DI FELIPPO, A.; DIAS-DA-SILVA, B. C. Extração de informações lógico-conceituais de dicionários para a elaboração de léxicos computacionais. *Revista Intercâmbio*, volume XV. São Paulo: LAEL/PUC-SP, ISSN 1806-275X, 2006.

EXTRAÇÃO DE INFORMAÇÕES LÓGICO-CONCEITUAIS DE DICIONÁRIOS PARA A ELABORAÇÃO DE LÉXICOS COMPUTACIONAIS

Ariani DI FELIPPO (Universidade Estadual Paulista/ Campus Araraquara)
Bento Carlos DIAS-DA-SILVA (Universidade Estadual Paulista/ Campus Araraquara)

ABSTRACT: *The paper focuses on the description of the process of "mining" lexical semantic information from published dictionaries of Brazilian Portuguese language. Specifically, it is described the manual approach of compiling and filtering hyperonymy/hyponymy and holonymy/meronym logic-conceptual relations of concrete nouns. These relations, filtering from the dictionaries, will be use to organize part of the nouns of Wordnet.Br lexical database.*

KEYWORDS: *lexical semantics; concrete nouns; dictionaries; computational linguistics.*

0. Introdução

O conhecimento lexical, ou seja, o conhecimento individual sobre os itens lexicais, é essencial para todos os sistemas computacionais de processam línguas naturais, como *sistemas de tradução automática*, *sistemas de sumarização automática*, *sistemas* de *correção gramatical e ortográfica*, entre outros. Essa necessidade decorre do fato de que grande parte das informações necessárias para o processamento automático de uma língua é fornecida pelo "léxico" ("megarquivo") desses sistemas.

Com o desenvolvimento de sistemas que processam textos reais, surge a necessidade de léxicos que sejam (i) *manipuláveis* pelo sistema do qual fazem parte, isto é, léxicos cujas informações sejam explicitamente especificadas por meio de um esquema de representação formal (ou formalismo) e (ii) *lingüisticamente motivados*, tanto do ponto de vista da robustez (isto é, léxicos que contenham uma quantidade de unidades compatível com o léxico de uma língua natural), quanto da qualidade das informações associadas às entradas lexicais (Handke, 1995; Grishman, Calzolari, 1998). Em outras palavras, pode-se dizer que surge a necessidade de léxicos sofisticados, que armazenam, em especial,

informações semânticas sobre os itens da língua (Saint-Dizier, Viegas, 1995; Palmer, 1999).

Para a construção de léxicos computacionais que armazenam informações semânticas, os pesquisados da área do Processamento Automático das Línguas Naturais (PLN) têm lançado mão de vários métodos para a extração dessas informações, os quais podem ser assim agrupados: (i) extração manual, semi-automática ou automática de ocorrências a partir de córpus (Grefenstette, 1994; Klavans, Tezoukermann, 1996); (ii) extração semi-automática ou automática a partir de "dicionários legíveis por máquina" (do inglês, "machine readable dictionaries")[1] (Copestake, 1990; Roventini Et Al., 1998; Chugur et al., 2001); (iii) extração manual a partir de dicionários impressos (Sparck Jones, 1960, *apud* Grefenstette, 1994).

Para a construção de redes *wordnets* (um tipo especial de léxico computacional), os pesquisadores do PLN têm utilizado os métodos descritos em (i), (ii) e (iii), tanto isoladamente quanto em combinação. Especificamente na construção da rede wordnet para o português do Brasil (a Wordnet.Br), adotou-se, para a montagem dos conjuntos de sinônimos, o processo manual de extração de informações semânticas a partir de dicionários monolíngües impressos (e de versões em CD-ROM). Alguns fatos justificaram essa escolha. Como ilustração, citam-se dois deles: a inexistência de "dicionários legíveis por máquina" disponíveis para o português Brasil e o fato de que a compilação de informações a partir de córpus demanda tempo e uma grande equipe de pesquisadores especializados (Dias-Da-Silva Et Al., 2003; Dias-Da-Silva; Moraes, 2003).

Para a compilação das relações lógico-conceituais de hiperonímia/hiponímia e holonímia/meronímia que caracterizam a categoria dos nomes e que serão inseridas na base da WordNet.Br em uma etapa futura, objetiva-se utilizar especificamente dois métodos de extração de informações: (a) semi-automático a partir de *corpus* e (b) manual a partir (das versões digitais) de dicionários.

Neste trabalho, discute-se, do ponto de vista lingüístico, a aplicação do método de extração manual a partir (das versões digitais) de dicionários. Para tanto, enfatiza-se, na Seção 1, a relevância deste trabalho para a expansão da Wordnet.Br. Na Seção 2, descreve-se o conjunto de dicionários escolhido para a realização deste trabalho. Na Seção 3, são brevemente exemplificadas as tarefas de extração e filtragem das informações dos dicionários. Por fim, na Seção 4, são apresentadas

algumas considerações finais sobre a tarefa de extração manual de informações lógico-conceituais a partir de dicionários.

1. Das redes *wordnets*

Os pesquisadores do Laboratório de Ciência Cognitiva da Universidade de Princeton, EUA, embora estivessem preocupados em emular o léxico mental, suprem parte da necessidade do PLN ao desenvolverem a WordNet para o inglês americano[2] (Miller, Fellbaum, 1991). Do ponto de vista formal, uma wordnet estrutura-se em termos de conjuntos de sinônimos, os "synsets", que visam a representar o conceito lexicalizado pelas unidades sinônimas que o compõem. Do ponto de vista da topologia da rede, cada synset constitui um nó e as ligações entre os diferentes nós, feita por meio de arcos rotulados, visam a exprimir a relação léxico-semântica de antonímia e as relações lógico-conceituais de hiponímia, troponímia, meronímia, causa e acarretamento. Por exemplo, WordNet (versão 1.5), o conceito lexicalizado pelo synset unitário {bottle} ("garrafa") está associado ao conceito expresso pelo synset unitário {vessel} por meio da relação lógico-conceitual de hiperonímia, já que {bottle} é um tipo de {vessel} ("vasilha"). A WordNet de Princeton registra, ainda, outros tipos de informações: (i) para cada synset, há um índice numérico de identificação e uma glosa que especifica o conceito por ele lexicalizado; (ii) para cada unidade lexical, há uma ou mais frases-exemplo que serve para ilustrar o seu contexto de uso (Fellbaum, 1998).

Diante da relevância lingüística e das implicações tecnológicas desse tipo de base para o desenvolvimento dos mais variados tipos de sistemas de PLN, iniciou-se, em 2001, o empreendimento de construção da wordnet para o português do Brasil, a **Wordnet.Br**[3]. A base da Wordnet.Br contém aproximadamente 44 mil unidades lexicais (17 mil substantivos, 15 mil adjetivos, 11 mil verbos e mil advérbios), estruturadas em função das relações de sinonímia e antonímia. A especificação das demais relações e informações já referidas também está prevista no desenvolvimento da Wordnet.Br. Além disso, objetiva-se relacionar os synsets semanticamente equivalentes das bases das duas redes, a brasileira e a americana, com vistas à implementação de uma futura base bilíngüe (inglês-português) (Dias-Da-Silva et al. 2002).

É pensando, então, na inserção das relações lógico-conceituais de hiperonímia/ hiponímia e holonímia/ meronímia que se

estabelecem entre os conceitos lexicalizados por nomes, discute-se, neste trabalho, o processo manual de extração dessas relações a partir de dicionários monolíngües. Para tanto, focaliza-se um subconjunto dos nominais, mais especificamente, aquele formado por nomes (concretos)[4] que referenciam "recipientes" (em inglês, "containers") (Lherer, 1974).

2. Da delimitação do *corpus* de referência

Ao conjunto de dicionários a partir do qual as informações semânticas e lógico-conceituais podem ser extraídas, é dado o nome *corpus de referência* (CR). Para compor o CR deste trabalho, foram escolhidas as seguintes obras lexicográficas monolíngües do português do Brasil:

- Dic. Eletrônico Houaiss da Língua Portuguesa (doravante H) (Houaiss, 2001);
- Dic. Aurélio Eletrônico (doravante A) (Ferreira, 1998);
- Dic. Michaelis Português (doravante M) (Weizflog, 1998).

Essa escolha justifica-se basicamente por duas razões: a primeira, de ordem teórica, diz respeito ao fato de que essas obras são inegavelmente fontes de conhecimento lexical e a segunda, de ordem prática, diz respeito ao fato que essas obras estão em meio digital, o que agiliza a extração das informações.

3. As tarefas de extração e filtragem das informações nos dicionários do CR

Nesta Seção, propõe-se que a tarefa de extração de informações léxico-conceituais dos nomes concretos a partir de dicionários seja guiada por dois critérios: (a) observações empíricas e (b) hipóteses teóricas.

As observações empíricas (a) consistem na identificação de regularidades sistemáticas e similaridades de itens lexicais e padrões definicionais entre os dicionários do CR. Para tanto, parte-se da tipologia das definições de nomes concretos (Biderman, 1993). Segundo Biderman (1993), as definições para os nomes concretos podem ser: sinonímica, hiperonímica, metonímica, antonímica, enumerativa e por aproximação. Devido à organização hierárquica dos conceitos nominais (Collins; Quillian, 1969), a definição comumente dada aos nomes é a hiperonímica (ou seja, "gênero próximo mais diferença específica"). Biderman (1993), aliás, ressalta que a definição hiperonímica constitui o modelo ideal, pois é o caso em que se utiliza um hiperônimo como classificador básico em

cuja classe se inclui o nome. A consideração dessa tipologia definicional auxilia a identificação de similaridades entre os dicionários do CR.

Quanto às hipóteses teóricas (b), parte-se do construto da *estrutura qualia* de Pustejovsky (1996) como guia para a identificação das informações lógico-conceituais a serem extraídas dos dicionários. A *estrutura qualia*, proposta por Pustejovsky no âmbito do modelo do Léxico Gerativo (Pustejovsky, 1996), é responsável pela especificação dos "modos de significação" dos nomes, ou seja, ela apresenta os atributos e valores de um objeto em função dos quale: FORMAL (*o que é* ***x***), CONSTITUTIVO (*de que* ***x*** *é feito*), TÉLICO (*função de* ***x***) e AGENTIVO (*como* ***x*** *surge*).

A seguir, são exemplificadas as tarefas de extração e filtragem das informações lógico-conceituais dos nomes concretos a partir dos dicionários que compõem o CR segundo os critérios (a) e (b).

Para tanto, considera-se o item lexical garrafa e suas respectivas definições nos dicionários do CR.

- recipiente de gargalo e boca estreitos, geralmente de vidro, cristal ou louça e sem alça(s), destinado a conter líquido (H)
- vaso, comumente de vidro, com gargalo estreito, e destinado a conter líquidos (A)
- vaso geralmente de vidro, de gargalo estreito, destinado a líquidos (M)

Partindo-se das similaridades entre os dicionários que compõem o CR, observa-se que as definições elaboradas para garrafa são do tipo hiperonímica, sendo que uma glosa identificadora para o conceito codificado por garrafa pode ser assim resumida: "recipiente de gargalo e boca estreitos, geralmente de vidro, usado para conter líquido".

Com base na *estrutura qualia*, são identificadas, extraídas e representadas (e/ou organizadas) as seguintes informações sobre o item garrafa:

FORMAL	= *é_um* (<recipiente>, <vaso>)
CONSTITUTIVO	= *tem_como_membro* (<gargalo>, <boca>)
	= *feito_de* (<vidro>)
TÉLICO	= *usado_para* (<conter_líquido>)

Dos "gêneros próximos" expressos nas definições dos dicionários, foi possível identificar os itens lexicais que são hiperônimos de garrafa, representados no *quale* FORMAL, no caso, recipiente e vaso. Dos trechos dos textos definitórios em que constam as "diferenças

específicas", foi possível identificar as informações representadas pelos *quale* CONSTITUTIVO e TÉLICO, sendo que, no *quale* CONSTITUTIVO, estão representadas as relações lógico-conceituais de holonímia/ meronímia; no caso, gargalo e boca são partes de garrafa. Com base em uma versão mais alargada da *estrutura qualia*, proposta por Zavaglia (2003), foi possível ainda associar ao *quale* CONSTITUTIVO a informação *feito_de* (<vidro>). É claro que por meio da *estrutura qualia* é possível sistematizar não só as relações de hiperonímia/ hiponímia e holonímia/ meronímia mas também outros tipos de informações, como a que está associada ao *quale* TÉLICO, ou seja, *usado_para* (<conter_líquido>).

Feito isso, ou seja, identificadas e sistematizadas as informações referentes ao item garrafa, passa-se à inserção das mesmas na base da rede Wordnet.Br.

Como resultado da análise do item garrafa nos dicionários que compõem o CR, conclui-se que o synset unitário {garrafa}, na base da rede Wordnet.Br, deva ser associado a:

(i) o synset que contém recipiente e vaso, ou seja, {recipiente; receptáculo; vasilha; vaso;}, por meio da relação de hiperonímia;

(ii) os synsets que contêm gargalo e boca, ou seja, {gargalo, pescoço} e {abertura; ádito; boca3; entrada;}, respectivamente, por meio da relação de meronímia.

Para a inserção das informações *feito_de* e *usado_para*, outros tipos de relações lógico-conceituais deverão ser elaboradas.

4. Considerações finais

Neste trabalho, discutiu-se a tarefa de extração (e filtragem) de informações lógico-conceituais dos nomes concretos ("recipientes") a partir de um *corpus* de referência composto por obras lexicográficas monolíngües do português do Brasil. Para essa tarefa, foram propostos dois critérios fundamentas, são eles: observações empíricas e hipóteses teóricas. Por meio dessa metodologia, é possível extrair e organizar as informações desse subconjunto dos nomes concretos de tal modo que as relações lógico-conceituais (hiperonímia/ hiponímia, holonímia/ meronímia, entre outras) possam ser definidas e inseridas na base da rede Wordnet.Br.

Vale ressaltar, no entanto, que a análise das ocorrências dos itens lexicais em *corpus* é etapa complementar à extração das informações dos dicionários, pois a análise de córpus pode auxiliar na delimitação dos conceitos, atestando (ou não) os sentidos registrados nos dicionários.

DI FELIPPO, A.; DIAS-DA-SILVA, B. C. Extração de informações lógico-conceituais de dicionários para a elaboração de léxicos computacionais. *Revista Intercâmbio*, volume XV. São Paulo: LAEL/PUC-SP, ISSN 1806-275X, 2006.

NOTAS:

[1] Dicionários cujos arquivos-fonte são codificados e armazenados de tal forma que informações semânticas possam ser automaticamente extraídas por meio de regras (BOGURAEV, BRISCOE, 1989).

[2] O nome da rede americana é grafado com "N" maiúsculo para diferenciá-la das demais, caracterizando-a, como diz Fellbaum (1998), como "a mãe de todas as Wordnets", construída para essa variante do inglês. A base da WordNet de Princeton e informações diversas relativas a esse projeto estão disponíveis no endereço http://wordnet.princeton.edu/.

[3] CNPq 09/2001- Processo Nº 552057/01-0. Observe-se também que, na fase atual de desenvolvimento da base da Wordnet.Br, estão sendo realizadas as seguintes tarefas: (i) análise da consistência semântica dos synsets; (ii) coleta e seleção das frases-exemplo; (iii) especificação de glosas para uma parcela dos verbos.

[4] Segundo Lyons (1977), os nomes concretos "codificam" conceitos cujos *referentes* (no mundo real ou imaginário) são objetos físicos, localizáveis no tempo e no espaço, e com propriedades perceptíveis diretamente observáveis.

REFERÊNCIAS BIBLIOGRÁFICAS

BIDERMAN, M.T.C. A definição lexicográfica. *Terminologia –Projeto Termisul.* Cadernos do Instituto de Letras. Porto Alegre: Universidade Federal de Mato Grosso do Sul, 1993.

BOGURAEV, B., BRISCOE, T. (Eds.). *Computational lexicography for natural language processing*. London: Longman, 1989.

CHUGUR, I., PENAS, A., GONZALO, J., VERDEJO, F. Monolingual and bilingual dictionary approaches to the enrichment of the Spanish WordNet with adjectives. In: NAACL 2001 - WORKSHOP ON WORDNET AND OTHER LEXICAL RESOURCES, 2001, Pittsburgh. *Proceedings.*… Pittsburgh, 2001.

COLLINS, A., QUILLIAN, M. R. Retrieval time from semantic memory. *Journal of Verbal Behavior and Verbal Learning*. vol. 8, p. 240-7, 1969.

COPESTAKE, A. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. INTERNATIONAL WORKSHOP ON INHERITANCE IN

NATURAL LANGUAGE PROCESSING, 1990, Tilburg. *Proceedings…* Tilburg, 1990, p. 19-29.

DIAS-DA-SILVA, B. C.; OLIVEIRA, M. F.; MORAES, H. R. Groundwork for the development of the Brazilian Portuguese Wordnet. In: Ranchold, E.M.; Mamede, N.J. (Eds.). *Advances in natural language processing*. Berlin: Springer-Verlag, 2002, p.189-196.

DIAS-DA-SILVA, B. C. MORAES, H. R. A construção de thesaurus eletrônico para o português do Brasil. *Alfa (São Paulo)*. v. 47(2), p. 101-115, 2003.

DIAS-DA-SILVA, B., OLIVEIRA, M.F., MORAES, H.R. Reusability of dictionary in the compilation of NLP lexicons. In: Mamede, N.J.; Baptista, J.; Trancoso, I. (Eds.) *Computational processing of the Portuguese language*. Berlim: Springer-Verlag, 2003, p. 78-85.

FELLBAUM, C. A. (Ed.). *Wordnet: an electronic lexical database*. Cambridge: The MIT Press, 1998.

FERREIRA, A.B.H. *Dicionário Aurélio eletrônico: novo dicionário Aurélio – século XXI* (versão 3.0). São Paulo: Lexicon, 1999. CD-ROM.

GREFENSTETTE, G. *Explorations in automatic thesaurus discovery*. Boston/London/Dordrecht: Kluwer Academic Publishers, 1994.

GRISHMAN, R., CALZOLARI, N. Lexicons. In: COLE, R.A. (Ed.). *Survay of the state of the art in Human Language Technology*. Cambridge, Mass.: Cambridge University Press, 1998, p. 392-5.

HANDKE, J. *The structure of the lexicon:* human versus machine. Berlin: Mouton de Gruyter, 1995.

HOUAISS, A. *Dicionário eletrônico Houaiss da Língua Portuguesa* - versão 1.0. Editora Objetiva, 2001.

KLAVANS, J.; TZOUKERMANN, E. Dictionaries and corpora: combining corpus and machine-readable dictionary data for building bilingual lexicons. *Machine Translation*, vol. 10(3-4), p. 1-34, 1996.

LEHRER, A. Semantic fields and lexical structure. Amsterdam: North-Holland, 1974.

LYONS, J. *Semantics.* vol. 2. Cambridge: Cambridge University Press, 1977.

MILLER, G. A., FELBAUM, C. Semantic networks of English. *Cognition.* vol. 41, p. 197 – 229, 1991.

PALMER, M. Multilingual resources – Chapter 1. In: HOVY, E. et al. (Eds.) *Multilingual information management: current levels and future*

abilities, 1999. Disponível em <http://www.cs.cmu.edu/~ref/mlim/> Acesso em 1 abril de 2005.

PUSTEJOVSKY, J. *The generative lexicon.* 2ª ed. Cambridge: Mass.: The MIT Press, 1996.

SAINT-DIZIER, P., VIEGAS, E. *Computational lexical semantics*. Cambridge: Cambridge University Press, 1995.

ROVENTINI A., PETERS C., CALZOLARI N., BERTAGNA F. Building a semantic network for Italian using existing lexical resources. In: Rubio A. et al. (Eds.). INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES & EVALUATION, 1, 1998, Granada, Spain. *Proceedings.*… Granada, Spain, 1998, p. 377-383.

WEISZFLOG, W. (ed) *Michaelis Português*: *moderno dicionário da língua portuguesa* (versão 1.1). São Paulo: DTS Software Brasil Ltda, 1998.

ZAVAGLIA, C. A homonímia no português: tratamento semântico segundo a estrutura qualia de Pustejovsky com vistas a implementações computacionais. *Alfa (São Paulo)*. v. 47(2), p. 77-99, 2003.