First release of the Multilingual Central Repository of MEANING*

Luís Villarejo, Jordi Atserias, Gerard Escudero, German Rigau

TALP Research Center
Jordi Girona Salgado, 1-3.
08034 Barcelona
{luisv,batalla,escudero}@talp.upc.es

IXA Group
Euskal Herriko Unibertsitatea
Donostia.
{rigau}@si.ehu.es

Resumen: Este artículo contine una breve descripción de los principales componentes software del Multilingual Central Repository de MEANING y su contenido inicial.
Palabras clave: Wordnet, EuroWordnet, MultiWordnet, Adquisición Multilingüe

Abstract: This paper provides a brief description of the main software components of the Multilingual Central Repository of MEANING and their initial content.
Keywords: Wordnet, EuroWordnet, MultiWordnet

1 Introduction

The MEANING project (Rigau et al., 2002) [1] plans to perform three consecutive cycles of large-scale WSD and acquisition processes in five European languages.

The knowledge acquired for each language will be consistently upload and integrated into the respective local wordnets, and then ported and distributed across the rest of wordnets, balancing resources and technological advances across languages.

The Multilingual Central Repository (MCR) will grant the consistency and integrity of all the semantic knowledge produced by MEANING, acting as a multilingual interface for integrating and distributing all the knowledge acquired in MEANING.

1.1 MCR

The MCR follows the model proposed by the EuroWordNet (Vossen, 1998) project. EuroWordNet is a multilingual lexical database with wordnets for several European languages, which are structured as the Princeton WordNet (Fellbaum, 1998).

The first version of the MCR includes:

- ILI
  - WordNet 1.6
  - EuroWordNet Base Concepts
  - EuroWordNet Top Ontology
  - MultiWordNet Domains
- Local wordnets
  - English WordNet 1.5, 1.6, 1.7.1
  - Basque, Catalan, Italian and Spanish wordnets
- Large collections of semantic preferences
  - Acquired from SemCor
  - Acquired from BNC
- Instances
  - Named Entities

The Multilingual Central Repository Database plans to represent most of the WordNet and EuroWordNet data and properties, including *language independent information* (ILIs, Top Ontology, domains, etc.) as well as *language dependent data* (synsets, variants, etc). The first release of the MCR database contains four European wordnets and three different English versions and their corresponding links to the ILI [2]

In order to upload all this information, local wordnets based on WordNet1.5 has been mapped to Wordnet 1.6. As well as the original set of Base Concepts and the Top Concept Ontology. Moreover the EuroWordNet Top Concept Ontology has been expanded top-down through the WordNet1.6 structure.

1.2 Web Interface to the MCR

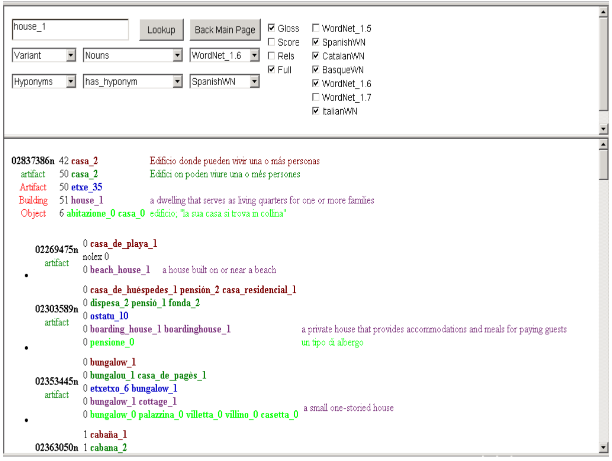
The MCR web interface is based on the Web EuroWordNet Interface (WEI)[3] (Benítez et al., 1998). The interface provides consulting and editing facilities of the data content of the MCR. The basic aim of this tool is to

* This research has been partially funded by the European Comission (MEANING IST-2001-34460)

[1] http://www.lsi.upc.es/~nlp/meaning/meaning.html

[2] http://www.lsi.upc.es/~nlp/tools/mapping.html

[3] http://nipadio.lsi.upc.es/wei.html



house_1

Variant ▾ Nouns ▾ WordNet_1.6 ▾

Hyponyms ▾ has_hyponym ▾ SpanishWN ▾

☑ Gloss
☐ Score
☐ Rels
☑ Full

☐ WordNet_1.5
☑ SpanishWN
☑ CatalanWN
☑ BasqueWN
☑ WordNet_1.6
☐ WordNet_1.7
☑ ItalianWN

```
02837386n 42 casa_2                  Edificio donde pueden vivir una o más personas
  artifact 50 casa_2                 Edifici on poden viure una o més persones
  Artifact 50 etxe_35
  Building 51 house_1                a dwelling that serves as living quarters for one or more families
  Object    6 abitazione_0 casa_0    edificio; "la sua casa si trova in collina"

     02269475n 0 casa_de_playa_1
               nolex 0
     artifact  0 beach_house_1   a house built on or near a beach
•
               0 casa_de_huéspedes_1 pensión_2 casa_residencial_1
     02303589n 0 dispesa_2 pensió_1 fonda_2
     artifact  0 ostatu_10
               0 boarding_house_1 boardinghouse_1            a private house that provides accommodations and meals for paying guests
               0 pensione_0                                  un tipo di albergo
•
               0 bungalow_1
     02353445n 0 bungalou_1 casa_de_pagès_1
     artifact  0 etxetxo_6 bungalow_1
               0 bungalow_1 cottage_1
               0 bungalow_0 palazzina_0 villetta_0 villino_0 casetta_0  a small one-storied house
•
               1 cabaña_1
     02363050n 1 cabana_2
```

Figura 1: Querying for relationships.

provide a flexible access to our multilingual lexical knowledge databases. The Web EuroWordNet Interface (WEI) is a tool that provides to the user all the lexico-semantic information contained in all uploaded WordNets.

WEI allows the user to consult the MCR using a powerful but very intuitive user interface. WEI provides facilities for a flexible querying of the MCR. First, the user can select how to enter to the MCR by providing a word or a variant or a synset of any wordnet uploaded into the MCR. Then, the user must choose one of the wordnets to navegate through some of its semantic relations. Finally, the user select which information and from which wordnet whats to obtain the result of the consultation.

Consulting the MCR using WEI can be done in three different ways, introducing in the search field either: the word (e.g. house), the synset identifier (e.g. 02837386n), or the variant (e.g. house_1).

The information that can be displayed is:

- The *ILI number*
- The *Offset number*
- The *WordNet semantic file* linked to the synset.
- The *Top Ontology Labels.*
- The *MultiWordnet Domain Labels.*
- *Gloss:* A brief description or exemple of the synset.
- *Score:* The confidence index shows whether the synset has been manually validated (99%) o it is the result of an automatic method (this percentage is different depending on the method used).
- *Rels:* Shows the total number and type of relationships for the synset.
- *Full:* Lets showing either the directly related synsets (e.g. direct hyponyms) or all the transitively related synsets (e.g. the full hyponym tree of a synset).
- The seven remaining checkboxes enable/disable the visualization of each WordNet.

2 Conclusions

The first version of the MCR integrates now into the same EuroWordNet framework (using an upgraded release of Base Concepts and Top Ontology and MultiWordNet Domains) five local wordnets (with three English WordNet versions) with hundreds of thousand of new semantic relations, instances and properties fully expanded. After the first MEANING porting, all wordnets gained some kind of new knowledge comming from other wordnets by means of the first porting process. In fact, the resulting MCR is the largest and richest multilingual lexical knowledge base ever build.

References

Benítez, L., S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. 1998. Methods and tools for building the catalan wordnet. In *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages*, Granada, Spain.

Fellbaum, C., editor. 1998. *WordNet. An Electronic Lexical Database.* MIT Press.

Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll. 2002. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of COLLING Workshop*, Taipei, Taiwan.

Vossen, P., editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers .