

Towards the MEANING Top Ontology: Sources of Ontological Meaning

Jordi Atserias*, Salvador Climent**, German Rigau†

* TALP Research Center. Universitat Politècnica de Catalunya. batalla@talp.upc.es

** Universitat Oberta de Catalunya. scliment@uoc.edu

† IXA Group. University of the Basque Country. rigau@si.ehu.es

Abstract

This paper describes the initial research steps towards the Top Ontology for the Multilingual Central Repository (MCR) built in the MEANING project. The current version of the MCR integrates five local wordnets plus four versions of Princeton's English WordNet, three ontologies and hundreds of thousands of new semantic relations and properties automatically acquired from corpora. In order to maintain compatibility among all these heterogeneous knowledge resources, it is fundamental to have a robust and advanced ontological support. This paper studies the mapping of main Sources of Ontological Meaning onto the wordnets and, in particular, the current work in mapping the EuroWordNet Top Concept Ontology.

1. Introduction: the MEANING Project

MEANING¹ (Rigau et al., 2002) is a UE-funded project (IST-2001-34460) which has as one of its major goals the integration of several large-scale knowledge resources. MEANING has designed a Multilingual Central Repository (MCR) to act as a multilingual interface for integrating and distributing all the knowledge acquired in the project (Atserias et al., 2004). The MCR follows the model proposed by the EuroWordNet project (EWN): a multilingual lexical database with wordnets for several languages.

The EWN architecture includes the Inter-Lingual-Index (ILI), a preliminary Domain Ontology (DO) and a Top Concept Ontology (TCO) (Vossen, 1998). The ILI consists of a flat list of records that interconnect synsets across wordnets. During the EWN Project around 1000 ILI-Records were selected as Base Concepts (BC) and connected to the TCO.

Using the ILI, wordnets in the MCR are interconnected so that it is possible to go from word meanings in one language or particular wordnet to their equivalents in other languages or wordnets.

In EWN, the ILI was enhanced, enriched and structured by two separate ontologies:

- the **Top Concept ontology** (TCO), which is a hierarchy of language-independent concepts, reflecting important semantic distinctions, e.g. Object, Substance, Location, Dynamic;
- the **Domain ontology** (DO), which is a hierarchy of domain labels, which are knowledge structures grouping meanings in terms of topics or scripts, e.g. Transport, Sports, Medicine, Gastronomy;

The main purpose of the TCO is to provide a common framework for all the wordnets. It consists of 63 basic semantic distinctions that classify a set of ILI-records connected to EWN BC which represents the most important concepts in the different wordnets.

The current version of the MCR uses the set of Princeton WordNet (WN) 1.6 synsets as ILI. Initially most of the knowledge to be uploaded into the MCR has been derived from WN (automatic selection preferences acquired from SemCor and BNC) and the Italian wordnet and the MultiWordNet Domains, both developed at IRST are using WordNet 1.6 as ILI (Bentivogli et al., 2002; Magnini and Cavagli, 2000). This option also minimises side effects with other European initiatives (Balkanet, EuroTerm, etc.) and wordnet developments around Global WordNet Association. However, the ILI for Spanish, Catalan and Basque wordnets, the EWN TCO and the associated BC were based on WordNet 1.5 (Atserias et al., 1997; Benítez et al., 1998).

After this short introduction, section 2. describes the main ontological resources used in MEANING. Section 3. presents the inheritance mechanism used to expand the TCO properties. In section 4. we present the semi-automatic approach we plan to follow to perform consistency checking of the TCO related to the diverse conceptual information used in MEANING. Finally, section 5. provides some concluding remarks.

2. MEANING's Ontological Meaning

Although wordnets and ontologies are both graphs connecting concepts, they are different in nature: while wordnets build concepts upon lexical units of a particular language, nodes in ontologies are claimed to be language-independent concepts. Wordnets can be

¹<http://www.lsi.upc.es/~nlp/meaning>

straightforwardly used for NLP tasks. On the contrary, ontologies, although being meaningful constructs, can not be straightforwardly used for NLP unless they are associated to linguistic units and structures.

Moreover, different ontologies usually are designed using different theoretical grounds; e.g. while SUMO incorporates previous ontologies and insights by Sowa, Pierce, Russell and Norvig and others, the TCO is based on more linguistic grounds: Lyons, Vendler, Verkuyl and Pustejovsky. Therefore, although different ontologies can be comparable, it would take a great theoretical effort to merge all of them in a unique standard and comprehensive construct to be consistently associated to WN.

For this reason, in MEANING we intend to adopt a hybrid and simple approach: to build the MEANING TO, different Sources of Ontological Meaning (SOM) are assigned to language-independent ILI-records so that they can be mapped to WN concepts and expanded throughout them using its internal semantic relations. The different SOM do not need to be equivalent nor even compatible as they will stand as independent information. Besides, no claim of completion will be made.

Currently, MCR integrates through ILI different SOM:

1. An upgraded version of the EWN Base Concepts (BC)
2. An upgraded version of the EWN Top Concept Ontology (TCO)
3. The WordNet Domains Ontology (DO) (Magnini and Cavagli, 2000), a hierarchy of 165 domain labels
4. The Suggested Upper Merged Ontology, SUMO (Niles and Pease, 2001)
5. WN Semantic Files (SF), corresponding to lexicographical files from wordnet, e.g. noun.animal, verb.possession, etc.

The integration of all these SOM into a single platform both demands and allows for cross-checking. For instance, we can improve SUMO labels and WordNet Domains mappings by merging and comparing them.

To illustrate how we can detect errors and inconsistencies between different types of SOM, we can see in the example in table 1 that the nouns corresponding to the SUMO process Breathing has been labeled with ANATOMY domain, some verbs with MEDICINE and some adjectives with FACTOTUM, when in fact, all these senses correspond to different Part-of-Speech of the same concept.

Synset	Word	SUMO	Domain
00003142v	exhale	Breathing	medicine
00899001a	exhaled	Breathing	factotum
00263355a	exhaling	Breathing	factotum
00536039n	expiration	Breathing	anatomy
02849508a	expiratory	Breathing	anatomy
00003142v	expire	Breathing	medicine

Table 1: SUMO vs. Domain labels

In MEANING the TCO has been uploaded in four steps (see (Atserias et al., 2003) for further details):

1. Upgrading the WN1.5 BC to WN1.6
2. TCO properties have been assigned to WN1.6 synsets through the WN 1.5 to 1.6 mapping (Daudé et al., 2001).
3. For those WN1.6 Tops (synsets without any parent) that do not have any assigned property through the mapping, we assigned to them the TCO properties via a table of equivalence between TCO and SF.
4. The resulting properties were propagated top-down through the WN hierarchy

The original set of BC from EWN based on WN1.5 totaled 1,030 ILI-records. Now, the BC from WN1.5 have been mapped to WN1.6. After a manual revision and expansion to all WN1.6 top beginners, the resulting BC for WN1.6 totaled 1,601 ILI-records. In that way, the new version of BC covers the complete hierarchy of ILI-records.

3. Expanding TCO properties

The EWN project only performed a complete validation of the consistency of the TCO at the BC level.

Assuming (as the builders of SUMO and DO have done) that the ontological properties have been correctly assigned to particular synsets and that WN defines coherent subsumption chains, an automatic process can consistently inherit all the properties through the whole hierarchy of WN - no matter the ontology they come from.

MEANING have performed an automatic expansion of the TCO properties assigned to the BC. That is, we enriched the complete ILI structure with features coming from the BC by inheriting the Top Concept features following the hyponymy relationship.

This way, once ontological properties are exported to the ILI and inherited through the whole WN Hierarchy, all concepts in a WN will result to be assigned

with a set of semantic features as in the following example.

lentil_1	
<i>WD</i>	gastronomy
<i>SF</i>	food
<i>SUMO</i>	FruitOrVegetable
<i>TCO</i>	Comestible ; Plant

In order to provide consistency to the inheritance process we used the following basic incompatibilities among TCO properties which were defined inside the EWN project:

- substance - object
- plant - animal - human - creature
- natural - artifact
- solid - liquid - gas

As the classification of WN is not always consistent with the TCO, these incompatibilities impeded the full automatic top-down propagation of the TCO properties. That semi-automatic process resulted in a number of synsets showing non-compatible information. Specifically:

- Sticking to TCO and according to the set of incompatibilities, some TCO properties assigned by hand appeared to be incompatible with either (a) inherited information, (b) information assigned via equivalence to *SF* or/and even (c) other TCO properties assigned by hand.
- TCO properties, either original or inherited, are suspicious to be incompatible with other SOM.

By examining a subset of synsets, we realised that there are at least the following main sources of errors:

- Erroneous hand-made TCO mappings
- Erroneous statements of equivalence between TCO properties and *SFs*
- Erroneous ISA links in WN -which causes erroneous inheritance (Guarino and Welty, 2000)
- Multiple inheritance within WN can cause incompatibilities in inheritance of properties

The following example has incompatible information. 3rdOrderEntity can not coexist with properties only attributable to Events:

00660718 process_1	
<i>MWD</i>	factotum
<i>WN16SF</i>	act
<i>SUMO</i>	IntentionalProcess
<i>EWNTO</i>	3rdOrderEntity;Cause;Mental;Purpose

4. Consistency checking

The procedure we will apply to solve the TCO incompatibilities is the following:

1. Hand-fixing TCO mappings where appearing incompatible properties
2. Setting inheritance-blocking-points and hand-fixing TCO mappings around these points (i.e. all involved hypernyms and hyponyms)
3. Recalculating the inheritance according to the information obtained in (1) and (2)
4. Reexamining the involved subtrees to check whether re-calculation of the inheritance produce new incompatibilities
5. Exporting the mappings and blocking-point information to the ILI.

It should be noticed that it is important to export also blocking-point information to the ILI in order to ease future correct exportation of SOM's information to other wordnets, i.e. to prevent incorrect expansion of properties by inheritance. Inside a particular wordnet, when reaching a blocking point, a subsumption link can be considered as broken for ontological purposes -therefore, it will be assumed that the conceptual chain only proceeds upwards consistently to the SOM (not to the hypernym synsets), via the ILI-records.

This process can be applied iteratively looking for suspicious synsets in WN. In the first round we will check the list of 38 synsets which show incompatibility between hand-assigned TCO properties. In the second one we will check the set of WN top beginners which only bear information mapped via the TCO-SF table of equivalence. Third, we will check synsets showing incompatibility between information directly mapped via TCO and information mapped via the TCO-SF table of equivalence. Last, we will check the remaining cases of incompatibility between TCO manual and inherited information.

Being more precise, for each synset in any of the subsets we will proceed as follows:

1. Fixing the properties of those synsets having contradictory TCO properties: TCO assignments are fixed in the synset and its immediate relatives (mainly hypernym and hyponyms). All these synsets will be marked as "hand-checked". The result will be correct TCO information assigned to several synsets as in the following example where, originally, non-agentive and non-

intentional **00661612 stiffening_1** was inheriting all of the **00660718 process_1** properties:

00660718 process_1	
EWNT0	Dynamic;Agentive;Purpose
00661612 stiffening_1	
EWNT0	Dynamic;Cause

2. For those synsets having false WN subsumptions, we will introduce a blocking point between a pair of synsets. The result will be a list of blocking points, e.g.: between 00661612n and 00660718n.
3. We keep record of TCO–SF erroneous equivalences, since they will be useful in the future to detect more synsets with erroneous mappings. The result will be a list of suspicious TCO–SF equivalences, e.g.: [TCO:Agentive–SF:ACT]
4. To study TCO–SUMO equivalences in such synsets. As in the previous step, they can be useful in the future to detect more synsets with mistaken mappings. The result will be a list of incompatible TCO–SUMO concepts, e.g.: [TCO:3rdOrderEntity–SUMO:Physical]
5. To inspect as well WN Domain assignments. The result will be a list of doubtful WN Domain assignments, e.g. 00364173n#play_3:ENTERPRISE

Following an iterative and incremental approach, the inheritance will be re-calculated, the resulting data will be re-examined, and the eventual correct information will be again uploaded into the MCR thus overwriting the pre-existent one

Although such hand-checking is extremely complex and delicate, therefore slow and needing of sound semantic expertise to carry it on, we expect the task is affordable since critical conflicts seem to concentrate in a workable layer of synsets close to the higher part of the WN hierarchy.

5. Conclusions

In order to maintain compatibility among all the heterogeneous resources uploaded into the MCR, it is fundamental to have a robust and advanced ontological support. This paper studied the mapping of the main Sources of Ontological Meaning onto the MCR and, in particular, the current work with the Top Concept Ontology.

6. References

- Atserias, J., S. Climent, X. Farreres, G. Rigau, and H. Rodríguez, 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In *Proceeding of RANLP'97*. Bulgaria.
- Atserias, Jordi, Luís Villarejo, and German Rigau, 2003. Integrating and porting knowleges across languages. In *RANLP'03*. Borovets, Bulgaria.
- Atserias, Jordi, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen, 2004. The meaning multilingual central repository. In *Second International WordNet Conference-GWC 2004*. Brno, Czech Republic. ISBN 80-210-3302-9.
- 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, First International Conference on Language Resources & Evaluation*. Granada, Spain.
- Bentivogli, L., E. Pianta, and C. Girardi, 2002. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*. Mysore, India.
- Daudé, J., L. Padró, and G. Rigau, 2001. A complete wn1.5 to wn1.6 mapping. In *Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*. Pittsburg, PA, United States.
- Guarino, Nicola and Christopher A. Welty, 2000. A formal ontology of properties. In *Proceedings of ECAI'2000 Workshop on Knowledge Acquisition, Modeling and Management*.
- Magnini, B. and G. Cavagli, 2000. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*. Athens. Greece.
- Niles, I. and A. Pease, 2001. Towards a standard upper ontology. In *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Chris Welty and Barry Smith, eds.
- Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll, 2002. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING'2002 Workshop on A Roadmap for Computational Linguistics*. Taipei, Taiwan.
- Vossen, P. (ed.), 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers .