Sistema de Aquisição semi-automática de Ontologias

Gabriel Gonçalves[1], **Rodrigo Wilkens**[1], **Aline Villavicencio**[1,2]

[1]Instituto de Informática, Universidade Federal do Rio Grande do Sul (Brasil)

[2]CSAIL, MIT (EUA)

gabrielgonc@gmail.com, {rwilkens,avillavicencio}@inf.ufrgs.br

Abstract. *This paper presents an ongoing work on ontology learning from text, focusing on the acquisition of concepts and relations. In order to do that, this work investigates approaches for ontology learning, and presents a proposal based on graphs metrics to identify concepts, and text analysis to find relations between the concepts.*

Resumo. *Este artigo apresenta um trabalho em andamento na área de aprendizado de ontologias a partir de texto, focando na identificação de conceitos e relações. Para isto, este trabalho investiga abordagems para o aprendizado de ontologias e apresenta uma proposta baseada métricas de grafos para identificar conceitos, e análise do texto com os conceitos encontrados para obter relações.*

1. Introdução

Em alguns sistemas computacionais como sistemas de perguntas e repostas e agentes conversacionais, para suprir as necessidades de informações de usuários, pode ser necessário utilizar informações não-estruturadas, como as disponíveis na web, e realizar um processamento dessas informações. Para tanto, diversas linguagens e padrões vem sendo desenvolvidos, tais como *Resource Description Framework* [3] e *Web Ontology Language* [1], que permitem a definição de conceitos e a descrição de suas relações e propriedades. Segundo o W3C (World Wide Web Consortium) [13], para sistemas que precisam compartilhar conhecimentos do mesmo domínio (por exemplo, medicina, mercado imobiliário e petróleo) é necessário o uso de ontologias para unificar este conhecimento. Contudo, o processo de criação de ontologias de forma manual é custoso em termos de tempo e recursos e exige um especialista do domínio. Desta forma, algumas tarefas desse processo tem sido automatizadas em sistemas computacionais, como mostrado em [16], [18], [11] e [6].

Em geral o aprendizado automático de ontologias é visto como a aquisição de conhecimento a partir de textos, onde grande parte do trabalho utiliza como base áreas da computação como processamento de linguagem natural, inteligência artificial e aprendizado de máquina [2]. Para Yang e Jamie [18] o processo de construção de ontologias ocorre em quatro passos: (1) detectar candidatos a conceitos; (2) agrupar conceitos similares; (3) encontrar um nome para cada grupo; (4) formar uma árvore para representar a ontologia.

Para muitas línguas e domínios o aprendizado de ontologias tem que ser realizado a partir de poucos recursos linguísticos disponíveis. Nesse contexto, este trabalho objetiva

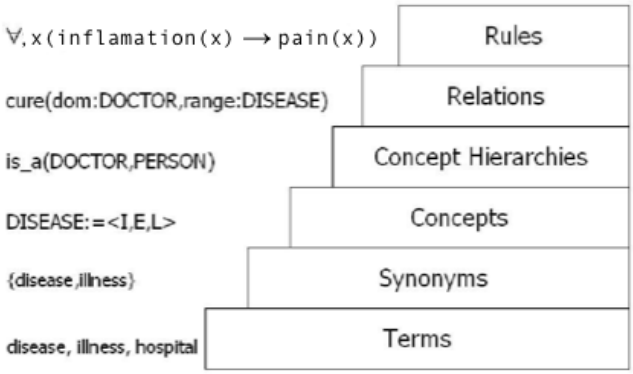

Figura 1. Hierarquia dos processos de aprendizado de ontologia [2]

investigar dois aspectos do aprendizado de ontologias, a identificação de conceitos e de relações entre conceitos, focando na identificação de conceitos simples e na identificação de elementos que indicam relações entre termos. Para tanto esse trabalho inicia com uma revisão do estado da arte, na seção 2. A seguir, na seção 3 são apresentadas as técnicas utilizados na abordagem proposta. Na seção 4 são discutidas as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Gruber [10] define uma ontologia como uma especificação formal e explícita de uma conceitualização compartilhada por um domínio de interesse, onde formal significa que a ontologia deve ser interpretável por computador e aceita por um grupo ou comunidade da área que a ontologia modela. Além disso, deve ser restrita a um dado domínio de interesse e, portanto, modelar conceitos e relações relevantes a uma tarefa ou aplicação particular do domínio [2]. Atualmente não há um consenso sobre os métodos para o aprendizado automático de ontologias, que segundo [2], podem ser divididos em seis níveis: termos, sinônimos, conceitos, hierarquias de conceitos, relações e regras. A hierarquia dessas tarefas no processo de aprendizado de ontologias é mostrada na Figura 1.

A aquisição de termos consiste em encontrar automaticamente palavras que representem conceitos de um domínio. Este é o passo inicial do aprendizado de ontologias, sendo seus resultados usados em todas as etapas posteriores. As técnicas mais utilizadas para tanto são a indexação de termos, análise de frequência, coocorrência e uma combinação dos dois métodos anteriores [14]. Segundo Buitelaar [2], a extração de conceitos é uma etapa controversa, por não estar claro o que exatamente é um conceito. Nesta etapa podem ser considerados como conceitos uma definição, instancias de um conceito ou um conjunto multilíngue de termos, dependendo do uso que o pesquisador da ontologia gerar.

A identificação de sinônimos visa a aquisição semântica de variantes de termos, ou seja, encontrar entre os termos de um texto aqueles que compartilham funções semânticas. Para tanto, o estado da arte mapeia a semântica de cada palavra e identifica as palavras que possuem intersecção, sendo este mapeamento comumente realizado pelo contexto dos termos [3] ou diretamente pala semântica dos termos [17].

A extração de taxonomias busca identificar uma organização hierárquica entre

os conceitos, sendo comum o uso de listas de termos que indicam tais relações, o que gera uma boa precisão na identificação, mas devido ao fato destes padrões serem muito específicos esta abordagem apresenta uma baixa cobertura das relações existentes [11]. Outra abordagem é a hipótese de distribuição, onde são derivadas automaticamente as hierarquias de termos a partir do texto usando análise de conceitos formais [8] (ex. [4], [7], [9]). A comunidade de recuperação de informação trata esta tarefa a partir da avaliação da distribuição e relevância dos termos nos documentos, como mostrado por Sanderson e Croft em [15].

A extração de outras relações não hierárquicas entre conceitos (por exemplo, relações entre sintomas, doenças e drogas) tem sido feita a partir de textos, em geral procurando por relações entre pares de conceitos com mesma classe gramatical.

Por fim, a extração de regras, discutida em [12] e [5], é a área pesquisada menos abordada em aprendizado de ontologias [2]. O objetivo deste passo é encontrar regras gramaticais que rejam as relações das ontologias.

Dentro desse contexto, esse trabalho é similar ao de [3] no uso de mutual information para a extração de sinonimia, com a difierença de que utilizamos esta métrica sobre um grafo do texto, e não diretamente sobre ele, e a [16] que verificam relações, diferindo por generalizarmos os padrões encontrados.

3. Metodologia

O objetivo deste trabalho é gerar automaticamente ontologias a partir de um corpus do domínio, com foco na identificação de conceitos e relações do domínio, discutidos respectivamente nas seções 3.1 e 3.2.

3.1. Aquisição de Termos e Conceitos

Neste trabalho não diferenciamos termos e conceitos no processo de aquisição devido à natureza próxima destes, assim tornando o resultante do sistema mais próximo de uma ontologia linguística de domínio. O processo inicia com a geração de um grafo a partir do corpus, onde as palavras são os nós, que são ligados uns aos outros quando as palavras que formam os nós encontram-se na mesma sentença, como ilustrado na Figura 2. Nas Figuras 2.i e 2.ii, as frases "João e Maria foram ao parque domingo"e "Domingo o parque estava lotado", respectivamente, são transformadas em grafos. As duas frases unidas geram um grafo, cujas arestas são pesadas de acordo com o número de vezes que cada par de nós coocorre no texto. (Figura 2.iii). Sobre este grafo utilizamos as seguintes métricas de grafos para gerar candidatos a conceitos:

- **centralidade** para verificar a importância do nó no grafo,
- **grau**, que representa o numero de ligações de um nó e
- ***closeness***, que verifica a média dos caminhos mínimos para se chegar ao nó.

3.2. Aquisição de Relações

Para a obtenção das relações não hierárquicas realizamos uma análise do corpus para identificar possíveis expressões que indiquem alguma relação entre os termos. Este processo foi dividido em três etapas sequenciais: extração de relações; generalização das relações para obter padrões; e re-extração das relações utilizando os padrões encontrados.

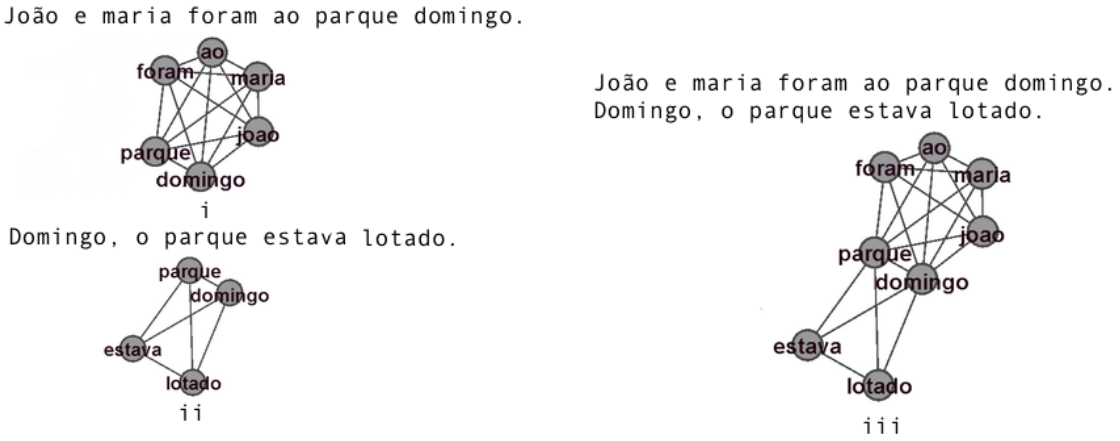

Figura 2. Exemplo de texto transformado em grafo.

Para a extração de relações o sistema identifica no corpus todos os conceitos e segmenta as palavras que ocorrem entre eles.[1] Todas as palavras que se encontram entre um par de conceitos são consideradas candidatas a relação. Estas relações candidatas são filtradas, permanecendo apenas palavras cujas classes gramaticais são permitidas (neste ponto utilizamos filtros que combinam informações lexicais e morfosintáticas para uma extração mais direcionada). Desta forma é obtida a primeira lista de relações entre conceitos (este processo é exemplificado na Figura 3, onde duas relações distintas são encontradas para a frase[2] entre os conceitos *obras* e *licensas*, e *distribuição* e *trabalhos*).

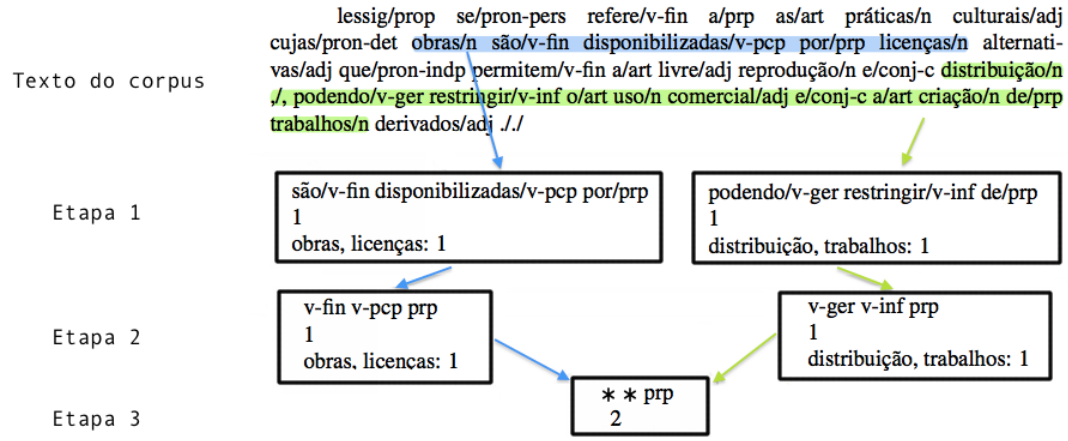

Figura 3. Extração de relações entre conceitos.

Na segunda etapa, generalização das relações, consideramos as relações apenas como uma sequência de classes gramaticais (no exemplo da Figura 3, etapa 2, as palavras são substituídas por suas classes gramaticais). Estas relações formam uma segunda lista, onde estão as relações compostas de classes gramaticais e suas respectivas frequências.

[1] Assume-se que não pode haver um conceito entre um par de conceitos.

[2] A frase está anotada com suas classes gramaticais (prop: nome próprio, pron-pers: pronome pessoal, v-fin: verbo finito, prp: preposição, art: artigo, adj: adjetivo, pron-det: pronome determinado, n: substantivo, v-pcp: verbo no particípio, pron-ind: pronome indeterminado, conj-c: conjunção coordenada, v-ger: verbo no gerúndio, v-inf: verbo no infinitivo).

Neste ponto, as relações são generalizados de acordo com seu número de palavras e de classes gramaticais que compartilham a mesma posição. Na Figura 3, etapa 3, as duas relações têm o mesmo tamanho e compartilham o mesmo elemento na posição três, gerando uma nova relação genérica contendo três elementos, restringindo apenas o terceiro.

O objetivo da primeira etapa é mostrar as relações que ocorrem diretamente no corpus, enquanto a segunda etapa objetiva criar padrões genéricos de identificação. Com estas informações, a terceira etapa, re-extração das relações, utiliza a lista gerada pela etapa 2 como modelo para identificar novas relações no corpus, ou seja, relações que não foram identificadas na primeira etapa.

4. Conclusões e Trabalhos Futuros

O aprendizado de ontologias é um campo interdisciplinar, que abrange diversas áreas da computação, como processamento de linguagem natural. As propostas para aprendizado semi-automático de ontologias permitem diminuir consideravelmente o custo e esforço envolvidos na construção de ontologias.

Dentro desse contexto, esse trabalho apresentou uma abordagem baseada em grafos para a identificação de termos e relações a partir de corpora. Essa abordagem permite extrair de forma recursiva novas expressões que PODEM indicar relações entre termos.

Como trabalhos futuros se prevê uma avaliação sistemática dos resultados obtidos, por cada etapa do processo, por um especialista do domínio. Os trabalhos futuros envolvem ainda a aquisição de sinônimos e aquisição de relações hierárquicas, assim permitindo alem da identificação das relações gerais, aquelas relações mais específicas (por exemplo, “tipo de”, “é um”). Pretendemos também validar os resultados obtidos com o sistema utilizando corpus de diferentes domínios, como o corpus GENIA [3] do domínio de biologia.

Agradecimentos

Esta pesquisa tem apoio dos projetos COMUNICA (FINEP/SEBRAE 1194/07), CAPES-COFECUB (707/11) e CNPq (479824/2009-6, 202007/2010-3 e 309569/2009-5).

Referências

[1] Resource description framework (rdf) model and syntax, 2011.

[2] P. Buitelaar, P. Cimiano, and B. Magnini. Ontology learning from text: An overview. *Ontology learning from text: Methods, evaluation and applications*, 123:3–12, 2005.

[3] A. Chotimongkol and A.I. Rudnicky. Automatic concept identification in goal-oriented conversations. In *Seventh International Conference on Spoken Language Processing*, 2002.

[4] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24(1):305–339, 2005.

[5] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. *Machine Learning Challenges*, pages 177–190, 2006.

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

[6] E. Drymonas. Ontology learning from text based on multi-word term concepts: The ontogain method. Master's thesis, Department of Electronic and Computer Engineering, Technical University of Crete, Greece, 2009.

[7] D. Faure and C. Nédellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*, pages 707–728. Citeseer, 1998.

[8] B. Ganter and R. Wille. Formal concept analysis. *WISSENSCHAFTLICHE ZEITSCHRIFT-TECHNISCHEN UNIVERSITAT DRESDEN*, 45:8–13, 1996.

[9] G. Grefenstette. *Explorations in automatic thesaurus discovery*. Springer, 1994.

[10] T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, 43(5):907–928, 1995.

[11] M.A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.

[12] D. Lin and P. Pantel. Discovery of inference rules from text, April 5 2001. US Patent App. 09/826,355.

[13] D.L. McGuinness, F. Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10:2004–03, 2004.

[14] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval 1. *Information processing & management*, 24(5):513–523, 1988.

[15] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213. ACM, 1999.

[16] F.M. Suchanek, G. Ifrim, and G. Weikum. Leila: Learning to extract information by linguistic analysis. In *Proceedings of the ACL-06 Workshop on Ontology Learning and Population*, pages 18–25, 2006.

[17] F. Venant. Semantic visualization and meaning computation. In *22nd International Conference on on Computational Linguistics: Demonstration Papers*, pages 185–188. Association for Computational Linguistics, 2008.

[18] H. Yang and J. Callan. Metric-based ontology learning. In *Proceeding of the 2nd international workshop on Ontologies and information systems for the semantic web*, pages 1–8. ACM, 2008.