Automatic Assignment of Domain Labels to WordNet*

Mauro Castillo[1], Francis Real[1], and German Rigau[2]

[1] Universitat Politècnica de Catalunya
Jordi Girona Salgado 13, 08034 Barcelona, Spain
Email: castillo@lsi.upc.es, fjreal@lsi.upc.es
[2] Basque Country University
649 Posta Kutxa, 20080 Donostia, Spain
Email: rigau@si.ehu.es

Abstract. This paper describes a process to automatically assign domain labels to WordNet glosses. One of the main goals of this work is to show different ways to enrich sistematically and automatically dictionary definitions (or gloses of new WordNet versions) with MultiWordNet domains. Finally, we show how this technique can be used to verify the consistency of the current version of MultiWordNet Domains.

1 Introduction

Although the importance of WordNet (WN) has widely exceeded the purpose of its creation [12], and it has become an essential semantic resource for many applications [11,1], at the moment is not rich enough to directly support advanced semantic processing [6].

The development of wordnets large and rich enough to semantically process non-restricted text keeps on being a complicated work that may only be carried out by large research groups during long periods of time [4,2,3].

One of the main motivations of this work is to semantically enrich WN (or other lexic resources like dictionaries, etc.) with the semantic domain labels of *MultiWordNet Domains (*MWND*)* [8]. This resource has proved his utility in word domain disambiguation [7].

The work presented in this paper explores the automatic and sistematic assignment of domain labels to glosses and dictionary definitions.

This methodology may be also used to correct and verify the suggested labeling. It may also provide new cues to assign domain labels in dictionary definitions or in free texts.

This paper is organized as follows: section 2 introduces MWND, section 3 summarizes the experimental work carried out, section 4 is devoted to the the evaluation and results of the experiments and section 5 provides an in deep analisys of the experimental results. Finally, in section 6 some concluding remarks and future work are presented.

2 Semantic Resources

MWND [7] is a lexical resource developed in ITC-IRST where the *synsets* have been annotated semiautomatically with one or more domain labels. These domain labels are

* This paper had partially financed by the European Comition (MEANING IST-2001-34460), Generalitat de Catalunya (2002FI 00648) y UTEM-Chile.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 75–82.

organized hierarchically. These labels group meanings in terms of topics or scripts, e.g. Transport, Sports, Medicine, Gastronomy. which were partially derived from the Dewey Decimal Classification[3]. The version we used in these experiments is a hierarchy of 165 Domain Labels associated to WN1.6. Information brought by Domain Labels is complementary to what is already in WN. First of all Domain Labels may include synsets of different syntactic categories: for instance MEDICINE groups together senses from nouns, such as *doctor* and *hospital*, and from verbs such as *to operate*. Second, a Domain Label may also contain senses from different WN subhierarchies. For example, SPORT contains senses such as **athlete**, deriving from *person*, **game equipment**, from *artifact*, **sport** from *act*, and **playing field**, from *location*.

The table 1 shows the distribution of the number of domain labels per *synset*. This table also shows that most of the *synsets* have only one domain label.

Table 1. Distribution of domain labels for *synset* and distribution of *synset* with and without the domain label factotum in WN

domain labels for synset					
#	**noun**	**verb**	**adj**	**adv**	**%**
1	56458	11287	16681	3460	88.202
2	8104	743	1113	109	10.105
3	1251	88	113	6	1.4632
4	210	8	8	0	0.2268
5	2	1	0	0	0.0030

distribution synsets with CF and without SF factotum			
POS	**CF**	**SF**	**%fact**
noun	66025	58252	11.77
verb	12127	4425	63.51
adj	17915	6910	61.42
adv	3575	1039	70.93

On average, a noun synset has 1.170 domain labels assigned, a verbal synset 1.078, an adjectival synset 1.076 and and adverb synset 1.033.

When building MWND, any labels were assigned in high levels of the WN hierarchy and were automaticaly spread across the hypernym and troponym hierarchy. To our knowledge, a complete verification has not been made, neither automatic nor manual, of the whole set of assignments of domains to synsets.

The domain label *Factotum* includes two types of *synsets*:

Generic synsets: Used to mark the senses of WN that do not have a specific domain, for instance: person, dates, etc.

Stop Senses: The *synsets* that appear frequently in different contexts, for instance: numbers, colours, etc...

Table 1 shows the percentage of factotum labels for nouns, verbs, adjectives and adverbs in WN1.6. There is a high percentage of *synsets* labeled as factotum, except in nouns.

Recently, Domain information has been proven to be useful in many semantic applications. For instance, in Word Sense Disambiguation task (WSD), [5] emphasize the rol of domains. [9] introduce Word Domain Disambiguation (WDD) as a variant of WSD where for each word in a text the domain label is selected instead of the sense label (synset). In addition, MWND have been also used [10] in tasks such as “Text Categorization” (TC).

[3] `http://www.oclc.org/dewey`

3 Experiments

Even though MWND is a useful resource, it was semiautomatically constructed and it needs to be either manually or automatically validated. This validation would allow to study the domain label assignments to synsets of WN1.6 and acquire the implicit models of the domain assignment to glosses. With these models others resources as dictionaries or other WN versions without domains may be labeled. The main goals of the experiments described in this paper were:

- To study new automatic, consistent and robust procedures to assign domains labels to the WN1.6 glosses (or other versiones of WN), or to other definitions of generic dictionaries.
- To study new validation procedures of the consistency of the domain assignment in WN1.6, and especially, the automatic assigment of the factotum labels.

For the experiments, an small set of synsets (around 1%) was randomly selected as a test set and the other synsets were used as a training set (647 noun with 11.9% factotum and 121 verb with 60.33% factotum)

3.1 Labeling methodology

As a first attempt, we studied the performance of the automatic labeling metodology described in [13]. Rigau et al. used WN and a Spanish/English bilingual dictionary to automatically label a Spanish mononlolingual dictionary with WN Semantic Fields (or Lexicographic files).

We can use different similarity measures to obtain the importance (or saliency) of each word with respect each domain.

Using the salient words per domain gathered in the previous step, we can label each gloss again. When any of the salient words of a domain appears in a gloss, there is evidence that the gloss belongs to a particular domain. If several of these words appear, the evidence for that domain grows. Adding together their weights, over all words in a gloss, a program can determines the domain for which the sum is greatest. Thus, this automatic process depends on:

- The **similarity measure** used to assign domain weights to words 3.2. The words that form the synsets of the training data (variants, synonyms and gloss) are used to decide the frecuency of each word with respect to the domain labels that the synset has. Using different similarity measures, a weighted vector of Domains is generated for each word. For instance, table 2 shows a part of a weighted array for the nouns *soccer* (monosemous) and *orange* (polysemous).
- The **parameter filtering** applied in the experiment. Among others, the different weights for each part of information considered: *variants* (70%), words in the gloss (30%). The vectors obtained for each synset were normalized and only labels in the top 15% were considered (range [0.85..1]).

3.2 Measures

To estimate the weights of the words assigned to the domains 3 different formulas have been studied:

Table 2. Weighted array for nouns with factotum (CF)

word	weight	label	weight	label	word
soccer	2.826	soccer	8.181	botany	orange
soccer	2.183	play	5.129	gastronomy	orange
soccer	1.987	football	3.019	color	orange
soccer	1.917	sport	1.594	entomology	orange
soccer	0.998	rugby	1.205	jewellery	orange
...	...	...	...	...	...

M1: Square root formula	M2: Association Ratio	M3: Logarithm formula
$\frac{count(w,D)-\frac{1}{Ncount(w)count(D)}}{\sqrt{count(w,D)}}$	$Pr(w/D)log_2(\frac{Pr(w/D)}{Pr(w)})$	$log_2(\frac{Ncount(w,D)}{count(w)count(D)})$

4 Evaluation and Results

We studied the performance of the different labelling procedures by means of the following evaluation measures:

MiA measures the success of each formula (M1, M2 or M3) when the first proposed label is a correct one.

MiD measures the success of each formula (M1, M2 or M3) when the first proposed label is a correct one (or subsumed by a correct one in the domain hierarchy). For instance, if the proposed label is *Zoology* and the correct answer is *Biologogy* it is considered a correct answer.

Accuracy for the first proposed label $AP = \frac{\text{success of the first label}}{\text{total of synsets}}$	*Accuracy* for all the proposed labels $AT = \frac{\text{success of all the labels}}{\text{total of synsets}}$
Precision $P = \frac{\text{(proposed and correct labels)}}{\text{(total proposed labels)}}$	*Recall* $R = \frac{\text{(proposed and correct labels)}}{\text{total correct labels}}$

For nouns, different experiments were carried out. On average, the method assigns 1.23 domain labels per nominal synset and 1.20 domain labels per verbal synset.

The results when training with factotum and testing with factotum are shown in table 3; and presents the results when making the training and test without factotum. The best average results were obtained with the M1 measure. It must be emphasized that more than 70% of the first labels agree with MWND.

Table 4 presents the results obtained when training and testing for verbs with factotum, and shows the results obtained when training and testing verbs without factotum. In both cases the results are worst than the results obtained for the nouns. One of the reasons may be the high number of verbal *synsets* labeled with factotum domain(see table 1). However, in the case of verbs without factotum, the correct labeling at first proposal are fairly close to 70%.

Table 3. Results for nouns with (CF) and without factotum (SF)

CF					
N	**AP**	**AT**	**P**	**R**	**F1**
M1A	70.94	79.75	64.74	68.25	66.45
M1D	74.50	84.85	68.88	72.62	70.70
M2A	45.75	50.39	42.73	43.12	42.92
M2D	52.09	57.50	48.75	49.21	48.98
M3A	66.77	74.50	60.86	63.76	62.27
M3D	71.56	81.45	66.54	69.71	68.09

SF					
N	**AP**	**AT**	**P**	**R**	**F1**
M1A	73.95	81.82	66.81	68.68	67.73
M1D	78.50	87.24	71.24	73.24	72.23
M2A	52.45	57.52	49.32	48.24	48.77
M2D	59.44	65.21	55.94	54.71	55.32
M3A	74.48	82.69	68.41	69.41	68.91
M3D	78.85	88.64	73.33	74.41	73.87

Table 4. Results for verbs with (CF) and without factotum (SF)

CF					
V	**AP**	**AT**	**P**	**R**	**F1**
M1A	51.24	57.02	47.26	50.74	48.94
M1D	51.24	57.02	47.26	50.74	48.94
M2A	13.22	14.88	12.68	13.24	12.95
M2D	16.53	19.83	16.90	17.65	17.27
M3A	23.14	28.10	21.94	25.00	23.37
M3D	24.79	29.75	23.23	26.47	24.74

SF					
V	**AP**	**AT**	**P**	**R**	**F1**
M1A	69.77	76.74	64.71	55.93	60.00
M1D	74.72	83.72	69.23	61.02	64.86
M2A	20.93	25.58	19.64	18.64	19.13
M2D	41.86	51.16	38.60	37.29	37.93
M3A	41.86	55.81	39.34	40.68	40.00
M3D	53.49	67.44	46.77	49.15	47.93

From these tables, we can also observe that, M1 measure has better F1 than M2 and M3 and the behaviour of M1 and M3 is similar for nouns (CF and SF).

As espected, the method performs better for nouns than for verbs, because nouns have more and (maybe) more clear domain assigments.

For nouns, using the domain hierarchy, the performance increases, achieving 70.94% accuracy when assigning the first domain. However, using the domain hiearchy, it seems that for verbs only increases consistently when testing without factotum. In this case, for verbs the method obtains 51.24% accuracy when assigning the first domain.

Table 5. Training with factotum for nouns using the M1 measure

	Train CF			
	Test CF		Test SF	
	P	R	P	R
M1A	64.74	68.25	86.15	82.35
M1D	68.88	72.62	89.23	85.29

On table 5 there is a comparison for nouns using measure M1 and training with factotum and testing with (CF) and without factotum (SF).

For nouns, the best results are obtained training with factotum and testing without factotum, achieving a 86.15% of precision in the first assignment. One possible reason could be that labels, different than factotum, seems to be better assigned.

5 Discussion

Although the results are quite good, a more accurate analysis of the errors in the automatic assignments will show that the proposed labels are quite similars. It suggests a lack of systematicity in the semi-automatic assignment.

To illustrate possible errors, we show different examples where the proposed label has been considered a mistake in the evaluation.

1. **Monosemic words.** These words may help to find the correct domain.
   credit_application#n#1 (an application for a line of credit)
   Labeled with SCHOOL; proposal 1: Banking and proposal 2: Economy
   OBS: line_of_credit#n#1 is monosemous and is labeled as Banking.
   plague_spot#n#1 (a spot on the skin characteristic of the plague)
   Labeled with ARCHITECTURE; proposal 1: Physiology and proposal 2: Medicine
   OBS: plague#n#1 is monosemic and is labeled as Physiology-Medicine. In addition, skin#n has 6 senses as noun labeled with Anatomy, Transport and Factotum.
2. **Relations between labels.** Exists a direct relation in the domain hierarchy between the proposed labels and correct labels.
   academic_program#n#1 (a program of education in liberal arts and sciences (usually in preparation for higher education))
   Labeled with PEDAGOGY; proposal 1: School and proposal 2: University
   OBS: Pedagogy is the father of School and University.
   shopping#n#1 (searching for or buying goods or services)
   Labeled with ECONOMY; proposal 1: Commerce
   OBS: In the domain hierarchy, Commerce and Economy depend directly on Social_science.
   fire_control_radar#n#1 (radar that controls the delivery of fire on a military target)
   Labeled with MERCHANT_NAVY; proposal 1: Military
   OBS: Merchant_navy depends on Transport and Military and Transport depends on Social_science.
3. **Relations in** WN. Sometimes the *synsets* are related to words in the gloss.
   bowling#n#2 (a game in which balls are rolled at an object or group of objects with the aim of knocking them over play)
   Labeled with BOWLING; proposal 1: Play
   OBS: game#n#2 is hypernym and is labeled as Play. In addition, play#n#16 labeled as Play-Sport is related with holonym with game#n#2. In the domain hierarchy, Play and Sport are sibling and Bowling depends on Sport.
   cost_analysis#n#1 (breaking down the costs of som e operation and reporting on each factor separately)
   Labeled with FACTOTUM; proposal 1: Economy
   OBS: The word “cost” of the gloss have 3 senses labeled with Economy, Money and Quality.

4. **Uncertain cases.** There are cases where the proposed label is not represented by any pattern, but they may be considered as a correct label.
   birthmark#n#1 (a blemish on the skin formed before birth)
   Labeled with QUALITY; proposal 1: Medicine
   bardolatry#n#1 (idolization of William Shakespeare)
   Labeled with RELIGION; proposal 1: history and proposal 2: literature

Further analysis of these cases can help to obtain a validation method of the semi–automatic assigment of domains to synsets. A complete methodology should consider the addition, the removal or substitution of domains.

6 Conclusions and Further Work

The procedure to assign domain labels to WN gloss is very promising, especially because it is a difficult problem for the polysemy of WN and the semi-automatic process to generate the domain labels, using the WN hierarchy.

The proposal process is very reliable with the first proposal labels, reaching more that 70% on accuracy when testing without factotum.

We provided also an study of the typology of the errors. This suggest that in certain cases it is possible to add new correct labels or validate the old ones. In addition, other suggestion is that a lot of words labeled as factotum may be labeled with concrete domain label.

As future work we consider to make improvements, adaptations in the algorithm and test new methods to label other versions of WN.

References

1. E. Agirre and D. Martinez. Integrating selectional preferences in wordnet. In *Proceedings of the first International WordNet Conference in Mysore*, India, 21–25 January 2002.
2. J. Atserias, S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. Combining multiple methods for the automatic construction of multilingual wordnets. In *Procceeding of RANLP'97*, pages 143–149, Bulgaria, 1997. Also to appear in a Book.
3. L. Bentivogli, E. Pianta, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India, 2002.
4. C. Fellbaum, editor. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.
5. J. Gonzalo, F. Verdejo, C. Peters, and N. Calzolari. Applying eurowordnet to cross-language text retrieval. *Computers and Humanities*, 1998.
6. S. Harabagiu, M. Pasca, and S. Maiorano. Experiments with open-domain textual question answering. In *Proceedings of COLING-2000*, Saarbruken Germany, 2000.
7. B. Magnini and G. Cavagliá. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens. Greece, 2000.
8. B. Magnini and C. Strapparava. User modelling for news web sites with content-based techniques. In *Proceedings WWW-2002, the Eleventh International World Wide Web Conference, Poster session*, Honululu, Hawaii, USA,, 2002.
9. B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. Using domain information for word sense disambiguation. In *Proceedings of* 2^{nd} *International Workshop "Evaluating Word Sense Disambiguation Systems", SENSEVAL-2*, Toulouse, France, 2001.

10. B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. Comparing Ontology-Based and Corpus-Based Domain Annotations in WordNet. *In Proceedings of First International WordNet Conference*, 2002.
11. D. McCarthy. *Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, 2001.
12. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five Papers on WordNet. *Special Issue of International Journal of Lexicography*, 3(4), 1990.
13. G. Rigau, H. Rodríguez M., and E. Agirre. Building accurate semantic taxonomies from monolingual mrds. In *In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics COLING-ACL '98*, Montreal, Canada, 1998.