A proposal for improving WordNet Domains

Aitor González-Agirre, Mauro Castillo*, German Rigau

IXA group UPV/EHU, Donostia Spain, Donostia Spain
agonzalez278@ikasle.ehu.com, german.rigau@ehu.com
*UTEM, Santiago de Chile, Chile, mcast@informatica.utem.cl

Abstract

WordNet Domains (WND) is a lexical resource where synsets have been semi-automatically annotated with one or more domain labels from a set of 165 hierarchically organized domains. The uses of WND include the power to reduce the polysemy degree of the words, grouping those senses that belong to the same domain. But the semi-automatic method used to develop this resource was far from being perfect. By cross-checking the content of the Multilingual Central Repository (MCR) it is possible to find some errors and inconsistencies. Many are very subtle. Others, however, leave no doubt. Moreover, it is very difficult to quantify the number of errors in the original version of WND. This paper presents a novel semi-automatic method to propagate domain information through the MCR. We also compare both labellings (the original and the new one) allowing us to detect anomalies in the original WND labels.

Keywords: WordNet, WordNet Domains, Lexical Semantics

1. Introduction

Building large and rich knowledge bases is a very costly effort which involves large research groups for long periods of development. For instance, hundreds of person-years have been invested in the development of wordnets for various languages (Vossen, 1998).

WordNet Domains[1] (WND) is a lexical resource where synsets have been semi-automatically annotated with one or more domain labels from a set of 165 hierarchically organized domains (Magnini and Cavaglià, 2000; Bentivogli et al., 2004). WND allows to reduce the polysemy degree of the words, grouping those senses that belong to the same domain (Magnini et al., 2002).

But the semi-automatic method used to develop this resource was not free of errors and inconsistencies. By cross-checking the ontological content of the MCR it is possible to find some of these problems. For instance, noun synset <diver_1 frogman_1 underwater_diver_1> defined as *someone who works underwater* has domain *history* because it inherits from its hypernym <explorer_1 adventurer_2>.

We suggest a novel graph-based approach for improving WND. As a result we obtained a new semantic resource derived from WordNet Domains and aligned to WordNet 3.0.

After this short introduction, Section 2. describes a very simple method of inheritance used to fill the gaps that have arisen due to the porting process from WordNet 1.6 to 3.0. In section **??** we describe our novel graph-based method, based on the UKB algorithm, used to generate new domain labels aligned to WordNet 3.0. Section proposes a discussion about the new labelling. Finally, section 4. presents an example of how to evaluate in a semi-automatic way the quality of the domain labels assigned in the original WND.

2. Domain inheritance

WND was developed using WordNet 1.6. One consequence of the automatic mapping that we used to upgrade version 1.6 to 3.0 is that many synsets were left unlabeled (because there are new synsets, changes in the structure, etc.).

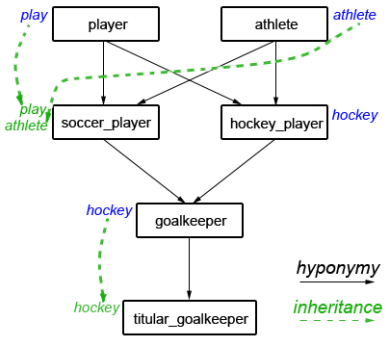

Figure 1: Example of inheritance of domain labels.

One of the first tasks undertaken has been to fill these gaps. For them, we have carried out a propagation of the labels by inheritance of nominal and verbal synsets. In WordNet, the adjectives are organized in terms of binary oppositions (antonymy) and similarity of meaning (synonymy). The structure of WordNet for adjectives and adverbs makes this spread not trivial. Therefore this simple process has been not carried out neither for adjectives nor for adverbs.

Consider the example shown in Figure 1. For nouns and verbs, we have worked on the assumption that synsets are mostly correctly labeled, and therefore we have worked exclusively on those synsets that had no labels at all. We inherited the label or labels from its hypernyms. If a synset has more than one hypernym, the domain labels are taken from all of them. During this phase has been taken into account the incompatibility between domains, preventing the same synset can be both *factotum* and *biology*.

This process increased our domain information by nearly a 18-19%, as shown in Tables 1 and 2:

However, this process may also have propagated innapropriate domain labels to unlabeled synsets. It remains for

[1] http://wndomains.fbk.eu/

PoS	Before	After	Increase
Nouns	66,595	83,286	+25%
Verbs	12,219	14,224	+16%
All	100,315	119,011	+19%

Table 1: Number of synsets with domain labels.

PoS	Before	After	Increase
Nouns	87,938	108,665	+24%
Verbs	13,026	15,051	+16%
All	124,551	146,899	+18%

Table 2: Total number of domain labels.

future research an accurate evaluation of this new resource. In the next section we present some examples using a new graph-based method for propagating domain labels through WordNet. Additionaly, the method can also be used to detect anomalies in the original WND labels.

3. A new graph based method

UKB[2] algorithm (Agirre and Soroa, 2009) applies personalized PageRank on a graph derived from a wordnet. This algorithm has proven to be very competitive on Word Sense Disambiguation tasks and it is easily portable to other languages that have a wordnet (Agirre et al., 2010). Now, we present a novel use of the UKB algorithm for propagating information through a wordnet structure.

Given an input context, '*ukb_ppv*' (*Personalized PageRank Vector*) algorithm outputs a ranking vector over the nodes of a graph, after applying a *Personalized PageRank* over it. We just need to use a wordnet as a knowledge base and pass to the application the contexts we want to process, performing a kind of *spreading activation* through the structure of a wordnet.

As a context we used those synsets labelled with a particular domain. Thus, for each of the 169[3] domain labels included in the MCR we generated a context. Each file contains the list of offsets corresponding to those synsets with a particular domain label. After creating the context file, we just need to execute '*ukb_ppv*' that will return a ranking of the weights for each wordnet synset with respect to that particular domain.

Once made the process for all domains we will have the weight of each synset for each of the domains. Therefore, we know which are the highest weights for each domain and the highest weights for each synset. This allows us to estimate which synsets are more representative of each domain (those who have more weight in the ranking) and which domains are best for each synset (those who have attained a higher weight for that synset).

Basically, what we do is to mark some synsets with a domain (using the labels we already know from the original porting process) and use the wordnet graph to propagate the new labelling. We work on the assumption that a synset directly related to several synsets labelled with a particular domain (i.e *biology*) would itself possibly be also related somehow to that domain (i.e. *biology*). Therefore, it makes no sense to use the domain *factotum* for this technique.

3.1. Propagating domain labels

We have generated two different knowledge bases. The first one only contains the original WordNet relations. The second one, also contains the relationships between glosses, increasing the size and richness of the knowledge base. Instructions for preparing the binary databases for UKB using WordNet relations are inside the downloadable file[4] of the UKB package.

It has been necessary to generate a context file for each domain. Generating a context is as simple as creating a text file with the synset offsets that have the domain label. An example of a context file for the *rugby* domain can be seen in Figure 2. We can see a list of offsets representing synset of the Table 3.

Synset	Variants
eng-30-00136876-n	goal-kick
eng-30-00242146-n	scrum, scrummage
eng-30-00470966-n	rugby, rugby_football, rugger
eng-30-00471277-n	knock_on
eng-30-01148101-v	hack
eng-30-01148199-v	hack
eng-30-04118538-n	rugby_ball

Table 3: List of synset with "*rugby*" as domain label.

One of the problems that comes up when analyzing the results is that the own domain labels of a synset have an unbalanced weight on the final ranking of that synset. Almost always the own labels of a synset appear in the top positions. In order to avoid this undesired effect, we generated new contexts, specific to each synset, and each domain. Thus, a synset can not vote for its own domains and only the rest of synsets decide the final weights of the ranking.

3.2. Post-processing

Once generated the context files, the UKB algorithm is executed. The result is a list with the weight for each synset for a domain. The next step is to sort the file by weight, highlighting those synsets that are more representative of the domain (Figure 3).

Furthermore, we can sort the result by synset. This allows us to, once we have a file for each domain, put them together in a matrix. Each line of this matrix will represent a synset, and the columns will be weights corresponding to each domain. The highest values of a line (synset) will be the more representative domains for that synset.

Table 4 shows the first ten domains and weights resulting from the application of this method on synset $<\text{diver}^1_n$ frogman^1_n $\text{underwater\_diver}^1_n>$ originally labeled as *history*, which seem to be incorrect. The suggestions of the algorithm seems to improve the current labeling because it suggests *sub* (possibly the best one) and *diving* (possibly, the second best option). Moreover, the method suggests the wrong label with a much lower weight.

[2] `http://ixa2.si.ehu.es/ukb/`

[3] Excluding *factotum* labels.

[4] http://ixa2.si.ehu.es/ukb/

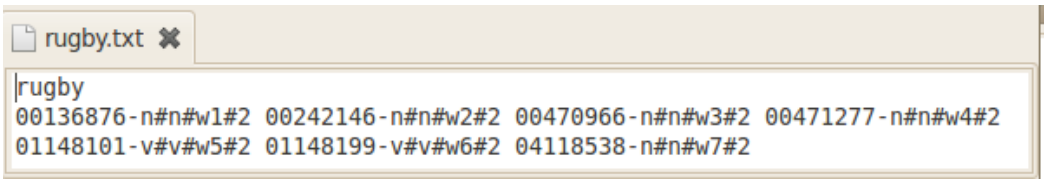

Figure 2: View of the format of a context file.

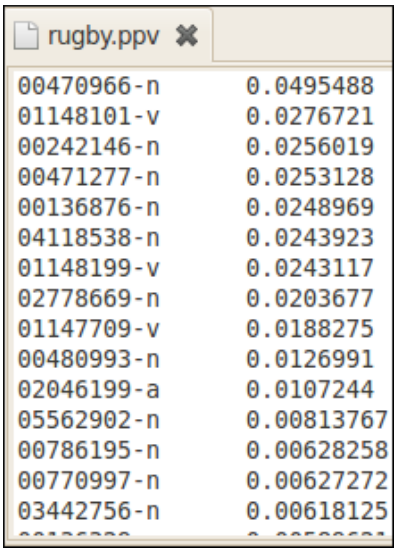

Figure 3: Result of a PPV ranking sorted by weight (only the first lines are shown).

Weight	Domain
0.0144335:	sub
0.0015939:	diving
0.0001725:	swimming
0.0001297:	history
0.0000557:	nautical
0.0000529:	fashion
0.0000412:	jewellery
0.0000315:	ethnology
0.0000274:	archaeology
0.0000204:	gas

Table 4: PPV weight rankings for sense $diver^1_n$.

4. Analyzing ranking changes

It seems that the algorithm is able to generate a ranking in which the most appropriate labels obtain larger weights and also that avoiding the own labels of a synset reduces the weights for incorrect domain labels.

In the next experiment we study how to evaluate in a semi-automatic way the quality of the original labelling. To do that we check the domain labels of the synsets, taking into account the position they occupy in the weight vector. If a synset has '*n*' domain labels, the displacement is calculated for every label. For example, if a synset has two labels and one of the domains occupies the first position and the other the third one, they receive an offset of +0 and +1 respectively. That is, we calculate how many positions they moved from its original place. All those labels with an offset of six or greater are considered in the same group. Possibly, this test will allow us to discover wrong labeled synsets (or at least delimit the search) or to create a group of labels with a high value of reliability.

Therefore we tested the process for each PoS. The results obtained are in the Table 5.

Detecting the labels that have been displaced six or more positions (Table 5) allows us to recognize possible synset that have been labeled incorrectly. An example can be seen in Table 6.

Results for 'ili-30-00747215-n':

- **Variants:** pornography_1 porno_1 porn_1 erotica_1 smut_5
- **Gloss:** creative activity (writing or pictures or films etc.) of no literary or artistic value other than to stimulate sexual desire
- **Domains:** law

Method WN+gloss	
Weight	**Domain**
0.000123453:	sexuality
0.000112444:	cinema
0.000077780:	theatre
0.000075525:	painting
0.000062377:	telecommunication
0.000060640:	publishing
0.000050370:	psychological_features
0.000047003:	photography
0.000046853:	artisanship
0.000040458:	graphic_arts

Table 6: Method WN+gloss: UKB weight rankings for sense 1 of "*porno*".

The example in Table 6 shows how the label *law* (incorrectly assigned) disappears from the first ten positions of the list. Instead, the algorithm suggests *sexuality* and *cinema*, which in this case seems to be much more appropriate.

5. Discussion

After aplying our novel method for propagating domain information through WordNet, we obtained a new distribution for the domain labels. We present the distribution of domain labels for the original WND (Figure 4), applying

	Offset						
PoS	**0**	**1**	**2**	**3**	**4**	**5**	**6+**
Nouns	55.52%	18.51%	10.06%	5.19%	1.95%	1.95%	6.82%
Verbs	40.46%	15.95%	13.39%	7.69%	4.56%	0.85%	17.09%
Adjectives	51.04%	21.35%	8.85%	2.60%	1.56%	4.17%	10.42%
Adverbs	60.40%	13.86%	5.94%	0.99%	4.95%	2.97%	10.89%
Total	**54.48%**	**18.60%**	**10.07%**	**5.04%**	**2.06%**	**2.10%**	**7.65%**

Table 5: Method WN+gloss: Displacement of domain labels regarding their current position (separated by PoS).

the PPV algorithm using WN as a graph (Figure 5), and applying the PPV algorithm using WN+gloss as a graph (Figure 6). Since some of the percentages are too high, the tables only presents distributions of 5% at most. The domains that have percentajes exceeding 5% are the following: *biology* (14.72% for WND), *animals* (6.59% for WND), *plants* (6.07% for WND) and *plastic arts* (6.42% for PPV using WN).

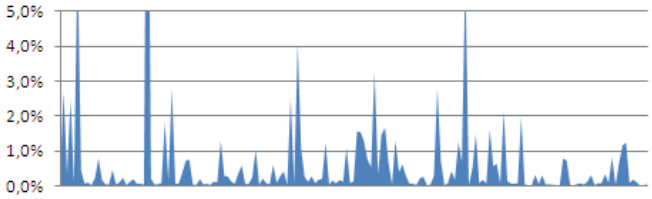

Figure 4: Distribution of original WND labels in alphabetical order (left to right).

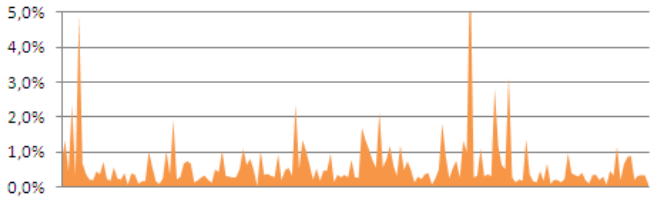

Figure 5: Distribution of new domain labels using PPV with WN as KB, in alphabetical order (left to right).

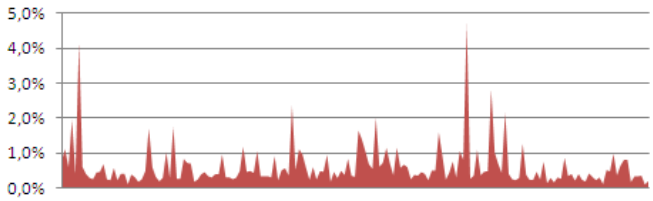

Figure 6: Distribution of new domain labels using PPV WN enriched with gloss relations as KB, in alphabetical order (left to right).

Aparently, the number of synsets per domain has been smoothly distributed across the domains in both PPV propagations. Some of the most frequent domains are now much less frequent. In contrast, many domains with very few synsets are now much better represented. This effect is a consecuence of the PPV algorithm which normalizes de page rank vector: the sum of all vector weights is one for every domain.

We can also analize which domains have modified its representation. Figures 7 (applying PPV using WN as KB) and 8 (applying PPV using WN enriched with glosses as KB) show the total percentage of increment/decrement of representation for each domain with respect to the original domains. As can be seen, in both cases, most of the domains appear to be in the positive side (top), indicating that they increased their representation.

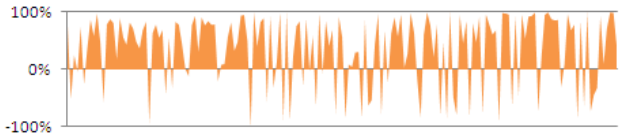

Figure 7: Percentage increment/decrement for domains applying PPV using WN, compared to the original domains.

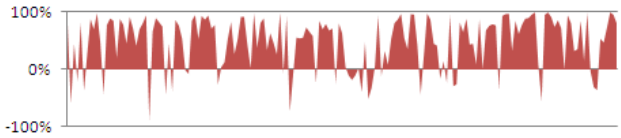

Figure 8: Percentage increment/decrement for domains applying PPV using WN+gloss, compared to the original domains.

If most of the domains have increased their representation, necessarily a few domains have decreased much of their representation. Figures 9 (for PPV and WN) and 10 (for PPV and WN+gloss) show respectively the five domains that have increased more their representation, and the five ones that have decreased more their representacion.

Using WN as a knowledge base, the five domains showing the largest increment of representation are *plastic_arts*, *veterinary*, *sub*, *topology* and *philately*. Possibly, these domains are now over represented. The five domains showing the largest decrement are *gastronomy*, *geography*, *religion*, *biology*, *fashion*. Possibly, these domains were over represented in the original labelling.

A very similar behaviour is observed using WN enriched with the gloss relations. The five domains showing the

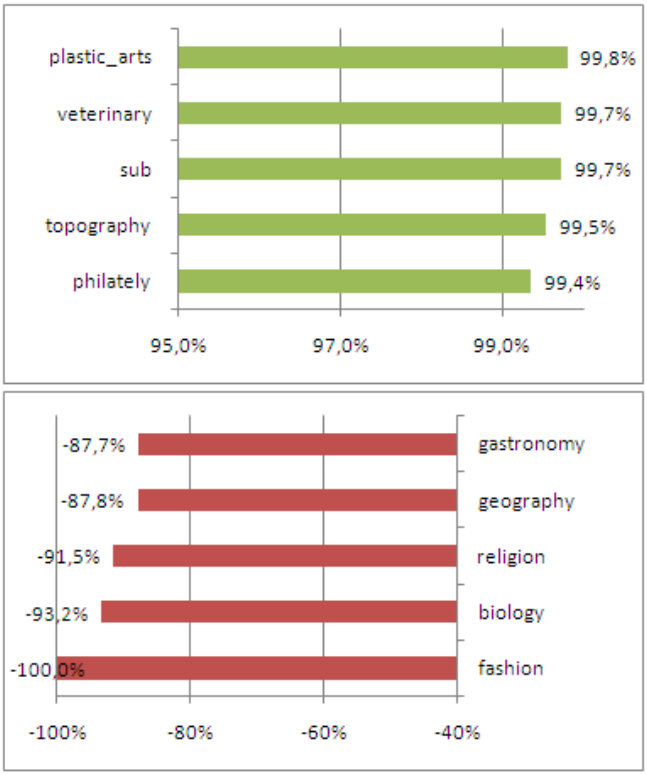

Figure 9: Percentage of increment (green) and decrement (red) for the five more affected domains using PPV and WN as KB

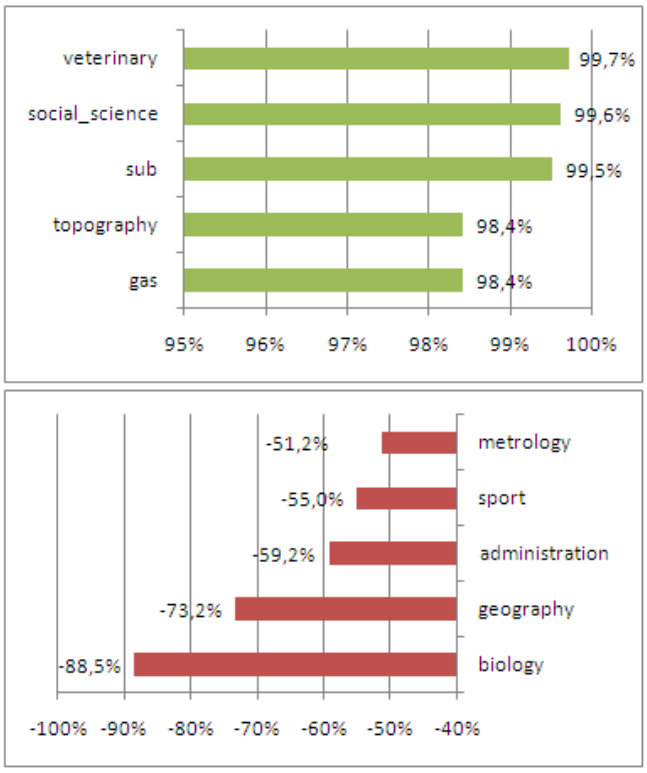

Figure 10: Percentage of increment (green) and decrement (red) for the five more affected domains using PPV and WN enriched with gloss relations as KB

largest increment of representation are *veterinary*, *social_science*, *sub*, *topography* and *gas*. Possibly, these domains are now over represented. In fact, some of them seems to be over represented also in the previous labelling (*veterinary* and *sub*). The five domains showing the largest decrement are *metrology*, *sport*, *administration*, *geography*, *biology*. Possibly, these domains were over represented in the original labelling. In fact, some of them seems to be over represented also in the original labelling (*geography* and *biology*).

These imbalances should need to be further studied in a near future.

6. Concluding Remarks

We have presented a new robust graph-based method which propagates domain information through WordNet. Firstly, we described a simple inheritance mechanism to complete unlabelled synsets from WordNet 3.0. Secondly, we provide some examples of the new domain labellings focussing on those synsets which provided larger variations.

After these initial qualitative tests, we drawn some preliminary conclusions:

1. The propagation method seems to provide some interesting results which deserve more research.

2. The gloss relations seems to provide useful knowledge for the propagation of domain labels through WordNet.

Obviously, some improvements and further investigation are needed with these new resources. For instance, we need to develop an automatic method to select which label or labels finally assign to a particular synset. Moreover, not all domains affect in the same way due to its initial distribution through the WordNet structure. We also need to investigate different combinations of relations for creating the knowledge base used by UKB. For instance, using only gloss relations, or a particular subset of WordNet relations.

We also plan to try different combinations of methods and resources to improve the final result. For instance, we also plan to derive domain information from Wikipedia by exploiting WordNet++ (Navigli and Ponzetto, 2010).

We already have carried out an empirical evaluation in a common Word Sense Disambiguation task. On this task, the new labeling clearly outperforms by a large margin the original WordNet Domains (Gonzalez-Agirre et al., 2012). Additionaly, we need to empirically evaluate the new WordNet Domains in some additional semantic tasks.

7. Acknowledgements

We thank the IXA NLP group from the Basque Country University. This work was been possible thanks to its support withing the framework of the KNOW2 (TIN2009-14715-C04-04) and PATHS (FP7-ICT-2009-6-270082) projects. We also wish to thank the reviewers for their valuable comments.

8. References

Agirre, E., Cuadros, M., Rigau, G., and Soroa, A. (2010). Exploring Knowledge Bases for Similarity. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10). European Language Resources Association (ELRA). ISBN: 2-9517408-6-7. Pages 373–377.".*

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In *in Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources*, pages 101–108.

Gonzalez-Agirre, A., Castillo, M., and Rigau, G. (2012). A graph-based method to improve WordNet Domains. In *Proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING'12)*, New Delhi, India.

Magnini, B. and Cavaglià, G. (2000). Integrating subject field codes into WordNet. In *Proceedings of the Second Internatgional Conference on Language Resources and Evaluation (LREC)*, Athens. Greece.

Magnini, B., Satrapparava, C., Pezzulo, G., and Gliozzo, A. (2002). The Role of Domains Informations. In *In Word Sense Disambiguation*, Treto, Cambridge.

Navigli, R. and Ponzetto, S. P. (2010). Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

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