Evaluating large-scale Knowledge Resources across Languages

Montse Cuadros
TALP Research Center
Universitat Politécnica de Catalunya
Barcelona, Spain
cuadros@lsi.upc.edu

German Rigau
IXA NLP Group
Euskal Herriko Unibersitatea
Donostia, Spain
german.rigau@ehu.es

Mauro Castillo
Departamento de Computación e Informática
Universidad Tecnológica Metropolitana
Santiago de Chile, Chile
mcast@utem.cl

Abstract

This paper presents an empirical evaluation in a multilingual scenario of the semantic knowledge present on publicly available large-scale knowledge resources. The study covers a wide range of manually and automatically derived large-scale knowledge resources for English and Spanish. In order to establish a fair and neutral comparison, the knowledge resources are evaluated using the same method on two Word Sense Disambiguation tasks (Senseval-3 English and Spanish Lexical Sample Tasks). First, this study empirically demonstrates that the combination of the knowledge contained in these resources surpass the most frequent sense classifier for English. Second, we also show that this large-scale topical knowledge acquired from one language can be successfully ported to other languages.

Keywords

Large-scale knowledge resources, lexical semantics, evaluation, WordNet, Word Sense Disambiguation

1 Introduction

Using large-scale knowledge bases, such as WordNet (WN) [9], has become a usual, often necessary, practice for most current Natural Language Processing (NLP) systems. Even now, building large and rich enough knowledge bases for broad–coverage semantic processing takes a great deal of expensive manual effort involving large research groups during long periods of development. In fact, hundreds of person-years have been invested in the development of wordnets for various languages [21]. For example, in more than ten years of manual construction (from 1995 to 2006, that is from version 1.5 to 3.0), WN passed from 103,445 semantic relations to 235,402 semantic relations[1]. That is, around one thousand new relations per month. But this data does not seems to be rich enough to support advanced concept-based NLP applications directly. It seems that applications will not scale up to working in open domains without more detailed and rich general-purpose (and also domain-specific) semantic knowledge built by automatic means.

Fortunately, during the last years the research community has devised a large set of innovative methods and tools for large-scale automatic acquisition of lexical knowledge from structured and unstructured corpora. Among others we can mention eXtended WN [17], large collections of semantic preferences acquired from SemCor [2, 3] or acquired from British National Corpus (BNC) [15], large-scale Topic Signatures for each synset acquired from the web [1] or acquired from the BNC [6]. Obviously, all these semantic resources have been acquired using a very different set of processes, tools and corpora, resulting on a different set of new semantic relations between synsets. In fact, each semantic resource has different volume and accuracy figures when evaluated in a common and controlled framework [7]. However, as far as we know, no empirical study has been carried out trying to see how these semantic resources complement each other.

Furthermore, since this knowledge is language independent (knowledge represented at the semantic level as relations between synsets), to date no empirical evaluation has been performed showing to which extent these large-scale semantic resources acquired from one language (in this case English) could be of utility for another (in this case Spanish).

This paper is organized as follows. First, we introduce the multilingual semantic resources compared in the evaluation. In section 3 we present the multilingual evaluation framework used in this study. Section 4 describes the results when evaluating these large-scale semantic resources on English and section 5 on Spanish. Finally, section 6 presents some concluding remarks and future work.

2 Multilingual Knowledge Resources

Our evaluation covers a wide range of large-scale semantic resources: WordNet (WN) [9], eXtended WordNet [17], large collections of semantic preferences ac-

[1] Symmetric relations are counted only once.

Source	#relations
Princeton WN1.6	138,091
Selectional Preferences from SemCor	203,546
New relations from Princeton WN2.0	42,212
Gold relations from eXtended WN	17,185
Silver relations from eXtended WN	239,249
Normal relations from eXtended WN	294,488
Total English	934,771
Total Spanish	517,279

Table 1: *Semantic relations uploaded in the* MCR

quired from SemCor [2, 3] or acquired from the BNC [15], large-scale Topic Signatures for each synset acquired from the web [1] or SemCor [11].

Although these resources have been derived using different WN versions, using the technology for the automatic alignment of wordnets [8], most of these resources have been integrated in a common resource called Multilingual Central Repository (MCR) [4] maintaining the compatibility among all the knowledge resources which use a particular WN version as a sense repository. Furthermore, these mappings allow to port the knowledge associated to a particular WN version to the rest of WN versions.

2.1 Multilingual Central Repository

The Multilingual Central Repository (MCR) [4] is a result of the 5th Framework MEANING project[2]. The MCR follows the model proposed by the EuroWordNet project. EuroWordNet [21] is a multilingual lexical database with wordnets for several European languages, which are structured as the Princeton WN.

The MCR constitutes a natural multilingual large-scale linguistic resource for a number of semantic processes that need large amounts of multilingual knowledge to be effective tools. The MCR also integrates WN Domains [14], new versions of the Base Concepts and the Top Concept Ontology, and the SUMO ontology [20]. The current version of the MCR contains 934,771 semantic relations between synsets, most of them acquired by automatic means. This represents almost four times larger than the Princeton WN (235,402 unique semantic relations in WN 3.0). Table 1 shows the number of semantic relations between synset pairs in the MCR. As the current version of the Spanish Wordnet do not have translation equivalents for all the English synsets[3], the total number of ported relations is around a half of the English ones.

Hereinafter we will refer to each resource as follows:

WN [9]: This resource uses the direct relations encoded in WN1.6 and WN2.0 (for instance, tree#n#1–hyponym–>teak#n#2). We also tested WN^2 (using relations at distance 1 and 2), WN^3 (using relations at distances 1 to 3) and WN^4 (using relations at distances 1 to 4).

XWN [17]: This resource uses the direct relations encoded in eXtended WN (for instance, teak#n#2–gloss–>wood#n#1).

[2] http://nipadio.lsi.upc.es/~nlp/meaning

[3] Currently, the Spanish WN has translation equivalents to English for 62,720 synsets.

political_party#n#1	2.3219
party#n#1	2.3219
election#n#1	1.0926
nominee#n#1	0.4780
candidate#n#1	0.4780
campaigner#n#1	0.4780

Table 2: *Topic Signatures for party#n#1 obtained from Semcor (6 out of 719 total word senses)*

.

WN+XWN: This resource uses the direct relations included in WN and XWN. We also tested $(WN+XWN)^2$ (using either WN or XWN relations at distances 1 and 2, for instance, tree#n#1–related–>wood#n#1).

spBNC [15]: This resource contains 707,618 selectional preferences acquired for subjects and objects from BNC.

spSemCor [3]: This resource contains the selectional preferences acquired for subjects and objects from SemCor (for instance, read#v#1–tobj–>book#n#1).

MCR [4]: This resource uses the direct relations included in MCR but in the experiments below we excluded spBNC because of its poor performance. Thus, MCR contains the direct relations from WN, XWN, and spSemCor but not the indirect relations of $(WN+XWN)^2$. We also tested MCR^2 (using relations at distance 1 and 2), which also integrates $(WN+XWN)^2$ relations.

2.2 Topic Signatures

Topic Signatures (TS) are word vectors related to a particular topic [13].

For this study, we use two different large-scale TS. The first constitutes one of the largest available semantic resource with around 100 million relations (between synsets and words) acquired from the web [1]. The second has been derived directly from SemCor.

TSWEB[4]: Inspired by the work of [12], these TS were constructed using monosemous relatives from WN (synonyms, hypernyms, direct and indirect hyponyms, and siblings), querying Google and retrieving up to one thousand snippets per query (that is, a word sense), extracting the words with distinctive frequency using TFIDF. For these experiments, we used at maximum the first 700 words.

Since this is a semantic resource between word-senses and words, it is not possible to port these relations to Spanish without introducing a large amount of noise.

TSSEM: These TS have been constructed using the part of SemCor having all words tagged by PoS, lemmatized and sense tagged according to WN1.6 totalizing 192,639 words. For each word-sense appearing in SemCor, we gather all sentences for that word sense, building a TS using TFIDF for all word-senses co-occurring in those sentences.

In table 2, there is an example of the first word-senses we calculate from party#n#1.

[4] `http://ixa.si.ehu.es/Ixa/resources/sensecorpus`

The total number of relations between WN synsets acquired from SemCor is 932,008. In this case, due to the smaller size of the Spanish WN, the total number of ported relations is 586,881.

3 Evaluation framework

In order to compare the knowledge resources described in the previous section, we evaluated all these resources as Topic Signatures (TS).This simple representation tries to be as neutral as possible with respect to the resources used.

All knowledge resources are evaluated on a WSD task. In particular, in section 4 we used the noun-set of Senseval-3 English Lexical Sample task which consists of 20 nouns and in section 5 we used the noun-set of the Senseval-3 Spanish Lexical Sample task which consists of 21 nouns. For Spanish, the MiniDir dictionary was specially developed for the Senseval-3 task. Most of the MiniDir word senses have links to WN1.5 (which in turn are linked by the MCR to the Spanish WN). All performances are evaluated on the test data using the fine-grained scoring system provided by the organizers. We use the noun-set only because TSWEB is available only for nouns, and the English Lexical Sample uses the WordSmyth dictionary [18] as a sense repository for verbs instead of WN.

Furthermore, trying to be as neutral as possible with respect to the resources studied, we applied systematically the same disambiguation method to all of them. Recall that our main goal is to establish a fair comparison of the knowledge resources rather than providing the best disambiguation technique for a particular knowledge base.

A common WSD method has been applied to all knowledge resources. A simple word overlapping counting is performed between the TS and the test example[5]. The synset having higher overlapping word counts is selected. In fact, this is a very simple WSD method which only considers the topical information around the word to be disambiguated. Finally, we should remark that the results are not skewed (for instance, for resolving ties) by the most frequent sense in WN or any other statistically predicted knowledge.

4 English evaluation

4.1 Baselines for English

We have designed a number of basic baselines in order to establish a complete evaluation framework for comparing the performance of each semantic resource on the English WSD task.

RANDOM: For each target word, this method selects a random sense. This baseline can be considered as a lower-bound.

SemCor MFS (SEMCOR-MFS): This method selects the most frequent sense of the target word in SemCor.

WordNet MFS (WN-MFS): This method selects the most frequent sense (the first sense in WN1.6) of the target word. WN word-senses were ranked using SemCor and other sense-annotated corpora. Thus, WN-MFS and SemCor-MFS are similar, but not equal.

TRAIN-MFS: This method selects the most frequent sense in the training corpus of the target word.

Train Topic Signatures (TRAIN): This baseline uses the training corpus to directly build a TS using TFIDF measure for each word sense. Note that in WSD evaluation frameworks, this is a very basic system, a baseline. However, in our evaluation framework, this "WSD baseline" should be considered as an upper-bound.

[5] We also consider multiword terms.

Baselines	P	R	F1
TRAIN	65.1	65.1	65.1
TRAIN-MFS	54.5	54.5	54.5
WN-MFS	53.0	53.0	53.0
SEMCOR-MFS	49.0	49.1	49.0
RANDOM	19.1	19.1	19.1

Table 3: *P, R and F1 results for English Lexical Sample Baselines*

Table 3 presents the precision (P), recall (R) and F1 measure (harmonic mean of recall and precision) of the different baselines. In this table, TRAIN has been calculated with a vector size of at maximum 450 words. As expected, RANDOM baseline obtains the poorest result. The most frequent senses obtained from SemCor (SEMCOR-MFS) and WN (WN-MFS) are both below the most frequent sense of the training corpus (TRAIN-MFS). However, all of them are far below to the TS acquired using the training corpus (TRAIN).

4.2 Evaluating each resource on English

Table 4 presents ordered by F1 measure, the performance of each knowledge resource and its average size of the Topic Signature per word-sense. The average size of a knowledge resource is the lenght of the word list associated to a synset on average. Obviously, the best resources would be those obtaining better performances with a smaller number of related words per synset. The best results for precision, recall and F1 measures are shown in bold. We also mark in italics those derived resources applying non-direct relations. Surprisingly, the best results are obtained by TSSEM (with F1 of 52.4). The lowest result is obtained by the knowledge directly gathered from WN mainly because of its poor coverage (recall of 18.4 and F1 of 26.1). Also interesting, is that the knowledge integrated in the MCR although partly derived by automatic means performs much better in terms of precision, recall and F1 measures than using them separately (F1 with 18.4 points higher than WN, 9.1 than XWN and 3.7 than spSemCor).

Despite its small size, the resources derived from SemCor obtain better results than its counterparts using much larger corpora (TSSEM vs. TSWEB and spSemCor vs. spBNC).

Regarding the baselines, all knowledge resources surpass RANDOM, but none achieves neither WN-MFS, TRAIN-MFS nor TRAIN. Only TSSEM obtains

KB	P	R	F1	Av. Size
TSSEM	**52.5**	**52.4**	**52.4**	103
MCR[2]	45.1	45.1	45.1	26,429
MCR	45.3	43.7	44.5	129
spSemCor	43.1	38.7	40.8	56
(WN+XWN)[2]	38.5	38.0	38.3	5,730
WN+XWN	40.0	34.2	36.8	74
TSWEB	36.1	35.9	36.0	1,721
XWN	38.8	32.5	35.4	69
WN[3]	35.0	34.7	34.8	503
WN[4]	33.2	33.1	33.2	2,346
WN[2]	33.1	27.5	30.0	105
spBNC	36.3	25.4	29.9	128
WN	44.9	18.4	26.1	14

Table 4: *P, R and F1 fine-grained results for the resources evaluated individually on English.*

better results than SEMCOR-MFS and is very close to the most frequent sense of WN (WN-MFS) and the training (TRAIN-MFS).

Regarding other expansions and combinations, the performance of WN is improved using words at distances up to 2 (F1 of 30.0), and up to 3 (F1 of 34.8), but it decreases using distances up to 4 (F1 of 33.2). Interestingly, none of these WN expansions achieve the results of XWN (F1 of 35.4). Finally, (WN+XWN)^2 performs better than WN+XWN and MCR^2 slightly better than MCR[6].

4.3 Combining resources

In order to evaluate more deeply the contribution of each knowledge resource, we also provide some results of the combined outcomes of several resources. The combinations are performed following three different basic strategies [5].

Direct Voting (DV): Each semantic resource has one vote for the predominant sense of the word to be disambiguated and the sense with most votes is chosen.

Probability Mixture (PM): Each semantic resource provides a probability distribution over the senses of the word to be disambiguated. These probabilities (normalized scores) are summed, and the sense with the highest score is chosen.

Rank-Based Combination (Rank): Each semantic resource provides a ranking of senses of the word to be disambiguated. For each sense, its placements according to each of the methods are summed and the sense with the lowest total placement (closest to first place) is selected.

4.3.1 Combining two resources

Table 5 presents the F1 measures with respect these three methods when combining two different resources. The combinations are ordered by the result of the rank-based combination. The best result which corresponds to the rank-based combination of MCR and TSSEM[7] is shown in bold.

[6] No further distances have been tested.

[7] Note that in this case, some information appearing in SemCor could be counted twice, as we are not removing duplicated relations

KB	PM	DV	Rank
MCR+TSSEM	52.3	45.4	**52.7**
MCR+(WN+XWN)^2	47.8	37.8	51.5
(WN+XWN)^2+TSSEM	51.0	41.7	50.5
TSSEM+TSWEB	51.0	42.2	49.4
MCR+TSWEB	48.9	37.6	48.6
(WN+XWN)^2+TSWEB	41.5	34.3	45.4

Table 5: *F1 fine-grained results for the 2 system-combinations*

KB	PM	DV	Rank
MCR+TSSEM+(WN+XWN)^2	52.6	37.9	**54.6**
MCR+TSWEB+TSSEM	54.1	37.2	53.3
MCR+TSWEB+(WN+XWN)^2	49.8	33.3	52.1
(WN+XWN)^2+TSSEM+TSWEB	51.5	36.1	51.5

Table 6: *F1 fine-grained results for the 3 system-combinations*

Regarding the combination method applied, the probability-mixture and the rank-based methods behave similarly (each method wins in three of the six combinations), and obtaining better results than the direct-voting method. Hereinafter, we use the rank-based measure for comparing results.

Interestingly, only in two cases the ensemble of resources makes worse the individual results. Both cases involve TSSEM (F1 of 52.4) when combined with TSWEB (F1 of 49.4) and (WN+XWN)^2 (F1 of 50.5). However, for the rest of the cases, it seems that each resource provides some kind of knowledge not provided by the others. For instance, the knowledge contained in (WN+XWN)^2 seems to be not represented in the MCR. Furthermore, despite (WN+XWN)^2+TSWEB obtains the lower results (F1 of 45.4) when combining two resources, the individual contribution to the ensemble is impressive (5.4 points with respect (WN+XWN)^2) and (9.4 points with respect to TSWEB). However, the larger increment corresponds to MCR+(WN+XWN)^2 (F1 of 51.5, 6.0 points higher than MCR and 13.25 higher than (WN+XWN)^2), indicating that both resources contain complementary knowledge. In fact, there is some knowledge contained in the MCR not present in TSSEM (because the small increment of 0.3 points with respect TSSEM alone).

Regarding the baselines, none of the combinations achieves the most frequent sense of WN (WN-MFS with F1 of 53.0). However, several of them surpass the most frequent sense of SemCor (SEMCOR-MFS with F1 of 49.1). In particular, the combinations including information from SemCor (TSSEM or MCR).

4.3.2 Combining three resources

Table 6 presents the F1 measure results with respect these three methods when combining three different semantic resources. The combinations are ordered by the result of the rank-based combination. The best result which corresponds to the rank-based combination of MCR (WN+XWN+spSemCor), TSSEM and (WN+XWN)^2 is presented in bold.

KB	PM	DV	Rank
MCR+(WN+XWN)^2+TSWEB+TSSEM	53.1	32.7	**55.5**

Table 7: *F1 fine-grained results for the 4 system-combinations*

Baselines	P	R	F1
TRAIN	81.8	68.0	74.3
MiniDir-MFS	67.1	52.7	59.2
RANDOM	21.3	21.3	21.3

Table 8: *P, R and F1 fine-grained results for Spanish Lexical Sample Baselines*

Regarding the combination method applied, the rank-based method seems to be similar to probability-mixture (winning in two of the four combinations, losing in one and having a tie in one). Again, both strategies are superior to the direct-voting method.

Considering only the rank-based combination, in general, the combination of three knowledge resources obtains slightly better results than using only two or one resource. In this case, only one ensemble of resources makes worse the individual results. This case involves again TSSEM (F1 of 52.4) when combined with (WN+XWN)^2+TSWEB (F1 of 45.4). However, for the rest of the cases, again it seems that the combination of resources integrates some knowledge not provided by the resources individually. In this case, the larger increase corresponds to MCR+TSWEB+(WN+XWN)^2 (F1 of 52.1, 16.1 points higher than TSWEB, 12.1 points higher than (WN+XWN)^2, and 7.6 points higher than MCR). Furthermore, there is some knowledge contained in the MCR+(WN+XWN)^2 not present in TSSEM (because an small increment of 2.2 points with respect TSSEM alone).

In fact, all these combinations outperform the most frequent sense of SemCor (F1 of 49.1), and two combinations of three resources surpass the most frequent sense of WN (WN-MFS with F1 of 53.0): MCR+TSWEB+TSSEM (F1 of 53.3) and MCR+TSSEM+(WN+XWN)^2 (F1 of 54.6), and the later is also slightly over the most frequent sense of the training (F1 of 54.5). Obviously, this result should be highlighted since in the all-words tasks most current supervised approaches rarely surpass the simple heuristic of choosing the most frequent sense in the training data, despite taking local context into account [10].

4.3.3 Combining four resources

Table 7 presents the F1 measure results with respect the three methods when combining the four different semantic resources. In bold is presented the best result which corresponds to the rank-based combination of MCR, TSSEM, TSWEB and (WN+XWN)^2.

It seems that the rank-based has better behavior than direct-voting or probability-mixture methods.

Considering only the rank-based combination, as expected, the combination of the four knowledge resources obtains better results than using only three, two or one resource. Again, it seems that the combination of resources provides some kind of knowledge not provided by each of the resources individually. In this case, 19.5 points higher than TSWEB, 17.25 points higher than (WN+XWN)^2, 11.0 points higher than MCR and 3.1 points higher than TSSEM.

Regarding the baselines, this combination outperforms the most frequent sense of SemCor (SEMCOR-MFS with F1 of 49.1), WN (WN-MFS with F1 of 53.0) and, the training data (TRAIN-MFS with F1 of 54.5). This fact indicates that the resulting combination of large-scale resources encodes the knowledge necessary to behave as a most frequent sense tagger for English. Furthermore, it is also worth mentioning that the most frequent synset for a word, according to the WN sense ranking is very competitive in WSD tasks, and it is extremely hard to improve upon even slightly [16].

5 Spanish evaluation

5.1 Spanish Baselines

As well as for English, we have designed a number of basic baselines in order to establish a complete evaluation framework for comparing the performance of each semantic resource when evaluated on the Spanish WSD task.

RANDOM: For each target word, this method selects a random sense. This baseline can be considered as a lower-bound.

Minidir MFS (Minidir-MFS): This method selects the most frequent sense (the first sense in Minidir) of the target word. Since Minidir is a special dictionary built for the task, the word-sense ordering corresponds to their frequency in the training data. Thus, for Spanish, Minidir-MFS is equal to TRAIN-MFS.

Train Topic Signatures (TRAIN): This baseline uses the training corpus to directly build a Topic Signature using TFIDF measure for each word sense. As for English, this baseline can be considered as an upper-bound of our evaluation.

Note that the Spanish WN do not encodes word-sense frequency information and for Spanish there is no all-words sense tagged corpora available of the style of Italian[8].

In the Spanish evaluation only sense–disambiguated relations can be ported without introducing extra noise. For instance, TSWEB has not been tested on the Spanish side. TSWEB relate synsets to words, not synsets to synsets. As this resource is not word-sense disambiguated, when translating the English words to Spanish, a large amount of noise would be introduced (Spanish words not related to the particular synset).

Table 8 presents the precision (P), recall (R) and F1 measure of the different baselines. As for English, TRAIN has been calculated with a vector size of at maximum 450 words. As expected, RANDOM baseline obtains the poorest result and the most frequent sense obtained from Minidir (Minidir-MFS, and also TRAIN-MFS) is far below the TS acquired using the training corpus (TRAIN).

[8] http://multisemcor.itc.it/

Knowledge Bases	P	R	F1	Av. Size
MCR	46.1	**41.1**	**43.5**	66
WN^2	56.0	29.0	42.5	51
$(WN+XWN)^2$	41.3	41.2	41.3	1,892
TSSEM	33.6	33.2	33.4	208
XWN	42.6	27.1	33.1	24
WN	**65.5**	13.6	22.5	8

Table 9: *P, R and F1 fine-grained results for the resources evaluated individually on Spanish.*

5.2 Evaluating each resource on Spanish

Table 9 presents ordered by F1 measure, the performance of the knowledge resources and its average size per word-sense. In bold appear the best results for precision, recall and F1 measures. WN obtains the highest precision (P of 65.5) but due to its poor coverage (R of 13.6), the lowest result (F1 of 22.5). Also interesting, is that the knowledge integrated in the MCR outperforms in terms of precision, recall and F1 measures the results of TSSEM, possibly indicating that the knowledge currently uploaded in the MCR is more robust than TSSEM and that the topical knowledge gathered from a sense-annotated corpus of one language can not be directly ported to another language. Possible explanations of these low results could be the smaller size of the resources (approximately a half size) and the differences in the evaluation frameworks, including the dictionary, sense distinctions and mappings.

Regarding the baselines, all knowledge resources surpass RANDOM, but none achieves neither Minidir-MFS (equal to TRAIN-MFS) nor TRAIN.

6 Conclusions and further work

To our knowledge, this is the first time to show that a very simple WSD system based on topical knowledge gathered from several semantic resources outperforms the Most Frequent Sense classifiers in the SensEval-3 English lexical-sample task. Obviously, more sophisticated approaches could be devised [19]. Furthermore, since these resources represent semantic relations at the conceptual level, can be also successfuly ported to and evaluated in other languages.

It is our belief, that accurate WSD systems would rely not only on sophisticated algorithms but on knowledge intensive approaches. The results presented in this paper suggests that much more research on acquiring and using large-scale semantic resources should be addressed.

It seems that the combination of publicly available large-scale resources encodes the knowledge necessary to behave as a most frequent sense tagger for English. We plan to empirically validate this hypothesis in all-words tasks.

Further experiments in the cross-lingual scenario are also needed to clarify the different behaviours of the MCR and TSSEM, maybe using the Italian WN (also integrated in the MCR) and MultiSemCor.

7 Acknowledgements

We want to thank the valuable comments of the anonymous reviewers. This work has been partially supported by the projects KNOW (TIN2006-15049-C03-01) and ADIMEN (EHU06/113).

References

[1] E. Agirre and O. L. de Lacalle. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of LREC*, Lisbon, Portugal, 2004.

[2] E. Agirre and D. Martinez. Learning class-to-class selectional preferences. In *Proceedings of CoNLL*, Toulouse, France, 2001.

[3] E. Agirre and D. Martinez. Integrating selectional preferences in wordnet. In *Proceedings of GWC*, Mysore, India, 2002.

[4] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. The meaning multilingual central repository. In *Proceedings of GWC*, Brno, Czech Republic, 2004.

[5] S. Brody, R. Navigli, and M. Lapata. Ensemble methods for unsupervised wsd. In *Proceedings of COLING-ACL*, pages 97–104, 2006.

[6] M. Cuadros, L. Padró, and G. Rigau. Comparing methods for automatic acquisition of topic signatures. In *Proceedings of RANLP*, Borovets, Bulgaria, 2005.

[7] M. Cuadros and G. Rigau. Quality assessment of large scale knowledge resources. In *Proceedings of EMNLP*, 2006.

[8] J. Daudé, L. Padró, and G. Rigau. Validation and Tuning of Wordnet Mapping Techniques. In *Proceedings of RANLP*, Borovets, Bulgaria, 2003.

[9] C. Fellbaum, editor. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.

[10] V. Hoste, W. Daelemans, I. Hendrickx, and A. van den Bosch. Evaluating the results of a memory-based word-expert approach to unrestricted word sense disambiguation. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 95–101, 2002.

[11] S. Landes, C. Leacock, and R. Tengi. Building a semantic concordance of english. In *WordNet: An electronic lexical database and some applications. MIT Press, Cambridge,MA., 1998*, pages 97–104, 2006.

[12] C. Leacock, M. Chodorow, and G. Miller. Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–166, 1998.

[13] C. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, 2000. Strasbourg, France.

[14] B. Magnini and G. Cavaglià. Integrating subject field codes into wordnet. In *Proceedings of LREC*, Athens. Greece, 2000.

[15] D. McCarthy. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, 2001.

[16] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant senses in untagged text. In *Proceedings of ACL*, pages 280–297, 2004.

[17] R. Mihalcea and D. Moldovan. extended wordnet: Progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA, 2001.

[18] R. Mihalcea, T.Chlovski, and A.Killgariff. The senseval-3 english lexical sample task. In *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona, 2004.

[19] R. Navigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7):1063–1074, 2005.

[20] I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 17–19. Chris Welty and Barry Smith, eds, 2001.

[21] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* . Kluwer Academic Publishers , 1998.