WordNet.PT new directions

Palmira MARRAFA, Raquel AMARO, Rui Pedro CHAVES, Susana LOUROSA,
Catarina MARTINS, Sara MENDES
Group for the Computation of Lexical and Grammatical Knowledge,
Center of Linguistics, University of Lisbon
Avenida Professor Gama Pinto, 2
1649-003 Lisboa, Portugal
{palmira.marrafa,ramaro,rui.chaves,slourosa,cmartins,sara.mendes}@clul.ul.pt

Abstract

This paper reports the current Portuguese WordNet (WordNet.PT) research and development directions, which mainly regard the enrichment of the WordNet model with event and argument structures (section 1), the codification of cross-part-of speech relations (section 2) and the exploitation of WordNet.PT in concrete applications (section 3).

Introduction

WordNet.PT is being built since July 1999 at the University of Lisbon (Centre of Linguistics) as a project of the *Group for the Computation of Lexical and Grammatical Knowledge (CLG).*

WordNet.PT is being developed within the general approach of EuroWordNet (Vossen (1998, 1999)). Therefore, like each wordnet in EuroWordNet, WordNet.PT has a conceptual architecture structured along the lines of the Princeton WordNet (Miller et al. (1990); Fellbaum (1998)).

Since the linguistic resources available for Portuguese are not suitable enough for the purpose of building a wordnet automatically, this project is being carried out mainly on the basis of manual work.

Aiming at using the Portuguese WordNet in language learning applications, among others, the starting point for the specification of a fragment of the Portuguese lexicon, at the first phase of the project (1999-2003), was the selection of a set of semantic domains covering concepts with high productivity in the daily life communication. Concerning lexical categories, the emphasis was on nouns. The encoding of language-internal relations followed a mixed top-down/bottom-up strategy for the extension of small local nets.

In mid 2004 a new phase of the project started, aiming at an integrated increment of the database, concerning several new semantic domains and all the main POS.

In order to better define the semantic domains and to capture the corresponding relevant concepts, at first we have treated semantic domains separately, as exhaustively as possible.

This methodology has led us to build semantic domain oriented wordnets. Such approach has made apparent the need for enriching the WordNet model with event and argument structures information, as described in section 1, and for encoding new relations, in particular cross-part-of-speech relations, referred to in section 2. The coherence of the results is evaluated on the basis of using WordNet.PT in Language Engineering (LE) applications. Section 3 describes a Question-Answer (QA) system prototype that uses WordNet.PT as a knowledge base (KB). We conclude this report with a few concluding remarks.

1 Enriching the WordNet model with event and argument structure information

Basic research, in particular on verbs of movement (Amaro (2005a, 2005b)), has rendered evident the need for enriching the WordNet lexical entries with qualia information. The incorporation of qualia information into the lexical entries together with the assumption of a qualia unification operation makes it possible to predict co-troponyms co-occurrence restrictions and to deal with non-exclusive co-hyponyms (see Mendes & Chaves (2001) for nouns).

Moreover, a decompositional approach of troponymy taking into account argument structure and Aktionsart seems to have strong motivation.

Research carried out on resultative predicates, on the other hand, has made it apparent that certain verbs, although denoting a final state, do not include in their denotation the set of content properties of the final state, as it is the case of *tornar* ("make"). This informational gap has to be filled with a resultative expression to avoid ungrammaticality. Thus, such verbs are conceptually deficitary (cf. Marrafa 2004 and 2005). Since concepts are the basic unit of WordNets, their descriptive adequacy cannot be preserved if concepts are not fully represented.

Besides the facts pointed out, a deeper analysis on these and other issues is being carried out, in order to provide an integrated approach on this matter.

2 Encoding cross-part-of-speech relations

Extending WordNet.PT to all the main POS has involved a revision of certain commonly used relations and the specification of several cross-part-of-speech relations.

With regard to adjectives, for instance, instead of being linked amongst themselves by a similarity relation, all adjectives modifying the same attribute are linked to the noun that lexicalizes this attribute. This way we obtain the cluster effect, argued by Fellbaum et al. (1993) and Miller (1998) to be the basis of the organization of adjectives, without having to encode it directly in the network (Mendes (2005)).

Concerning conceptual opposition, it seems that it does not have to be explicitly encoded, as well, since it is possible to make it emerge from the combination of *synonymy* and *antonymy* relations (Mendes op. cit).

Unlike other wordnets, WordNet.PT encodes all adjectives in the same file, avoiding having to decide beforehand whether an adjective is relational or descriptive, for instance. Rather, membership to these classes emerges from the relations expressed in the database.

A new relation encodes salient characteristics of nouns expressed by adjectival expressions: *is a characteristic of / has as a characteristic*. Despite the fact that we can object the status of this relation is not clear, concerning the lexical knowledge, it regards crucial information for many wordnet-based applications, namely those using inference systems.

Just to give a last example, a telic state relation is also created to encode the relation between resultative verbs that incorporate the telic (final) state and the adjectives that express it (e.g. *entristecer/triste* ("make_sad"/"sad")).

3 Using WordNet.PT in LE applications

Exploratory work aiming at the evaluation of WordNet.PT contents for LE applications purposes has been carried out. In this context, a wordnet-based QA system prototype (INQUER) has been built (Marrafa et al. (2004)). INQUER allows users to interact with the WordNet.PT database by means of a natural language interface. This system is not constrained to template queries. It rather provides a natural language interface that allows a user-friendly interaction with WordNet.PT through natural language questions. Direct natural language answers are given to the user's questions by applying inference and information extraction mechanisms. A syntactic-semantic analyzer is applied, not only to analyze the question but also to build a first-order logic semantic representation. Inference and information extraction mechanisms are then applied (on the fly) to extract the relevant information from the KB. In a last step, a natural language answer is generated, based both on the representation of the question and on the information extracted.

INQUER does not rely on the probabilistic extraction of answers from large collections of text. Instead, it uses WordNet.PT both as a lexical database required to analyze questions and generate natural language answers and as a semantic knowledge base required to obtain the requested information via an inference engine.

For now, the inference engine mainly involves ISA and part/whole relations and glosses as informal definitions. Nevertheless, its extension to deal with more complex questions should take into account other semantic relations and the internal analysis of glosses.

Final remarks

As pointed out, at the current stage WordNet.PT is being developed in an integrated approach that provides a very fine-grained characterization of lexical meaning. On the other hand, the results are being evaluated in concrete applications. This way, both meaning modelling motivation and usefulness for Computational Linguistics and Language Engineering purposes are guaranteed.

References

Amaro, R. (2005a), *Semantic Incorporation in a Portuguese WordNet of Verbs of Movement: on Aktionsart shifting*, in Proceedings of the Third International Workshopon Generative Approaches to the Lexicon, École de Traduction et d'Interpretation, University of Genéve, pp.1-9.

Amaro, R. (2005b), *Wordnet as a base lexicon model for the computation of verbal predicates*, paper accepted at the GWA06 Conference.

Fellbaum, C., D. Gross & K. Miller (1993) *Adjectives in WordNet* , in Miller et al. "Five Papers in WordNet", Technical Report, Cognitive Science Laboratory, Princeton University, pp. 26-39.

Fellbaum, C. (1998a), *A Semantic Netwok of English: The Mother of All WordNets*, in P. Vossen (ed.), "EuroWordNet: A Multilingual Database with Lexical Semantic Networks", Dordrecht: Kluwer Academic Publishers, pp. 137-148.

Fellbaum, C. (1998b), *A Semantic Netwok of English Verbs*, in C. Fellbaum (ed.), "WordNet: An Electronic Lexical Database", MA: The MIT Press, pp. 69-104.

Marrafa, P. (2004), *Extending WordNets to Implicit Information*, in Proceedings of LREC 2004, International Conference on Language Resources and Evaluation, Lisboa, Portugal, pp. 1135-1138.

Marrafa, P. (2005), *The Representation of Complex Telic Predicates in WordNets: the Case of Lexical-Conceptual Structure Deficitary Verbs*, Research on Computing Science, Vol. 12, pp. 109-116.

Marrafa, P. C. Ribeiro & R. Santos (2004), *Gathering Information from a Relational Lexical-Conceptual Database: A Natural Language Question-Answering System*, in Proceedings of The 8th World Multi-Conference on Systemics, Cybernetics and Informatics, Orlando, Florida, USA.

Mendes, S. & R. P. Chaves (2001), *Enriching WordNet with Qualia information*, in Proceedings of the Workshop on WordNets and Other Lexical Resources at NAACL 2001 Conference, Pittsburgh, pp. 108-112.

Mendes, S. (2005), *Adjectives in WordNet.PT*, paper accepted at the GWA06 Conference.

Miller, K. J. (1998) *Modifiers in WordNet*, in Fellbaum, C. (ed.) "WordNet: an electronic lexical database", Cambridge, MA: The MIT Press, pp. 47-68.

Ribeiro, C., R. Santos, J. Correia, R. P. Chaves, & P. Marrafa (2004), *Inquer: a WordNet-based Question-Answering Application* in Proceedings of LREC 2004, International Conference on Language Resources and Evaluation, Lisboa, Portugal, pp. 1947-1950.

Vossen, P. (1998) (ed.), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Dordrecht: Kluwer Academic Publishers

Vossen, P. (1999), *EuroWordNet General Document*, University of Amsterdam.