Ferramentas para aprendizagem de ontologias a partir de textos

Faruk Mustafa Zahra

Mestre em Tecnologia em Saúde pelo Programa de Pós-Graduação em Tecnologia em Saúde (PPGTS) da Pontifícia Universidade Católica do Paraná (PUCPR), graduado em Análise de Sistemas pela Pontifícia Universidade Católica do Paraná e especialista e certificado na linguagem de programação JAVA. Arquiteto de Soluções na Lume Tecnologia

Andreia Malucelli

Doutora em Engenharia Eletrotécnica e de Computadores pela Faculdade de Engenharia da Universidade do Porto (FEUP), Portugal e mestre em Engenharia Elétrica e Informática Industrial pela Universidade Tecnológica Federal do Paraná (UTFPR). Professora titular da Pontifícia Universidade Católica do Paraná (PUCPR), Programa de Pós-Graduação em Informática (PPGIa)

Ademir Roberto

Doutor em Engenharia Elétrica e Informática Industrial pela Universidade Tecnológica Federal do Paraná (UTFPR). coordenador do Curso de Licenciatura em Informática e professor na área de Informática

César Augusto Tacla

Doutor em Informática pela Université de Technologie de Compiègne, França. Atua nos Programas de Pós-Graduação em Engenharia Elétrica e Informática Industrial e em Computação Aplicada da UTFPR – Curitiba

Há muito interesse na construção de ontologias, porém, há poucos trabalhos que apontam para uma utilização de ontologias em larga escala. Algumas das razões para isso são o tempo, custo e recursos associados ao desenvolvimento. Com o objetivo de reduzir estes esforços, métodos de aprendizagem de ontologias têm sido desenvolvidos. Para isso, é necessária a automatização do processo de aquisição de conhecimento e diversas abordagens têm sido sugeridas. A utilização de

textos como fonte de aquisição de conhecimento parece ser um caminho interessante, visto que a linguagem é a primeira forma de transferência de conhecimento entre os seres humanos, além de haver uma diversidade de documentos digitais disponíveis. Para dar suporte à construção semiautomática de ontologias a partir de textos, várias ferramentas foram desenvolvidas, cada uma utilizando técnicas e métodos diferentes. Este artigo tem como objetivo apresentar algumas das ferramentas disponíveis, assim como suas características principais e usabilidade do ponto de vista de um usuário não versado em computação.

Palavras-chave: *Representação do conhecimento; Ontologia; Aprendizagem de ontologias.*

Tools for ontology learning from texts

There is much interest in building ontologies, however, there are few works pointing to a large-scale use of ontologies. Some reasons for this are time, cost, and resources associated with the development. With the goal of reducing such efforts, ontology learning methods have been developed. Therefore, it is necessary to automate the process of acquiring knowledge and different approaches have been suggested. The use of texts as a source of knowledge acquisition appears to be a correct path, since language is the first manner of knowledge transferring among human beings. Besides, there are a lot of digital documents available. To support semi-automatic construction of ontologies from texts, several tools have been developed, each one using different techniques and methods. Due to the importance of using these tools, this paper describes the available tools, their main features, and usability from the perspective of a regular user.

Keywords: *Knowledge representation; Ontology; Ontology learning.*

Recebido em 13.09.2011 Aceito em 16.07.2013

1 Introdução

Muitas pesquisas em Inteligência Artificial (IA) têm sido dedicadas ao desenvolvimento de sistemas computacionais que incorporam conhecimento sobre determinado domínio, permitindo inferências, raciocínios e tomada de decisões. Estes sistemas mantêm uma representação explícita e simbólica do conhecimento. Tal representação tem a vantagem de estar separada dos aspectos procedurais relacionados à aplicação, podendo, desta forma, ser reusado por outros sistemas. Para realizar esta tarefa, faz-se necessário organizar o conhecimento de maneira formal e disponibilizá-lo em uma linguagem padrão para que possa ser compartilhado, pois computadores são, essencialmente, máquinas processadoras de símbolos e precisam de instruções claras sobre como manipular estes símbolos, de maneira a atribuir-lhes significado (CIMIANO, 2006).

Neste contexto, as ontologias podem ser utilizadas para representar o domínio em questão, pois permitem representar vocabulários formais que descrevem as premissas básicas de um determinado domínio (FREITAS; SCHULZ, 2009).

A definição de ontologia, mais amplamente difundida e utilizada, diz que "ontologia é uma especificação explícita de uma conceitualização" (GRUBER, 1993, p. 2). Esta definição foi complementada por Studer, Benjamins e Fensel (1998, p. 184), como: "ontologia é uma especificação explícita e formal de uma conceitualização compartilhada de um domínio de interesse". Nesta definição, compartilhada significa que uma ontologia deve capturar o conhecimento consensual e formal refere-se ao fato que a ontologia deve ser declarativamente definida, legível e interpretável por máquina.

Processar informações, usando ontologias, que provêm excelente contexto para dar significado às informações, tanto para usuários humanos quanto para agentes de *software*, vem se tornando uma tendência em várias áreas e tipos de aplicações (NOY; MUSEN, 2002). A motivação para o desenvolvimento de ontologias é encontrada nas diversas aplicações e benefícios obtidos pelo seu uso.

Há muito interesse na construção de ontologias, porém, há pouco trabalho desenvolvido e pouca utilização de ontologias em larga escala. Algumas das razões para isso são o tempo, custo e recursos associados neste desenvolvimento. Para a criação de uma ontologia é necessário que um especialista do domínio transfira o conhecimento consensual de sua área para a ontologia, sendo que este processo é extremamente custoso. Além disso, as ontologias devem ser compartilhadas por um grupo de pessoas ou por uma comunidade, sendo o processo de construção dificultado por envolver diferentes pessoas com pontos de vistas, muitas vezes, divergentes (CIMIANO, 2006). Por estas razões, a comunidade de engenharia de ontologias está explorando novos métodos e técnicas para reduzir tempo e esforço necessários no processo de aquisição do

conhecimento, facilitando a construção de novas ontologias (GÓMEZ-PÉREZ, MANZANO-MACHO, 2003).

Na aquisição de conhecimentos, os engenheiros/especialistas possuem dificuldades em capturar e representar o conhecimento, mais precisamente, elicitar conhecimentos a serem representados na ontologia e, finalmente, formalizar e implementar este conhecimento em uma linguagem interpretável por máquina. Deste modo, devido ao tamanho, complexidade e dinamicidade de um determinado domínio, métodos de aprendizagem de ontologias têm sido desenvolvidos com o objetivo de reduzir esforços na aquisição de conhecimentos.

O processo de aprendizagem de ontologias (*ontology learning)* é formado por um conjunto de métodos e técnicas de construção semiautomática de novas ontologias ou para enriquecer ontologias já existentes (GÓMEZ-PÉREZ; MANZANO-MACHO, 2003).

Porém, para construir uma ontologia de maneira semiautomática, é necessária a automatização do processo de aquisição de conhecimento e, para isso, diversas abordagens foram sugeridas (MAEDCHE; STAAB, 2004). Estas abordagens podem ser classificadas de acordo com o tipo de entrada. Logo, há abordagens para a construção de ontologias a partir de textos, dicionários, bases de conhecimento, esquemas semiestruturados e bases de dados relacionais.

Considerando a grande quantidade de documentos digitais disponíveis na *Web* e nas organizações, o melhor seria considerá-los como um recurso útil na aquisição de conhecimento para a construção de ontologias. Segundo Buitelaar; Magnini (2005), a utilização de textos como fonte de aquisição de conhecimento é um caminho interessante, visto que a linguagem é a primeira forma de transferência de conhecimento entre os seres humanos.

Para dar suporte ao processo de aquisição do conhecimento e para a construção semiautomática de ontologias a partir de textos, várias ferramentas foram desenvolvidas, cada uma utilizando técnicas e métodos diferentes. Devido à importância do uso destas ferramentas, faz-se necessário conhecê-las, entender quais componentes da ontologia que podem ser construídos em cada ferramenta, assim como quais estão disponíveis para uso e a sua facilidade.

Este artigo apresenta uma visão geral das ferramentas para a construção semiautomática de ontologias, a partir de textos existentes, e minimiza o trabalho de pesquisadores interessados em utilizar estas ferramentas, sendo possível saber quais estão realmente disponíveis para uso e possuem fácil usabilidade, principalmente para quem não é da área da computação. Nota-se que das dezoito ferramentas encontradas na literatura para a língua estrangeira, apenas três estavam disponíveis para *download* e das três ferramentas encontradas para língua portuguesa, apenas duas estavam disponíveis para *download*. Ainda assim, nem todas que estavam disponíveis puderam ser utilizadas.

Neste artigo, é possível observar que ainda há uma distância entre o que é proposto na literatura científica e o que está disponibilizado na

prática. Profissionais de áreas diferentes da ciência da computação precisam de ferramentas disponíveis para uso em suas aplicações e que não seja necessário desenvolver programas de computador específicos ou adaptar as ferramentas computacionais para que seja possível o seu uso.

2 Aprendizagem de ontologias a partir de textos

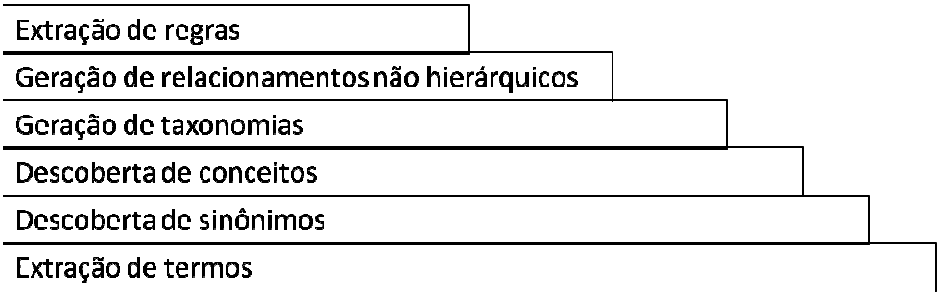
A Aprendizagem de Ontologias (*Ontology Learning)*, também conhecida como processo de Geração de Ontologias (*Ontology Generation*) ou Extração de Ontologias (*Ontology Extraction*), é definida como um conjunto de métodos e técnicas para a construção de uma ontologia ou enriquecimento de uma já existente (GÓMEZ-PÉREZ; MANZANO-MACHO, 2003), utilizando-se de diversos tipos de fontes de informação, como dicionários, textos ou bases de dados.

Segundo Maedche e Staab (2004), as abordagens de aprendizagem de ontologias podem ser classificadas de acordo com as entradas que incluem:

> a)textos: partem de um corpus de um domínio e utilizam técnicas de processamento de linguagem natural e aprendizado de máquina para aprender termos e relacionamentos;
>
> b)dicionários: utilizam as definições das entradas léxicas e utilizam as relações existentes entre elas (ex. sinonímia);
>
> c)bases de conhecimento: utilizam as regras existentes nas bases de conhecimento;
>
> d)esquemas semiestruturados: utilizam estruturas pré-definidas de dados e de relacionamentos como os existentes em arquivos *eXtended Markup Language* (XML); e
>
> e)bancos de dados relacionais: extraem conceitos e relações relevantes do conhecimento representados nos registros e nos *schemas* dos bancos de dados.

Para cada um dos tipos de abordagem, há técnicas específicas e ferramentas relacionadas. A tarefa de construção de ontologias a partir de textos abrange basicamente as seguintes etapas (Figura 1): extração de termos, descoberta dos sinônimos, geração de taxonomias, geração dos relacionamentos (não hierárquicos) e extração de regras (BUITELAAR; MAGNINI, 2005). As etapas podem ser vistas como camadas, sendo que as superiores exigem técnicas mais complexas de aprendizado.

Figura 1 – Etapas para construção de ontologias a partir de textos



Fonte: Adaptado de BUITELAAR; MAGNINI (2005).

A *extração dos termos* é um pré-requisito para todos os outros passos da aprendizagem de ontologias a partir de textos (CIMIANO, 2006). Um termo é uma palavra ou um conjunto de palavras com algum significado associado ao domínio de interesse. Os termos podem ser recuperados através de técnicas de recuperação de informação (*information retrieval*), termos indexados, terminologias e processamento de linguagem natural (PLN).

Na etapa seguinte, *descoberta dos sinônimos,* são recuperadas as variações semânticas dos termos. Muitos trabalhos sugeridos desta área usam o léxico disponível em *WordNet* (MILLER *et al.*, 1990; MILLER, 1995) para encontrar os sinônimos em inglês e o EuroWordNet (VOSSEN, 1998), para sinônimos de outras línguas do continente europeu.

A descoberta de *conceitos* possui alguma sobreposição com a camada anterior de descoberta de sinônimos. As técnicas utilizadas nestas camadas (sinônimos e conceitos) têm por objetivo descobrir palavras que compartilham algum dos seus significados.

Para a etapa de *geração de taxonomias,* existem diversos técnicas baseadas em informações léxico-sintáticas, em termos complexos e no agrupamento (*clusterization*) hierárquico dos termos.

A técnica de (HEARST, 1992), baseada em informações léxico-sintáticas, tem por objetivo extrair relações de hiponímia (relação "é-um") dos textos. Em Morin e Jacquemin (2004), é proposta uma técnica semelhante a de Hearst (1992), mudando-se apenas, em alguns casos, as estruturas extraídas e a língua alvo de inglês para francês.

Na abordagem baseada em termos complexos, também chamados de compostos ou multi-palavras, a técnica analisa a estrutura interna dos termos, portanto, quando um termo está lexicamente contido em outro, ele é considerado seu hipônimo. Os trabalhos que utilizam esta técnica foram desenvolvidos por Buitelaar; Magnini (2005); Ryu; Choi (2006); e Baségio (2007).

O agrupamento hierárquico dos termos consiste na construção de estruturas taxonômicas a partir do agrupamento conceitual e hierárquico dos termos. Faure e Nedellec (1998) propõem um método em que os agrupamentos básicos são formados por palavras que ocorrem com o mesmo verbo após a mesma preposição. Calcula-se, então, a similaridade semântica entre os agrupamentos básicos e agrupam-se os mais

similares, formando novos agrupamentos. Este processo é repetido até restar somente um único *agrupamento,* no qual todos os termos estão presentes.

Para a etapa de *geração dos relacionamentos* não hierárquicos, vários trabalhos sugerem combinar mineração de textos (*text mining*), análises estatísticas e linguísticas, explorando as estruturas sintáticas para extrair relacionamentos entre os conceitos da ontologia (CIARAMITA *et al*., 2005). Além de identificar as propriedades, também é importante identificar as hierarquias de propriedades. Porém poucas abordagens possuem este propósito, entre elas destaca-se a de Staab, Erdmann e Maedche (2001).

Segundo Cimiano (2006), as regras dizem respeito, entre outras, à disjunção e à equivalência de conceitos. Disjunção é o fato de dois conceitos diferentes não admitirem instâncias em comum. Uma das técnicas utilizadas para descoberta de disjunção (HAASE; VÖLKER, 2005) se baseia em padrões de expressões que indicam uma provável disjunção entre conceitos (ex. "os verdes são opostos aos industriais"). As regras se referem também às relações entre conceitos. Por exemplo, (LIN; PANTEL, 2001) busca relações inversas entre conceito tal como: "X é autor de Y" então "Y é produzido por X".

3 Ferramentas de construção de ontologias a partir de textos

Buscou-se na *Web* por ferramentas de aprendizagem de ontologias a partir de textos. Das dezoito ferramentas encontradas para língua estrangeira, apenas três estavam disponíveis para *download* e instalação, as quais foram instaladas e testadas, utilizando-se um *corpus* com 15 artigos científicos sobre aprendizagem de ontologias, na língua inglesa. Das três ferramentas encontradas para a língua portuguesa, apenas duas estavam disponíveis para *download* e instalação.

O Quadro 1 apresenta um comparativo entre as ferramentas para aprendizagem de ontologias para a língua portuguesa. Entre as ferramentas para o português estavam disponíveis: ONTOLP (RIBEIRO JUNIOR, 2008) e PORONTO (ZAHRA, 2009; ZAHRA; CARVALHO; MALUCELLI, 2013).

Quadro 1 – Comparativo entre ferramentas para aprendizagem de ontologias para língua portuguesa

Nome	Objetivo e Escopo	Fonte de dados	Intervenção do usuário	Idioma do corpus	Fonte
Baségio	Extração semiautomática de estruturas ontológicas, extração de termos e relações taxonômicas.	Texto analisado morfologicamente, o processo é manual.	Seleção de termos extraídos	Português	(Baségio, 2007)
ONTOLP	Extração semiautomática de estruturas ontológicas, extração de termos e relações taxonômicas.	Texto analisado morfossintaticamente, no padrão XCES, utilizando a ferramenta VISL.	Seleção de termos extraídos e de grupos semânticos	Português	(Ribeiro Junior, 2008)
PORONTO	Extração semiautomática de estruturas ontológicas, extração de termos e relações taxonômicas.	Texto analisado morfologicamente, utilizando o TreeTagger.	Seleção de termos válidos extraídos	Português	(Zahra, 2009)

Fonte: Dados da pesquisa.

O Quadro 2, adaptado de Gómez-Pérez, Manzano-Macho (2003), apresenta um comparativo entre as ferramentas para aprendizagem de ontologias para línguas estrangeiras encontradas na literatura. Dentre as ferramentas para línguas estrangeiras, estavam disponíveis para *download:* KEA (*Keyphrases Extraction Algorithm*) (JONES; PAYNTER, 2002), TERMINAE (BIÉBOW; SZULMAN; CLÉMENT, 1999) e *Text-To-Onto* (MAEDCHE; STAAB, 2004).

Quadro 2 – Comparativo entre ferramentas para aprendizagem de ontologias para línguas estrangeiras

Nome	Objetivo e Escopo	Fonte de dados	Intervenção do usuário	Idioma do *corpus*	Fonte
ASIUM	Objetivo do ASIUM é encontrar relacionamentos taxonômicos entre textos em linguagem natural na língua francesa.	Texto analisado morfossintaticamente	Todo o processo	Francês	(FAURE; POIBEAU, 2000)
Caméléon	Objetivo do Caméléon é identificar padrões léxico-sintático ou utilizar o padrão de Hearst para encontrar relacionamentos não hierárquicos e	Texto analisado morfossintaticamente	Validação ou definição de novas regras léxicas sintáticas	Inglês e Francês	(AUSSENAC-GILLES; SEGUELA, 2000)
CORPORUM-Ontobuilder	Objetivo do CORPORUM-Ontobuilder é encontrar relacionamentos taxonômicos entre	Texto em linguagem natural	Não é necessário	Inglês	(ENGELS, 2001)
DOE: Differential Ontology Editor	DOE é um editor de ontologias que permite o engenheiro criar ontologias em três passos de acordo com a	Texto em linguagem natural	Todo o processo	Inglês	(BACHIMON, 1996)
KEA: Keyphrases Extraction Algorithm	KEA é uma ferramenta que extrai automaticamente palavras-chave dado um conjunto de textos utilizando	Texto em linguagem natural	Avaliação do resultado final	Inglês	(JONES; PAYNTER, 2002)
LTG (Language Technology Group) Text Processing Workbench	LTG é um conjunto de ferramentas que tem o objetivo de descobrir estruturas internas de textos na língua	Texto em linguagem natural	Todo o processo	Inglês	(MIKHEEV; FINCH, 1997)
Mo'K Workbench	Mo'K é uma ferramenta que tem como objetivo extrair conceitos taxonômicos utilizando anotações	Texto analisado morfossintaticamente	Todo o processo	Francês	(BISSON et al., 2000)
OntoLearn Tool	OntoLearn é uma ferramenta que tem como objetivo extrair termos de domínio de um texto, relacioná-los com conceitos adequados para uma ontologia e extrair relacionamentos	Texto em linguagem natural	Avaliação do resultado final	Inglês	(VELARDI et al., 2002)
Prométhée	Prométhée é uma ferramenta que tem como objetivo extrair e refinar padrões léxico-sintáticos relativos a	Texto em um padrão restrito	Todo o processo	Francês	(MORIN, 1999)
SOAT: a Semi-Automatic Domain Ontology Acquisition Tool	SOAT permite extrair de forma semiautomática ontologias de domínio usando como fonte *corpus* de domínio, identificando palavras-chave	Texto em linguagem natural	Informações não disponíveis no artigo	Chinês	(WU et al.,2002)
SubWordNet Engineering Process tool	SubWordNet é uma ferramenta que tem como objetivo construir e manter línguas alternativas para o WordNet,	Texto em linguagem natural	Todo o processo	Inglês	(GUPTA et al., 2002)
SVETLAN'	SVETLAN' é uma ferramenta que cria clusters de palavras que aparecem dos textos fornecidos,	Texto em linguagem natural	Avaliação do resultado final	Francês	(CHAELANDAR; GRAU, 2000)
TFIDF-based term classification system	TFIDF-based term classification system tem como objetivo extrair termos de domínio e relacionamentos entre eles, utilizando como técnica	Texto em linguagem natural	Avaliação do resultado final	Alemão	(XU et al.,2002)
TERMINAE	TERMINAE tem como objetivo auxiliar o engenheiro de ontologias na criação de uma ontologia, integrando	Texto em linguagem natural	Avaliação do resultado final	Francês e Inglês	(BIÉBOW; SZULMAN, 1999)
Text-To-Onto	Text-To-Onto tem como objetivo encontrar relacionamentos taxonômicos e não taxonômicos, utilizando técnicas como analise estatística, podas e regras de associação.	Texto em linguagem natural, dicionários e ontologias	Avaliação do resultado final	Inglês	(MAEDCHE; STAAB, 2003)
TextStorm and Clouds	O objetivo TextStorm and Clouds é a criação semiautomática de uma rede semântica, onde contem apenas conceitos e seus relacionamentos, utilizando textos de um domínio especifico.	Texto em linguagem natural	Todo o processo	Inglês	(NOVAK; GOWIN, 1984)
Welkin	Welkin é uma ferramenta que tem como objetivo gerar automaticamente material para um e-learning, utilizando técnicas como TFIDF e qui-quadrado.	Ontologia de domínio e WordNet	Não é necessário	Inglês	(ALFONSECA; RODRÍGUEZ, 2002)
WOLFIE (Word Learning From Interpreted Examples)	O objetivo do Wolfie é a aprendizagem léxico-semântica de um corpus de frases com representações de seus significados.	Corpus pré-processado com representações de seus significados	Avaliação do resultado final	Inglês	(THOMPSON; MOONEY, 1997)

Fonte: Adaptado de GÓMEZ-PÉREZ, MANZANO-MACHO (2003).

4 Experimentos

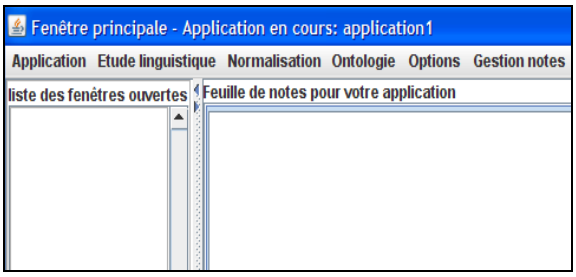
As ferramentas utilizadas no experimento foram as seguintes: KEA, Text-To-Onto e PORONTO. Algumas das ferramentas disponíveis para *download* não foram testadas, devido à ausência de ferramentas auxiliares (TERMINAE – ausência do LEXER - e OntoLP – ausência de uma ferramenta de análise linguística). Cada ferramenta testada funciona de maneira diferente, não sendo possível utilizar o mesmo *corpus* para os experimentos. A KEA necessita de uma lista de palavras-chaves selecionadas manualmente por um especialista de domínio; a Text-To-Onto foi desenvolvida para ser utilizada com *corpora* em inglês; e a PORONTO foi desenvolvida para *corpora* em português. Os experimentos realizados são descritos a seguir.

4.1 TERMINAE

Esta ferramenta está disponível sob a forma de um programa executável (extensão jar, que indica ser um programa codificado na linguagem Java). Para inicializar a ferramenta, é necessário executar o comando "JAVA –jar <arquivo>.jar", adicionando o nome do arquivo cuja extensão é jar. A ferramenta se propõe a fazer a extração semiautomática de termos, gerando uma lista de termos candidatos, que podem ser selecionados manualmente pelo usuário.

Foi realizado o *download*, instalação e execução da ferramenta. A ferramenta está disponível apenas em Francês (Figura 2), dificultando o uso por pessoas que não conhecem esta língua. Há um manual em francês sobre a sua utilização. Para ser possível testar a ferramenta, é preciso instalar uma ferramenta auxiliar ao processo de extração de termos, denominada LEXTER. Entretanto, esta última não está disponível para o *download*, limitando, assim, o uso do TERMINAE.

Figura 2 – Interface gráfica do TERMINAE



Fonte: AUTOR

4.2 ONTOLP

ONTOLP é um *plugin* para o ambiente de construção de ontologias Protégé[1]. É de fácil instalação, por ser um *plugin*, sendo necessário apenas desempacotá-lo no diretório de instalação do Protégé. Seu uso é fácil e

[1] Disponível em: <HTTP://protege.stanford.edu/>. Acesso em: 10/02/2009

extrai os termos rapidamente. A principal dificuldade reside na necessidade de anotar o *corpus* em outra ferramenta que não é gratuita, antes de utilizá-lo no ONTOLP.

O método proposto utiliza como entrada um *corpus* anotado no padrão XCES (*Xtra Computer Equipament Services*), um padrão baseado em XML para codificar *corpora*. Para a identificação dos termos candidatos, Ribeiro Junior (2008) definiu as seguintes etapas:

a)extração dos grupos semânticos (opcional);

b)filtragem dos grupos semânticos irrelevantes feita pelo usuário (opcional);

c)extração de termos simples considerando apenas aqueles pertencentes aos grupos selecionados;

d)exclusão dos termos simples irrelevantes realizada pelo usuário;

e)extração dos termos complexos considerando apenas aqueles que possuem no mínimo uma palavra presente na lista final de termos simples e que pertençam a um grupo selecionado; e

f)exclusão dos termos complexos irrelevantes, feita pelo usuário.

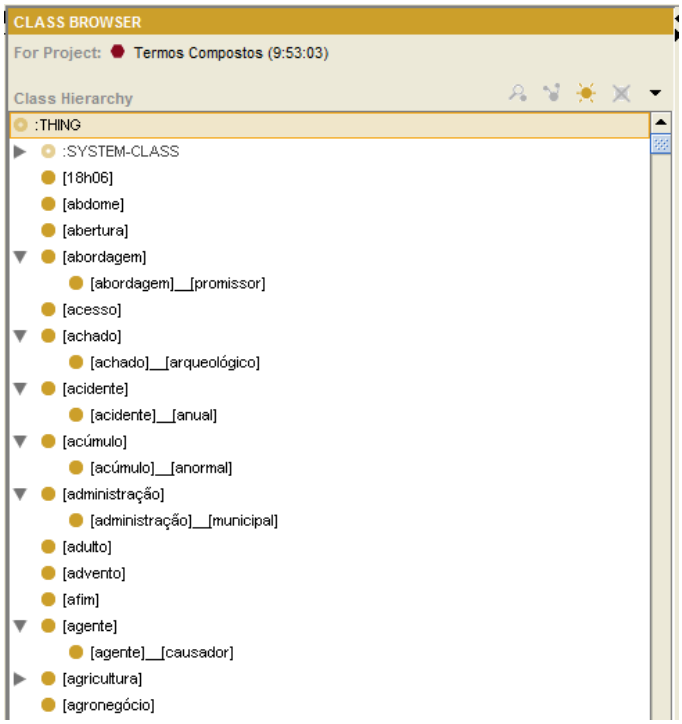
Para o cálculo de relevância dos termos simples, foram implementados os métodos: *FR* (MANNING; SCHÜTZE, 1999), *tf-idf* (MANNING; SCHÜTZE, 1999), *NC-Value* (FRANTZI; ANANIADOU; TSUJII, 1998) e para o cálculo de relevância dos termos complexos, foram utilizadas os métodos *FR, tf-idf, C-Value e NC-Value* (FRANTZI; ANANIADOU; TSUJII, 1998).

Para a tarefa de identificação de relações taxonômicas, são utilizados os termos extraídos e avaliados anteriormente. Com base nesses termos são aplicados três métodos diferentes: identificação de relações taxonômicas com base em termos complexos, identificação de relações taxonômicas através dos padrões de (HEARST, 1992*)* e identificação de relações taxonômicas através dos padrões de (MORIN; JACQUEMIN, 2004).

Foi realizado o *download* e a instalação do *plugin* no Protégé. É disponibilizado um *corpus* já anotado no padrão XCES para testes, mas não é disponibilizada a ferramenta utilizada para realizar este procedimento, impossibilitando a execução de testes com outros *corpus*.

O primeiro passo realizado foi a importação do *corpus* disponibilizado. Após a inserção do *corpus*, foram extraídos 1.272 termos simples e 531 compostos. Não foi possível calcular a taxa de acerto, pois não se sabe o número de palavras no *corpus.* O terceiro passo foi gerar a taxonomia (Figura 3), não sendo possível avaliar o resultado, pois não se sabe qual o domínio do *corpus*, o nome "CIENCIA" do arquivo é muito abrangente.

Figura 3 – Taxonomia no ONTOLP



Fonte: AUTOR

4.3 KEA: Keyphrases Extraction Algorithm

Esta ferramenta está disponível como um projeto do IDE Eclipse, assim, pode-se importá-la no IDE e executá-la. A ferramenta se propõe a fazer a extração automática de termos para a construção de uma ontologia. Para fins de comparação com os resultados produzidos pela ferramenta, foi utilizado um conjunto de documentos com termos selecionados manualmente por um especialista do domínio como referência e a medida *tfidf* como medida de extração.

Foi realizado o *download*, o desempacotamento e a importação do KEA no Eclipse. Sendo o projeto *open source* e sendo fornecidos os arquivos necessários para a importação no eclipse, o processo foi facilitado. Após esta etapa, a ferramenta foi executada conforme o manual que é fornecido junto à ferramenta.

Além do código fonte, o KEA disponibiliza arquivos para realizar os testes da ferramenta. O Quadro 3 ilustra o resultado da execução da ferramenta com um dos arquivos teste fornecido, no qual a primeira coluna apresenta os termos selecionados automaticamente pela ferramenta e a segunda, os termos selecionados manualmente por um especialista do domínio.

Quadro 3 – Termos extraídos pelo KEA

Extraídas Automaticamente	Selecionadas Manualmente
Bangladesh	Bangladesh
Cooperative farming	Development aid
Family planning	Famine
Famine	Food supply
First aid	Poverty
Food aid	Rural environment
Jute	Rural population
Landlessness	Social structure
Landowners	Tenure
Merchants	Villages

Fonte: AUTOR

A ferramenta extrai termos, mas a sua utilização é deficiente por dois motivos: a falta de uma interface gráfica e a necessidade de um especialista para criar previamente uma lista de termos de um determinado domínio, dificultando o uso por usuários menos experientes.

4.4 TEXT-TO-ONT

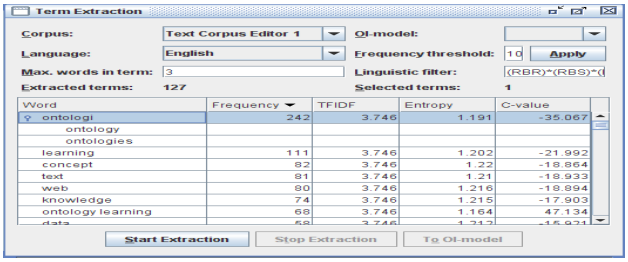
Foi realizado o *download* da ferramenta, o desempacotamento e a execução. A execução da ferramenta é fácil, pois estão disponíveis com o pacote de instalação dois arquivos executáveis, um de extensão *sh* (para Linux) e um de extensão *bat* (para Windows).

A ferramenta se propõe a fazer a extração semiautomática de termos, extraindo uma lista de termos candidatos e disponibilizando para o usuário selecionar os termos corretos, a descoberta de relacionamentos (não hierárquicos) e a geração da taxonomia dos termos extraídos.

Neste experimento, foi utilizado um *corpus* com 15 artigos científicos na língua inglesa sobre aprendizagem de ontologias. A extração dos termos foi realizada utilizando quatro medidas: frequência, *tf-idf*, entropia e *c-value*.

Após selecionar o *corpus*, a ferramenta extrai os termos, descobre as relações taxonômicas e descobre os relacionamentos não-hierárquicos. Com os termos extraídos (Figura 4), pode-se acionar outra funcionalidade chamada de *Taxo Builder* para criar automaticamente uma taxonomia com os termos processados anteriormente.

Figura 4 – Termos extraídos no Text-To-Onto



Fonte: AUTOR

A funcionalidade de criar uma taxonomia apresenta duas abordagens, a *FCA-based* e a *combination-based*, pelas quais a ferramenta combina o padrão de *Hearst* com padrões heurísticos. A opção *FCA-based* apresentou erro ao ser executada, sendo, então, utilizada a opção *combination-based,* a qual também apresentou problemas, não sendo possível avaliar o resultado da geração de relações taxonômicas. Outra funcionalidade analisada foi a de importar/exportar ontologias em outros formatos. A ferramenta disponibiliza apenas a opção de importar/exportar em formato RDF. Para o experimento, foi importada uma ontologia de exemplo disponibilizada pelo editor de ontologias Protégé.

A ferramenta é intuitiva, sendo de fácil aprendizado, mas apresenta problemas de desempenho, mesmo com um *corpus* com apenas três documentos. Cabe ressaltar que Text-To-Onto é uma ferramenta que agrega vários algoritmos de aprendizagem de ontologias. Estes algoritmos identificam conceitos, relações taxonômicas e não taxonômicas. Porém, não especialistas possuem dificuldades em selecionar o algoritmo, bem como identificar cada uma das respectivas funcionalidades.

4.5 PORONTO

A ferramenta disponibiliza um WAR, este deve ser inserido em um servidor WEB JAVA e sua instalação é feita de maneira automática. É uma ferramenta para construção semiautomática de ontologias em português, desenvolvida em tecnologia de código aberto e gratuita. O processo de *construção semiautomática de ontologias foi dividido em duas etapas: a criação do corpus e a criação da ontologia*.

A *criação do corpus* é dividida em cinco fases:

1)o usuário envia os artigos que deseja processar em formato PDF;

2)os artigos são transformados em arquivos texto limpo (sem os marcadores padrões de arquivos PDF;

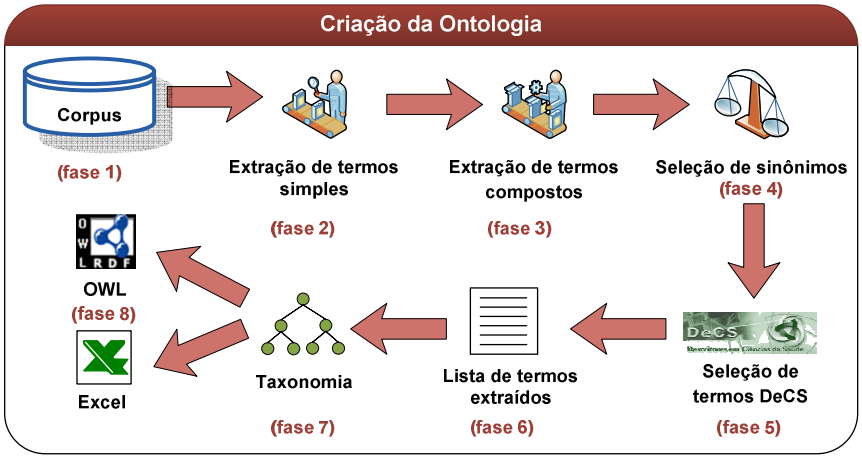
3)ocorre um pré-processamento dos textos, onde o texto é dividido por espaços em branco para posteriormente ser feito o processamento de anotação linguística com a ferramenta *TreeTagger*;

4)as *stopwords* são removidas; e

5)os textos são processados com o *TreeTagger*.

Com o *corpus* construído, a etapa de *criação da ontologia pode* ser iniciada, a qual é dividida em oito fases (Figura 5):

Figura 5 – Processo da criação da ontologia do PORONTO



Fonte: ZAHRA (2009)

1)o usuário preenche os filtros;

os termos simples são extraídos aplicando as medidas frequência, *tfidf* e entropia;os termos compostos são extraídos com base em regras expressas no Quadro 4 e aplicam-se as medidas de frequência, *tfidf* e entropia;

Quadro 4 – Regras de identificação de termos compostos no PORONTO

Regras de sequência morfológica
Substantivo + Adjetivo
Substantivo + Preposição + Substantivo
Substantivo + Preposição + Adjetivo + Substantivo
Substantivo + Preposição + Substantivo + Preposição + Substantivo

Fonte: ZAHRA (2009)

1)uma busca por sinônimos dos termos é realizada na lista do *OpenThesaurusPT* (OPENTHESAURUSPT, 2010), para facilitar o critério de seleção do termo pelo usuário;

2)é realizada uma pesquisa para verificar se os termos extraídos possuem correspondência na lista de Descritores em Ciência da Saúde (DeCS), novamente para facilitar o critério de seleção do termo pelo usuário;

3)os termos simples e os compostos, extraídos pela ferramenta, assim como alguns sinônimos para estes termos e se o termo está ou não incluído na lista de descritores, são apresentados para o usuário, para que seja feita a seleção dos termos mais relevantes que devem ser inseridos na ontologia;

4)após selecionar os termos relevantes, o usuário pode optar por solicitar a organização destes termos em uma taxonomia; e

5)o usuário pode exportar o resultado para o formato XLS ou OWL.

Neste experimento, foi utilizado um *corpus* com 24 artigos científicos na língua portuguesa sobre câncer de mama. A extração dos termos foi realizada seguindo as oito fases descritas anteriormente. A ferramenta é intuitiva, sendo fácil sua utilização, porém não identifica relações não taxonômicas.

5 Conclusões

A aprendizagem de ontologia, a partir de textos, pode ser útil para a construção de uma ontologia ou para o enriquecimento de uma já existente, para diferentes finalidades. No entanto, o objetivo de se construir uma ontologia automaticamente está longe de ser alcançado.

Há várias ferramentas para aprendizagem de ontologias a partir de textos. Muitas delas resultam de trabalhos acadêmicos e poucas delas estão disponíveis para *download* e continuam a receberem incrementos. Além disso, algumas ferramentas não são de fácil utilização, principalmente quando empregadas por pessoas que não são da área da computação.

Tais ferramentas requerem o entendimento de algoritmos, métricas e fórmulas específicas e são destinadas principalmente a realização de experimentos científicos e, em última análise, a engenheiros do conhecimento (mas não a usuário que são especialistas do domínio). Outro entrave é que as ferramentas disponíveis representam apenas algumas etapas de todo o processo e desenvolvimento de uma ontologia, não estando inseridas em uma metodologia maior, destinada ao desenvolvimento de ontologias.

A maioria dos métodos apresentados para aprendizagem de ontologias a partir de textos baseia-se principalmente em técnicas de processamento de linguagem natural e medidas estatísticas. Todos estes métodos requerem a participação de um engenheiro do conhecimento e de um especialista do domínio para avaliar a ontologia final e, também, o andamento do processo. Não existem métodos automáticos que avaliem com exatidão o processo de aprendizagem. A avaliação de ontologias é um tópico por si só complexo e normalmente é medido pela capacidade que a ontologia oferece em responder a um conjunto de questões as quais ela pretensamente deveria trazer respostas.

Mesmo sendo uma área que ainda necessita de muito avanço, o uso destas ferramentas auxilia os atores envolvidos na construção/enriquecimento de ontologias, reduzindo o tempo e o custo do desenvolvimento por meio da extração de conhecimento dos textos. Porém, no estágio de desenvolvimento atual destas técnicas, sempre caberá aos atores humanos a decisão final de incluir um conceito ou de se fazer algum relacionamento entre conceitos na ontologia.

Referências

BASÉGIO, T. Uma abordagem semiautomática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil.

05/01/1997. 124 p. Dissertação (Mestrado em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Porto Alegre, 05/011997.

BIÉBOW, B; SZULMAN S. CLÉMENT, Av. J. B. TERMINAE: a linguistic-based tool for the building of a domain ontology. In: EKAW- International Conference on Knowledge Engineering and Knowledge Management, 11th, 1999, Dagstuhl, Proceedings of the 11th European Workshop on Knowledge Acquisition, Modelling and Management, Berlin, Springer, 1999, p. 49-66.

BUITELAAR, P.; MAGNINI, B. Ontology learning from text: an overview. In BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. (Eds.), *Artificial Intelligence and Applications Series*. The Netherlands: IOS Press, v.123, 2005. p. 3-12.

CIMIANO, P. Ontology learning and population from text: algorithms, evaluation and applications. New York: Springer, 2006. 347 p.

CIARAMITA, M.; GANGEMI, A.; RATSCH, E.; SARIC, J.; ROJAS, I. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In: International Joint Conference on Artificial Intelligence, 19th, 2005, Barcelona, Proceedings of the 19th International Joint Conference on Artificial Intelligence, San Francisco, CA, USA, Morgan Kaufmann Publishers, 2005. p. 1-6.

FAURE, D.; NEDELLEC, C. A corpus-based conceptual clustering method for verb

frames and ontology acquisition. In: Velardi, P. (Ed.), Proceedings LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications, 1st, 1998, Granada, Espanha, Proceedings of the First Workshop on adapting lexical and corpus resources to sublanguages and applications, 1998, p. 5–12.

FRANTZI, K. T.; ANANIADOU, S.; TSUJII, J. The c-value/nc-value method of automatic recognition for multi-word terms. In: European Conference on Research and Advanced Technology for Digital Libraries, 2nd, 1998, Heraklion, Crete, Grécia, Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, London, UK, Springer, 1998. p. 585-604.

FREITAS, F; SCHULZ; S. Ontologies, semantic Web and health. *RECIIS, R. Eletr. de Com. Inf. Inov. Saúde*. Rio de Janeiro, v. 3, n. 1, p. 4-7, mar. 2009.

GÓMEZ-PÉREZ, A.; MANZANO-MACHO D. *A survey of ontology learning methods and Techniques.* Espanha: Universidad Politécnica de Madrid, 2003.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, v. 5, p. 199-220, 1993.

HAASE, P.; VÖLKER, J. Ontology learning and reasoning- dealing with uncertainty and inconsistency. In: Workshop on Uncertainty Reasoning of the Semantic Web, Galway, Irlanda, Proceedings of the Workshop on Uncertainty Reasoning of the Semantic Web, Berlin, Springer, 2005, p. 45-55.

HEARST, M. A. Automatic acquisition of hyponyms from large text corpora. In: Conference on Computational linguistics, 14th, 1992, Nantes, França, Proceedings of the 14th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics, 1992, p. 539-545.

JONES, S.; PAYNTER, G. W. Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. *Journal of the American Society for Information Science and Technology (JASIST)*, v.53, n.8, p. 653-677, 2002.

LIN, D.; PANTEL, P. Discovery of inference rules from text. In: Conference on Knowledge Discovery and Data Mining, 7th, 2001, São Francisco, CA, USA, Proceedings of

the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, USA, ACM, 2001, p. 323-328.

MAEDCHE, A.; STAAB, S. Ontology learning. In: STAAB, S.; STUDER, R. (Eds.). *Handbook on ontologies*. Berlin: Springer, 2004. p. 173-189.

MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing.* Cambridge, Massachusetts: The MIT Press, 1999. Disponível em: <citeseer.ist.psu.edu/635422.html>. Acesso em: 20/10/2008.

MIKHEEV, A.; FINCH, S. A workbench for finding structure in texts. In: Fifth Conference on Applied Natural Language Processing, 5th, 1997, Washington, DC. Proceedings of the Fifth Conference on Applied Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics, 1997, p. 372-379

MILLER, G. A. *et al.* Introduction to WordNet: an on-line lexical database. *International*

Journal of Lexicography, v. 3, n. 4, p. 235-244, 1990.

MILLER, G. A. WordNet: a lexical database for English. *Communications of*

*the ACM***,** v. 38, n. 11, p. 39-41, 1995.

MORIN, E.; JACQUEMIN, C. Automatic acquisition and expansion of hypernym links. *Computers and the Humanities*, v. 38, n. 4, p. 363-396, 2004.

NOY, N. F.; MUSEN, M. A. Promptdiff: a fixed-point algorithm for comparing ontology versions. In: 18th National Conference on Artificial Intelligence, 18th, 2002, Edmonton, Canada. Proceedings of the 18th National Conference on Artificial Intelligence, Palo Alto, California, AAAI PRESS, 2002, p.1-7.

OPENTHESAURUSPT. *Um projeto Open Source para a construção de um Dicionário de Sinônimos para a língua portuguesa.* Disponível em: <http://openthesaurus.caixamagica.pt >. Acesso em: 26 abr. 2010.

RIBEIRO JUNIOR, L. C. OntoLP: construção semiautomática de ontologias a partir de Textos da Língua Portuguesa. 2008. 158 p. Dissertação (Mestrado Computação Aplicada) - Universidade do Vale do Rio dos Sinos - UNISINOS, 2008.

RYU, P.; M.; CHOI, K.S. Taxonomy learning using term specificity and similarity. In: Workshop on Ontology Learning and Population, 2nd, 2006, Sydney, Australia, Proceedings of the 2nd Workshop on Ontology Learning and Population, Stroudsburg, PA, USA, Association for Computational Linguistics, 2006, p. 41–48.

STAAB, S.; ERDMANN, M.; MADCHE, A. Engineering ontologies using semantic patterns. In: IJCAI Workshop on E-Business and Intelligent Web, 1st, 2001, Seattle, Washington, USA, Proceedings of the IJCAI Workshop on E-Business and Intelligent Web, Palo Alto, California, AAAI PRESS, 2001, p. 174-185.

STUDER, R.; BENJAMINS, R.; FENSEL, D. Knowledge engineering: principles and methods. *Data and Knowledge Engineering*, v. 25, p. 161-197, 1998.

VELARDI, P.; NAVIGLI R.; MISSIKOFF M. *Integrated approach for Web ontology learning and engineering*. Los Alamitos, CA: IEEE Computer Society Press, 2002.

VOSSEN, P. *EuroWordNet: a multilingual database with lexical semantic networks*.

Dordrecht, The Netherlands: Kluwer Academic Publishers, 1998.

ZAHRA, F. M. PORONTO: ferramenta para construção semiautomática de ontologias em português. 92 p. 17/12/2009. Dissertação (Mestrado Ciência da Computação) - Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba, 17/12/2009.

ZAHRA, F. M.; CARVALHO, D. R.; MALUCELLI, A. PORONTO: ferramenta para construção semiautomática de ontologias em português. *Journal of Health Informatics*,. v. 5, n. 2, abr-jun. 2013. http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/232/167. 22/07/2013.