ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge

Luisa Bentivogli[1], Andrea Bocco[2], and Emanuele Pianta[1]

[1] ITC-irst, Via Sommarive 18
38050 Povo – Trento, Italy
{bentivo,pianta}@itc.it
[2] Politecnico di Torino – Dipartimento di Casa Città, viale Mattioli 39
10125 Torino, Italy
{andrea.bocco}@polito.it

Abstract. Linguistic resources with domain-specific coverage are crucial for the development of concrete application systems, especially when integrated with domain-independent resources. In this paper we present our experience in the creation of ArchiWordNet, a specialized WordNet for the architecture and construction domain which is being created according to the WordNet model and integrated with WordNet itself. Problematic issues related to the creation of a domain-specific wordnet and its integration with a general language resource are discussed, and practical solutions adopted are described.

1 Introduction

The ArchiWordNet (ArchiWN) project is a joint effort between ITC-irst and the Turin Polytechnic aiming at building a thesaurus for the architecture domain to be used within Still Image Server (SIS), an architecture image archive available at the Polytechnic.

SIS was created for educational purposes, with the aim of making accessible to Architecture students and researchers the huge iconographic heritage available in different departments, thus contributing to the preservation and development of the heritage itself. The digitized images are catalogued and organized in a database that can be queried through a web interface accessible within the Polytechnic Intranet. During the cataloguing phase, several keywords are assigned to each image. To make the use of the keywords more systematic and facilitate the retrieval of the images, it is necessary to constrain the keywords used by both indexers and end users through a thesaurus. However, up to now an exhaustive thesaurus for the architecture domain able to meet the needs of the image archive has not been available and thus we decided to create ArchiWN, a bilingual WordNet-like English/Italian thesaurus to be integrated into WordNet itself.

In this paper our experience in the creation of ArchiWN is presented. Section 2 describes the motivations behind the decision of building a WordNet-like thesaurus and its distinguishing features. In sections 3 and 4 some problematic issues related to the creation of a domain-specific WordNet and its integration with a general language resource are discussed, and the practical solutions adopted are presented. Finally, Section 5 outlines ArchiWN future enhancements and new application fields.

2 ArchiWordNet: a WordNet-like thesaurus

The main characteristic of ArchiWN is that, while exploiting as much as possible information from already existing architecture thesauri and other specialized sources, it is structured according to the WordNet model [4] and fully integrated into it. More specifically, as we aim at creating a bilingual English/Italian resource, we decided to work within the MultiWordNet (MultiWN) framework. MultiWN [7] is a multilingual lexical database in which the Italian WordNet is strictly aligned with Princeton's English WordNet.

ArchiWN will differ from traditional thesauri with respect to both concepts and relations [2]. Thesauri usually represent concepts using a controlled vocabulary where many synonyms are missed. Also, they include few relations (such as "broader term", "narrower term", "used for", and "related to") whose semantics is rather informal. On the contrary, concepts in WordNet are represented by sets of synonymous words actually occurring in the real language, and WordNet relations are explicit and encoded in a homogeneous way, enabling transitivity and thus inheritance. Given these differences, we decided to adopt the WordNet model for a number of reasons. On the one side, the more rigorous structure of WordNet allows for a more powerful and expressive retrieval mechanism. On the other side, it makes ArchiWN more suitable for educational purposes, as it provides conceptual frameworks which can support learning: its well-structured hierarchies can be browsed to form both a general idea of the architecture domain and a structured knowledge of specific topics.

ArchiWN will differ from traditional thesauri not only in its structure but also in the fact that it is fully integrated with MultiWN. From a theoretical point of view, MultiWN offers a general and multilingual framework for the specialized knowledge contained in ArchiWN. From a practical point of view, the possibility of integrated access allows more flexible retrieval of the information. Moreover, given the huge cost in terms of human effort involved in the construction of such a resource, the integration is particularly useful as information already existing in the generic WordNet can be exploited in the creation of the specialized one.

Throughout the ArchiWN creation phase, we have been faced with the tension between the diverging aims of two different disciplines such as computational linguistics and architecture. More specifically, we had to find a trade off between the necessity of creating a linguistically motivated formalized resource, suitable also for Natural Language Processing applications, and building an application-oriented tool geared to meet the practical needs of specialists in the field. This interdisciplinary cooperation turned out to be an added value. In fact, with respect to other specialized thesauri, ArchiWN has the advantage of having a formalized structure and of inheriting linguistic oriented information from the generic WordNet; with respect to other lexical resources, it has the advantage that many synsets will be associated with images representing the concept.

Another distinguishing characteristic of ArchiWN with respect to other existing WordNet-like lexical resources is the fact that the synonyms will be ordered on the basis of their representativeness with respect to the concept they express: given a synset, the first synonym will be the word which is most commonly used by domain experts to express that concept.

In the creation of ArchiWN we had to face a number of problematic issues related both to the adoption of the MultiWN model and to the integration with MultiWN itself. In the following Sections we discuss the different steps that have to be undertaken in order to build such a resource.

3 Adopting and adapting the MultiWordNet model

Two basic criteria have been followed in the construction of ArchiWN. First, we referred as much as possible to already existing and widely accepted specialized sources for the architecture and construction domain. Second, MultiWN information is exploited whenever possible to create those hierarchies for which a complete and well structured domain-specific terminology is not available.

With regard to domain-specific sources, various specialized materials have been used to create both the synsets and the hierarchies of ArchiWN, among which the *Art and Architecture Thesaurus* (AAT) [6], the *Construction Indexing Manual* of CI|SfB [8], the international and national standardization rules (ISO, CEN, UNI), the *Lessico per la descrizione delle alterazioni e degradazioni macroscopiche dei materiali lapidei* created by the NORMAL commission, and other scientific literature in the area, technical dictionaries included. Both English and Italian sources are being used and correspondences between the two languages have to be found to create the bilingual synsets of ArchiWN. From the analysis of these sources, it turned out that very often they are not compatible with the MultiWN model. Either they are not structured on the basis of the ISA relation or they present mixed hierarchies where different levels are not homogeneous and relations between concepts are underspecified and ambiguous. On the contrary, relations in WordNet are explicit and information is encoded in a homogeneous way. Thus, it is necessary to reorganize these sources to make them compatible with the WordNet model. An example is given by the reorganization of the AAT hierarchy for the term “metal”, an excerpt of which is shown in Figure 1.

To make AAT compatible with the ArchiWN model, we had to interpret its spurious relations by disambiguating the type of relation connecting superordinate and subordinate concepts and by deciding how to manage intermediate “artificial” nodes which are not relevant from the point of view of the ISA hierarchy. As it can be seen in Figure 1, the artificial nodes have been eliminated and only the ISA relations have been maintained. The concepts previously connected to “metal” by a “form” relation have been modified, put in their appropriate ISA hierarchy, and connected to “metal” with the HAS-SUBSTANCE WordNet relation.

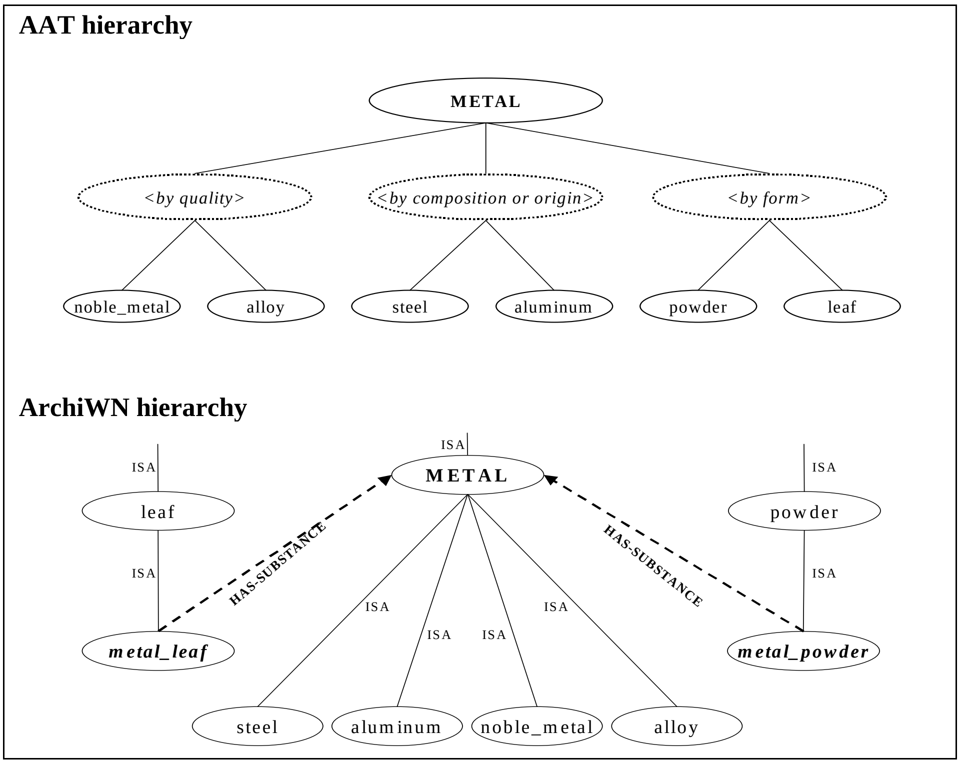

Fig. 1. Reorganization of the AAT hierarchy for "metal" according to the WordNet model

The second main source for the creation of ArchiWN, mainly used when a complete and structured domain-specific terminology is not available, is MultiWN itself. Synsets already existing in MultiWN which are considered appropriate by the domain experts are included into ArchiWN. However, this methodology cannot always be applied straightforwardly. In fact, as MultiWN synsets represent general language while ArchiWN must represent a specialized language, it is possible that both MultiWN synsets and relations are not always completely suitable for representing the architecture and construction domain. When included into ArchiWN, MultiWN synsets can undergo three different kinds of modification.

First, in those cases where the criterion for synonymy suitable for MultiWN is inadequate for ArchiWN, it is possible to add or delete synonyms to MultiWN synsets. This can happen as words that are considered synonyms in everyday usage may not be synonyms in the architecture domain. Second, when general language definitions are not compatible with a technical definition, it is possible to modify MultiWN definitions of the synsets. Third, it must be possible to delete and add relations between synsets. When included into ArchiWN, a synset can maintain all or some or none of its original MultiWN relations, depending on their appropriateness to the architecture domain. Moreover, other relations can be added to encode further information relevant to the specialized domain.

Finally, three new semantic relations, missing in MultiWN but useful to define concepts in the architecture domain, have been introduced in ArchiWN:

- HAS FORM (n/n) {tympanum} HAS FORM {triangle, trigon, trilateral}
- HAS ROLE (n/n) {metal section} HAS ROLE {upright, vertical}
- HAS FUNCTION[1] (n/v) {beam} HAS FUNCTION {to hold, to support, …}

4 Integrating ArchiWordNet with MultiWordNet

To integrate ArchiWN with MultiWN, a first list of 5,000 terms has been created relying on the specialized sources described above and on the direct experience of the domain experts. Then, the majority of such terms have been grouped in 13 semantic areas, as shown in Table 1. These semantic areas correspond to the main hierarchies to be represented in ArchiWN.

After the identification of the MultiWN nodes where to insert the ArchiWN hierarchies, the integration procedure requires (i) the actual inclusion of ArchiWN hierarchies in MultiWN, and (ii) the handling of the overlapping between terms present in both MultiWN and ArchiWN. This latter requirement is due to the fact that, unlike other domains characterized by a very specialized terminology, the architecture domain includes a significant amount of terms commonly used in the general language.

In the literature, different approaches are presented to address the problem of linking existing lexical/semantic hierarchies [3] and of integrating the information of a generic lexical resource with domain-specific information [9, 5, 1]. The methodology we developed to realize the integrated wordnet takes as a basis the "plug-in" approach proposed in [5] with some basic differences and extentions. In [5] two existing, independently created, wordnets are connected whereas ArchiWN is created so as to maximize the integration with MultiWN. Thus, to meet our needs, some existing procedures were extended and new procedures were created, especially for maximizing the exploitation of MultiWN information.

Our methodology consists of *basic operations* that can be performed on single MultiWN synsets and that constitute the basis of *complex procedures (plug-in)* which apply to entire hierarchies. The basic operations allow us to:

a. eclipse a synset
b. tag a synset with the "architecture and construction" domain label

[1] The HAS ROLE and HAS FUNCTION relations can be compared to the EuroWordNet [10] (EuroWN) INVOLVED/ROLE relation which connects second-order entities (i.e. nouns and verbs expressing properties, acts, processes, states, events) to first-order entities (i.e. concrete nouns referring to physical things). However, in EuroWN, the INVOLVED/ROLE relation is used for encoding information on arguments/adjuncts that are strongly implied in the meaning of a second-order verb/noun. For example, "to hammer" INVOLVED "hammer" and "hammer" ROLE "to hammer". On the contrary, given the specialized nature of ArchiWN, we are more interested in adding *encyclopaedic* information, concerning the usage of concrete entities within the architecture field. The HAS ROLE and HAS FUNCTION relations are used to encode the function of an entity; such function is not necessarily inherent in the semantics of the word designating the entity.

c. add or delete relations to a synset
d. add or delete synonyms in a synset
e. modify the synset definition.

The eclipsing operation (a) removes a certain MultiWN synset and all relations originating from that synset. It is used to avoid overlappings when a specialized synset has been created in ArchiWN and a similar synset already exists in MultiWN but it is not considered suitable to be included into ArchiWN. The labeling operation (b) has the effect of including a MultiWN synset in ArchiWN, when this is considered suitable for the architecture and construction domain. It is used to avoid overlappings exploiting MultiWN information. Removing and adding relations to synsets (c) are the fundamental integration operations. Merging ArchiWN and MultiWN always requires adding one or more new relations to a synset (the root of the hierarchy in the case of complex procedures) and sometimes removing all or some of its original relations.

Finally, to customize MultiWN synsets to the architecture and construction domain operations of type (d) and (e) can be carried out (see Section 3).

To operate on ArchiWN and MultiWN hierarchies, we devised four complex procedures, able to cope with different integration requirements:

- *Substitutive plug-in*. A hierarchy from ArchiWN substitutes a MultiWN sub-hierarchy. This procedure, involving the eclipsing of all synsets in the MultiWN hierarchy, is used when an ArchiWN hierarchy is rich and well structured while the corresponding MultiWN one is not.
- *Integrative plug-in*. The two hierarchies are merged. The root of the ArchiWN sub-hierarchy substitutes the MultiWN one and the MultiWN hyponyms relevant to the architecture domain are included in ArchiWN through a labeling operation. This plug-in procedure is used when MultiWN has a well structured hierarchy and thus it is useful to integrate this information with the specialized one.
- *Hyponymic plug-in*. An ArchiWN hierarchy is connected as a hyponym of a MultiWN synset.
- *Inverse plug-in*. A MultiWN sub-hierarchy (possibly part of an eclipsed sub-hierarchy) is moved from MultiWN and connected to ArchiWN as a hyponym of an ArchiWN synset. This procedure is mainly used to exploit portions of MultiWN hierarchies which are considered relevant to the architecture and construction domain but are not in a correct position in MultiWN.

Given this methodology, we identified for each ArchiWN hierarchy one or more plug-in nodes in MultiWN and the complex procedures to be applied. As summarized in Table 1, some hierarchies can be directly plugged in MultiWN, while others required reorganizing MultiWN hierarchies. The results obtained in the integration phase are quite encouraging, showing not only that it is possible to integrate ArchiWN with MultiWN, but also that MultiWN can be widely exploited in the creation of ArchiWN hierarchies. In fact, for eight ArchiWN hierarchies we could exploit an integrative plug-in, while a substitutive plug-in was necessary for only three ArchiWN hierarchies. Finally, two ArchiWN hierarchies ("components of buildings" and "single buildings and buildings complexes") required a reorganization of some MultiWN sub-hierarchies, involving some plug-hyponymies, large synset eclipsing, but also a number of inverse plug-ins, which means the reuse of some MultiWN sub-hierarchies.

Table 1. Integration of ArchiWN hierarchies with MultiWN

ArchiWN hierarchies	**MultiWN Plug-in nodes** (lemma/sense number)	**Type of plug-in**
Architectural styles	architectural_style/1	substitutive
Materials	material/1, substance/1	substitutive
Construction products	building_material/1	substitutive
Techniques	technique/1	integrative
Tools	tool/1	integrative
Components of buildings	structure/1, component/3, region/1	hyponymic
Single buildings and building complexes	structure/ArchiWN building/1, building_complex/1	hyponymic inverse
Physical properties	physical_property/1	integrative
Conditions	condition/1	integrative
Disciplines	discipline/1	integrative
People	person/1	integrative
Documents	document/1	integrative
Drawings and representations	drawing/2, representation/2	integrative

As regards the population of ArchiWN, up to now the “Simple buildings and building complexes” sub-hierarchy has been populated with about 900 synsets, containing in most cases both Italian and English synonyms along with an accurate definition.

This work has been done manually, using the MultiWN graphical interface which allows the user both to modify existing synsets and relations and to create new synsets.

During the creation of the bilingual synsets, we had to deal with the issue of lexical gaps, i.e. cases in which a language expresses a concept with a lexical unit whereas the other language does not. For example, the English synset for the word “kirk” (a Scottish church) has not an Italian correspondent and, viceversa, the Italian synset for “trullo” (a typical rural construction from Apulia, Italy) has not an English correspondent. However, this kind of idiosyncrasy does not represent a significant problem as it does not involve mismatches in the hierarchies. Moreover, as the specialized architecture lexicon mainly refers to objects and physical phenomena, in general we think that also for the remaining ArchiWN hierarchies we will not be faced with particularly problematic cross-linguistic idiosyncrasies.

5 Conclusion and future work

In this paper we have presented our experience in the creation of ArchiWN. The analysis of the problematic issues that arose, and the development and integration work carried out up to now show both that it is possible to integrate ArchiWN with MultiWN and that MultiWN itself can be considered a useful resource to be exploited in the creation of ArchiWN hierarchies.

With regard to future work, we will go on enriching the “Simple buildings and building complexes” hierarchy and populating the remaining hierarchies.

Moreover, we received a request from the Italian Architectural aluminium and steel manufacturers association (UNCSAAL) to create a multilingual specialized lexicon of approximately 1,000 synsets specifically referring to the window and curtain wall industry. In order to meet the needs of this industrial application, a further development of some of the hierarchies is planned, together with the extension of the resource to other languages such as German, French, and possibly Spanish.

ArchiWN's range of applications will be twofold: it will be a thesaurus for cataloguing images within the SIS archive, and a useful integrated resource for Natural Language Processing applications. Moreover, an important achievement is represented by an agreement which is under way for the future usage of ArchiWN by the institutions in charge of cataloguing the architectural cultural heritage of the Piemonte region.

Acknowledgements

This project is being carried out under the precious scientific direction of Professor Gianfranco Cavaglià (Turin Polytechnic) and Fabio Pianesi (ITC-irst). Enrica Bodrato and Antonella Perin of the Turin Polytechnic have being compiling the thesaurus with competence and accuracy.

References

1 Buitelaar, P. and Sacaleanu, B.: Extending Synsets with Medical Terms. In: *Proceedings of the First International Conference on Global WordNet*, Mysore, India (2002)

2 Clark, P., Thompson, J., Holmback, H.and Duncan, L.: Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search. In: *Proceedings of AAAI/IAAI 2000*, Austin, Texas (2000)

3 Daudé, J., Padro, L. and Rigau, G.: Mapping WordNets Using Structural Information. In: *Proceedings of ACL 2000*, Hong Kong (2000)

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

5 Magnini, B. and Speranza, M.: Integrating Generic and Specialized Wordnets. In: *Proceedings of the Euroconference RANLP 2001*, Tzigov Chark, Bulgaria (2001)

6 Petersen, T. (director): *Art and Architecture Thesaurus*, Oxford University Press, New York-Oxford (1994). http://www.getty.edu/research/tools/vocabulary/aat/

7 Pianta, E. Bentivogli, L. and Girardi, C.: MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India (2002)

8 Ray-Jones, A. and Clegg, D.: *CI|SfB. Construction Indexing Manual 1976*, RIBA Publications, London (1991)

9 Turcato, D., Popowich, F., Toole, J., Fass, D., Nicholson, D., and Tisher, G.: Adapting a Synonym Database to Specific Domains. In: *Proceedings of ACL 2000 Workshop on Information Retrieval and Natural Language Processing*, Hong Kong (2000)

10 Vossen, P. (ed.): *Computers and the Humanities,* Special Issue on EuroWordNet, Volume 32, Nos. 2-3 (1998)