The Basque lexical-sample task

Eneko Agirre, Itziar Aldabe, Mikel Lersundi, David Martinez, Eli Pociello, Larraitz Uria[*]
IxA NLP group, Basque Country University
649 pk. 20.080 Donostia, Spain
eneko@si.ehu.es

Abstract

In this paper we describe the Senseval 3 Basque lexical sample task. The task comprised 40 words (15 nouns, 15 verbs and 10 adjectives) selected from the Basque WordNet. 10 of the words were chosen in coordination with other lexical-sample tasks. The examples were taken from newspapers, an in-house balanced corpus and Internet texts. We additionally included a large set of untagged examples, and a lemmatised version of the data including lemma, PoS and case information. The method used to hand-tag the examples produced an inter-tagger agreement of 78.2% before arbitration. The eight competing systems attained results well above the most frequent baseline and the best system from Swarthmore College scored 70.4% recall.

1 Introduction

This paper reviews the Basque lexical-sample task organized for Senseval 3. Each participant was provided with a relatively small set of labelled examples (2/3 of 75+15*senses+7*multiwords) and a comparatively large set of unlabelled examples (roughly ten times more when possible) for around 40 words. The larger number of unlabelled data was released with the purpose to enable the exploration of semi-supervised systems. The test set comprised 1/3 of the tagged examples. The sense inventory was taken from the Basque WordNet, which is linked to WordNet version 1.6 (Fellbaum, 1998). The examples came mainly from newspaper texts, although we also used a balanced in-house corpus and texts from Internet. The words selected for this task were coordinated with other lexical-sample tasks (such as Catalan, English, Italian, Romanian and Spanish) in order to share around 10 of the target words.

The following steps were taken in order to carry out the task:

1. set the exercise
   a. choose sense inventory from a pre-existing resource
   b. choose target corpora
   c. choose target words
   d. lemmatize the corpus automatically
   e. select examples from the corpus
2. hand-tagging
   a. define the procedure
   b. revise the sense inventory
   c. tag
   d. analyze the inter-tagger agreement
   e. arbitrate

This paper is organized as follows: The following section presents the setting of the exercise. Section 3 reviews the hand-tagging, and Section 4 the details of the final release. Section 5 shows the results of the participant systems. Section 6 discusses some main issues and finally, Section 7 draws the conclusions.

2 Setting of the exercise

In this section we present the setting of the Basque lexical-sample exercise.

2.1 Basque

As Basque is an agglutinative language, the dictionary entry takes each of the elements necessary to form the different functions. More specifically, the affixes corresponding to the determinant, number and declension case are taken in this order and independently of each other (deep morphological structure). For instance, '*etxekoari emaiozu*' can be roughly translated as '*[to the one in the house] [give it]*' where the underlined sequence of suffixes in Basque corresponds to '*to the one in the*'.

2.2 Sense inventory

We chose the Basque WordNet, linked to WordNet 1.6, for the sense inventory. This way, the hand tagging enabled us to check the sense coverage and overall quality of the Basque WordNet, which is under construction. The Basque WordNet is available at http://ixa3.si.ehu.es/wei3.html.

[*] Authors listed in alphabetic order.

2.3 Corpora used

Being Basque a minority language it is not easy to find the required number of occurrences for each word. We wanted to have both balanced and newspaper examples, but we also had to include texts extracted from the web, specially for the untagged corpus. The procedure to find examples from the web was the following: for each target word all possible morphological declensions were automatically generated, searched in a search-engine, documents retrieved, automatically lemmatized (Aduriz et al. 2000), filtered using some heuristics to ensure quality of context, and finally filtered for PoS mismatches. Table 1 shows the number of examples from each source.

2.4 Words chosen

Basically, the words employed in this task are the same words used in Senseval 2 (40 words, 15 nouns, 15 verbs and 10 adjectives), only the sense inventory changed. Besides, in Senseval 3 we replaced 5 verbs with new ones. The reason for this is that in the context of the MEANING project[1] we are exploring multilingual lexical acquisition, and there are ongoing experiments that focus on those verbs. (Agirre et al. 2004; Atserias et al. 2004).

In fact, 10 words in the English lexical-sample have translations in the Basque, Catalan, Italian, Romanian and Spanish lexical tasks: *channel*, *crown*, *letter*, *program*, *party* (nouns), *simple* (adjective), *play*, *win*, *lose*, *decide* (verbs).

2.5 Selection of examples from corpora

The minimum number of examples for each word according to the task specifications was calculated as follows:

$$N=75+15*senses+7*multiwords$$

As the number of senses in WordNet is very high, we decided to first estimate the number of senses and multiwords that really occur in the corpus. The taggers were provided with a sufficient number of examples, but they did not have to tag all. After they had tagged around 100 examples, they would count the number of senses and multiwords that had occurred and computed the N according to those counts.

The context is constituted of 5 sentences, including the sentence with the target word appearing in the middle. Links were kept to the source corpus, document, and to the newspaper section when applicable.

The occurrences were split at random in training set (2/3 of all occurrences) and test set (1/3).

	Total	(N)	(B)	(I)
# words	40			
# senses	316			
# number of tagged examples	7362	5695	924	743
# number of untagged examples	62498	-	-	62498
# tags	9887			

Table 1: Some figures regarding the task. N, B and I correspond to the source of the examples: newspaper, balanced corpus and Internet respectively.

3 Hand tagging

Three persons, graduate linguistics students, took part in the tagging. They are familiar with word senses, as they are involved in the development of the Basque WordNet. The following procedure was defined in the tagging of each word.

- Before tagging, one of the linguists (the editor) revised the 40 words in the Basque WordNet. She had to delete and add senses to the words, specially for adjectives and verbs, and was allowed to check the examples in the corpus.
- The three taggers would meet, read the glosses and examples given in the Basque WordNet and discuss the meaning of each synset. They tried to agree and clarify the meaning differences among the synsets. For each word two hand-taggers and a referee is assigned by chance.
- The number of senses of a word in the Basque WordNet might change during this meeting; that is, linguists could agree that one of the word's senses was missing, or that a synset did not fit with a word. This was done prior to looking at the corpus. Then, the editor would update the Basque WordNet according to those decisions before giving the taggers the final synset list. Overall (including first bullet above), 143 senses were deleted and 92 senses added, leaving a total of 316 senses. This reflects the current situation of the Basque WordNet, which is still under construction.
- Two taggers independently tagged all examples for the word. No communication was allowed while tagging the word.
- Multiple synset tags were allowed, as well as the following tags: the lemma (in the case of multiword terms), U (unassignable), P (proper noun), and X (incorrectly lemmatized). Those with an X were removed from the final release. In the case of proper nouns and multiword terms no synset tag was assigned. Sometimes the U tag was used for word senses which are not in the Basque WordNet. For instance, the sense of *kanal* corresponding to *TV channel*, which is the most frequent sense in the

[1] http://www.lsi.upc.es/~nlp/meaning/meaning.html

examples, is not present in the Basque WordNet (it was not included in WordNet 1.6).

- A program was used to compute agreement rates and to output those occurrences where there was disagreement. Those occurrences were grouped by the senses assigned.
- A third tagger, the referee, reviewed the disagreements and decided which one was the correct sense (or senses).

The taggers were allowed to return more than one sense, and they returned 9887 tags (1.34 per occurrence). Overall, the two taggers agreed in at least one tag 78.2% of the time. Some words attained an agreement rate above 95% (e.g. nouns *kanal* or *tentsio*), but others like *herri – town/people/nation–* attained only 52% agreement. On average, the whole tagging task took 54 seconds per occurrence for the tagger, and 20 seconds for the referee. However, this average does not include the time the taggers and the referee spent in the meetings they did to understand the meaning of each synset. The comprehension of a word with all its synsets required 45.5 minutes on average.

4 Final release

Table 1 includes the total amount of hand-tagged and untagged examples that were released. In addition to the usual release, the training and testing data were also provided in a lemmatized version (Aduriz et al. 2000) which included lemma, PoS and case information. The motivation was twofold:

- to make participation of the teams easier, considering the deep inflection of Basque.
- to factor out the impact of different lemmatizers and PoS taggers in the system comparison.

5 Participants and Results

5 teams took part in this task: Swarthmore College (swat), Basque Country University (BCU), Instituto per la Ricerca Scientifica e Tecnologica (IRST), University of Minnesota Duluth (Duluth) and University of Maryland (UMD). All the teams presented supervised systems which only used the tagged training data, and no other external resource. In particular, no system used the pointers to the full texts, or the additional untagged texts. All the systems used the lemma, PoS and case information provided, except the BCU team, which had additional access to number, determiner and ellipsis information directly from the analyzer. This extra information was not provided publicly because of representation issues.

	Prec.	Rec.	Attempted
basque-swat_hk-bo	71.1	70.4	99.04 %
BCU_Basque_svm	69.9	69.9	100.00 %
BCU_-_Basque_Comb	69.5	69.5	100.00 %
swat-hk-basque	67.0	67.0	100.00 %
IRST-Kernels-bas	65.5	65.5	100.00 %
swat-basque	64.6	64.6	100.00 %
Duluth-BLSS	60.8	60.8	100.00 %
UMD_SST1	65.6	58.7	89.42 %
MFS	55.8	55.8	100.00 %

Table 2: Results of systems and MFS baseline, ordered according to Recall.

We want to note that due to a bug, a few examples were provided without lemmas.

The results for the fine-grained scoring are shown in Table 2, including the Most Frequent Sense baseline (MFS). We will briefly describe each of the systems presented by each team in order of best recall.

- Swat presented three systems based in the same set of features: the best one was based on Adaboost, the second on a combination of five learners (Adaboost, maximum entropy, clustering system based on cosine similarity, decision lists, and naïve bayes, combined by majority voting), and the third on a combination of three systems (the last three).
- BCU presented two systems: the first one based on Support Vector Machines (SVM) and the second on a majority-voting combination of SVM, cosine based vectors and naïve bayes.
- IRST participated with a kernel-based method.
- Duluth participated with a system that votes among three bagged decision trees.
- UMD presented a system based on SVM.

The winning system is the one using Adaboost from Swat, followed closely by the BCU system using SVM.

6 Discussion

These are the main issues we think are interesting for further discussion.

Sense inventory. Using the Basque WordNet presented some difficulties to the taggers. The Basque WordNet has been built using the translation approach, that is, the English synsets have been 'translated' into Basque. The taggers had some difficulties to comprehend synsets, and especially, to realize what makes a synset different from another. In some cases the taggers decided to group some of the senses, for instance, in *herri – town/people/nation–* they grouped 6 senses. This explains the relatively high number of tags per occurrence (1.34). The taggers think that the tagging would be much more satisfactory if they had defined the word senses directly from the corpus.

Basque WordNet quality. There was a mismatch between the Basque WordNet and the corpus: most of the examples were linked to a specific genre, and this resulted in i) having a handful of senses in the Basque WordNet that did not appear in our corpus and ii) having some senses that were not included in the Basque WordNet. Fortunately, we already predicted this and we had a preparation phase where the editor enriched WordNet accordingly. Most of the deletions in the preliminary part were due to the semi-automatic method to construct the Basque WordNet. All in all, we think that tagging corpora is the best way to ensure the quality of the WordNets and we plan to pursue this extensively for the improvement of the Basque WordNet.

7 Conclusions and future work

5 teams participated in the Basque lexical-sample task with 8 systems. All of the participants presented supervised systems which used lemma, PoS and case information provided, but none used the large amount of untagged senses provided by the organizers. The winning system attained 70.4 recall. Regarding the organization of the task, we found that the taggers were more comfortable grouping some of the senses in the Basque WordNet. We also found that tagging word senses is essential for enriching and quality checking of the Basque WordNet.

Acknowledgements

The work has been partially funded by the European Commission (MEANING project IST-2001-34460). Eli Pociello has a PhD grant from the Basque Government.

References

I. Aduriz, E. Agirre, I. Aldezabal, I. Alegria, X. Arregi, J.M. Arriola, X. Artola, K. Gojenola, A. Maritxalar, K. Sarasola, M. Urkia. 2000. A Word-grammar Based Morphological Analyzer for Agglutinative Languages. In *Proceedings of the International Conference on Computational Linguistics* (COLING). Saarbrucken, Germany.

E. Agirre, A. Atutxa, K. Gojenola, K. Sarasola. 2004. Exploring portability of syntactic information from English to Basque. In *Proceedings of the 4rd International Conference on Languages Resources and Evaluations* (LREC). Lisbon, Portugal.

J. Atserias, B. Magnini, O. Popescu, E. Agirre, A. Atutxa, G. Rigau, J. Carroll and R. Koeling 2004. Cross-Language Acquisition of Semantic Models for Verbal Predicates. In *Proceedings of the 4rd International Conference on Languages Resources and Evaluations* (LREC). Lisbon, Portugal.

C. Fellbaum. 1998. *WordNet: An electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts.