Towards an Automatic Strategy for Acquiring the WordNert.Br Hierarchical Relations

Ariani Di Felippo[1,2], Bento Carlos Dias-da-Silva[1,2]

[1]Centro de Estudos Lingüísticos e Computacionais da Linguagem (CELiC)
Faculdade de Ciências e Letras – Universidade Estadual Paulista (UNESP/Ar.)
Caixa Postal 174 – 14800-270 – Araraquara – SP – Brazil

[2]Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-270 – São Carlos – SP – Brazil

arianidf@uol.coim.br, bento@fclar.unesp.br

Abstract. *This paper outlines an "automatic strategy with post-editing" for specifying the Brazilian Portuguese WordNet (WordNet.Br) hierarchical lexical-conceptual relations (taxonomies and meronymies). It relies on a two-step analysis: (a) the computer-aided linking (lexical-conceptual alignment) of the WordNet.Br synsets to the their semantically equivalent ones in the Princeton WordNet; (b) the automatic transfer with post-editing of the relevant hierarchical relations spotted by the previous step.*

1. Introduction

A revolutionary development of the 1990s was the Princeton WordNet (PWN) [Fellbaum 1998], an online reference lexical database built for English that combines the design of a dictionary and a thesaurus with a rich ontological potential. PWN contains information about nouns, verbs, adjectives and adverbs and is organized around the notion of *synsets* (i.e. sets of word-forms with the same part-of-speech that lexicalize the same concept), e.g. {car, auto, automobile, machine, motorcar}. The synsets are related to one another by lexical-conceptual relations, such as *antonymy*, *hyponymy/hypernymy*, *meronymy/holonymy*, *cause* and *entailment*. There is no doubt that the PWN has become a *de facto* standard for a wide range of NLP applications [Morato et al. 2004] and it has determined the emergence of several projects that aim at the construction of WordNets for other languages than English[1] or the development of multilingual WordNets (such as EuroWordNet [Vossen 1998]). Inspired by the impact that the availability of PWN had on NLP researches, it was definitely launched in 2003 the *Brazilian Portuguese WordNet* (WordNet.Br or WN.Br) [Dias-da-Silva et al 2002; Dias-da-Silva 2003]. Currently, the WN.Br database contains around 11,000 verbs, 17,000 nouns, 15,000 adjectives, and 1,000 adverbs. These word-forms are organized into 18,500 synsets, and, where relevant, part of the synsets is connected via the antonymy relation. Among other bits of information, the verb synset database of the WB.Br is being augmented by the hierarchical relations[2] of hyponymy/hypernymy and meronymy/holonymy between synsets [Dias-da-Silva et al 2006].

[1] For details of WordNets around the world, see http://www.globalwordnet.org/gwa/wordnet_table.htm.

[2] Two main sorts of hierarchical relations are distinguished: (i) taxonomic (hipernymy/hyponymy) and (ii) meronomic (holonymy/meronymy) [Cruse 2004].

With regard to the hierarchical relations, it is well-known that manual specification is a costly and highly time-consuming task. Consequently, several approaches to obtain such relations in a (semi)automatic way have been carried out taking advantage of available structured and unstructured lexical resources [Rigau 1998]. The most widely used structured resources are monolingual *machine-readable dictionaries* (MRDs) (i.e., machine-readable versions of a standard dictionaries). The main approach for the extraction of such relational information from MRDs is parsing the dictionary definitions. The acquisition of taxonomies is based on the observation that, at least for nouns and verbs, the syntactic head(s) of a definition (or *genus term*) is(are) usually the hypernym(s) of the word being defined (e.g., [Bruce and Guthrie 1991], [Rigau et al 1998], [Matsumoto 2003], etc.). The most widely used unstructured lexical resources for (semi)automatic extraction of hierarchical relations are very large *corpora* (100 million words or more) [Hearst 1998]. The main techniques for extraction these relations from *corpora* are based on (i) the frequency of co-occurrence of content word to create clusters of semantic similar word-forms [Church and Hanks 1990] and (ii) the searching for lexical-syntactic patterns [Hearst 1992]. As there are no available monolingual MRDs and large *corpora* for Brazilian Portuguese, we describe, in Section 2, an alternative method that has been developed for automatically acquiring the hierarchical relations, with post-editing[3], from the alignment of the PWN and the WN.Br databases. In Section 3, we present some comments on the automatic acquisition method as proposed here and list future works.

2. WordNet.Br Strategy

The previous considerations on the (semi)automatic acquisition methods proposed in the aforementioned literature led WordNet.Br developers to propose an alternative approach, which actually consists of automatically acquiring of the hierarchical relations from a computer-aided alignment of the WN.Br and the PWN synset databases. In Subsection 2.1, a general notion of the alignment process is introduced, which needs to be carried out before the detailed description of the alternative acquisition strategy can be properly appreciated.

2.1. The Computer-Aided Alignment

The overall computer-aided method for the lexical-conceptual alignment of verb synsets of both databases, the WN.Br and the PWN databases, implies the use of monolingual (Brazilian Portuguese-Brazilian Portuguese, BP-BP; English-English, En-En) and bilingual (Brazilian Portuguese-English, BP-En) dictionaries, corpora (Portuguese and English texts), language specific knowledge, and the supervision of the linguist. The task comprises a six-step process:

1. Manual selection of a WN.Br word-form *x*;
2. Automatic searching of every possible En word-forms that corresponds to *x;*
3. Manual selection of one En word-form *y*;
4. Automatic searching of all PWN synsets containing *y;*
5. Manual analysis of the PWN synsets containing *y* and the ones in the WN.Br containing *x* (target synsets), and identification of the appropriate equivalence relation that holds between them;
6. Drag-and-drop linking of the appropriate synsets of the PWN database to the target synsets in the WN.Br database.

[3] Post-editing is absolutely necessary for taxonomies and meronymies that are language-dependent.

The alignment process starts with a BP word-form being manually selected in the WN.Br database. In other words, the linking procedure input is a BP word-form already encoded in the WN.Br (Step 1). For instance, let us consider the BP word-form "arriscar" (En: "risk"). Equipped with a bilingual BP-En MRD[4], all possible En word-forms that are equivalent to the BP word-form are automatically generated by the WordNet.Br Editor: "risk", "endanger" and "jeopardize" (Step 2). After analyzing them, the linguist manually selects one of those English forms (Step 3). Let us say: "risk". As soon as the translation is selected, all the PWN synsets containing the En word-form are automatically searched and identified by the editor: {risk; put on the line; lay on the line} and {gamble; chance; risk; hazard; take chances; adventure; run a risk; take a chance} (Step 4). Then, the linguist analyzes the relevant types of equivalence links between the PWN synsets previously identified and the WN.Br synsets containing "arriscar": {arriscar; expor}, {arriscar; aventurar; malparar}, and {apostar; arriscar; jogar; pôr}. The analysis shows that the synsets {risk, put on the line, lay on the line} and {arriscar; expor} exhibit a specific equivalence relation between (Step 5). According to the set of inter-lingual equivalence relations identified by [Vossen 1998], these two synsets are linked by means of the EQ-SYNONYM label, and, accordingly, the synset {arriscar; expor} in the WN.Br inherits all information originally attached to the En synset to which it is linked (Step 6). This semi-automatic procedure is being performed to all verb word-forms already encoded in WN.Br, one by one. In the Subsection 2.2, we stress how the aligned synsets can be used to automatically obtain the hierarchical relations.

2.2. The Automatic Acquisition of the Hierarchical Relations

Let us consider that the PWN synset {try; seek; attempt; essay; assay} and the WN.Br synset {tentar; ensaiar; experimentar} are already linked. In the PWN, the concept represented by {try; seek; attempt; essay; assay} is a hypernym of the concept represented by {risk, put on the line, lay on the line}, which is, in turn, one of its hyponyms. If the synsets {tentar; ensaiar; experimentar} and {arriscar; expor} are, respectively, their corresponding aligned synsets in the WN.Br, it is possible to automatically identify the hipernymy/hyponymy relations and import them from the alignment (Figure 1). So, in the WN.Br, a directional "hypernym pointer" is created from the synset {arriscar; expor} to the synset {tentar; ensaiar, experimentar}, and an opposite pointer labeled "hyponym" is created from the synset {tentar; ensaiar; experimentar} to the synset {arriscar; expor}. The automatic acquisition of the hierarchical relations is possible due to the fact that if two synsets S_1 and S_2 are linked by a semantic relation R in PWN and if S'_1 and S'_2 are the corresponding aligned synsets in the WN.Br, then S'_1 and S'_2 can be linked by the relation R.

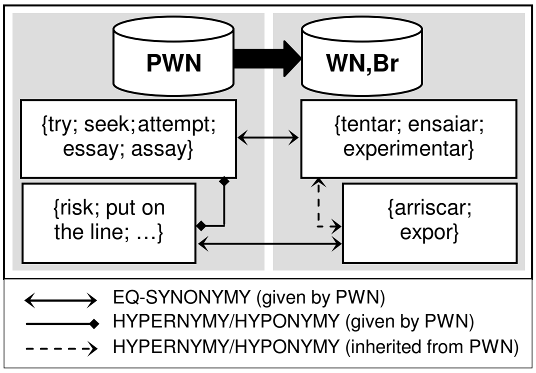
PWN
WN,Br
{try; seek;attempt; essay; assay}
{tentar; ensaiar; experimentar}
{risk; put on the line; ...}
{arriscar; expor}
EQ-SYNONYMY (given by PWN)
HYPERNYMY/HYPONYMY (given by PWN)
HYPERNYMY/HYPONYMY (inherited from PWN)

Figure 1. A sample of an automatic acquisition of hyponymy/hypernymy.

[4] This task is automatically performed with the help of the online free version of the Babylon Dictionary (http://www1.uol.com.br/babylon/).

3. Final Remarks

We have described an automatic approach that intends to make the WN.Br hierarchical relations specification easier and quicker. Considering that there are no available structured nor unstructured wide range lexical resources for BP, the method has been developed to be applied to the resulting links created by the ongoing computer-aided alignment of the PWN and the WB.Br databases. The alignment is supervised by linguists and it makes possible to establish high accuracy links between the WordNets, and, hence, to inherit the appropriate hierarchical relations, despite the need of post-editing, for, as note 3 warns, hierarchical lexical-conceptual relations are language dependent. At the current stage of the WN.Br development, an outstanding computational-linguistic initiative [Dias-da-Silva et al. 2006], the routines (i.e., the sequences of computer instructions) for automatic acquisition of such hierarchical relations have been implemented in an original software tool (the WordNet.Br Editor) that supports the WN.Br construction. Following such extension, the WN.Br Editor GUI (its "graphical user interface") will be also augmented by new wizards to make the post-editing of the automatically acquired hierarchical relations transparent and quicker.

References

Bruce, R., Guthrie L.(1991). "Building a Noun Taxonomy from a Machine Readable Dictionary", Research Report MCCS-91-207. Computing Research Laboratory, New Mexico State University, Las Cruces.

Church, K., Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, v. 16, n. 1, pages 22-29.

Cruse, A. (2004), Meaning in Language. Oxford, Oxford University Press.

Dias-da-Silva, B.C. (2003). "Human language technology research and the development of the brazilian portuguese wordnet", In: Proceedings of the 17th International Congress of Linguists, Edited by E. Hajičová, A. Kotěšovcová, and J. Mírovský, Prague, Matfyzpress, p. 1-12.

Dias-da-Silva, B.C., Oliveira, M.F., Moraes, H.R. (2002). "Groundwork for the development of the Brazilian Portuguese Wordnet", In: Proceedings of the 3rd PorTAL, Faro, Springer-Verlag, p. 189-196.

Dias-da-Silva, B.C., Di Felippo, A., Hasegawa, R. (2006). "Methods and Tools for Encoding the WordNet.Br Sentences, Concept Glosses, and Conceptual-Semantic Relations", In: Proceedings of the 7th PROPOR'06, Rio de Janeiro, Springer-Verlag, p. 120-130.

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

Hearst, M. (1992) "Automatic Acquisition of Hyponyms from Large Text Corpora". In: Proceedings of the 14th COLING'92. Nantes, France.

Hearst, M.A. (1998). Automated discovery of wordnet relations. In: WordNet: An Electronic Lexical Database, Edited by C. Fellbaum, Cambridge, MA, MIT Press, p.131-151.

Matsumoto, Y. (2003). Lexical Knowledge Acquisition, In: The Oxford Handbook of Computational Linguistics, Edited by R. Mitkov, Oxford, Oxford University Press, p. 395-413.

Morato, J., Marzal, M.A., Llorens, J., Moreiro, J. (2004) "WordNet applications", In: Proceedings of the 2nd Global WordNet Conference, Brno, p. 270-278.

Rigau, G. (1998). Automatic Acquisition of Lexical Knowledge from MRDs. Tesis doctoral, Departament de Llenguatges i Sistemes Informàtics, UPC, Barcelona.

Rigau, G., H. Rodriguez, Agirre, E. (1998). "Building Accurate Semantic Taxonomies from Monolingual MRDs", In: Proceedings of COLING-ACL '98, Montréal, Canada.

Vossen, P. (1998). Introduction to EuroWordNet. In *Computers and the Humanities*, v. 32, pages 73-89.