Exploring the Automatic Selection of Basic Level Concepts

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

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

Abstract

We present a very simple method for selecting Base Level Concepts using 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 [11]. 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) [7].

WordNet 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 WordNet as a sense repository, the organizers of the English all-words task at SensEval-3 reported an inter-annotation agreement of 72.5% [17]. Interestingly, this result is difficult to outperform by current 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 [8] [14] [12] [1] and on using predefined sets of sense-groupings for learning class-based classifiers for WSD [16] [4] [18] [5] [3]. 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 sense–groupings such as WordNet Domains [10], SUMO labels [13], EuroWordNet Base Concepts [19] or Top Concept Ontology labels [2]. 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 [15] 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 (WN)[2] [7] 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

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

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

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 (EWN)[3] [19] 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 WordNet. By means of the ILI, synsets and words or different languages are connected, allowing advanced multilingual natural language applications [20].

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;
- Having many relations to other concepts;

Thus, the BC are the fundamental building blocks for establishing the relations in a wordnet and give information about the dominant lexicalization patterns in languages. Thus, the Lexicografic Files (or Supersenses) of WN could be considered the most basic set of BC.

Basic Level Concepts [15] (hereinafter BLC) should not be confused with Base Concepts. BLC are the result of a compromise between two conflicting principles of characterization:

- Represent as many concepts as possible;
- Represent as many features as possible;

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

WordNet 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.

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

2.2 EuroWordNet Base Concepts

Within EuroWordNet, a set of Base Concepts was selected to reach maximum overlap an compatibility across wordnets in different languages following the two main criteria described above.

- A high position in the semantic hierarchy
- Having many relations to other concepts

Initially, a set of 1,024 Common Base Concepts from WN1.5 (concepts that acts as BC in at least two languages) was selected, only considering English, Dutch, Spanish and Italian wordnets.

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 model proposed by the EWN project to build the Multilingual Central Repository (MCR) [2]. 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:

- All: considers all types of relations encoded in WN
- Hyp: considers only the hyponymy relations encoded in WN

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

[4] http://www.ceid.upatras.gr/Balkanet
[5] http://www.globalwordnet.org/gwa/5000_bc.zip
[6] http://www.lsi.upc.es/~nlp/meaning

#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*

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.

Instead of the highly related concepts, we also considered highly frequent concepts as possible indicators for of large set of features. Following the same basic algorithm, we also considered the relative frequency of the synsets in the hypernymy chain. That is, we derived two other different sets of BLC depending on the source of relative frequencies considered:

- FreqSemCor: considers the frequency counts of SemCor
- FreqWN: considers the frequency counts appearing in WordNet

The frequency of a synset has been obtained summing up the frequencies of its word senses. In fact, WordNet word-senses were ranked using SemCor and other sense-annotated corpora. Thus, the frequencies appearing in SemCor and in WordNet 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 hyponym), sense frequencies 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).

Threshold	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*

As expected, when increasing the threshold, the total number of automatic BLC 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

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

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 considered (WN or SemCor).

Threshold	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*

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 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) sense frequencies from SemCor and from 7.44 (no threshold) to 4.41 (threshold 50) using sense frequencies from WN.

As expected, the verbal part of WN behave differently than nouns. In fact, 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.

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).

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

Table 5: MEANING *Base Concepts using WN1.6*

In contrast, the BC derived from the Lexicographic Files of WN, represent a more coarse-grained set (26 categories for nouns and 15 for verbs). The synsets corresponding to these categories are also called the unique beginners of WN, being at the top of the hierarchies.

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 Base Concepts, MEANING Base Concepts, Automatic Base Level Concepts and Lexicographic Files) using the English All–words task of SensEval–3[8]. 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 effectivity when testing 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 SemCor1.6[9] [9] 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 Most Frequent Classifier.

For WN senses, MEANING BC, the automatic BLC, and Lexicographic Files of WN, we used WN1.6. For BALKANET BC we used the synset mappings provided by [6][10], 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 the WN senses, BLC-A for the automatic BLC derived using a threshold of 20 and all relations, BLC-S for the automatic BLC derived using a threshold of 20 and frequencies from Sem-

[8] http://www.senseval.org
[9] Annotated using WN1.6
[10] http://www.lsi.upc.edu/ nlp/

Cor and SS for the SuperSenses or the Lexicographic Files of WN. As expected, while increasing the abstraction level (from the sense level to the SuperSense level, passing to an intermediate level of representation) the polysemy degree decreases. For instance in SensEval-3, at the sense level, the polysemy degree for nous is 4.92 (4.92 senses per word), while at the SuperSense level, the polysemy degree for nouns is 3.01 (3.01 classes per word). Notice that the reduction is dramatic for verbs (from 11.0 to only 1.03). Notice also, that when using the Base Level Concept representations a high degree of polysemy is maintained for nouns and verbs.

	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*

Tables 7 and 8 presents for polysemous words only the performance in terms of F1 measure of the different sense-groupings using the relations 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 7, we present the results of using all relations for selecting the Base Level Concepts.

Comparing both sets of BC, 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, for nouns, the best result is obtained for BLC using a threshold of 20 with an F1 of 67.44. In fact, all types of BLC surpass the rest of class–groupings (except verbal Supersenses).

We can observe that better results are obtained when using positive thresholds. It seems that, the restriction over the minimum number of concepts for a Base Level Concept has a positive impact in the generalization selection.

It also seems that Base Concept Levels are more appropriate abstractions for representing the semantic classes of the different word senses. 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). We should highlight this result, nominal BLC obtain better WSD performance while maintaining more information of the original synsets.

In table 8, we present the results of using hyponymy relations for selecting the BLC. In this case, the best results for nouns are obtained for the Base Level Concept using a threshold of 50. We can also observe that in general, using hyponymy relations we obtain slightly lower performances. 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. Also in this case, all types of BLC surpass the rest of class–groupings (except verbal Supersenses).

Class	Nouns	Verbs
Senses	63.69	47.36
Balkanet	65.15	52.71
Meaning	65.28	53.05
BLC–0	66.36	54.30
BLC–10	66.31	54.45
BLC–20	**67.64**	54.60
BLC–30	67.03	55.60
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*

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	**56.54**
BLC–50	67.19	**56.54**
SuperSenses	**73.05**	**76.41**

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

Tables 9 and 11 presents for polysemous words only the performance in terms of F1 measure of the different sense-groupings using the frequency criteria 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. Interestingly, the best results for nouns are obtained for the Base Level Concept using a threshold of 50 (with only 94 BLC). In this case, not all nominal BLC surpass the rest of class–groupings.

In this case, verbal BLC obtain slightly lower results than using the relations criteria (both all and hypo).

Class	Nouns	Verbs
Senses	63.69	47.36
Balkanet	65.15	52.71
Meaning	65.28	53.05
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 11, we present the results of using frequencies from WN for selecting the BLC. In this case, the best results for nouns are obtained using a threshold of

40 (with only 132 BLC). Again, not all nominal BLC surpass the rest of class–groupings. However, note that BLC-40 using 132 classes achieves an accuracy of 68.95%, while SuperSenses using a much smaller set (26 classes) only achieves 66.00%.

Class	Nouns	Verbs
Senses	63.69	47.36
Balkanet	65.15	52.71
Meaning	65.28	53.05
BLC–0	64.95	51.75
BLC–10	65.59	53.29
BLC–20	66.30	53.44
BLC–30	66.77	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 only 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 for nouns the criteria of using all relations achieves a high level of accuracy while maintaining an adequate level of abstraction (with hundreds of BLC). 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

In the last SensEval-3 edition, the results of top systems (all of them using supervised techniques) presented very small differences in performance for the English lexical–sample task. This suggests that a plateau has been reached for this design of task with this kind of techniques. The results of the best system (72.9% accuracy) are way ahead of the Most–Frequent–Sense baseline (55.2% accuracy). These results present a significant improvement from the previous Senseval edition, which could be due, in part, to the change in the verb sense inventory (Wordsmyth instead of WordNet).

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 "WordNet first sense" baseline would achieve 62.4%[11]. Furthermore, it is also worth mentioning that in this edition there were a few systems above the "WordNet first sense" baseline (4 out of 26 systems).

To our knowledge, the best results for class–based WSD are those reported by [3]. 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 improvemement 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. Usually, this baseline is very competitive in WSD tasks, and it is extremely hard to improve upon even slightly.

Table 11 presents for polysemous words only the performance in terms of F1 measure of the different sense-groupings when using the WN first sense heuristic and testing on SensEval–3. That is, for each polysemous word in SensEval–3 the Most Frequent Class is obtained according to the WN sense ranking. The results of BLC are those of using all relations. Best results are marked using bold.

In this case, the best results are obtained for SuperSenses both for nouns and verbs. Considering more fine grained sense–groupings, again, the best results for nouns are those obtained by BLC using a threshold of 20 with an F1–score of 64.80. For verbs, BLC with a threshold of 40 achieves an F1–score of 55.51.

Class	Nouns	Verbs
Senses	62.14	49.78
Balkanet	62.90	51.24
Meaning	63.45	52.65
BLC–0	63.08	53.82
BLC–10	63.35	54.12
BLC–20	**65.10**	54.27
BLC–30	64.77	54.27
BLC–40	64.66	**54.60**
BLC–50	64.91	**54.60**
SuperSenses	**68.60**	**75.78**

Table 11: *F1 measure for polysemous words using all relations for BLC*

Possibly, the origin of the discrepancies between our results and those reported by [3] 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.

Tables 12 and 13 presents for monosemous and polysemous nouns and verbs the F1 measures of the different sense-groupings when training the class–frequencies on SemCor and testing on SensEval–3. That is, for each word in SensEval–3 the Most Frequent Class is obtained from SemCor. Best results for BLC are marked using bold. Table 12 presents the results using all relations criteria and table 13 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 12: *F1 measure for nouns and verbs using all relations for BLC*

Obviously, higher accuracy figures are obtained when incorporating also monosemous words. For

[11] Depending on the treatment of multiwords and hyphenated words

nouns, BLC selected using all relations and a threshold of 20 achieve 75.22, while for verbs the best results are obtained using a threshold of 40 reaching 61.07.

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 13: *F1 measure for nouns and verbs using WN frequencies for BLC*

When using frequencies instead of relations, BLC even achieve higher results while reducing expressive power. For nouns, BLC selected using WN frequencies and a threshold of 40 achieve 77.24, while for verbs the best results are obtained using a threshold of 50 reaching 58.69.

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 75.22%, while using BLC-40 (selected from WN frequencies, 132 semantic labels) the system achieves 77.24%. Finally, using SuperSenses for verbs (15 semantic labels) this naive system scores 75.51%.

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 like data sparseness just because the lack of adequate and enough 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 comparison between the current sets of automatically selected BLC and previous sets manually selected. 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 class–based WSD. We suspect that using this approach higher accuracy figures for WSD could be expected.

References

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

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

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

[4] M. Ciaramita and M. Johnson. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the Conference on Empirical methods in natural language processing (EMNLP'03)*, pages 168–175. ACL, 2003.

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

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

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

[8] M. Hearst and H. Schütze. Customizing a lexicon to better suit a computational task. In *Proceedingns of the ACL SIGLEX Workshop on Lexical Acquisition*, Stuttgart, Germany, 1993.

[9] H. Kučera and W. N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, USA, 1967.

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

[11] L. Màrquez, G. Escudero, D. Martínez, and G. Rigau. Supervised corpus-based methods for wsd. In *E. Agirre and P. Edmonds (Eds.) Word Sense Disambiguation: Algorithms and applications.*, volume 33 of *Text, Speech and Language Technology*. Springer, 2006.

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

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

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

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

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

[17] B. Snyder and M. Palmer. The english all-words task. In R. Mihalcea and P. Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July 2004. Association for Computational Linguistics.

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

[19] P. Vossen, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, and W. Peters. The eurowordnet base concepts and top ontology. Technical report, Paris, France, France, 1998.

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