On the Quality of Lexical Resources for Word Sense Disambiguation

Lluís Màrquez[1], Mariona Taulé[2], Lluís Padró[1],
Luis Villarejo[1], and Maria Antònia Martí[2]

[1] TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
{lluism,padro,luisv}@lsi.upc.es
[2] Centre de Llenguatge i Computació (CLiC)
Universitat de Barcelona
{mtaule,amarti}@ub.edu

Abstract. Word Sense Disambiguation (WSD) systems are usually evaluated by comparing their absolute performance, in a fixed experimental setting, to other alternative algorithms and methods. However, little attention has been paid to analyze the lexical resources and the corpora defining the experimental settings and their possible interactions with the overall results obtained. In this paper we present some experiments supporting the hypothesis that the quality of lexical resources used for tagging the training corpora of WSD systems partly determines the quality of the results. In order to verify this initial hypothesis we have developed two kinds of experiments. At the linguistic level, we have tested the quality of lexical resources in terms of the annotators' agreement degree. From the computational point of view, we have evaluated how those different lexical resources affect the accuracy of the resulting WSD classifiers. We have carried out these experiments using three different lexical resources as sense inventories and a fixed WSD system based on Support Vector Machines.

1 Introduction

Natural Language Processing applications have to face ambiguity resolution problems at many levels of the linguistic processing. Among them, semantic (or lexical) ambiguity resolution is a currently open challenge, which would be potentially very beneficial for many NLP applications requiring some kind of *language understanding*, e.g., Machine Translation and Information Extraction/Retrieval systems [1].

The goal of WSD systems is to assign the correct semantic interpretation to each word in a text, which basically implies the automatic identification of its sense. In order to be able to address the WSD task, electronic dictionaries and lexicons, and semantically tagged corpora are needed. We assume that these linguistic resources are fundamental to successfully carry out WSD.

One of the approaches to WSD is the *supervised*, in which statistical or Machine Learning (ML) techniques are applied to automatically induce, from se-

J. L. Vicedo et al. (Eds.): EsTAL 2004, LNAI 3230, pp. 291–302, 2004.

mantically annotated corpora, a classification model for sense disambiguation. This approach is typically confronted with the *knowledge-based* approach (also referred sometimes as *unsupervised*[1]) in which some external knowledge sources (e.g., WordNet, dictionaries, parallel corpora, etc.) are used to devise some heuristic rules to perform sense disambiguation, avoiding the use of a manually annotated corpus. Despite the appeal of the unsupervised approach, it has been observed through a substantial body of comparative work, carried out mainly in the Senseval exercises[2], that the ML-based supervised techniques tend to overcome the results of the knowledge–based approach when enough training examples are available. In this paper we will concentrate on the quality of the resources needed to train supervised systems.

We consider that there are two critical points in the supervised WSD process which have been neglected, and are determinant when good results want to be reached: first, the quality of the lexical sources and, second, the quality of the manually tagged corpora. Moreover, the quality of these corpora is determined, to a large extent, by the quality of the lexical source used for carry out the tagging process. Our research has focused both on the evaluation of three different lexical sources: *Diccionario de la Real Academia Española* (DRAE, [2]), MiniDir (MD, [3]), and Spanish WordNet (SWN, [4]), and on how these resources determine the results of the machine learning-based methods for word sense disambiguation.

The methodology followed for the evaluation of the lexical sources is based on the parallel tagging of a single corpus by three different annotators for each lexical source. The annotators' agreement degree will be used for measuring the lexical source quality: the more agreement there is, the more quality the source will have. Thus, a high agreement would indicate that the senses in the lexical source are clearly defined and have a wide coverage. This methodology guarantees objectivity in the treatment of senses.

For measuring the influence of lexical sources in supervised WSD systems, we trained and tested a system based on Support Vector Machines (SVM, [5, 6]) using each of the lexical resources. Results are compared both straightforwardly and after a sense clustering process which intends to compensate for the advantage of disambiguating against a fine-grained resource such as WordNet lexical database or DRAE dictionary.

The rest of the paper is divided into two main parts. The first one is devoted to the analysis of the quality of lexical sources (section 2) and the second one aims at testing whether the best results in the first phase correlate with the best results obtained by the supervised word sense disambiguation system (section 3).

[1] This term is rather confusing since in machine learning terminology, *unsupervised* refers to a learning scenario from unnanotated examples (in which the class labels are omitted). In that case, the goal is to induce clusters of examples, representing the underlying classes.

[2] Senseval is a series of evaluation exercises for Word Sense Disambiguation organized by the ACL-SIGLEX. See `http://www.senseval.org` for more information.

Finally, in section 4 we present the main conclusions drawn and some lines for future work.

2 Lexical Resources Evaluation

Several authors have carried out studies with the aim of proposing specific models and methodologies for the elaboration of lexical sources oriented to WSD tasks. A very outstanding proposal is that of Véronis [7], in which the validity of traditional lexical representation of senses is questioned. This author proposes a model of lexical source suitable for WSD based mainly on syntactic criteria. Kilgarriff [8] developed an experiment on semantic tagging, with the aim to define the upper-bound in manual tagging. In that paper, the upper bound was established at 95% of annotators' agreement. Krishnamurthy and Nichols [9] analyze the process of the gold-standard corpus tagging for Senseval-2, highlighting the most common inconsistencies of dictionaries: incorrect sense division, definition errors, etc. Fellbaum et al. [10] analyze the process of semantic tagging with a lexical resource such as WordNet, but focusing on those features they consider as a source of difficulty: the lexical category, the order of the senses in the lexical source, and the annotators' profile. All the authors highlight the importance of the lexical source as an essential factor in order to obtain quality results. The aim of our research has been to evaluate the quality of lexical resources and test its influence in the quality of results of WSD based on machine learning techniques.

The methodology followed in this work for the evaluation of the lexical resources consists in the manual semantic tagging of a single corpus with three different lexical sources: DRAE, MiniDir, and Spanish WordNet. The tagging process has been carried out by different annotators. This methodology allows us to analyze comparatively the results obtained for each of the lexical sources and, therefore, to determine which of them is the most suitable for WSD tasks. Our starting point is the hypothesis that the annotator agreement degree is proportional to the quality level of the lexical resource: the more agreement there is the more quality has the lexical source.

The evaluated lexical sources present very different characteristics and have been selected for different reasons. Firstly, we have used the *Diccionario de la Real Academia Española*, as it is the reference and normative dictionary of Spanish language. Secondly, *MiniDir-2.1* is a lexicon designed specifically for automatic WSD. This lexical source contains a limited number of entries (50) which have been elaborated specifically as a resource for the Senseval-3 Spanish Lexical Sample Task[3]. Finally, we have also used *Spanish WordNet* as sense repository, since WordNet is one of the most used lexical resources for WSD.

We have performed all the evaluation and comparative experiments using the following subset of ten lexical entries (see the most common translations into English between parentheses). Four nouns: *columna* (column), *corazón* (heart),

[3] See `www.lsi.upc.es/~nlp/senseval-3/Spanish.html` for more information.

SOURCE:MiniDir-2.1; LEMMA:*columna*; POS:ncmfs; SENSE:**1**; DEFINITION:*figura arquitectónica de forma cilíndrica que sirve como soporte o elemento decorativo*; EXAMPLE:*una gran columna de hormigón; una antigua columna del tiempo de los romanos*; SYNONYMS:*manejar*; COLLOCATIONS: *columna_corintia, columna_de_bronce, columna_de_mármol, columna_de_piedra, columna_dórica, columna_griega, columna_jónica*; SYNSETS:02326166n/02326665n/02881716n; DRAE:1

SOURCE:MiniDir-2.1; LEMMA:*columna*; POS:ncmfs; SENSE:**4**; DEFINITION:*forma cilíndrica que toman algunos fluidos o gases cuando ascienden o cuando están contenidos en un cilindro*; EXAMPLE:*una densa columna de humo*; SYNONYMS:?; COLLOCATIONS:*columna_de_agua, columna_de_humo*; SYNSETS: 08508248n; DRAE:3/5

Fig. 1. Example of two Minidir-2.1 lexical entries for *columna*

letra (letter), and *pasaje* (passage). Two adjectives: *ciego* (blind) and *natural* (natural). Four verbs: *apoyar* (to lean/rest; to rely on), *apuntar* (to point/aim; to indicate; to make a note), *explotar* (to exploit; to explode), and *volar* (to fly; to blow up). See more information on these words in table 2.

2.1 The Lexical Sources

In the development of MiniDir-2.1 we have basically taken into account information extracted from corpora. We have used the corpora from the newspapers *El Periódico* and *La Vanguardia*, with a total of 3.5 million and 12.5 million words, respectively, and also Lexesp [11]. The latter is a balanced corpus of 5.5 million words, which includes texts on different topics (science, economics, justice, literature, etc.), written in different styles (essay, novel, etc.) and different language registers (standard, technical, etc.). The corpora provide quantitative and qualitative information which is essential to differentiate senses and to determine the degree of lexicalization. As regards the information of the entries of the dictionary, every sense is organized into the nine following lexical fields: LEMMA, POS CATEGORY[4], SENSE, DEFINITION, EXAMPLE, SYNONYMS (plus ANTONYMS in the case of adjectives), COLLOCATIONS, SYNSETS, DRAE. Figure 1 shows an example of the first and fourth senses of the lexical entry *columna* (column) in MiniDir-2.1. As Minidir-2.1 has a low granularity, in general, its senses correspond to multiple senses in Spanish WordNet. For instance, we can observe that the sense *columna_1* corresponds to three Spanish WordNet synsets (02326166n, 02326665n, and 02881716n).

Because of MiniDir2.1 is a lexical resource build up taking into account WSD, it includes additional information like examples and collocations. Such information, which is not present in the other sources, is potentially very useful for performing word sense disambiguation.

[4] The lexical category is represented by the Eagle tags (Eureka 1989-1995) which have been abridged.

SOURCE:DRAE; LEMMA:*columna*; POS:ncmfs; SENSE:**3**; DEFINITION:*forma que toman algunos fluidos, en su movimiento ascendente. Columna de fuego, de humo*; ... SYNSETS:08508248n; MiniDir-2.1:4

SOURCE:DRAE; LEMMA:*columna*; POS:ncmfs; SENSE:**5**; DEFINITION:*porción de fluido contenido en un cilindro*; ... SYNSETS:08508248n; MiniDir-2.1:4

Fig. 2. Two simplified DRAE lexical entries for the word *columna*

DRAE is a normative dictionary of Spanish language which has not been designed for the computational treatment of language nor word sense disambiguation. Entries have been adapted to the format required by the semantic tagging editor [12] used in the manually semantic tagging. DRAE presents also a high level of granularity and overlapping among definitions. Many senses belong to specific domains and it is also frequent to find outdated senses. Figure 2 contains an example of DRAE entries for senses 3 and 5 of the word *columna* (columna).

The third lexical source we have used is the Spanish WordNet lexical database. It was developed inside the framework of EuroWordNet [4] and includes paradigmatic information (hyperonymy, hyponymy, synonymy, and meronymy). As it is well known, this lexical knowledge base is characterized by its fine granularity and the overlapping of senses, which makes more difficult the annotation process. Spanish WordNet was developed following a semiautomatic methodology [4], which took as reference the English version (WordNet 1.5). Since there is not a one to one correspondence between the senses of both languages, some mismatches appeared in the mapping process. In spite of Spanish WordNet has been checked many times, some mismatches remain and this explains the lack of some senses in Spanish and the excessive granularity for others.

2.2 The Tagging Process

The annotated corpus used for evaluating the different lexical sources (DRAE, MiniDir 2.1 and Spanish WordNet) is the subset of the MiniCors [13] corpus corresponding to the ten selected words. MiniCors was compiled from the corpus of the EFE Spanish News Agency, which includes 289,066 news spanning from January to December of 2000[5], and it has been used as source for the Senseval-3 Spanish Lexical Sample task [14]. The MiniCors corpus contains a minimum of 200 examples for each of the represented words. The context considered for each word is larger than a sentence, as the previous and the following sentences were also included. For each word, the goal was to collect a minimum of 15 occurrences per sense from available corpora, which was not always possible. At the end, only the senses with a sufficient number of examples were included in the final version of the corpus.

The tagging process was carried out by experienced lexicographers and it was developed individually, so as to avoid interferences. Also, the authors of the

[5] The size of the complete EFE corpus is 2,814,291 sentences, 95,344,946 words, with an average of 33.8 words per sentence.

dictionary did not participate in the tagging process. In order to systematize and simplify the annotation process to the utmost, a tagging handbook specifying annotation criteria was designed in an initial phase [12], and a graphical Perl-Tk interface was programmed in order to assist the tagging process. See [14] for more details on the construction and annotation of the MiniCors corpus.

The 10–word subset of MiniCors treated in this paper has been annotated with the senses of DRAE and Spanish WordNet, in addition to the MiniDir-2.1 original annotations. Again, each word has been annotated by three different expert lexicographers in order to facilitate the manual arbitration phase, which was reduced only to cases of disagreement. The annotators could assign more than one tag to the same occurrence in order to reflect more precisely the different agreement degrees.

2.3 Evaluation and Arbitration

Once the corpus has been tagged, we have carried out a comparative study among the different annotations and the subsequent evaluation of the results in order to obtain a disambiguated corpus to begin with the evaluation of the lexical sources. Since each word has been tagged three times for each lexical source, the subsequent process of arbitration has been reduced to those cases of disagreement among the three annotators.

We distinguish 4 different situations of agreement/disagreement between annotators: *total agreement*, *partial agreement*, *minimum agreement*, and *disagreement.* Total agreement takes place when the three annotations completely match (e.g.: 1, 1, 1 $\Rightarrow$ 1). When not all the annotations match but there is a individual sense assigned by all annotators we get partial agreement (e.g.: 1, 1, 1/2 $\Rightarrow$ 1; 1/2, 1/2, 1 $\Rightarrow$ 1). Minimum agreement occurs when two annotations match but the other one is different (e.g.: 1, 1, 2 $\Rightarrow$ 1). Finally, disagreement is produced when none of the annotations match. These agreement cases, either total, partial or minimum, are validated automatically according to the pattern we have previously defined. Only cases of disagreement undergo a manual arbitration phase. We have considered also the pairwise agreements between annotators for the analysis of results. The measure Pairwise Agreement counts the average of the agreement levels between each pair of annotators. In this case, we distinguish among *Minimum Pairwise Agreement* (cases of total agreement among every pair of annotators) and *Maximum Pairwise Agreement* (cases of partial agreement among each pair of annotators).

Table 1 shows the results obtained on each of the previous measures for each sense repository and for each POS category. *NumSenses* is the average number of senses assigned by the annotators.

2.4 Analysis of the Results

The tagging experiments presented in table 1 show that the lexical source which has been designed with specific criteria for WSD, MiniDir-2.1, reaches much higher Total Agreement levels in the manual tagging of corpus than Spanish

Table 1. Per POS category and global annotation agreements using Spanish WordNet, MiniDir-2.1, and DRAE sources

Nouns	SWN	MD-2.1	DRAE	**Adjectives**	SWN	MD-2.1	DRAE
TotAgr	0.64	0.83	0.57	TotAgr	0.15	0.67	0.24
PartAgr	0.12	0.03	0.18	PartAgr	0.42	0.06	0.51
MinAgr	0.20	0.14	0.23	MinAgr	0.33	0.26	0.23
DisAgr	0.04	0.00	0.02	DisAgr	0.10	0.01	0.02
MaxPairAgr	0.83	0.90	0.83	MaxPairAgr	0.70	0.81	0.84
MinPairAgr	0.72	0.88	0.70	MinPairAgr	0.32	0.77	0.69
NumSenses	1.10	1.02	1.08	NumSenses	1.56	1.03	1.12

Verbs	SWN	MD-2.1	DRAE	**Overall**	SWN	MD-2.1	DRAE
TotAgr	0,34	0,66	0,53	TotAgr	0,42	0,72	0,45
PartAgr	0,30	0,08	0,08	PartAgr	0,25	0,06	0,25
MinAgr	0,34	0,25	0,36	MinAgr	0,28	0,21	0,28
DisAgr	0,02	0,01	0,03	DisAgr	0,05	0,01	0,02
MaxPairAgr	0,78	0,83	0,74	MaxPairAgr	0,77	0,85	0,80
MinPairAgr	0,47	0,76	0,67	MinPairAgr	0,50	0,80	0,69
NumSenses	1,53	1,03	1,05	NumSenses	1,39	1,03	1,08

WordNet or DRAE, which stand for lexical sources of common use. The worst results have been obtained by Spanish WordNet, being slightly worse than those of DRAE. We can also analyze the results obtained through three related dimensions: the disagreement measure, the overlapping degree between senses, and the number of senses per entry.

Regarding the disagreement measure, Spanish WordNet has the highest score, 0.05, in front of the 0.02 from DRAE and 0.01 from MiniDir-2.1. That means that the arbitration phase in MiniDir-2.1 and DRAE has been done almost automatically, whereas in the case of Spanish WordNet more manual intervention has been applied. In Spanish WordNet and DRAE we find a high level of overlapping between senses because these dictionaries are very fine grained. These characteristics are reflected in the high numbers for the Partial Agreement measure (compared to MiniDir-2.1) and in the big differences between Maximum and Minimum Pairwise Agreement. This is partially a consequence of the fact that the 1.39 average number of senses assigned to each example in Spanish WordNet is the highest one compared to 1.08 from DRAE and 1.03 from MiniDir-2.1.

If we evaluate the results according to lexical categories, nouns achieve the highest levels of agreement probably because of their referents are more stable and clearly identifiable. As regards adjectives and verbs, the levels of agreement are lower, specially in Spanish WordNet.

The annotation with MiniDir-2.1 reaches results considerably acceptable (with an overall agreement higher than 80% if we sum Total and Partial Agreement cases) that prove their adequacy for WSD tasks. Among the MiniDir-2.1 characteristics that could explain the better results in the annotators agreement degree we should point out the fact that it contains both syntagmatic and co-occurrence information, that constitute determining factors in order to help annotators to decide the correct sense, as it can be seen in the entries for *columna* presented in figure 1.

3 Automatic Disambiguation Experiments

A supervised word sense disambiguation system based on Support Vector Machines has been trained and tested using each of the three lexical resources. This system is the core learning component of two participant systems to the Senseval-3 English Allwords and Lexical Sample tasks, which obtained very competitive results [6, 15].

Support Vector Machines is a learning algorithm for training linear classifiers. Among all possible separating hyperplanes, SVM selects the hyperplane that separates with maximal distance the positive examples from the negatives, i.e., the maximal margin hyperplane. By using kernel functions SVMs can be used also to efficiently work in a high dimensional feature space and learn non-linear classification functions. In our WSD setting, we simply used a linear separator, since some experiments on using polynomial kernels did not provide better results. We used the SVM^{light} freely available implementation by Joachims [5] and a simple one–vs–all binarization scheme to deal with the multiclass classification WSD problem.

Regarding feature representation of the training examples, we used the Feature Extraction module of the TALP team in the Senseval-3 English Lexical Sample task. The feature set includes the classic window–based pattern features extracted from a ±3-token local context and the "bag–of–words" type of features taken from a broader context. It also contains a set of features representing the syntactic relations involving the target word, and some semantic features of the surrounding words extracted from the Multilingual Central Repository of the Meaning project. See [15, 6] for more details about the learning algorithm and the feature engineering used.

We have been working with a total of 1,536 examples, which are the examples in the intersection of the three annotation sources. That means that some examples had to be eliminated from the original Senseval-3 sets, since they could not be assigned to any sense either in the DRAE or Spanish WordNet sense repositories. The training and test partitions have been obtained by randomly selecting 2/3 and 1/3 of the total number of examples, respectively. The total number of training examples is 1,094, while the number of test examples is 543. The number of observed senses for these 10 words (ambiguity rate) range from 3 to 13 depending on the lexical source. Note that, though the DRAE and Spanish WordNet are much more fine-grained than MiniDir-2.1, the difference in the number of senses actually observed in the examples is not dramatic (7.9 and 7.8 versus 5.7). Moreover, the average number of senses according to DRAE and Spanish WordNet are almost identical. See more information about the individual words in table 2.

Table 2. Basic information about the 10 selected words for training and evaluating the SVM-based WSD system

		Number of senses			examples	
word	POS	DRAE	MD-2.1	SWN	#train	#test
apoyar	v	5	3	6	140	51
apuntar	v	8	9	7	124	54
ciego	a	8	5	7	83	49
columna	n	7	8	9	127	63
corazón	n	8	6	8	113	58
explotar	v	6	5	7	131	53
letra	n	10	5	7	92	63
natural	a	9	6	13	92	46
pasaje	n	11	4	7	87	53
volar	v	7	6	7	105	53
avg./total	-	7.9	5.7	7.8	1,094	543

The multiplicity of labels in examples (see the 'NumSenses' row in table 1) has been addressed in the following way. When training, the examples have been replicated, one for each sense label. When testing, we have considered a correct prediction whenever the proposed label is any of the example labels.

The overall and per-word accuracy results obtained are presented in table 3. For each lexical source we include also the results of the baseline Most Frequent Sense classifier (MFS). It can be seen that the MFS results are fairly similar for all three annotation sources (from 46.78% to 47.88%), while the SVM-based systems clearly outperforms the MFS classifier in all three cases. The best results are obtained when using the MiniDir-2.1 lexical source (70.90%), followed by DRAE (67.22%) and Spanish WordNet (66.67%). This accuracy represents an increase of 24.12 percentage points over MFS and an error reduction of 45.32%.

Table 3. WSD results using all three sense repositories: DRAE, MD-2.1, and SWN. Columns 3, 5, and 7 contain the results of the MFS baseline (mosty-frequent sense classifier). Columns 4, 6, and 8 contain the results of the SVM–based system

		DRAE		MD-2.1		SWN	
word	POS	MFS	%ACC.	MFS	%ACC.	MFS	%ACC.
apoyar	v	92.16%	92.16%	88.24%	84.31%	80.39%	68.63%
apuntar	v	55.56%	66.67%	46.30%	68.52%	59.26%	85.19%
ciego	a	57.14%	71.43%	61.22%	75.51%	48.98%	71.43%
columna	n	22.22%	74.60%	20.63%	79.37%	38.10%	74.60%
corazón	n	37.93%	58.62%	43.10%	67.24%	46.55%	65.52%
explotar	v	43.40%	50.94%	43.40%	69.81%	41.51%	64.15%
letra	n	39.68%	61.90%	34.92%	60.32%	41.27%	53.97%
natural	a	58.70%	73.91%	47.83%	65.22%	34.78%	50.00%
pasaje	n	35.85%	60.38%	39.62%	77.36%	37.74%	64.15%
volar	v	47.17%	64.15%	52.83%	62.26%	41.51%	67.92%
average	-	47.88%	67.22%	46.78%	70.90%	46.78%	66.67%

Compared to the other lexical sources, the differences in favor of MiniDir-2.1 are statistically significant with a confidence level of 90% (using a z–test for the difference of two proportions). The difference between MiniDir-2.1 and Spanish WordNet is also significant at 95%. These results provide some empirical evidence which complements the one presented in the previous section. Not only human annotators achieve a higher agreement when using MiniDir, but also a supervised WSD system obtains better results when using this source for training.

Nevertheless, the advantage could be due to the fact that MiniDir-2.1 (5.7 senses/word in average) is a bit coarser grained than DRAE (7.9 senses/word) and WordNet (7.8) on the ten considered words. To compare the lexical resources on a more fair basis, it seems that a new evaluation metric is needed able to compensate for the difference on the number of senses. As a first approach, we clustered together the senses from all lexical sources, following the coarsest of the three (MiniDir-2.1). That is, each DRAE and Spanish WordNet sense was mapped to a MiniDir-2.1 sense, and any answer inside the same cluster was considered correct. This procedure required some manual work in the generation of the mappings between lexical sources. Some ad-hoc decisions were taken in order to correct inconsistencies induced by the more natural mappings between the three sources.

The evaluation according to the sense clusters leaded to some disappointing results. The best overall accuracy results were obtained by DRAE (72.62%), followed by Spanish WordNet (71.19%) and MiniDir-2.1 (70.48%). However, it is worth noting that none of this differences is statistically significant (at a confidence level of 90%). It remains to be studied if this lack of actual differences is due to the small number of examples used in our experiments, or to the fact that the dictionary used is not really affecting very much the achievable performance of supervised machine learning WSD systems. The way in which we addressed the problem of the multiple sense labels per example (see table 1 and above) may tend to favor the evaluation of the most fine-grained lexical sources (Spanish WordNet and DRAE), and partly explaining the lack of differences observed. We think that the design of other evaluation measures, independent of the number of senses and able to isolate the contribution of the lexical sources, deserves also further investigation.

4 Conclusions

In this study we have evaluated different lexical sources in order to determine the most adequate one for WSD tasks. The evaluation has consisted of the tagging of a single corpus with three different dictionaries and different annotators. The agreement degree among the annotators has been the determining criteria to establish the quality of the lexical source.

According to our experiments, MiniDir-2.1, the lexical source designed with specific criteria for WSD, reaches much higher agreement levels (above 80%) in the manual tagging of the corpus than Spanish WordNet or DRAE. The MiniDir-2.1 specific features that help explaining these differences are the following: 1) MiniDir-2.1 is coarser grained than DRAE and Spanish WordNet, avoiding to some extent the overlapping of senses; 2) It contains both syntagmatic and co-occurrence information, which help the annotators to decide the correct senses.

The evaluation of a SVM–based WSD classifier, trained on the three different lexical resources, seems to indicate that a reference dictionary with a higher agreement degree produces also better results for automatic disambiguation.

We also provide results of a first attempt in trying to evaluate the WSD systems with independence of the average number of senses per word, by means of a sense mapping and clustering across lexical sources. Unfortunately, these results showed no significant differences among lexical sources. Up to now, it remains unclear whether the increase in performance produced by the use of a lexical source specifically designed for WSD is mainly explained by the the higher quality of the lexical source or by the decrease on sense granularity. This is an issue that requires further research, including experiments on bigger corpora to produce statistically significant results and a careful design of the evaluation metrics used.

Acknowledgments

This research has been possible thanks to the support of Spanish and European research projects: BFF2002-04226-C03 (X-TRACT2), TIC2000-0335-C03-02 (Hermes), and IST-2001-34460 (Meaning). The team of annotators was composed by members of the research group CLiC of the *Universitat de Barcelona* (UB) and by members of Natural Language Group of the *Universidad Nacional de Educación a Distancia* (UNED).

References

1. Ide, N., Véronis, J.: Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics, Special issue on Word Sense Disambiguation **24** (1998) 1–40
2. Real Academia Española: *Diccionario de la lengua española*, 22nd edition, Madrid, Spain (2001)
3. Artigas, N., García, M., Martí, M., Taulé, M.: *Diccionario MiniDir-2.1.* Technical Report XTRACT2-WP-03/08, Centre de Llenguatge i Computaci (CLiC), Universitat de Barcelona (2003)
4. Vossen, P., ed.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1999)
5. Joachims, T.: Making large–scale SVM learning practical. In Schölkopf, B., Burges, C.J.C., Smola, A.J., eds.: Advances in Kernel Methods — Support Vector Learning. MIT Press, Cambridge, MA (1999) 169–184
6. Villarejo, L., Màrquez, L., Agirre, E., Martínez, D., Magnini, B., Strapparava, C., McCarthy, D., Montoyo, A., Suárez, A.: The "Meaning" system on the english all-words task. In: Proceedings of the Senseval-3 ACL-SIGLEX Workshop, Barcelona, Spain (2004)
7. Véronis, J.: Sense tagging: does it make sense? In: Proceedings of the Corpus Linguistics'2001 Conference, Lancaster, U.K. (2001)

8. Kilgarriff, A.: 95% replicability for manual word sense tagging. In: Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, EACL'99, Bergen, Norway (1999)
9. Krishnamurthy, R., Nicholls, D.: Peeling an onion: The lexicographer's experience of manual sense-tagging. Computers and the Humanities. Special Issue on Evaluating Word Sense Disambiguation Programs **34** (2000) 85–97
10. Fellbaum, C., Grabowsky, J., Landes, S.: Analysis of a hand-tagging task. In: Proceedings of the ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington D.C., USA (1997)
11. Sebastián, N., Martí, M.A., Carreiras, M.F., Gómez, F.C.: *Lexesp, léxico informatizado del español.* Edicions de la Universitat de Barcelona, Barcelona (2000)
12. Artigas, N., García, M., Martí, M., Taulé, M.: *Manual de anotacin semántica.* Technical Report XTRACT2-WP-03/03, Centre de Llenguatge i Computaci (CLiC), Universitat de Barcelona (2003)
13. Taulé, M., Civit, M., Artigas, N., García, M., Màrquez, L., Martí, M., Navarro, B.: Minicors and cast3lb: Two semantically tagged soanish corpora. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC-2004, Lisbon, Portugal (2004)
14. Màrquez, L., Taulé, M., Martí, M.A., García, M., Artigas, N., Real, F., Ferrés, D.: Senseval-3: The spanish lexical sample task. In: Proceedings of the Senseval-3 ACL Workshop, Barcelona, Spain (2004)
15. Escudero, G., Màrquez, L., Rigau, G.: TALP system for the english lexical sample task. In: Proceedings of the Senseval-3 ACL Workshop, Barcelona, Spain (2004)