Wordnet.Br: An Exercise of Human Language Technology Research

Bento Carlos Dias-da-Silva
CELiC - Faculdade de Ciências e Letras, Universidade Estadual Paulista
Rodovia Araraquara-Jau Km 1
14800-901 Araraquara, São Paulo, Brazil,
bento@fclar.unesp.br

Abstract

This paper reports the ongoing project (since 2002) of developing a wordnet for Brazilian Portuguese (Wordnet.Br) from scratch. In particular, it describes the process of constructing the Wordnet.Br core database, which has 44,000 words organized in 18,500 synsets Accordingly, it briefly sketches the project overall methodology, its lexical resourses, the synset compilation process, and the Wordnet.Br editor, a GUI (graphical user interface) which aids the linguist in the compilation and maintenance of the Wordnet.Br. It concludes with the planned further work.

Introduction

Assuming a compromise between Human Language Technology and Linguistics, and based on the Artificial Intelligence notion of Knowledge Representation Systems (Hayes-Roth, 1990, Durkin, 1994), this project applies a three-domain approach methodology to the development of the Brazilian Portuguese (BP) WordNet (Wordnet.Br).[1] This approach claims that the linguistic-related information to be computationally modelled, like a rare metal, must be "mined", "molded", and "assembled" into a computer-tractable system (Dias-da-Silva, 1998). Accordingly, the processes of designing and implementing the Wordnet.Br lexical database are being developed in the following complementary domains:

- *The Linguistic-related Domain*, where the lexical resources (dictionaries and text corpuses), the lexical-conceptual relations (synonymy, antonymy, hyponymy, meronymy, entailment, cause), and a sort of natural language ontology of concepts ("Base Concepts" and "Top Ontology")[2] are mined;
- *The Representational Domain*, where the overall information selected and organized in the preceeding domain is molded into a computer-tractable representation (the "synsets", the "lexical matrix", and the wordnet "lexical database" itself)[3];
- *The Computational Domain*, where the computer-tractable representations are assembled by means of utilities (the Wordnet.Br editor).

This paper, in particular, reports the first part of the project where in a two-year span the effort of three linguists and a computer scientist, each working in his respective domain, managed to compile the Wordnet.Br core database: 44,000 BP words organized in 18,500 synsets. In other words, the core database is a thesaurus-like lexical database.

1 The Linguistic-related Domain

1.1 Synonymy in Context

The Wordnet.Br core database architecture conforms to the two key representations of the Princeton WordNet (Fellbaum, 1998): the synset and the lexical matrix.

Its synsets are built on the basis of the notion of "synonymy in context", i.e. word interchangeability in context (Miller, 1998). Antonymy is checked either against morphological properties of words and their dictionary lexicographical information. The notion of lexical matrix (Miller and Fellbaum, 1991) is intended to capture the "many to many" associations between form and meaning.

1.2 The Reference Corpus

Given the team of three linguists, the unavailability of reusing machine-readable dictionaries and other existing wordnets,[4] and a two-year deadline to present large-acale results, the Wordnet.Br developers manually reused, merged, and tuned synonymy and antonymy information registered in five outstanding published dictionaries of BP:[5] Ferreira (1999), Weiszflog (1998), Barbosa (1999), Nascentes (1981), and Borba (1990). BP texts available in the *NILC Corpus* (CETENFolha, 2004) and in the web complemented the project reference corpus.

To understand how the linguits "mined" for synsets into the reference corpus, let us follow an example. Weiszflog (1998) distinguishes seven senses of the verb *lembrar* (English: "to remember"). After collecting the synonyms, and

[1] This project was supported in part by contract 552057/01, with funding provided by The National Council for Scientific and Technological Development (CNPq); in part by grant 2003/03623-7 from The State of São Paulo Research Foundation (FAPESP).

[2] Rodríguez et al. (1998).

[3] Fellbaum (1998).

[4] Copyright reasons prevented us from reusing or adopting existing wordnet databases and utilities.

[5] The dictionaries were chosen for their pervasive use of synonyny and antonymy to define word senses. In a way, this choice dictated the way to proceed the work alphabetically, instead of working by semantic fields.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 301–303. © Masaryk University, 2005

disregarding definitions, the following synsets can be initially compiled:

1. {*lembrar*, *recordar*}
   (English: {“to remember”, “to recall”})
2. {*lembrar*, *advertir*, *notar*}
   (English: {“to remember”, “to warn”, “to notify”})
3. {*lembrar*, *sugerir*}
   (English: {“to suggest”, “to evoke”, “to hint”})
4. {*lembrar*, *recomendar*}
   (English: {“to remember”, “to commend”})

   Their consistency can be checked further by looking up the dictionary synonym entries for each of the remaining verbs.

   Weiszflog (1998) distinguishes five senses of *recordar*. Two of them are related to *lembrar*. One of them is given by the paraphrase *trazer à memória* (English; “to call back to memory”), the other is given by its synonym *lembrar*. This information validates (1).

   The analysis of the verb *esquecer* (English: “to forget”), the canonical antonym for *lembrar*, validated (5).
5. {esquecer, olvidar}

   The dictionary mining process allowed the linguist to compile (6); further corpus checking reduced it to (7).
6. {comemorar, ementar, lembrar, memorar, reconstituir, recordar, relembrar, rememorar, rever1, revisitar, reviver, revivescer, ver}
7. {lembrar, recordar, relembrar, rememorar, reviver, revivescer, }

The very same process can be repeated to compile all other synsets, and the analytical cycle starts over by collecting the synonyms from the next dictionary entry in the alphabetical order.

2 The Representational Domain

2.1 The Wordnet.Br Core Database Design

Each Wordnet.Br entry consists of the template in Fig. 1, where n is the entry identification number; X is a noun, verb, adjective, or adverb; and $n.1 \ldots n.m$ are sense identification numbers of the entry n.

[<Headword> n (<X>)
Sense $n.1$ [{Synset}; {Antonym Synset}]
. . .
Sense $n.m$ [{Synset}; {Antonym Synset}]]

Figure 1: The entry template

From the logical point of view, the overall structure of the database is made up of two lists: an Entry List (EL) of word forms ordered alphabetically and the Synset List (SL) of synsets. Each element of a synset is necessarily an element of the EL. Each EL entry is specified for a particular Sense Specification (SS). Each SS is indexed by three pointers: the “synonymy pointer” to a particular synset in the SL; the “antonymy pointer” to a particular synset in the SL which is the antonym of a particular synset; and the “sense pointer” to the particular entry in the EL to which synsets are associated.

Table 1: The Wordnet.Br Core Statistics

CATEGORY	LEXICAL UNITS	SYNSETS
Verbs	11,000	4.000
Nouns	17,000	8.000
Adjectives	15,000	6,000
Adverbs	1,000	500
TOTAL	44,000	18,500

3 The Computational Domain

3.1 The Wordnet.Br Editing Utility

The editing tool is a Windows®-based GUI where the linguists enter synsets, sample-sentences, glosses, and generates different lists (synsets listed by syntactic category, number of elements, degree of homonymy and polysemy, and list of sample sentences) and statistics.

Its main functionalities include: the storage of general information of the Wordnet.Br core database and its bookkeeping.

Currently, Wordnet.Br core database presents the figures in Table 1.

Further Work

The existing Wordnet.Br database can now be further refined, augmented and updated. Future steps will involve the specification of the following information:

- a concept gloss for each synset;
- a sample sentence for each word form;
- the cross-linguistic alignment of the Wordnet.Br core database with an Inter-Lingual-Index (Vossen, 1998), including correspondence links with the European Portuguese WordNet. (Marrafa, 2001).
- the semiautomatic conceptual relations of meronymy and hyponymy with the help of the preceding alignments.

Acknowledgements

My thanks go to the GWA-2006 referees, who helped make this paper better.

References

Barbosa, O. (1999) *Grande Dicionário de Sinônimos e Antônimos*. Ediouro, Rio de Janeiro, 550 p.

Borba, F. S., coord. (1990) *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil*. Editora da Unesp, São Paulo, 600 p.

CETENFolha (2004) *Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo*. `http://www.linguateca.pt/`.

Dias-da-Silva, B. C. (1998) *Bridging the Gap Between Linguistic Theory and Natural Language Processing*. In "16th International Congress of Linguists", B. Caron, ed., Pergamon-Elsevier Science, Oxford, 10 p.

Durkin, J. (1994) *Expert Systems: Design and Development*. Prentice Hall International, London, 800 p.

Fellbaum, C., ed. (1998) *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, Mass., 423 p.

Ferreira, A. B. H. (1999) *Dicionário Aurélio Eletrônico Século XXI*. Lexicon, São Paulo, CD-ROM.

Hayes-Roth, F. (1990) *Expert Systems*. In "Encyclopedia of Artificial Intelligence", E. Shapiro, ed., Wiley, New York, pp. 287–298.

Marrafa, P. (2001) *WordNet do Português: u ma base de dados de conhecimento lingüístico*. Instituto Camões, Lisboa, 77 p.

Miller, G. (1998) *Nouns in WordNet*. In "WordNet: An Electronic Lexical Database", C. Fellbaum, ed., The MIT Press, Cambridge, Mass., pp. 47–46.

(1991) *Semantic Networks of English*. Cognition, 41, pp. 197-229.

Miller, G., Fellbaum, C. (1991) *Semantic Networks of English*. Cognition, 41, pp. 197–229.

Nascentes, A. (1981) *Dicionário de Sinônimos*. Nova Fronteira, Rio de Janeiro, 485 p.

Rodríguez, H. et al. (1998) *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*. Computers and the Humanities, 32/2,3, pp. 117–152.

Vossen, P. (1998) *Introduction to EuroWordNet*. Computers and the Humanities, 32/2,3, pp. 73–89.

Weiszflog, W., ed. (1998) *Michaelis Português – Moderno Dicionário da Língua Portuguesa*. DTS Software Brasil Ltda, São Paulo, CD-ROM.