Semi-Automatic UNL Dictionary Generation using WordNet.PT

Catarina Ribeiro, Ricardo Santos, Rui Pedro Chaves and Palmira Marrafa

Universidade de Lisboa, CLUL
CLG – Computation of Lexical and Grammatical Knowledge Research Group
Av. Prof. Gama Pinto, 2
1649-003 Lisboa, Portugal
{catarina.ribeiro,ricardo.santos,rui.chaves}@clul.ul.pt, Palmira.Marrafa@netcabo.pt

Abstract

The increase of Internet users all over the world and the subsequent growth of available multilingual information on the Web have brought new challenges to machine translation systems. The Universal Networking Language (UNL) is a meta-language developed for conveying linguistic expressions in order to encode websites information into a standard representation. In order to integrate Portuguese into this platform it is necessary to develop both a dictionary and a grammar. This paper focuses on the development of a PT-UNL dictionary using the WordNet.PT knowledge base. UNL makes use of lexical-semantic relations that have some correspondence in WordNet.PT. Since building the dictionary is a very time-consuming task, a semi-automatic process to generate it has been developed and is described here. The reuse of manually built lexical resources is of great relevance both for the economy of the project and for the reliability of the results.

Keywords: Universal concepts/words, Linguistic Knowledge Base, Lexical-semantic relations, Inheritance.

1. Introduction

In the past few years, machine translation techniques have been applied to web environments. The growing amount of available multilingual information on the WWW, as well as the increase of Internet users has led to a justifiable interest on this area. The UNL system main goal is to provide Internet users access to multilingual websites using a common representation. This will allow users to visualize websites in their own language, whether it has been built under a different language or not. This has a growing relevance since the usage of the WWW is generalized across cultural and linguistic barriers. Many languages such as Arabic, French, Russian, Spanish, Italian, English, Chinese or Brazilian Portuguese have already been included in the UNL platform. Our aim is to introduce European Portuguese into this system. In order to implement this project with the lowest time and human effort costs, we will reuse linguistic resources already available as much as possible. This paper focuses on the semi-automatic generation of the dictionary.

Lexical knowledge representation is a critical issue in natural language processing systems. Recently, the development of large-scale lexica with specific formats capable of being used by several different kinds of applications has been given special focus, in particular to multilingual systems.

The manual development of such resources is a very time--consuming and expensive task. In this context, the reuse of available resources is a major goal in the large domain of Language Engineering.

This paper describes an application to port information from the Portuguese WordNet database to the Portuguese UNL Dictionary, in a semi-automatic way.

This tool will take advantage of the similarity between the WordNet.PT's lexical semantic relations and the UNL's ones. Moreover, semantic attributes can be derived from the hierarchical structure of WordNet.PT. Inheritance mechanisms are applied to deduce and extract such knowledge.

Similar previous work has been done by Verma & Bahattacharyya (2003) for English, Hindi and Marathi. Such method is fully automatic and aims at generating document specific UNL Dictionaries for such languages. Our approach does not make use of documents; rather it encodes WordNet.PT's information directly in a supervised way. A semi-automatic process allows the encoding of morphological rules in UNL dictionary entries.

In section 2, a general description of WordNet.PT is presented and, in section 3, an overview of UNL is given. The UNL dictionary and its specifications are introduced in section 4 and section 5 describes the tool developed and a semi-automatic process of PT-UNL dictionary generation. Finally, section 5 contains concluding remarks and future work.

2. WordNet.PT

WordNet.PT is a lexical database developed under the EuroWordNet framework (Vossen, 1999). EuroWordNet is a multilingual network with several European wordnets interrelated by Inter-Lingual-Index (ILI) based on the Princeton WordNet (Miller et al., 1990; Fellbaum, 1998). All the individual wordnets are structured under the same lines as the Princeton WordNet.

Lexicalised concepts are the basic units of a wordnet and can be represented by several word forms that are grouped in a 'synset' (set of synonyms). The meaning of a lexical unit is derived from the lexical-semantic relations it establishes with other members of the synset as well as with other synsets. Synonymy is the most basic semantic relation in these networks.

Each synset is linked via ILI to its equivalent on the Princeton WordNet. This external relation allows the translation of concepts in different languages.

A wordnet ontology is hierarchically structured via the hyponymy/hyperonymy relation. This semantic relation is fundamental on the network and can be defined as follows: A is a hyponym of B (B hyperonym of A) iff A is

a kind of B and B is not a kind of A. For instance, *cat* is a hyponym (kind) of *feline* and *feline* is a hyperonym of *cat*.
Other major relation is meronymy/holonym, the part-whole relation, that, such as the hyponymy/hyperonymy relation, comes in inverse pairs (i.e. meronymy and holonymy are symmetric relations). So, if A is a holonym of B, B is the meronym of A. For instance, *cat* is a holonym (whole) of *tail* and, consequently, *tail* is a meronym (part) of *cat*.
Event and thematic relations are also included on this database.
WordNet.PT has been manually developed and contains around 13K lexical items.

3. Universal Networking Language

UNL is a formalism that allows the processing of information across linguistic barriers (cf. UNL Specifications 2003). This artificial language has been developed to convey linguistic expressions of natural languages for machine translation purposes. Such information is expressed in an unambiguous way through a semantic network with hyper-nodes. Nodes (that represent concepts) and arcs (that represent relations between concepts) compose the network.
UNL contains three main elements: (i) Universal Words (UW) (cf. UW Manual 2003) which represent concepts and are inter-linked with other UWs to form sentences; (ii) relation labels that express relations between UWs and specify its role within a sentence; and (iii) attribute labels that express linguistic properties standing for contextual restrictions. These elements are combined in order to establish a hierarchical Knowledge Base (UNLKB) that defines unambiguously the semantics of UWs. The UNLKB considered during the development of the system contains 21413 UWs.
The UNL Development Set provides tools that enable the semi-automatic conversion of natural language into UNL and vice-versa. Two of such tools are the EnConverter and the DeConverter. The EnConverter (cf. EnConverter Specifications, 2002) main role is to translate natural language sentences into UNL sentences. This tool implements a language independent parser that provides a framework for morphological, syntactic and semantic analysis synchronously. This allows morphological and syntactical ambiguities resolution. The DeConverter (cf. DeConverter Specifications, 2002), on the other hand, is a language independent generator of natural language sentences from UNL expressions. It also includes a morphological and syntactical framework and word selection for collocations.

4. UNL Dictionary

The UNL dictionary entries require a specific encoding. A general entry is presented in (1):

(1) `[lex] “UW” (ATTRIBUTE,…) <L,F,R>`

The label `lex` represents a lexical expression that corresponds to a lexicalized concept in the source language and to an `UW`. The `UW` identifies the meaning of `lex` and contains both an English translation (headword) and a constraint list. The `UW` allows lexical translation into different languages. A list of labels (`ATTRIBUTEs`) is also required in order to encode syntactic attributes (such as category), semantic attributes (which stand for linguistic contextual restrictions) and “pointers” to morphological rules (applicable to `lex`). Such labels refer to `lex` and not to UWs.
The last component of the lexical entry is a set of tags that correspond to, respectively, the source language flag (`L`), the frequency of `lex` (`F`), and its priority (`R`). An example of a PT-UNL dictionary entry is given in (2):

(2) `[abelha] "bee(icl>insect)" (ATTR,…) <P,0,0>;`

5. Semi-Automatic Generation of PT-UNL Dictionary

To introduce European Portuguese into UNL system both a dictionary and a grammar (which includes syntactic, semantic and morphological rules) are required.
However, the construction of such resources is a time consuming and very expensive task.
This paper focuses on the development of the dictionary through a supervised process. Our goal is to port already available resources, namely from the Portuguese wordnet ontology (WordNet.PT), into the UNL dictionary. WordNet.PT seems a logical choice since it contains information that is useful for UNL Dictionary encoding. Both UNL and wordnets make use of comparable lexical semantic relations to build a structured knowledge base.
A PT-UNL dictionary entry includes a lexical expression in Portuguese (cf. `lex`), its English translation (cf. headword) and a set of constrains that restrict its meaning.
A lexical expression in the PT-UNL dictionary corresponds to a variant of a synset in WordNet.PT. Moreover, the headword can roughly be seen as the ILI record. The constraint list includes, amongst others, the *icl* (includes) and the *pos* (part-of) relations. Such UNL relations are equivalent, respectively, to the hyperonymy/hyponymy and the meronymy/holonymy relations in WordNet.PT. So, on the generation of the constraint list, both knowledge bases (WordNet.PT and UNLKB) can considered.
In what concerns the attribute list generation, syntactic attributes can be obtained from WordNet.PT. Some of the semantic features can be automatically extracted via an inheritance mechanism, since some semantic properties are inherited throughout the WordNet hierarchy. Moreover, the applicable morphological rules can be automatically identified through the analysis of the lexical expression ending. The correct morphological rule is then selected in a supervised process.
The general structure of the developed tool to support the PT-UNL dictionary generation is presented in Figure 1. This system includes the UNLKB and WordNet.PT as main databases (stored in *mysql*), attribute rules and morphological rules databases and an inference engine. This tool has a web interface that allows the supervision of the PT-UNL dictionary encoding.

5.1. Generating UWs

The first step of the whole process is the extraction of a lexical expression from WordNet.PT and the corresponding translation via ILI (step ❶ in Figure 1). As mentioned, a lexical expression is an element (variant) of a *synset* which is linked to its equivalent in Princeton Wordnet. Using a SQL query it is possible to extract a variant ($Term_3$) and a set of translations. This set is

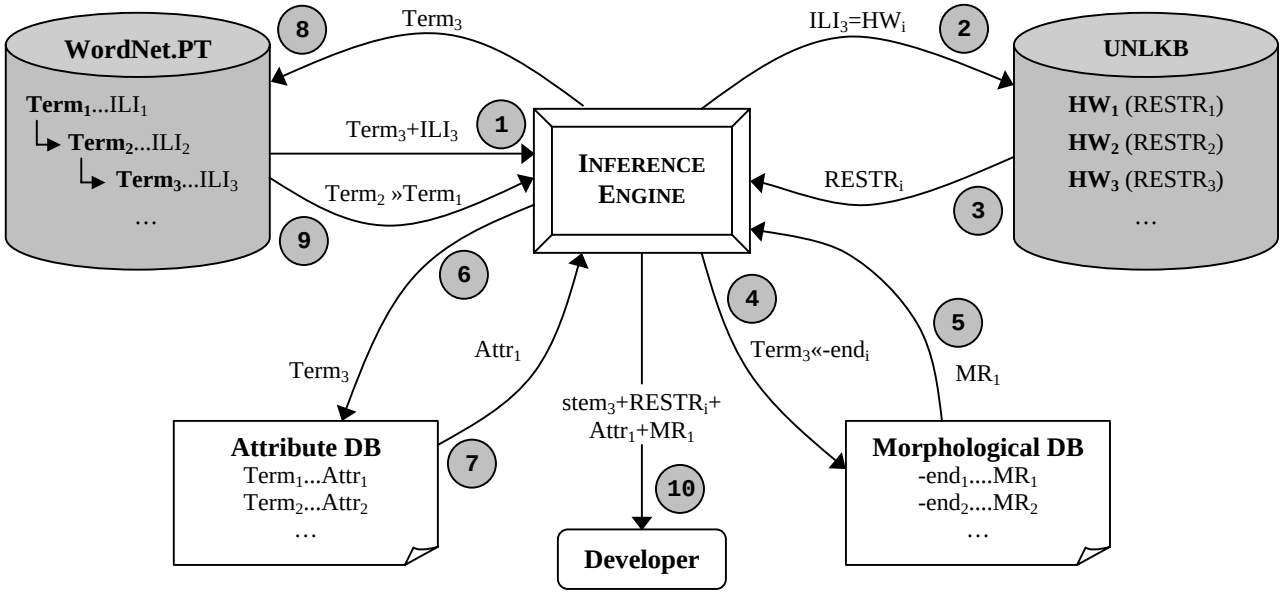

Figure 1 – General structure of the PT-UNL Dictionary supporting tool

decomposed into its basic elements (ILI_3) which are compared to headwords (HW_i) in the UNLKB ❷. If there is at least one successful match, the corresponding constraint list is extracted ($RESTR_i$) ❸. If that is not the case, a new SQL query is submitted in order to obtain $Term_3$ direct hyperonyms and meronyms (and their translations – ILI_2). The search process is then repeated (❷ and ❸) using such translations. If there is a match, a constraint list is generated considering ILI_3, the *icl* or *pof* relation and ILI_2. For example, let $Term_3$ be *abelha*, ILI_3 *bee*, $Term_2$ *insecto*, ILI_2 *insect* and let $Term_2$ be the direct hyperonym of $Term_3$: if *bee* does not match any headword in the UNLKB but *insect* does, the generated UW would be `bee(icl>insect)`.

A recursive climbing mechanism is applied: steps ❶, ❷ and ❸ are repeated until either a UW is generated or no more hyperonyms/meronyms exist. If the latter occurs the generated UW contains only the headword (ILI_3). In such cases, the developer must manually create the correct UW. The proposed UW generation process profits from the information in WordNet.PT and UNLKB. The latter's contents is preferred when available. On the other hand, new concepts are correctly placed in the network.

5.2. Identifying the Applicable Morphological Rules

In order to reduce the dictionary size to a minimum and to promote resource optimization, the UNL system contains a mechanism capable of generating all the inflected forms of a given stem. Morphological rules are encoded and applied so that the system's efficiency can be increased.

Each dictionary entry includes a "pointer" (in the attribute list) to the morphological rule that allows the generation of all inflected forms of the considered lexical expression.

The supporting tool makes use of a Portuguese lexical endings database in which each entry contains an ending and its respective morphological rule. For identical endings, different morphological rules can be applied; so, more than one entry per ending can be found. For instance, word forms whose singular ending is *ão*, can inflect in *ões*, *ães* or *ãos* for plural.

In our system the inference engine analyses the word form ($Term_3$) and identifies its ending. Using a *mysql* query the inference engine interacts with the lexical endings database (❹) and selects all the applicable morphological rules (❺). The correct rule is chosen on a supervised process. For example, considering the word form *gato* (cat), the selected morphological rules would be (3):

(3) i. MR1: generates *gato*/*gato**s*** (ms/mp)
ii. MR2: generates *gato*/**gato**es*** (ms/mp)
iii. MR3: generates *gat**o***/*gat**os***/*gat**a***/*gat**as*** (ms/mp/fs/fp)

(where *m*-masculine, *f*-feminine, *s*-singular and *p*-plural)

After the supervised selection, the generated dictionary entry would include MR3 in its attribute list (4):

(4) `[gat] "cat(icl>feline)" (MR3,…) <P,0,0>;`

At present the system makes use of 31 morphological rules.

5.3. Determining Semantic Attributes

The next step of the process is the generation of the semantic attributes list. The association of semantic attributes to certain concepts of WordNet.PT can be conducted in order to convey semantic properties inherent to those concepts. Only the most general concepts (top hyperonyms of those which possess a certain property) are labeled and inserted into the attribute database. The nature of the hyperonymy relation legitimates the straightforward application of a transitive inheritance mechanism that assign a concept missing attributes from its hyperonyms. For example, let $Term_3$ be *bee*, $Term_2$ *insect* and $Term_1$ *animal* and let $Term_1$ be the direct hyperonym of $Term_2$ which is the direct hyperonym of $Term_3$: if $Term_1$ possess the "animate" property, it is possible to automatically assign such property to $Term_3$ through this inheritance mechanism.

In our system, the inference engine interacts with the attribute database through a SQL query (❻) in order to extract (if any) the lexical expression attribute ($Attr_1$) – ❼. The inheritance mechanism is then applied. The WordNet.PT architecture is exploited so that all the attributes of $Term_3$ are deducted and automatically included on its attribute list. The inference engine interacts with WordNet.PT in order to obtain the hyperonyms subtree whose leaf node is $Term_3$ (❽ and ❾). Foreach hyperonym a new SQL query is submitted to the attribute database so that all semantic attributes of $Term_3$ are selected (steps ❻ and ❼ are repeated).
The extracted semantic attributes are automatically added to the lexical entry of $Term_3$. However, not all of the required semantic attributes can be deducted. In this case, they have to be manually chosen.
At present 133 WordNet.PT's concepts are associated to semantic attributes.

5.4. Ambiguities: Selecting One from Several Options

The last step of the process is the presentation of all the ambiguous information extracted to a human capable of selecting the correct one. In order to do so, we have developed a web interface that presents the information extracted throughout the previous steps and allows the restrict selection of the correct information to be included on the dictionary entry. The supervised process is reduced to few constrained decisions.

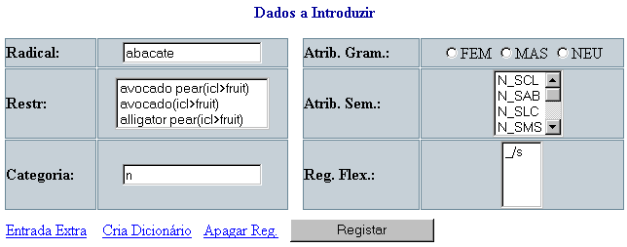

Figure 2 – Interface of the supporting tool

In Figure 2 a screenshot of the interface of the supporting tool is provided. On the left the lexical expression (*radical*), the UW (*restr.*) and the category (*categoria*) are presented. On the right, gender (*atrib. gram.*), semantic attributes (*atrib. sem.*) and morphological rules (*reg. flex.*) fields are provided. The developer must select the correct options.

6. Conclusions and Future Work

In this paper we have presented a semi-automatic process of generation of a PT-UNL dictionary through the reuse of available resources – WordNet.PT and UNLKB. The fact that we can extract information from these lexical databases in a semi-automatic fashion reduces the problem of the UNL dictionary building to a few restricted decisions. A computational tool to assist this process has been developed. Although the entire process is not fully automatic, it can greatly benefit from this application.
Since the approach proposed here is language independent, it can be easily adopted on the development of UNL dictionaries for all languages encoded into a wordnet-like framework.
In the future, we aim to extend the semi-automatic generation of PT-UNL dictionary to verbal concepts. In order to do so, thematic and event WordNet relations can also be exploited on the generation of UWs. For example, a relation such as `involved_agent(bark,dog)` indicates that *bark* needs a subject *dog* which can be used on the generation of the verbal UW `bark(agt>dog)`.
Moreover, statistical methods will be applied in order to reduce and rank the range of possible choices to be presented to the developer. This will ease and accelerate the PT-UNL dictionary generation process.

References

(2002a). DeConverter Specifications, version 2.7. UNL Center.
(2002b). EnConverter Specifications, version 3.3. UNL Center.
(2003a). The Universal Networking Language (UNL) Specifications. United Nations University. Available at URL: http://www.unl.ias.unu.edu/unlsys/.
(2003b). UW Manual. UNL Center.
Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database. In International Journal of Lexicography, Vol.3, Nº 4 (pp.235—244).
Fellbaum, C. (1998). A Semantic Network of English: The Mother of All WordNets. In P. Vossen (ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.
Verma, N. & Bhattacharyya, P. (2003). Automatic Generation of Multilingual Lexicon by Using WordNet. In Proceedings of Convergences'03, International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies, Alexandria, Egipt.
Vossen, P. (1999). EuroWordNet General Document. University of Amsterdam.