Complete and Consistent Annotation of WordNet using the Top Concept Ontology

Javier Álvez*, Jordi Atserias**, Jordi Carrera***, Salvador Climent***, Egoitz Laparra*, Antoni Oliver*** and German Rigau*

* Basque Country University. ** Universitat Pompeu Fabra. *** Open University of Catalonia.

scliment@uoc.edu, jibalgij@si.ehu.es, jordi.atserias@upf.edu, jcarrerav@uoc.edu, ego.laparra@gmail.com, aoliverg@uoc.edu, german.rigau@ehu.es

Abstract

This paper presents the complete and consistent ontological annotation of the nominal part of WordNet. The annotation has been carried out using the semantic features defined in the EuroWordNet Top Concept Ontology and made available to the NLP community. Up to now only an initial core set of 1,024 synsets, the so-called Base Concepts, was ontologized in such a way.

The work has been achieved by following a methodology based on an iterative and incremental expansion of the initial labeling through the hierarchy while setting inheritance blockage points. Since this labeling has been set on the EuroWordNet's Interlingual Index (ILI), it can be also used to populate any other wordnet linked to it through a simple porting process.

This feature-annotated WordNet is intended to be useful for a large number of semantic NLP tasks and for testing for the first time componential analysis on real environments. Moreover, the quantitative analysis of the work shows that more than 40% of the nominal part of WordNet is involved in structure errors or inadequacies.

1. Introduction

Componential semantics has a long tradition in Linguistics since the work of Hjelmslev in the thirties or Katz and Fodor (1963) among generativists. There is common agreement that this kind of lexical-semantic information can be extremely valuable for making complex linguistic decisions. Nevertheless, according to Simone (1990), componential analysis cannot be actually achieved due to three main reasons (being the first the most important): (1) the vocabulary of a language is too large, (2) each word needs several features for its semantics to be adequately represented and (3) semantic features should be organized in several levels.

Wordnets are large lexical resources freely-available and widely used by the NLP community. Currently, they serve a wide number of tasks involving some degree of semantic processing. In most of these tasks, wordnets are used to generalize or abstract a set of synsets to a subsuming one by following the WordNet hierarchy up. The main problem is finding the right level of generalization; that is, finding the concept which optimally subsumes a given set of concepts; but it could be the case that the class which would optimally capture the generalization is not lexical (i.e. a synset), but abstract –thus being better represented by non-lexical semantic features. It can also be the case that wordnet simply is not the kind of taxonomy required; this can be due to several reasons: incompleteness, incorrect structuring, or perhaps that WordNet's structure should be arranged differently for particular NLP tasks.

For many tasks, it seems that using a feature-annotated lexicon seems more appropriate than using the WordNet tree-structure, since (i) the WordNet hierarchy is not consistently structured (Guarino, 1998) and (ii) a feature-annotated lexicon allows to make predictions based on measures of similarity even for words that, being sparsely distributed in WordNet, can only be generalized by reaching common hypernyms in levels too high in the hierarchy. Besides, a multiple-feature design allows to naturally depict semantically complex concepts, such as the so-called dot-objects (Pustejovsky, 1995), –intrinsically polysemic words as for instance “letter”, since a letter is something that can both be destroyed and carry information (as in “I burnt your love letter”).

Our work provides a good solution to all these questions, since 65,989 noun concepts from WordNet 1.6 (WN1.6) (Fellbaum, 1998) corresponding to 116,364 noun lexemes (variants) have been consistently annotated with an average of 6.47 features per synset, being those features organized in a multilevel hierarchy. WN1.6 was adopted in EuroWordNet (EWN) as an Inter-Linguistic Index (ILI). Therefore, this resource might allow componential semantics to be tested and applied in real world situations probably for the first time, thus contributing to a wide number of NLP tasks involving semantic processing: Word Sense Disambiguation, Syntactic Parsing using selectional restrictions, Semantic Parsing or Reasoning.

Despite its wide scope, the work presented here is envisaged to be the first stage of an incremental and iterative process, as we do not assume that the current version of the EWN Top Concept Ontology (TCO) covers

the optimal set of features for the aforementioned tasks. Currently, a second phase has started within the framework of the KNOW Project[1] in which the first version of the enriched lexicon is being used to label a corpus in order to abstracting semantic properties of verbs. This will lead, presumably, to a reformulation of the TCO features and structure, probably following Vossen (2001).

This paper is organized as follows. After a brief summary of the EWN Top Concept Ontology (section §2), we present our methodology for annotating the nominal part of EWN (section §3). Section §4 summarizes a qualitative discussion, §5 a quantitative account and finally section §6 provides some concluding remarks.

2. The EuroWordNet Top Concept Ontology

The TCO (Alonge et al., 1998) was not primarily designed to be used as a repository of lexical semantic information but for clustering, comparing and exchanging concepts across languages in the EWN Project (Vossen, 1998). Nevertheless, most of its semantic features (e.g. *Human*, *Instrument*, etc.) have a long tradition in theoretical lexical semantics so they have been usually postulated as semantic components of meanings.

The TCO consists of 63 features and it is primarily organized, following Lyons (1977), in three disjoint types of entities:

- *1stOrderEntity* (physical things)
- *2ndOrderEntity* (events, states and properties)
- *3rdOrderEntity* (unobservable entities)

1st Order entities are further distinguished in terms of four ways of conceptualizing things (Pustejovsky, 1995):

- *Form*: as an amorphous substance or as an object with a fixed shape (*Substance* or *Object*)
- *Composition*: as a group of self-contained wholes or as a necessary part of a whole (*Group* or *Part*)
- *Origin*: the way in which an entity has come about (*Artifact* or *Natural*).
- *Function*: the typical activity or action is associated to the entity (*Comestible*, *Furniture*, *Instrument*, etc.)

Concepts can be classified in terms of any combination of these four categories. As such, the Top Concepts can be seen more as features than as ontological classes. Nevertheless, most of their subdivisions are disjoint categories: a concept cannot be both *Object* and *Substance*, or both *Natural* and *Artifact*. As explained below, feature disjunction plays an important role in our methodology.

2ndOrderEntitiy lexicalizes nouns and verbs (as well as adjectives and adverbs) denoting static or dynamic situations. All of the 2nd Order entities are classified using two different classification schemes:

- *SituationType*
- *SituationComponent*

SituationType represents a basic classification in terms of the Aktionsart properties of nouns and verbs, as described for instance in Vendler (1967). *SituationType* can be *Static* or *Dynamic*, further subdivided in *Property* and *Relation* on the one side, and *UnboundedEvent* and *BoundedEvent* on the other.

SituationComponent subtypes (e.g. *Location*, *Existence*, *Cause*) emerged empirically when selecting verbal and deverbal Base Concepts (BCs) in EWN. They resemble the cognitive components that play a role in the conceptual structure of events as in Talmy (1985).

Each *2ndOrderEntity* concept can be classified in terms of a mandatory but unique *SituationType* and any number of *SituationComponent* subtypes.

Last, *3rdOrderEntity* was not further subdivided.

The TCO has been redesigned twice, first by the EAGLES expert group (Sanfilippo et al., 1999) and then by Vossen (2001). EAGLES expanded the original ontology by adding 74 concepts while the latter made it more flexible, allowing, for instance, to cross-classify features between the three orders of entities.

3. Methodology

Our methodology for annotating the ILI with the TCO follows the strategy defined in Atserias et al. (2004) and it is based on the common assumption that hyponymy corresponds to feature set inclusion (Cruse, 2002) and in the observation that, since wordnets are taken to be crucially structured by hyponymy "(…) by augmenting important hierarchy nodes with basic semantic features, it is possible to create a rich semantic lexicon in a consistent and cost-effective way after inheriting these features through the hyponymy relations" (Sanfilippo et al., 1999).

Nevertheless, performing such operation is not straightforward, as (i) wordnets are not consistently structured by hyponymy (Guarino, 1998), and (ii) they allow multiple inheritance. Notwithstanding, these drawbacks, instead of hindering our work, have been situations we have taken advantage of.

As told above, within the EWN project, a limited set of lexical BCs was annotated with TCO features. Despite being largely general in meaning, this set did not cover all of the upper level nodes in the wordnets. Thus, the first step of our work consisted of annotating the gaps up the hierarchy, from the BCs to the unique beginners. This was made semiautomatically: synsets were assigned a TCO feature via a table of expected equivalence between TCO nodes and WN1.6 Semantic Files, e.g.:

[1] See the acknowledgements section.

04 noun.act => Agentive
05 noun.animal => Animal
06 noun.artifact => Artifact
07 noun.attribute => Property
08 noun.body => Object; Natural
09 noun.cognition => Mental

This made WN1.6 ready to be fully populated with at least one feature per synset. Nevertheless, in many cases, synsets got more than one feature, for one or more of the following reasons:

- They were BCs, so they had been manually annotated with several features
- In addition to their own manual annotation, they inherited features from one or more of their hypernyms
- They inherited features from different hypernyms, either located at different levels in a single track of hierarchy or by the effect of multiple inheritance

The work has been based on TCO feature incompatibilities. Throughout the process, co-occurrences of pairs of incompatible features in a synset have been automatically detected. The axiomatic incompatibilities are the following:

- 1stOrderEntity - 2ndOrderEntity
- 1stOrderEntity - 3rdOrderEntity
- 3rdOrderEntity - 2ndOrderEntity [except for SituationComponent]
- 3rdOrderEntity - Mental
- Object - Substance
- Gas - Liquid - Solid
- Artifact - Natural
- Animal - Creature - Human - Plant
- Dynamic - Static
- BoundedEvent - UnboundedEvent
- Property - Relation
- Physical - Mental
- Agentive - Phenomenal - Stimulating

The first round of feature expansion caused the following number of conflicts:

- 214 feature conflicts in 49 synsets caused by incompatible hand annotation
- 2,247 feature conflicts in 743 synsets caused by hand annotation incompatible with inherited features
- 225,447 feature conflicts in 26,166 synsets caused by incompatibility between inherited features

The first type of conflicts usually pointed to synsets causing ontological doubts to the annotators of the EWN project (e.g., “skin”, it is an object or a substance?). The third type usually reveals errors in WordNet structure – such as *ISA overloading* (Guarino, 1998) or other kinds of inconsistencies. The second type might be caused by either or both reasons.

Manual checking of feature incompatibilities led to (i) adding or deleting ontological features, and (ii) setting inheritance blockage points. A blockage point is an annotation in WN1.6 which breaks the ISA relation between two synsets, thus no information can be passed through it by inheritance.

When a case of feature incompatibility occurred, the synset involved, together with its structural surroundings (hypernyms, hyponyms), was analyzed in detail. If the problem was due to a WN1.6 subsumption error, the corresponding link was blocked and synsets below the blockage point were annotated with new TCO features.

Changes in the annotation were made and blockage points were set until all conflicts were resolved. Then, following an iterative and incremental approach, inheritance was being re-calculated and the resulting data was re-examined several times.

Despite the large number of conflicts to solve, the task ended up being feasible because working on the topmost origin of one conflict usually results in fixing many levels of hyponyms.

Regarding the completion of the work, the possibility that some areas in the WordNet hierarchy have remained unexamined cannot be completely excluded, although it should be noticed that more than 13,000 manual changes have been made and that, when removing links or features to fix errors, all hyponymy lines involved by the action have been checked again and newly annotated in order not to loss information.

The task has been carried out using application interfaces, which allowed accessing the synsets and their glosses in three languages at the same time: English, Spanish and Catalan. The following information was used in order to make decisions:

- Relational information regarding the synset under study and its neighbors; i.e. the WN1.6 structure
- The nature of the conflict (any of the three types of incompatibility aforementioned)
- Synsets' glosses as provided by EWN
- Glosses, descriptions and examples of the TCO features as provided in Alonge et al. (1998)
- Usual word-substitution tests that acknowledge hyponymy, as in Cruse (1986 pp. 88-92).

The task finished when finally a re-expansion of properties did not result in new conflicts. Then, two final steps were applied. First, as the TCO is itself a hierarchy, for every synset, its resulting annotation was expanded up-feature – e.g. when a synset bore the feature *Animal* it was also labeled *Living*, *Natural*, *Origin* and

1stOrderEntity. Second, the whole noun hierarchy was checked for consistency using formal Theorem Provers like Vampire (Riazanov and Voronkov, 2002) and E-prover (Schulz, 2002). This step resulted in a number of new conflicts which were re-analyzed and eventually fixed.

As stated in Atserias et al. (2005) this procedure can be seen as a shallow ontologization of WN1.6. That is, all WN1.6 *Tops* and all synsets under a blockage point are assigned one or more TCO nodes. This amounts to pruning WordNet's branches and linking them to an upper ontology. It constitutes a pragmatic solution to the difficulty of a complete wordnet's ontologization. In this sense, our work will probably be the second one to ontologize the whole WordNet (for nouns) after SUMO (Niles and Pease, 2003). The difference is that our annotation is (i) multiple thus more flexible (SUMO links each synset to only one label of the ontology), and (ii) more workable, since it uses a more intuitive and simple ontology (SUMO is a very large and complex one).

4. Qualitative discussion

Several examples showing our methodology at work can be seen in Atserias et al. (2005) and Álvez et al. (2008). A simple but very typical case is the following, in which the conflict comes from the combination of multiple inheritance and the incorrect use of hyponymy instead of meronymy in WN1.6[2]:

{Bandung_1[3] [*Artifact*+ *Natural*+]}]
---> {Java_1 [*Natural*+]}
---> {island_1 [*Natural*+]}
---> {city_1 [*Artifact*=]}

Clearly, Bandung is a city, but it *is not a* Java (though it is a *part of* Java). This case is revealed thanks to incompatibility between features *Natural* and *Artifact*. It is fixed by blocking the subsumption link between Bandung_1 and Java_1:

{Bandung_1 [*Artifact*+]}]
-x-> {Java_1 [*Natural*+]}
---> {island_1 [*Natural*+]}
---> {city_1 [*Artifact*=]}

At this point, it is worth noticing that in the original typology of ontological miscategorizations first established by Guarino (1998), four main sources of taxonomic inconsistencies were described: (a) overgeneralization, (b) reduction of sense, (c) confusion of senses and (d) suspect type-to-role relationships. During our research we set aside the last type, for it was regarded as carrying information which is both truthful and still useful in a lexical network. Moreover, we did uncover three new types of pervasive misconceptions: (i) extensional ambiguities, (ii) conflicts between *3rdOrderEntity* versus *Mental* 2nd Order entities and (iii) technical inconsistencies of different sorts.

Let us now see some examples of what our methodology has been able to uncover.

4.1. Overgeneralization

Overgeneralization takes place whenever one synset has as its hyponyms one or more synsets whose meaning is not entailed by the meaning of the so-called hypernym. "Accolade", for example, is a tangible entity intended to express approval. Its hyponyms, however, include both events ("citation", "mention") and concepts (social constructs, e.g. "academic degree"). Therefore, the synset has been overgeneralized: it comprises hyponyms which do not belong in its semantic class. Conflicts arise subsequently when semantic features associated with the hypernym pass on incorrectly to its hyponyms.

4.2. Reduction of sense

A reduction of sense occurs whenever a hypernym accounts for a part of the meaning of one of its hyponyms, while failing to express some other crucially relevant semantic component. Take the case of {counterfoil_1, stub_4}, which is a piece of paper which conveys specific information having to do with money transfers. As such, it should have been labeled as *1stOrderEntity* and, particularly, *Money Representation*. However, all of its hypernyms refer to “information, content”, and go up to {abstraction_1}, from which this whole taxonomic path derives. That is, no single ancestor accounts for the fact that a counterfoil is a piece a paper (all of them have to do with the information the counterfoil conveys).

4.3. Confusion of senses

A confusion of senses occurs when two conflicting lines of inheritance converge into one single synset. Thus, in the case of {ID_1}, EWN uses two semantically disjoint lines of inheritance (as a physical entity, a badge, and as the information regarding somebody’s identity) to express two different meaning components. As the current design of the TCO still doesn't allow such cross-order semantic conflation, this was solved blocking one of the lines of inheritance (that corresponding to the information meaning component).

4.4. Extensional ambiguities

With this term we refer to the fact that some entities seem to be objects in some sense but substances in yet another sense, which is why they cannot be properly labeled as

[2] Noun synsets are represented by one of their variants enclosed in curly brackets and TCO features by its name in italics, capitalized and enclosed in square brackets. Inherited features are marked ‘+’ while manually assigned features are marked ‘=’. Indentations stand for ISA relations. The symbol ‘x’ as in '-x->' means that the relation has been blocked.

[3] A city in the island of Java.

either. “Layer”, for example, refers just to an amount of matter in a homogeneous disposition over some surface. Therefore, a layer is made of some substance. On the other hand, however, it is not that substance: a layer lies over something else, so that it has at least one limiting boundary, which is one of the characteristic features based on which objects are distinguished from substances (i.e. having definite boundaries). At the same time, however, a layer as such is not an object proper, either: since it only lies relative to some other object, its limits are not intrinsic, but relative to this other object. One would never say, “there are two objects on the table, one vase and a layer of wood”. So, synsets of this kind were left unspecified for extension, whereas previous feature inheritance usually resulting in their being labeled either as objects or substances.

4.5. Conflicts between 3rd Order entities and *Mental* 2nd Order entities

When the time came to assign semantic features to synsets such as {unit_of_measurement_1}, it was not altogether clear whether these were to be better labeled as *3rdOrderEntity* (i.e. concepts) or as *Mental* 2nd Order entities (in this case, relations). Any unit of measurement expresses a relation between an entity and the reference used to measure that entity but, as far as the relation itself is concerned, is it abstract? That is, could a relation be so without being abstract? This would seem rather weird, for what is “specific” about a relation are the related entities, not the relation itself. In fact, relations are built on commonalities and commonalities cannot, by definition, be specific. Therefore, we arrived at a point at which a synset could only be a 2nd Order *Relation* if it was at the same time a 3rd Order entity, which simply exceeded the current theoretical framework. As for these cases, most were labeled by default as *3rdOrderEntity*.

5. Quantitative analysis

The whole process has provided a complete and consistent annotation of the nominal part of WN1.6, which consists of 65,989 nominal synsets, with 116,364 variants or senses.

All 227,908 initial incompatibilities were solved by manually adding or removing 13,613 TCO features and establishing 359 blockage points.

The final resource has 207,911 synset-feature pairs (an average of 2.66 TCO features per synset), expanded to 427,460 pairs when applying the inheritance of features consistently (an average of 6.48 TCO features per synset). In fact, the synset public_relations_1 has the maximum number of directly assigned features with nine, followed by ballyhoo_1 with eight features. Every TCO feature has been assigned on average 3,300 times. Ranging from Object which is the most widely assigned TCO feature (with 24,905 assignments) to Origin which was only assigned once.

The blockage points appear to be distributed along most WordNet levels. However, levels 6, 7 and 8 concentrate most of them (67% of the total, with 86, 87 and 67 blocking points respectively).

Interestingly, every blockage point affects a large number of synsets. Every blockage point subsumes an average of 120.16 synsets. In fact, 28,123 synsets have at least one blockage point in their hypernymy chain (i.e., from itself to WordNet's top). That is, following the TCO ontological incompatibilities, more than 40% of the nominal part of WordNet is involved in structural errors or inadequacies. However, it seems that most of the them are concentrated in small subparts of the WordNet hierarchy. 18,284 synsets inherit only one blocking point while only 9,839 synsets inherit more than one. On the other side, for instance, 62 synsets inherit 11 blocking points (most of them because of the structural problems of academic_degree_1).

6. Conclusions and further work

In this paper we have presented the full annotation of the nouns on the EuroWordNet (EWN) Interlingual Index (ILI) with those semantic features constituting the EWN Top Concept Ontology (TCO). This goal has been achieved by following a methodology based on an iterative and incremental expansion of the initial labeling through the hierarchy while setting inheritance blockage points. Since this labeling has been set on the ILI and it is defined as language-independent, it can be also used to populate any other wordnet linked to it through a simple porting process.

Moreover, the work shows that more than 40% of the nominal part of WordNet is involved in WordNet's structure errors or inadequacies. This fact poses significant challenges for relying in WordNet's hierarchy as the unique resource for abstracting semantic classes for NLP.

The TCO-annotated WordNet which we have presented here is intended to be useful for a large number of semantic NLP tasks and for testing for the first time componential analysis on real environments. Bearing these goals in mind, further work will focus on the annotation of a corpus oriented to the acquisition of selectional preferences. These selectional preferences will be compared to state-of-the-art synset-generalization semantic preferences. As a result, we expect a qualitative evaluation of the resource. As a side effect, we expect to gain some knowledge for designing an enhanced version of the TCO more suitable for semantically-based NLP.

The resource developed by this work can be downloaded from http://lpg.uoc.edu/ > *Results* > *Software and resources* and it can be browsed at http://adimen.si.ehu.es/cgi-bin/wei5/public/wei.consult.perl

7. Acknowledgements

This research has been partially funded by the following projects: *KNOW. Developing large-scale multilingual technologies for language understanding*. Ministerio de Educación y Ciencia. TIN2006-15049-C03-02; and KYOTO: Knowledge Yielding Ontologies for Transition-based Organization. EC. ICT-2007-211423.

8. References

Alonge A., Bertagna F., Bloksma L., Climent S., Peters W., Rodríguez H., Roventini A. and Vossen P. (1998) The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In Piek Vossen (ed.) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Álvez J., Atserias J., Carrera J., Climent S., Oliver A. and Rigau G.. (2008) Consistent annotation of EuroWordNet with the Top Concept Ontology. In *Proceedings of The 4th Global Wordnet Association Conference*. Szeged. Hungary

Atserias J., Climent S. and Rigau G. (2004) Towards the MEANING Top Ontology: Sources of Ontological Meaning. *Proceedings of the LREC 2004*. Lisbon

Atserias J., Climent S., Moré J. and Rigau G. (2005) A Proposal for a Shallow Ontologization Of Wordnet. *Procesamiento del Lenguaje Natural* 35 pp. 161-167.

Cruse D.A. (1986) *Lexical Semantics*. Cambridge University Press. NY

Cruse D.A. (2002) Hyponymy and Its Varieties. In: R. Green, C.A. Bean, & S. H. Myaeng (eds.) *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer Verlag.

Fellbaum C. (ed.) (1998) *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge MA.

Guarino N. (1998) Some Ontological Principles for Designing Upper Level Lexical Resources. Proceedings of the 1st International Conference on Language Resources and Evaluation. Granada

Katz J.J. and Fodor J.A. (1963) *The Structure of a Semantic Theory*. Language, 39: 170-210

Lyons J. (1977) *Semantics*. Cambridge University Press. Cambridge, UK

Niles I. and Pease A. (2003) Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Model Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering*. Las Vegas, USA.

Pustejovsky J. (1995) *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

Riazanov A. and Voronkov A. 2002 The Design and implementation of Vampire. *Journal of AI Communications*. 15(2). IOS Press.

Sanfilippo A., Calzolari N., Ananiadou S., et al. (1999) *Preliminary Recommendations on Lexical Semantic Encoding*. Final Report. EAGLES LE3-4244

Schulz, S. 2002. A Brainiac Theorem Prover. *Journal of AI Communications* 15(2/3): IOS Press.

Simone R. (1990) *Fondamenti di Linguistica.* Laterza & Figli. Bari-Roma. Trad. Esp.: Ariel, 1993

Talmy L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In Shopen 1985 ed. *Language typology and syntactic description: Grammatical categories and the lexicon*. Vol. 3. 57–149. Cambridge University Press.

Vendler Z., (1967): *Linguistics in philosophy*. Ithaca, N.Y.: Cornell University Press.

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

Vossen P. (2001) Tuning Document-Based Hierarchies with Generative Principles. In *Proceedings of GL'2001 First International Workshop on Generative Approaches to the Lexicon*. Geneva.