Procesamiento del lenguaje Natural, nº 41 (2008), pp. 121-128 recibido 7-05-2008; aceptado 16-06-2008

Multilingual Evaluation of KnowNet

Evaluación Multilíngüe de KnowNet

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

Resumen: Este artículo presenta un nuevo método totalmente automático de construcción de bases de conocimiento muy densas y precisas a partir de recursos semánticos preexistentes. Básicamente, el método usa un algoritmo de Interpretación Semántica de las palabras preciso y de amplia cobertura para asignar el sentido más apropiado a grandes conjuntos de palabras de un mismo tópico que han sido obtenidas de la web. KnowNet, la base de conocimiento resultante que conecta grandes conjuntos de conceptos semánticamente relacionados es un paso importante hacia la adquisición automática de conocimiento a partir de corpus. De hecho, KnowNet es varias veces más grande que cualquier otro recurso de conocimiento disponible que codifique relaciones entre sentidos, y el conocimiento que KnowNet contiene supera cualquier otro recurso cuando es empíricamente evaluado en un marco multilingüe común.
Palabras clave: Bases de Conocimiento de amplia cobertura, Interpretación Semántica de las Palabras, Adquisición de Conocimiento.

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.
Keywords: Large-Scale Knowledge Resources, Word Sense Disambiguation, Knowledge Acquisition

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 grew from 103,445 to 235,402 semantic relations(Symmetric relations are counted only once). 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 (WN) 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.,

ISSN: 1135-5948

2008). It contains manually coded information about English nouns, verbs, adjectives and adverbs 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" 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 Lacalle, 2004) or knowledge about individuals from Wikipedia (Suchanek, Kasneci, and Weikum, 2007). Obviously, all these semantic resources have been acquired using a very different set of processes, tools and corpora. As expected, each semantic resource has different volume and accuracy figures when evaluated in a common and controlled framework (Cuadros and Rigau, 2006).

However, not all these 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 acquired from the web (Agirre and de Lacalle, 2004). This is one of the largest semantic resources ever build with around one hundred million relations between synsets and semantically related words (http://ixa.si.ehu.es/Ixa/resources/sensecorpus).

A knowledge net or KnowNet (KN), is an extensible, large and accurate knowledge base, which has been derived by semantically disambiguating small portions of 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.

Knowledge Resources	**#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
New KnowNet-5 (es)	144,493
New KnowNet-10 (es)	447,317
New KnowNet-15 (es)	922,256
New KnowNet-20 (es)	1,606,893

Table 1: Number of synset relations

Table 1 compares the different volumes of semantic relations between synset pairs of available knowledge bases and the newly created KnowNets in English and its ported relations to Spanish (es)(These KnowNet versions are available at http://adimen.si.ehu.es).

Variying 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. In fact, KnowNet is several times larger than WordNet, and when evaluated empirically across languages, the knowledge it contains outperforms any other semantic resource.

After this introduction, section 2 describes the Topic Signatures acquired from the web. Section 3 presents the approach we followed for building highly dense and accurate knowledge bases from the Topic Signatures. In section 4, we present the evaluation framework used in this study and we describe the results when evaluating in a multilingual framework different versions of KnowNet for English and Spanish. Finally, section 5 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

tammany#n	0.0319
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
conservative#j	0.0148
socialist#n	0.0140

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

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 selecting 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 semantic resource available with around 100 million relations (between synsets and words) (Agirre and de Lacalle, 2004). Inspired by the work of (Leacock, Chodorow, and Miller, 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 WN1.6. 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 presents the first words (lemmas and part-of-speech) and weights of the Topic Signature acquired for party#n#1(This format stands for word#pos#sense).

3 Building highly connected and dense knowledge bases

We acquired by fully automatic means highly connected and dense knowledge bases by disambiguating small portions of the Topic Signatures obtained from the web, increasing the total number of semantic relations from less than one million (the current number of available relations) to millions of new and accurate semantic relations between synsets. We applied a knowledge–based all–words Word Sense Disambiguation algorithm to the Topic Signatures for deriving a sense vector from each word vector.

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 (Cuadros and Rigau, to appear 2008). The SSI algorithm is very simple and consists of an initialization step and a set of iterative steps (Navigli and Velardi, 2005).

Given W, an ordered list of words to be disambiguated, the SSI algorithm performs as follows. During the initialization 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 selected, 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 (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). However, when disambiguating a TS of a word sense s (for instance party#n#1), the list I already includes s.

In order to measure the proximity of one synset to the rest of synsets of I, we use part of the knowledge already available to build a very large connected graph with 99,635 nodes (synsets) and 636,077 edges. This graph includes the set of direct relations between synsets gathered from WordNet and eXtended WordNet. On that graph, we used a very efficient graph library, BoostGraph(http://www.boost.org) to compute the Dijkstra algorithm. The Dijkstra algorithm is a greedy algorithm for computing 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.

SSI-Dijkstra has very interesting properties. For instance, always provides the minimum distance between two synsets. That is, the algorithm always provides an answer being the minimum distance close or far. In contrast, the original SSI algorithm not always provides a path distance because it depends on a predefined grammar of semantic relations. 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).

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

3.2 Building KnowNet

We developed KnowNet (KN), a large-scale and extensible knowledge base by applying SSI-Dijkstra to each topic signature from TSWEB.

We have generated four different versions of KnowNet applying SSI-Dijkstra to only 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 KN 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 relations from the topic (a word sense) to the disambiguated word senses of the TS (for instance, party#n#1 –> federalist#n#1), but also the indirect relations between disambiguated words from the TS (for instance, federalist#n#1 –> republican#n#1). Finally, we removed symmetric and repeated relations.

Table 3 shows the percentage of the overlaping 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 8,5% 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. As expected, each KnowNet is very large, ranging from hundreds of thousands to millions of new semantic relations among increasing sets of synsets.

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

4 *Evaluation framework*

In order to empirically establish the relative quality of these new semantic resources, we used the evaluation framework of task 16 of SemEval-2007: Evaluation of wide coverage knowledge resources (Cuadros and Rigau, 2007).

All knowledge resources are evaluated on a WSD task. In particular, in section 4.5 we used the noun-set of Senseval-3 English Lexical Sample task which consists of 20 nouns and in section 4.6 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 (Atserias et al., 2004)). 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 (Mihalcea, T.Chlovski, and A.Killgariff, 2004) as a sense repository for verbs instead of WordNet.

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. All knowledge bases are evaluated as topic signatures. That is, word vectors with weights associated to a particular synset which are obtained by collecting those word senses appearing in the synsets directly related to the topics. This simple representation tries to be as neutral as possible with respect to the resources used.

A common WSD method has been applied to all knowledge resources. A simple word overlapping counting is performed between the topic signature representing a word sense and the test example (We also consider those multiword terms appearing in WN). 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.

4.1 KnowNet Evaluation

We evaluated KnowNet using the same framework explained in section 4. That is, the noun part of the test set from the English and Spanish Senseval-3 lexical sample tasks.

4.2 English Baselines

We have designed a number of baselines in order to establish a complete evaluation framework for comparing the performance of each semantic resource when evaluated 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 and selecting at maximum the first 450 words. 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 his own annotated corpus.

4.3 Spanish Baselines

As well as for English, we have designed a number of 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. Again, this baseline can be considered as a lower-bound.

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: 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 (http://multisemcor.itc.it/).

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 original synset).

4.4 Other 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.

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 integrates the direct relations of WN, XWN and spSemCor.

TSSEM (Cuadros, Rigau, and Castillo, 2007): These Topic Signatures have been constructed using SemCor. 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.

4.5 Evaluating each resource in English

Table 4 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 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

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

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 when increasing the window size of processed words of TSWEB. As expected, KnowNet-5 obtains 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).

These initial results seem to be very promising. If we do not consider the re-

KB	P	R	F1	Av. S
TRAIN	*81.8*	*68.0*	*74.3*	450
MiniDir-MFS	*67.1*	*52.7*	*59.2*	
KnowNet-15	54.7	48.9	**51.6**	176
KnowNet-20	51.8	**49.6**	50.7	319
KnowNet-10	53.5	43.1	47.7	81
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
KnowNet-5	58.5	26.9	36.8	22
TSSEM	33.6	33.2	33.4	208
XWN	42.6	27.1	33.1	24
WN	**65.5**	13.6	22.5	8
RANDOM	21.3	21.3	21.3	

Table 5: P, R and F1 fine-grained results for the resources evaluated individually on Spanish.

sources 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 labeled senses).

4.6 Evaluating each resource on Spanish

Table 5 presents ordered by F1 measure, the performance of each knowledge resource on the Senseval-3 Spanish Lexical Sample task and its average size of the TS per word-sense. Obviously, the average size in this case is also different with respect the English evaluations. The best results for precision, recall and F1 measures are shown in bold. We also mark in italics the results 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).

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), 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.

Interestingly, the knowledge contained into the MCR (F1 of 43.5), partially derived by automatic means and ported from English resources, almost doubles the results of the original Spanish WN (F1 of 22.5).

Regarding the KnowNet versions ported to Spanish, KnowNet-5 performs better than WN, XWN and the TS acquired from SemCor. 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). Interestingly, the best result is obtained by the ported relations of KnowNet-15 which performs slightly better than KnowNet-20 (while having much less relations).

5 *Conclusions and future research*

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

The initial results obtained for the different versions of KnowNet seem to be a major step towards the autonomous acquisition of knowledge from raw corpora, since they are several times larger than the available knowledge resources which encode relations between synsets, and the knowledge they contain outperform any other resource when is empirically evaluated in a common multilingual framework. In fact, when comparing the ranking of the different knowledge resources, the different versions of KnowNet seem to be more robust and stable across languages.

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 Lacalle. 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.

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. 2007. Semeval-2007 task 16: Evaluation of wide coverage knowledge resources. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).*

Cuadros, M. and G. Rigau. to appear 2008. KnowNet: Building a large net of knowledge from the web. In *COLING'08.*

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

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.

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

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.

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.

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 .