A Proposal of Automatic Selection of Coarse-grained Semantic Classes for WSD*

Rubén Izquierdo & Armando Suárez
GPLSI. Departament de LSI. UA. Alacant, Spain.
{ruben,armando}@dlsi.ua.es

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

Resumen: Presentamos un método muy simple para seleccionar conceptos base (Base Level Concepts) usando algunas propiedades estructurales básicas de WordNet. Demostramos empíricamente que el conjunto de Base Level Concepts obtenido agrupa sentidos de palabras en un nivel de abstracción adecuado para la desambiguación del sentido de las palabras basada en clases. De hecho, un sencillo clasificador basado en el sentido más frecuente usando las clases generadas, es capaz de alcanzar un acierto próximo a 75% para la tarea de etiquetado semántico.

Palabras clave: WordNet, Sentidos de las palabras, niveles de abstracción, Desambiguación del Sentido de las Palabras

Abstract: We present a very simple method for selecting Base Level Concepts using some basic structural properties of WordNet. We also empirically demonstrate that these automatically derived set of Base Level Concepts group senses into an adequate level of abstraction in order to perform class-based Word Sense Disambiguation. In fact, a very naive Most Frequent classifier using the classes selected is able to perform a semantic tagging with accuracy figures over 75%.

Keywords: WordNet, word-senses, levels of abstraction, Word Sense Disambiguation

1 Introduction

Word Sense Disambiguation (WSD) is an intermediate Natural Language Processing (NLP) task which consists in assigning the correct semantic interpretation to ambiguous words in context. One of the most successful approaches in the last years is the *supervised learning from examples*, in which statistical or Machine Learning classification models are induced from semantically annotated corpora (Màrquez et al., 2006). Generally, supervised systems have obtained better results than the unsupervised ones, as shown by experimental work and international evaluation exercises such as Senseval[1]. These annotated corpora are usually manually tagged by lexicographers with word senses taken from a particular lexical semantic resource –most commonly WordNet (WN) (Fellbaum, 1998).

WN has been widely criticised for being a sense repository that often offers too fine–grained sense distinctions for higher level applications like Machine Translation or Question & Answering. In fact, WSD at this level of granularity, has resisted all attempts of infering robust broad-coverage models. It seems that many word–sense distinctions are too subtle to be captured by automatic systems with the current small volumes of word–sense annotated examples. Possibly, building class-based classifiers would allow to avoid the data sparseness problem of the word-based approach. Recently, using WN as a sense repository, the organizers of the English all-words task at SensEval-3 reported an inter-annotation agreement of 72.5% (Snyder and Palmer, 2004). Interestingly, this result is difficult to outperform by state-of-the-art fine-grained WSD systems.

Thus, some research has been focused on deriving different sense groupings to overcome the fine–grained distinctions of WN (Hearst and Schütze, 1993) (Peters, Peters, and Vossen, 1998) (Mihalcea and Moldovan, 2001) (Agirre, Aldezabal, and Pociello, 2003) and on using predefined sets of sense-groupings for learning class-based classifiers for WSD (Segond et al., 1997) (Ciaramita and Johnson, 2003) (Villarejo, Màrquez, and Rigau, 2005) (Curran, 2005) (Ciaramita and Altun, 2006). However, most of the later approaches used the original Lexicographical Files of WN (more recently called Supersenses) as very coarse–grained sense distinctions. However, not so much attention has been paid on learning class-based classifiers from other available

* This paper has been supported by the European Union under the project QALL-ME (FP6 IST-033860) and the Spanish Government under the project Text-Mess (TIN2006-15265-C06-01) and KNOW (TIN2006-15049-C03-01)

[1] http://www.senseval.org

sense–groupings such as WordNet Domains (Magnini and Cavaglia, 2000), SUMO labels (Niles and Pease, 2001), EuroWordNet Base Concepts (Vossen et al., 1998) or Top Concept Ontology labels (Atserias et al., 2004). Obviously, these resources relate senses at some level of abstraction using different semantic criteria and properties that could be of interest for WSD. Possibly, their combination could improve the overall results since they offer different semantic perspectives of the data. Furthermore, to our knowledge, to date no comparative evaluation have been performed exploring different sense–groupings.

We present a very simple method for selecting Base Level Concepts (Rosch, 1977) using basic structural properties of WN. We also empirically demonstrate that these automatically derived set of Base Level Concepts group senses into an adequate level of abstraction in order to perform class-based WSD.

This paper is organized as follows. Section 2 introduce the different levels of abstraction that are relevant for this study, and the available sets of semi-automatically derived Base Concepts. In section 3, we present the method for deriving fully automatically a number of Base Level Concepts from any WN version. Section 4 reports the resulting figures of a direct comparison of the resources studied. Section 5 provides an empirical evaluation of the performance of the different levels of abstraction. In section 6 we provide further insights of the results obtained and finally, in section 7 some concluding remarks are provided.

2 Levels of abstraction

WordNet[2] (WN) (Fellbaum, 1998) is an online lexical database of English which contains concepts represented by synsets, sets of synonyms of content words (nouns, verbs, adjectives and adverbs). In WN, different types of lexical and semantic relations interlink different synsets, creating in this way a very large structured lexical and semantic network. The most important relation encoded in WN is the subclass relation (for nouns the hyponymy relation and for verbs the troponymy relation). The last version of WN, WN 3.0, was released on december 2006. It contains 117,097 nouns and 11,488 verbs, organized into 81,426 noun synsets and 13,650 verb synsets.

EuroWordNet[3] (EWN) (Vossen et al., 1998) is a multilingual database than contains wordnets for several languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). Each of these single wordnets represent a unique language-internal system of lexicalizations, and it is structured following the approach of English wordnet: synsets and relations between them. Different wordnets are linked to the Inter-Lingual-Index (ILI), based on Princeton English WN. By means of the ILI, synsets and words or different languages are connected, allowing advanced multilingual natural language applications (Vossen et al., 2006).

The notion of Base Concepts (hereinafter BC) was introduced in EuroWordNet. The BC are supposed to be the concepts that play the most important role in the various wordnets of different languages. This role was measured in terms of two main criteria: a high position in the semantic hierarchy and having many relations to other concepts. Thus, the BC are the fundamental building blocks for establishing the relations in a wordnet. In that sense, the Lexicografic Files (or Supersenses) of WN could be considered the most basic set of BC.

Basic Level Concepts (Rosch, 1977) (hereinafter BLC) should not be confused with Base Concepts. BLC are a compromise between two conflicting principles of characterization: a) to represent as many concepts as possible (abstract concepts), and b) to represent as many distinctive features as possible (concrete concepts).

As a result of this, Basic Level Concepts typically occur in the middle of hierarchies and less than the maximum number of relations. BC mostly involve the first principle of the Basic Level Concepts only. BC are generalizations of features or semantic components and thus apply to a maximum number of concepts. Our work focuses on devising simple methods for selecting automatically an accurate set of Basic Level Concepts from WN.

2.1 WordNet Base Concepts

WN synsets are organized in forty five Lexicographer Files, or **SuperSenses**, based on syntactic categories (nouns, verbs, adjectives and adverbs) and logical groupings, such as person, phenomenon, feeling, location, etc. There are 26 basic categories for nouns, 15 for verbs, 3 for adjectives and 1 for adverbs. For instance, the Supersenses corresponding to the four senses of the noun *church* in WN1.6 are *noun.group* for the first *Christian Church* sense, *noun.artifact* for the second *church_building* sense and *noun.act* for the third *church_service* sense.

2.2 EuroWordNet Base Concepts

Within EuroWordNet, a set of Base Concepts was selected to reach maximum overlap and compatibility across wordnets in different languages following the two main criteria described above: a high position in the semantic hierarchy and having many relations to other concepts. Initially, a set of 1,024 Common Base Concepts from WN1.5 (concepts acting as BC in at least two languages) was selected, only considering English, Dutch, Spanish and Italian wordnets.

[2] http://wordnet.princeton.edu

[3] http://www.illc.uva.nl/EuroWordNet/

2.3 Balkanet Base Concepts

The Balkanet project[4] followed a similar approach to EWN, but using other languages: Greek, Romanian, Serbian, Turkish and Bulgarian. The goal of Balkanet was to develop a multilingual lexical database for the new languages following the guidelines of EWN. Thus, the Balkanet project selected his own list of BC extending the original set of BC of EWN to a final set of 4,698 ILI records from WN2.0[5] (3,210 nouns, 1,442 verbs and 37 adjectives).

2.4 MEANING Base Concepts

The MEANING project[6] also followed the architectural model proposed by the EWN to build the Multilingual Central Repository (Mcr) (Atserias et al., 2004). In this case, BC from EWN based on WN1.5 synsets were ported to WN1.6. The number of BC finally selected was 1,535 (793 for nouns and 742 for verbs).

3 Automatic Selection of Base Level Concepts

This section describes a simple method for deriving a set of Base Level Concepts (BLC) from WN. The method has been applied to different WN versions for nouns and verbs. Basically, to select the appropriate BLC of a particular synset, the algorithm only considers the relative number of relations of their hypernyms. We derived two different sets of BLC depending on the type of relations considered: a) all types of relations encoded in WN (All) and b) only the hyponymy relations encoded in WN (Hypo).

The process follows a bottom-up approach using the chain of hypernym relations. For each synset in WN, the process selects as its Base Level Concept the first local maximum according to the relative number of relations. For synsets having multiple hypernyms, the path having the local maximum with higher number of relations is selected. Usually, this process finishes having a number of “fake” Base Level Concepts. That is, synsets having no descendants (or with a very small number) but being the first local maximum according to the number of relations considered. Thus, the process finishes checking if the number of concepts subsumed by the preliminary list of BLC is higher than a certain threshold. For those BLC not representing enough concepts according to a certain threshold, the process selects the next local maximum following the hypernym hierarchy. Thus, depending on the type of relations considered to be counted and the threshold established, different sets of BLC can be easily obtained for each WN version.

An example is provided in table 1. This table shows the possible BLC for the noun “church” using WN1.6. The table presents the hypernym chain for each synset together with the number of relations encoded in WN for the synset. The local maxima along the hypernym chain of each synset appears in bold. For **church_1** the synset with 12 total relations *faith_3* will be selected. The second sense of church, **church_2** is a local maximum with 19 total relations. This synset will be selected if the number of descending synsets having **church_2** as a Base Level Concept is higher than a predefined threshold. Finally, the selected Base Level Concept for **church_3** is *religious_ceremony_1*. Obvioulsy, different criteria will select a different set of Base Level Concepts.

#rel.	**synset**
18	group_1,grouping_1
19	social_group_1
37	organisation_2,organization_1
10	establishment_2,institution_1
12	faith_3,religion_2
5	Christianity_2,**church_1**,Christian_church_1

#rel.	**synset**
14	entity_1,something_1
29	object_1,physical_object_1
39	artifact_1,artefact_1
63	construction_3,structure_1
79	building_1,edifice_1
11	place_of_worship_1, ...
19	**church_2**,church_building_1

#rel.	**synset**
20	act_2,human_action_1,human_activity_1
69	activity_1
5	ceremony_3
11	religious_ceremony_1,religious_ritual_1
7	service_3,religious_service_1,divine_service_1
1	**church_3**,church_service_1

Table 1: Possible Base Level Concepts for the noun *Church* in WN1.6

Instead of highly related concepts, we also considered highly frequent concepts as possible indicator of a large set of features. Following the same basic algorithm, we also used the relative frequency of the synsets in the hypernym chain. That is, we derived two other different sets of BLC depending on the source of relative frequencies considered: a) the frequency counts in SemCor (FreqSC) and b) the frequency counts appearing in WN (FreqWN). The frequency of a synset has been obtained summing up the frequencies of its word senses. In fact, WN word-senses were ranked using SemCor and other sense-annotated corpora. Thus, the frequencies of SemCor and WN are similar, but not equal.

4 Comparing Base Level Concepts

Different sets of Base Level Concepts (BLC) have been generated using different WN versions, types of relations (All and Hypo), sense frequencies (FreqSC and FrecWN) and thresholds.

Table 2 presents the total number of BLC and its average depth for WN1.6[7] varying the threshold and the type of relations considered (All or Hypo).

As expected, when increasing the threshold, the total number of automatic BLC and its ave-

[4] http://www.ceid.upatras.gr/Balkanet

[5] http://www.globalwordnet.org/gwa/5000_bc.zip

[6] http://www.lsi.upc.es/~nlp/meaning

[7] WN1.6 have 66,025 nominal and 12,127 verbal synsets.

Thres.	Rel.	PoS	#BLC	Av. depth.
0	all	Noun	3,094	7.09
		Verb	1,256	3.32
	hypo	Noun	2,490	7.09
		Verb	1,041	3.31
10	all	Noun	971	6.20
		Verb	719	1.39
	hypo	Noun	993	6.23
		Verb	718	1.36
20	all	Noun	558	5.81
		Verb	673	1.25
	hypo	Noun	558	5.80
		Verb	672	1.21
50	all	Noun	253	5.21
		Verb	633	1.13
	hypo	Noun	248	5.21
		Verb	633	1.10

Table 2: Automatic Base Level Concepts for WN1.6 using All or Hypo relations

Thres.	Rel.	PoS	#BLC	Av. depth.
0	SemCor	Noun	34,865	7.44
		Verb	3,070	3.41
	WN	Noun	34,183	7.44
		Verb	2,615	3.30
10	SemCor	Noun	690	5.74
		Verb	731	1.38
	WN	Noun	691	5.77
		Verb	738	1.40
20	SemCor	Noun	339	5.43
		Verb	659	1.22
	WN	Noun	340	5.47
		Verb	667	1.23
50	SemCor	Noun	94	4.35
		Verb	630	1.12
	WN	Noun	99	4.41
		Verb	631	1.12

Table 3: Automatic Base Level Concepts for WN1.6 using SemCor or WN frequencies

rage depth decrease. For instance, using all relations on the nominal part of WN, the total number of BLC ranges from 3,094 (no threshold) to 253 (threshold 50). Using hyponym relations, the total number of BLC ranges from 2,490 (no threshold) to 248. However, although the number of total BLC for nouns decreases dramatically (around 10 times), the average depth of the synsets selected only ranges from 7.09 (no threshold) to 5.21 (threshold 50) using both types of relations (All and Hypo). This fact, possibly indicates the robustness of the approach.

Also as expected, the verbal part of WN behave differently. For verbs and using all relations, the total number of BLC ranges from 1,256 (no threshold) to 633 (threshold 50). Using hyponym relations, the total number of BLC ranges from 1,041 (no threshold) to 633 (threshold 50). In this case, since the verbal hierarchies are much shorter, the average depth of the synsets selected ranges from 3.32 (no threshold) to only 1.13 (threshold 50) using all relations, and from 3.31 (no threshold) to 1.10 (threshold 50) using hypo relations.

Table 3 presents the total number of BLC and its average depth for WN1.6 varying the threshold and the type of frequency (WN or SemCor).

In general, when using the frequency criteria, we can observe a similar behaviour than when using the relation criteria. That is, when increasing the threshold, the total number of automatic BLC and its average depth decrease. However, now the effect of the threshold is more dramatic, specially for nouns. For instance, the total number nominal BLC ranges from around 34,000 with no threshold to less than 100 nominal BLC with threshold equal to 50 descendants. Again, although the number of total BLC for nouns decreases dramatically, the average depth of the synsets selected only ranges from 7.44 (no threshold) to 4.35 (threshold 50) using sense frequencies from SemCor and from 7.44 (no threshold) to 4.41 (threshold 50) using sense frequencies from WN.

As expected, verbs behave differently than nouns. The number of BLC (for both SemCor and WN frequencies) reaches a plateau of around 600. In fact, this number is very close to the verbal top beginners.

Table 4 summarizes the Balkanet Base Concepts including the total number of synsets and their average depth.

PoS	#BC	Av. depth.
Noun	3,210	5.08
Verb	1,442	2.45

Table 4: Balkanet Base Concepts using WN2.0

In a similar way, table 5 presents the Meaning Base Concepts including the total number of synsets and their average depth.

PoS	#BC	Av. depth.
Noun	793	4.93
Verb	742	1.36

Table 5: Meaning Base Concepts using WN1.6

For nouns, the set of Balkanet BC is four times larger than the Meaning BC, while the average depth is similar in both sets (5.08 vs. 4.93 respectively). The verbal set of Balkanet BC is twice larger than the Meaning one, while contrary to the nominal subsets, their average depth is quite different (2.45 vs. 1.36). However, when comparing these sets of BC to the automatically selected BLC, it seems clear that for similar volumes, the automatic BLC appear to be deeper in the hierarchies (both for nouns and verbs).

In contrast, the BC derived from the Lexicographic Files of WN (or Supersenses), represent a much more coarse-grained set (26 categories for nouns and 15 for verbs).

5 *Sense–groupings as semantic classes*

In order to study to what extend the different sense–groupings could be of the interest for class–based WSD, we present a comparative evaluation of the different sense–groupings in a controlled framework. We tested the behaviour of the different sets of sense–groupings (WN senses, Balkanet BC, Meaning BC, automatic BLC and SuperSenses) using the English all–words task of SensEval–3. Obviously, different sense–groupings would provide different abstractions of the semantic content of WN, and we expect a different behaviour when disambiguating nouns and verbs. In fact, the most common baseline used to test the performance of a WSD system, is the Most Frequent Sense Classifier. In this study, we will use this simple but robust heuristic to compare the performances of the different sense–groupings. Thus, we will use SemCor[8] (Kuĉera and Francis, 1967) to train for Most Frequent Classifiers for each word and sense–grouping. We only used brown1 and brown2 parts of SemCor to train the classifiers. We used standard Precision, Recall and F1 measure (harmonic mean between Precision and Recall) to evaluate the performance of each classifier.

For WN senses, Meaning BC, the automatic BLC, and Lexicographic Files, we used WN1.6. For Balkanet BC we used the synset mappings provided by (Daudé, Padró, and Rigau, 2003)[9], translating the BC from WN2.0 to WN1.6. For testing the Most Frequent Classifiers we also used these mappings to translate the sense–groupings from WN1.6 to WN1.7.1.

Table 6 presents the polysemy degree for nouns and verbs of the different words when grouping its senses with respect the different semantic classes on SensEval–3. Senses stand for WN senses, BLC-A for automatic BLC derived using a threshold of 20 and all relations, BLC-S for automatic BLC derived using a threshold of 20 and frequencies from SemCor and SS for the SuperSenses. As expected, while increasing the abstraction level (from the sense level to the SuperSense level, passing to intermediate levels) the polysemy degree decreases. For instance in SensEval–3, at the sense level, the polysemy degree for nous is 4.93 (4.93 senses per word), while at the SuperSense level, the polysemy degree for nouns is 3.06 (3.06 classes per word). Notice that the reduction is dramatic for verbs (from 11.0 to only 4.08). Notice also, that when using the Base Level Concept representations a high degree of polysemy is maintained for nouns and verbs.

Tables 7 and 8 presents for polysemous words the performance in terms of F1 measure of the different sense-groupings using the relation criteria (All and Hypo) when training the class–frequencies on SemCor and testing on SensEval–3. That is, for each polysemous word in SensEval–3 the Most Frequent Class is obtained from SemCor. Best results are marked using bold.

	Senses	**BLC-A**	**BLC-S**	**SS**
Nouns	4.93	4.07	4.00	3.06
Verbs	11.00	8.64	8.72	4.08
N + V	7.66	6.13	6.13	3.52

Table 6: Polysemy degree over SensEval–3

Class	**Nouns**	**Verbs**
Senses	63.69	49.78
Balkanet	65.15	50.84
Meaning	65.28	53.11
BLC–0	66.36	54.30
BLC–10	66.31	54.45
BLC–20	**67.64**	54.60
BLC–30	67.03	54.60
BLC–40	66.61	55.54
BLC–50	67.19	**55.69**
SuperSenses	**73.05**	**76.41**

Table 7: F1 measure for polysemous words using all relations for BLC

In table 7, we present the results of using all relations for selecting BLC. As expected, SuperSenses obtain very high F1 results for nouns and verbs with 73.05 and 76.41, respectively. Comparing the BC from Balkanet and Meaning, the best results seems to be achieved by Meaning BC for both nouns and verbs. Notice that the set of BC from Balkanet was larger than the ones selected in Meaning, thus indicating that the BC from Meaning provide a better level of abstraction.

Interestingly, all sets of automatic BLC perform better than those BC provided by Balkanet or Meaning. For nouns, the best result is obtained for BLC using a threshold of only 20 with an F1 of 67.64. We should highlight this result since this set of BLC obtain better WSD performance than the rest of automatically derived BLC while maintaining more information of the original synsets. Interestingly, BLC-20 using 558 classes achieves an F1 of 67.64, while SuperSenses using a much smaller set (26 classes) achieves 73.05.

For verbs, it seems that the restriction on the minimum number of concepts for a Base Level Concept has a positive impact in the generalization selection.

These results suggest that intermediate levels of representation such as the automatically derived Base Concept Levels could be appropriate for learning class-based WSD classifiers. Recall that for nouns SuperSenses use only 26 classes, while BLC–20 uses 558 semantic classes (more than 20 times larger).

In table 8, we present the results of using hyponymy relations for selecting the BLC. Again,

[8] Annotated using WN1.6.

[9] http://www.lsi.upc.edu/˜nlp/

all sets of automatically derived BLC perform better than those BC provided by Balkanet or Meaning. In this case, the best results for nouns are obtained again for BLC using a threshold of 20 (F1 of 67.28 with 558 classes). We can also observe that in general, using hyponymy relations we obtain slightly lower performances than using all relations. Possibly, this fact indicates that a higher number of hyponymy relations is required for a Base Level Concept to compensate minor (but richer) number of relations.

Class	**Nouns**	**Verbs**
Senses	63.69	49.78
Balkanet	65.15	50.84
Meaning	65.28	53.11
BLC–0	65.76	54.30
BLC–10	65.86	54.45
BLC–20	**67.28**	54.60
BLC–30	66.72	54.60
BLC–40	66.77	**55.54**
BLC–50	67.19	**55.54**
SuperSenses	**73.05**	**76.41**

Table 8: F1 measure for polysemous words using hypomym relations for BLC

Tables 9 and 10 presents for polysemous words the performance in terms of F1 measure of the different sense-groupings using the frequency criteria (FreqSC and FreqWN) when training the class–frequencies on SemCor and testing on SensEval–3. That is, for each polysemous word in SensEval–3 the Most Frequent Class is obtained from SemCor. Best results are marked using bold.

In table 9, we present the results of using frequencies from SemCor for selecting the BLC. In this case, not all sets of automatic BLC surpass the BC from Balkanet and Meaning. For nouns, the best result for automatic BLC is obtained when using a threshold of 50 (F1 of 68.84 with 94 classes), while for verbs, the best result is obtained when using a threshold of 40. However, in this case, verbal BLC obtain slightly lower results than using the relations criteria (both all and hypo).

Class	**Nouns**	**Verbs**
Senses	63.69	49.78
Balkanet	65.15	50.84
Meaning	65.28	53.11
BLC–0	64.45	52.27
BLC–10	64.98	53.21
BLC–20	65.73	53.97
BLC–30	66.46	54.15
BLC–40	68.46	**54.63**
BLC–50	**68.84**	**54.63**
SuperSenses	**73.05**	**76.41**

Table 9: F1 measure for polysemous words using frequencies from SemCor for BLC

In table 10, we present the results of using frequencies from WN for selecting the BLC. Again, not all automatic sets of BLC surpass the BC from Balkanet and Meaning. For nouns, the best result for automatic BLC is obtained when using a threshold of 40 (F1 of 69.16 with 132 classes), while for verbs, the best result is obtained when using a threshold of 50. We can also observe that in general, using SemCor frequencies we obtain slightly lower performances than using WN frequencies. Again, verbal BLC obtain slightly lower results than using the relations criteria (both all and hypo).

Class	**Nouns**	**Verbs**
Senses	63.69	49.78
Balkanet	65.15	50.84
Meaning	65.28	53.11
BLC–0	64.95	51.75
BLC–10	65.59	53.29
BLC–20	66.30	53.44
BLC–30	66.67	53.61
BLC–40	**69.16**	54.22
BLC–50	69.11	**54.63**
SuperSenses	**73.05**	**76.41**

Table 10: F1 measure for polysemous words using frequencies from WN for BLC

These results for polysemous words reinforce our initial observations. That is, that the method for automatically deriving intermediate levels of representation such the Base Concept Levels seems to be robust enough for learning class-based WSD classifiers. In particular, it seems that BLC could achieve high levels of accuracy while maintaining adequate levels of abstraction (with hundreds of BLC). In particular, the automatic BLC obtained using the relations criteria (All or Hypo) surpass the BC from Balkanet and Meaning. For verbs, it seems that even the unique top beginners require an extra level of abstraction (that is, the SuperSense level) to be affective.

6 *Discussion*

We can put the current results in context, although indirectly, by comparison with the results of the English SensEval–3 all–words task systems. In this case, the best system presented an accuracy of 65.1%, while the "WN first sense" baseline would achieve 62.4%[10]. Furthermore, it is also worth mentioning that in this edition there were a few systems above the "WN first sense" baseline (4 out of 26 systems). Usually, this baseline is very competitive in WSD tasks, and it is extremely hard to improve upon even slightly.

Tables 11 and 12 presents for monosemous and polysemous nouns and verbs the F1 measures of the different sense-groupings obtained

[10] This result could be different depending on the treatment of multiwords and hyphenated words.

with all relations criteria when training the class–frequencies on SemCor and testing on SensEval–3. Best results are marked using bold. Table 11 presents the results using all relations criteria and table 12 presents the same results but using the WN frequency criteria.

Class	Nouns	Verbs	Nouns+Verbs
Senses	71.79	52.89	63.24
Balkanet	73.06	53.82	64.37
Meaning	73.40	56.40	65.71
BLC–0	74.80	58.32	67.35
BLC–10	74.99	58.46	67.52
BLC–20	76.12	58.60	68.20
BLC–30	75.99	58.60	68.14
BLC–40	75.76	59.70	68.51
BLC–50	**76.22**	**59.83**	**68.82**
SuperSenses	**81.87**	**79.23**	**80.68**

Table 11: F1 measure for nouns and verbs using all relations for BLC

Obviously, higher accuracy figures are obtained when incorporating also monosemous words. Note this naive system achieves for Senses an F1 of 63.24, very similar to those reported in SensEval–3, and for SuperSenses a very high a F1 of 80.68. Regarding the automatic BLC, the best results are obtained for BLC–50, but all of them outperform the BC from Balkanet and Meaning. However, for nouns, BLC–20 (with 558 classes) obtain only slightly lower F1 figures than BLC–50 (with 253 classes).

Class	Nouns	Verbs	Nouns+Verbs
Senses	71.79	52.89	63.24
Balkanet	73.06	53.82	64.37
Meaning	73.40	56.40	65.71
BLC–0	72.99	55.33	65.01
BLC–10	74.60	57.08	66.69
BLC–20	75.62	57.22	67.31
BLC–30	76.10	57.63	67.76
BLC–40	**78.03**	58.18	69.07
BLC–50	**78.03**	**58.87**	**69.38**
SuperSenses	**81.87**	**79.23**	**80.68**

Table 12: F1 measure for nouns and verbs using WN frequencies for BLC

When using frequencies instead of relations, BLC even achieve higher results. Again, the best results are obtained for BLC–50. However, in this case, not all of them outperform the BC from Balkanet and Meaning.

Surprisingly, these naive Most frequent WSD systems trained on SemCor are able to achieve very high levels of accuracy. For nouns, using BLC-20 (selected from all relations, 558 semantic labels) the system reaches 75-62, while using BLC-40 (selected from WN frequencies, 132 semantic labels) the system achieves 78.03. Finally, using SuperSenses for verbs (15 semantic labels) this naive system scores 79.23.

To our knowledge, the best results for class–based WSD are those reported by (Ciaramita and Altun, 2006). This system performs a sequence tagging using a perceptron–trained HMM, using SuperSenses, training on SemCor and testing on the SensEval–3. The system achieves an F1–score of 70.74, obtaining a significant improvement from a baseline system which scores only 64.09. In this case, the first sense baseline is the SuperSense of the most frequent synset for a word, according to the WN sense ranking.

Possibly, the origin of the discrepancies between our results and those reported by (Ciaramita and Altun, 2006) is twofold. First, because they use a BIO sequence schema for annotation, and second, the use of the brown-v part of SemCor to establish sense–frequencies.

In order to measure the real contribution of the automatic BLC on the WSD task, we also performed a final set of experiments. Once trained on SemCor the Most Frequent Class of a word, we tested on SensEval–3 the first sense appearing in WN of the word for that Class. In that way, we developed a very simple sense tagger which uses the frequency counts of more coarse-grained sense–groupings. Table 13 presents the F1 measures for all nouns and verbs of this naive class–based sense tagger when using WN frequencies for building the automatic BLC. Note that these results are different from the rest since are evaluated at a sense level.

Class	Nouns	Verbs	Nouns+Verbs
Senses	71.79	52.89	63.24
Balkanet	72.35	52.48	63.36
Meaning	72.01	53.17	63.49
BLC–0	72.35	52.89	63.55
BLC–10	72.24	53.03	63.55
BLC–20	72.47	53.03	63.68
BLC–30	**72.70**	53.03	63.80
BLC–40	**72.70**	**53.31**	**63.93**
BLC–50	72.58	**53.31**	63.86
SuperSenses	72.47	53.03	63.68

Table 13: F1 measure for nouns and verbs of the class–based sense tagger.

Surprisingly, all these oportunistic class–based sense taggers surpass the Most Frequent Sense tagger. Interestingly, the results of all automatic BLC using threshold higher than 10 obtain equal or better performance than SuperSenses. In fact, the best results for nouns are those obtained using BLC–30 while for verbs those obtained by BLC–40. That is, the sense-groupings seem to stablish more robust sense frequencies.

7 Conclusions and further work

The WSD task seems to have reached its maximum accuracy figures with the usual framework. Some of its limitations could come from the sense–granularity of WordNet (WN). WN has been often criticised because its fine–grained sense distinctions. Nevertheless, other problems arise for supervised systems like data sparseness just because the lack of adequate and enough training examples. Moreover, it is not clear how WSD can contribute with the current result to improve other NLP tasks.

Changing the set of classes could be a solution to enrich training corpora with many more examples. In this manner, the classifiers generalize among an heterogeneous set of labeled examples. At the same time these classes are more easily learned because there are more clear semantic distinctions between them. In fact, our most frequent naive systems are able to perform a semantic tagging with accuracy figures over 75%.

Base Level Concepts (BLC) are concepts that are representative for a set of other concepts. In the present work, a simple method for automatically selecting BLC from WN based on the hypernym hierarchy and the number of stored frequencies or relationships between synsets have been shown. Although, some sets of Base Concepts are available at this moment (e.g. EuroWordNet, Balkanet, Meaning), a huge manual effort should be invested for its development. Other sets of Base Concepts, like WN Lexicographer Files (or SuperSenses) are clearly insufficient in order to describe and distinguish between the enormous number of concepts that are used in a text. Using a very simple baseline, the Most Frequent Class, our approach empirically shows a clear improvement over such other sets. In addition, our method is capable to get a more or less detailed sets of BLC without losing semantic discrimination power. Obviously, other selection criteria for selecting BLC should be investigated.

We are also interested in the direct comparison between automatically and manually selected BLC. An in depth study of their correlations deserves more attention.

Once having defined an appropriate level of abstraction using the new sets of BLC, we plan to use them for supervised class–based WSD. We suspect that using this approach higher accuracy figures for WSD could be expected.

References

Agirre, E., I. Aldezabal, y E. Pociello. 2003. A pilot study of english selectional preferences and their cross-lingual compatibility with basque. En *Proceedings of the International Conference on Text Speech and Dialogue (TSD'2003)*, CeskBudojovice, Czech Republic.

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, y P. Vossen. 2004. The meaning multilingual central repository. En *Proceedings of Global WordNet Conference (GWC'04)*, Brno, Czech Republic.

Ciaramita, M. y Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, páginas 594–602, Sydney, Australia. ACL.

Ciaramita, M. y M. Johnson. 2003. Supersense tagging of unknown nouns in wordnet. En *Proceedings of the Conference on Empirical methods in natural language processing (EMNLP'03)*, páginas 168–175. ACL.

Curran, J. 2005. Supersense tagging of unknown nouns using semantic similarity. En *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*, páginas 26–33. ACL.

Daudé, J., Ll. Padró, y G. Rigau. 2003. Validation and tuning of wordnet mapping techniques. En *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'03)*, Borovets, Bulgaria.

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

Hearst, M. y H. Schütze. 1993. Customizing a lexicon to better suit a computational task. En *Proceedingns of the ACL SIGLEX Workshop on Lexical Acquisition*, Stuttgart, Germany.

Kuĉera, H. y W. N. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, USA.

Magnini, B. y G. Cavaglia. 2000. Integrating subject fields codes into wordnet. En *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*.

Màrquez, Ll., G. Escudero, D. Martínez, y G. Rigau. 2006. Supervised corpus-based methods for wsd. En *E. Agirre and P. Edmonds (Eds.) Word Sense Disambiguation: Algorithms and applications.*, volumen 33 de *Text, Speech and Language Technology*. Springer.

Mihalcea, R. y D. Moldovan. 2001. Automatic generation of coarse grained wordnet. En *Proceding of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburg, USA.

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

Peters, W., I. Peters, y P. Vossen. 1998. Automatic sense clustering in eurowordnet. En *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Rosch, E. 1977. Human categorisation. *Studies in Cross-Cultural Psychology*, I(1):1–49.

Segond, F., A. Schiller, G. Greffenstette, y J. Chanod. 1997. An experiment in semantic tagging using hidden markov model tagging. En *ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. ACL, New Brunswick, New Jersey, páginas 78–81.

Snyder, Benjamin y Martha Palmer. 2004. The english all-words task. En Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, páginas 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

Villarejo, L., L. Màrquez, y G. Rigau. 2005. Exploring the construction of semantic class classifiers for wsd. En *Proceedings of the 21th Annual Meeting of Sociedad Espaola para el Procesamiento del Lenguaje Natural SEPLN'05*, páginas 195–202, Granada, Spain, September. ISSN 1136-5948.

Vossen, P., L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, y W. Peters. 1998. The eurowordnet base concepts and top ontology. Informe técnico, Paris, France, France.

Vossen, P., G. Rigau, I. Alegria, E. Agirre, D. Farwell, y M. Fuentes. 2006. Meaningful results for information retrieval in the meaning project. En *Proceedings of the 3rd Global Wordnet Conference*, Jeju Island, Korea, South Jeju, January.