Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 161-167 recibido 02-05-2005; aceptado 01-06-2005

A Proposal for a Shallow Ontologization of Wordnet

Jordi Atserias
Universitat Politècnica de Catalunya
TALP Research Center
batalla@lsi.upc.edu

Salvador Climent
Universitat Oberta de Catalunya
GTL-UOC Group
scliment@uoc.edu

Joaquim Moré
Universitat Oberta de Catalunya
GTL-UOC Group
jmore@uoc.edu

German Rigau
University f the Basque Country
IXA Group
rigau@si.ehu.es

Resumen: En este artículo se presenta el trabajo que se está realizando para la llamada *ontologización superficial* de WordNet, una estructura orientada a superar muchos de los problemas estructurales de la popular base de conocimiento léxico. El resultado esperado es un recurso multilingüe más apropiado que los ahora existentes para el procesamiento semántico a gran escala.
Palabras clave: WordNet, Ontologías, Recursos Léxico-Semánticos

Abstract: This paper presents the work carried out towards the so-called *shallow ontologization* of WordNet, which is argued to be a way to overcome most of the many structural problems of the widely used lexical knowledge base. The result shall be a multilingual resource more suitable for large-scale semantic processing.
Keywords: WordNet, Ontologies, Lexical-Semantic Resources

1. Introduction

Using large-scale lexical-semantic knowledge bases has become a usual practice for most current NLP systems. Building appropriate resources of this nature for broad-coverage semantic processing is very a hard and expensive task. For this reason it is not surprising that most of the recent efforts on this research area reuse already existing large-scale semantic resources, primarily WordNet (Fellbaum, 1998). In order to enrich this resource with new information, some of these efforts use the WordNet hierarchy to expand in a full automatic way semantic properties assigned to a reduced number of high level synsets. This approach has been used for instance for building WordNet Domains (Magnini & Cavaglia, 2000) or the MEANING Top Concept Ontology (Atserias et al. 2004). However, this semiautomatic process (the information is assigned by hand to a limited number of synsets and inherited top-down by automatic means to the rest of WordNet) is not completely correct. By examining a subset of synsets, we realised that there are the following main sources of errors:

- Erroneous hand-made assigments
- Erroneous ISA links which causes erroneous inheritance
- Multiple inheritance cause incompatibilities (Guarino & Welty, 2000)

Martin (2003a) argues that the semantic web will not succeed without a large natural language ontology allowing to share meaning between ontologies. Given available resources and works it seems reasonable to use WordNet as the base for building that natural language ontology. Working in this direction, Martin (2003a) has merged WN 1.7's top level to several top-level ontologies: DOLCE, SUO, Sowa's and DAML. Then he has carried out different types of restructuring in WN 1.7, such as intuitive identifier generation, instance-category discrimination, the fixing of 315 links to overcome lexical-semantic problems and the addition of 161 new links. This is a superb work on WordNet ontologization. Nevertheless, as himself declares, it doesn't go further than that which others have done before: to insert WordNet's top level into another better-structured top level – while the rest, the bulk of WordNet, remains badly structured and showing the set of problems which he and others have pointed out (e.g. Oltramari et al., 2002).

ISSN: 1135-5948

2. Problems in the Structure of WordNet

Taking into account only taxonomic relations, the main problems in the structure of WordNet are the following:

1. There is no distinction between instances and categories;
2. Some specializations (hyponyms) contradict their categories' (hypernyms) nature
3. There are heterogeneous levels of specialization
4. There is no distinction between types and roles
5. The ISA link is used to code other types of relations (e.g. *similar* or *place*); and
6. Exclusivity between categories is not always clear (unclear multiple inheritance).

Martin (2003b) concludes that such lack of structure might be a problem for applications and that fixing it can be as difficult as building up a new WordNet from scratch.

We believe that such an inconsistent WordNet hierarchy is a useless tool for knowledge engineering, since it will cause erroneous transmission of information or from another point of view, erroneous retrieving of concepts.

For instance, we can see that when expanding properties down through the WordNet hierarchy, drug_1 and its subcategories (as anesthetic_1) result to be objects while leaf_1 and subcategories (e.g. dandelion_green_1) are classified as substances.

3. How to Structure Wordnet? Our Approach

Contrary to Martin (2003b) we do think that a more structured version of the WordNets can be achieved. In order to turn WordNet into a more useful tool for NLP applications, we will concentrate on the more serious structural problems (1, 2 and 5), since they violate the nature of the ISA relationship, causing problems during propagation of relations or inferencing.

We are performing a shallow restructuring of WordNet. It is based on blocking inheritance in those edges where subsumption errors show up and then linking chopped off branches to a basic Top Ontology. Multiple links are allowed, as EAGLES (Sanfilippo et al.1999) recommend and Vossen (2001) does. We term it a shallow ontologization as it doesn't reassign links inside WordNet but instead it chops WordNet branches off and links them to a Top Ontology. This is a pragmatic solution to face the problem of the difficulty or impossibility of ontologizing WordNet. We hypothesize that: (a) in many cases such classification would be sufficient for specific purposes of knowledge engineering (as Vossen, 2001 intend to show); or (b) it can help those researchers aiming to embark on full WordNet's ontologization, as they could select coherent groups of their branches as steps to their goal.

We also take advantage of the declaration and classification of Base Concepts: currently, 1600 concepts classified by means of such Top Ontology. It must be noticed that the Base Concepts are a set of relevant synsets not only belonging to WordNet top level. Therefore, the work on classifying the Base Concepts is a way to enter deep into the reorganization of the whole WordNet.

The Top Ontology (subsequently TO) has the advantage of both having definitions for their categories and definitions of their internal incompatibilities. Besides, by the fact of being based on the work of Pustejovsky (1995), it allows to express different facets of the word meaning, since lexicalization links standing between WordNet synsets and the TO can as well be seen as concept features. This way, being "fruit" linked to the TO nodes Comestible, Object and Plant, it can as well be seen and represented as a feature structure:

fruit: *Function: Comestible*
Form: Object
Origin: Natural: Plant

Moreover, this design allows to naturally code dot objects, that is, inherently polysemous words such as 'letter': something that can both be destroyed and carry information ("I burnt the letter of condolence"):

letter: *Function: LanguageRepresentation*
Form: Object

The TO is deliberately simple; it only incorporates distinctions which are basic, intuitive and grounded on notions widely used in linguistic semantics. We think that simplicity and intuitiveness is a prerequisite for an ontology to be actually used by the community. Another virtue of the TO is that it springs up from the lexicon. That is, it has not been

created in advance by theorists but it emerges from lexical concepts' clustering. Therefore, we hypothesize that it can reflect better how the lexicon is and, consequently, be more useful for NLP than other kinds of sets of logical categories – probably more suitable for abstract information categorization.

In the future, once we have fully classified WordNet according to the TO, we intend to evolve to a new TO release incorporating Sanfilippo et al. (1999) expansion and Vossen (2001) theoretic reshuffle which allows for attributing qualia features to 2nd and 3rd Order entities. Sanfilippo's TO is better than the original EuroWordNet TO because it is richer (74 concepts more). Vossen's TO is also better because it is more flexible: for instance it allows to cross-classifying mental entities and situations with features which in the TO you can only attribute to physical entities.

We have not started the work using already an enhanced TO because:

1. We have a large number of synsets encoded by means of TO, 1600, while Sanfilippo *et al.* (1999) only have 164 and there are no public distribution of the work of Vossen (2001).

2. Different to the aforementioned other models, most nodes in the TO are well defined or at least exemplified (Vossen, 1998);

3. Category disjunctions and incompatibilities are explicitly declared in the TO; this allows for finding synsets bearing contradictory information, therefore, presumably, WordNet subsumption failings.

4. The TO is expected to be easily mapped to both EAGLES and Vossen's TO.

In some sense, we prefer not to migrate to a new TO yet to prevent introducing noise into the process; we think it's better to complete and improve the current classification of synsets first and then, once having taken advantage of the experience, decide about the design of the new TO and map the old to the new. Anyway, in our current work we have already been careful to incorporate or maintain compatibility with the EAGLES 164 synset classifications (Sanfilippo,1999, pp. 218-222).

The 3rd reason is key for our work, since it allows us to take advantage now of existent coding contradictions to detect those synsets where we have to place blockings and recoding. The same way Martin's (2003b) 'exclusion links' have led him to the detection of several inconsistencies in WordNet, the TO coding of the Base Concepts allows us to find out many more, and much more deep into WordNet:

- 214 feature conflicts in 49 synsets caused by mistaken hand annotation
- 2247 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 TO features

In spite of such a large number of conflicts, working on the topmost origin of the contradiction results on fixing a lot of cases. For instance, leaf_1, ("the main organ of photosynthesis and transpiration in higher plants") subcategorizes 66 kinds of leaves. It was categorized as Substance, but it seems clear that such TO concept can not apply to it. Consequently, removing the link from leaf_1 to Substance results in 66 conflicts fixed.

In the EWN Project, some mistakes where made when linking Base Concepts to the TO. Normally, such errors where due to false intuitions. However, in many cases they also correspond to inconsistencies in WordNet. As told above, we precisely are exploiting such contradictions to reach our goals. Checking contradictions in the coding allows for detecting those synsets where ontological doubts or WordNet classification problems show up. Basically, there are three types of feature conflicts:

a) internal: incompatible manual classifications
b) caused by simple inheritance
c) caused by multiple inheritance

Usually (a) points out to synsets causing ontological doubts – e.g. "skin", is it Object or Substance? On the other hand, (b) and (c) usually show classification mistakes in WordNet.

TO feature incompatibilities are a powerful tool for detecting structural inconsistencies in WordNet. For instance, artifact_1 is glossed as "a man-made object" thus it is quite obvious to find it classified the TO concept Object. Its hyponym drug_1 thus inherits Object but, besides, it was (correctly) hand-classified as Substance. The incompatibility between

Object and Substance, allows for the automatic detection of this structural inconsistency of WordNet –which is corroborated by the fact that drug_1 subcategorizes substances, e.g. aborticide_1 or anesthetic_1.

Our work will probably be the second one to ontologize all WordNet after that of Niles and Pease (2003) with SUMO. However, our coding: (i) will not be simple but multiple (SUMO links every synset to only one node of the ontology); and (ii) uses a more intuitive, simple and workable TO.

4. The Case of Body_Covering

In this section a case of complete restructuring of a semantic field using the procedure described above is presented. It is the case of the hyponyms of **body_covering_**1.

In 4.1 we'll see first the taxonomy as it is presented in WN1.6[1]. Then in 4.2 we'll see how we use blocking on the TO classification to get a more accurate representation of the meaning of each synset. Last, in 4.3 we'll show several examples of concept clustering can be obtained by using the resulting shallow ontology.

4.1. The WN1.6 Taxonomy

This classification (see Figure 1) shows many problematic and controversial hyponymy links. First of all, this taxonomy embeds both bounded and unbounded concepts. For instance, **down_1** and **sickle feather_1** are cohyponyms, and **feather_1** and **plumage_1** are supposed to be synonyms. We think that in these cases the correct interpretation would be to have the bounded concepts as some kind of parts of the unbounded (feathers part of plumage, prepuce part of skin) not as subtypes or synonyms. The bounded-unbounded distinction is intended to be captured by the exclusive TO features *Object* and *Substance*. According to the definitions of the EuroWordNet Top Ontology, *Object* lexicalizes "Any conceptually-countable concrete entity with an outer limit"; and *Substance*: "All stuff without boundary or fixed shape, considered from a conceptual point of view not from a linguistic point of view". A taxonomy mixing both sorts of concepts also mixes their corresponding features thus spreading them down by inheritance and causing multiple contradictions.

[1] For the sake of simplicity synsets are represented only by one variant or synonym.

Another important problem of the body_covering_1 taxonomy is that it mixes natural objects and artifacts. It is counterintuitive to have both dewlap_1 and heel_4 (in the sense of the part of a shoe) coexisting as subtypes of skin, while it seems that pieces of leather are skin in the same sense than a table is wood. They are not subtypes but some kind of made-of.

Last, it is also shocking to have hairdo_1 and hyponyms as subtypes of hair_1 as long as in fact they are ways of arranging it.

```
body_covering_1
    skin_4
    plumage_1 feather_1
            down_1
            sickle_feather_1
    protective_covering_2
    skin_1
            pellicle_1
            dewlap_1
            prepuce_2
            scalp_1
            animal_skin_1
                    parchment_2
                    leather_1
                            piece_of_leather_1
                                    heel_4
                                    toe_2
                            cordovan_1
                    fur_1
                            bearskin_1
                            lapin_1
    hair_1
            coat_3
            hairball_2
            mane_1
            beard_3
            postiche_1
            hairdo_1
                    afro_1
            pubic_hair_1
            eyebrow_1
            eyelash_1
```

Figure 1: Taxonomy of **body_covering** in WN1.6

4.2 Blocking points and classification using EWNTO

See in Figure 2 the resulting taxonomy using blocking points (notated 'x') and new TO feature assignments. The symbol "+" stands for inherited and the symbol "=" for assigned.

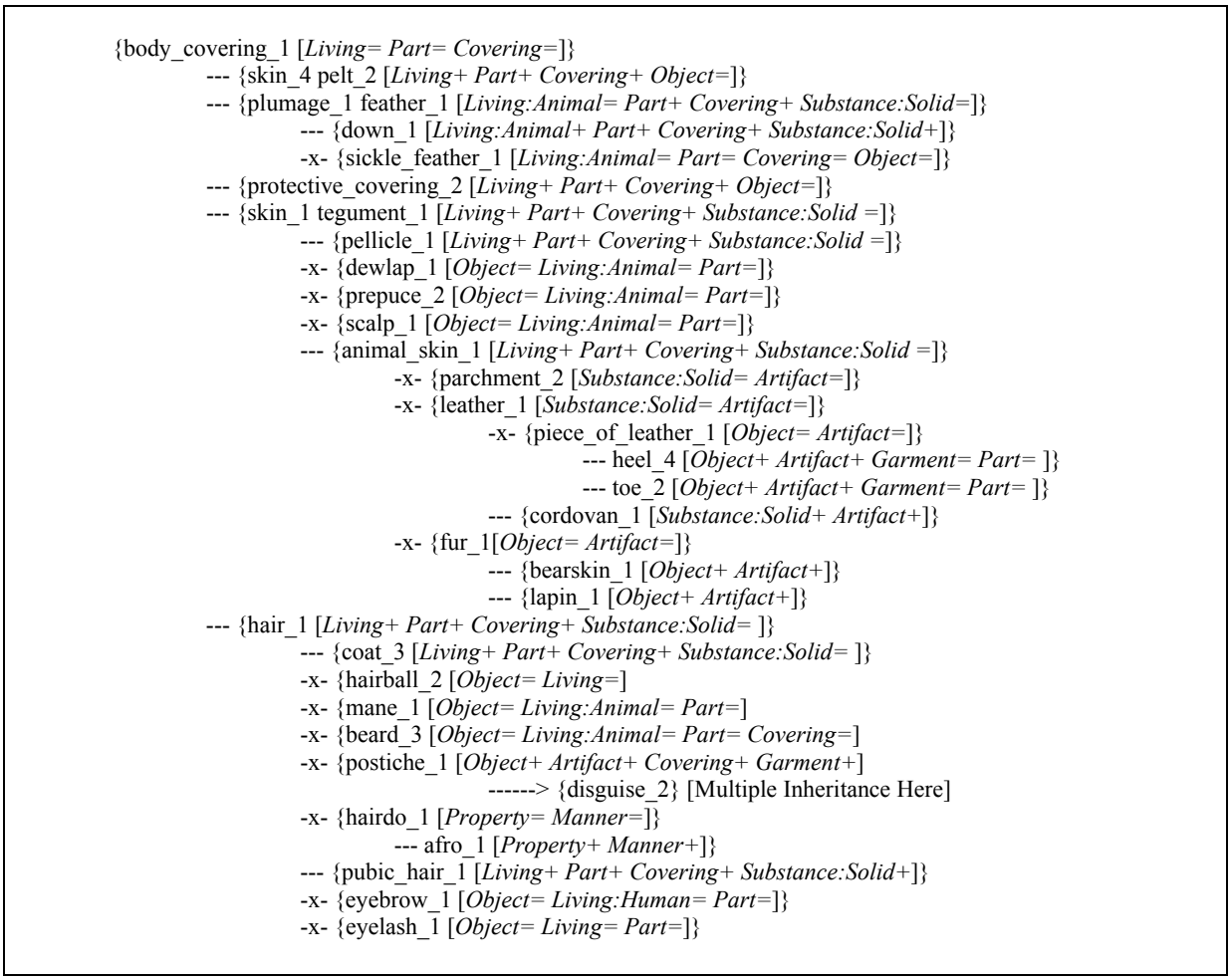

Figure 2: Resulting taxonomy for *body_covering*

The resulting taxonomy is product of the following main decisions.

- The top of the taxonomy, **body_covering_1**, is left underspecified for *Object* or *Substance* and as well underspecified for being part of either animal or plant. Its hyponyms will further incorporate such distinctive information.

- **{plumage_1 feather_1}** is not correct, since it joins as synonyms a mass whole and its count part, i.e., as told above, feathers are parts of the plumage, and therefore the concepts are not synonyms. Looking at the gloss and most of the synset relations, we assume for the synset the mass meaning. When a hyponym is countable, as **sickle_feather_1**, we will block the subsumption relation and code the hyponym as *Object*.

- The synset structure of the two "skins", **{skin_1 tegument_1}** and **{skin_4 pelt_1}** forces the interpretation of the former as mass and the later as countable. Consequently, countable hyponyms of **skin_1** need a blocking.

- There are only made-of-skin artifacts (**leather_1**, **parchment_2**) under **animal_skin_1**. This forces massive blocking of the relationship between it and all their hyponyms. Indeed the solution looks forced, but it is unavoidable as long as the glosses and other clues show clearly that **animal_skin_1** is not conceived by WordNet as an artifact.

- **Piece_of_leather_1** is subsumed by **leather_1**, in another case of false subsumption, since the relationship should be undoubtfully 'made-of'.

- **heel_4** and **toe_2** inherit the properties of **piece_of_leather_1**, and add to it further information capturing the idea that they are a part of a piece of garment

4.3 Examples of clustering and retrieval

As mentioned before, multiple classification using a TO allows for tailoring WordNet, i.e. retrieving clusters or branches according to specific needs. We show in Figure 3 some examples using the work presented in the previous section. At the top, in italics, there is the TO node or set of TO nodes used for the retrieval. Below, the subtrees linked to it or them.

As it can be seen, retrieving the hyponyms of body_covering which are simultaneously linked to Living Part and Covering gives those parts of living beings that serve for covering them. Retrieving those linked to Artifact gives things which, being parts of animals in origin, have been treated by men to obtain artifacts. By using the features Artifact, Garment and Part we can retrieve those synsets that are a part of a piece of garment. Object, Living and Part gives bounded covering parts of the body.

This is an example of how multiple TO classification becomes a tool to tailoring WordNet in different ways, thus getting coherent clusters, and to attributing semantic features to concepts and words in a way that they can be more useful for semantic processing purposes than using the original WordNet taxonomy.

5. Some Figures

We started to work using a WN1.6 version annotated by hand with 2.696 TO features which expanded by inheritance to 253.003 features. At this moment, as we have worked on 47 Base Concepts we have 2.756 hand-coded features which expand to 276.384. Currently, 52 blocking points have been assigned [2].

Comparing both versions:

1- Both versions share

- 2.676 hand-coded labels (corresponding to 1.013 different synsets)

- 51.043 expanded labels (corresponding to 36.289 different synsets)

2- Differences

- The initial version had 201.960 expanded labels belonging to 75.052 synsets which now are not present. The new version has 225.341 new expanded labels, belonging to 75.295 synsets.

Living Part Covering
- skin_4
- plumage_1 feather_1
  - down_1
- protective_covering_2
  - armor_2
  - scute_1
  - sclerite_1
- hair_1
  - coat_3
  - pubic_hair_1
- skin_1
  - pellicle_1
  - animal_skin_1
- sickle_feather_1
- dewlap_1
- prepuce_2
- scalp_1

Artifact
- parchment_2
- leather_1
  - cordovan_1
- piece_of_leather_1
  - heel_4
  - toe_2
- fur_1
  - bearskin_1
  - lapin_1

Artifact Garment Part
- heel_4
- toe_2

Object Living Part
- skin_4
- sickle_feather_1
- protective_covering_2
- dewlap_1
- prepuce_2
- scalp_1
- mane_1
- beard_3
- eyebrow_1
- eyelash_1

Property Manner
- hairdo_1
  - afro_1

Figure 3: Examples of semantic clusters obtained by applying the method

[2] The current state of the work can be consulted at: http://nipadio.lsi.upc.es/cgi-bin/wei4/public/wei.consult.perl

6. Conclusions and Further Work

This work is still too preliminary to be quantitatively assessed. Nevertheless, it appears to be clear that exploiting feature contradictions is an effective method for detecting and fixing ontological mistakes in the deep of WordNet, while most of the previous works have only worked in the ontologization of its upper levels – the top of the iceberg.

From now on our main goal is to go on improving WordNet until eliminating all TO feature conflicts. We will also use the 315 WordNet inconsistencies detected by Martin (2003b). Moreover, we also intend to mark up the difference between instances and categories, using Martin's (2003b) list of individuals.Once achieved we plan to work on an enhanced design of a new TO in order to envisage which semantic features are or not really useful for NLP applications.

Another work to carry on in the future is to draw a new set of Base Concepts. In many cases, current Base Concepts only qualify for a few top concepts rather than for rich combinations, as they necessarily have to be abstract and general. Probably, new Base Concepts will belong to lower points in the taxonomy, but will maximize the concentration of semantic information, i.e. TO features. After that, we will assess the impact of using the enhanced classification on real NLP tasks.

REFERENCES

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

Cruse D.A. 1986 *Lexical Semantics*. Cambridge University Press. Cambridge.

Fellbaum Ch. (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Guarino N. 1998. Some Ontological Principles for Designing Upper Level Lexical Resources. In *Proceedings of the First International Conference on Language Resources and Evaluation*, *LREC*, Granada: 527-534.

Guarino, N. and C. A. Welty 2000. A formal ontology of properties. In *Proceedings of ECAI'2000 Workshop on Knowledge Acquisition, Modeling and Management*

Magnini, B. & G. Cavaglia. 2000. Integrating subject field codes into Wordnet. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC*, Athens.

Martin P. 2003a Knowledge Representation, Sharing and Retrieval on the Web. In N. Zhong, J. Lin, Y. Yao Eds., *Web Intelligence*. Springer-Verlag.

Martin Ph. 2003b. Correction and Extension of WordNet 1.7. In *Proceedings of ICCS 2003, 11th International Conference on Conceptual Structures* (Springer Verlag, LNAI 2746, pp. 160-173), Dresden.

Niles, I., and Pease, A., 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pp 412-416.

Oltramari A., A. Gangemi, N. Guarino C. Masolo 2002 Restructuring WordNet's Top-Level: The OntoClean approach. In *Proceedings of LREC 2002* (OntoLex workshop). Las Palmas, Spain.

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

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

Vossen P. (Ed.) 1998 *EUROWORDNET: A multilingual Database with Lexical Semantic Networks*. Kluwer, Dordrecht

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