainda segundo Sowa, a diferença entre uma ontologia

terminológica e uma ontologia formal é mais de grau do que de conteúdo: quanto mais axiomas são adicionados a uma ontologia terminológica mais próxima ela fica de uma ontologia formal. A figura 3 [Biemann 2005] ilustra um exemplo de Ontologia Terminológica.

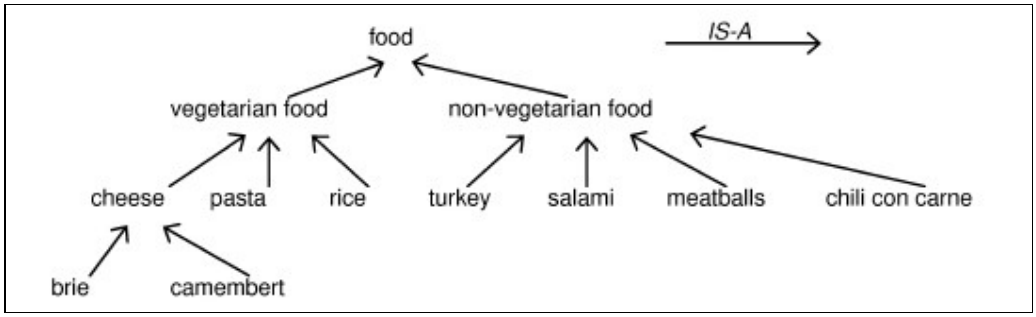

Figura 3: Exemplo de Ontologia Terminológica

2.2. Categorias de sistemas de aprendizagem de ontologias

Sistemas de aprendizagem de ontologias podem ser classificados em seis categorias (dimensões) e em algumas subcategorias [Mehrnoush & Barforoush 2003]:

- **Elementos aprendidos** (conceitos, relações, axiomas, regras, instâncias, categorias sintáticas, e papeis temáticos);
- **Ponto de partida** (conhecimento prévio, e tipo e língua de entrada);
- **Pré-processamento** (processamento lingüístico, tais como compreensão aprofundada ou processamento superficial de texto);
- **Método de aprendizagem** consistindo de:
  - **Categoria de aprendizagem** (supervisionado x não supervisionado, online x offline);
  - **Abordagem de aprendizagem** (estatística x simbólica, lógica, baseada em lingüística, casamento de padrão, guiada por template, métodos híbridos);
  - **Tarefa de aprendizagem** (classificação, clusterização, aprendizagem de regras, formação de conceitos, população de ontologias);
  - **Grau de automação** (manual, semi-automática, cooperativa, completamente automática, tipo e quantidade de intervenção do usuário)
- **Resultado** (ontologia x estruturas intermediárias, no primeiro caso, as características da ontologia construída, tais como grau de cobertura, uso e propósito, tipo de conteúdo, estrutura, topologia e linguagem de representação);
- **Métodos de avaliação** (avaliação do método de aprendizagem ou avaliação da ontologia resultante)

2.2.1. Elementos aprendidos

Os elementos aprendidos podem ser apenas conhecimento ontológico ou léxico também. Os principais elementos léxicos, aprendidos pelos sistemas, são palavras, e os elementos ontológicos são conceitos, relações e axiomas. Existem

alguns sistemas que aprendem meta conhecimento de como extrair conhecimento ontológico da entrada.

2.2.1.1. Palavras

Embora a maioria dos sistemas de aprendizagem use léxicos pré-definidos, alguns deles aprendem conhecimento léxico sobre palavras também. A forma como tratar palavras desconhecidas e o tipo de conhecimento léxico a ser aprendido sobre palavras são diferentes em diferentes sistemas de aprendizagem. Por exemplo, o SYNDIKATE, adivinha as classes de palavras para itens léxicos desconhecidos a partir de uma hierarquia que cobre todas as classes relevantes de palavras para um idioma particular.

2.2.1.2. Conceitos

Um conceito pode ser qualquer coisa que é dita sobre algo, pode ser abstrato ou concreto, elementar ou composto, real ou fictício, a descrição de uma tarefa, função, ação, estratégia, processo de raciocínio. Conceitos são representados por nós em grafos de ontologia e podem ser aprendidos por sistemas de aprendizagem. Eles podem ser extraídos da entrada ou serem criados durante uma fase de refinamento a partir de outros conceitos. Em outras palavras eles podem ou não ter elementos correspondentes na entrada. Em uma aquisição terminológica (baseada em termos) de conceitos, um nó do grafo (conceito) será criado em correspondência a um termo extraído que podem ser palavras ou mesmo frases de um texto, enquanto que aquisição conceitual (baseada em semântica) que é usualmente feita na fase de refinamento, o conceito será construído de acordo com suas características (atributos / valores), sua funcionalidade, etc. Ou seja, pode não ter nenhum correspondente na entrada (nenhuma palavra correspondente ou frase no texto de entrada).

2.2.1.3. Instâncias

Alguns sistemas de aprendizagem de ontologias aproveitam-se de alguma ontologia existente e simplesmente a populam com instâncias de classes e relações. Muitos desses sistemas não aprendem novos conceitos (classes) e apenas aprendem instâncias de classes existentes.

2.2.1.4. Relações entre conceitos

Relações podem ser estudadas de duas formas:

- Uma relação como um nó na ontologia, então a relação é tratada como um conceito e pode ser aprendida como outros conceitos;
- Uma relação se liga a dois ou mais conceitos e então ela deve ser aprendida como um subconjunto de um produto de n conceitos ($n > 1$).

2.2.1.5. Relações Taxonômicas (É-UM)

Taxonomias são largamente utilizadas para organizar conhecimento ontológico através do uso de relações de generalização/especialização. Conhecimento taxonômico é aprendido por alguns sistemas de aprendizagem.

2.2.1.6. Relações não taxonômicas

Relações não taxonômicas entre conceitos referem-se a qualquer relação entre conceitos exceto as relações É-UM.

2.2.1.7. Axiomas

Axiomas são usados para modelar sentenças que são sempre verdade. Eles podem ser incluídos em ontologias para muitos propósitos, tais como restringir a informação contida em uma ontologia, verificar sua corretude ou deduzir novas informações.

2.2.1.8. Meta conhecimento

Além de sistemas que aprendem conhecimento ontológico, existem sistemas que aprendem como extrair conhecimento ontológico. Eles aprendem meta conhecimento tais como regras para extrair instâncias ou padrões para extrair conhecimento a partir de textos.

2.2.2. Ponto de partida

Sistemas de aprendizagem de ontologias utilizam conhecimento a priori e adquirem novo conhecimento a partir dos textos de entrada. A qualidade e quantidade desse conhecimento prévio, e o tipo, estrutura e idioma dos textos de entrada a partir dos quais o sistema aprende diferem de um sistema para o outro. Esse Conhecimento prévio varia tanto em tipo como em volume para diferentes sistemas.

O conhecimento pode se representado por lingüística (lexical, gramatical, templates, etc) ou representado por recursos ontológicos. Em muitos sistemas existe um léxico pré-definido usado para processar textos. O tamanho e cobertura da ontologia base é outro fator que distingui e que pode variar desde uma quase vazia, a exemplo da ontologia base do HASTI, ou uma ontologia fornecida pelo usuário, pequeno número de palavras sementes que representam conceitos, até ontologias gigantes de senso comum como o Cyc [Lenat 1995].

2.2.3. Pré-processamento

O pré-processamento mais popular utilizado em extração de ontologias é o processamento lingüístico. Deep understanding pode fornecer relações específicas entre conceitos enquanto que técnicas superficiais podem fornecer conhecimento genérico a respeito de conceitos. Como o deep understanding usualmente diminui a velocidade do processo de construção de ontologia a maioria dos sistemas existentes usam técnicas superficiais de processamento de texto tais como tokenização, part-of-speech (POS) tagging, análise sintática entre outras para extrair estruturas essenciais a partir de textos de entrada.

2.2.4. Métodos de aprendizagem

Métodos de extração de conhecimento se subdividem em abordagens estatísticas e simbólicas. Das abordagens simbólicas o texto tratará da abordagem lógica, baseada em lingüística e guiadas por template. Métodos eurísticos podem ser usados para facilitar cada abordagem. Existem também abordagens híbridas, que combina duas ou mais das abordagens mencionadas acima como forma de empregar seus benefícios e eliminar suas limitações.

2.2.4.1. Estatística

Nesta abordagem, análise estatística será executada em dados coletados dos textos de entrada, é baseada na freqüência de ocorrência de palavras e sua distribuição pelo corpus. O WEB→KB, por exemplo, é uma sistema que classifica documentos da web baseado em um modelo probabilístico que classes rotuladas por dados de treinamento que servem para classificar páginas nunca vistas com a classe mais provável dada a evidência de palavras que descrevem essa página.

2.2.4.2. Lógico

Métodos de aprendizagem baseados em lógica podem descobrir novo conhecimento por dedução ou indução e representam conhecimento por proposições, lógica de primeira ou mais alta ordem. Sistemas baseados em dedução, tais como o HASTI, exploram regras de dedução e inferência tais como resolução para deduzir novo conhecimento a partir de conhecimento existente enquanto que sistema baseados em indução, induzem hipóteses a partir de observações (exemplos) e sintetizam novo conhecimento a partir de experiência.

2.2.4.2.1. Lingüístico

Abordagens lingüísticas são usadas para extrair conhecimento ontológico de textos.

Em sua maioria são dependentes de idioma e usualmente executam pré-processamento no texto de entrada para extrair estruturas que servem de base para construção das ontologias.

2.2.4.2.2. Baseado em padrões

Casamento de padrão é uma abordagem amplamente utilizada no campo de extração de informação e também é herdada pela aprendizagem de ontologias. Os padrões podem ser gerais, tais como os usados pelo HASTI ou não específicos, dependentes de um domínio / aplicação particular a exemplo dos usados por Assadi (1999).

2.2.4.2.3. Heurístico (AD HOC)

Heurísticas podem ser usadas em combinação com qualquer uma das abordagens já citadas. Em outras palavras, métodos heurísticos não são independentes e completos, eles devem ser usados para assistir outras abordagens. O Text-to-Onto,

por exemplo, usa uma heurística que relaciona todos os conceitos contidos no corpo de um documento HTML com os conceitos presentes no titulo desse documento.

2.2.4.2.4. Multi estratégia

Muitos sistemas, que aprendem mais de um tipo de elemento de ontologias, usam abordagens combinadas. Eles aplicam aprendizado multi estratégia para aprender diferentes componentes da ontologia utilizando-se de diferentes algoritmos, por exemplo, o HASTI aplica uma combinação de abordagens lógicas, baseadas em linguística e em padrão, além de métodos heurísticos.

2.2.5. Tarefa de aprendizagem

Métodos de aprendizagem podem ser categorizados com base na tarefa que desempenham. Nesta categoria, classificação, clusterização, aprendizagem de regras, análise formal de conceitos, e população de ontologias são algumas das tarefas, que pode ser feitas em cada uma das abordagens já mencionadas.

2.2.6. Grau de automação

A fase de aquisição de conhecimento pode ser executada de forma manual a completamente automática. Em sistemas de aprendizagem semi-automática e cooperativa, o papel do usuário pode variar bastante, mas basicamente ele pode propor uma ontologia inicial e validar/mudar versões diferentes propostas pelo sistema.

2.2.7. Resultado

O primeiro passo para essa categorização é distinguir aprendizagem de ontologias suporte a construção de ontologia. Muitos dos sistemas apresentados aqui aprendem ontologias (estruturas ontológicas), mas alguns deles apenas suportam usuários, especialistas e outros sistemas a aprenderem ontologias. Em outras palavras alguns sistemas são sistemas autônomos de aprendizagem enquanto outros são módulos que executam uma tarefa e resultam em um conjunto de dados intermediários que serão usados para construir a ontologia.

2.2.8. Avaliação

Encontrar métodos formais e padronizados para avaliar sistemas de aprendizagem de ontologias é um problema em aberto. Para avaliação de tais sistemas existem duas abordagens:

- Avaliar métodos de aprendizagem;
- Avaliar a ontologia resultante;

Como comparar a precisão de técnicas de aprendizagem de ontologias não é uma tarefa trivial, a primeira abordagem preocupada em medir a corretude das técnicas de aprendizagem é menos utilizada.

2.3. Ferramentas de suporte e Sistemas de aprendizagem

Não há duvida que a construção de uma ontologia seja um processo que demanda tempo, além de requerer o esforço de especialista tanto em engenharia de ontologia como no domínio de interesse. Para algumas aplicações a utilização de uma abordagem totalmente manual para construção de ontologia é aceitável, mas alguma forma de abordagem semi-automática é indispensável, principalmente quando se trabalha com grandes volumes de informação como é o caso da internet.

Durante a última década muitas abordagens de aprendizagem e sistemas foram propostas. Alguns deles são sistemas de aprendizagem autônoma de ontologias enquanto outros simplesmente suportam ferramentas para construção de ontologias. Nesta sessão será discutida uma seleção dos dois tipos de sistemas.

2.3.1. ASIUM

ASIUM aprende a subcategorizar verbos e ontologias a partir de análise sintática de textos técnicos em linguagem natural (Francês). As entradas do ASIUM resultam dessa análise sintática, ele forma clusters ao redor de substantivos que ocorrem com um mesmo verbo e depois de uma mesma preposição (ou outra estrutura com mesmo valor sintático). ASIUM agrega sucessivamente os clusters para formar novos conceitos na forma de um grafo de generalidades que representa a ontologia do domínio. Os verbos são aprendidos em paralelo. O método ASIUM é baseado em clusterização conceitual. ASIUM propõe um método de aprendizagem de máquina cooperativa, que provê o usuário com uma visão global da tarefa de aquisição e também com ferramentas de aquisição como separação automática de conceitos, geração de exemplos, e uma visão da ontologia. Etapas de validação utilizando essas características são entrelaçadas com etapas de aprendizagem de forma que o usuário valida os conceitos conforme eles são aprendidos.

2.3.2. DODDLE II

DODDLE II (Domain Ontology Rapid Development Enviroment) provê um ambiente para construção de ontologias de domínio com relacionamentos taxonômicos e não- taxonômicos entre conceitos, explorando um dicionário legível por máquina (WordNet) e textos de domínio específico. Ele assiste o usuário na construção de ontologias de domínio. Os relacionamentos taxonômicos são oriundos do WordNet e contam com a interação de um especialista do domínio. Os relacionamentos não-taxonômicos vem de textos de domínio específico com a análise estatística de léxicos (coocurrence) baseada em quão próximos semanticamente eles estão um do outro.

2.3.3. HASTI

HASTI é um sistema de construção automática de ontologias, que constrói ontologias dinâmicas a partir do zero. HASTI aprende conhecimento léxico e ontológico de textos (Persas) de linguagem natural. Seu léxico é aproximadamente vazio inicialmente e vai crescendo gradualmente por aprendizagem de novas palavras. A ontologia no HASTI é um pequeno núcleo no começo. HASTI aprende conceitos,

relações taxonômicas não-taxonômicas entre conceitos e axiomas a fim de construir ontologias sobre o núcleo existente. A abordagem de aprendizagem em HASTI é uma abordagem simbólica híbrida, uma combinação de métodos lingüísticos, lógicos e heurísticos. Ele executa algoritmos de clusterização online e offline para organizar sua ontologia.

2.3.4. SYNDIKATE

É um sistema para aquisição automática de conhecimento a partir textos (Alemão) do mundo real, e para transferir seus conteúdos para estruturas de representação formal que constituem uma base de conhecimento. Ele integra requisitos vindos da análise de sentenças isoladas, assim como conjunto de sentenças que formam textos coesos. Além dos mecanismos centrados em análise de discurso para anáforas, SYNDIKATE é provido com um modulo de aprendizagem para boot-strapping automático do seu conhecimento de domínio conforme a análise do texto procede. A abordagem de aprender novos conceitos como um resultado da compreensão do texto constrói duas diferentes fontes de evidência: o conhecimento prévio do domínio tratado pelos textos, e construções gramaticais em que itens léxicos desconhecidos ocorrem nos textos. Uma dada ontologia é atualizada de forma incremental conforme novos conceitos são adquiridos de textos do mundo real. O processo de aquisição é centrado em lingüística e no conceito de "qualidade" de várias formas de evidencia que fundamentam a geração e refinamento de hipóteses conceituais.

2.3.5. SVETLAN'

É uma ferramenta de suporte a construção de ontologias. É um sistema para classificar substantivos no contexto. É capaz de aprender categorias de substantivos a partir de textos, seja qual for seu domínio. Palavras são aprendidas considerando o seu uso no contexto para evitar confusão no seu significado. SVETLAN' é uma ferramenta de suporte. Seus dados de entrada são domínios e a saída é o domínio estruturado aprendido contendo as classificações dos substantivos com suas relações com verbos. É baseado em uma abordagem distribuída: substantivos exercendo o mesmo papel sintático com um verbo em sentenças conectadas a um mesmo tópico (mesmo domínio) são agregados na mesma classe.

2.3.6. TEXT-TO-ONTO

TEXT-TO-ONTO é um ambiente de aprendizagem de ontologias, baseado em uma arquitetura geral para descobrimento de estruturas conceituais a partir de texto. Ele aceita tanto a aquisição de estruturas conceituais como mapeamento de recursos lingüísticos às estruturas adquiridas. Sua nova versão que suporta aprendizado de ontologias a partir de documentos da Web, permite a importação de dados semi-estruturados e estruturados como entrada. Ele também é composto de uma biblioteca de métodos de aprendizagem que são usados sob demanda. Seu método de aprendizagem é multi-estratégico, combinando vários métodos para várias entradas e várias tarefas.

2.3.7. WEB→KB

O objetivo do WEB→KB é criar de forma automática uma base de conhecimento da Web que seja compreensível por computador, cujo conteúdo seja um espelho da Web. Sua abordagem é desenvolver um sistema treinável de extração de informação que receba duas entradas: (1) uma base de conhecimento composta de uma ontologia definida por classes e relações de interesse, e opcionalmente, instâncias de algumas dessas classes e relações. (2) exemplos de treinamento da Web que descrevam instâncias dessas classes e relações. Dado essas entradas, o sistema determina ações gerais capazes de extrair instancias adicionais dessas classes e relações navegando pelo resto da internet. As saídas devem ser instancias classificadas e regras para extrair novas instâncias, regras para classificar paginas e regras para reconhecer relações entre muitas páginas. WEB→KB usa algoritmos de aprendizagem lógica e estatística para execução dessas tarefas.

No quadro 1 são mencionadas as características dos sistemas apresentados que os diferenciam dos demais sistemas de sua categoria [Shamsfard & Barforoush 2003].

Nome do sistema	Características
ASIUM	Aprendizagem de frames verbais e conhecimento taxonômico, baseado em análise estatística de parsers sintáticos de textos em francês.
DODDLE II	Ferramenta de suporte para aprendizagem de relações taxonômicas e não-taxonômicas utilizando-se de métodos estatísticos (co-occurrence analysis), explorando um dicionário (WordNet) e texto de domínio específico.
HASTI	Aprendizagem de conceitos, relações e axiomas por modos incremental e não incremental, começando de um pequeno kernel (aprendizagem a partir do zero), utilizando uma abordagem simbólica híbrida, uma combinação de métodos lógicos, métodos baseados em lingüística, e métodos heurísticos.
SVETLAN’	Ferramenta de suporte para construção de ontologias por aprendizagem de hierarquia de substantivos. Recebe domínios semânticos com unidades temáticas e constrói domínios estruturados para classificar substantivos conforme suas relações com verbos iguais.
SYNDIKATE	Aprendizagem incremental de conceitos e relações baseadas em compreensão do texto, usando lingüística e o conceito “qualidade” de várias formas de que fundamentam a geração e refinamento de hipóteses conceituais.
TEXT-TO-ONTO	Aprendizagem de conceitos e relações a partir de dados não estruturados, semi-estruturados e estruturados, usando um método multi-estratégico, uma combinação de regras de associação, análise conceitual formal e clusterização.
WEB→KB	Combinação métodos estatísticos (Bayes) e lógicos para aprendizagem de instancias e regras de extração de instâncias de documentos da Web.

Quadro 1: Resumo dos sistemas selecionados

2.4. Considerações finais

Para praticamente todas as categorias de sistemas de aprendizagem existem exemplos a serem citados, apesar de a grande maioria focar na aprendizagem de conceitos e se utilizarem de um abordagem de aprendizagem híbrida, garantindo assim mais flexibilidade, a variedade, para qualquer categoria que seja, é bem grande. Entre os sistemas selecionados, dois são ferramentas de suporte para aprendizagem de ontologias e cinco são sistemas de aprendizagem autônoma de ontologias. As ferramentas de suporte (SVETLAN' e DODDLEII) extraem estruturas essenciais de textos para fazer com que um sistema de aprendizagem autônoma seja capaz de construir uma ontologia. Já os sistemas de aprendizagem combinam várias abordagens de aprendizagem e em sua maioria, como já comentado, prestam maior atenção na aprendizagem de conceitos.

3. Técnicas e Ferramentas para aprendizagem

Aprendizagem de ontologias utiliza métodos das mais diversas áreas, tais como, aprendizagem de máquina, aquisição de conhecimento, processamento de linguagem natural, recuperação de informação, inteligência artificial, banco de dados, etc.

Este capítulo apresenta uma breve descrição das técnicas utilizadas pelo método de aprendizagem proposto no capítulo seguinte, além das ferramentas utilizadas para aplicação de tais técnicas e construção e avaliação da ontologia. Finalmente, são feitas algumas considerações a respeito do que foi apresentado no capítulo.

3.1. Técnicas de Processamento de Texto

É possível perceber que certas expressões fornecem-nos muita informação sobre a natureza das classes denotadas por nomes, portanto a abordagem utilizada para aprendizagem de ontologias a partir de texto deve explorar esse fato. Dessa forma, se faz necessário a aplicação de técnicas de processamento de linguagem natural de maneira automática. Tais técnicas são reconhecidamente empregadas em sistemas de Extração de Informação [Maynard 2006] e serão detalhadas a seguir.

3.1.1. Tokenization

Normalmente, uma etapa inicial do processamento é dividir o texto em unidades chamadas tokens onde cada token consiste de uma palavra ou alguma outra estrutura como um número ou um sinal de pontuação. O propósito dessa fase, portanto, é detectar limites de sentenças e de palavras. Em geral, o emprego da pontuação, representa uma dificuldade encontrada nessa fase. Um ponto, por exemplo, pode denotar o fim de uma sentença, o final de uma abreviação, ou até mesmo pode ser utilizado para expressar datas, horas, números de telefone, preços, etc. Um problema comum principalmente em textos extraídos da internet, e que diz respeito ainda ao emprego de sinais de pontuação, é a utilização dos populares "smiles", coisas como ":-)" [Manning & Schuetze 1999]. Outro problema que pode ser citado é que, por exemplo, espaços em branco nem sempre indicam os limites de uma palavra, como é o caso de "Centro de Informática". Quando isso se configura, a técnica de Tokenization pode ser utilizada em conjunto com Named Entity Recognition, técnica que é capaz de identificar nomes que se referem a entidades únicas do mundo real [Cimiano 2006].

3.1.2. Part-of-Speech (POS) Tagging

A lingüística agrupa as palavras de uma língua em classes (conjuntos) que apresentam comportamento sintático similar, e quase sempre um tipo semântico. Estas classes de palavras são conhecidas por classe gramatical, ou mais tecnicamente por parts-of-speech (POS). Três importantes part-of-speech são substantivo, verbo e adjetivo. Substantivos tipicamente referem-se a pessoas, animais, conceitos e coisas. Verbos são usados para expressar ações em uma sentença. Adjetivos descrevem propriedades de nomes. Muitas palavras têm várias parts-of-speech: a palavra "jogo"

pode ser um verbo, como em “eu jogo xadrez muito bem”, ou um substantivo como em “o jogo de xadrez estava muito bom” [Manning & Schuetze 1999].

Sistemas tradicionais de part-of-speech distinguem as palavras em um número restrito de classes, entretanto a lingüística normalmente utiliza classificações mais detalhadas de classes de palavras. Existem conjuntos de abreviações bem estabelecidas para nomear estas classes, usualmente conhecidas como POS tags. Algumas delas pode ser vistas no quadro 2.

Tag	Part-of-Speech
CC	Coordinating Conjunction
CD	Cardinal Number
DT	Determiner
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
NP	Proper noun
NPS	Common noun
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper Noun, singular
NNPS	Proper Noun, plural
POS	Possessive Ending
PRP	Personal Pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
WDT	Wh-determiner

Quadro 2: Algumas POS Tags

O POS Tagging, então, diz respeito à tarefa de associar a cada token sua classificação gramatical, tal como substantivo, adjetivo, verbo, etc. Os taggers que executam tal tarefa são baseados em vários métodos: árvores de decisão [Schmid 1994I], redes neurais [Schmid 1994II], modelo de Markov [Church 1998], regras ou transformação [Brill 1994], entre tantos outros. Nesse trabalho o tagger adotado foi desenvolvido por Mark Hepple [Hepple 2000] que tem em média 97% de acerto sob o corpus e é o POS tagger padrão do GATE [Bontcheva & Cunningham 2002].

3.1.3. Lematization

Lematization é um passo importante de pré-processamento não apenas para processamento de linguagem natural, mas para muitas aplicações de mineração de texto. Também é considerada uma forma produtiva de gerar palavras-chaves genéricas para engenhos de busca ou rótulos para mapas de conceitos.

Lematization é similar a técnica de Stemming, mas não requer a produção da raiz da palavra, mas apenas substituir o sufixo da mesma a fim de obter sua forma normalizada. Por exemplo, os sufixos das palavras em inglês: working, works, worked, devem ser substituídos para gerar a forma normalizada da palavra, o infinitivo: work, neste caso, a forma normalizada coincide com a raiz. Às vezes a forma normalizada difere da raiz da palavra. Por exemplo, as palavras em inglês:

computes, computing, computed pela técnica de Stemming devem ser reduzidas a palavra comput, mas a sua forma normalizada é o infinitivo do verbo: compute [Plisson et al 2001].

3.2. Ferramentas para construção de ontologias

Nesta seção são apresentadas as ferramentas utilizadas para instanciar o modelo que é proposto no próximo capítulo. Todas as técnicas de processamento de texto descritas até então são aplicadas por meio dessas ferramentas assim como as expressões regulares definidas para tratar a natureza dos textos.

3.2.1. GATE

O GATE [Bontcheva & Cunningham 2002] (a General Architecture for Text Engineering) é uma, ambiente de desenvolvimento que permite a criação de ferramentas para compilação e manutenção distribuída de corpus entre outras ferramentas que dão suporte ao processamento de linguagem natural. Foi lançada em 1996, quando então foi completamente reformulada, sendo relançada em 2002. Hoje, o GATE é um dos sistemas do seu gênero mais utilizados no mundo todo.

Características chaves do GATE [Cunningham 2007]:

- Desenvolvimento baseado em componentes o que reduz o overhead da integração de sistemas em pesquisas colaborativas;
- Medição automática de performance de componentes de engenharia de linguagem que promove avaliação comparativa quantitativa;
- Distinção entre tarefas de baixo-nível tais como armazenamento e visualização de dados, descoberta e carga de componentes e tarefas de alto nível de processamento de linguagem;
- Separação clara entre estrutura de dados e algoritmos de processamento de linguagem natural;
- Uso consistente de mecanismos padronizados para componentes para comunicar dados sobre linguagem, e uso de padrões abertos como Unicode e XML.
- Oferta de um conjunto base de componentes de engenharia que podem ser estendidos ou substituídos de acordo com as necessidades do usuário.

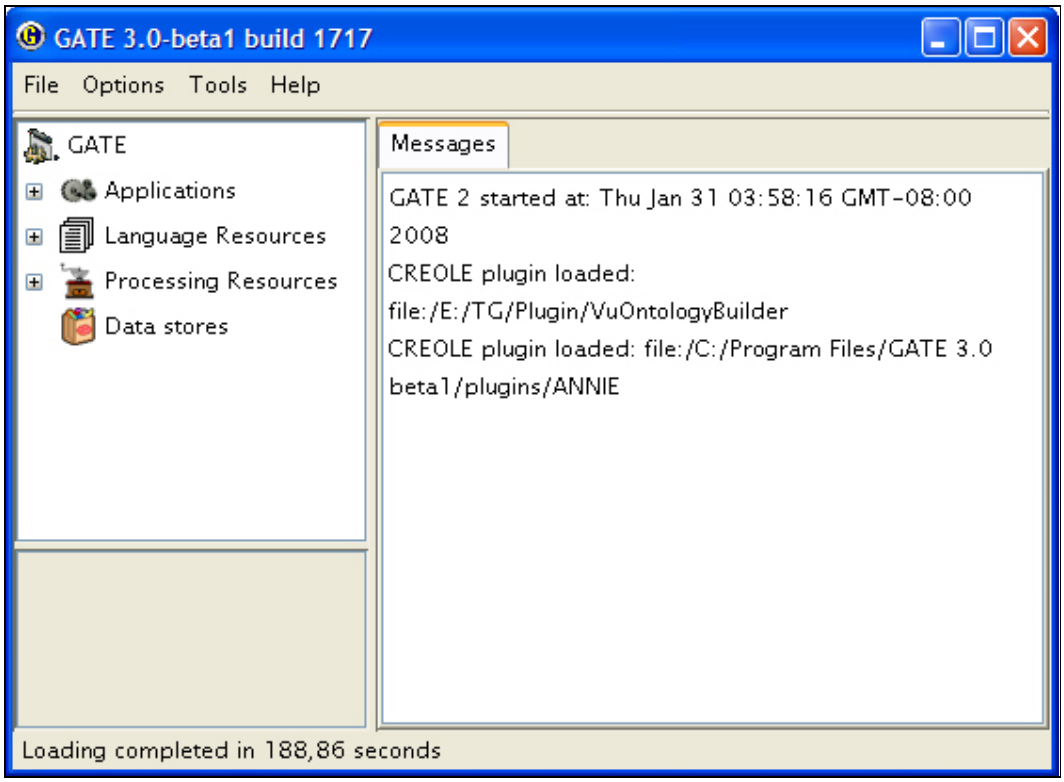

Figura 4: Ferramenta GATE

3.2.2. ANNIE

O GATE é distribuído com um conjunto de componentes de extração de informação chamado ANNIE (A Nearly-New IE system). Atualmente o ANNIE consiste do seguinte conjunto de módulos (que podem ser usados individualmente ou acoplados juntamente com novos módulos com o intuito de criar novas aplicações): tokeniser, sentence splitter, POS tagger, gazetter, finite state transducer, orthomatcher, e pronominal coreference resolution.

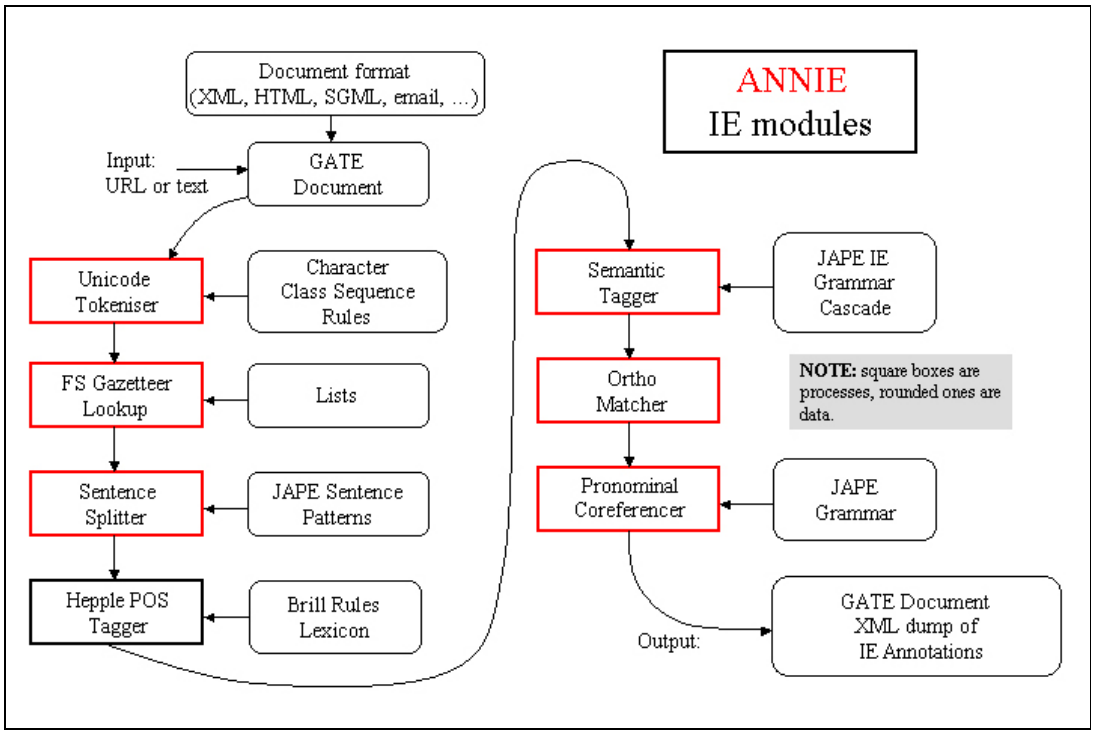

Figura 5: Arquitetura do ANNIE [Cunningham et al 2007]

Os módulos se comunicam via GATE's API annotation, que consiste basicamente de uma camada do sistema que serializa qualquer documento do corpus para uma representação interna em XML, essa representação consiste em um grafo de anotações que fazem referências a partes do texto do documento.

3.2.3. JAPE - Java Annotation Patterns Engine

O JAPE permite o reconhecimento de expressões regulares em anotações de documentos. Uma gramática JAPE consiste de um conjunto de fases, e cada fase consiste de um conjunto de regras padrões/ações. As fases executam em seqüência e constituem uma cadeia de máquinas de estado finito sobre as anotações. O lado esquerdo das regras consiste de um padrão de anotação que pode conter operadores de expressões regulares (|, *, ?, +). Já o lado direito das regras consiste de instruções para manipulação de anotações. Anotações que casem com o lado esquerdo de uma regra podem ser referenciados no lado direito por meio de rótulos [Tablan et al 2004].

A ferramenta é utilizada para identificar um conjunto de padrões potencialmente interessantes para construção da ontologia.

3.2.4. Benchmark Evaluation Tool

O Benchmark Evaluation Tool é uma ferramenta do próprio GATE, que permite calcular as métricas de recall-precision (mais detalhes no capítulo 5) com base no confronto dos termos extraídos automaticamente pelo sistema de aprendizagem com termos extraídos manualmente pelo engenheiro especialista no domínio.

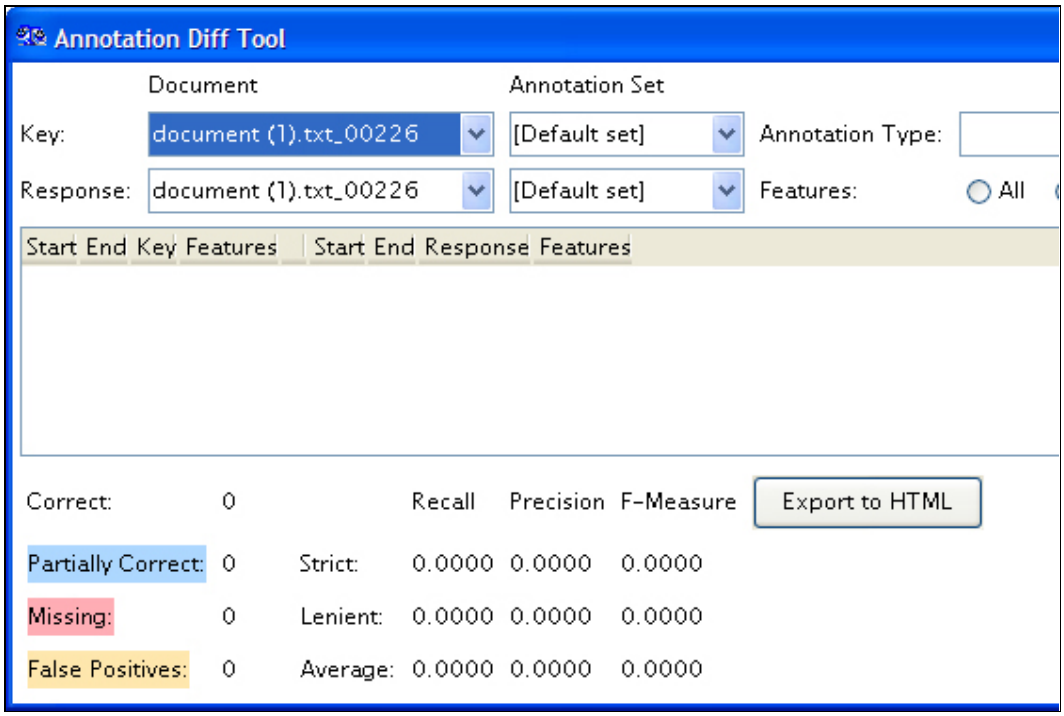

Figura 6: Benchmark Evaluation Tool

3.3. Considerações Finais

Neste capítulo foram apresentadas todas as técnicas utilizadas pelo sistema de aprendizagem proposto no próximo capítulo. As técnicas apresentadas aqui foram escolhidas por representarem um conjunto mínimo de técnicas utilizadas para a tarefa de aprendizagem de ontologias, segundo a literatura pesquisada. As ferramentas selecionadas para aplicação dessas técnicas e para a avaliação do sistema foram escolhidas com base em critérios que abragem desde facilidade de uso à maturidade e ampla aceitação das mesmas seja no mercado ou no meio acadêmico.

4. O Sistema de Aprendizagem

Esta seção contém a descrição do sistema proposto para o problema de aprendizagem de ontologias. O sistema proposto para extração de ontologias, em termos dos elementos aprendidos, pode ser classificado na categoria daqueles que aprendem conceitos. No que diz respeito ao método de aprendizagem ele pode ser classificado como híbrido por utilizar mais que uma abordagem de aprendizagem (baseada em lingüística e casamento de padrões).

Neste capítulo trataremos de como se deu a implementação do sistema. Será apresentado o domínio a parir do qual o sistema aprendeu assim como os detalhes de como o corpus foi coletado. Será descrito todas as fases do processo de extração da ontologia, construção e pruning. Finalmente, são feitas algumas considerações a respeito do que foi apresentado no capítulo.

4.1. A formação do Corpus

O corpus para o sistema foi extraído dos comentários Javadoc de arquivos de código fonte JAVA de um componente feito para um sistema desenvolvido pelo autor desse trabalho na empresa em que trabalha. Trata-se de um sistema para controle de transações e escalonamento de requisições a Web Services. Javadoc é uma ferramenta distribuída pela Sun em conjunto com sua a máquina virtual, e que se utiliza desses comentários para gerar documentação em formato HTML, esses comentários estão presentes para cada método definido nas classes presentes nos arquivos de código fonte como ilustrado na figura 7.

```
/**
 * Abort the current transaction and abandon any changes in progress.
 *
 * @param text  the text of the tool tip
 */
public void abortCurrentTransaction(){
```

Figura 7: Comentário JavaDoc

Cada comentário é constituído de uma descrição para o seu método, além dos parâmetros recebidos e tipo retornado por ele. Apenas a descrição é aproveitada para geração do documento. A descrição é extraída através da seguinte expressão regular que é exibida na figura 8.

```
Pattern p =
    Pattern.compile("(?<=\\/\\*\\*[\\n\t\\*\\s]{1,50}+)" +
                    "(.+?)" +
                    "(?=[\\n\t\\*\\s]{1,50}+@)", Pattern.DOTALL);
```

Figura 8: Expressão regular p/ extração da descrição do comentário

O subsistema de geração do corpus ainda é composto de um gerador de documentos que apenas tem a função de estruturar essa informação em documentos de texto distintos para métodos diferentes de cada classe do sistema, ou seja, para cada método de uma classe do sistema um documento do corpus é criado.

4.2. Extração da Ontologia

A extração da ontologia consiste de quatro fases, a fase de Tokenization (primeira fase), assim como a fase de Lematization (fase intermediária entre Padrões e Construção) foi omitida da figura abaixo para efeito de simplificação.

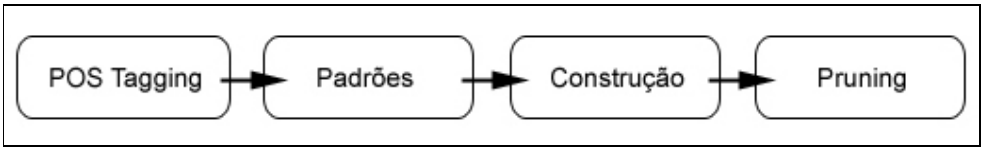

Figura 9: Pipeline de extração

4.2.1. POS Tagging

Durante a primeira fase o corpus é anotado pelo POS Tagger com informação referente à categoria sintática de cada token. A figura 10, que foi extraída do GATE ilustra as anotações feitas a um documento do corpus durante o processo de extração.

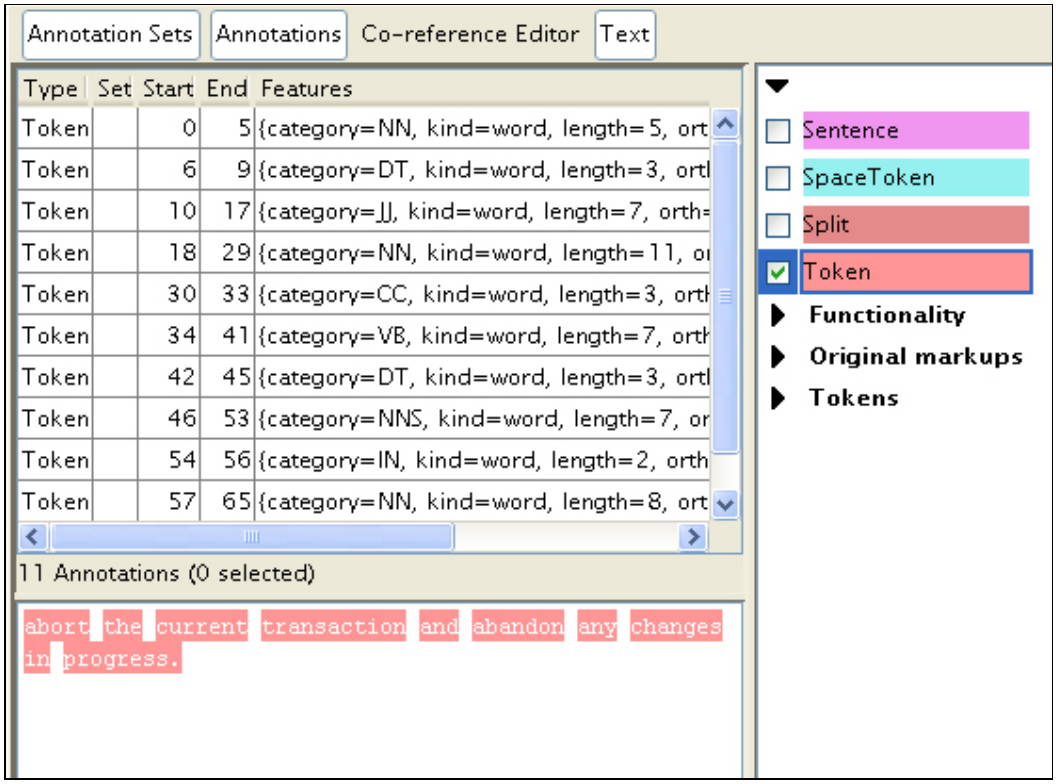

Figura 10: Anotações da Fase de POS Tagging

4.2.2. Padrões

Para o sistema foram definidos, basicamente dois padrões do JAPE para identificação dos conceitos do domínio.

Baseado no fato de que conceitos são representados por substantivos em um corpus para o primeiro padrão foram extraídas as orações substantivas. Orações

substantivas, são aquelas que desempenham as funções sintáticas próprias do substantivo, são sentenças construídas ao redor de um substantivo. Mais formalmente, uma oração substantiva consiste de um substantivo precedido por um número arbitrário (zero ou mais) de substantivos ou adjetivos conhecidos como seus modificadores. Como já citado, para executar essa extração foi utilizado o JAPE. Esta é a regra do JAPE que identifica orações substantivas:

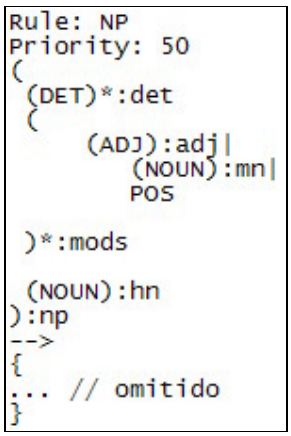

Figura 11: Padrão para orações substantivas

O padrão no lado esquerdo da regra identifica todas as seqüências de palavras começando com zero ou mais determinantes (ex., the, a), zero ou mais adjetivos, substantivos ou pronomes possessivos em qualquer ordem e obrigatoriamente terminando com um substantivo. DET, ADJ, NOUN e POS são macros, como ilustrado na figura 12, para outras regras que identificam palavras que são partes dessas classes gramaticais. O lado direito da regra, que foi omitido para efeito de simplificação, anota a seqüência identificada como uma oração substantiva [Sabou 2005|].

```
Macro: NOUN
(
  {Token.category == NN, Token.kind == word} |
  {Token.category == NN, Token.kind == punctuation} |
  {Token.category == NNS, Token.kind == word}|
  {Token.category == NNP, Token.kind == word} |
  {Token.category == NNPS, Token.kind == word} |
  {Token.category == NP, Token.kind == word} |
  {Token.category == NPS, Token.kind == word} |
  {Token.category == CD, Token.kind == word}
)

Macro: DET
(
  {Token.category == DT, Token.kind == word} |
  {Token.category == PRP, Token.kind == word} |
  {Token.category == WDT, Token.kind == word}
)

Macro: ADJ
(
  {Token.category == JJ, Token.kind == word} |
  {Token.category == JJR, Token.kind == word} |
  {Token.category == JJS, Token.kind == word}
)

Macro: POS
(
  {Token.category == POS, Token.kind == word}
)
```

Figura 12: Macros para classes gramaticais

Além do padrão para identificação dos conceitos do domínio através das orações substantivas, a extração também identifica as funcionalidades oferecidas pelo sistema. Em geral essas funcionalidades estão atreladas aos verbos que estão presentes na descrição dos comentários Javadoc, assim sendo, também foi utilizado um padrão para extrair os verbos e as orações substantivas que os seguem. As regras utilizadas, que compõem o padrão, foram as seguintes:

```
Rule:VerbID
(
  {Token.category=="VB"}|
  {Token.category=="VBZ"}
):vb

-->
{
... // omitido
}
```

Figura 13: Padrão para funcionalidades parte 1 de 2

```
Rule:Functionality
(
  ({VB}):vb
  ({Token.category == "IN",Token.kind == "word",Token.string == "for"})*
  ({NP}):np
):funct
-->
{
... // omitido
}
```

Figura 14: Padrão para funcionalidades parte 2 de 2

A equação 3 anota os verbos identificados da saída do POS Tagging (vb), enquanto que a regra ilustrada pela equação 4 apenas seqüencia os verbos das orações substantivas, como já comentado anteriormente, e os anota como uma anotação Functionality.

É possível observar as anotações feitas pela gramática definida no JAPE a um documento do corpus na seqüência de imagens capturadas do GATE durante a extração:

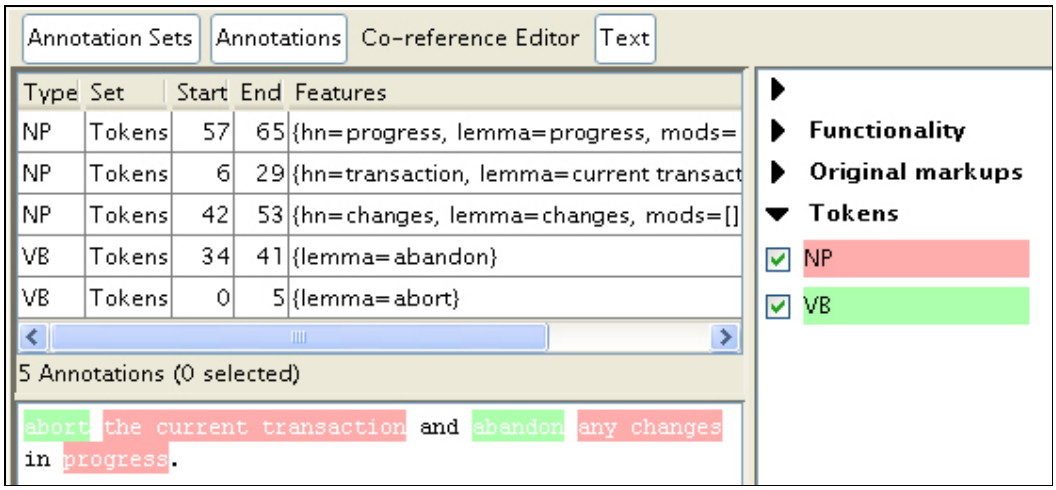

Figura 15: Anotações da Fase de Padrões parte 1 de 2

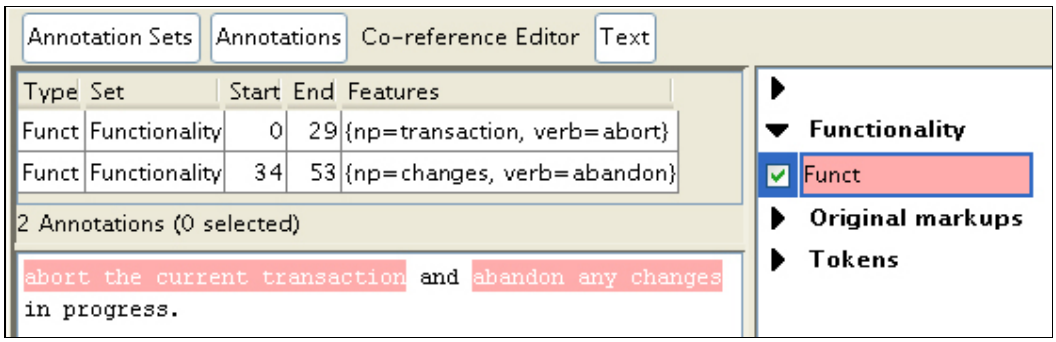

Figura 16: Anotações da Fase de Padrões parte 2 de 2

4.2.3. Construção da Ontologia

O estágio de construção da ontologia coleta os resultados da fase anterior baseada em expressões regulares. Os termos extraídos são usados para construir dois aspectos diferentes da ontologia de domínio. As orações substantivas são base para derivar a estrutura hierárquica de dados, que para o domínio explorado nesse trabalho podem representar conceitos relativos a estruturas de dados (listas, arrays, etc.) ou classes/objetos do sistema. E uma hierarquia de funcionalidades que representa os conceitos referentes a métodos do sistema, essa hierarquia foi construída a partir dos termos anotados como funcionalidade pela gramática definida para o JAPE. Os termos passaram pela técnica de Lematization, antes serem usados para construção da ontologia. Para geração do esquema RDFS foi utilizado um plugin disponível em [Sabou 2005II]. A figura 17 ilustra um fragmento para as duas hierarquias da ontologia aprendida.

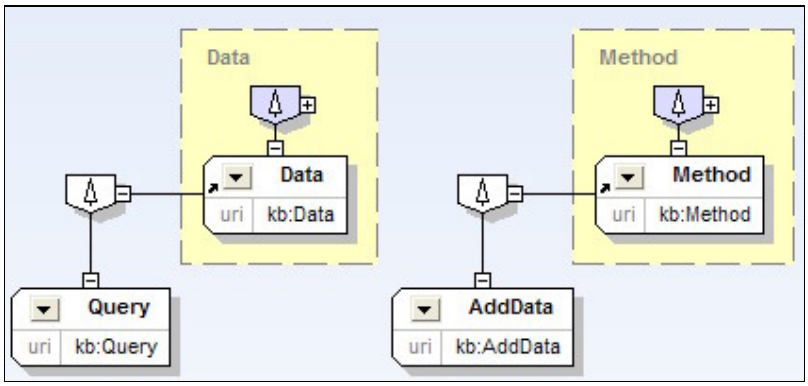

Figura 17: Fragmento das duas hierarquias da ontologia aprendida

4.2.4. Pruning

A fase de Pruning filtra conceitos irrelevantes da ontologia extraída, foi empregada uma estratégia de que termos com ocorrência freqüente no corpus denotam conceitos do domínio enquanto que termos menos freqüentes conduzem para conceitos que podem ser eliminados da ontologia com segurança [Maedche 2002]. Foi considerada a freqüência média dos termos como um valor de limiar e descartados todos os conceitos que apresentassem uma freqüência menor que esse valor. Outra heurística utilizada é que as orações substantivas incluídas em uma anotação de funcionalidade são mais prováveis de denotar conceitos do domínio, então se uma oração substantiva está associada a uma anotação de funcionalidade que não foi eliminada, a mesma também não é eliminada mesmo no caso em ocorrer poucas vezes no texto. O mesmo plugin utilizado na fase de construção da ontologia foi empregado na fase de Pruning.

4.3. Considerações Finais

Neste capítulo foi descrito em detalhes todo o processo de aprendizagem utilizado pelo sistema proposto, desde uma breve descrição do domínio e formação do corpus a extração, construção e pruning da ontologia resultante. Algumas dificuldades ficaram por conta das regras definidas para o JAPE, devido a falta de documentação disponível para essa ferramenta, além da dificuldade de configuração do ambiente do GATE para trabalhar como plugin utilizado para fase de construção e pruning.

5. Avaliação

Avaliação de aprendizagem de ontologias é extremamente importante, entretanto é um problema de difícil solução [Buitelaar et al 2004]. Tipicamente dois estágios de avaliação são executados na avaliação de um método de aprendizagem de ontologias. Primeiro: avaliação de termos que estima a performance da extração de termos relevantes ao domínio do corpus. Segundo: um estágio de avaliação da qualidade da ontologia estima a qualidade da ontologia extraída.

Neste capítulo serão introduzidos os tipos de avaliações mais comuns para aplicações de aprendizagem de ontologias, assim como os resultados baseados em duas dessas abordagens. Finalmente, algumas considerações a respeito do que foi apresentado no capítulo são feitas.

5.1. Tipos de Avaliação

Enquanto que a avaliação de termos pode ser executa usando métricas de precisão bem estabelecidas, avaliação de qualidade da ontologia é mais delicada e não existe método padrão para fazê-la. Uma abordagem é comparar uma ontologia extraída de forma automática com uma ontologia Gold Standard que é uma ontologia manualmente construída do mesmo domínio [Reinberger & Spyns 2004], frequentemente refletindo o conhecimento existente no corpus usado para a extração [Cimiano et al 2003]. O objetivo desta abordagem é avaliar o grau com que a ontologia cobre o domínio analisado. Outra abordagem é avaliar a adequação de uma ontologia para uma dada tarefa. Experimentos iniciais com avaliações baseadas em tarefas de ontologias são reportados em [Porzel & Malaka 2004]. Como uma terceira abordagem, uma avaliação por conceito por um especialista no domínio do corpus do qual a ontologia foi extraída.

Tais abordagens de avaliação cobrem aspectos importantes e complementares como [Sabou 2005|]:

Número	Problema	Métrica
1	Qual é a performance do algoritmo de aprendizagem?	Performance de extração
2	A ontologia extraída é uma boa base para construção de ontologia?	Avaliação de especialista
3	A ontologia extraída cobre todo o domínio analisado?	Avaliação de especialista
4	A ontologia extraída dá suporte a determinada tarefa?	Adequação a tarefa

Quadro 3: Tipos de Avaliação

5.2. Resultados

O método de extração apresentado nesse trabalho foi aplicado, como já comentado, aos comentários Javadoc de arquivos de código fonte JAVA do sistema. Os resultados foram avaliados segundo os critérios apresentados na seção anterior. No total foram cem documentos extraídos, sendo um documento para método encontrado no arquivo de código fonte.

Para o estudo de caso proposto foram executados dois dos quarto tipos de avaliações descritas acima, Primeiro foi executada a avaliação de termos (1). E em seguida confiou-se a ontologia extraída a uma avaliação conceito por conceito pelo desenvolvedor do sistema com base no modelo de análise de sistema (artefato de engenharia de software) (2).

5.2.1. Performance da Extração

Para medir a performance do modulo de extração, foram identificados manualmente todos os termos relevantes a serem extraídos do corpus. Então utilizando a ferramenta de benchmark (Benchmark Evaluation Tool) oferecia pelo GATE, e já apresentada no capítulo 3, foi possível comparar este conjunto de termos com aqueles que foram identificados na extração baseada em padrões. Foi utilizado o Term Recall (TRecall) para quantificar a taxa de termos relevantes (manualmente classificados) que foram extraídos do corpus analisado (correctextracted) sobre todos os termos a serem extraídos do corpus (allcorpus). Term Precision (TPrecision) denota a taxa de termos corretamente extraídos sobre todos os termos (allextracted).

$$TRecall = \frac{correct_{extracted}}{all_{corpus}} \quad ; \quad TPrecision = \frac{correct_{extracted}}{all_{extracted}}$$

Figura 18: Cálculo do TRecall

Para a ontologia extraída os valores encontrados foram:

TRecall	TPrecision
0.75	0.67

Quadro 4: TRecall e TPrecision calculados

Uma inspeção mais minuciosa nos termos extraídos pelo método revelou que verbos no início de sentenças eram confundidos com substantivos pelo POS Tagger o que diminuiu a quantidade de termos relevantes para descrever as funcionalidades.

5.2.2. Avaliação de Especialista

Avaliar se a ontologia extraída oferece uma base útil para construção de uma ontologia de domínio é importante uma vez que o principal objetivo deste trabalho é suportar a tarefa de construção de ontologia. Durante a análise conceito por conceito da ontologia extraída o especialista no domínio classifica os conceitos como CORRECT se tais conceitos foram úteis para a construção da ontologia e se já estavam inclusos na ontologia Gold Standard. Conceitos que são relevantes para o domínio, mas não foram considerados durante a construção manual da ontologia são classificados como NEW. Finalmente, conceitos irrelevantes, que não puderam ser usados, foram marcados como SPURIOUS. Quanto maior a taxa entre todos os conceitos relevantes e todos os conceitos da ontologia, melhor a extração no suporte a construção de ontologia. Essa taxa é expressa como (OPrecision):

$$OPrecision = \frac{correct + new}{correct + new + spurious}$$

Figura 19: Cálculo do OPrecision

Para a ontologia extraída o valor encontrado foi:

OPrecision
0.55

Quadro 5: OPrecision calculado

5.3. Considerações Finais

A avaliação do sistema de aprendizagem de ontologias proposto sugere que o método de extração é preciso, tendo em vista a extração de vários termos importantes do corpus (métricas precision-recall), ou seja, o método proporciona a construção de uma ontologia que contém muitos conceitos relevantes ao domínio. A ontologia extraída contém uma boa parte dos conceitos identificados manualmente além de sugerir vários outros conceitos complementares que acabaram sendo esquecidos por quem criou manualmente os conceitos a partir do artefato de análise do sistema, já mencionado antes.

6. Conclusão

Para que tecnologias apropriadas a dar suporte a Web Semântica, reuso de informação, integração de dados, Web Services, automatização do raciocínio em geral, possam emergir, algo precisa se feito no sentido de criar tais ontologias para o grande volume de informação disponível na Web. Esse objetivo pode ser alcançado com a aprendizagem de ontologias a partir desse volume de informação presente na rede.

Para o sistema proposto nesse trabalho em particular, a avaliação feita permite sugerir que o método de extração é preciso, tendo em vista a extração de vários termos importantes do corpus (métricas precision-recall), ou seja, a construção de uma ontologia que contém muitos conceitos relevantes ao domínio. O método ainda pode ser empregado como uma ferramenta de construção, uma vez que, como observado no capítulo anterior, pôde sugerir vários outros conceitos complementares que acabaram sendo esquecidos por quem criou manualmente os conceitos a partir do corpus.

O método definitivamente não é independente de domínio, considerando obviamente a necessidade em reescrever os padrões do JAPE de acordo com a natureza do texto a partir do qual será extraída a ontologia, mas isso não se configura em um problema muito grande, pois o processo de escrita das regras não é uma tarefa muito custosa, muito menos complexa.

6.1. Trabalhos Futuros

Devido a restrições de tempo e acréscimo de complexidade que pouco contribuiria para uma compreensão inicial do problema, o domínio de aplicação explorado por esse trabalho se limitou a documentação em língua inglesa de arquivos de código fonte, do contrário seria necessária implementação ou reutilização de um tokeniser em português, assim como um POS Tagger também da língua portuguesa. No entanto, um entendimento mais profundo da tarefa de aprendizagem de ontologias a partir de texto seria beneficiado com a não limitação do domínio a qualquer língua, portanto estender o sistema nesse sentido é um ponto importante a ser considerado.

Da mesma forma, a utilização da abordagem definida nesse trabalho em outros estudos de casos, a fim de testar a escalabilidade do método proposto na extração de ontologias de diferentes domínios, considerando obviamente a necessidade em reescrever os padrões do JAPE, seria uma opção a ser considera.

Seria válida também a aplicação de outros recursos no pipeline do GATE, tais como o Gazetteer, que consiste de um conjunto de listas de termos, que teria a função de um Stop List, ou seja, teria a responsabilidade de eliminar termos que não representassem conceitos para a ontologia a ser construída, mas que aparecessem com freqüência no corpus.

Explorar as funcionalidades providas pela API do GATE com o objetivo de customizar uma aplicação JAVA que venha a oferecer os mesmos recursos da configuração utilizada pela abordagem atual, acrescido de características que possibilitem uma utilização mais eficiente e em larga escala da metodologia sugerida no sentido da formação do corpus. Ou seja, prever a integração do módulo do sistema que extrai os comentários e das regras definidas para o JAPE.

Referências

[Berners-Lee 2001] Berners-Lee, T., Hendler, J., Lassila, O. “The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities” http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&pageNumber=6&catID=2. Publicado em 2001.

[Biemann 2005] Biemann, C., Ontology Learning from Text: A Survey of Methods. LDV-FORUM. Publicado em 2005.

[Blackburn 1996] Blackburn, S., The Oxford Dictionary of Philosophy. Oxford University Press, USA; New Ed edition (Março 30, 1996). ISBN-13: 978-0192831347

[Bontcheva & Cunningham 2002] Bontcheva, K., Cunningham H., GATE - a General Architecture for Text Engineering. Proceedings of the ACL-02 Demonstrations Session, Philadelphia, Julho 2002. Association for Computational Linguistics.

[Bontcheva et al 2002] Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., Cunningham, H. Shallow Methods for Named Entity Coreference Resolution. Department of Computer Science, University of Sheffield, 2002.

[Brill 1994] Brill, E. 1994. Some advances in transformation-based part of speech tagging. In Proceedings of the National Conference on Artificial Intelligence (AAAI).

[Buitelaar et al 2004] Buitelaar, P., Handschuh, S., Magnini, B. ECAI Workshop on Ontology Learning and Population: Towards Evaluation of Text-based Methods in the SemanticWeb and Knowledge Discovery Life Cycle. Valencia, Spain, Agosto de 2004.

[Church 1998] Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the Applied Natural Language Processing Conference (ANLP).

[Cimiano 2006] Cimiano, P. Ontology Learning and Population from Text Algorithms, Evaluation and Applications. University of Karlsruhe, Germany. ISBN-13:978-0-387-30632-2. 2006 Springer.

[Cimiano et al 2003] Cimiano, P., Staab, S., Tane, J. Automatic Acquisition of Taxonomies from Text: FCA meets NLP. In Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, Cavtat–Dubrovnik, Croatia, 2003.

[Cimiano et al 2005] Cimiano, P., Buitelaar, P., Magnini, B. 2005. Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press. ISBN-13: 978-1586035235

[Cunningham 2007] Cunningham, H. Infrastructure for Human Language Technology - GATE. University of Sheffield. 2007. Disponível em: http://www.gate.ac.uk/sale/gate-flyer/2007/gate-flyer-4-page.pdf. Acessadoo em 15 de Janeiro de 2008.

[Cunningham et al 2007] Cunningham, H. et al. Developing Language Processing Components with GATE Version 4 (a User Guide)The University of Sheffield. Julho de 2007. Disponível em http://gate.ac.uk/sale/tao/. Acessado em 22 de Janeiro de 2008.

[Genesereth & Nilsson 1987] Genesereth, M., Nilsson, L., Logical foundation of AI. San Francisco, Los Altos, Califórnia : Morgan Kaufman, 1987.

[Gove 2002] Gove, P., Webster's Third New International Dictionary. Unabridged. New York: Merriam-Webster, 2002. 2.783 p.

[Gruber 1993] Gruber, T., Towards principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993, later published in International Journal of Human-Computer Studies, Vol. 43, Issues 4-5, Novembro 1995

[Guarino & Giaretta 1995] Guarino, N., Giaretta, P., Ontologies and Knowledge Bases: Towards a Terminological Clarification. In N. Mars (ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing 1995. IOS Press, Amsterdam: 25-32.

[Guarino 1998] Guarino, N. "Formal ontology and information systems". In: N. Guarino (ed.), Formal Ontology in Information Systems. Proceedings of the First International Conference, Trento, Italy, 6-8 June 1998. IOS Press, 1998 p. 4

[Hepple 2000] Hepple, M. 2000. Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based Part-of-Speech Taggers. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000). Hong Kong, Outubro 2000.

[Lenat 1995] Lenat, B., CYC: A Large-Scale Investment in Knowledge Infrastructure, Communications of the ACM 38(11), 33-38, 1995.

[Maedche & Staab 2002] Maedche, A., Staab, S. Measuring similarity between ontologies. In Proceedings of European Knowledge Acquisition Workshop (EKAW). Springer, 2002.

[Maedche 2002] Maedche. A. Ontology Learning for the Semantic Web. Kluwer Academic Publishers, 2002.

[Mahmoud 2005] Mahmoud, Q., Service-Oriented Architecture (SOA) and Web Services: The Road to Enterprise Application Integration (EAI) [Online]. 2005. Disponível em: http://java.sun.com/developer/technicalArticles/WebServices/soa/. Acessado em 12 de Janeiro de 2008.

[Manning & Schuetze 1999] Manning, C., Schuetze, H. Foundations of Statistical Natural Language Processing. The MIT Press; 1 edition (June 18, 1999). ISBN-13: 978-0262133609.

[Maynard 2006] Maynard, D. GATE Training course 2006. GATE GUI. Sheffield University. Disponível em: http://videolectures.net/gate06_maynard_gg/. Acessado em 12 de Janeiro de 2008.

[Mehrnoush & Barforoush 2003] Mehrnoush, S., Barforoush, A. The State of the Art in Ontology Learning: A Framework for Comparison. 2003.

[Plisson et al 2001] Plisson, J., Lavrac, N., Mladenic, D. A Rule based Approach to Word Lemmatization. Department of Knowledge Technologies. Jožef Stefan Institute, 2001.

[Porzel & Malaka 2004] Porzel, R., Malaka, R. A Task-based Approach for Ontology Evaluation. In ECAI Workshop on Ontology Learning and Population, Valencia, Spain, 2004.

[Reinberger & Spyns 2004] Reinberger, M., Spyns, P. Discovering Knowledge in Texts for the Learning of DOGMA-Inspired Ontologies. In ECAI Workshop on Ontology Learning and Population, Valencia, Spain, Agosto de 2004.

[Sabou 2005I] Sabou M. Learning Web Service Ontologies: an Automatic Extraction Method and its Evaluation. Department of Artificial Intelligence, Vrije Universiteit Amsterdam. IOS Press, 2005.

[Sabou 2005II] Sabou, M. Learning Domain Ontologies for Web Service Descriptions - resource page. http://kmi.open.ac.uk/people/marta/experiments/extraction.html. Acessado em 17 de Janeiro de 2008.

[Schmid 1994I] Schmid, H. Part-of-Speech Tagging with Neural Networks. Proceedings of the 15th International Conference on Computational Linguistics (COLING-94). Agosto de 1994.

[Schmid 1994II] Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. Setembro de 1994.

[Shamsfard & Barforoush 2003] Shamsfard, M., Barforoush, A. The State of the Art in Ontology Learning: A Framework for Comparison. Intelligent Systems Laboratory, Computer Engineering Dept., Amir Kabir University of Technology, Hafez ave., Tehran, Iran. 2003.

[Sowa 1999] Sowa, J., Building, sharing and merging ontologies. Tutorial. [S. 1.: s. n.], 1999. Disponível em: http://users.bestweb.net/~sowa/ontology/ontoshar.htm. Acesso em: 7 de Janeiro de 2008.

[Sowa 2003] Sowa, J. F. (2003). Ontology. Disponível em: http://www.jfsowa.com/ontology/. Acessado em: 26 de Janeiro de 2008.

[Staab & Studer 2004] Staab, S., Studer, R., (Eds). 2004. Handbook on Ontologies. International Handbooks on Information Systems. Springer: ISBN 3-540-40834-7.

[Tablan et al 2004] Tablan, V., Maynard, D., Bontcheva, K., Cunningham, H. GATE - An Application Developer's Guide. Department of Computer Science, University of Sheffield, UK. 19 de Julho 2004.

Assinaturas

Frederico Luiz Gonçalves de Freitas
Orientador

Flávia de Almeida Barros
Avaliadora

Zinaldo Araujo Barros Jr
Aluno