Senseval-3: The Spanish Lexical Sample Task

L. Màrquez[†], M. Taulé[‡], M.A. Martí[‡], M. García[‡], N. Artigas[‡], F.J. Real[†], D. Ferrés[†]

†TALP Research Center, Software Department Universitat Politècnica de Catalunya {lluism,fjreal,dferres}@lsi.upc.es ‡CLiC, Centre de Llenguatge i Computació Universitat de Barcelona {mtaule,amarti}@ub.edu,{nuripa,mar}@clic.fil.ub.es

1 Introduction

In this paper we describe the Spanish Lexical Sample task. This task was initially devised for evaluating the role of unlabeled examples in supervised and semi-supervised learning systems for WSD and it was coordinated with five other lexical sample tasks (Basque, Catalan, English, Italian, and Rumanian) in order to share part of the target words.

Firstly, we describe the methodology followed to develop the linguistic resources necessary for the task: the MiniDir-2.1 lexicon and the MiniCors corpus. Secondly, we summarize the participant systems, the results obtained, and a comparative analysis. Participant systems include pure supervised, semi–supervised, and unsupervised learning.

2 The Spanish Lexicon: MiniDir-2.1

Due to the enormous effort needed for rigorously developing lexical resource and manually annotated corpora, we limited our work to the treatment of 46 words of three syntactic categories: 21 nouns, 7 adjectives, and 18 verbs. The selection was made trying to maintain the core words of the Senseval-2 Spanish task and sharing around 10 of the target words with Basque, Catalan, English, Italian, and Rumanian lexical tasks. Table 1 shows the set of selected words.

We used the MiniDir-2.1 dictionary as the lexical resource for corpus tagging, which is a subset of the broader MiniDir¹. MiniDir-2.1 was designed as a resource oriented to WSD tasks, i.e., with a granularity level low enough to avoid the overlapping of senses that commonly characterizes lexical sources. Regarding the words selected, the average number of senses per word is 5.33, corresponding to 4.52 senses for the nouns subgroup, 6.78 for verbs and 4 for adjectives (see table 1, right numbers in column '#senses').

The content of MiniDir-2.1 has been checked and refined in order to guarantee not only its consis-

#LEMMA:conducir #POS:VM #SENSE:2 #DEF.: Manejar un vehículo para desplazarse #EXAMPLE: conducir un camión; conduce bien #SYNONYMS: manejar #COLLOC.: carné de conducir; permiso de conducir #SYNSETS: 01100152v;01099937v;01101463v;01176439v

Figure 1: Example of a Minidir-2.1 lexical entry

tency and coverage but also the quality of the gold standard. Each sense in Minidir-2.1 is linked to the corresponding synset numbers in EuroWordNet (Vossen, 1999) and contains syntagmatic information as collocates and examples extracted from corpora². Regarding the dictionary entries, every sense is organized in nine lexical fields. See figure 1 for an example of one sense of the lexical entry *conducir* ('to drive').

3 The Spanish Corpus: MiniCors

MiniCors is a semantically tagged corpus according to the Senseval lexical sample setting, labeled with the MiniDir-2.1 sense repository. The MiniCors corpus is formed by 12,625 tagged examples, covering 35,875 sentences and 1,506,233 words. The context considered for each example includes the target sentence, plus the previous and the following ones. All the examples have been extracted from the year-2000 corpus of the Spanish EFE News Agency, which includes 289,066 news (2,814,291 sentences and 95,344,946 words) spanning from January to December of 2000.

For every word, a minimum of 200 examples have been manually tagged by three independent expert human annotators and disagreement cases have been resolved by another lexicographer (assigning a unique sense to each example). The annotation process has been assisted by a graphical Perl-Tk interface specifically designed for this task, and a

¹MiniDir is a dictionary under development by the CLiC research group, http://clic.fil.ub.es.

 $^{^{2}}$ We have used corpora from newspapers, *El Periódico* (3.5 million words), *La Vanguardia* (12.5 million words), and the *Lexesp* corpus (Sebastián et al., 2000), a balanced corpus of 5.5. million words.

word.POS	#senses	#train / test / unlab	%MFS
actuar.v	3 / 4	133 / 67 / 1,500	73.13
apoyar.v	3/4	259 / 128 / 1,500	92.97
apuntar.v	4/9	213 / 106 / 1,500	50.94
arte.n	3/4	251 / 121 / 1,500	95.87
autoridad.n	2/4	268 / 132 / 1,500	96.97
bajar.v	3/5	235 / 115 / 1,500	84.35
banda.n	4 / 7	230 / 114 / 1,500	70.18
brillante.a	2/2	126 / 63 / 1,369	88.89
canal.n	4/6	262 / 131 / 1,500	65.65
canalizar.v	2/3	253 / 126 / 700	96.83
ciego.a	3/5	102 / 52 / 390	59.62
circuito.n	4 / 5	261 / 132 / 1,500	56.82
columna.n	7/8	129 / 64 / 1,258	20.31
conducir.v	4/5	134 / 66 / 1,094	45.45
corazón.n	3/6	123 / 62 / 1,500	45.16
corona.n	3 / 4	124 / 64 / 916	68.75
duplicar.v	2/2	254 / 126 / 1,500	96.03
explotar.v	5/5	212 / 103 / 1,500	45.63
ganar.v	3 / 8	237 / 118 / 1,500	90.68
gracia.n	3 / 5	72 / 38 / 1,209	50.00
grano.n	3 / 4	117 / 61 / 524	60.66
hermano.n	2/3	128 / 66 / 1,500	90.91
jugar.v	3 / 5	236 / 117 / 1,500	90.60
letra.n	5 / 5	226 / 114 / 1,251	34.21
masa.n	3 / 4	172 / 85 / 1,151	43.53
mina.n	2/4	134 / 66 / 1,458	51.52
natural.a	5/6	215 / 107 / 1,500	46.73
naturaleza.n	3 / 4	258 / 128 / 1,500	67.19
operación.n	3/4	134 / 66 / 1,500	54.55
órgano.n	2/3	263 / 131 / 1,500	85.50
partido.n	2/2	133 / 66 / 1,500	56.06
pasaje.n	4/4	220 / 111 / 375	45.95
perder.v	4 / 11	218 / 106 / 1,500	60.38
popular.a	3/3	133 / 67 / 1,500	44.78
programa.n	3/3	267 / 133 / 1,500	75.19
saltar.v	8 / 15	200 / 101 / 1,117	29.70
simple.a	3 / 4	117 / 61 / 1,500	70.49
subir.v	3 / 5	231 / 114 / 1,500	74.56
tabla.n	3 / 6	130 / 64 / 1,500	76.56
tocar.v	6/13	158 / 78 / 1,500	28.21
tratar.v	3 / 12	143 / 72 / 1,235	43.06
usar.v	2/3	263 / 130 / 1,500	97.69
vencer.v	3 / 7	134 / 65 / 1,500	80.00
verde.a	2 / 5	69 / 33 / 1,500	60.61
vital.a	2/3	131 / 65 / 1,500	75.38
volar.v	3 / 6	122 / 60 / 705	53.33
avg/total	3.30 / 5.33	8,430 / 4,195 / 61,252	67.72

Table 1: Information about Spanish datasets

tagging handbook for the annotators. The interannotator complete agreement achieved was 90% for nouns, 83% for adjectives, and 83% for verbs. These are the best results obtained in a comparative study (Taulé et al., 2004) with other dictionaries used for tagging the same corpus. The senses corresponding to multi-word expressions were eliminated since they are not considered in MiniDir-2.1.

The initial goal was to obtain for each word at least 75 examples plus 15 examples per sense. For the words below these figures we performed a second round by labeling up to 200 examples more. After that, senses with less than 15 occurrences (\leq 3.5% of the examples) have been simply discarded from the datasets. See table 1, left numbers in column '#senses', for the final ambiguity rates. We know that this is a quite controversial decision that leads to a simplified setting. But we preferred to maintain the proportions of the senses naturally appearing in the EFE corpus rather than trying to artificially find examples of low frequency senses by mixing examples from many sources or by getting them with specific predefined patterns. Thus, systems trained on the MiniCors corpus are intended to discriminate between the typical word senses appearing in a news corpus.

4 Resources Provided to Participants

Participants were provided with the complete Minidir-2.1 dictionary, a training set with 2/3 of the labeled examples, a test set with 1/3 of the examples and a complementary big set of unlabeled examples, limited to 1,500 for each word (when available). Each example is provided with a non null list of category-labels marked according to two annotation schemes: ANPA and IPTC³.

Aiming at helping teams with few resources on the Spanish language, sentences in all datasets were tokenized, lemmatized and POS tagged, using the Spanish linguistic processors developed at TALP– $CLiC^4$, and provided as complementary files. Table 1 contains information about the sizes of the datasets and the proportion of the most-frequent sense for each word (MFC). The baseline MFC classifier obtains a high accuracy of 67.72% due to the moderate number of senses considered.

5 The Participant Systems

Seven teams took part on the Spanish Lexical Sample task, presenting a total of nine systems. We will refer to them as: IRST, UA-NSM, UA-NP, UA-SRT, UMD, UNED, SWAT, Duluth-SLSS, and CSUSMCS. From them, seven are supervised and two unsupervised (UA-NSM, UA-NP). Only one of the participant systems uses a mixed learning strategy that allows to incorporate the knowledge from the unlabeled examples, namely UA-SRT. It is a Maximum Entropy–based system, which makes use of a re-training algorithm (inspired by Mitchell's cotraining) for iteratively relabeling unannotated examples with high precision and adding them to the training of the MaxEnt algorithm.

³All the datasets of the Spanish Lexical Sample task and an extended version of this paper are available at: http://www.lsi.upc.es/~nlp/senseval-3/Spanish.html.

⁴http://www.lsi.upc.es/~nlp/freeling.

Regarding the supervised learning approaches applied, we find Naive Bayes and Decision Lists (SWAT), Maximum Entropy (UA-SRT), Decision Trees (Duluth-SLSS), Support Vector Machines (IRST), AdaBoost (CSUSMCS), and a similarity method based on co-occurrences (UNED). Some systems used a voted combination of these basic learning algorithms to produce the final WSD system (SWAT, Duluth-SLSS). The two unsupervised algorithms apply only to nouns and target at obtaining high precision results (the annotations on adjectives and verbs come from a supervised MaxEnt system). UA-NSM method is called Specification Marks and uses the words that co-occur with the target word and their relation in the noun WordNet hierarchy. UA-NP bases the disambiguation on syntactic patterns and unsupervised corpus, relying on the "one sense per pattern" assumption.

All supervised teams used the POS and lemmatization provided by the organization, except Duluth-SLSS, which only used raw lexical information. A few systems used also the category labels provided with the examples. Apparently, none of them used the extra information in MiniDir (examples, collocations, synonyms, WordNet links, etc.), nor syntactic information. Thus, we think that there is room for substantial improvement in the feature set design. It is worth mentioning that the IRST system makes use of a kernel including semantic information within the SVM framework.

6 Results and System Comparison

Table 2 presents the global results of all participant systems, including the MFC baseline (most frequent sense classifier) and sorted by the combined F_1 measure. The COMB row stands for a voted combination of the best systems (see the last part of the section). As it can be seen, IRST and UA-SRT are the best performing systems, with no significant differences between them⁵.

All supervised systems outperformed the MFC baseline, with a best overall improvement of 16.48 points (51.05% relative error reduction)⁶. Both unsupervised systems performed below MFC.

It is also observed that the POS and lemma information used by most supervised systems is relevant, since Duluth-SLSS (based solely on raw lexical information) performed significantly worse than the rest of supervised systems⁷.

System	prec.	recall	cover.	$F_{\beta=1}$
IRST	84.20%	84.20%	100.0%	84.20
UA-SRT	84.00%	84.00%	100.0%	84.00
UMD	82.48%	82.48%	100.0%	82.48
UNED	81.76%	81.76%	100.0%	81.76
SWAT	79.45%	79.45%	100.0%	79.45
D-SLSS	74.29%	75.02%	100.0%	74.65
CSUSMCS	67.84%	67.82%	99.9%	67.83
UA-NSM	61.93%	61.93%	100.0%	61.93
UA-NP	84.31%	47.27%	56.1%	60.58
MFC	67.72%	67.72%	100.0%	67.72
COMB	85.98%	85.98%	100.0%	85.98

Table 2: Overall results of all systems

Detailed results by groups of words are showed in table 3. Word groups include part–of–speech, intervals of the proportion of the most frequent sense (%MFS), intervals of the ratio *number of examples per sense* (ExS), and the words in the retraining set used by UA-SRT (those with a MFC accuracy lower than 70% in the training set). Each cell contains precision and recall. Bold face results correspond to the best system in terms of the F_1 score. Last column, Δ -error, contains the best F_1 improvement over the baseline: absolute difference and error reduction (%).

As in many other previous WSD works, verbs are the most difficult words (13.07 improvement and 46.7% error reduction), followed by adjectives (19.64, 52.1%), and nouns (20.78, 59.4%). The gain obtained by all methods on words with high MFC (more than 90%) is really low, indicating the difficulties of supervised ML algorithms at acquiring information about non-frequent senses). On the contrary, the gain obtained on the lowest MFC words is really good (44.3 points and 62.5% error reduction). This is a very good property of the Spanish dataset and the participant systems, which is not always observed in other empirical studies using other WSD corpora (e.g., in the Senseval-2 Spanish task values of 29.9 and 43.1% were observed). The two unsupervised systems failed at achieving a performance on nouns comparable to the baseline classifier. UA-NP has the best precision but at a cost of an extremely low recall (below 5%).

It is also observed that participant systems are quite different along word groups, being the best performances shared between IRST, UA-SRT, UMD, and UNED systems. Interestingly, IRST is the best system addressing the words with less examples per sense, suggesting that SVM is a good learning algorithm for training on small datasets, but loses this advantage for the words with more

⁵Statistical significance was tested with a z-test (0.95 confidence level) for the difference of two proportions.

⁶These improvement figures are better than those observed in the Senseval-2 Spanish lexical sample task: 17 points but only 32.69% of error reduction.

⁷With the exception of CSUSMCS, which according to ta-

ble 3 shows a non-regular behavior with abnormal low results on some groups of words.

	IRST	UA-SRT	UMD	UNED	SWAT	D-SLSS	CSU	UA-NSM	UA-NP	MFC	Δ -error
adjs (prec)	81.92	81.25	74.78	75.67	74.55	71.27	66.74	81.47	81.47	62.28	19.64
(rec)	81.92	81.25	74.78	75.67	74.55	71.43	66.74	81.47	81.47	62.28	52.1%
nouns	83.89	84.25	85.79	85.58	80.25	73.60	67.42	36.38	88.68	65.01	20.78
	83.89	84.25	85.79	85.58	80.25	74.65	67.42	36.38	4.82	65.01	59.4%
verbs	85.09	84.43	80.81	79.14	79.81	75.80	68.56	84.76	84.76	72.02	13.07
	85.09	84.43	80.81	79.14	79.81	76.31	68.56	84.76	84.76	72.02	46.7%
%MFS	97.17	97.17	96.69	96.69	96.69	96.69	97.17	83.31	96.90	96.69	0.48
(95,100)	97.17	97.17	96.69	96.69	96.69	96.69	97.17	83.31	64.09	96.69	14.5%
%MFS	92.77	92.54	91.38	91.84	91.38	91.61	65.50	90.68	92.56	91.38	1.39
(90,95)	92.77	92.54	91.38	91.84	91.38	91.61	65.50	90.68	78.32	91.38	16.1%
%MFS	89.04	90.11	86.36	90.37	86.10	85.71	83.16	63.10	88.89	84.76	5.61
(80,90)	89.04	90.11	86.36	90.37	86.10	86.63	83.16	63.10	57.75	84.76	36.8%
%MFS	83.82	88.51	80.91	85.11	78.64	75.08	75.89	59.06	85.59	73.62	14.89
(70,80)	83.82	88.51	80.91	85.11	78.64	75.08	75.89	59.06	46.12	73.62	56.4%
%MFS	79.73	78.59	80.31	80.88	69.98	66.22	57.93	48.37	80.25	64.44	16.4
(60,70)	79.73	78.59	80.31	80.88	69.98	66.35	57.93	48.37	24.86	64.44	46.2%
%MFS	81.06	76.11	80.38	77.82	76.96	67.91	60.34	44.54	70.04	54.27	26.8
(50,60)	81.06	76.11	80.38	77.82	76.96	68.26	60.34	44.54	28.33	54.27	58.6%
%MFS	78.75	75.33	72.07	66.57	71.03	61.18	56.02	57.80	78.31	45.17	33.6
(40,50)	78.75	75.33	72.07	66.57	71.03	61.81	56.02	57.80	48.29	45.17	61.3%
%MFS	68.35	73.39	71.43	64.71	62.75	49.35	37.54	49.30	65.92	29.13	44.3
(0,40)	68.35	73.39	71.43	64.71	62.75	52.94	37.54	49.30	33.05	29.13	62.5%
ExS	92.55	93.42	91.17	92.55	90.39	89.47	88.92	68.57	95.64	89.26	4.16
>120	92.55	93.42	91.17	92.55	90.39	89.78	88.92	68.57	47.53	89.26	38.7%
ExS	86.70	88.32	86.04	88.41	83.38	80.44	64.77	68.00	91.61	75.21	13.2
(90,120)	86.70	88.32	86.04	88.41	83.38	80.44	64.77	68.00	54.99	75.21	53.2%
ExS	79.39	78.00	77.71	74.58	74.07	65.32	58.89	54.55	77.34	53.97	25.42
(60,90)	79.39	78.00	77.71	74.58	74.07	65.99	58.85	54.55	39.04	53.97	55.2%
ExS	74.92	72.31	70.68	66.12	64.17	56.04	53.42	55.54	70.42	45.11	29.81
(30,60)	74.92	72.31	70.68	66.12	64.17	58.14	53.42	55.54	51.95	45.11	54.3%
retrain	78.34	76.79	77.05	73.86	71.59	62.75	55.46	48.42	74.42	50.73	27.61
	78.34	76.79	77.05	73.86	71.59	63.78	55.44	48.42	32.80	50.73	56.0%

Table 3: Results of all participant systems on some selected subsets of words

examples. These facts opens the avenue for further improvements on the Spanish dataset by combining the outputs of the best performing systems. As a first approach, we conducted some simple experiments on system combination by considering a voting scheme, in which each system votes and the majority sense is selected (ties are decided favoring the best method prediction). From all possible sets, the best combination includes the five systems with the best precision figures: UA-NP, IRST, UMD, UNED, and SWAT. The resulting F₁ measure is 85.98, 1.78 points higher than the best single system (see table 2). This improvement comes mainly from the better F₁ performance on nouns: from 83.89 to 87.28.

We also calculated the agreement rate and the Kappa statistic between each pair of systems. The agreement ratios ranged from 40.93% to 88.10%, and the Kappa values from 0.40 to 0.87. It is worth noting that the system relying on the simplest feature set (Duluth-SLSS) obtained the most similar output to the most frequent sense classifier.

7 Acknowledgements

This work has been supported by the Spanish research projects: XTRACT-2, BFF2002-04226-C03-03; FIT-150-500-2002-244; and HERMES, TIC2000-0335-C03-02. Francis Real holds a research grant by the Catalan Government (2002FI-00648). Authors want to thank the linguists of CLiC and UNED who collaborated in the annotation task.

References

- N. Sebastián, M. A. Martí, M. F. Carreiras, and F. Cuetos Gómez. 2000. *Lexesp, léxico informatizado del español*. Edicions de la Universitat de Barcelona.
- M. Taulé, M. Civit, N. Artigas, M. García, L. Márquez, M. A. Martí, and B. Navarro. 2004. MiniCors and Cast3LB: Two Semantically Tagged Spanish Corpora. In *Proceedings of the* 4th LREC, Lisbon.
- P. Vossen, editor. 1999. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht.