Evaluating large-scale Knowledge Resources across Languages

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

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.

1 Introduction

Using large-scale knowledge bases, such as WordNet (Fellbaum, 1998), 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, dozens of person-years have been invested in the development of wordnets for various languages (Vossen, 1998). For example, in more than ten years of manual construction (from version 1.5 to 2.1), WordNet passed from 103,445 semantic relations to 245,509 semantic relations[1]. That is, around one thousand new relations per month. But this data do 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 WordNet (Mihalcea and Moldovan, 2001), large collections of semantic preferences acquired from SemCor (Agirre and Martinez, 2001; Agirre and Martinez, 2002) or acquired from British National Corpus (BNC) (McCarthy, 2001), large-scale Topic Signatures for each synset acquired from the web (Agirre and de la Calle, 2004) or acquired from the BNC (Cuadros et al., 2005). 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 (Cuadros and Rigau, 2006). However, as far as we know, no empirical study has been carried out trying to see how these large-scale semantic resources complement each other.

[1] Symmetric relations are counted only once.

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 extend 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

The evaluation presented here covers a wide range of large-scale semantic resources: WordNet (WN) (Fellbaum, 1998), eXtended WordNet (Mihalcea and Moldovan, 2001), large collections of semantic preferences acquired from SemCor (Agirre and Martinez, 2001; Agirre and Martinez, 2002) or acquired from the BNC (McCarthy, 2001), large-scale Topic Signatures for each synset acquired from the web (Agirre and de la Calle, 2004) or SemCor (Landes et al., 2006).

Although these resources have been derived using different WN versions, using the technology for the automatic alignment of wordnets (Daudé et al., 2003), most of these resources have been integrated into a common resource called Multilingual Central Repository (MCR) (Atserias et al., 2004) 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)[2] follows the model proposed by the EuroWordNet project. EuroWordNet (Vossen, 1998) is a multilingual lexical database with wordnets for several European languages, which are structured as the Princeton WordNet. The Princeton WordNet contains information about nouns, verbs, adjectives and adverbs in English and is organized around the notion of a *synset*. A synset is a set of words with the same part-of-speech that can be interchanged in a certain context. For example, <*party*, *political_party*> form a synset because they can be used to refer to the same concept. A synset is often further described by a gloss, in this case: ”an organization to gain political power”. Finally, synsets can be related to each other by semantic relations, such as hyponymy, meronymy, cause, etc.

The current version of the MCR (Atserias et al., 2004) is a result of the 5th Framework MEANING project. The MCR integrates into the same EuroWordNet framework wordnets from five different languages, including Spanish (together with four English WN versions). The wordnets are currently linked via an Inter-Lingual-Index (ILI) allowing the connection from words in one language to translation equivalent words in any of the other languages. In that way, the MCR constitutes a natural multilingual large-scale linguistic resource for a number of semantic processes that need large amount of multilingual knowledge to be effective tools. The MCR also integrates WordNet Domains (Magnini and Cavaglià, 2000), new versions of the Base Concepts and the Top Concept Ontology, and the SUMO ontology (Niles and Pease, 2001). 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 WordNet (245,509 unique semantic relations in WordNet 2.1).

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 semantic resource as follows:

WN (Fellbaum, 1998): This resource uses the direct relations encoded in WN1.6 or WN2.0. We also tested WN^2 (using relations at distance 1 and 2), WN^3 (using relations at distances 1 to 3) and WN^4

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

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

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 into the MCR

(using relations at distances 1 to 4).

XWN (Mihalcea and Moldovan, 2001): This resource uses the direct relations encoded in eXtended WN.

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).

spBNC (McCarthy, 2001): This resource contains 707,618 selectional preferences acquired from BNC.

spSemCor (Agirre and Martinez, 2002): This resource contains the selectional preferences acquired from SemCor.

MCR (Atserias et al., 2004): This resource uses the direct relations included in MCR. We also tested $(MCR)^2$ (using relations at distance 1 and 2).

2.2 Topic Signatures

Topic Signatures (TS) are word vectors related to a particular topic (Lin and Hovy, 2000). Topic Signatures are built by retrieving context words of a target topic from large corpora. In our case, we consider word senses as topics. Basically, the acquisition of TS consists of A) acquiring the best possible corpus examples for a particular word sense (usually characterizing each word sense as a query and performing a search on the corpus for those examples that best match the queries), and then, B) building the TS by deriving the context words that best represent the word sense from the selected corpora.

For this study, we use two different large-scale Topic Signatures. The first constitutes one of the largest available semantic resource with around 100 million relations (between synsets and words) acquired from the web (Agirre and de la Calle, 2004). The second has been derived directly from SemCor.

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
regime#n#1	0.3414
identification#n#1	0.3414
government#n#1	0.3414
designation#n#3	0.3414
authorities#n#1	0.3414

Table 2: Topic Signatures for party#n#1 obtained from Semcor (11 out of 719 total word senses)

.

TSWEB[4]: Inspired by the work of (Leacock et al., 1998), these Topic Signatures were constructed using monosemous relatives from WordNet (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.

In this case, being this a semantic resource between word-senses and words, it is not possible to port this large amount of relations to Spanish.

TSSEM: These Topic Signatures 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.

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). That is, word vectors with weights associated to a particular

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

synset which are obtained by collecting those word senses appearing in the synsets directly related to them. 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 task. Most of the MiniDir word senses have links to WN1.5 (which in turn are linked by the MCR to the Spanish WordNet). All performances are evaluated on the test data using the fine-grained scoring system provided by the organizers. The interest in only the noun-set is for comparison purpouses since TSWEB is only available for nouns.

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 Topic Signature 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.

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

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.

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 Topic Signature using TFIDF measure for each word sense. Note that this baseline can be considered as an upper-bound of our evaluation.

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 Topic Signatures 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 per word-sense. 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 (R of 18.4 and F1 of 26.1). Also interesting, is that the knowledge integrated into the MCR although partly derived by automatic means performs much better in terms of precision, recall and F1 measures than us-

[5] We also consider multiword terms.

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	38.25	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.

ing 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 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 (Brody et al., 2006).

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

KB	Sum	Direct	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

Probability Mixture: 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: 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.

Regarding the combination method applied, the direct-voting and the rank-based methods behave similarly (each method wins in three of the six combinations), and obtaining better results than the probability-mixture 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 into the MCR. Furthermore, despite (WN+XWN)^2+TSWEB obtains the lower results

[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	Sum	Direct	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

(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 into the MCR not present into 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. Regarding the combination method applied, the rank-based method seems to be similar to direct-voting (winning in two of the four combinations, losing in one and having a tie in one). Again, both strategies are superior to the probability-mixture 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).

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

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

For instance, the knowledge contained in (WN+XWN)^2 seems to be not represented into 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 11.5 higher than (WN+XWN)^2), indicating that the three resources contain complementary knowledge. Furthermore, there is some knowledge contained into the MCR+(WN+XWN)^2 not present into 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 o 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 (Hoste et al., 2002).

4.3.3 Combining four resources

Table 7 presents the F1 measure results with respect these 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 $(WN+XWN)^2$.

Again, the rank-based method 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 (McCarthy et al., 2004).

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 (McCarthy et al., 2004).

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. Being Minidir 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. Note that this baseline can be considered as an upper-bound of our evaluation.

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

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.

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).

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 Topic Signatures acquired using the training corpus (TRAIN).

5.2 Evaluating each resource on Spanish

Table 9 presents ordered by F1 measure, the performance of knowledge resource 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 into the MCR outperforms in terms of precision, recall and F1 measures the results of TSSEM, possi-

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

bly indicating that the knowledge currently uploaded into 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), the differences in the evaluation frameworks, including the dictionary (sense distinctions and mappings), etc.

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 using only large amounts of topical knowledge gathered from several resources outperforms the Most Frequent Sense classifiers in the SensEval-3 English lexical-sample task. Obviously, more sophisticated approaches could be devised (Navigli and Velardi, 2005). 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 needed to clarify the different behaviours of the MCR and TSSEM, maybe using the Italian WN (also integrated into the MCR) and MultiSemCor.

References

E. Agirre and O. Lopez de la Calle. 2004. Publicly available topic signatures for all wordnet nominal senses. In *Proceedings of LREC*, Lisbon, Portugal.

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

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

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

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

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

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

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

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

V. Hoste, W. Daelemans, I. Hendrickx, and A. van den Bosch. 2002. 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.

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

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

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

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

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

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

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

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

I. Niles and A. Pease. 2001. 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.

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