The MEANING Multilingual Central Repository

J. Atserias[1], L. Villarejo[1], G. Rigau[2], E. Agirre[2],
J. Carroll[3], B. Magnini[4], P. Vossen[5]

[1] TALP Research center, Universitat Politécnica de Catalunya. Catalonia
{batalla, luisv}@talp.upc.es
[2] IXA Group, University of the Basque Country, Computer Languages and Systems
{rigau, eneko}@si.ehu.es
[3] University of Sussex, Cognitive and Computing Sciences. UK
J.A.Carroll@sussex.ac.uk
[4] ITC-IRST Italy
magnini@itc.it
[5] Irion Technologies B.V. The Netherlands
Piek.Vossen@irion.nl

Abstract. This paper describes the first version of the Multilingual Central Repository, a lexical knowledge base developed in the framework of the MEANING project. Currently the Mcr integrates into the EuroWordNet framework five local wordnets (including four versions of the English WordNet from Princeton), an upgraded version of the EuroWordNet Top Concept ontology, the MultiWordNet Domains, the Suggested Upper Merged Ontology (SUMO) and hundreds of thousand of new semantic relations and properties automatically acquired from corpora. We believe that the resulting Mcr will be the largest and richest Multilingual Lexical Knowledge Base in existence.

1 Introduction

Building large and rich knowledge bases takes a great deal of expensive manual effort; this has severely hampered Knowledge-Technologies and HLT application development. For example, dozens of person-years have been invested into the development of wordnets (Wns) [4] for various languages [9] [3], but the data in these resources is still not sufficiently rich to support advanced multilingual concept-based HLT applications directly. Furthermore, resources produced by introspection usually fail to register what really occurs in texts.

The Meaning project [8] [6] identifies two complementary and intermediate tasks which are crucial in order to enable the next generation of intelligent open domain HLT application systems: Word Sense Disambiguation (WSD) and large-scale enrichment of Lexical Knowledge Bases (LKBs). Advances in these two areas will allow large-scale acquisition of shallow meaning from texts, in the form of relations between concepts.

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

However, progress is difficult due to the following interdependence: (i) in order to achieve accurate WSD, we need far more linguistic and semantic knowledge than is available in current LKBs (e.g. current WNS); (ii) in order to enrich existing LKBs we need to acquire information from corpora accurately tagged with word senses.

MEANING proposes an innovative bootstrapping process to deal with this inter–dependency between WSD and knowledge acquisition exploiting a multilingual architecture based on EuroWordNet (EWN) [9]. The project plans to perform three consecutive cycles of large-scale WSD and acquisition processes in five European languages including Basque, Catalan, English, Italian and Spanish. As languages realize meanings in different ways, some semantic relations that can be difficult to acquire in one language can be easy to capture in other languages. The knowledge acquired for each language during the three consecutive cycles will be consistently upload and integrated into the respective local WNS, and then ported and distributed across the rest of WNS, balancing resources and technological advances across languages.

This paper describes the first version of the Multilingual Central Repository produced after the first MEANING cycle. Section 2 presents the MCR structure, content and associated software tools. Section 3 describes the first uploading process, and section 4 the porting process. Section 5 and 6 conclude and discusse directions for future work.

2 Multilingual Central Repository

The Multilingual Central Repository (MCR) ensures the consistency and integrity of all the semantic knowledge produced by MEANING. It acts as a multilingual interface for integrating and distributing all the knowledge acquired in the project. The MCR follows the model proposed by the EWN project, whose architecture includes the **Inter–Lingual–Index** (ILI), a **Domain ontology** and a **Top Concept ontology** [9].

The first version of the MCR includes only conceptual knowledge. This means that only semantic relations among synsets have been acquired, uploaded and ported across local WNS. The current MCR integrates: (i) the ILI based in WN1.6, includes EWN Base Concepts, EWN Top Concept ontology, MultiWordNet Domains (MWND), Suggested Upper Merged Ontology (SUMO); (ii) Local WNS connected to the ILI, including English WN 1.5, 1.6, 1.7, 1.7.1, Basque, Catalan, Italian and Spanish WN; (iii) Large collections of semantic preferences, acquired both from SemCor and from BNC; Instances, including named entities.

The MCR provides a web interface to the database based on Web EuroWordNet Interface [7]. Three different APIs have been also developed to provide flexible access to the MCR: first, a SOAP API to allow users to interact with the MCR, an extension of the WNQUERY Perl API to the MCR and a C++ API for high performance software.

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

3 Uploading Process

Uploading consists of the correct integration of every piece of information into the MCR. That is, linking correctly all this knowledge to the ILI. This process involves a complex cross–checking validation process and usually a complex expansion/inference of large amounts of semantic properties and relations through the WN semantic structure (see [2] fot further details).

3.1 Uploading WNs

To date, most of the knowledge uploaded into the MCR has been derived from WN1.6 (or SemCor); the Italian WN and the MWND, both use WN1.6 as ILI. However, the ILI for Spanish, Catalan and Basque WNs was WN1.5, as well as the EWN Top Concept ontology and the associated Base Concepts. To deal with the gaps between versions and to minimize side effects with other international initiatives (Balkanet, EuroTerm, eXtended WN) and WN developements around Global WordNet Association, we used a set of improved mappings between all involved resources [8].

3.2 Uploading Base Concepts

The original set of **Base Concepts** from EWN based on WN1.5 contained a total of 1,030 ILI-records. Now, the Base Concepts from WN1.5 have been mapped to WN1.6. After a manual revision and expansion to all WN1.6 top nodes, the resulting Base Concepts for WN1.6 total 1,535 ILI-records. In this way, the new version of Base Concepts covers the complete hierarchy of ILI-records (only nouns and verbs).

3.3 Uploading the Top Ontology

The purpose of the EWN **Top Concept ontology** was to enforce more uniformity and compatibility of the different WN developments. The EWN project only performed a complete validation of the consistency of the **Top Concept ontology** of the Base Concepts.

Although the classification of WN is not always consistent with the **Top Concept ontology**, we performed an automatic expansion of the **Top Concept** properties assigned to the Base Concepts. That is, we enriched the complete ILI structure with features coming from the Base Concepts by inheriting the Top Concept features following the hyponymy relationship. The **Top Concept ontology** has been uploaded in three steps:

1. Properties are assigned to WN1.6 synsets through the mapping.
2. For those WN1.6 Tops (synsets without any parent) that do not have any property assigned through the mapping, we assigned to them the Top Concept ontology properties by hand.

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

3. The properties are propagated top–down through the WN hierarchy

The following incompatibilities inside the **Top Concept ontology** have been used to block the top–down propagation of the **Top Concept** properties:

- *1stOrderEntity - 2ndOrderEntity - 3rdOrderEntity*,
- *substance - object*,
- *plant - animal - human - creature*,
- *natural - artifact*,
- *solid - liquid - gas.*

Thus, when detecting that any of the current **Top Concept ontology** properties of a synset is incompatible with other inherited (due possibly to multiple inheritance), this property is not assigned to the synset and the propagation to the synset's descendants stops.

3.4 Uploading SUMO

The Suggested Upper Merged Ontology (SUMO) [7] is an upper ontology created at Teknowledge Corporation and proposed as starting point for the IEEE Standard Upper Ontology Working group.

SUMO provides definitions for general purpose terms and is the result of merging different free upper ontologies (e.g. Sowa's upper ontology, Allen's temporal axioms, Guarino's formal mereotopology, etc.) with WN1.6. Currently only the SUMO labels and the SUMO ontology hyperonym relations are loaded into the MCR. We plan to cross–check the **Top Concept ontology** expansion and the **Domain ontology** with the SUMO ontology.

3.5 Uploading Selectional Preferences

A total of 390,549 weighted Selectional Preferences (SPs) obtained from two different corpora and using different approaches have been uploaded into the MCR. The first set [6] of weighted SPs was obtained by computing probability distributions over the WN1.6 noun hierarchy derived from the result of parsing the BNC. The second set [1] was obtained from generalizations of grammatical relations extracted from Semcor.

The SPs are included in the MCR as ROLE noun–verb relations [9]. Although we can distinguish subjects and objects, all of them have been included as a more general ROLE relation.

4 Porting Process

In the first porting process all the knowledge integrated into the MCR has been ported (distributed) directly to the local WNs (no extra semantic knowledge

[9] In EWN , INVOLVED and ROLE relationships are defined symmetrically.

	Spanish		English		Italian	
Relations	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**
be_in_state	1,302	=	1,300	+2	364	+2
causes	240	=	224	+19	117	+15
near_antonym	7,444	=	7,449	+221	3,266	=
near_synonym	10,965	=	21,858	+19	4,887	+54
role	106	=	0	+106	0	+46
role_agent	516	=	0	+516	0	+227
role_instrument	291	=	0	+291	0	+151
role_location	83	=	0	+83	0	+39
role_patient	6	=	0	+6	0	+3
xpos_fuzzynym	37	=	0	+37	0	+23
xpos_near_synonym	319	=	0	+319	0	+181
Other relations	31,644	=	29,120	+2,627	9,541	+22
Total	53,272	=	59,951	+4,246	18,175	+763
role_agent-semcor	0	+52,394	69,840	=	0	+41,910
role_agent-bnc	0	+67,109	95,065	=	0	+40,853
role_patient-semcor	0	+80,378	110,102	=	0	+41,910
role_patient-bnc	0	+79,443	115,102	=	0	+50,264
Role	0	+279,324	390,109	=	0	+174,937
Instances	0	+1,599	0	+2,198	791	=
Proper Nouns	1,806	=	17,842	=	2,161	=
Base Concepts	1,169	=	1,535	=	0	+935
Domains Links	0	+55,239	109,621	=	35,174	=
Domains Synsets	0	+48,053	96,067	=	30,607	=
Top Ontology Links	3,438	=	0	+4,148	0	+2,544
Top Ontology Synsets	1,290	=	0	+1,554	0	+946

Table 1. PORT0 Main figures for Spanish, English and Italian

has been inferred in this process). Table 1 summarises the main results before (UPLOAD0) and after the whole porting process (PORT0) for Spanish, English and Italian. In this table, relations do not consider hypo/hyperonym relations and *links* stands for total number of Domains or Top Concept ontology properties ported (before application of the top-down expansion process).

4.1 An example

When uploading coherently all this knowledge into the MCR, we added consistently a large set of explicit knowledge about each sense which can be used to differentiate and characterize better their particular meanings. We will illustrate the current content of the MCR, after porting, with a simple example: the Spanish noun *pasta*.

<table>
<tr>
<td>Domain: chemistry-pure_science
Semantic File: 27-Substance
SUMO:
Substance-SelfConnectedObject-Object-Physical-Entity

Top Concept ontology
Natural-Origin-1stOrderEntity
Substance-Form-1stOrderEntity

pasta#n#7 10541786-n
paste#1
gloss: any mixture of a soft and malleable consistency</td>
<td>Domain: money-economy-soc.science
Semantic File: 21-MONEY
SUMO:
CurrencyMeasure-ConstantQuantity-PhysicalQuantity-Quantity-Abstract-Entity
Top Concept ontology
Artifact-Origin-1stOrderEntity
Function-1stOrderEntity
MoneyRepresentation-Representation-Function-1stOrderEntity

pasta#n#6 09640280-n
dough#2,bread#2,loot#2, ...
gloss: informal terms for money</td>
</tr>
<tr>
<td colspan="2">Domain: gastronomy-alimentation-applied_science
Semantic File: 13-FOOD
Top concept ontology
Comestible-Function-1stOrderEntity
Substance-Form-1stOrderEntity</td>
</tr>
<tr>
<td>Top Concept ontology
Natural-Origin-1stOrderEntity

Top Concept ontology
Part-composition-1stOrderEntity
pasta#n#4 05886080-n
spread#5,paste#3
gloss: a tasty mixture to be spread on bread or crackers</td>
<td>pasta#n#1 05671312-n
pastry#1,pastry_dough#1
gloss: a dough of flour and water and shortening

pasta#n#3 05739733-n
pasta#1,alimentary_paste#1
gloss: shaped and dried dough made from flour and water & sometimes egg

pasta#n#5 05889686-n dough#1
gloss: a dough of flour and water and shortenings</td>
</tr>
<tr>
<td>Top Concept ontology
Artifact-Origin-1stOrderEntity
Group-Composition-1stOrderEntity</td>
<td>pasta#n#2 05671439-n
pie_crust#1,pie_shell#1
gloss: pastry used to hold pie fillings</td>
</tr>
</table>

Table 2. Food senses for the Spanish word *pasta*

The word *pasta* (see table 2) illustrates how all the different classification schemes uploaded into the MCR: Semantic File, MWND, Top Concept ontology, etc. are consistent and makes clear semantic distinctions between the money sense (pasta_6), the general/chemistry sense (pasta_7) and the food senses (all the rest). The food senses of *Pasta* can now be further differentiate by means of explicit EWN Top Concept ontology properties. All the food senses are descendants of *substance_1* and *food_1* and inherits the Top Concept attributes *Substance* and *Comestible* respectively.

Selectional Preferences can also help to distinguish between senses, e.g only the money sense has the following preferences as object: *1.44 01576902-v* {*raise#4*}, *0.45 01518840-v* {*take_in#5, collect#2*} or *0.23 01565625-v* {*earn#2, garner#1*}.

We will investigate new inference facilities to enhance the uploading process. After full expansion (**Realization**) of the EWN Top Concept ontology properties, we will perform a full expansion through the noun part of the hierarchy of the selectional preferences acquired from SemCor and BNC (and possibly other implicit semantic knowledge currently available in WN such as meronymy information).

We plan further investigation to perform full bottom-up expansion (**Generalization**), rather than merely expanding knowledge and properties top-down. In this case, different knowledge and properties can collapse on particular Base Concepts, Semantic Files, Domains and/or Top Concepts.

5 Future Work

Having all these types of different knowledge and properties coming from different sources, methods, and completely expanded through the whole MCR, a new set of inference mechanisms can be devised to further infer new relations and knowledge inside the MCR. For instance, new relations could be generated when detecting particular *semantic patterns* occurring for some synsets having certain ontological properties, for a particular Domain, etc. That is, new relations could be generated when combining different methods and knowledge. For instance, creating new explicit relations (regular polysemy, nominalizations, etc.) when several relations derived in the integration process have confidence scores greater than certain thresholds, occurring between certain ontological properties, etc.

Obviously, new research is also needed for porting the various types of knowledge across languages. For instance, new ways to validate the ported knowledge in the target languages.

6 Conclusions

The first version of the MCR integrates into the same EWN framework (using an upgraded release of Base Concepts and Top Concept ontology and MWND) five local WNS (with four English WN versions) with hundreds of thousands of new semantic relations, instances and properties fully expanded. All WNS have gained some kind of knowledge coming from other WNS by means of the first porting process. We believe that the resulting MCR is the largest and richest multilingual LKB in existence.

We intend this version of the MCR to be a natural multilingual large-scale knowledge resource for a number of semantic processes that need large amounts of linguistic knowledge to be effective tools (e.g. Semantic Web ontologies).

When uploading coherently all this knowledge into the MCR a full range of new possibilities appears for improving both Acquisition and WSD tasks in the next two MEANING rounds.

Future versions of the MCR may include language dependent data, such as syntactic information, subcategorization frames, diathesis alternations, Dorr's Lexical Conceptual Structures, complex semantic relations [5], etc. The information will be represented following current standards (e.g. EAGLES), where these exist.

Regarding the *porting process*, we will investigate inference mechanisms to infer new explicit relations and knowledge (regular polysemy, nominalizations, etc). Finally, more research is needed to verify the correctness of the various types of semantic knowledge ported across languages.

Acknowledgments

This research has been partially funded by the Spanish Research Department (HERMES TIC2000-0335-C03-02) and by the European Commission (MEANING IST-2001-34460).

References

1. E. Agirre and D. Martinez. Integrating selectional preferences in wordnet. In *Proceedings of the first International WordNet Conference in Mysore*, India, 21-25 January 2002.
2. J. Atserias, L. Villarejo, and G. Rigau. Integrating and porting knowledge across languages. In *Procceeding of Recent Advances in Natural Language Processing (RANLP'03)*, pages 31–37, Bulgaria, 2003.
3. L. Bentivogli, E. Pianta, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India, 2002.
4. C. Fellbaum, editor. *WordNet. An Electronic Lexical Database.* The MIT Press, 1998.
5. D. Lin and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):343–360, 2001.
6. D. McCarthy. *Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences.* PhD thesis, University of Sussex, 2001.
7. I. Niles and A. Pease. Towards a standard upper ontology. In *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 17–19. Chris Welty and Barry Smith, eds, 2001.
8. G. Rigau, B. Magnini, E. Agirre, P. Vossen, and J. Carroll. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLLING Workshop 'A Roadmap for Computational Linguistics'*, Taipei, Taiwan, 2002.
9. P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers , 1998.