Mapping Concrete Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and Results

Adriana Roventini, Nilda Ruimy, Rita Marinelli, Marisa Ulivieri, Michele Mammini
Istituto di Linguistica Computazionale – CNR
Via Moruzzi,1 – 56124 – Pisa, Italy
{adriana.roventini,nilda.ruimy,rita.marinelli,
marisa.ulivieri,michele.mammini}@ilc.cnr.it

Abstract

This paper describes a work in progress aiming at linking the two largest Italian lexical-semantic databases ItalWordNet and PAROLE-SIMPLE-CLIPS. The adopted linking methodology, the software tool devised and implemented for this purpose and the results of the first mapping phase regarding 1stOrderEntities are illustrated here.

1 Introduction

The mapping and the integration of lexical resources is today a main concern in the world of computational linguistics. In fact, during the past years, many linguistic resources were built whose bulk of linguistic information is often neither easily accessible nor entirely available, whereas their visibility and interoperability would be crucial for HLT applications.

The resources here considered constitute the largest and extensively encoded Italian lexical semantic databases. Both were built at the CNR Institute of Computational Linguistics, in Pisa.

The ItalWordNet lexical database (henceforth IWN) was first developed in the framework of EuroWordNet project and then enlarged and improved in the national project SI-TAL[1]. The theoretical model underlying this lexicon is based on the EuroWordNet lexical model (Vossen, 1998) which is, in its turn, inspired to the Princeton WordNet (Fellbaum, 1998).

PAROLE-SIMPLE-CLIPS (PSC) is a four-level lexicon developed over three different projects: the LE-PAROLE project for the morphological and syntactic layers, the LE-SIMPLE project for the semantic model and lexicon and the Italian project CLIPS[2] for the phonological level and the extension of the lexical coverage. The theoretical model underlying this lexicon is based on the EAGLES recommendations, on the results of the EWN and ACQUILEX projects and on a revised version of Pustejovsky's Generative Lexicon theory (Pustejovsky 1995).

In spite of the different underlying principles and peculiarities characterizing the two lexical models, IWN and PSC lexicons also present many compatible aspects and the reciprocal enhancements that the linking of the resources would entail were illustrated in Roventini et al., (2002); Ruimy & Roventini (2005). This has prompted us to envisage the semi-automatic link of the two lexical databases, eventually merging the whole information into a common representation framework. The first step has been the mapping of the 1stOrderEntities which is described in the following.

This paper is organized as follows: in section 2 the respective ontologies and their mapping are briefly illustrated, in section 3 the methodology followed to link these resources is described; in section 4 the software tool and its workings are explained; section 5 reports on the results of the complete mapping of the 1stOrderEntities. Future work is outlined in the conclusion.

2 Mapping Ontology-based Lexical Resources

In both lexicons, the backbone for lexical representation is provided by an ontology of semantic types.

[1] *Integrated System for the Automatic Language Treatment.*

[2] *Corpora e Lessici dell'Italiano Parlato e Scritto.*

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 161–164,
Prague, June 2007. ©2007 Association for Computational Linguistics

The IWN Top Ontology (TO) (Roventini et al., 2003), which slightly differs from the EWN TO[3], consists in a hierarchical structure of 65 language-independent Top Concepts (henceforth TCs) clustered in three categories distinguishing 1st OrderEntities, 2ndOrderEntities and 3rdOrder Entities. Their subclasses, hierarchically ordered by means of a subsumption relation, are also structured in terms of (disjunctive and non-disjunctive) opposition relations. The IWN database is organized around the notion of *synset*, i.e. a set of synonyms. Each synset is ontologically classified on the basis of its hyperonym and connected to other synsets by means of a rich set of lexical-semantic relations. Synsets are in most cases cross-classified in terms of multiple, non disjoint TCs, e.g.: *informatica* (computer science): [Agentive, Purpose, Social, Unboundedevent]. The semantics of a word sense or *synset variant* is fully defined by its membership in a synset.

The SIMPLE Ontology (SO)[4], which consists of 157 language-independent semantic types, is a multidimensional type system based on hierarchical and non-hierarchical conceptual relations. In the type system, multidimensionality is captured by *qualia roles* that define the distinctive properties of semantic types and differentiate their internal semantic constituency. The SO distinguishes therefore between *simple* (one-dimensional) and *unified* (multi-dimensional) semantic types, the latter implementing the principle of *orthogonal inheritance.* In the PSC lexicon, the basic unit is the word sense, represented by a 'semantic unit' (henceforth, *SemU*). Each SemU is assigned one single semantic type (e.g.: *informatica*: [Domain]), which endows it with a structured set of semantic information.

A primary phase in the process of mapping two ontology-based lexical resources clearly consisted in establishing correspondences between the conceptual classes of both ontologies, with a view to further matching their respective instances.

The mapping will only be briefly outlined here for the 1stOrderEntity. More information can be found in (Ruimy & Roventini 2005; Ruimy, 2006).

The IWN 1stOrderEntity class structures concrete entities (referred to by concrete nouns). Its main cross-classifying subclasses: Form, Origin, Composition and Function correspond to the four Qualia roles the SIMPLE model avails of to express orthogonal aspects of word meaning. Their respective subdivisions consist of (mainly) disjoint classes, e.g. Natural vs. Artifact. To each class corresponds, in most of the cases, a SIMPLE semantic type or a type hierarchy subsumed by the Concrete_entity top type. Some other IWN TCs, such as Comestible, Liquid, are instead mappable to SIMPLE distinctive features: e.g. Plus_Edible, Plus_Liquid, etc.

3 Linking Methodology

Mapping is performed on a semantic type-driven basis. A semantic type of the SIMPLE ontology is taken as starting point. Considering the type's SemUs along with their PoS and 'isa' relation, the IWN resource is explored in search of linking candidates with same PoS and whose ontological classification matches the correspondences established between the classes of both ontologies.

A characteristic of this linking is that it involves lexical elements having a different status, i.e. semantic units and synsets.

During the linking process, two different types of data are returned from each mapping run:
1) A set of matched pairs of word senses, i.e. SemUs and synset variants with identical string, PoS and whose respective ontological classification perfectly matches. After human validation, these matched word senses are linked.
2) A set of unmatched word senses, in spite of their identical string and PoS value. Matching failure is due to a mismatch of the ontological classification of word senses existing in both resources. Such mismatch may be originated by:
a) an incomplete ontological information. As already explained, IWN synsets are cross-classified in terms of a combination of TCs; however, cases of synsets lacking some meaning component are not rare. The problem of incomplete ontological classification may often be overcome by relaxing the mapping constraints; yet, this solution can only be applied if the existing ontological label is informative enough. Far more problematic to deal with are those cases of incomplete or little informative ontological labels, e.g. 1stOrderEntities as different as *medicinale, anello, vetrata* (medicine, ring, picture window) and only classified as 'Function';

[3] A few changes were in fact necessary to allow the encoding of new syntactic categories.

[4] http://www.ilc.cnr.it/clips/Ontology.htm

b) a different ontological information. Besides mere encoding errors, ontological classification discrepancy may be imputable to:

i) a different but equally defensible meaning interpretation (e.g.: *ala* (aircraft wing) : [Part] vs. [Artifact Instrument Object]). Word senses falling into this category are clustered into numerically significant sets according to their semantic typing and then studied with a view to establishing further equivalences between ontological classes or to identify, in their classification schemes, descriptive elements lending themselves to be mapped.

ii) a different level of specificity in the ontological classification, due either to the lexicographer's subjectivity or to an objective difference of granularity of the ontologies.

The problems in ii) may be bypassed by climbing up the ontological hierarchy, identifying the parent nodes and allowing them to be taken into account in the mapping process.

Hyperonyms of matching candidates are taken into account during the linking process and play a particularly determinant role in the resolution of cases whereby matching fails due to a conflict of ontological classification. It is the case for sets of word senses displaying a different ontological classification but sharing the same hyperonym, e.g. *collana*, *braccialetto* (necklace, bracelet) typed as [Clothing] in PSC and as [Artifact Function] in IWN but sharing the hyperonym *gioiello* (jewel). Hyperonyms are also crucial for polysemous senses belonging to different semantic types in PSC but sharing the same ontological classification in IWN, e.g.: SemU1595*viola* (violet) [Plant] and SemU1596*viola* (violet) [Flower] vs. IWN: *viola*1 (has_hyperonym *pianta*1 (plant)) and *viola*3 (has_hyperonym *fiore*1 (flower)), both typed as [Group Plant].

4 The Linking Tool

The LINKPSC_IWN software tool implemented to map the lexical units of both lexicons works in a semiautomatic way using the ontological classifications, the 'isa' relations and some semantic features of the two resources. Since the 157 semantic types of the SO provide a more fine-grained structure of the lexicon than the 65 top concepts of the IWN ontology, which reflect only fundamental distinctions, mapping is PSC → IWN oriented. The mapping process foresees the following steps:

1) Selection of a PSC semantic type and definition of the loading criteria, i.e. either all its SemUs or only those bearing a given information;

2) Selection of one or more mapping constraints on the basis of the correspondences established between the conceptual classes of both ontologies, in order to narrow the automatic mapping;

3) Human validation of the automatic mapping and storage of the results;

4) If necessary, relaxation/tuning of the mapping constraints and new processing of the input data.

By human validation of the automatic mapping we also intend the manual selection of the semantically relevant word sense pair(s) from the set of possible matches automatically output for each SemU. A decision is taken after checking relevant information sources such as hyperonyms, SemU/synset glosses and the IWN-ILI link.

Besides the mapping results, a list of unmatched word senses is provided which contains possible encoding errors and polysemous senses of the considered SemUs (e.g., *kiwi* (fruit) which is discarded when mapping the 'Animal' class). Some of these word senses proceed from an extension of meaning, e.g. People-Human: *pigmeo, troglodita* (pygmy, troglodyte) or Animal-Human *verme, leone* (worm, lion) and are used with different levels of intentionality: either as a semantic surplus or as dead metaphors (Marinelli, 2006).

More interestingly, the list of unmatched words also contains the IWN word senses whose synset's ontological classification is incomplete or different w.r.t. the constraints imposed to the mapping run. Analyzing these data is therefore crucial to identify further mapping constraints. A list of PSC lexical units missing in IWN is also generated, which is important to appropriately assess the lexical intersection between the two resources.

5 Results

From a quantitative point of view three main issues are worth noting (cf. Table 1): first, the considerable percentage of linked senses with respect to the linkable ones (i.e. words with identical string and PoS value); second, the many cases of multiple mappings; third, the extent of overlapping coverage.

SemUs selected	27768	
Linkable senses	15193	54,71%
Linked senses	10988	72,32%
Multiple mappings	1125	10,23%
Unmatched senses	4205	27,67%

Table 1 summarizing data

Multiple mappings depend on the more fine grained sense distinctions performed in IWN. The eventual merging of the two resources would make up for such discrepancy.

During the linking process, many other possibilities of reciprocal improvement and enrichment were noticed by analyzing the lists of unmatched word-senses. All the inconsistencies are in fact recorded together with their differences in ontological classification, or in the polysemy treatment that the mapping evidenced. Some mapping failures have been observed due to a different approach to the treatment of polysemy in the two resources: for example, a single entry in PSC corresponding to two different IWN entries encoding very fined-grained nuances of sense, e.g. *galeotto*1 (galley rower) and *galeotto*2 (galley slave).

Other mapping failures are due to cases of encoding inconsistency. For example, when a word sense from a multi-variant synset is linked to a SemU, all the other variants from the same synset should map to PSC entries sharing the same semantic type, yet in some cases it has been observed that SemUs corresponding to variants of the same synset do not share a common semantic type.

All these encoding differences or inconsistencies were usefully put in the foreground by the linking process and are worthy of further in-depth analysis with a view to the merging, harmonization and interoperability of the two lexical resources.

6 Conclusion and Future Work

In this paper the PSC-IWN linking of concrete entities, the methodology adopted, the tool implemented to this aim and the results obtained are described. On the basis of the encouraging results illustrated here, the linking process will be carried on by dealing with 3rdOrder Entities. Our attention will then be devoted to 2ndOrderEntities which, so far, have only been object of preliminary investigations on Speech act (Roventini 2006) and Feeling verbs. Because of their intrinsic complexity, the linking of 2ndOrderEntities is expected to be a far more challenging task.

References

James Pustejovsky 1995. *The generative lexicon*. MIT Press.

Christiane Fellbaum (ed.) 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

Piek Vossen (ed.) 1998. EuroWordNet: *A multilingual database with lexical semantic networks*. Kluwer Academic Publishers.

Adriana Roventini et al. 2003. ItalWordNet: *Building a Large Semantic Database for the Automatic Treatment of Italian*. Computational Linguistics in Pisa, Special Issue, XVIII-XIX, Pisa-Roma, IEPI. Tomo II, 745--791.

Nilda Ruimy et al. 2003. *A computational semantic lexicon of Italian: SIMPLE*. In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo II, 821-864.

Adriana Roventini, Marisa Ulivieri and Nicoletta Calzolari. 2002 *Integrating two semantic lexicons, SIMPLE and ItalWordNet: what can we gain?* LREC Proceedings, Vol. V, pp. 1473-1477.

Nilda Ruimy and Adriana Roventini. 2005 *Towards the linking of two electronic lexical databases of Italian*, In Zygmunt Veutulani (ed.), L&T'05 -

Nilda Ruimy. 2006. *Merging two Ontology-based Lexical Resources*. LREC Proceedings, CD-ROM, 1716-1721.

Adriana Roventini. 2006. *Linking Verbal Entries of Different Lexical Resources*. LREC Proceedings, CD-ROM, 1710-1715.

Rita Marinelli. 2006. *Computational Resources and Electronic Corpora in Metaphors Evaluation*. Second International Conference of the German Cognitive Linguistics Association, Munich, 5-7 October.