The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet

Aldo Gangemi[1], Roberto Navigli[2], Paola Velardi[2]

[1]Laboratory for Applied Ontology, ISTC-CNR,
viale Marx 15, 00137 Roma, Italy
gangemi@ip.rm.cnr.it

[2] Dipartimento di Informatica, University of Roma "La Sapienza"
via Salaria 113, 00198 Roma, Italy
{navigli,velardi}@dsi.uniroma1.it

Abstract. In this paper we present a progress report of the OntoWordNet project, a research program aimed at achieving a formal specification of WordNet. Within this program, we developed a hybrid bottom-up top-down methodology to automatically extract association relations from WordNet, and to interpret those associations in terms of a set of conceptual relations, formally defined in the DOLCE foundational ontology. Preliminary results provide us with the conviction that a research program aiming to obtain a consistent, modularized, and axiomatized ontology from WordNet can be completed in acceptable time with the support of semi-automatic techniques.

1. Introduction

The number of applications where WordNet (WN) is being used as an ontology rather than as a mere lexical resource seems to be ever growing. Indeed, WordNet contains a good coverage of both the lexical and conceptual palettes of the English language. However, WordNet is serviceable as an ontology (in the sense of a *theory* expressed in some *logical language*) if some of its lexical links are interpreted according to a formal semantics that tells us something about the way we use a lexical item in some context for some purpose. In other words, we need a *formal specification of the conceptualizations that are expressed by means of WordNet's synsets*[1]. A formal specification requires a clear semantics for the primitives used to export WordNet

---

[1] Concept names in WordNet are called *synsets*, since the naming policy for a concept is a set of synonym words, e.g. for sense 1 of car: { car, auto, automobile, machine, motorcar }.

information into an ontology, and a methodology that explains how WordNet information can be bootstrapped, mapped, refined, and modularized.

The formal specification of WordNet is the objective of the so-called OntoWordNet research program, started two years ago at the ISTC-CNR, and now being extended with other partners, since collaborations have been established with the universities of Princeton, Berlin and Roma. The program is detailed in section 2, where we outline the main objectives and current achievements.

In this paper we describe a joint ongoing work of ISTC-CNR and the University of Roma that has produced a methodology and some preliminary results for adding *axioms* (DAML+OIL "restrictions") to the concepts derived from WordNet synsets. The methodology is hybrid because it employs both top-down techniques and tools from formal ontology, and bottom-up techniques from computational linguistics and machine learning. Section 3 presents a detailed description of the methodology.

The preliminary results, presented in section 4, seem very encouraging, and provide us with the conviction that a research program aiming to obtain a consistent, modularized, and axiomatized ontology from WordNet can be completed in acceptable time with the support of semi-automatic techniques.

2. Objectives of the OntoWordNet research program.

The OntoWordNet project aims at producing a formal specification of WordNet as an axiomatic theory (an *ontology*). To this end, WordNet is reorganized and enriched in order to adhere to the following commitments:

- *Logical commitment*. WordNet synsets are transformed into logical types, with a formal semantics for lexical relations. The WordNet lexicon is also separated from the logical namespace.
- *Ontological commitment*. WordNet is transformed into a general-purpose ontology library, with explicit categorial criteria, based on formal ontological distinctions (Gangemi et al. 2001). For example, the distinctions enable a clear separation between (kinds of) concept-synsets, relation-synsets, meta-property-synsets, and enable the instantiation of individual-synsets. Moreover, such formal ontological principles facilitate the axiomatic enrichment of the ontology library.
- *Contextual commitment*. WordNet is modularized according to knowledge-oriented domains of interest. The modules constitute a partial order.
- *Semiotic commitment*. WordNet lexicon is linked to text-oriented (or speech act-oriented) domains of interest, with lexical items ordered by preference, frequency, combinatorial relevance, etc.

In (Gangemi et al 2001,2002) these commitments are discussed in detail and working hypotheses and first achievements are presented. This paper is concerned mainly with the ontological commitments.

WordNet's *ontological commitments* are more demanding to be explicitated, but many results are already available. For example, an incremental methodology has been adopted, in order to revise or to reorganize WordNet synset taxonomies and relations (see also paragraph 3.2.1). Substantial work has been done on the refinement of the *hyponym/hyperonym* relations, which have being investigated since several years. The hyperonymy relation in WN is basically interpreted as *formal subsumption*, although hyperonymy for concepts referring to individuals (geographical names, characters, some techniques, etc.) is interpreted as *instantiation*. This will be referred as *assumption A1* ("hyperonymy as synset subsumption").

WordNet *synonymy* is a relation between words, not concepts, therefore we should assume that the synonymy relation (*synsets* in WordNet) is an *equivalence class* of words (or phrases), sharing the same *meaning* within an ontology. Consequently, two words are synonyms when their intended meaning in WordNet is the same. This will be referred to as *assumption A2* ("synset as meaning equivalence class").

However, we have no formal definition of words in WordNet that allows us to create equivalence classes (synsets) analytically (i.e., to state *semantic equivalences*). Instead, we have pre-formal synsets that have been validated by lexicographers with an intuition that *could* be formalized as semantic equivalence. Part of this intuition is conveyed by textual definitions (called *glosses*). This will be referred as *assumption A3* ("glosses as axiomatizations"). We are trying to formalize such intuition.

A related assumption that we make is that words in glosses are used in a way consistent to the WordNet synsets. This will be referred as *assumption A4* ("glosses are synset-consistent"). A4 lets us assume also that the informal theory underlying synsets, hyperonymy relations, and glosses, can be formalized against a finite signature (the set of WN synsets), and a set of axioms derived from the associations (*A-links*) between any synset *S* and the synsets that can be associated to the words used in the gloss of *S*. This is dependent on A3 and A4, and may be referred as *assumption A5* ("*A-links* as conceptual relations").

3. Semi-automatic axiomatization of WordNet

The task of axiomatizing WordNet, starting from assumptions A1-A5 outlined in the previous section, requires that the informal definition in a synset gloss be transformed in a logical form. To this end, first, words in a gloss must be disambiguated, i.e. replaced by their appropriate synsets. This first step provides us with pairs of generic semantic associations (A-links) between a synset and the synsets of its gloss. Secondly, A-links must be interpreted in terms of more precise, formally defined semantic relations. The inventory of semantic relations is selected or specialized from the foundational ontology DOLCE, as detailed later, since in WordNet only a limited set of relations are used, that are partly ontological, partly lexical in nature. For example, part_of (*meronymy* in WordNet) and kind_of (*hyponymy* in WordNet) are typical semantic relations, while *antonymy* (e.g. *liberal* and *conservative*) and *pertonymy* (e.g *slow* and *slowly*) are lexical relations. Furthermore, WordNet relations are not axiomatized, nor are they used in a fully consistent way.

To summarize, the objective of the method described in this section is to:

- automatically extract a number of semantic relations implicitly encoded in WordNet, i.e. the relations holding between a synset and the synsets in its gloss.
- (semi)-automatically interpret and axiomatize these relations.

For example, sense 1 of *driver* has the following gloss "the operator of a motor vehicle". The appropriate sense of *operator* is #2: *operator, manipulator* ("an agent

that operates some apparatus or machine"), while motor vehicle is monosemous: *motor vehicle*, *automotive vehicle* ("a self-propelled wheeled vehicle that does not run on rails").
After automatic sense disambiguation, we (hopefully) learn that there exists an A-link between driver#1 and operator#2, and between *driver#1* and *motor vehicle#1*. Subsequently, given a set of axiomatized semantic relations in DOLCE, we must select the relation that best fits the semantic restrictions on the relation universes (domain and co-domain, or range). For example, given an A-link between *driver#1* and *motor vehicle#1*, the best fitting relation is *agentive-co-participation* (Figure 1), whose definition is:

$$\mathrm{AG\_CO\_PCP}(x,y) =_{\mathrm{df}} \mathrm{CO\_PCP}(x,y) \wedge \mathrm{Agentive\_Physical\_Object}(x) \wedge \\ \wedge \mathrm{Non\_Agentive\_Functional\_Object}(y)$$

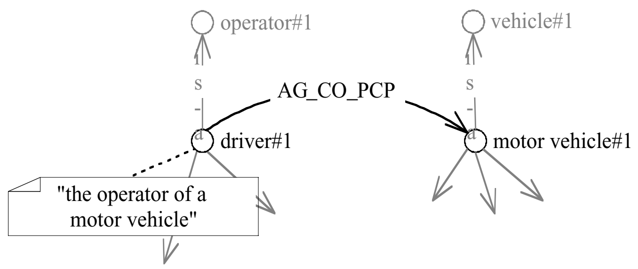

Figure 1. An example of semantic relation.

The definition says that *agentive co-participation* is a relation of mutual participation (participation of two objects in the same event), with the domain restricted to "Agentive_Physical_Object" and the range restricted to "Non_Agentive_Functional_Object".
Domain and range in a conceptual relation definition are established in terms of the DOLCE ontology. Consequently, another necessary step of our method is to re-link at least some of the higher level nodes in WordNet with the DOLCE upper ontology.
In the following sub-sections we detail the procedures for gloss disambiguation, WordNet re-linking, and selection of conceptual relations.

3.1 Bottom-up learning of association links.

The first step is a bottom-up procedure that analyses the NL definitions (glosses) in WordNet and creates the A-links.
For each gloss (i.e., linguistic concept definition), we perform the following automatic tasks:
a) POS-tagging of glosses (using the ARIOSTO NL processor) and extraction of *relevant* words;
b) Disambiguation of glosses by the algorithm described hereafter;
c) Creation of explicit "association" links (A-links) from synsets found in glosses to synsets to which glosses belong.

3.1.1 Description of the gloss disambiguation algorithm

We developed a greedy algorithm for gloss disambiguation that relies on a set of heuristic rules and is based on multiple, incremental iterations. The algorithm takes as input the synset *S* whose gloss *G* we want to disambiguate.

Two sets are used, *P* and *D*. *D* is a set of disambiguated synsets, initially including only the synset S. *P* is a set of terms to be disambiguated, initially containing all the terms from gloss *G* and from the glosses {G'} of the direct hyperonyms of *S*. As clarified later, adding {G'} provides a richer context for semantic disambiguation. The term list is obtained using our NL processor to lemmatize words, and then removing irrelevant words. We use standard information retrieval techniques (e.g stop words) to identify irrelevant terms.

When, at each iteration of the algorithm, we disambiguate some of the terms in *P*, we remove them from *P* and add their interpretation (i.e. synsets) to the set *D*. Thus, at each step, we can distinguish between *pending* and *disambiguated* terms (respectively the sets *P* and *D*). Notice again that *P* is a set of terms, while *D* contains synsets.

<u>Step a) Find monosemous terms</u>

The first step of the algorithm is to remove monosemous terms from *P* (those with a unique synset) and include their unique interpretation in the set *D*.

<u>Step b) Disambiguate polysemous terms</u>

Then, the core iterative section of the algorithm starts. The objective is to detect *semantic relations* between some of the synsets in *D* and some of the synsets associated to the terms in *P*. Let *S'* be a synset in *D* (an already chosen interpretation of term *t'*) and *S"* one of the synsets of a polysemous term $t'' \in P$ (i.e., *t"* is still ambiguous). If a semantic relation is found between *S'* and *S"*, then *S"* is added to *D* and *t"* is removed from *P*.

To detect semantic relations between *S'* and *S"*, we apply a set of heuristics grouped in two classes, *Path* and *Context*, described in what follows. Some of these heuristics have been suggested in (Milhalcea, 2001),

<u>*Path heuristics*</u>

The heuristics in class Path seek for *semantic patterns* between the node *S'* and the node *S"* in the WordNet semantic network. A *pattern* is a chain of nodes (synsets) and arcs (directed semantic relations), where *S'* and *S"* are at the extremes.

Formally, we define $S' \xrightarrow{R}{}^{n} S''$ as $S' \xrightarrow{R} S_1 \xrightarrow{R} \ldots \xrightarrow{R} S_n \equiv S''$, that is a chain of *n* instances of the relation *R*. We also define $\xrightarrow{R_1,R_2}$ as $\xrightarrow{R_1} \cup \xrightarrow{R_2}$.

The symbols: $\xrightarrow{@}, \xrightarrow{\sim}, \xrightarrow{\#}, \xrightarrow{\%}, \xrightarrow{\&}$ respectively represent the following semantic relations coded in WordNet 1.6: *hyperonimy* (kind_of), *hyponymy* (has kind) *meronymy* (part_of), *holonymy* (has_part), and *similarity*. Similarity is a generic relation including near synonyms, adjectival clusters and antonyms. Finally, the *gloss* relation $S \xrightarrow{gloss} T$ indicates that the gloss of *S* includes a term *t*, and *T* is one of the synsets of *t*.

The following is an example of heuristics that we use to identify semantic paths ($S' \in D, S'' \in Synsets(t''), t'' \in P$):

Hyperonymy/Meronymy path: if $S' \xrightarrow{@,\#}{}^{n} S''$ choose S'' as the right sense of t'' (e.g., $archipelago\#1 \xrightarrow{\#} island\#1$);

<u>Context heuristics</u>

The Context heuristics use several available computational linguistic resources to detect co-occurrence patterns in sentences and contextual clues to determine a semantic proximity between *S'* and *S"*. The following heuristics are defined:

<u>Semantic co-occurrences</u>: word pairs may help in the disambiguation task if they always co-occur with the same senses within a tagged corpus. We use three resources in order to look for co-occurrences, namely:

- the *SemCor corpus*, a corpus where each word in a sentence is assigned a sense selected from the WordNet sense inventory for that word
- the *LDC corpus*, a corpus where each document is a collection of sentences having a certain word in common. The corpus provides a sense tag for each occurrence of the word within the document.
- *gloss examples:* in WordNet, besides glosses, examples are sometimes provided containing synsets rather than words. From these examples, as for the LDC Corpus, a co-occurrence information can be extracted.
- As we said above, only the SemCor corpus provides a sense for each word in a pair of adjacent words occurring in the corpus, while LDC and gloss examples provide the right sense only for one of the terms.

In either case, we can use this information to choose the synset *S"* as the interpretation of *t"* if the pair *t' t"* occurs in the gloss and there is an agreement among (at least two of) the three resources about the disambiguation of the pair *t' t"*. For example:

[…] Multnomah County may be short of general assistance money in its budget to handle an unusually high ***summer#1 month#1****'s need […].*

Later#1, Eckenfelder increased#2 the efficiency#1 of treatment#1 to between 75 and 85 percent#1 in the ***summer#1 months#1****.*

are sentences respectively from the LDC Corpus and SemCor. Since there is a full agreement between the resources, one can easily disambiguate *summer* and *months* in the gloss of *summer_camp#1*: "*a site where care and activities are provided for children during the* ***summer months***".

<u>Common domain labels</u>: Domain labels are the result of a semiautomatic methodology described in (Magnini and Cavaglia, 2000) for assigning domain labels (e.g. *tourism, zoology, sport*..) to WordNet synsets[2]. This information can be exploited to disambiguate those terms with the same domain labels of the start synset *S*.

<u>Step c) Update *D* and *P*</u>

During each iteration, the algorithm applies all the available heuristics in the attempt of disambiguating some of the terms in *P*, using all the available synsets in *D*. The heuristics are applied in a fixed order reflecting their importance, that has been

[2] Domain labels have been kindly made available by the IRST to our institution for research purposes.

experimentally determined. For example, Context heuristics are applied after Path heuristics 1-5. At the end of each iterative step, new synsets are added to *D*, and the correspondent terms are deleted from *P*. The next iteration makes use of these new synsets in order to possibly disambiguate other terms in *P*. Eventually, either *P* becomes empty, or no new semantic relations can be found.

When the algorithm terminates, $D \setminus \{ S \}$ can be considered a first approximation of a *semantic definition of* S. For mere gloss disambiguation purposes, the tagged terms in the hyperonyms' gloss are discarded, so that the resulting set (*GlossSynsets*) now contains only interpretations of terms extracted from the gloss of *S*. At this stage, we can only say that there is a semantic relation (A-link) between *S* and each of the synsets in *GlossSynsets* .

A second, more precise approximation of a sound ontological definition for *S* is obtained by determining the nature of the A-links connecting *S* with each concept in $D \setminus \{ S \}$. This is an ongoing task and is discussed in Section 4.

3.1.2 A running example

In the following, we present a sample execution of the algorithm on sense 1 of *retrospective*. Its gloss defines the concept as "*an exhibition of a representative selection of an artist's life work*", while its hyperonym, *art exhibition#1*, is defined as "*an exhibition of art objects (paintings or statues)*". Initially we have:

D = { *retrospective#1* }

P = { *work, object, exhibition, life, statue, artist, selection, representative, painting, art* }

The application of the monosemy step a) gives the following result:

D = { *retrospective#1, statue#1, artist#1* }

P = { *work, object, exhibition, life, selection, representative, painting, art* }

because *statue* and *artist* are monosemous terms in WordNet. During the first iteration, the algorithm finds three matching paths:

retrospective#1 $\overset{@}{\longrightarrow}^{2}$ *exhibition#2*, *statue#1* $\overset{@}{\longrightarrow}^{3}$ *art#1* and *statue#1* $\overset{@}{\longrightarrow}^{6}$ *object#1*

this leads to:

D = { *retrospective#1, statue#1, artist#1, exhibition#2, object#1, art#1* }

P = { *work, life, selection, representative, painting* }

During the second iteration, an hyponymy/holonymy path is found:

art#1 $\overset{\sim}{\longrightarrow}^{2}$ *painting#1* (painting is a kind of art)

D = { *retrospective#1, statue#1, artist#1, exhibition#2, object#1, art#1, painting#1* }

P = { *work, life, selection, representative* }

Since no new paths are found, the third iteration makes use of the LDC Corpus to find the co-occurrence "*artist life*", with sense 12 of *life* (*biography*, *life history*):

D = { *retrospective#1, statue#1, artist#1, exhibition#2, object#1, art#1, painting#1, life#12* }

P = { *work, selection, representative* }

Notice that, during an iteration, the context heuristics are used only if the path heuristics fail.

The algorithm stops because there are no additional matches. The chosen senses concerning terms contained in the hyperonym's gloss were of help during disambiguation, but are now discarded. Thus we have:
GlossSynsets(retrospective#1) = { *artist#1, exhibition#2, life#12* }
Figure 2 shows in dark gray the A-links between *retrospective#1* and the synset of its glosses, while in the light gray area are shown the synsets of the hyperonyms.

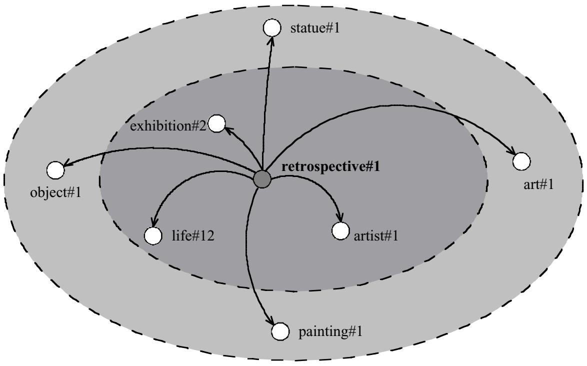

Figure 2. A first approximation of a semantic definition of *retrospective#1*.

3.2 Top-down learning: formal ontologies and WordNet "sweetening"

In the top-down phase, the A-links extracted in the bottom-up phase are refined. A-links provide just a *clue* of relatedness between a synset and another synset extracted from the gloss analysis, but this relatedness must be explicit, in order to understand if it is a hyperonymy relation, or some other conceptual relation (e.g. part, participation, location, etc.).

First of all, we need a shared set of conceptual relations to be considered as candidates for A-links explicitation, otherwise the result is not easily reusable. Secondly, these relations must be formally defined. In fact, as already pointed out at the beginning of section 3, not only are A-links vague, but they also lack a formal semantics: for example, if we decide (which seems reasonable) to represent associations as binary relations –like DAML+OIL "properties"– is an association symmetric? Does it hold for every instance, or only for some of the instances of the classes derived from the associated synsets? Is it just a constraint on the applicability of a relation to that pair of classes? Is the relation set a flat list, or there is a taxonomic ordering?

To answer such questions, the shared set of relations should be defined in a logical language using a formal semantics.

Since WordNet is a general-purpose resource, the formal shared set of relations should also be general enough, based on *domain-independent* principles, but still flexible, in order to be easily maintained and negotiated.

3.2.1 The DOLCE descriptive ontology

A proposal in this direction is provided by the WonderWeb[3] project Foundational Ontology Library (WFOL), which will contain a library including both compatible and alternative modules including domain-independent concepts and relations. A recently defined module that accomplishes the abovementioned requirements is DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering).

DOLCE is expressed in an S5 modal logic (Masolo et al. 2002), and has counterparts in computational logics, such as KIF, LOOM, RACER, DAML+OIL, and OWL. The non-KIF counterparts implement a reduced axiomatization of DOLCE, called DOLCE-Lite. DOLCE-Lite has been extended with some *generic plugins* for representing information, communication, plans, ordinary places, and with some *domain plugins* for representing e.g. legal, tourism, biomedical notions. The combination of DOLCE-Lite and the existing plugins is called DOLCE-Lite+. The current version 3.6 of DOLCE-Lite+ without domain plugins contains more than 300 concepts and about 150 relations (see http://www.loa-cnr.it for DOLCE versioning).

DOLCE assumes that its categories (top classes) constitute an *extensionally* closed set on any possible *particular* entity, i.e., entities that cannot be further instantiated within the assumptions of the theory (cf. Masolo et al. 2002, Gangemi et al. 2001). Of course, DOLCE does not assume an *intensionally* closed set, thus allowing for alternative ontologies to co-exist. Such assumptions will be referred to as *A6_D* ("extensional total coverage of DOLCE"). Consequently, we also assume that WN globally can be tentatively considered a (extensional) subset of DOLCE, after its formalization. Since we cannot practically obtain a complete formalization of WN, we will be content with incrementally approximating it.

A trivial formalization of WN might consist in declaring formal subsumptions for all *unique beginners* (top level synsets) under DOLCE categories, but this proved to be impossible, since the intension of unique beginners, once they are formalized as classes, is not consistent with the intension of DOLCE categories. Then we started (Gangemi et al. 2002) deepening our analysis of WN synsets, in order to find synsets that could be subsumed by a DOLCE category without being inconsistent.

In our previous OntoWordNet work, WordNet 1.6 has been analyzed, and 809 synsets have been relinked to DOLCE-Lite+ in order to harmonize ("sweeten") WN taxonomies with DOLCE-Lite+. A working hypothesis (*A7_D*) has been that the taxonomy branches of the relinked synsets are ontologically consistent with the DOLCE-Lite+ concepts, to which the relinking is targeted. After some additional work, the current linking of 809 synsets seems acceptable, but it needs refinement, since some subsumptions are debatable, and it must be considered that some extensions of DOLCE-Lite+ are still unstable.

3.2.2 Disambiguation of association links

Assumptions A4 and A5 (section 2), together with A6_D (in previous sub-section), make it possible to exploit the axiomatized relations in DOLCE-Lite+. Such relations

[3] http://wonderweb.semanticweb.org

are formally characterized by means of *ground axioms* (e.g. symmetry, transitivity, etc.), *argument restrictions* (qualification of their *universe*), *existential axioms, links to other primitives, theorems,* etc. (Masolo et al. 2002).

By looking at the A-links, a human expert can easily decide which relation from DOLCE-Lite+ is applicable in order to disambiguate the A-link, for example, from:

1. A-link(*car#1*, *engine#1*)

we may be able to infer that cars have engines as components:

$\forall x.\ \mathrm{Car}(x) \rightarrow \exists y.\ \mathrm{Engine}(y) \wedge \mathrm{Component}(x,y)$

or that from

2. A-link(*art_exhibition#1*, *painting#1*)

we can infer that exhibitions as collections have paintings as members:

$\forall \mathrm{x}.\ \mathrm{Art\_exhibition}(x) \rightarrow \exists y.\ \mathrm{Painting}(y) \wedge \mathrm{Member}(x,y)$

But this is an intellectual technique. We are instead interested, at least for the sake of bootstrapping a preliminary axiomatization of synsets, in a (semi) *automatic classification technique*. From this viewpoint, the only available structure is represented by the concepts to which the A-links apply. Such synsets can be assumed as the *argument restrictions* of a conceptual relation implicit in the association. For example, given (A-link(S_1, S_2)), where S_1, S_2 are synsets, we can introduce the argument restrictions for a conceptual relation $R^{a\text{-}link}_i(x,y) \rightarrow S_1(x) \wedge S_2(y)$. Then, from A5 and its depend-on assumptions, we have a good heuristics for concluding that $S_1(x) \rightarrow \exists y.\ R^{a\text{-}link}_i(x,y) \wedge S_2(y)$. In other words, we formalize the association existing between a synset and another synset used in its gloss. This leaves us with the question of what is the intension of $R^{a\text{-}link}_i(x,y)$, beyond its argument restrictions: e.g. what does it mean to be a relation between *art exhibitions* and *paintings*? And are we allowed to use this heuristics to conclude that art exhibitions are related to at least one painting?

Assuming A6_D, we can claim that some $R_i(x,y)$ from DOLCE-Lite+ subsumes $R^{a\text{-}link}_i(x,y)$. Since the relations from DOLCE-Lite+ have a total extensional coverage on any domain, we can expect that at least one relation from DOLCE has a universe subsuming that of $R^{a\text{-}link}_i(x,y)$. For example: Member(x,y) from DOLCE-Lite+ can subsume $R^{a\text{-}link}_i(x,y)$ when Art_exhibition(x) and Painting(y), since the domain and range of “Member” subsume “Art_exhibition” and “Painting” respectively.

These subsumptions are easily derivable by using a description-logic classifier (e.g. LOOM, MacGregor, 1993, or RACER, Moeller, 2001) that computes the applicable relations from DOLCE-Lite+ to the training set of A-links.

For example, an “ABox” query like the following can do the job in LOOM:

<u>ABox-1</u>

```
(retrieve (?x ?R ?y)  (and (get-role-types ?x ?R ?y) (min-cardinality ?x ?R 1) (A-link ?x ?y)))
```

i.e., provided that A-links have been defined on DOLCE-Lite+ classes (i.e. that WN synsets ?x ?y are subsumed by DOLCE-Lite+ classes), the relation “get-role-types” will hold for all the relations in DOLCE-Lite+ that are applicable to those classes, with a cardinality≥1. For example, given the previous example (2) of A-link, the classifier uses some of the DOLCE-Lite+ axioms to suggest the right conceptual relation. In fact, the WordNet synset *art_exhibition#1* is a (indirect) sub-class of the DOLCE class “unitary collection”, a category for which the following axiom holds:

$\forall x.\ \mathrm{Unitary\_Collection}(x) \rightarrow \exists y.\ \mathrm{Physical\_Object}(y) \wedge \mathrm{Member}(x,y)$

Furthermore, since *painting#1* is a (indirect) sub-class of “physical object”, and the axiom holds with a cardinality≥1, the correct relation and axiom are guessed.

In other cases, ABox-1 does not retrieve an existing appropriate relation. For example, given:

3. A-link(*boat#1*,*travel#1*)

with *boat#1* subsumed by Physical_Object and *travel#1* subsumed by Situation in DOLCE+WordNet, and the relation "Setting" holding between physical objects and situations, we have no axiom like the following in DOLCE-Lite+:

$* \ \forall x. \text{Physical\_Object}(x) \rightarrow \exists y. \text{Situation}(y) \wedge \text{Setting}(x,y)$

then the relation $R^{a\text{-}link}_{i}$ formalizing the A-link between *boat* and *travel* cannot be automatically classified and proposed as subsumed by the relation "Setting" in DOLCE-Lite+. In other words, in general *it is not true* that "for any physical object there is at least a situation as its possible "*setting*": we can figure out physical objects in general, without setting them anywhere, at least within the scope of a computational ontology.

The above examples show that axioms representing generally acceptable intuitions in a foundational ontology may prove inadequate in a given application domain, where certain axiomatizations need an ad-hoc refinement.

The solution presented here exploits a partition of argument restrictions for the gloss axiomatization task. For this solution, we need a partition Π of relation universes, according to the 25 valid pairs of argument restrictions that can be generated out of the five top categories of DOLCE-Lite+ (*Object*, *Event*, *Quality*, *Region*, and *Situation*), which on their turn constitute a partition on the domain of entities for DOLCE-Lite+. This enables us to assign one of the 25 relations to the A-link whose members are subsumed by the domain and range of that relation. For example, from: $(\text{Boat}(x) \rightarrow \text{Object}(x))$, and $(\text{Travel}(y) \rightarrow \text{Situation}(y))$, we can infer that some $R_{<\text{Object,Situation}>}$ holds for the pair $\{x,y\}$.

However, in DOLCE-Lite+, existing relations are based on primitives adapted from the literature, covering some basic intuitions and that are axiomatized accordingly. Therefore, the current set of DOLCE-Lite+ relations $\Pi\delta$ is not isomorphic with Π, while the same extensional coverage is supported. For example, the DOLCE-Lite+ relation "part" corresponds to a *subset* of the union of *all* the argument pairs in Π that include only the same category (e.g., <Event, Event>). $\Pi\delta$ is inadequate to perform an automatic learning of conceptual relations, because we cannot distinguish between "part" and other relations with the same universe (e.g. "connection"). Similarly, we cannot distinguish between different pairs of argument restrictions *within* the "part" universe (e.g. <Event, Event> vs. <Object, Object>).

The choice of axioms in DOLCE-Lite+ is motivated by the necessity of *grounding* the primitive relations in human intuition, for example in so-called *cognitive schemata* that are established during the first steps of an organism's life by interacting with its environment and using its specific abilities to react to the stimuli, constraints, and affordances provided by the context (Johnson 1987). In fact, without that grounding, the meaning of relations cannot be figured out at all (even though they are correct from a logical viewpoint).

There is also another reason for the inadequacy of $\Pi\delta$. A conceptual relation in DOLCE-Lite+ can be "mediated", e.g. defined through a *composition* (called also *chaining*, or *joining* in the database domain). For example, two objects can be related

because they participate in a same event, for example, *engine* and *driver* can "co-participate" because they both *participate in driving*.
In brief: we cannot use $\Pi\delta$, since it does not discriminate at the necessary level of detail, and because it is not a partition at all, if we take into account mediated relations. On the other hand, we cannot use Π, because it is cognitively inadequate.
Consequently, we have evolved a special partition $\Pi\delta+$ that keeps both worlds: a real partition, and cognitive adequacy. $\Pi\delta+$ denotes a partition with a precise mapping to $\Pi\delta$. In appendix 2, the current state of $\Pi\delta+$ is shown.
For example, by using $\Pi\delta+$, the proposed relation for the *car/engine* example is *(Physical-)Mereotopological-Association* (PMA), defined as the union of some DOLCE-Lite+ primitive relations: part, connection, localization, constituency, etc., holding only within the *physical object* category. In fact, many possible relational paths can be walked from an instance of *physical object* to another, and only a wide-scope relation can cover them all. Formally:

$$\begin{aligned}\mathrm{PMA}(x,y) =_{df} (&\mathrm{Part}(x,y) \vee \mathrm{Overlaps}(x,y) \vee \mathrm{Strong\text{-}Connection}(x,y) \vee \\ &\vee \mathrm{Weak\text{-}Connection}(x,y) \vee \mathrm{Successor}(x,y) \vee \mathrm{Constituent}(x,y) \vee \\ &\vee \mathrm{Approximate\text{-}Location}(x,y)) \wedge \\ &\wedge \mathrm{Physical\_Object}(x) \wedge \mathrm{Physical\_Object}(y)\end{aligned}$$

Starting from $\Pi\delta+$, other relations have been defined for subsets of the domains and ranges of the relations in $\Pi\delta+$.
By means of $\Pi\delta+$, the query function ABox-1 can be adjusted as follows:

<u>ABox-2</u>

```
(retrieve (?x ?r ?y)
  (and (A-Link ?x ?y) (Superrelations ?x Physical_Object) (Superrelations ?y Physical_Object)
  (not  (and (Superrelations?x Unitary_Collection) (Superrelations?y Physical_Object)))
  (not  (and (Superrelations?x Amount_of_Matter) (Superrelations?y Physical_Body)))
  (not (subject ?x dolce)) (not (subject ?y dolce))
  (not (Superrelations ?x ?y)) (not (Superrelations ?y ?x))
  (min-cardinality ?x ?r 1)))
```

The query approximately reads "if two synsets subsumed by *physical object* (provided that the first is not an amount of matter or a collection, and that they are not related by hyperonymy), are linked by an A-link, tell me what relations in DOLCE+WordNet are applicable between those synsets with a cardinality of at least 1".
In this way, we are able to learn all the relations that are applicable to the classes ?x and ?y involved in the A-Link tuples. For example, applied to the synset *car#1* that has an A-link to the synset *engine#1*, the query returns:

$$R_{PMA}(car\#1, engine\#1)$$

that, on the basis of known assumptions, is used to propose an axiom on *car#1*, stating that cars have a "physical mereotopological association" with an *engine*, because a DOLCE-Lite+ ancestor of both *car#1* and *engine#1* ("*physical object*") defines the universe of the relation PMA with a cardinality of at least 1 on the range. This heuristics supports the logical axiom:

$$\forall x.\ \mathrm{Car}(x) \rightarrow \exists y.\ \mathrm{Engine}(y) \wedge \mathrm{PMA}(x,y)$$

Notice that at this level of generality, the classifier cannot infer the "component" relation that we intellectually guessed at the beginning of section 3.2. A more specific relation can be approximated, if we define more specialised relations and axioms. For example, a "functional co-participation" can be defined with a universe of only “functional objects”, which are lower in the DOLCE-Lite+ taxonomy, but still higher than the pair of synsets associated by the A-link. Functional co-participation (“FCP”) is defined by composing two participation relations with a common event (in the example, a common event could be “car running”):

$$\text{FCP}(x,y) =_{\text{df}} \exists z.\ \text{Participant\_in}(x,z) \wedge \text{Participant}(y,z) \wedge \text{Event}(z)$$

FCP is closer to the “component” intuition. The last can be precisely inferred if we feed the classifier with “core” domain relations. For example, we may define a domain relation holding for vehicles and functional objects, provided that the functional object plays the role of system component for vehicles:

$$\textit{vehicles}\text{\^{}Component}(x,y) =_{\text{df}} \text{FCP}(x,y) \wedge \text{Vehicle}(x) \wedge \text{Functional\_Object}(y) \wedge \wedge \exists z.\ \text{Plays}(y,z) \wedge \text{Vehicle\_System\_Component}(z)$$

In other words, by increasing the specificity of the domain (tourism in the examples discussed so far), we may assume that relations should be specified accordingly. As discussed in this section, this process is triggered by the observation of some A-link, and proceeds semi-automatically until a reasonable coverage is reached.

4. Experimental results and discussion

The gloss disambiguation algorithm and the A-link interpretation methods have been evaluated on two sets of glosses: a first set of 100 general-purpose glosses[4] and a second set of 305 glosses from a tourism domain. This allows us to evaluate the method both on a restricted domain and a non-specialized task.

For each term in a gloss, the appropriate WordNet sense has been manually assigned by two annotators, for over 1000 words.

To assess the performance of the gloss disambiguation algorithm we used two common evaluation measures: *recall* and *precision*. Recall provides the percentage of right senses with respect to the overall number of terms contained in the examined glosses. In fact, when the disambiguation algorithm terminates, the list P may still include terms for which no relation with the synsets in D could be found. Precision measures the percentage of right senses with respect to the retrieved gloss senses. A baseline precision is also computed, using the “first sense choice” heuristic. In WordNet, synsets are ordered by probability of use, i.e. the first synset is the most likely sense. For a fair comparison, the baseline is computed only on the words for which the algorithm could retrieve a synset.

Table 1 gives an overview of the results. Table 1a provides an overall evaluation of the algorithm, while table 1b computes precision and recall grouped by morphological category. The precision is quite high (well over 90% for both general and domain glosses) but the recall is around 40%. Remarkably, the achieved improvement in precision with respect to the baseline is much higher for general glosses than for domain glosses. This is motivated by the fact that general glosses include words that are more ambiguous than those in domain glosses. Therefore, the general gloss baseline is quite low. This means also that the disambiguation task is far more

[4] The 100 generic glosses have been randomly selected among the 809 glosses used to re-link WordNet to DOLCE-Lite+.

complex in the case of general glosses, where our algorithm shows particularly good performance.

An analysis of performance by morphological category (Table 1b) shows that noun disambiguation has much higher recall and precision. This is motivated by the fact that, in WordNet, noun definitions are richer than for verbs and adjectives. The WordNet hierarchy for verbs is known as being more problematic with respect to nouns. In the future, we plan to integrate in our algorithm verb information from FRAMENET[5], a lexico-semantic knowledge base providing rich information especially for verbs.

Domains	# glosses	# words	# disamb. words	# of which ok	Recall	Precision	Baseline Precision
Tourism	305	1345	636	591	47,28%	92,92%	82,55%
Generic	100	421	173	166	41,09%	95,95%	67,05%

Domains	noun recall	noun precision	adj recall	adj precision	verb recall	verb precision	# tot nouns	# tot adj	# tot verbs
Tourism	64,52%	92,86%	28,72%	89,29%	9,18%	77,78%	868	195	294
Generic	58,27%	95,95%	28,38%	95,24%	5,32%	80%	254	74	94

Table 1a) performance of the gloss disambiguation algorithm b) performance by morphological category.

In Table 2 we summarize the efficacy of the A-link semi-automatic axiomatization, after the partly manual creation of a domain view $\Pi\delta+$ as discussed in section 3.2.

As a preventive measure, we have excluded the A-links that include either an adjective or a verb, since these synsets have not been integrated yet with DOLCE-Lite+. Another measure excluded the A-links that imply a subsumption (sub-class) link, since these are already formalized. This filter has been implemented as a simple ABox query that uses relations that range on classes:

ABox-3

(retrieve (?x ?y) (and (A-Link ?x ?y) (Superrelations ?x ?y)))

These measures reduced the amount of A-Links from the experimental set to 582 (435+147). We have used these tuples to run the revised query ABox-2.

The revised query produced 711 (569+142) candidate axioms by using all the pruned relations defined for the experiment in $\Pi\delta+$. Table 3 shows the resulting axioms ordered by generality of the relation universes (domain and range).

The most relevant results (see Table 4 for relation data) are:

One third of the A-Links from the tourism domain are actually subsumption links, while only 20% from the mixed generic set is a subsumption. This could be explained by the fact that glosses for generic synsets are less informative, or because generic words are not defined, in WN, in terms of more generic ones.

The correct subset of axioms learnt for the tourism domain is about 4 to 6% larger than for the generic one with reference to the whole sets.

[5] http://www.icsi.berkeley.edu/~framenet/

Domains	Synsets	A-links	Noun-only	Subsumptions	Filtered A-links	Axioms generated	Correct
Tourism	305	725	644	209	435	569	511
Generic	100	212	187	40	147	142	121

Table 2. Axiomatizations for the A-links. "Best arrangement" data refer to results in Table 3.

	Tourism	Tourism correct	Generic	Generic correct
Total amount of axioms	569	511 (89.80%)	142	121 (85.21%)
Axioms with generic universes	540	490 (90.74%)	139	121 (87.05%)
Axioms with some specific universes	545	507 (93.02%)	136	118 (86.76%)
Axioms with only topmost universes	375	356 (94.93%)	110	98 (89.09%)

Table 3. Axiomatizations ordered by generality.

Relation taxonomy	**Tourism**	***Tou. corr.***	**Generic**	***Gen. corr.***
Conceptual_Relation (Entity, Entity)	*top: correct by A5*			
: Descriptive_Association (Object, S-Description)	7	*6*	5	*4*
: Functional_Association (Object, Functional-Role)	72	*68*	22	*19*
: : Functional_Role_Co_Participation (F-Role,F-Role)	21	*21*	13	*12*
: Physical_Location_Of (Geographical-Entity, Physical-Object)	2	*2*	2	*2*
: Mereotopological_Association (Physical-Object, Physical-Object)	140	*140*	29	*29*
: : Functional_CoParticipation (Functional-Object, Functional-Object)	98	*94*	1	*1*
: : Has_Member (Collection, Object)	4	*4*		
: : Provides (Functional-Object, Functional-Matter)	22	*17*	3	*0*
: : Biological_Part_Of (Biological-Object, Organism)			4	*4*
: : Has_Material_Constituent (Physical-Object, Amount-Of-Matter)	24	*4*	6	*3*
: : Used_By_Co_Pcp (Functional-Object, Agentive-Physical-Object)	7	*4*		
: : Member_Of (Object, Collection)	1	*0*		
: Participant (Event, Object)	14	*14*		
: : Agentive_Participant (Event, Agentive-Object)	3	*3*		
: Participant_In (Object, Event)	14	*13*	6	*6*
: Setting_For (Situation, Entity)	18	*17*		
: Setting (Entity, Situation)	21	*21*	8	*7*

Table 4. An excerpt of the experimental set of relations $\Pi\delta+$. Argument restrictions into brackets, with assignment numerosity (correct ones in italics).

We have tried to use some relations that are in principle "less precise". For example, a universe composed of *physical objects* and *amounts of matter* has a basic intuition of

“constituency”, and the relation *has_n_constituent* has been defined to such purpose. This relation has proved very inefficient though: in the generic set, only 50% of learnt axioms are correct, while in the tourism domain, only 16% are correct. We could expect that domains like *earth science* and *physics* can be more appropriate for constituency relations. For this reason, we have included a relation with a functional flavor in the experimental set of relations (including $\Pi\delta+$ and its specializations), called “provides”, and defined on *functional objects* and *functional matters* (this universe is a meaningful subset of the previous one). This relation proved quite efficient in the tourism domain, just as expected, with about 78% of correct axioms, while it is useless in the generic set, with 0%. This is an example of “provides” axioms: $\forall x$. Brasserie(x) $\rightarrow \exists$y. Beer(y) $\wedge$ Provides(x,y).
This, and similar examples, confirm our expectations about the importance of developing dedicated sets of relations for different domains or tasks.
In 8 cases, the axioms were not definable with a cardinality≥1, although they could be used in more restricted domains or for subclasses of the universe.
Some indirect A-links can be investigated as well (though our first strategy has been to disregard indirect links, as explained in section 3.1). For example in the *retrospective#1* example of Figure 2, two synsets (*painting#1* and *statue#1*) are learnt as “indirect” synsets (they are learnt from the glosses relative to the hyperonyms of *retrospective#1*). But paintings and statues are not always found in exhibitions, then we are not allowed to infer an axiom with cardinality ≥1. In these cases, the algorithm could be refined to propose an axiom that includes a common parent to both *painting#1* and *statue#1*, i.e. *art#1*, which incidentally is another “indirect” A-link to *retrospective#1*. In Figure 5 the refined A-links for *retrospective#1* are shown: a *retrospective* in WordNet 1.6 has the intended meaning of a (unitary) collection in DOLCE-Lite+, which is a kind of non-agentive functional object. This lets the classifier infer:

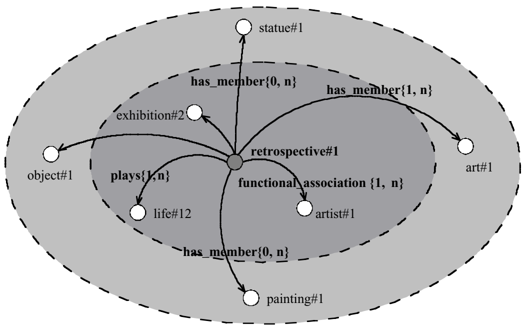

Figure 5. Interpretation of A-links for *retrospective#1*.

- a “functional association” to *artist#1*, because an artist is a functional role;
- a more precise “plays” relation to *life#12*, since an artistic biography is a functional role as well, and a collection of art works plays just the role of an artistic biography;
a subsumption of *retrospective#1* by *exhibition#2*;
- three “has_member” relationships to the indirect A-links: *art#1*, *painting#1*, and *statue#1*. These are correct, since a collection can have functional objects (art works)

as members. But while the first has a meaningful cardinality 1 to n, the others have a logically irrelevant cardinality of 0 to n.

Conclusions

In this paper we have presented some preliminary results of OntoWordNet, a large-scale project aiming at the "ontologization" of WordNet. We presented a two step methodology: during the first, automatic phase, natural language word sense glosses in WordNet are parsed, generating a first, approximate definition of WN concepts (originally called synsets). In this definition, generic associations (A-links) are established between a concept and the concepts that co-occur in its gloss.
In a second phase, the foundational top ontology DOLCE (in the *DOLCE-Lite+* version), including few hundreds formally defined concepts and conceptual relations, is used to interpret A-links in terms of axiomatised conceptual relations. This is a partly automatic technique that involves generating solutions on the basis of the available axioms, and then creating a specialized partition of the axioms (the set $\Pi\delta+$ and its specializations) in order to capture more domain-specific phenomena.
Overall, the experiments that we conducted show that a high performance may be obtained through the use of automatic techniques, significantly reducing the manual effort that would be necessary to pursue the objective of the OntoWordNet project.

References

(Basili et al. 1996) Basili R., Pazienza M.T. and Velardi P. *An Empirical Symbolic Approach to Natural Language Processing*, Artificial Intelligence, n. 85, (1996).

(Berners-Lee, 1998) Berners-Lee T., Semantic Web Road map, http://www.w3.org/DesignIssues/Semantic.html, 1998.

(Fellbaum 1995) Fellbaum, C. *WordNet: an electronic lexical database*, Cambridge, MIT press, (1995).

(Gangemi et al. 2001) Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. 2001. Understanding top-level ontological distinctions. In *Proceedings of IJCAI-01 Workshop on Ontologies and Information Sharing*. Seattle, USA, AAAI Press: 26-33. http://SunSITE.Informatik.RWTHAachen. DE/Publications/CEUR-WS/Vol-47/

(Gangemi et al. 2002) Gangemi A., Guarino N. Masolo C. Oltramari A., Schneider L. "Sweetening Ontologies with DOLCE", Proc. Of EKAW02 http://citeseer.nj.nec.com/cache/papers/cs/26864/http

(Harabagiu and Moldovan 1999) Harabagiu S. and Moldovan D. *Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text*. AAAI/MIT Press, (1999).

(Karkaletsis V. Cucchiarelli A Paliouras G. Spyropolous C. Velardi P. "Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods" 23rd annual ACM-SIGIR, Athens, June 2000.

(Johnson 1987) Johnson, Mark. 1987. *The Body in the Mind*. Chicago: University of Chicago Press.

(Mac Gregor, 1993) MacGregor, R. M. 1993. Using a Description Classifier to Enhance Deductive Inference. In *Proceedings of Seventh IEEE Conference on AI Applications*: 141-147.

(Masolo et al. 2002) Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, Alessandro Oltramari, and Luc Schneider. The WonderWeb Library of Foundational Ontologies. WonderWeb Deliverable 17, 2002.

(Magnini and Caviglia 2000) Magnini, B. and Cavaglia, G.: Integrating Subject Field Codes into WordNet. Proceedings of the 2nd International Conference on Language resources and Evaluation, LREC2000, Atenas .

(Miller et al. 1993) G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross & K. Miller; "Introduction to WordNet: An On-Line Lexical Database"; http://www.cosgi.princeton.edu/~wn; August 1993.

(Mihalcea and Moldovan, 2001) Milhalcea, R. and Moldovan. D. *eXtended WordNet: progress report*. NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsbourgh, June (2001).

(Moeller, 2001) Volker Haarslev, Ralf Möller Description of the RACER System and its Applications Proceedings International Workshop on Description Logics (DL-2001), Stanford, USA, 1.-3. August 2001

(Morin, 1999) Morin E., *Automatic Acquisition of semantic relations between terms from technical corpora*, Proc. of 5th International Congress on Terminology and Knowledge extraction, TKE-99, (1999).

(Navigli et al. 2003) R. Navigli, P. Velardi and A. Gangemi, *Ontology Learning and its Application to Automated Terminology Translation*, IEEE Intelligent Systems, vol. 18, n.1, pp. 22-31, January 2003

(Searle, 1985) Searle, J.R and Vanderveken, D. *Foundations of Illocutionary Logics*, Cambridge UP, (1985).

(Smith and Welty, 2001) Smith, B. and Welty, C. *Ontology: towards a new synthesis*, Formal Ontology in Information Systems, ACM Press, (2001).

(Vossen 2001) Vossen P. *Extending, Trimming and Fusing WordNet for technical Documents*, NAACL 2001 workshop on WordNet and Other Lexical Resources, Pittsburgh, July (2001).

Web Sites

DAML+OIL http://www.daml.org/2001/03/daml+oil-index

LDC corpus http://www.ldc.upenn.edu/

FRAMENET http://www.icsi.berkeley.edu/~framenet/

WordNet 1.6 http://www.cogsci.princeton.edu/~wn/w3wn.html

Semcor http://engr.smu.edu/~rada/semcor/

WonderWeb http://wonderweb.semanticweb.org