KnowNet: A proposal for building highly connected and dense knowledge bases from the web

Montse Cuadros
TALP Research Center, UPC
Barcelona, Spain
cuadros@lsi.upc.edu

German Rigau
IXA NLP Group, UPV/EHU
Donostia, Spain
german.rigau@ehu.es

Abstract

This paper presents a new fully automatic method for building highly dense and accurate knowledge bases from existing semantic resources. Basically, the method uses a wide-coverage and accurate knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate senses to large sets of topically related words acquired from the web. KnowNet, the resulting knowledge-base which connects large sets of semantically-related concepts is a major step towards the autonomous acquisition of knowledge from raw corpora. In fact, KnowNet is several times larger than any available knowledge resource encoding relations between synsets, and the knowledge KnowNet contains outperform any other resource when is empirically evaluated in a common multilingual framework.

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, hundreds 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 1995 to 2006, that is from version 1.5 to 3.0), WordNet passed from 103,445 to 235,402 semantic relations[1]. 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. Obviously, this fact has severely hampered the state-of-the-art of advanced NLP applications.

However, the Princeton WordNet is by far the most widely-used knowledge base (Fellbaum, 1998). In fact, WordNet is being used world-wide for anchoring different types of semantic knowledge including wordnets for languages other than English (Atserias et al., 2004), domain knowledge (Magnini and Cavaglià, 2000) or ontologies like SUMO (Niles and Pease, 2001) or the EuroWordNet Top Concept Ontology (Álvez et al., 2008). It contains manually coded information about nouns, verbs, adjectives and adverbs in English and is organised 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 organisation to gain political power" and by explicit semantic relations to other synsets.

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 knowledge about individuals from Wikipedia (Suchanek et al., 2007). Obviously, all these semantic resources have been acquired using a very different set of processes (Snow et al., 2006), tools and corpora. In fact, each semantic resource has different volume and

[1] Symmetric relations are counted only once.

Source	**#relations**
Princeton WN3.0	235,402
Selectional Preferences from SemCor	203,546
eXtended WN	550,922
Co-occurring relations from SemCor	932,008
New KnowNet-5	231,163
New KnowNet-10	689,610
New KnowNet-15	1,378,286
New KnowNet-20	2,358,927

Table 1: Number of synset relations

accuracy figures when evaluated in a common and controlled framework (Cuadros and Rigau, 2006).

However, not all available large-scale resources encode semantic relations between synsets. In some cases, only relations between synsets and words have been acquired. This is the case of the Topic Signatures(Agirre et al., 2000) acquired from the web (Agirre and de la Calle, 2004). This is one of the largest semantic resources ever built with around one hundred million relations between synsets and semantically related words [2].

A knowledge net or KnowNet, is an extensible, large and accurate knowledge base, which has been derived by semantically disambiguating the Topic Signatures acquired from the web. Basically, the method uses a robust and accurate knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate senses to the topic words associated to a particular synset. The resulting knowledge-base which connects large sets of topically-related concepts is a major step towards the autonomous acquisition of knowledge from raw text. In fact, KnowNet is several times larger than WordNet and the knowledge contained in KnowNet outperforms WordNet when empirically evaluated in a common framework.

Table 1 compares the different volumes of semantic relations between synset pairs of available knowledge bases and the newly created KnowNets[3].

Varying from five to twenty the number of processed words from each Topic Signature, we created automatically four different KnowNets with millions of new semantic relations between synsets.

After this introduction, section 2 describes the Topic Signatures acquired from the web. Section 3 presents the approach we plan to follow for building highly dense and accurate knowledge bases. Section 4 describes the methods we followed for building KnowNet. In section 5, we present the evaluation framework used in this study. Section 6 describes the results when evaluating different versions of KnowNet and finally, section 7 presents some concluding remarks and future work.

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:

- acquiring the best possible corpus examples for a particular word sense (usually characterising each word sense as a query and performing a search on the corpus for those examples that best match the queries)
- building the TS by deriving the context words that best represent the word sense from the selected corpora.

The Topic Signatures acquired from the web (hereinafter TSWEB) constitutes one of the largest available semantic resource acquired from the web with around 100 million relations (between synsets and words) (Agirre and de la Calle, 2004). Inspired by the work of (Leacock et al., 1998), TSWEB was 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 salient words with distinctive frequency using TFIDF. Thus, TSWEB consist of a large ordered list of words with weights associated to each of the polysemous nouns of WordNet 1.6. The number of constructed topic signatures is 35,250 with an average size per signature of 6,877 words. When evaluating TSWEB, we used at maximum the first 700 words while for building KnowNet we used at maximum the first 20 words.

For example, table 2 present the first words (lemmas and part-of-speech) and weights of the Topic Signature acquired for party#n#1.

3 A proposal for building highly connected and dense knowledge bases

It is our belief, that accurate semantic processing (such as WSD) would rely not only on so-

[2] Available at http://ixa.si.ehu.es/Ixa/resources/sensecorpus

[3] These KnowNet versions can be downloaded from http://adimen.si.ehu.es

tammany#n	0.0319
alinement#n	0.0316
federalist#n	0.0315
whig#n	0.0300
missionary#j	0.0229
Democratic#n	0.0218
nazi#j	0.0202
republican#n	0.0189
constitutional#n	0.0186
organization#n	0.0163

Table 2: TS of party#n#1 (first 10 out of 12,890 total words)

phisticated algorithms but on knowledge intensive approaches. In fact, the cycling arquitecture of the MEANING[4] project demonstrated that acquiring better knowledge allow to perform better Word Sense Disambiguation (WSD) and that having improved WSD systems we are able to acquire better knowledge (Rigau et al., 2002).

Thus, we plan to acquire by fully automatic means highly connected and dense knowledge bases from large corpora or the web by using the knowledge already available, increasing the total number of relations from less than one million (the current number of available relations) to millions.

The current proposal consist of:

- to follow (Cuadros et al., 2005) and (Cuadros and Rigau, 2006) for acquiring highly accurate Topic Signatures for all monosemous words in WordNet (for instance, using InfoMap (Dorow and Widdows, 2003)). That is, to acquire word vectors closely related to a particular monosemous word (for instance, airport#n#1) from BNC or other large text collections like GigaWord, Wikipedia or the web.

- to apply a very accurate knowledge–based all–words disambiguation algorithm to the Topic Signatures in order to obtain sense vectors instead of word vectors (for instance, using a version of Structural Semantic Interconnections algorithm (SSI) (Navigli and Velardi, 2005)).

For instance, consider the first ten weighted words (with Part-of-Speech) appearing in the Topic Signature (TS) of the word sense airport#n#1 corresponding to the monosemous word airport, as shown in table 3. This TS has been obtained from BNC using InfoMap. From the ten words appearing in the TS, two of them do not appear in WN (corresponding to the proper names heathrow#n and gatwick#n), four words are monosemous (airport#n, airfield#n, travelling#n and passenger#n) and four other are polysemous (flight#n, train#n, station#n and ferry#n).

word+pos	weight	#senses
airport#n	1.000000	1
heathrow#n	0.843162	0
gatwick#n	0.768215	0
flight#n	0.765804	9
airfield#n	0.740861	1
train#n	0.739805	6
travelling#n	0.732794	1
passenger#n	0.722912	1
station#n	0.722364	4
ferry#n	0.717653	2

Table 3: First ten words with weigths and number of senses in WN of the Topic Signature for airport#n#1 obtained from BNC using InfoMap

3.1 SSI-Dijkstra

We have implemented a version of the Structural Semantic Interconnections algorithm (SSI), a knowledge-based iterative approach to Word Sense Disambiguation (Navigli and Velardi, 2005). The SSI algorithm is very simple and consists of an initialisation step and a set of iterative steps. Given W, an ordered list of words to be disambiguated, the SSI algorithm performs as follows. During the initialisation step, all monosemous words are included into the set I of already interpreted words, and the polysemous words are included in P (all of them pending to be disambiguated). At each step, the set I is used to disambiguate one word of P, selecting the word sense which is closer to the set I of already disambiguated words. Once a sense is disambiguated, the word sense is removed from P and included into I. The algorithm finishes when no more pending words remain in P.

Initially, the list I of interpreted words should include the senses of the monosemous words in W, or a fixed set of word senses[5]. However, in this case, when disambiguating a TS derived from a monosemous word *m*, the list I includes since the beginning at least the sense of the monosemous word *m* (in our example, airport#n#1).

In order to measure the proximity of one synset (of the word to be disambiguated at each step) to a set of synsets (those word senses already in-

[4] http://www.lsi.upc.edu/˜nlp/meaning

[5] If no monosemous words are found or if no initial senses are provided, the algorithm could make an initial guess based on the most probable sense of the less ambiguous word of W.

Synsets	Distance
4	6
4530	5
64713	4
29767	3
597	2
20	1
1	0

Table 4: Minimum distances from airport#n#1

terpreted in I), the original SSI uses an in-house knowledge base derived semi-automatically which integrates a variety of online resources (Navigli, 2005). This very rich knowledge-base is used to calculate graph distances between synsets. In order to avoid the exponential explosion of possibilities, not all paths are considered. They used a context-free grammar of relations trained on SemCor to filter-out inappropriate paths and to provide weights to the appropriate paths.

Instead, we use part of the knowledge already available to build a very large connected graph with 99,635 nodes (synsets) and 636,077 edges (the set of direct relations between synsets gathered from WordNet and eXtended WordNet). On that graph, we used a very efficient graph library to compute the Dijkstra algorithm. The Dijkstra algorithm is a greedy algorithm that computes the shortest path distance between one node an the rest of nodes of a graph. In that way, we can compute very efficiently the shortest distance between any two given nodes of a graph. We call this version of the SSI algorithm, SSI-Dijkstra.

For instance, table 4 shows the minimum distances from airport#n#1 to the rest of the synsets of the graph. Interestingly, from airport#n#1 all synsets of the graph are accessible following paths of at maximum six edges. While there is only one synset at distance zero (airport#n#1) and twenty synsets directly connected to airport#n#1, 95% of the total graph is accessible at distance four or less.

SSI-Dijkstra has very interesting properties. For instance, it always provides the minimum distance between two synsets. That is, the Dijkstra algorithm always provides an answer being the minimum distance close or far[6]. In fact, the SSI-Dijkstra algorithm compares the distances between the synsets of a word and all the synsets already interpreted in I. At each step, the SSI-Dijkstra algorithm selects the synset which is closer to I (the set of already interpreted words).

Table 5 presents the result of the word–sense disambiguation process with the SSI-Dijkstra algorithm on the TS presented in table 3[7]. Now, part of the TS obtained from BNC using InfoMap have been disambiguated at a synset level resulting on a word–sense disambiguated TS. Those words not present in WN1.6 have been ignored (heathrow and gatwick). Some others, being monosemous in WordNet were considered already disambiguated (travelling, passenger, airport and airfield). But the rest, have been correctly disambiguated (flight with nine senses, train with six senses, station with four and ferry with two).

This sense disambiguated TS represents seven direct new semantic relations between airport#n#1 and the first words of the TS. It could be directly integrated into a new knowledge base (for instance, airport#n#1 –related–> flight#n#9), but also all the indirect relations of the disambiguated TS (for instance, flight#n#9 –related–> travelling#n#1). In that way, having n disambiguated word senses, a total of $(n^2 - n)/2$ relations could be created. That is, for the ten initial words of the TS of airport#n#1, twenty-eight new direct relations between synsets could be created.

This process could be repeated for all monosemous words of WordNet appearing in the selected corpus. The total number of monosemous words in WN1.6 is 98,953. Obviously, not all these monosemous words are expected to appear in the corpus. However, we expect to obtain in that way several millions of new semantic relations between synsets. This method will allow to derive by fully automatic means a huge knowledge base with millions of new semantic relations.

Furthermore, this approach is completely language independent. It could be repeated for any language having words connected to WordNet.

It remains for further study and research, how to convert the relations created in that way to more specific and labelled relations.

4 Building KnowNet

As a proof of concept, we developed KnowNet (KN), a large-scale and extensible knowledge base obtained by applying SSI-Dijkstra to each topic signature from TSWEB. That is, instead of using InfoMap and a large corpora for acquiring new Topic Signatures for all the monosemous terms

[6] In contrast, the original SSI algorithm not always provides a path distance because it depends on the grammar.

[7] It took 4.6 seconds to disambiguate the TS on a modern personal computer.

word	**offset-WN**	**weight**	**gloss**
flight#n	00195002n	0.017	a scheduled trip by plane between designated airports
travelling#n	00191846n	0	the act of going from one place to another
train#n	03528724n	0.012	a line of railway cars coupled together and drawn by a locomotive
passenger#n	07460409n	0	a person travelling in a vehicle (a boat or bus or car or plane or train etc) who is not operating it
station#n	03404271n	0.019	a building equipped with special equipment and personnel for a particular purpose
airport#n	02175180n	0	an airfield equipped with control tower and hangers as well as accommodations for passengers and cargo
ferry#n	02671945n	0.010	a boat that transports people or vehicles across a body of water and operates on a regular schedule
airfield#n	02171984n	0	a place where planes take off and land

Table 5: Sense disambiguated TS for airport#n#1 obtained from BNC using InfoMap and SSI-Dijkstra

in WN, we used the already available TSWEB. We have generated four different versions of KonwNet applying SSI-Dijkstra to the first 5, 10, 15 and 20 words for each TS. SSI-Dijkstra used only the knowledge present in WordNet and eXtended WordNet which consist of a very large connected graph with 99,635 nodes (synsets) and 636,077 edges (semantic relations).

We generated each KnowNet by applying the SSI-Dijkstra algorithm to the whole TSWEB (processing the first words of each of the 35,250 topic signatures). For each TS, we obtained the direct and indirect relations from the topic (a word sense) to the disambiguated word senses of the TS. Then, as explained in section 3, we also generated the indirect relations for each TS. Finally, we removed symmetric and repeated relations.

Table 6 shows the percentage of the overlapping between each KnowNet with respect the knowledge contained into WordNet and eXtended WordNet, the total number of relations and synsets of each resource. For instance, only an 8,6% of the total relations included into WN+XWN are also present in KnowNet-20. This means that the rest of relations from KnowNet-20 are new. This table also shows the different KnowNet volumes.

As expected, each KnowNet is very large, ranging from hundreds of thousands to millions of new semantic relations between synsets among increasing sets of synsets. Surprisingly, the overlapping between the semantic relations of KnowNet and the knowledge bases used for building the SSI-Dijkstra graph (WordNet and eXtended WordNet) is very small, possibly indicating disjunct types of knowledge.

KB	**WN+XWN**	**#relations**	**#synsets**
KN-5	3.2%	231,164	39,837
KN-10	5.4%	689,610	45,770
KN-15	7.0%	1,378,286	48,461
KN-20	8.6%	2,358,927	50,705

Table 6: Size and percentage of overlapping relations between KnowNet versions and WN+XWN

overlapping	**KN-5**	**KN-10**	**KN-15**	**KN-20**
KN-5	100	93,3	97,7	97,2
KN-10	31,2	100	88,5	88,9
KN-15	16,4	44,4	100	97.14
KN-20	9,5	26,0	56,7	100

Table 7: Percentage of overlapping relations between KnowNet versions

Table 7 presents the percentage of overlapping relations between KnowNet versions. The upper triangular part of the matrix presents the overlapping percentage covered by larger KnowNet versions.That is most of the knowledge from KnowNet-5 is also in larger versions of KnowNet. Interestingly, the knowledge contained into KnowNet-10 is only partially covered by KnowNet-15 and KnowNet-20. The lower triangular part of the matrix presents the overlapping percentage covered by smaller KnowNet versions.

5 Evaluation framework

In order to empirically establish the relative quality of these KnowNet versions with respect already available semantic resources, we used the noun-set of Senseval-3 English Lexical Sample task which consists of 20 nouns.

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 resource. Thus, all the semantic resources studied are evaluated as Topic Signatures. That is, word vectors with weights associated to a particular synset (topic) which are obtained by collecting those word senses appearing in the synsets directly related to the topics.

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[8]. 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. All performances are evaluated on the test data using the fine-grained scoring system provided by the organisers. 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.

5.1 Baselines

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: This baseline selects the most frequent sense of the target word in SemCor.

WN-MFS: This baseline is obtained by selecting the most frequent sense (the first sense in WN1.6) of the target word. WordNet 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 baseline selects the most frequent sense in the training corpus of the target word.

TRAIN: This baseline uses the training corpus to directly build a Topic Signature using TFIDF measure for each word sense. Note that in WSD evaluation frameworks, this is a very basic baseline. However, in our evaluation framework, this "WSD baseline" could be considered as an upper-bound. We do not expect to obtain better topic signatures for a particular sense than from its own annotated corpus.

5.2 Large-scale Knowledge Resources

In order to measure the relative quality of the new resources, we include in the evaluation a wide range of large-scale knowledge resources connected to WordNet.

WN (Fellbaum, 1998): This resource uses the different direct relations encoded in WN1.6 and 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 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 for subjects and objects from BNC.

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

MCR (Atserias et al., 2004): This resource uses the direct relations of WN, XWN and spSemCor (we excluded spBNC because of its poor performance).

TSSEM (Cuadros et al., 2007): 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.

6 KnowNet Evaluation

We evaluated KnowNet using the same framework explained in section 5. That is, the noun part of the test set from the Senseval-3 English lexical sample task.

6.1 Senseval-3 evaluation

Table 8 presents ordered by F1 measure, the performance in terms of precision (P), recall (R) and F1 measure (F1, harmonic mean of recall and precision) of each knowledge resource on Senseval-3 and its average size of the TS per word-sense. The different KnowNet versions appear marked in bold and the baselines appear in italics.

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

[8] We also consider the multiword terms.

training corpus (TRAIN-MFS). However, all of them are far below to the Topic Signatures acquired using the training corpus (TRAIN).

The best resources would be those obtaining better performances with a smaller number of related words per synset. 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). Interestingly, 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 better results than SEMCOR-MFS and is very close to the most frequent sense of WN (WN-MFS) and the training (TRAIN-MFS).

The different versions of KnowNet consistently obtain better performances as they increase the window size of processed words of TSWEB. As expected, KnowNet-5 obtain the lower results. However, it performs better than WN (and all its extensions) and spBNC. Interestingly, from KnowNet-10, all KnowNet versions surpass the knowledge resources used for their construction (WN, XWN, TSWEB and WN+XWN).

Regarding the integration of resources, WN+XWN+KN-20 performs better than MCR and similarly to MCR^2 (having less than 50 times its size). Also interesting is that WN+XWN+KN-20 have better performance than their individual resources, indicating a complementary knowledge. In fact, WN+XWN+KN-20 performs much better than the resources from which it derives (WN, XWN and TSWEB).

These initial results seem to be very promising. If we do not consider the resources derived from manually sense annotated data (spSemCor, MCR, TSSEM, etc.), KnowNet-10 performs better that any knowledge resource derived by manual or automatic means. In fact, KnowNet-15 and KnowNet-20 outperforms spSemCor which was derived from manually annotated corpora. This is a very interesting result since these KnowNet versions have been derived only with the knowledge coming from WN and the web (that is, TSWEB), and WN and XWN as a knowledge source for SSI-Dijkstra (eXtended WordNet only has 17,185 manually labelled senses).

KB	**P**	**R**	**F1**	**Av. Size**
TRAIN	*65.1*	*65.1*	*65.1*	450
TRAIN-MFS	*54.5*	*54.5*	*54.5*	
WN-MFS	*53.0*	*53.0*	*53.0*	
TSSEM	52.5	52.4	52.4	103
SEMCOR-MFS	*49.0*	*49.1*	*49.0*	
MCR^2	45.1	45.1	45.1	26,429
MCR	45.3	43.7	44.5	129
KnowNet-20	44.1	44.1	44.1	610
KnowNet-15	43.9	43.9	43.9	339
spSemCor	43.1	38.7	40.8	56
KnowNet-10	40.1	40.0	40.0	154
(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
KnowNet-5	35.0	35.0	35.0	44
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
RANDOM	*19.1*	*19.1*	*19.1*	

Table 8: P, R and F1 fine-grained results for the resources evaluated at Senseval-3, English Lexical Sample Task.

7 Conclusions and future research

The initial results obtained for the different versions of KnowNet seem to be very promising, since they seem to be of a better quality than the rest of available knowledge resources which encode relations between synsets.

In fact, this is a preliminary step towards obtaining new large-scale knowledge resources of a better quality from monosemous words using a tool like InfoMap as explained in section 3.

We also tested all these resources and the different versions of KnowNet on SemEval-2007 English Lexical Sample Task (Cuadros and Rigau, 2008a). When comparing the ranking of the different knowledge resources, the different versions of KnowNet seem to be more robust and stable across corpora changes than the rest of resources. Furthermore, we also tested the performance of KnowNet when ported to Spanish (as the Spanish WordNet is also integrated into the MCR). Starting from KnowNet-10, all KnowNet versions perform better than any other knowledge resource on Spanish derived by manual or automatic means (including the MCR) (Cuadros and

Rigau, 2008b).

Acknowledgments

We want to thank Aitor Soroa for his technical support and the anonymous reviewers for their comments. This work has been supported by KNOW (TIN2006-15049-C03-01) and KYOTO (ICT-2007-211423).

References

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

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

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

Agirre, E., O. Ansa, D. Martinez, and E. Hovy. 2000. Enriching very large ontologies with topic signatures. In *Proceedings of ECAI'00 workshop on Ontology Learning*, Berlin, Germany.

Álvez, J., J. Atserias, J. Carrera, S. Climent, A. Oliver, and G. Rigau. 2008. Consistent annotation of eurowordnet with the top concept ontology. In *Proceedings of Fourth International WordNet Conference (GWC'08)*.

Atserias, J., 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.

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

Cuadros, M. and G. Rigau. 2008a. KnowNet: Building a Large Net of Knowledge from the Web. In *Proceedings of COLING*.

Cuadros, M. and G. Rigau. 2008b. Multilingual Evaluation of KnowNet. In *Proceedings of SEPLN*.

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

Cuadros, M., G. Rigau, and M. Castillo. 2007. Evaluating large-scale knowledge resources across languages. In *Proceedings of RANLP*.

Dorow, B. and D. Widdows. 2003. Discovering corpus-specific word senses. In *EACL*, Budapest.

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

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

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

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

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

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

Navigli, R. 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.

Navigli, R. 2005. Semi-automatic extension of large-scale linguistic knowledge bases. In *Proc. of 18th FLAIRS International Conference (FLAIRS)*, Clearwater Beach, Florida.

Niles, I. 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.

Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll. 2002. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING'2002 Workshop on A Roadmap for Computational Linguistics*, Taipei, Taiwan.

Snow, R., D. Jurafsky, and A. Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING-ACL.*

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA. ACM Press.

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