SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

Senseval-3: The Catalan Lexical Sample Task

L. Màrquez†, **M. Taulé**‡, **M.A. Martí**‡, **M. García**‡, **F.J. Real**†, and **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, mar@clic.fil.ub.es

1 Introduction

In this paper we describe the Catalan 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 is the counterpart of the Spanish Lexical Sample task. It was coordinated also with other lexical sample tasks (Basque, English, Italian, Rumanian, and Spanish) in order to share part of the target words.

Firstly, we describe the methodology followed for developing the specific linguistic resources necessary for the task: the MiniDir-Cat lexicon and the MiniCors-Cat corpus. Secondly, we briefly describe the seven participant systems, the results obtained, and a comparative evaluation between them. All participant teams applied only pure supervised learning algorithms.

2 The Catalan Lexicon: MiniDir-Cat

Catalan language participates for the first time in the Senseval evaluation exercise. Due to the time constraints we had to reduce the initial expectations on providing annotated corpora for up to 45 words to the final 27 word set treated. We preferred to reduce the number of words, while maintaining the quality in the dictionary development, corpus annotation process, and number of examples per word. These words belong to three syntactic categories: 10 nouns, 5 adjectives, and 12 verbs. The selection was made by choosing a subset of the Spanish lexical sample task and trying to share around 10 of the target words with Basque, English, Italian, and Rumanian lexical sample tasks. See table 1 for a complete list of the words.

We used the MiniDir-Cat dictionary as the lexical resource for corpus tagging, which is a dictionary being developed by the CLiC research group[1]. MiniDir-Cat was conceived specifically as a resource oriented to WSD tasks: we have emphasized low granularity in order to avoid the overlapping of senses usually present in many lexical sources.

[1]http://clic.fil.ub.es.

#LEMMA:*banda* #POS:NCFS #SENSE:2
#DEFINITION: *Grup de persones que s'uneixen amb fins comuns, especialment delictius*
#EXAMPLE: *una banda que prostituïa dones i robava cotxes de luxe; la banda ultra de l'Atlètic de Madrid*
#SYNONYMS: *grup; colla*
#COLLOCATIONS: *banda armada; banda juvenil; banda de delinqüents; banda mafiosa; banda militar; banda organitzada; banda paramilitar; banda terrorista; banda ultra*
#SYNSETS: 05249044n; 05269792n

Figure 1: Example of a MiniDir-Cat entry

Regarding the polysemy of the selected words, the average number of senses per word is 5.37, corresponding to 4.30 senses for the nouns subset, 6.83 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 consistency and coverage but also the quality of the gold standard. Each sense in Minidir-2.1 is linked to the corresponding synset numbers in the semantic net EuroWordNet (Vossen, 1999) (zero, one, or more synsets per sense) and contains syntagmatic information as collocates and examples extracted from corpora[2]. Every sense is organized in the nine following lexical fields: LEMMA, POS, SENSE, DEFINITION, EXAMPLES, SYNONYMS, ANTONYMS (only in the case of adjectives), COLLOCATIONS, and SYNSETS. See figure 1 for an example of one sense of the lexical entry *banda* (noun 'gang').

3 The Catalan Corpus: MiniCors-Cat

MiniCors-Cat is a semantically tagged corpus according to the Senseval lexical sample setting, so one single target word per example is semantically labeled with the MiniDir-Cat sense repository. The MiniCors-Cat corpus is formed by 6,722 tagged examples, covering 45,509 sentences and 1,451,778 words (with an average of 31.90 words

[2]We have used a 3.5 million subset of the newspaper *El Periódico* in the Catalan version.

word.POS	#senses	#train / test / unlab	%MFS
actuar.v	2 / 3	197 / 99 / 2,442	80.81
apuntar.v	5 / 11	184 / 93 / 1,881	50.54
autoritat.n	2 / 2	188 / 93 / 102	87.10
baixar.v	3 / 4	189 / 92 / 1,572	59.78
banda.n	3 / 5	149 / 75 / 180	60.00
canal.n	3 / 6	188 / 95 / 551	56.84
canalitzar.v	2 / 2	196 / 99 / 0	79.80
circuit.n	4 / 4	165 / 83 / 55	46.99
conduir.v	5 / 7	198 / 101 / 764	63.37
cor.n	4 / 7	144 / 72 / 634	50.00
explotar.v	3 / 4	193 / 98 / 69	72.45
guanyar.v	2 / 6	184 / 92 / 2,106	76.09
jugar.v	4 / 4	115 / 61 / 0	57.38
lletra.n	5 / 6	166 / 86 / 538	30.23
massa.n	2 / 3	145 / 74 / 33	59.46
mina.n	2 / 4	185 / 92 / 121	90.22
natural.a	3 / 6	170 / 88 / 2,320	80.68
partit.n	2 / 2	180 / 89 / 2,233	95.51
passatge.n	2 / 4	140 / 70 / 0	55.71
perdre.v	2 / 8	157 / 78 / 2,364	91.03
popular.a	3 / 3	137 / 70 / 2,472	51.43
pujar.v	2 / 4	191 / 95 / 730	71.58
saltar.v	6 / 17	111 / 60 / 134	38.33
simple.a	2 / 3	148 / 75 / 310	85.33
tocar.v	6 / 12	161 / 78 / 789	37.18
verd.a	2 / 5	128 / 64 / 1,315	79.69
vital.a	3 / 3	160 / 81 / 220	60.49
avg/total	3.11 / 5.37	4,469 / 2,253 / 23,935	66.36

Table 1: Information on the Catalan datasets

per sentence). The context considered for each example includes the paragraph in which the target word occurs, plus the previous and the following paragraphs. All the examples have been extracted from the corpus of the ACN Catalan news agency, which includes about 110,588 news (January 2000–December 2003). This corpus has been tagged with POS. Following MiniDir-2.1, those examples containing the current word in a multiword expression have been discarded.

For every word, a total of 300 examples have been manually tagged by two independent expert human annotators, though some of them had to be discarded due to errors in the automatic POS tagging and multiword filtering. In the cases of disagreement a third lexicographer defined the definitive sense tags. All the annotation process has been assisted by a graphical Perl-Tk interface specifically designed for this task (in the framework of the Meaning European research project), and a tagging handbook for the annotators (Artigas et al., 2003). The inter–annotator agreement achieved was very high: 96.5% for nouns, 88.7% for adjectives, 92.1% for verbs, 93.16% overall.

The initial goal was to obtain, for each word, at least 75 examples plus 15 examples per sense. However, after the labeling of the 300 examples, senses with less than 15 occurrences were simply discarded from the Catalan 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 ACN 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-Cat corpus are only intended to discriminate between the most important word senses appearing in a general news corpus.

4 Resources Provided to Participants

Participants were provided with the complete Minidir-Cat dictionary, a training set with 2/3 of the labeled examples, a test set with 1/3 of the examples and a complementary large set of all the available unlabeled examples in the ACN corpus (with a maximum of 2,472 extra examples for the adjective *popular*). Each example is provided with a non null list of category-labels marked according to the newspaper section labels (politics, sports, international, etc.)[3]. Aiming at helping teams with few resources on the Catalan language, all corpora were tokenized, lemmatized and POS tagged, using the Catalan linguistic processors developed at TALP–CLiC[4], and provided to participants.

Table 1 contains information about the sizes of the datasets and the proportion of the most-frequent sense for each word (MFC). This baseline classifier obtains a high accuracy of 66.36% due to the small number of senses considered.

5 The Participant Systems

Five teams took part on the Catalan Lexical Sample task, presenting a total of seven systems. We will refer to them as: IRST, SWAT-AB, SWAT-CP, SWAT-CA, UNED, UMD, and Duluth-CLSS. All of them are purely supervised machine learning approaches, so, unfortunately, none of them incorporates the knowledge from the unlabeled examples. Most of these systems participated also in the Spanish lexical sample task, with almost identical configurations.

Regarding the supervised learning approaches applied, we find AdaBoost, Naive Bayes, vector–based cosine similarity, and Decision Lists (SWAT systems), Decision Trees (Duluth-CLSS), Support

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

[4] http://www.lsi.upc.es/∼nlp/freeling.

Vector Machines (IRST), and a similarity method based on co-occurrences (UNED). Some systems used a combination of these basic learning algorithms to produce the final WSD system. For instance, Duluth-CLSS applies a bagging–based ensemble of Decision Trees. SWAT-CP performs a majority voting of Decision Lists, the cosine–based vector model and the Bayesian classifier. SWAT-CA combines, again by majority voting, the previous three classifiers with the AdaBoost based SWAT-AB system. The Duluth-CLSS system is a replica of the one presented at the Senseval-2 English lexical sample task.

All teams used the POS and lemmatization provided by the organization, except Duluth-CLSS, 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 within the SVM framework, including semantic information. See IRST system description paper for more information.

6 Results and System Comparison

Table 2 presents the global results of all participant systems, including the MFC baseline (most frequent sense classifier), sorted by the combined F_1 measure. The COMB row stands for a voted combination of the best systems (see last part of the section for a description). As in the Spanish lexical sample task the IRST system is the best performing one. In this case it achieves a substantial improvement with respect to the second system (SWAT-AB)[5].

All systems obtained better results than the baseline MFC classifier, with a best overall improvement of 18.87 points (56.09% relative error reduction)[6]. For the multiple systems presented by SWAT, the combination of learning algorithms in the SWAT-CP and SWAT-CA did not help improving the accuracy of the basic AdaBoost–based system SWAT-AB. It is also observed that the POS and Lemma information used by most systems is relevant, since the system relying only on raw lexical information (Duluth-CLSS) performed significantly worse than the rest (confidence level 0.95).

System	prec.	recall	cover.	$F_{\beta=1}$
IRST	85.82%	84.64%	98.6%	85.23
SWAT-AB	83.39%	82.47%	98.9%	82.93
UNED	81.85%	81.85%	100.0%	81.85
UMD	81.46%	80.34%	98.6%	80.89
SWAT-CP	79.67%	79.67%	100.0%	79.67
SWAT-CA	79.58%	79.58%	100.0%	79.58
Duluth-CLSS	75.37%	76.48%	100.0%	75.92
MFC	66.36%	66.36%	100.0%	66.36
COMB	86.86%	86.86%	100.0%	86.86

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), and intervals of the ratio: number of examples per sense (ExS). 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 significantly more difficult (16.67 improvement and 49.3% error reduction) than nouns (23.46, 65.6%). The improvements obtained by all methods on words with high MFC (more than 90%) is generally low. This is not really surprising, since statistically–based supervised ML algorithms have difficulties at acquiring information about non-frequent senses. Notice, however, the remarkable 44.9% error reduction obtained by SWAT-AB, the best system on this subset. On the contrary, the gain obtained on the lowest MFC words is really good (34.2 points and 55.3% error reduction). This is a good property of the Catalan dataset and the participant systems, which is not always observed in other empirical studies using other WSD corpora. It is worth noting that even better results were observed in the Spanish lexical sample task.

Systems are quite different along word groups: IRST is globally the best but not on the words with highest (between 80% and 100%) an lowest (less than 50%) MFC, in which SWAT-AB is better. UNED and UMD are also very competitive on nouns but overall results are penalized by the lower performance on adjectives (specially UNED) and verbs (specially UMD). Interestingly, IRST is the best system addressing the words with few examples per sense, suggesting that SVM is a good algorithm for training on small datasets, but loses this advantage for the words with more examples.

All, these facts, open the avenue for further im-

[5]The difference is statistically significant using a z-test for the difference of two proportions with a confidence level of 0.90. If we raise the confidence level to 0.95 we lose significance by a short margin: $z = 1.93 < 1.96$.

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

	IRST	SWAT-ME	UNED	UMD	SWAT-CA	SWAT-CP	D-CLSS	MFC	Δ-error
adjs (prec)	**86.51**%	76.20%	79.10%	82.28%	83.33%	85.45%	79.63%	71.69%	14.82
(rec)	**86.51**%	71.16%	79.10%	82.28%	83.33%	85.45%	79.63%	71.69%	52.3%
nouns	**87.68**%	87.45%	86.61%	87.44%	82.87%	80.70%	78.38%	64.17%	23.46
	87.58%	87.45%	86.61%	87.33%	82.87%	80.70%	80.46%	64.17%	65.5%
verbs	**84.06**%	82.60%	79.06%	76.28%	75.62%	76.77%	71.43%	66.16%	16.67
	81.64%	82.60%	79.06%	74.09%	75.62%	76.77%	72.18%	66.16%	49.3%
%MFS	94.16%	**95.75**%	94.98%	94.16%	93.82%	93.82%	93.05%	92.28%	3.47
(90,100)	93.44%	**95.75**%	94.98%	93.44%	93.82%	93.82%	93.05%	92.28%	44.9%
%MFS	88.73%	**90.42**%	87.04%	85.63%	87.04%	87.61%	83.66%	83.38%	7.04
(80,90)	88.73%	**90.42**%	87.04%	85.63%	87.04%	87.61%	83.66%	83.38%	42.4%
%MFS	**89.86**%	85.71%	86.83%	83.11%	82.59%	84.60%	7.82%	75.67%	13.79
(70,80)	**89.06**%	85.71%	86.83%	82.37%	82.59%	84.60%	80.36%	75.67%	56.7%
%MFS	**85.17**%	81.09%	80.85%	80.62%	77.78%	78.25%	75.35%	60.76%	23.90
(59,70)	**84.16**%	81.09%	80.85%	79.67%	77.78%	78.25%	75.89%	60.76%	60.9%
%MFS	**82.93**%	77.98%	74.19%	75.05%	72.89%	71.80%	66.25%	53.58%	28.99
(50,59)	**82.21**%	73.75%	74.19%	74.40%	72.89%	71.80%	68.55%	53.58%	62.5%
%MFS	74.23%	**72.31**%	70.36%	73.88%	67.10%	65.15%	59.12%	38.11%	34.20
(0,50)	70.36%	**72.31**%	70.36%	70.03%	67.10%	65.15%	61.24%	38.11%	55.3%
ExS	91.77%	**91.96**%	91.35%	88.72%	88.62%	87.56%	85.05%	82.85%	9.11
>120	91.35%	**91.96**%	91.35%	88.32%	88.62%	87.56%	85.43%	82.85%	53.1%
ExS	87.66%	**88.08**%	86.38%	86.25%	83.90%	85.45%	80.03%	69.50%	18.58
(90,120)	86.84%	**88.08**%	86.38%	85.45%	83.90%	85.45%	81.27%	69.50%	60.9%
ExS	**83.06**%	76.11%	76.10%	77.42%	76.51%	75.70%	71.40%	61.04%	21.86
(60,90)	**82.73**%	72.29%	76.10%	77.11%	76.51%	75.70%	71.69%	61.04%	56.1%
ExS	**77.21**%	71.78%	67.78%	67.91%	64.00%	63.78%	59.40%	43.56%	31.89
(30,60)	**73.78**%	71.78%	67.78%	64.89%	64.00%	63.78%	61.78%	43.56%	56.5%

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

provements on the Catalan dataset by combining the outputs of the best performing systems, or by performing a selection of the best at word level. 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 of systems turned out to be: IRST, SWAT-AB, and UNED. The resulting F_1 measure is 86.86, 1.63 points higher than the best single system (see table 2). This improvement comes mainly from the better F_1 performance on noun and verb categories: from 87.63 to 90.11 and from 82.63 to 85.47, respectively.

Finally, see the agreement rates and the Kappa statistic between each pair of systems in table 4. Due to space restrictions we have indexed the systems by numbers: 1=MFC, 2=UMD, 3=IRST, 4=UNED, 5=D-CLSS, 6=SWAT-AB, 7=SWAT-CP, and 8=SWAT-CA. The upper diagonal contains the agreement ratios varying from 70.13% to 96.01%, and the lower diagonal contains the corresponding Kappa values, ranging from 0.67 and 0.95. It is worth noting that the system relying on the simplest feature set (Duluth-CLSS) obtains the most similar output to the most frequent sense classifier, and that the combination-based systems SWAT-CP and SWAT-CA generate almost the same output.

	1	2	3	4	5	6	7	8
1	–	78.96	72.79	74.43	87.84	70.13	82.25	84.42
2	0.77	–	82.78	85.66	83.95	82.51	84.55	87.31
3	0.70	0.81	–	81.05	81.98	80.74	85.13	85.18
4	0.71	0.84	0.79	–	79.87	82.56	81.76	83.31
5	0.86	0.82	0.80	0.77	–	77.19	86.77	88.88
6	0.67	0.81	0.79	0.81	0.75	–	79.09	82.69
7	0.80	0.83	0.83	0.80	0.85	0.77	–	96.01
8	0.82	0.86	0.83	0.81	0.87	0.81	0.95	–

Table 4: Agreement and Kappa values between each pair of systems

7 Acknowledgements

This work has been supported by the research projects: XTRACT-2, BFF2002-04226-C03-03; FIT-150-500-2002-244; HERMES, TIC2000-0335-C03-02; and MEANING, IST-2001-34460. Francis Real holds a predoctoral grant by the Catalan Government (2002FI-00648).

References

N. Artigas, M. García, M. Taulé, and M. A. Martí. 2003. Manual de anotación semántica. Working Paper XTRACT-03/03, CLiC, UB.

P. Vossen, editor. 1999. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.