Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006) - Poster Section, Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

# Metodologia Semi-Automática Baseada em *Corpus* para a Construção de Ontologias de Domínio

Ariani Di Felippo[1,2], Sandra M. Aluísio[1], Gladis M. B. Almeida[1,3], Leandro H. M. de Oliveira[1], Luiz Carlos Genoves Jr.[1], Lucas Antiqueira, Osvaldo Novais de Oliveira Jr.[1,4]

[1] Núcleo Interinstitucional de Lingüística Computacional (NILC/ICMC/USP), CP 668
13.560-970 São Carlos, SP, Brazil
{sandra}@icmc.usp.br,{lhmoliveira, genoves, lantiq}@gmail.com

[2] Universidade Estadual Paulista (CELiC/UNESP), CP 174
14.800-901 Araraquara, SP, Brazil
arianidf@uol.com.br

[3] Universidade Federal de São Carlos (GETerm/UFSCar), CP 676
13.565.905 São Carlos, SP, Brazil
gladis@power.ufscar.br

[4] Instituto de Física de São Carlos (IF/USP), CP 369
13560-970 São Carlos, SP, Brazil
chu@if.sc.usp.br

**Resumo.** Apresenta-se, neste artigo, uma metodologia semi-automática baseada em *corpus* para a construção de ontologias de domínio. A metodologia é baseada em princípios teóricos e práticos de duas áreas de pesquisa: da Lingüística, especificamente da Terminologia e do Processamento Automático das Línguas Naturais, mais especificamente da subárea preocupada com a extração de termos a partir de *corpus*. A proposta de metodologia semi-automática foi aplicada na construção de uma ontologia para a área de Nanociência e Nanotecnologia.

**Palavras-chave:** ontologia de domínio, *corpus*, extração automática de termos, terminologia, processamento de línguas naturais.

## 1 Introdução

Uma *ontologia* ou *estrutura conceitual* pode ser entendida como "uma especificação de uma conceitualização" [1], ou seja, um modelo comum ou estrutura conceitual sistematizada e de consenso que permite não só armazenar, mas também buscar e recuperar a informação sobre um determinado domínio do conhecimento. As ontologias podem ser classificadas de acordo com diferentes critérios. No âmbito do domínio a que se aplicam, as ontologias podem ser *gerais* (descrevem conceitos mais gerais como espaço, tempo, matéria, objeto, etc.) ou *de domínio restrito* (descrevem conceitos de um campo específico do conhecimento como Medicina, Geologia, etc.) [2]. As ontologias de domínio restrito (ou simplesmente *de domínio*), em especial, têm papel fundamental na confecção de glossários ou vocabulários e em aplicações ou sistemas computacionais que analisam, compreendem e/ou geram (traduzem) documentos técnicos [3]. A construção de ontologias de domínio, no entanto, é tarefa pouco trivial, já que os processos que a compõem (extração dos candidatos a termo e

posterior alocação dos mesmos em uma estrutura conceitual) são tradicionalmente manuais e, portanto, demorados [4].

É natural, portanto, que se busquem métodos para automatizar um ou mais dos processos. Neste trabalho, apresenta-se uma proposta metodológica semi-automática baseada em *corpus* para construção de ontologias de domínio. Para tanto, toma-se por base princípios teórico-metodológicos advindos da Terminologia [5], [6] e do Processamento Automático das Línguas Naturais (PLN) [7]. As etapas constitutivas da metodologia semi-automática, as estratégias/ferramentas adotadas em cada uma delas e o estudo de caso para a área de Nanociência e Nanotecnologia (N&N) são apresentados na Seção 2. A Seção 3 conclui e apresenta trabalhos futuros.

## 2 Proposta Metodológica Semi-automática

Seguindo a tendência atual dos estudos terminológicos descritivos e trabalhos desenvolvidos pelo Grupo de Estudos e Pesquisas em Terminologia (GETerm)[1], considera-se que a organização conceitual de uma área de especialidade pressupõe as atividades teórico-metodológicas apresentadas em (a)-(e). Essas atividades são descritas e exemplificadas tomando-se por base a área de N&N.

**a) Delimitação da área-objeto.** A construção de uma ontologia de domínio requer a escolha de uma área-objeto. A N&N é uma área emergente e, por isso, sua organização conceitual é importante para (i) a confecção de glossários e (ii) o tratamento computacional de documentos técnicos dessa área. Além disso, uma ontologia de N&N poderia subsidiar a concepção e organização do Portal da Rede de Nanotecnologia da USP (http://www.usp.br/prp/nanotecnologia/). Para a delimitação dessa área, optou-se, sob a orientação de alguns especialistas, pelo seguinte recorte: N&N lida essencialmente com *materiais*; para a produção desses materiais, são utilizados *métodos* e *processos* de fabricação; as *aplicações* dos materiais dependem da *caracterização* das *propriedades* dos mesmos, sendo que a caracterização é feita por *técnicas experimentais* e *métodos teóricos*. Com base nesse recorte, obteve-se uma organização global inicial composta pelos campos nocionais (1) síntese, processo e fabricação, (2) materiais, (3) propriedades e técnicas de caracterização, (4) máquinas e dispositivos; (5) teorias e métodos computacionais e (6) aplicações.

**b) Seleção e compilação do *corpus*-fonte.** Para a seleção do *corpus*-fonte, decidiu-se, por razões práticas, que a língua seria o inglês e que o *corpus* conteria apenas (i) textos (livros, artigos e resumos de artigos) e (ii) listas de palavras-chave e de tópicos disponíveis em formato eletrônico na web. Os artigos completos foram compilados dos periódicos *Journal of Nanoscience and Nanotechnology*[2], *Nanotechnology*[3], *Nature*[4] e *Science*[5] e os resumos e suas informações correspondentes (listas de palavras-chave e de tópicos) foram compilados

[1] O Grupo de Estudos e Pesquisas em Terminologia (GETerm) está sediado na Universidade Federal de São Carlos (UFSCar) e é coordenado pela Profa. Dra. Gladis Maria de Barcelos Almeida.

[2] http://www.aspbs.com/html/a0600frm.htm

[3] http://www.iop.org/EJ/S/3/451/YfSAujvDHdF2s7uXEE6cyg/journal/Nano

[4] http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v433/n7025/index.html

[5] http://www.sciencemag.org/

do portal *ISI Web of Knowledge*[6] (Web of Science). Para a compilação, foram adotados os métodos manual e automático. O método manual foi aplicado na coleta dos artigos em formato eletrônico dos referidos periódicos e portal. A coleta automática foi feita por meio do BootCaT[7] [8] ("Bootstrapping Corpora and Terms"), um extrator automático de *corpus* e de termos a partir de material disponível na web. Após a coleta, todos os textos do *corpus* foram uniformizados (manual ou automaticamente) para o formato texto e, em seguida, foram limpos e anotados, de modo a prepará-los para o processamento computacional. Ao final, obteve-se o *corpus* da área de N&N (Corpus N&N) com extensão de 2.570.792 palavras. O Corpus N&N é composto por 22 livros, 200 artigos de revistas e 9.982 resumos de artigos, com seus respectivos títulos, palavras-chave e lista de tópicos (no caso de livros).

**c) Extração automática de termos candidatos.** Essa tarefa pode ser feita por meio de métodos lingüísticos, estatísticos e híbridos [7]. No estudo de caso realizado, optou-se pelos estatísticos, mais especificamente, pelo Pacote NSP ("N-gram Statistics Package")[8] e pela ferramenta BootCaT. No que diz respeito às medidas estatísticas disponíveis no Pacote NSP, foram utilizadas: (a) *freqüência* (para a extração de unigramas); (b) *formação mútua*, *log-likelihood* e *coeficiente Dice* (para a extração de bigramas); (c) *informação mútua* e *log-likelihood* (para a extração de trigramas). Da lista de unigramas, a partir da qual, aliás, são geradas as de bigramas e trigramas, foram extraídas as unidades que não tinham valor terminológico (p.ex.: preposições, artigos, etc). Além disso, diante do tamanho do Corpus N&N, todos os n-gramas com freqüência inferior a 22 foram excluídos. Ao final, foram obtidos: 7645 unigramas, 1954 bigramas e 3216 trigramas. Para a geração de unigramas com o BootCaT, o Corpus N&N foi submetido aos processos de tokenização e *case folding*. Com o *corpus* tokenizado, a freqüência e a medida *log odd ratio* de cada palavra distinta foram calculadas. Para a aplicação da medida *log odd ratio*, extraiu-se uma lista de freqüência do *Brown Corpus*[9], definido aqui como *corpus* de referência em língua inglesa. Diante da medida *log odd ratio* de cada palavra do Corpus N&N, os 10% melhores unigramas classificados foram selecionados, totalizando 7343 candidatos a termo. Após a exclusão das *stopwords*, restaram 7297 unigramas, 6780 bigramas e 152 trigramas.

**d) Inserção manual e validação dos termos na ontologia.** Nesta fase, as listas de termos candidatos extraídas do Corpus N&N pelo Pacote NSP e pela ferramenta BootCaT foram enviadas a especialistas[10] da área-objeto para que os componentes dessas listas fossem identificados (ou não) como termos válidos da área de N&N e conceitualmente organizados, tomando-se como ponto de partida os campos nocionais descritos na Subseção 2 (a). Após a identificação e a alocação dos termos reais na estrutura conceitual, obteve-se uma ontologia validada da área de N&N com 1600 termos (em formato "txt"), denominada **OntoNano**.

**e) Edição semi-automática da ontologia.** A edição semi-automática de uma ontologia engloba os processos de inserção, exclusão, busca e visualização de termos auxiliados por

---

6 http://isi02.isiknowledge.com/portal.cgi/

7 O BooCaT envia um conjunto de sementes (termos relevantes da área-objeto fornecidas pelo usuário) ao Google e os textos das URLs que retornaram compõem um primeiro *corpus*. Desses textos, novas sementes são extraídas automaticamente e uma nova busca é feita. No final desse processo iterativo, obtém-se um *corpus* do qual serão extraídos os candidatos a termo.

8 http://www.d.umn.edu/~tpederse/nsp.html

9 clwww.essex.ac.uk/w3c/corpus_ling/ content/corpora/list/private/brown/brown.html

10 A tarefa de alocação dos candidatos a termos na ontologia coube a um grupo de especialistas da área de Física coordenado pelo Prof. Dr. Osvaldo Novais de Oliveira Jr. do IF/USP/São Carlos.

uma ferramenta computacional. No caso deste trabalho, os processos de inserção e exclusão dos termos na estrutura conceitual, no entanto, foram feitos manualmente (Subseção 2 (d)). Para a visualização da OntoNano, optou-se por utilizar o OntoEditor [4], que tinha a funcionalidade de converter a estrutura hierárquica da OntoNano em formato texto para estruturas arbórea e/ou hiperbólica.

## Considerações finais

A proposta metodológica ora apresentada caracteriza-se por unir pressupostos teóricos e práticos provenientes da Terminologia descritiva e do PLN. Caracteriza-se também por semi-automatizar alguns dos processos previstos na construção de ontologias de domínio, a saber: (i) seleção e compilação do corpus-fonte e (ii) extração de termos candidatos. É claro que essa semi-automatização não está livre de problemas. A ferramenta BootCaT, por exemplo, não retorna URLs cujos textos estão em formato doc ou pdf. Acreditamos que, no estudo de caso em N&N, esse problema foi satisfatoriamente contornado, pois a maior parte do material em formato doc e pdf disponível na web diz respeito aos artigos manualmente coletados dos periódicos descritos na Subseção 2(b). Além disso, salienta-se que a uniformização do corpus foi um processo complexo, posto que muitos textos em formato pdf, por exemplo, não puderam ser convertidos automaticamente, o que dificultou a montagem do corpus. Uma importante característica da metodologia é sua independência de domínio (utiliza métodos estatísticos e processos manuais bem definidos) o que permite a sua aplicação para diversas áreas do conhecimento. Por fim, salienta-se que a aplicação da metodologia na área de N&N gerou um corpus em inglês com aproximadamente 2500 milhões de palavras e uma ontologia com 1600 termos. A mesma metodologia está sendo atualmente aplicada na construção de um dicionário-piloto sobre N&N em português.

## Referências


1. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition, 5(2) (1993)
2. Guarino, N.: Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In: Pazienza, M. (eds.): Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, SCIE-97, Frascati, Italy (1997) 139-170
3. Jacquemin, C., Bourigault, D.: Term Extraction and Automatic Indexing. In: Mitkov, R. (ed.): The Oxford Handbook of Computational Linguistics. Oxford University Express, Oxford New York (2003) 599-615
4. Almeida, G. M. B; Oliveira, L. H. M.; Alusio, S. M.: A terminologia na era da informática. Ciencia e Cultura [online]. abr./jun. Vol.58, No.2, (2006) 42-45. Disponível em: http://cienciaecultura.bvs.br/scielo.php?pid=0009-6725&script=sci_serial
5. Cabré, M. T.: Theories of Terminology: their description, prescription and explanation. Terminology, Vol. 9, N. 2 (2003) 163-200
6. Almeida, G. M. B.: Teoria Comunicativa da Terminologia: uma aplicação. Araraquara, vol. I, 290 p.; vol. II, 86 p. Tese (Doutorado em Lingüística e Língua Portuguesa) – Faculdade de Ciências e Letras, Campus de Araraquara, Universidade Estadual Paulista (2000)
7. Teline, M. F.: Avaliação de Métodos de Extração Automática de Terminologia para textos em Português. ICMC-USP (Dissertação de Mestrado), São Carlos, SP, Fevereiro (2004)
8. Baroni, M.; Bernardini, S.: "BootCaT: Bootstrapping corpora and terms from the web". In: Proceedings of LREC 2004 (2004) 1313-1316