An Empirical Study on Class-based Word Sense Disambiguation*

Rubén Izquierdo & Armando Suárez
Deparment of Software and Computing Systems
University of Alicante. Spain
{ruben,armando}@dlsi.ua.es

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

Abstract

As empirically demonstrated by the last SensEval exercises, assigning the appropriate meaning to words in context has resisted all attempts to be successfully addressed. One possible reason could be the use of inappropriate set of meanings. In fact, WordNet has been used as a de-facto standard repository of meanings. However, to our knowledge, the meanings represented by WordNet have been only used for WSD at a very fine-grained sense level or at a very coarse-grained class level. We suspect that selecting the appropriate level of abstraction could be on between both levels. We use a very simple method for deriving a small set of appropriate meanings using basic structural properties of WordNet. We also empirically demonstrate that this automatically derived set of meanings groups senses into an adequate level of abstraction in order to perform class-based Word Sense Disambiguation, allowing accuracy figures over 80%.

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[2] (WN) (Fellbaum, 1998).

WN has been widely criticized for being a sense repository that often provides 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 inferring 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 sense-based WSD systems.

Thus, some research has been focused on deriving different word-sense groupings to overcome the fine–grained distinctions of WN (Hearst and Schütze, 1993), (Peters et al., 1998), (Mihalcea and Moldovan, 2001), (Agirre and LopezDeLaCalle, 2003), (Navigli, 2006) and (Snow et al., 2007). That is, they provide methods for grouping senses of the same **word**, thus producing coarser word sense groupings for better disambiguation.

Wikipedia[3] has been also recently used to overcome some problems of automatic learning methods: excessively fine–grained definition of meanings, lack of annotated data and strong domain dependence of existing annotated corpora. In this way, Wikipedia provides a new very large source of annotated data, constantly expanded (Mihalcea, 2007).

*This paper has been supported by the European Union under the projects QALL-ME (FP6 IST-033860) and KYOTO (FP7 ICT-211423), and the Spanish Government under the project Text-Mess (TIN2006-15265-C06-01) and KNOW (TIN2006-15049-C03-01)

[1] http://www.senseval.org
[2] http://wordnet.princeton.edu
[3] http://www.wikipedia.org

In contrast, some research have been focused on using predefined sets of sense-groupings for learning class-based classifiers for WSD (Segond et al., 1997), (Ciaramita and Johnson, 2003), (Villarejo et al., 2005), (Curran, 2005) and (Ciaramita and Altun, 2006). That is, grouping senses of different words into the same explicit and comprehensive semantic class.

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 (Magnini and Cavaglià, 2000), SUMO labels (Niles and Pease, 2001), EuroWordNet Base Concepts (Vossen et al., 1998), Top Concept Ontology labels (Alvez et al., 2008) or Basic Level Concepts (Izquierdo et al., 2007). 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 has been performed on SensEval data exploring different levels of abstraction. In fact, (Villarejo et al., 2005) studied the performance of class–based WSD comparing only SuperSenses and SUMO by 10–fold cross–validation on SemCor, but they did not provide results for SensEval2 nor SensEval3.

This paper empirically explores on the supervised WSD task the performance of different levels of abstraction provided by WordNet Domains (Magnini and Cavaglià, 2000), SUMO labels (Niles and Pease, 2001) and Basic Level Concepts (Izquierdo et al., 2007). We refer to this approach as class–based WSD since the classifiers are created at a class level instead of at a sense level. Class-based WSD clusters senses of different words into the same explicit and comprehensive grouping. Only those cases belonging to the same semantic class are grouped to train the classifier. For example, the coarser word grouping obtained in (Snow et al., 2007) only has one remaining sense for "church". Using a set of Base Level Concepts (Izquierdo et al., 2007), the three senses of "church" are still represented by *faith.n#3*, *building.n#1* and *religious_ceremony.n#1*.

The contribution of this work is threefold. We empirically demonstrate that a) Basic Level Concepts group senses into an adequate level of abstraction in order to perform supervised class–based WSD, b) that these semantic classes can be successfully used as semantic features to boost the performance of these classifiers and c) that the class-based approach to WSD reduces dramatically the required amount of training examples to obtain competitive classifiers.

After this introduction, section 2 presents the sense-groupings used in this study. In section 3 the approach followed to build the class–based system is explained. Experiments and results are shown in section 4. Finally some conclusions are drawn in section 5.

2 Semantic Classes

WordNet (Fellbaum, 1998) synsets are organized in forty five Lexicographer Files, more recetly called **SuperSenses**, based on open 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.

WordNet Domains[4] (Magnini and Cavaglià, 2000) is a hierarchy of 165 Domain Labels which have been used to label all WN synsets. Information brought by Domain Labels is complementary to what is already in WN. First of all a Domain Labels may include synsets of different syntactic categories: for instance MEDICINE groups together senses from nouns, such as doctor and hospital, and from Verbs such as to operate. Second, a Domain Label may also contain senses from different WordNet subhierarchies. For example, SPORT contains senses such as athlete, deriving from life form, game equipment, from physical object, sport from act, and playing field, from location.

SUMO[5] (Niles and Pease, 2001) was created as part of the IEEE Standard Upper Ontology Working Group. The goal of this Working Group is to develop a standard upper ontology to promote data interoperability, information search and retrieval, automated inference, and natural language processing. SUMO consists of a set of concepts, relations, and axioms that formalize an upper ontology. For these experiments, we used the complete WN1.6 mapping with 1,019 SUMO labels.

[4] http://wndomains.itc.it/

[5] http://www.ontologyportal.org/

Basic Level Concepts[6] (BLC) (Izquierdo et al., 2007) are small sets of meanings representing the whole nominal and verbal part of WN. BLC can be obtained by a very simple method that uses basic structural WN properties. In fact, the algorithm only considers the relative number of relations of each synset along the hypernymy chain. The process follows a bottom-up approach using the chain of hypernymy relations. For each synset in WN, the process selects as its BLC the first local maximum according to the relative number of relations. The local maximum is the synset in the hypernymy chain having more relations than its immediate hyponym and immediate hypernym. 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 preliminary BLC. Obviously, while ascending through this chain, more synsets are subsumed by each concept. 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 the threshold, the process selects the next local maximum following the hypernymy 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.

In this paper, we empirically explore the performance of the different levels of abstraction provided by Basic Level Concepts (BLC) (Izquierdo et al., 2007).

Table 1 presents the total number of BLC and its average depth for WN1.6, varying the threshold and the type of relations considered (all relations or only hyponymy).

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
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 1: BLC for WN1.6 using all or hyponym relations

[6] http://adimen.si.ehu.es/web/BLC

Classifier	Examples	# of examples
church.n#2 *(sense approach)*	church.n#2	58
building, edifice *(class approach)*	church.n#2	58
	building.n#1	48
	hotel.n#1	39
	hospital.n#1	20
	barn.n#1	17
	.....	.....
		TOTAL= 371 examples

Table 2: Examples and number of them in Semcor, for sense approach and for class approach

3 Class-based WSD

We followed a supervised machine learning approach to develop a set of class-based WSD taggers. Our systems use an implementation of a Support Vector Machine algorithm to train the classifiers (one per class) on semantic annotated corpora for acquiring positive and negative examples of each class and on the definition of a set of features for representing these examples. The system decides and selects among the possible semantic classes defined for a word. In the sense approach, one classifier is generated for each word sense, and the classifiers choose between the possible senses for the word. The examples to train a single classifier for a concrete word are all the examples of this word sense. In the semantic–class approach, one classifier is generated for each semantic class. So, when we want to label a word, our program obtains the set of possible semantic classes for this word, and then launch each of the semantic classifiers related with these semantic categories. The most likely category is selected for the word. In this approach, contrary to the word sense approach, to train a classifier we can use all examples of all words belonging to the class represented by the classifier. In table 2 an example for a sense of "church" is shown. We think that this approach has several advantages. First, semantic classes reduce the average polysemy degree of words (some word senses are grouped together within the same class). Moreover, the well known problem of acquisition bottleneck in supervised machine learning algorithms is attenuated, because the number of examples for each classifier is increased.

3.1 The learning algorithm: SVM

Support Vector Machines (SVM) have been proven to be robust and very competitive in many NLP tasks, and in WSD in particular (Màrquez et al., 2006). For these experiments, we used SVM-Light (Joachims, 1998). SVM are used to learn an hyperplane that separates the positive from the

negative examples with the maximum margin. It means that the hyperplane is located in an intermediate position between positive and negative examples, trying to keep the maximum distance to the closest positive example, and to the closest negative example. In some cases, it is not possible to get a hyperplane that divides the space linearly, or it is better to allow some errors to obtain a more efficient hyperplane. This is known as "soft-margin SVM", and requires the estimation of a parameter (C), that represent the trade-off allowed between training errors and the margin. We have set this value to 0.01, which has been proved as a good value for SVM in WSD tasks.

When classifying an example, we obtain the value of the output function for each SVM classifier corresponding to each semantic class for the word example. Our system simply selects the class with the greater value.

3.2 Corpora

Three semantic annotated corpora have been used for training and testing. SemCor has been used for training while the corpora from the English all-words tasks of SensEval-2 and SensEval-3 has been used for testing. We also considered SemEval-2007 coarse–grained task corpus for testing, but this dataset was discarded because this corpus is also annotated with clusters of word senses.

SemCor (Miller et al., 1993) is a subset of the Brown Corpus plus the novel *The Red Badge of Courage*, and it has been developed by the same group that created WordNet. It contains 253 texts and around 700,000 running words, and more than 200,000 are also lemmatized and sense-tagged according to Princeton WordNet 1.6.

SensEval-2[7] English all-words corpus (hereinafter SE2) (Palmer et al., 2001) consists on 5,000 words of text from three WSJ articles representing different domains from the Penn TreeBank II. The sense inventory used for tagging is WordNet 1.7. Finally, **SensEval-3**[8] English all-words corpus (hereinafter SE3) (Snyder and Palmer, 2004), is made up of 5,000 words, extracted from two WSJ articles and one excerpt from the Brown Corpus. Sense repository of WordNet 1.7.1 was used to tag 2,041 words with their proper senses.

[7] `http://www.sle.sharp.co.uk/senseval2`
[8] `http://www.senseval.org/senseval3`

3.3 Feature types

We have defined a set of features to represent the examples according to previous works in WSD and the nature of class-based WSD. Features widely used in the literature as in (Yarowsky, 1994) have been selected. These features are pieces of information that occur in the context of the target word, and can be organized as:

Local features: bigrams and trigrams that contain the target word, including part-of-speech (PoS), lemmas or word-forms.

Topical features: word–forms or lemmas appearing in windows around the target word.

In particular, our systems use the following **basic features**:

Word–forms and lemmas in a window of 10 words around the target word

PoS: the concatenation of the preceding/following three/five PoS

Bigrams and trigrams formed by lemmas and word-forms and obtained in a window of 5 words. We use of all tokens regardless their PoS to build bi/trigrams. The target word is replaced by *X* in these features to increase the generalization of them for the semantic classifiers

Moreover, we also defined a set of **Semantic Features** to explode different semantic resources in order to enrich the set of basic features:

Most frequent semantic class calculated over SemCor, the most frequent semantic class for the target word.

Monosemous semantic classes semantic classes of the monosemous words arround the target word in a window of size 5. Several types of semantic classes have been considered to create these features. In particular, two different sets of BLC (BLC20 and BLC50[9]), SuperSenses, WordNet Domains (WND) and SUMO.

In order to increase the generalization capabilities of the classifiers we filter out irrelevant features. We measure the relevance of a feature[10]. f for a class c in terms of the frequency of f. For each class c, and for each feature f of that class, we calculate the frequency of the feature within the class (the number of times that it occurs in examples

[9] We have selected these set since they represent different levels of abstraction. Remember that 20 and 50 refer to the threshold of minimum number of synsets that a possible BLC must subsume to be considered as a proper BLC. These BLC sets were built using all kind of relations.

[10] That is, the value of the feature, for example a *feature type* can be **word-form**, and a *feature* of that type can be "houses"

of the class), and also obtain the total frequency of the feature, for all the classes. We divide both values (classFreq / totalFreq) and if the result is not greater than a certain threshold *t*, the feature is removed from the feature list of the class *c*[11]. In this way, we ensure that the features selected for a class are more frequently related with that class than with others. We set this threshold *t* to 0.25, obtained empirically with very preliminary versions of the classifiers on SensEval3 test.

4 Experiments and Results

To analyze the influence of each feature type in the class-based WSD, we designed a large set of experiments. An experiment is defined by two sets of semantic classes. First, the semantic class type for selecting the examples used to build the classifiers (determining the abstraction level of the system). In this case, we tested: sense[12], BLC20, BLC50, WordNet Domains (WND), SUMO and SuperSense (SS). Second, the semantic class type used for building the semantic features. In this case, we tested: BLC20, BLC50, SuperSense, WND and SUMO. Combining them, we generated the set of experiments described later.

Test	pos	Sense	BLC20	BLC50	WND	SUMO	SS
SE2	N	4.02	3.45	3.34	2.66	3.33	2.73
	V	9.82	7.11	6.94	2.69	5.94	4.06
SE3	N	4.93	4.08	3.92	3.05	3.94	3.06
	V	10.95	8.64	8.46	2.49	7.60	4.08

Table 3: Average polysemy on SE2 and SE3

Table 3 presents the average polysemy on SE2 and SE3 of the different semantic classes.

4.1 Baselines

The most frequent classes (MFC) of each word calculated over SemCor are considered to be the baselines of our systems. Ties between classes on a specific word are solved obtaining the global frequency in SemCor of each of these tied classes, and selecting the more frequent class over the whole training corpus. When there are no occurrences of a word of the test corpus in SemCor (we are not able to calculate the most frequent class of the word), we obtain again the global frequency for each of its possible semantic classes (obtained from WN) over SemCor, and we select the most frequent.

4.2 Results

Tables 4 and 5 present the F1 measures (harmonic mean of recall and precision) for nouns and verbs respectively when training our systems on SemCor and testing on SE2 and SE3. Those results showing a statistically significant[13] positive difference when compared with the baseline are in marked bold. Column labeled as "Class" refers to the target set of semantic classes for the classifiers, that is, the desired semantic level for each example. Column labeled as "Sem. Feat." indicates the class of the semantic features used to train the classifiers. For example, class BLC20 combined with Semantic Feature BLC20 means that this set of classes were used both to label the test examples and to define the semantic features. In order to compare their contribution we also performed a "basicFeat" test without including semantic features.

As expected according to most literature in WSD, the performances of the MFC baselines are very high. In particular, those corresponding to nouns (ranging from 70% to 80%). While nominal baselines seem to perform similarly in both SE2 and SE3, verbal baselines appear to be consistently much lower for SE2 than for SE3. In SE2, verbal baselines range from 44% to 68% while in SE3 verbal baselines range from 52% to 79%. An exception is the results for verbs considering WND: the results are very high due to the low polysemy for verbs according to WND. As expected, when increasing the level of abstraction (from senses to SuperSenses) the results also increase. Finally, it also seems that SE2 task is more difficult than SE3 since the MFC baselines are lower.

As expected, the results of the systems increase while augmenting the level of abstraction (from senses to SuperSenses), and almost in every case, the baseline results are reached or outperformed. This is very relevant since the baseline results are very high.

Regarding nouns, a very different behaviour is observed for SE2 and SE3. While for SE3 none of the system presents a significant improvement over the baselines, for SE2 a significant improvement is obtained by using several types of seman-

[11] Depending on the experiment, around 30% of the original features are removed by this filter.

[12] We included this evaluation for comparison purposes since the current system have been designed for class-based evaluation only.

[13] Using the McNemar's test.

tic features. In particular, when using WordNet Domains but also BLC20. In general, BLC20 semantic features seem to be better than BLC50 and SuperSenses.

Regarding verbs, the system obtains significant improvements over the baselines using different types of semantic features both in SE2 and SE3. In particular, when using again WordNet Domains as semantic features.

In general, the results obtained by BLC20 are not so much different to the results of BLC50 (in a few cases, this difference is greater than 2 points). For instance, for nouns, if we consider the number of classes within BLC20 (558 classes), BLC50 (253 classes) and SuperSense (24 classes), BLC classifiers obtain high performance rates while maintaining much higher expressive power than SuperSenses. In fact, using SuperSenses (40 classes for nouns and verbs) we can obtain a very accurate semantic tagger with performances close to 80%. Even better, we can use BLC20 for tagging nouns (558 semantic classes and F1 over 75%) and SuperSenses for verbs (14 semantic classes and F1 around 75%).

Obviously, the classifiers using WordNet Domains as target grouping obtain very high performances due to its reduced average polysemy. However, when used as semantic features it seems to improve the results in most of the cases.

In addition, we obtain very competitive classifiers at a sense level.

Class	Sem. Feat.	SensEval2		SensEval3	
		Poly	All	Poly	All
Sense	baseline	59.66	70.02	64.45	72.30
	basicFeat	61.13	71.20	65.45	73.15
	BLC20	61.93	71.79	65.45	73.15
	BLC50	61.79	71.69	65.30	73.04
	SS	61.00	71.10	64.86	72.70
	WND	61.13	71.20	65.45	73.15
	SUMO	61.66	71.59	65.45	73.15
BLC20	baseline	65.92	75.71	67.98	76.29
	basicFeat	65.65	75.52	64.64	73.82
	BLC20	**68.70**	**77.69**	68.29	76.52
	BLC50	**68.83**	**77.79**	67.22	75.73
	SS	65.12	75.14	64.64	73.82
	WND	**68.97**	**77.88**	65.25	74.24
	SUMO	68.57	77.60	64.49	73.71
BLC50	baseline	67.20	76.65	68.01	76.74
	basicFeat	64.28	74.57	66.77	75.84
	BLC20	**69.72**	**78.45**	68.16	76.85
	BLC50	67.20	76.65	68.01	76.74
	SS	65.60	75.52	65.07	74.61
	WND	**70.39**	**78.92**	65.38	74.83
	SUMO	**71.31**	**79.58**	66.31	75.51
WND	baseline	78.97	86.11	76.74	83.8
	basicFeat	70.96	80.81	67.85	77.64
	BLC20	72.53	81.85	72.37	80.79
	BLC50	73.25	82.33	71.41	80.11
	SS	74.39	83.08	68.82	78.31
	WND	78.83	86.01	76.58	83.71
	SUMO	75.11	83.55	73.02	81.24
SUMO	baseline	66.40	76.09	71.96	79.55
	basicFeat	68.53	77.60	68.10	76.74
	BLC20	65.60	75.52	68.10	76.74
	BLC50	65.60	75.52	68.72	77.19
	SS	68.39	77.50	68.41	76.97
	WND	**68.92**	**77.88**	69.03	77.42
	SUMO	68.92	77.88	70.88	78.76
SS	baseline	70.48	80.41	72.59	81.50
	basicFeat	69.77	79.94	69.60	79.48
	BLC20	71.47	81.07	72.43	81.39
	BLC50	70.20	80.22	72.92	81.73
	SS	70.34	80.32	65.12	76.46
	WND	**73.59**	**82.47**	70.10	79.82
	SUMO	70.62	80.51	71.93	81.05

Table 4: Results for nouns

4.3 Learning curves

We also performed a set of experiments for measuring the behaviour of the class-based WSD system when gradually increasing the number of training examples. These experiments have been carried for nouns and verbs, but only noun results are shown since in both cases, the trend is very similar but more clear for nouns.

The training corpus has been divided in portions of 5% of the total number of files. That is, complete files are added to the training corpus of each incremental test. The files were randomly selected to generate portions of 5%, 10%, 15%, etc. of the SemCor corpus[14]. Then, we train the system on each of the training portions and we test the system on SE2 and SE3. Finally, we also compare the resulting system with the baseline computed over the same training portion.

Figures 1 and 2 present the learning curves over SE2 and SE3, respectively, of a class-based WSD system based on BLC20 using the basic features and the semantic features built with WordNet Domains.

Surprisingly, in SE2 the system only improves the F1 measure around 2% while increasing the training corpus from 25% to 100% of SemCor. In SE3, the system again only improves the F1 measure around 3% while increasing the training corpus from 30% to 100% of SemCor. That is, most of the knowledge required for the class-based WSD system seems to be already present on a small part of SemCor.

Figures 3 and 4 present the learning curves over SE2 and SE3, respectively, of a class-based WSD system based on SuperSenses using the basic features and the semantic features built with WordNet Domains.

Again, in SE2 the system only improves the F1

[14]Each portion contains also the same files than the previous portion. For example, all files in the 25% portion are also contained in the 30% portion.

Class	Sem. Feat.	SensEval2		SensEval3	
		Poly	All	Poly	All
Sense	baseline	41.20	44.75	49.78	52.88
	basicFeat	42.01	45.53	**54.19**	**57.02**
	BLC20	41.59	45.14	**53.74**	**56.61**
	BLC50	42.01	45.53	**53.6**	**56.47**
	SS	41.80	45.34	**53.89**	**56.75**
	WND	42.01	45.53	**53.89**	**56.75**
	SUMO	42.22	45.73	**54.19**	**57.02**
BLC20	baseline	50.21	55.13	54.87	58.82
	basicFeat	52.36	57.06	**57.27**	**61.10**
	BLC20	52.15	56.87	56.07	59.92
	BLC50	51.07	55.90	**56.82**	**60.60**
	SS	51.50	56.29	**57.57**	**61.29**
	WND	**54.08**	**58.61**	57.12	60.88
	SUMO	52.36	57.06	**57.42**	**61.15**
BLC50	baseline	49.78	54.93	55.96	60.06
	basicFeat	**53.23**	**58.03**	58.07	61.97
	BLC20	**52.59**	**57.45**	57.32	61.29
	BLC50	51.72	56.67	57.01	61.01
	SS	52.59	57.45	57.92	61.83
	WND	**55.17**	**59.77**	**58.52**	**62.38**
	SUMO	52.16	57.06	57.92	61.83
WND	baseline	84.80	90.33	84.96	92.20
	basicFeat	84.50	90.14	78.63	88.92
	BLC20	84.50	90.14	81.53	90.42
	BLC50	84.50	90.14	81.00	90.15
	SS	83.89	89.75	78.36	88.78
	WND	85.11	90.52	84.96	92.20
	SUMO	85.11	90.52	80.47	89.88
SUMO	baseline	54.24	60.35	59.69	64.71
	basicFeat	56.25	62.09	61.41	66.21
	BLC20	55.13	61.12	61.25	66.07
	BLC50	56.25	62.09	61.72	66.48
	SS	53.79	59.96	59.69	64.71
	WND	55.58	61.51	61.56	66.35
	SUMO	54.69	60.74	60.00	64.98
SS	baseline	62.79	68.47	76.24	79.07
	basicFeat	**66.89**	**71.95**	75.47	78.39
	BLC20	63.70	69.25	74.69	77.70
	BLC50	63.70	69.25	74.69	77.70
	SS	63.70	69.25	74.84	77.84
	WND	**66.67**	**71.76**	77.02	79.75
	SUMO	64.84	70.21	74.69	77.70

Table 5: Results for verbs

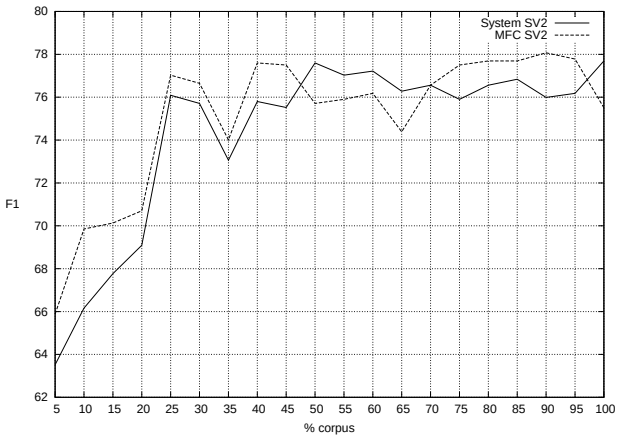

Figure 1: Learning curve of BLC20 on SE2

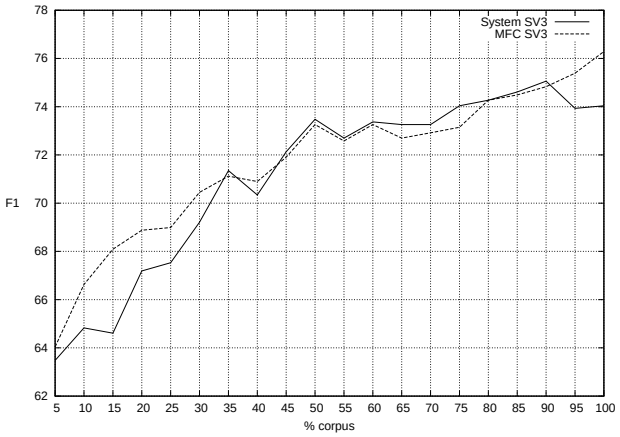

Figure 2: Learning curve of BLC20 on SE3

measure around 2% while increasing the training corpus from 25% to 100% of SemCor. In SE3, the system again only improves the F1 measure around 2% while increasing the training corpus from 30% to 100% of SemCor. That is, with only 25% of the whole corpus, the class-based WSD system reaches a F1 close to the performance using all corpus. This evaluation seems to indicate that the class-based approach to WSD reduces dramatically the required amount of training examples.

In both cases, when using BLC20 or SuperSenses as semantic classes for tagging, the behaviour of the system is similar to MFC baseline. This is very interesting since the MFC obtains high results due to the way it is defined, since the MFC over the total corpus is assigned if there are no occurrences of the word in the training corpus. Without this definition, there would be a large number of words in the test set with no occurrences when using small training portions. In these cases, the recall of the baselines (and in turn F1) would be much lower.

5 Conclusions and discussion

We explored on the WSD task the performance of different levels of abstraction and sense groupings. We empirically demonstrated that Base Level Concepts are able to group word senses into an adequate medium level of abstraction to perform supervised class–based disambiguation. We also demonstrated that the semantic classes provide a rich information about polysemous words and can be successfully used as semantic features. Finally we confirm the fact that the class–based approach reduces dramatically the required amount of training examples, opening the way to solve the well known acquisition bottleneck problem for supervised machine learning algorithms.

In general, the results obtained by BLC20 are not very different to the results of BLC50. Thus, we can select a medium level of abstraction, without having a significant decrease of the perfor-

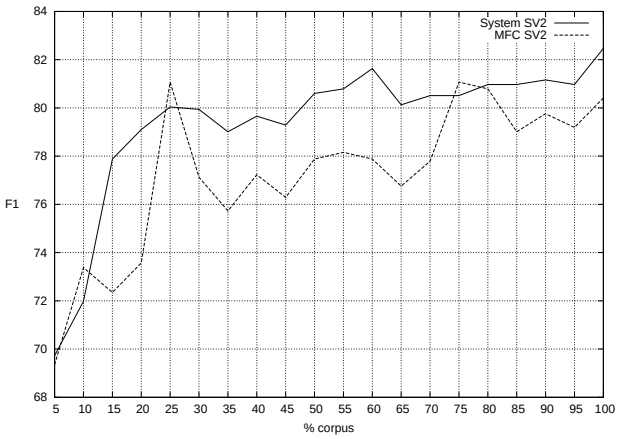

Figure 3: Learning curve of SuperSense on SE2

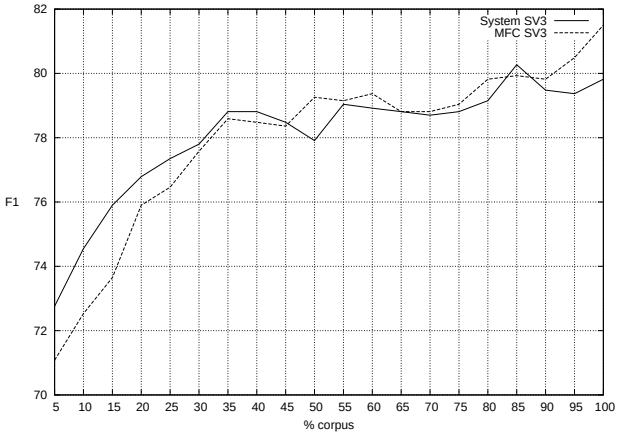

Figure 4: Learning curve of SuperSense on SE3

mance. Considering the number of classes, BLC classifiers obtain high performance rates while maintaining much higher expressive power than SuperSenses. However, using SuperSenses (46 classes) we can obtain a very accurate semantic tagger with performances around 80%. Even better, we can use BLC20 for tagging nouns (558 semantic classes and F1 over 75%) and SuperSenses for verbs (14 semantic classes and F1 around 75%).

As BLC are defined by a simple and fully automatic method, they can provide a user–defined level of abstraction that can be more suitable for certain NLP tasks.

Moreover, the traditional set of features used for sense-based classifiers do not seem to be the most adequate or representative for the class-based approach. We have enriched the usual set of features, by adding semantic information from the monosemous words of the context and the MFC of the target word. With this new enriched set of features, we can generate robust and competitive class-based classifiers.

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 SensEval3. The system achieves an F1–score of 70.54, 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. Although this result is achieved for the all words SensEval3 task, including adjectives, we can compare both results since in SE2 and SE3 adjectives obtain very high performance figures. Using SuperSenses, adjectives only have three classes (WN Lexicographic Files 00, 01 and 44) and more than 80% of them belong to class 00. This yields to really very high performances for adjectives which usually are over 90%.

As we have seen, supervised WSD systems are very dependent of the corpora used to train and test the system. We plan to extend our system by selecting new corpora to train or test. For instance, by using the sense annotated glosses from WordNet.

References

E. Agirre and O. LopezDeLaCalle. 2003. Clustering wordnet word senses. In *Proceedings of RANLP'03*, Borovets, Bulgaria.

J. Alvez, J. Atserias, J. Carrera, S. Climent, E. Laparra, A. Oliver, and G. Rigau G. 2008. Complete and consistent annotation of wordnet using the top concept ontology. In *6th International Conference on Language Resources and Evaluation LREC*, Marrakesh, Morroco.

M. Ciaramita and Y. Altun. 2006. 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. ACL.

M. Ciaramita and M. Johnson. 2003. 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.

J. Curran. 2005. 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.

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

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

R. Izquierdo, A. Suarez, and G. Rigau. 2007. Exploring the automatic selection of basic level concepts. In Galia Angelova et al., editor, *International Conference Recent Advances in Natural Language Processing*, pages 298–302, Borovets, Bulgaria.

T. Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

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

Ll. Màrquez, G. Escudero, D. Martínez, and G. Rigau. 2006. 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.

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

R. Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT 2007*.

G. Miller, C. Leacock, R. Tengi, and R. Bunker. 1993. A Semantic Concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*.

R. Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112, Morristown, NJ, USA. Association for Computational Linguistics.

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

M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, and H. Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of the SENSEVAL-2 Workshop. In conjunction with ACL'2001/EACL'2001*, Toulouse, France.

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

F. Segond, A. Schiller, G. Greffenstette, and J. Chanod. 1997. 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.

R. Snow, Prakash S., Jurafsky D., and Ng A. 2007. Learning to merge word senses. In *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1005–1014.

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

L. Villarejo, L. Màrquez, and G. Rigau. 2005. 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. ISSN 1136-5948.

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

D. Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL'94)*.