Framework and Results for the Spanish SENSEVAL

German Rigau, Mariona Taulé, Ana Fernandez and **Julio Gonzalo**
g.rigau@lsi.upc.es, TALP Research Center, Universitat Politècnica de Catalunya
mtaule@lingua.fil.ub.es, CLiC, Universitat de Barcelona
ana.fernandez@uab.es, CLiC, Universitat Autònoma de Barcelona
julio@lsi.uned.es, GPLN, Universidad Nacional de Educación a Distancia

Abstract

In this paper we describe the structure, organisation and results of the SENSEVAL exercise for Spanish. We present several design decisions we taked for the exercise, we describe the creation of the gold-standard data and finally, we present the results of the evaluation. Twelve systems from five different universities were evaluated. Final scores ranged from 0.56 to 0.65.

1 Introduction

In this paper we describe the structure, organisation and results of the Spanish exercise included within the framework of SENSEVAL-2.

Although we closely follow the general architecture of the evaluation of SENSEVAL-2, the final setting of the Spanish exercise involved a number of choices detailed in section 2. In the following sections we describe the data, the manual tagging process (including the inter-tagger agreement figures), the participant systems and the accuracy results (including some baselines for comparison purposes).

2 Design Decisions

2.1 Task Selection

For Spanish SENSEVAL, the lexical-sample variant for the task was chosen. The main reasons for this decision are the following:

- During the same tagging session, it is easier and quicker to concentrate only on one word at a time. That is, tagging multiple instances of the same word.
- The all-words task requires access to a full dictionary. To our knowledge, there are no full Spanish dictionaries available (with low or no cost). Instead, the lexical-sample task required only as many dictionary entries as words in the sample task.

2.2 Word Selection

The task for Spanish is a "lexical sample" for 39 words[1] (17 nouns, 13 verbs, and 9 adjectives). See table 1 for the complete list of all words selected for the Spanish lexical sample task. The words can belong only to one of the syntactic categories. The fourteen words selected to be translation-equivalents to English has been:

- Nouns: *arte* (=art), *autoridad* (= authority), *canal* (= channel), *circuito* (= circuit), and *naturaleza* (= nature).
- Verbs: *conducir* (= drive), *tratar* (= treat), and *usar* (= use).
- Adjectives: *ciego* (= blind), *local* (= local), *natural* (= natural), *simple* (= simple), *verde* (= green), and *vital* (= vital).

2.3 Corpus Selection

The corpus was collected from two different sources: "El Periódico"[2] (a Spanish newspaper) and LexEsp[3] (a balanced corpus of 5.5 million words). The length of corpus samples is the sentence.

2.4 Selection of Dictionary

The lexicon provided was created specifically for the task and it consists of a definition for each sense linked to the Spanish version of EuroWordNet and, thus, to the English WordNet 1.5. The syntactic category and, sometimes, examples and synonyms are also provided. The connections to EuroWordNet have been provided in order to have a common language independent conceptual structure. Neither proper nouns nor multiwords has been considered. We have also provided the complete mapping between WordNet 1.5 and 1.6 versions[4]. Each dictionary entry have been constructed consulting the cor-

[1] The noun "arte" was not included in the exercise because it was provided to the competitors during the trial phase.

[2] The working corpus of the HERMES project CICYT TIC2000-0335-C03-02. More details at http://http://terral.ieec.uned.es/hermes.

[3] Provided by LEXESPIII project DGICYT APC 99-0105

[4] http://www.lsi.upc.es/~nlp/mapping.html

pus and multiple Spanish dictionaries (including the Spanish WordNet).

2.5 Annotation procedure

The Spanish SENSEVAL annotation procedure was divided into three consecutive phases.

- Corpus and dictionary creation
- Annotation
- Referee process

All these processes have been possible thanks to the effort of volunteers from three NLP groups from Universitat Politècnica de Catalunya[5] (UPC), Universitat de Barcelona[6] (UB) and Universidad Nacional de Educación a Distancia[7] (UNED).

2.5.1 Corpus and Dictionary Creation

The most important and crucial task was carried out by the UB team of linguists, headed by Mariona Taulé. They were responsible for the selection of the words, the creation of the dictionary entries and the selection of the corpus instances. First, this team selected the polysemous words for the task consulting several dictionaries including the Spanish WordNet and a quick inspection to the Spanish corpus. For the words selected, the dictionary entries were created simultaneously with the annotation of all occurrences of the word. This allowed the modification of the dictionary entries (i.e. adapting the dictionary to the corpus) during the annotation and the elimination of unclear corpus instances (i.e. adapting the corpus to the dictionary).

2.5.2 Annotation

Once the Spanish SENSEVAL dictionary and the annotated corpus were created, all the data was delivered to the UPC and UNED teams, removing all the sense tags from the corpus. Having the Spanish SENSEVAL dictionary provided by the UB team as the unique semantic reference for annotation both teams performed in parallel and simultaneously a new annotation of the whole corpus. Both teams where allowed to provide comments/problems on the each of the corpus instances.

2.5.3 Referee Control

Finally, in order to provide a coherent annotation, a unique referee from the UPC team collate both annotated corpus tagged by the UPC and the UNED teams. This referee was not integrated in the UPC team in the previous annotating phase. The referee was in fact providing a new annotation for each instance when occurring a disagreement between the sense tags provided by the UPC and UNED teams.

[5] http://www.lsi.upc.es/∼nlp
[6] http://www.ub.es/ling/labing.htm
[7] http://rayuela.ieec.uned.es/

3 The Spanish data

3.1 Spanish Dictionary

The Spanish lexical sample is a selection of high, medium and low polysemy frequent nouns, verbs and adjectives. The dictionary has 5.10 senses per word and the polysemy degree ranges from 2 to 13. Nouns has 3.94 ranging from 2 to 10, verbs 7.23 from 4 to 13 and adjectives 4.22 from 2 to 9 (see table 1 for further details).

The lexical entries of the dictionary have the following form:

$$
\begin{array}{l}
< HEADWORD >\# \\
< POS >\# \\
< SENSENUMBER >\# \\
< GLOSS : EXAMPLEs >\# \\
SIN :< SINONYMWORDs >\# \\
< SYNSETNUMBERs >\#
\end{array}
$$

Figure 1: Dictionary entry format

For instance, the dictionary for noun headword *arte* (= art) is:

arte#NCMS#1#Actividad humana o producto de tal actividad que expresa simbólicamente un aspecto de la realidad: el arte de la música; el arte precolombino #SIN:?#00518008n/02980374n#

arte#NCMS#2#Sabiduría, destreza o habilidad de una persona en una actividad o conducta determinada: tiene mucho arte bailando; desplegó todo su arte para convencerle #SIN:?#03850627n#

arte#NCMS#3#Aparato que sirve para pescar#SIN:?#02005770n#

3.2 Spanish Corpus

We adopted, when possible, the guidelines proposed by the SENSEVAL organisers (Edmonds, 2000). For each word selected having n senses we provided at least $75 + 15n$ instances. For the adjective *popular* a larger set of instances has been provided to test performance improvement when increasing the number of examples. These data has been then ramdomly divided in a ratio of 2:1 between training and test set.

The corpus was structured following the standard SENSEVAL XML format.

3.3 Major problems during annotation

In this section we discuss the most frequent and regular types of disagreement between annotators.

In particular, the dictionary proved to be not sufficiently representative of the selected words to be annotated. Although the dictionary was built for the task, out of 48% of the problems during the second phase of the annotation where due to the lack of the appropriate sense in the corresponding dictionary entry. This portion includes 5% of metaphorical uses not explicitly described into the dictionary entry. Furthermore, 51% of the problems reported by the annotators were concentrated only on five words (*pasaje*, *canal*, *bomba*, *usar*, and *saltar*).

Selecting only one sentence as a context during annotation was the other main problem. Around 26% of the problems where attributed to insufficient context to determine the appropriate sense.

Other sources of minor problems included different Part-of-Speech from the one selected for the word to be annotated, and sentences with multiple meanings.

3.4 Inter-tagger agreement

In general, disagreement between annotators (and sometimes the use of multiple tags) must be interpreted as misleading problems in the definition of the dictionary entries. The inter-tagger agreement between UPC and UNED teams was 0.64% and the Kappa measure 0.44%.

4 The Systems

Twelve systems from five teams participated in the Spanish task.

- Universidad de Alicante (UA) combined a Knowledge-based method and a supervised method. The first uses WordNet and the second a Maximum Entropy model.
- John Hopkins University (JHU) presented a metalearner of six diverse supervised learning subsystems integrated via classifier. The subsystems included decision lists, transformation-based error-driven learning, cosine-based vector models, decision stumps and feature-enhanced naive Bayes systems.
- Stanford University (SU) presented a metalearner mainly using Naive Bayes methods, but also including vector space, n-gram, and KNN classifiers.
- University of Maryland (UMD) used a margin-based algorithm to the task: Support Vector Machine.
- University of Manitoba (d6-10,dX-Z) presented different combinations of classical Machine Learning algorithms.

5 The Results

Table 1 presents the results in detail for all systems and all words. The best scores for each word are highlighted in boldface. The best average score is obtained by the JHU system. This system is the best in 12 out of the 39 words and is also the best for nouns and verbs but not for adjectives. The SU system gets the highest score for adjectives.

The associated agreement and kappa measures for each system are shown in Table 2. Again JHU system scores higher in both agreement and Kappa measures. This indicates that the results from the JHU system are closer to the corpus than the rest of participants.

6 Conclusions and Further Work

Obviously, an in deep study of the strengths and weaknesses of each system with respect to the results of the evaluation must be carried out, including also further analysis comparing the UPC and UNED annotations against each system.

Following the ideas described in (Escudero et al., 2000) we are considering also to add a cross-domain aspect to the evaluation in future SENSEVAL editions, allowing the training on one domain and the evaluation on the other, and vice-versa.

In order to provide a common platform for evaluating different WSD algorithms we are planning to process the Spanish corpus tagged with POS using MACO (Carmona et al., 1998) and RELAX (Padró, 1998).

7 Acknowledgements

The Spanish SENSEVAL has been possible thanks to the effort of volunteers from three NLP groups from UPC, UB, and UNED universities.

References

J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. 1998. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the First International Conference on Language Resources and Evaluation, LREC*, Granada, Spain.

P. Edmonds. 2000. Designing a task for SENSEVAL-2. Draft, Sharp Laboratories, Oxford.

G. Escudero, L. Màrquez, and G. Rigau. 2000. A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the 4th Computational Natural Language Learning Workshop, CoNLL*, Lisbon, Portugal.

L. Padró. 1998. *A Hybrid Environment for Syntax-Semantic Tagging*. Phd. Thesis, Software Department (LSI). Technical University of Catalonia (UPC).

words	p	e	s	MF	UA	SU	JHU	UMD	d6	d7	d8	d9	d10	dX	dY	dZ
actuar	v	155	6	0.28	0.27	0.60	0.56	0.45	0.25	0.27	0.40	0.36	0.35	0.22	**0.67**	0.22
apoyar	v	210	4	0.64	0.63	**0.70**	0.68	0.67	0.64	0.63	0.67	0.64	0.66	0.66	0.64	0.64
apuntar	v	191	8	0.47	0.55	0.55	**0.65**	0.53	0.49	0.49	0.51	0.51	0.55	0.49	0.47	0.49
autoridad	n	122	6	0.49	**0.68**	0.50	0.53	0.47	0.50	0.56	0.56	0.47	0.62	0.47	0.62	0.50
bomba	n	113	2	0.71	0.27	0.70	0.68	0.73	0.78	0.71	0.79	**0.80**	0.74	0.78	0.59	**0.80**
brillante	a	256	2	0.52	0.63	0.76	**0.83**	0.76	0.81	0.76	0.81	0.76	0.78	0.73	0.78	0.78
canal	n	156	5	0.33	0.34	0.63	0.68	**0.76**	0.49	0.59	0.56	0.51	0.56	0.56	0.46	0.59
ciego	a	114	4	0.54	**0.71**	0.69	0.62	0.62	0.64	0.55	0.57	0.60	0.60	0.60	0.55	0.57
circuito	n	123	4	0.34	0.43	**0.59**	0.57	0.37	0.49	0.55	0.61	0.31	0.53	0.53	0.29	0.49
claro	a	204	7	0.83	0.82	**0.88**	0.82	0.83	0.83	0.85	0.85	0.83	0.86	0.85	0.85	0.85
clavar	v	131	9	0.44	0.50	0.64	0.48	0.64	0.61	**0.68**	0.64	0.52	0.61	0.57	0.57	0.57
conducir	v	150	9	0.35	0.35	0.43	0.44	**0.46**	0.41	0.43	0.43	0.35	0.41	0.37	0.41	0.41
copiar	v	147	8	0.32	0.42	0.55	0.45	0.47	0.45	0.40	0.42	0.53	0.43	0.38	**0.62**	0.42
corazon	n	146	5	0.36	0.23	0.53	0.77	0.68	0.66	0.74	**0.79**	0.53	0.77	0.64	0.68	0.62
corona	n	119	4	0.45	0.53	**0.80**	0.70	0.53	0.55	0.62	0.57	0.55	0.57	0.55	0.53	0.55
coronar	v	244	6	0.32	0.49	0.65	**0.70**	0.65	0.55	0.62	0.61	0.64	0.61	0.59	0.41	0.62
explotar	v	133	6	0.32	0.49	0.56	0.56	0.56	0.46	0.39	0.41	0.49	0.41	0.44	**0.61**	0.41
gracia	n	160	6	0.30	0.28	**0.79**	0.74	0.61	0.69	0.66	0.79	0.59	0.72	0.70	0.70	0.80
grano	n	78	3	0.44	0.37	0.32	0.50	0.45	0.36	0.50	0.32	0.32	0.45	0.36	**0.64**	0.36
hermano	n	135	5	0.61	0.74	0.58	0.74	0.72	0.70	0.74	0.74	0.70	**0.75**	0.70	0.74	0.74
local	a	139	3	0.74	0.84	0.78	**0.89**	0.75	0.76	0.84	0.85	0.73	0.84	0.78	0.82	0.82
masa	n	131	5	0.45	0.39	0.63	**0.68**	0.61	0.54	0.54	0.61	0.56	0.66	0.56	0.41	0.59
natural	a	137	6	0.25	0.34	0.48	**0.60**	0.45	0.36	0.41	0.40	0.31	0.47	0.41	0.38	0.41
naturaleza	n	167	10	0.44	0.45	0.66	0.59	0.54	0.64	0.70	0.66	0.52	**0.68**	0.57	0.64	0.59
operacion	n	142	5	0.35	**0.71**	0.60	0.55	0.49	0.43	0.45	0.40	0.57	0.45	0.47	**0.60**	0.47
organo	n	212	4	0.52	0.73	**0.83**	0.81	0.73	0.70	0.64	0.64	0.70	0.64	0.64	0.53	0.68
partido	n	159	2	0.55	0.81	0.84	**0.86**	0.81	0.74	0.74	0.74	0.67	0.75	0.72	0.67	0.77
pasaje	n	112	4	0.39	**0.83**	0.44	0.56	0.34	0.39	0.39	0.39	0.32	0.56	0.41	0.29	0.39
popular	a	661	3	0.65	0.77	**0.90**	0.83	0.75	0.77	0.78	0.80	0.71	0.77	0.77	0.68	0.75
programa	n	142	6	0.49	0.36	0.49	**0.64**	0.49	0.49	**0.64**	0.55	0.47	0.40	0.40	0.49	0.45
saltar	v	137	14	0.15	0.51	0.49	0.57	0.51	0.16	0.35	0.32	0.11	0.54	0.32	0.65	0.30
simple	a	217	5	0.61	0.67	**0.77**	0.63	0.65	0.68	0.70	0.72	0.65	0.72	0.67	0.67	0.65
tabla	n	119	3	0.51	**0.88**	0.73	0.66	0.71	0.66	0.59	0.73	0.76	0.68	0.73	0.59	0.76
tocar	v	236	12	0.31	0.51	0.61	**0.66**	0.59	0.41	0.51	0.49	0.39	0.47	0.42	0.34	0.42
tratar	v	192	13	0.21	0.39	0.46	**0.60**	0.56	0.27	0.39	0.37	0.30	0.43	0.30	0.24	0.34
usar	v	167	4	0.68	0.77	0.73	**0.79**	0.70	0.70	0.68	0.70	0.70	0.64	0.70	0.70	0.70
vencer	v	183	8	0.63	**0.72**	0.69	0.62	0.69	0.69	**0.72**	0.71	0.69	0.71	0.69	0.71	0.69
verde	a	109	9	0.37	0.48	0.61	0.52	0.64	0.58	0.58	0.61	0.61	**0.67**	0.48	0.55	**0.67**
vital	a	256	4	0.45	0.65	0.68	**0.77**	0.68	0.54	0.67	0.68	0.51	0.66	0.47	0.53	0.51
NOUNS	n	2336	4	0.45	0.55	0.63	**0.66**	0.59	0.58	0.61	0.61	0.55	0.62	0.58	0.56	0.60
VERBS	v	2276	7	0.40	0.51	0.59	**0.60**	0.58	0.47	0.5.	0.51	0.48	0.52	0.47	0.54	0.48
ADJS	a	2093	4	0.58	0.66	**0.73**	0.72	0.68	0.66	0.68	0.70	0.63	0.71	0.64	0.65	0.67
TOTAL	T	6705	5	0.48	0.56	0.64	**0.65**	0.61	0.56	0.59	0.60	0.55	0.61	0.56	0.57	0.57

Table 1: Evaluation of Spanish words. **p** stands for Part-of-Speech; **e** for the total number of examples (including train and test sets); **s** for the number of senses; **MF** for the Most Frequent Sense Classifier and the rest are the system acronyms.

words	UA	SU	JHU	UMD	d6	d7	d8	d9	d10	dX	dY	dZ
Agreement	0.51	0.63	**0.65**	0.61	0.55	0.57	0.59	0.53	0.59	0.55	0.51	0.57
Kappa	0.20	0.34	0.47	0.20	0.13	0.19	0.23	0.06	0.24	0.15	-0.03	0.15

Table 2: Agreement and Kappa measures