Um Processo Semi-Automático para o Povoamento de Ontologias a partir de Fontes Textuais

Carla Faria[1], Rosario Girardi[2]

[1] Instituto Federal de Educação, Ciência e Tecnologia do Maranhão - IFMA, Departamento de Informática, São Luís, Maranhão, Brasil

[2] Universidade Federal do Maranhão – UFMA, Departamento de Informática, São Luís, Maranhão, Brasil

carlafaria@ifma.edu.br, rosariogirardi@gmail.com

Abstract. *The knowledge acquisition is a costly, complex and expensive process that requires a domain expert. Therefore, it becomes essential to create a semi-automatic or automatic of this process. Ontology population is an approach for the semi-automatic or automatic instantiation of concepts, relationships and properties of an ontology. Ontology population with speed and low cost is crucial to the success of knowledge-based applications. This article proposes a process for semi-automatic population of ontologies (PSAPO) from textual resources. Some experiments using a legal corpus were conducted in order to evaluate it. Initial results are promising and indicate that our approach can extract instances with good effectiveness.*

Resumo. *A aquisição de conhecimento é um processo de alto custo, complexo e caro que requer um especialista de domínio. Por isso, torna-se fundamental uma semi-automatização ou automatização desse processo. O povoamento de ontologias constitui uma abordagem para automatizar ou semi-automatizar a instanciação de classes, propriedades e relacionamentos de ontologias. O povoamento de ontologias com rapidez e baixo custo é crucial para o sucesso de aplicações baseadas em conhecimento. Este artigo propõe um processo semi-automático para o Povoamento de Ontologias (PSAPO) a partir de fontes textuais. Experimentos foram conduzidos na área do direito de família para avaliar o processo proposto e os resultados iniciais foram promissores.*

1. Introdução

As ontologias constituem uma abordagem para a representação de conhecimento capaz de expressar um conjunto de entidades, seus relacionamentos, restrições e regras de um determinado domínio [Guarino, Masolo e Vetere 1999] [Nierenburg e Raskin 2004]. São utilizadas pelos modernos sistemas baseados em conhecimento para representar e compartilhar o conhecimento sobre um domínio de aplicação. Permitem o processamento semântico das informações e, através de interpretações mais precisas das informações, os sistemas apresentam maior efetividade e usabilidade. [Girardi 2010]

Povoamento de Ontologias é o termo usado para designar as técnicas utilizadas para extração e classificação de instâncias de classes, relacionamentos e propriedades de uma ontologia. O povoamento manual de ontologias por especialistas de domínio e engenheiros de conhecimento é uma tarefa cara, tediosa e demorada, razão pela qual é necessária a semi-automatização ou automatização desse processo.

Este artigo propõe um Processo Semi-Automático para o Povoamento de Ontologias (PSAPO) a partir de fontes textuais, baseado em técnicas de processamento de linguagem natural [Allen 1995] [Dale, Moisl e Somers 2000] e de extração de informação [Cowie e Wilks 2000] [Cunningham 2005].

O artigo está organizado da seguinte maneira. A seção 2 introduz a definição de ontologia utilizada pelo processo. A seção 3 apresenta o processo proposto. A seção 4 descreve os experimentos conduzidos para avaliação. A seção 5 resume os trabalhos relacionados e finalmente a seção 6 apresenta as considerações finais.

2. Uma definição de Ontologia

Uma ontologia é uma especificação formal explícita de uma conceituação compartilhada de um domínio de interesse [Guarino, Masolo e Vetere 1999] [Nierenburg e Raskin 2004].

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

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

onde,

$C = C_C \cup C_I$ é o conjunto de entidades do domínio sendo modelado. O conjunto C_C é formado por classes, ou seja, conceitos que representam entidades que descrevem um conjunto de objetos (por exemplo, "Mãe" $\in C_C$) enquanto que o conjunto C_I é formado por instâncias, ou seja, entidades únicas no domínio (por exemplo, "Anne Smith" $\in C_I$).

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

$I = \{é\_um\ (c_1,c_2) \mid c_1 \in C_I \wedge c_2 \in C_C\ \} \cup \{prop_K\ (c_i,valor) \mid c_i \in C_I\} \cup \{rel_K\ (c_1, c_2, ...c_n) \mid \forall i, c_i \in C_I\}$, é o conjunto de relacionamentos entre os elementos da ontologia e suas instâncias, por exemplo "é_um("Anne Smith", Mãe)", "data_de_nascimento ("Anne Smith", "12/02/1980")" e "mãe_de("Anne Smith", "Clara Smith")" são relacionamentos entre classes, relacionamentos, propriedades e suas instâncias.

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

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

$A = \{condition_x \Rightarrow conclusion_y\ (c_1,c_2,..., c_n)\ \forall j, c_j \in C_C\}$ é um conjunto de axiomas, regras que permitem checar a consistência da ontologia e deduzir novos conhecimentos através de algum mecanismo de inferência. O termo $condition_x$ é dado por: $condition_x = \{\ (cond_1,cond_2,\ldots,cond_n) \mid \forall_z, cond_z \in H \cup I \cup R\}$. Por exemplo, "Mãe, Filha1, Filha2, mãe_de(Mãe, Filha1), mãe_de(Mãe, Filha2) $\Rightarrow$ irmã_de(Filha1, Filha2)" é uma regra que indica que se duas filhas têm a mesma mãe, então as filhas são irmãs.

3. Processo Semi-Automático para o Povoamento de Ontologias

O Processo Semi-Automático para o Povoamento de Ontologias (PSAPO) a partir de fontes textuais proposto neste artigo utiliza técnicas de Processamento de Linguagem Natural (PLN) [Allen 1995] [Dale, Moisl e Somers 2000] e de Extração de Informação (EI) [Cowie e Wilks 2000] [Cunningham 2005]. Consiste de duas fases: "Extração e Classificação de Instâncias" e "Representação de Instâncias" (Figura 1).

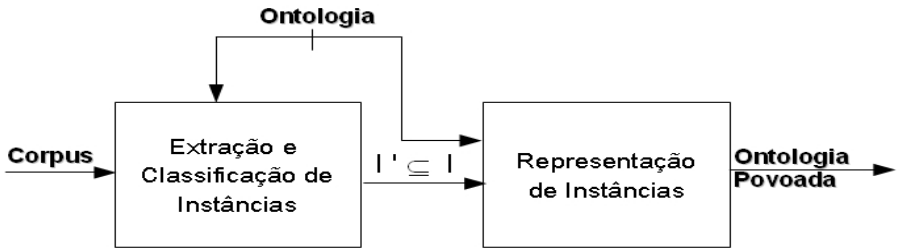

Figura 1. Processo Semi-Automático para o Povoamento de Ontologias.

A fase de "Extração e Classificação de Instâncias" visa extrair o subconjunto I' do conjunto I da definição de ontologia apresentada na seção 2, composto de instâncias de classes, relacionamentos e propriedades. Por exemplo, para um corpus na área do direito de família I' = {é_um(mãe,"Anne Smith"), mãe_de("Anne Smith", "Clara Smith"), data_de_nascimento("Anne Smith", "12/02/1980")}.

A fase de "Representação de Instâncias" visa à instanciação de classes, relacionamentos e propriedades da ontologia e sua especificação utilizando uma linguagem de representação de ontologias, como OWL [OWL 2010]. Como produto, temos a ontologia povoada.

3.1. Extração e Classificação de Instâncias

A fase de "Extração e Classificação de Instâncias" consiste de três tarefas: "Anotação do Corpus", "Construção de Regras de Extração e Classificação" e "Extração e Classificação de Instâncias" (Figura 2).

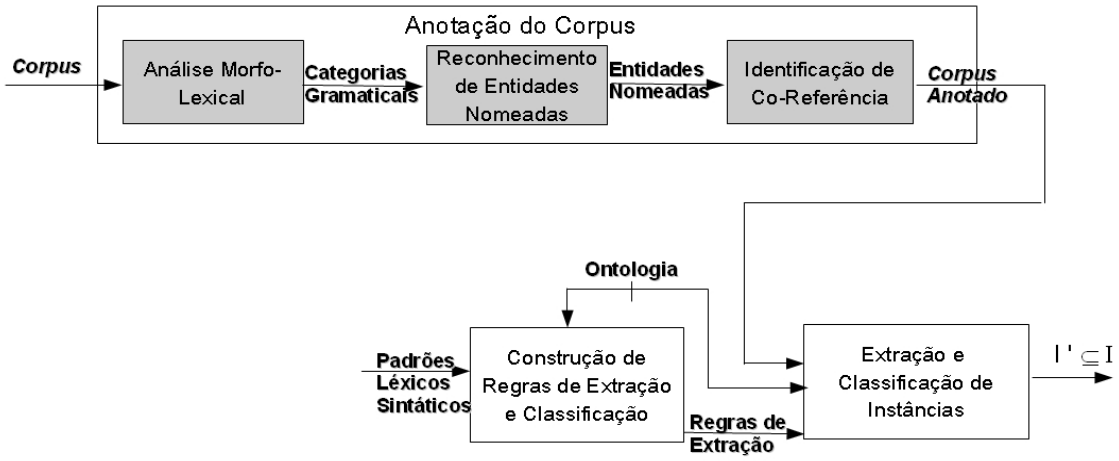

Figura 2. Extração e Classificação de Instâncias.

A primeira tarefa, "Anotação do Corpus", visa à aplicação de técnicas de processamento de linguagem natural sobre um conjunto de documentos textuais. Esta

primeira tarefa é realizada através de três atividades: “Análise Morfo-Lexical”, “Reconhecimento de Entidades Nomeadas” e “Identificação de Co-Referências”. A “Análise Morfo-Lexical” visa a identificação das categorias gramaticais. O “Reconhecimento de Entidades Nomeadas” busca detectar nomes que se referem a objetos exclusivos do mundo, prováveis instâncias de classes e de propriedades da ontologia. A “Identificação de Co-Referências” visa a identificação de co-referências pronominais e nominais, ou seja, nomes ou pronomes que se referem a uma mesma entidade descrita previamente no texto.

A segunda tarefa, “Construção de Regras de Extração e Classificação”, é realizada de forma independente da primeira tarefa e busca a construção de regras de extração e classificação baseada em padrões léxicos sintáticos e no conhecimento do domínio representado através de uma ontologia de domínio. Uma regra de extração e classificação é definida como:

se <condição> **então classificar** <classificação>

onde:

<condição> representa uma expressão regular a ser identificada no corpus anotado.

<classificação> representa a forma de classificar a(s) instância(s) identificada(s) na <condição>.

As regras de extração e classificação são construídas a partir de relacionamentos, classes e propriedades da ontologia de domínio e de padrões léxicos sintáticos. Uma regra de extração e classificação construída a partir de um relacionamento da ontologia de domínio a ser instanciada pode ser representada da seguinte forma:

se NomePróprio1 relacionamento NomePróprio2 **então classificar** NomePróprio1 e NomePróprio2 como instância do relacionamento.

Uma regra de extração e classificação construída a partir de uma classe da ontologia a ser instanciada pode ser representada da seguinte forma:

se NomePróprio (Classe) **então classificar** Nome Próprio como instância da Classe

Uma regra de extração e classificação construída a partir de um padrão léxico-sintático combinado com uma classe na ontologia a ser instanciada obedece à estrutura do padrão. Para o padrão “such as”, temos:

se Classe such as NomePróprio **então classificar** Nome Próprio como instância da Classe

A terceira tarefa, “Extração e Classificação de Instâncias”, envolve a aplicação das regras construídas na segunda tarefa “Construção de Regras de Extração e Classificação”, sobre o corpus anotado resultante da aplicação da primeira tarefa, “Anotação do Corpus”, gerando o conjunto $I' \subseteq I$. Para cada regra é realizada uma busca de expressões regulares que representem a condição estabelecida na regra. Caso essas expressões sejam encontradas então elas são extraídas e, em seguida, é executada a ação definida na regra produzindo instâncias de classes, relacionamentos ou propriedades.

Para uma melhor compreensão desta fase, cada uma das tarefas e atividades descritas é ilustrada através de exemplos. Para a tarefa “Anotação do Corpus”, considere como entrada o fragmento de texto de um documento no domínio do direito de família (Figura 3), que, por razão de simplicidade, contém apenas um parágrafo.

Petitioner Keith R. (Father) and real party in interest H.R. (Mother) were married in 2004. B.R. (Daughter) was born in 2005. fn. 2 After Father filed for divorce in September 2006, Mother asserted domestic violence allegations against Father, and requested sole custody of Daughter. Following an investigation and a hearing, the court (Judge Claudia Silbar) denied Mother's requests. In February 2007, the court (Judge Pollard) entered an order granting both parents joint legal and physical custody, and appointed a child custody evaluator, who recommended maintaining the current custody arrangements based on Daughter's parental attachments.

Figura 3. Fragmento de texto de um documento no domínio do direito de família [Macedo 2010].

A atividade “Análise Morfo-Lexical” consiste em atribuir a cada palavra um rótulo com a sua categoria sintática como, por exemplo, substantivo, verbo, adjetivo, dentre outros. O resultado desta atividade para o fragmento de texto da Figura 3 é mostrado na Figura 4. As tags utilizadas neste artigo são aquelas do conjunto de tags Penn Treebank [Marcus, Santorini, Marcinkiewicz 1993]. Note que tais tags referem-se à língua inglesa.

Petitioner [NNP] Keith [NNP] R [NNP]. (Father) [NNP] and [CC] real[JJ] party [NN] in [IN] interest [NN] H [NNP] . R [NNP]. (Mother) [NNP] were [VBD] married [VBN] in [IN] 2004 [CD]. B [NNP] . R [NNP] . (Daughter) [NNP] was [VBD] born [VBN] in [IN] 2005 [CD]. [NN]. 2 [LS] After [IN] Father [NNP] filed [VBD] for [JJS] divorce [NN] in [IN] September [NNP] 2006 [CD], Mother [NNP] asserted [VBD] domestic [JJ] violence [NN] allegations [NNS] against [IN] Father [NNP], and [CC] requested [VBD] sole [JJ] custody [NN] of [IN] Daughter [NNP]. Following [VBG] an [DT] investigation [NN] and [CC] a [DT] hearing [NN], the [DT] court [NN] (Judge [NNP] Claudia [NNP] Silbar [NNP]) denied [VBD] Mother [NNP] 's [POS] requests [NNS]. In [IN] February [NNP] 2007 [CD], the [DT] court [NN] (Judge [NNP] Pollard [NNP]) entered [VBD] an [DT] order [NN] granting [VBD] both [DT] parents [NNS] joint [JJ] legal [NN] and [CC] physical [JJ] custody [NN], and [CC] appointed [VBZ] a [DT] child [NN] custody [NN] evaluator [NN], who [WP] recommended [VBD] maintaining [VBG] the [DT] current [JJ] custody [NN] arrangements [NNS] based [VBN] on [IN] Daughter [NNP] 's [POS] parental [JJ] attachments [NNS].

Figura 4. Resultado da análise morfo-lexical do fragmento de texto da Figura 3.

O resultado da atividade “Reconhecimento de Entidades Nomeadas” para o fragmento de texto da Figura 3 é um conjunto de três entidades do tipo “Pessoa”: “Keith R.”, “B. R.” e “H. R.”.

O produto da tarefa “Construção de Regras de Extração e Classificação” são as regras. Para a construção das regras foram considerados a classe “casamento” da ontologia que representa parte do domínio do direito de família (Figura 5) e padrões léxicos sintáticos.

Marriage		
wife_member	Instance	Person
child_member	Instance*	Person
		Person
husband_member	Instance	Person
Dissolution_Date	String	
Constitutive_Date	String	

Figura 5. Classe casamento da ontologia do direito de família.

Por exemplo, considerando a propriedade data de constituição - "*constitutive_date*", a seguinte regra de extração e classificação é criada:

se *casados em data* **então classificar** data como instância da propriedade data de constituição - "constitutive_date"

Por exemplo, considerando a propriedade data de dissolução - "*dissolution_date*" a seguinte regra de extração e classificação é criada:

se divórcio em data **então classificar** data como instância da propriedade data de dissolução - *"dissolution_date"*

Por exemplo, considerando os relacionamentos esposa - "*wife_member*" e marido - "*husband_member*" a seguinte regra de extração e classificação é criada:

se NomePróprio1 e NomePróprio2 foram casados **então classificar** NomePróprio1 como instância do relacionamento marido - *"husband_member"* e o NomePróprio2 como instância do relacionamento esposa - *"wife_member"*

Por exemplo, considerando o relacionamento filho - "*child_member*" a seguinte regra de extração e classificação é criada:

se NomePróprio (Filha) **então classificar** NomePróprio como instância do relacionamento filho -*"child_member"*

O produto da tarefa "Extração e Classificação de Instâncias" considerando o corpus anotado resultante da tarefa "Anotação do Corpus" e as regras de extração e classificação construídas na tarefa "Construção de Regras de Extração e Classificação" é extraído o conjunto I' = {"*wife_member*"(" *Marriage1*", "H. R."), "*child_member*"(" *Marriage1*", "B. R"), "*husband_member*"("*Marriage1*", "Keith R."), "*dissolution_date*"("*Marriage1*", "2006"), "*constitutive_date*"("*Marriage1*", "2004")}.

3.2. Representação de Instâncias

A fase "Representação de Instâncias" consiste de duas tarefas: "Seleção de Instâncias" e "Instanciação" (Figura 6).

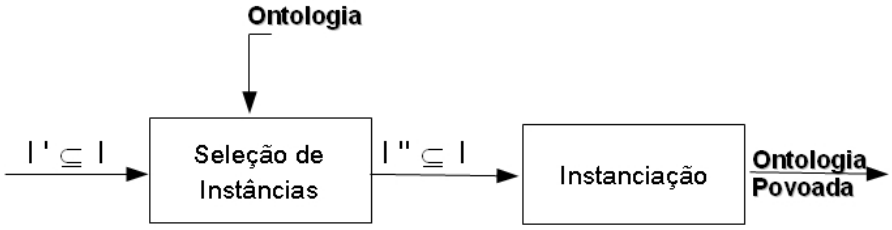

Figura 6. Representação de Instâncias.

A tarefa "Seleção de Instâncias" visa à eliminação de instâncias duplicadas do conjunto I', gerando o conjunto I''. Para cada instância do conjunto I' é realizada uma busca na ontologia para descobrir se a instância já existe ou não. Se a instância não existir na ontologia, a instância fará parte do conjunto I''. Se a instância já existir na ontologia a instância será descartada.

A tarefa “Instanciação” visa a efetiva criação das instâncias. Para cada instância do conjunto I’’ é feita uma busca pela classe, relacionamento ou propriedade ao qual a instância pertence, quando identificada a instanciação é feita gerando uma ontologia povoada especificada em uma linguagem de representação de ontologias.

Para uma melhor compreensão desta fase, cada uma das tarefas é ilustrada através de exemplos. Considere o conjunto I’ = {“*wife_member*”(“ *Marriage1*”, “H. R.”), “*child_member*”(“ *Marriage1*”, “B. R”), “*husband_member*”(“*Marriage1*”, “Keith R.”), “*dissolution_date*”(“ *Marriage1*”, “2006”), “*constitutive_date*”(“ *Marriage1*”, “2004”)} produto da fase de “Extração e Classificação de Instâncias” e a classe casamento da ontologia que representa parte do domínio do direito de família (Figura 5) contendo a classe C^C = {“*marriage*”}.

A tarefa “Seleção de Instâncias” faz uma busca na ontologia para a seleção das instâncias únicas gerando o conjunto I’’ = {“*wife_member*”(“ *Marriage1*”, “H. R.”), “*child_member*”(“ *Marriage1*”, “B. R”), “*husband_member*”(“*Marriage1*”, “Keith R.”), “*dissolution_date*”(“ *Marriage1*”, “2006”), “*constitutive_date*”(“ *Marriage1*”, “2004”)}.

A tarefa “Instanciação” tem como entrada o conjunto I’’ = {“*wife_member*”(“ *Marriage1*”, “H. R.”), “*child_member*”(“ *Marriage1*”, “B. R”), “*husband_member*”(“*Marriage1*”, “Keith R.”), “*dissolution_date*”(“ *Marriage1*”, “2006”), “*constitutive_date*”(“ *Marriage1*”, “2004”)} onde é primeiro localizado o relacionamento “*wife_member*” e é feita a instanciação de “H. R.”; em seguida é localizado o relacionamento “*child_member*” e é feita a instanciação de “B. R.”; em seguida é localizado o relacionamento “*husband_member*” e é feita a instanciação de “Keith R.”; em seguida é localizada a propriedade “*dissolution_date*” e é feita a instanciação de “2006”; por fim é localizada a propriedade “*constitutive_date*” e é feita a instanciação de “2004”. Como produto temos a classe “*marriage*” povoada da ontologia (Figura 7).

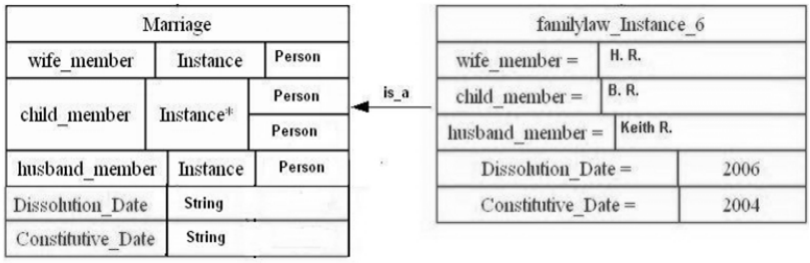

Figura 7. Exemplo da classe casamento instanciada da ontologia do direito de família.

4. Avaliação

Um estudo de caso na área do direito de família foi conduzido para uma avaliação preliminar da efetividade do processo proposto. Para tanto foi desenvolvido o protótipo de uma ferramenta para automatização do processo utilizando a ferramenta GATE [Cunningham, Maynard, Bontcheva e Tablan 2002]. As regras de extração e classificação foram especificadas na linguagem JAPE [Cunningham, Maynard, Bontcheva e Tablan 2002].

Foi utilizado no estudo de caso o corpus denominado FamilyJuris, composto de 919 documentos textuais, capturados a partir do site “family.findlaw.com” contendo casos de jurisprudência do direito de família norte-americano [Macedo 2010].

A ontologia FamilyLaw adotada no estudo de caso foi desenvolvida na ferramenta Protégé [Noy, Fergerson e Musen 2000] e descreve o conhecimento do Direito de Família. A FamilyLaw (Figuras 8 e 9), consiste de uma classe raiz “Direito de Família”, e as subclasses “Entidade Familiar” e “Pessoa”. A classe “Entidade Familiar” descreve as principais entidades que legalmente são consideradas família e a classe “Pessoa” discrimina os elementos pessoais constituintes de uma família.

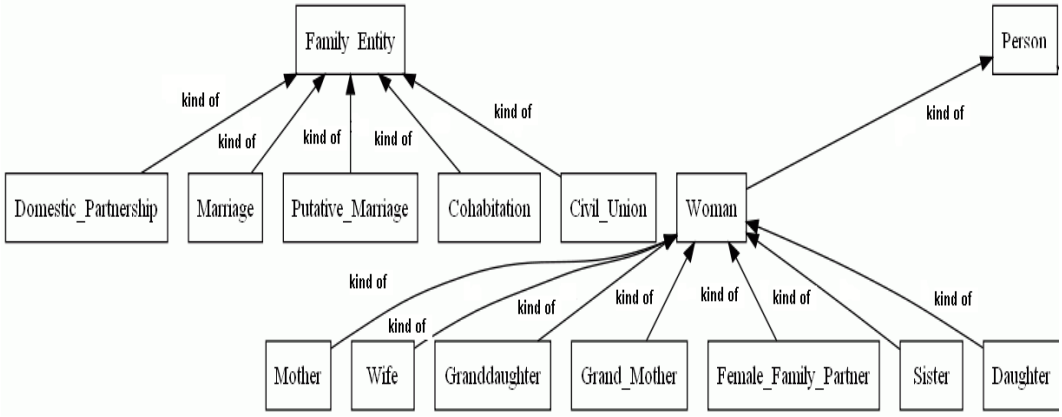

Figura 8. Parte 1 da estrutura de classes da ontologia *FamilyLaw*.

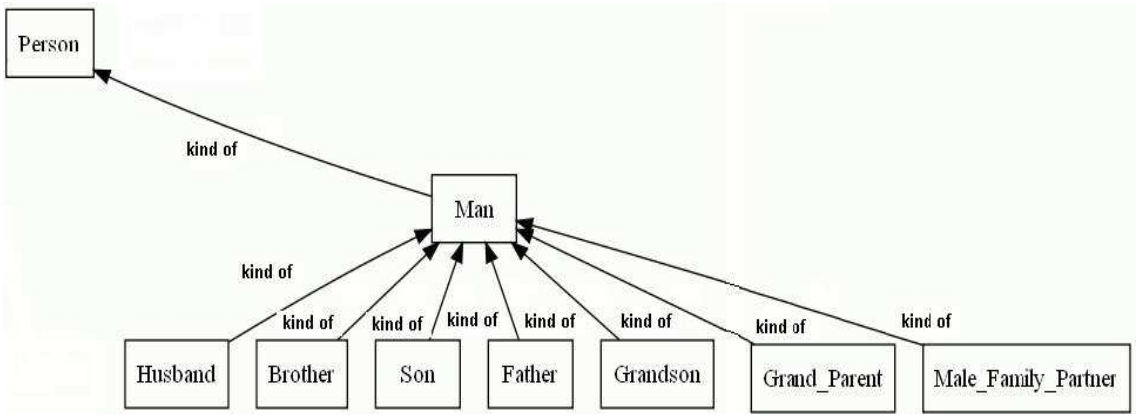

Figura 9. Parte 2 da estrutura de classes da ontologia *FamilyLaw*.

As regras de extração e classificação foram implementadas em uma gramática JAPE, que consiste em um conjunto regras < condição, ação >. Uma gramática JAPE tem dois lados: o Esquerdo e o Direito. O lado esquerdo da gramática contém uma expressão regular a ser detectada no conjunto de documentos. O lado direito descreve a ação a ser tomada sobre a expressão regular detectada. A Figura 10 ilustra dois exemplos de regras especificadas em JAPE com o uso da classe “mãe” e da classe filha da ontologia FamilyLaw. O lado esquerdo, que representa a expressão regular a ser detectada, contém instâncias da classe “mãe” na primeira regra e da classe “filha” na segunda regra, enquanto o lado direito estabelece que a ação a ser executada é classificar “Nomes Próprios” como instâncias da classe “mãe” na primeira regra e classificar “Nomes Próprios” como instâncias da classe “filha” na segunda regra.

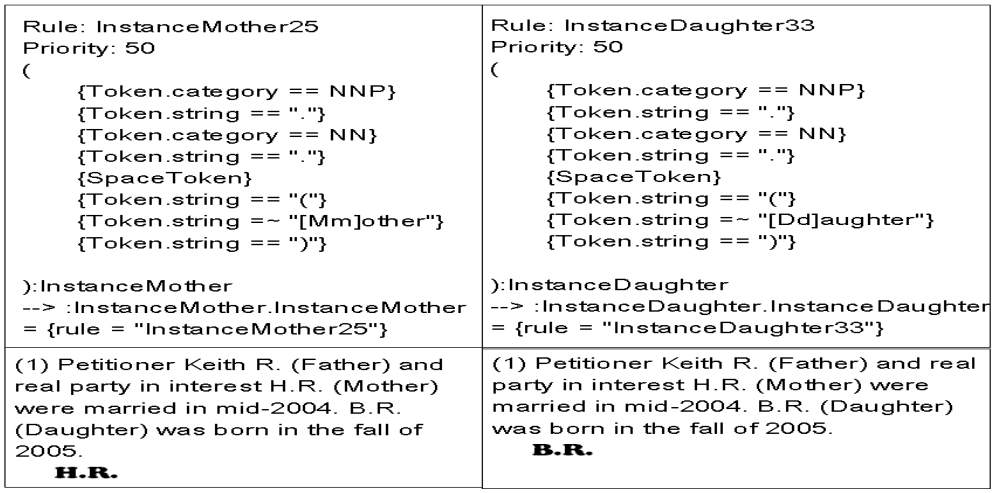

Figura 10. Regras JAPE para as classes mãe e filha.

Uma adaptação das medidas clássicas *precision* e *recall* da área de Recuperação de Informação [Dellschaft e Staab 2006] foram utilizados para avaliar a efetividade considerando o número de instâncias classificadas corretamente.

Precision é a razão entre o número de instâncias extraídas corretamente (NIEC) e o número de instâncias extraídas (NIE).

$$P = NIEC / NIE$$

Recall é a razão entre o número de instâncias extraídas corretamente (NIEC) e o número de instâncias no corpus (NIC).

$$C = NIEC / NIC$$

A taxa de *precision* dos resultados obtidos na aplicação do processo em uma parte do corpus (231 documentos aleatórios) foi de 95% e a taxa de recall foi de 85%.

5. Trabalhos Relacionados

A Tabela 1 mostra as principais abordagens automáticas e semi-automáticas para o povoamento de ontologias.

Tabela 1. Abordagens para o Povoamento Automático de Ontologias

Abordagens	**Técnicas**	**Ferramentas**	***Precision***
Cimiano e Volker	AM, PLN e EI	Pankow	36,82%
Cimiano et. al.	AM, PLN e EI	C-Pankow	74,37%
Craven et. al.	AM e PLN	WEB-KB	74%
Etzioni et. al.	AM, PLN e EI	KnowItAll	90%
Evans	AM e EI	NERO	92,25%
Fleischman e Hovy	AM e PLN	MenRun	70,4%
Giuliano e Gliozo	AM	-	62,3%
Karkaletsis et. al.	AM, PLN e EI	M-PIRO e NLG	-
Ruiz-Martinez et. al.	PLN e EI	GATE	92,5%
Tanev e Magnini	AM, PLN e EI	MiniPar	65%

Das abordagens analisadas para o povoamento de ontologias nota-se que são baseadas no Processamento da Linguagem Natural (PLN), Aprendizagem de Máquina (AM) e/ou Extração de Informação (EI).

As abordagens como Cimiano e Volker [Cimiano e Volker 2005], Cimiano et. al. [Cimiano, Ladwig e Staab 2005], Craven et. al. [Craven et. al. 2000], Etzioni et. al. [Etzioni et. al. 2005], Evans [Evans 2003], Fleischman e Hovy [Fleischman e Hovy 2002], Giuliano e Gliozo [Giuliano e Gliozo 2008], Karkaletsis et. al. [Karkaletsis, Valarakos e Spyropoulos 2006] e Tanev e Magnini [Tanev e Magnini 2006] utilizam algoritmos de aprendizagem de máquina (AM) para a classificação das instâncias descobertas, enquanto que o processo proposto e Ruiz Martinez et. al. [Ruiz-Martínez et. al. 2008] utilizam a Extração de Informação (EI) para a classificação das instâncias descobertas.

As abordagens como Cimiano e Volker [Cimiano e Volker 2005], Cimiano et. al. [Cimiano, Ladwig e Staab 2005], Craven et. al. [Craven et. al. 2000], Etzioni et. al. [Etzioni et. al. 2005], e Giuliano e Gliozo [Giuliano e Gliozo 2008] utilizam como fonte a Web. Enquanto que as abordagens como Evans [Evans 2003], Fleischman e Hovy [Fleischman e Hovy 2002], Karkaletsis et. al. [Karkaletsis, Valarakos e Spyropoulos 2006], Tanev e Magnini [Tanev e Magnini 2006], Ruiz Martinez et. al. [Ruiz-Martínez et. al. 2008] e o processo proposto utilizam como fonte Corpus.

As abordagens como Cimiano e Volker [Cimiano e Volker 2005], Cimiano et. al. [Cimiano, Ladwig e Staab 2005], Evans [Evans 2003], Fleischman e Hovy [Fleischman e Hovy 2002], Giuliano e Gliozo [Giuliano e Gliozo 2008] e Tanev e Magnini [Tanev e Magnini 2006] fazem a instanciação somente de classes. As abordagens como Craven et. al. [Craven et. al. 2000], Etzioni et. al. [Etzioni et. al. 2005] e Karkaletsis et. al. [Karkaletsis, Valarakos e Spyropoulos 2006] fazem a instanciação de classes e relacionamentos. Enquanto que a abordagem de Ruiz Martinez et. al. [Ruiz-Martínez et. al. 2008] faz a instanciação de classes, relacionamentos e propriedades.

Das abordagens analisadas somente Karkaletsis et. al. [Karkaletsis, Valarakos e Spyropoulos 2006] e o processo proposto fazem o Povoamento de Ontologias de forma semi-automática.

A efetividade de cada abordagem em termos de precisão é apresentada na Tabela 1. Entretanto, estes números não podem ser considerados para propósitos comparativos entre as abordagens, pois a avaliação de cada uma foi conduzida utilizando diferentes corpora e ontologias.

A vantagem do processo proposto neste artigo, em relação aos outros trabalhos apresentados na Tabela 1, é que ele providencia um processo sistemático para o Povoamento de Ontologias.

6. Considerações Finais

O povoamento de ontologias tem se tornado uma tarefa essencial para a construção de bases de conhecimento, uma vez que contribui decisivamente para a redução dos custos associados à sua construção.

O processo semi-automático para o povoamento de ontologias proposto neste artigo é baseado no processamento da linguagem natural e na extração de informação e

consiste de duas fases: “Extração e Classificação de Instâncias” e “Representação de Instâncias”.

O processo mostrou-se viável, posto que o estudo de caso conduzido com o corpus *FamilyJuris*, com o objetivo de povoar a ontologia *FamilyLaw*, apresentou bons resultados.

Os resultados obtidos demonstram que o uso de técnicas de processamento da linguagem natural em conjunto com a extração de informação representa uma abordagem promissora para o povoamento de ontologias.

Atualmente, está sendo abordada a geração automática de regras, com o intuito de reduzir o esforço na construção de regras de extração e classificação de instâncias e automatizar o Processo para o Povoamento de Ontologias. O desenvolvimento de uma ferramenta para a execução de todo o processo proposto também está sendo desenvolvida.

Ainda com o intuito de avaliar o processo proposto pretende-se desenvolver uma abordagem baseada em aprendizagem de máquina para a extração e classificação de instâncias de ontologia.

Referências

Allen J. (1995) “Natural Language Understanding”, Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc.

Cimiano P., Volker J. (2005) “Towards large-scale, open-domain and ontology-based named entity classification”, In: Proceedings of RANLP’05, p. 166–172, Borovets, Bulgaria.

Cimiano P., Ladwig G., Staab S. (2005) “Gimme 'the context: Context-driven automatic semantic annotation with C-PANKOW”, In: Proceedings of the 14th World Wide Web Conference (WWW), p. 332-341.

Cowie J., Wilks Y. (2000) “Information Extraction”, Handbook of Natural Language Processing, Robert Dale, Hermann Moisl and Harold Somers, p. 241–260.

Craven M., DiPasquo D., Freitag D., McCallum A., Mitchell T., Nigam K., Slattery S. (2000) “Learning to construct knowledge bases from the world wide web”, Artificial Intelligence, 118:69-113.

Cunningham H. (2005) “Information Extraction”, Encyclopedia of Language and Linguistics, 2nd Edition.

Cunningham H., Maynard D., Bontcheva K., Tablan V. (2002) “GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications”, In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July.

Dale R., Moisl H. and Somers H. L. (2000) “Handbook of natural language processing”, CRC.

Dellschaft K., Staab S. (2006) “On how to perform a gold standard based evaluation of ontology learning”, In: Proceedings of the 5th International Semantic Web Conference, p. 228 – 241, Athens. Springer.

Etzioni O., Cafarella M., Downey D., Popescu A. M., Shaked T., Soderland S., Weld D.,Yates A. (2005) “Unsupervised named-entity extraction from the web: An experimental study”, Artificial Intelligence, 165(1):91-134.

Evans R. (2003) “A framework for named entity recognition in the open domain”, In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), p. 137-144.

Fleischman M., Hovy E. (2002) “Fine Grained Classification of Named Entities”, In: Proceedings of COLING, Taipei, Taiwan, August.

Girardi, R. (2010) “Guiding Ontology Learning and Population by Knowledge System Goals”, In: Proceedings of International Conference on Knowledge Engineering and Ontology Development, Ed. INSTIIC, Valence, October, p. 480-484.

Giuliano C., Gliozzo A. (2008) “Instance-Based Ontology Population Exploiting Named-Entity Substitution”, In: Proceedings of the The 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, August, p. 18-22.

Guarino N., Masolo C., Vetere C. (1999) “Ontoseek: Content-based Access to the web”, IEEE Intelligent Systems, v. 14(3), p. 70-80.

http://www.w3.org/2001/sw/WebOnt/, Acessado em 30 de Agosto de 2010.

Karkaletsis V., Valarakos A., Spyropoulos C. D. (2006) “Populating ontologies in biomedicine and presenting their content using multilingual generation”, Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine, Italy, Genoa.

Macedo M. J. C. (2010) “Processamento da Linguagem Natural para Identificação de Classes e Instâncias de uma Ontologia”, Monografia de Graduação, CGCC-UFMA.

Marcus, M., Santorini, B., Marcinkiewicz, M. (1993) “Building a Large Annotaded Corpus of English: Penn TreeBank”, Computational linguistics: Special Issue on Using Large Corpora, [S. I.], v. 19, n.2, p. 313 – 330.

Nierenburg S., Raskin V. (2004) Ontological Semantics, MIT Press.

Noy N. F., Fergerson R., Musen M. A. (2000) “The knowledge model of Protege-2000: Combining interoperability and flexibility”, In: Proceedings of the 2th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000), Juan-les-Pins, France.

Ruiz-Martínez J. M., Miñarro-Giménez J. A., Guillén-Cárceles L., Castellanos-Nieves D., Valencia-García R., García-Sánchez F., Fernández-Breis J. T., Martínez-Béjar R. (2008) “Populating Ontologies in the eTourism Domain”, In: Proceedings of the 2008 IEEE/WIC/ACM international Conference on Web intelligence and intelligent Agent Technology - Volume 03. Web Intelligence & Intelligent Agent. IEEE Computer Society, Washington, DC, December 09 – 12, p. 316-319.

Tanev H., Magnini B. (2006) “Weakly Supervised Approaches for Ontology Population”, In: Proceedings of of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), p. 17-24.