Handling of missing values in lexical acquisition

Núria Bel
Universitat Pompeu Fabra
IULA, Roc Boronat 138, 08018 Barcelona, Spain
E-mail: nuria.bel@upf.edu

Abstract

We propose a strategy to reduce the impact of the sparse data problem in the tasks of lexical information acquisition based on the observation of linguistic cues. It justifies that the uncertainty created by missing values, i.e. non-observed cues, can be handled by estimating its likelihood of being observable. Because of the Zipfian distribution of words, instead of estimating the likelihood from the data, we exploit the correlation drawn from the fact that a lexical class is based on the observation of different cues. We obtained experimental results that show a clear benefit of the proposed approach.

1. Introduction

Lexical coverage is crucial to achieve the proper performance of any processing component for NLP applications that relies on lexical information. Research in automatic lexical acquisition tries to offer a solution to the creation and maintenance of large-coverage lexica for feeding these processing components.

Recent research with supervised Machine Learning methods shows that automatic lexical acquisition can be approached by assigning a word certain properties according to the linguistic information gathered from its occurrences in texts (Brent, 1993; Merlo and Stevenson, 2001; Baldwin and Bond, 2003; Baldwin, 2005; Joanis and Stevenson, 2003; Joanis et al. 2007). Words are represented in terms of a collection of attributes, which are taken as hints or cues for the properties to be assigned. Each attribute records the observation that the word has been found in a particular context or not. For instance, to induce whether a noun can be classified as a countable or as a mass noun, its co-occurrence with particular determiners and quantifiers will be taken as a cue: 'some/*many mud'. A learner is supplied with the information about the occurrences of pre-classified samples in these significant contexts. This linguistically-motivated information, registered as numerical values of vector attributes about matched and non-matched cues, is supposed to be enough to separate new data into the proposed classes. The work by Dorr and Jones (1996) leads us to think that if the linguistic cues are properly identified, the mapping to classes has to follow necessarily.

One of the problems of this approach (Joanis and Stevenson, 2003; Joanis et al. 2007; Korhonen et al. 2008), and the one we will concentrate on in this paper, is the high number of missing values, that is, zero values for most of the attributes that register the occurrence of the token in these particular contexts that are considered cues. In this framework, a zero value could mean either that the cue has not been observed because the word in question does not belong to the class, i.e. negative evidence, or that the word in question has not been observed in the cue sought, i.e. lack of evidence. This uncertainty creates problems to the learner, because zero values for incompatible labelled examples make the cue lose its predictive capacity and even though some samples display the sought context, it is not taken into account.
In this paper we present the results of our experiments to try to reduce this uncertainty by, as other authors do (Joanis et al. 2007, for instance), substituting zero values for pre-processed estimates. Here we present a first round of experiments that have been the basis for the estimates of linguistic information motivated by lexical classes. The results show a clear improvement with respect to other methods.

These experiments have addressed two cases that suffer from severe problems of missing values just to test this approach. The classification of these particular classes might be better solved with other approaches, however. Note, in addition, that there are other important problems in lexical acquisition, such as noise filtering, which are not under the focus of this paper (see for instance Korhonen, 2002).

In what follows, we first introduce the state of the art in linguistic cue-based lexical information acquisition, and we motivate our proposal of using the linguistic knowledge about correlations between cues to re-estimate the missing values. This is the subject of section 2. In section 3, we focus on the problems caused by missing values and we introduce our approach taking into account the characteristics of linguistic data. In section 4, we give details about the probability estimation and the production of what we call 'harmonized vectors'. Section 5 is reserved for giving methodological details of the experiments, whose results are presented in section 6, together with an evaluation. In Section 7, we present our conclusions and future work.

2. The use of cues in lexical information acquisition

According to the linguistic tradition, words that can be inserted in the same contexts can be said to belong to the same class (Harris, 1951). Thus, lexical classes are linguistic generalizations drawn from the characteristics of the contexts where a number of words tend to appear. One of the approaches to lexical information acquisition proposes classifying words by training a classifier with information about their occurrence in selected contexts

where words belonging to a class indeed occur, e.g. the class of transitive verbs will appear in passive constructions, while intransitive verbs will not, as expected. Thus, the whole set of occurrences (tokens) of a word are taken as cues for defining its class membership (the class of the type), either because the word is observed in a number of particular contexts or because it is not.

Different supervised methods of machine learning techniques have been applied to cue-based lexical acquisition. A learner is supplied with classified examples of words represented by numerical information about matched and not matched cues. The final exercise is to confirm that the data characterized by the linguistically-motivated cues indeed support the division into the proposed classes. This was the approach taken by Merlo and Stevenson (2001), who worked with a DT and selected linguistic cues to classify English verbs into three classes: *unaccusative*, *unergative* and *object-drop*. For instance, *animacy of the subject* is a significant cue[1] for the class of object-dropping verbs, in contrast to verbs in *unergative* and *unaccusative* classes. More general linguistic information was used by Joanis et al. (2007): i.e. frequency of filled syntactic positions or slots, tense and voice features, etc. to describe the whole system of English verbal classes.

Cue-based classification of nouns has been less addressed than that of verbs. Some selected references are Light (1996), who used information from derivational affixes to classify nouns; and Baldwin and Bond (2003), who induced mass/count information, from a parsed English corpus, using parallel supervised classifiers that took into account morpho-syntactic cues like head number, modifier number, subject-verb agreement, occurrence in 'N of N' constructions, etc. The accuracy of their system was measured in terms of F-score[2] of 0.89 in the classification of English mass nouns, with a gold standard test set that, however, accepted a double classification, i.e. a noun could be both mass and count. More recently, Bel et al. (2007) trained a Decision Tree (DT) to classify Spanish nouns as mass nouns (among other lexical features, such as subcategorized complements and bounded prepositions) with an accuracy of 0.67, although allowing only one class per word in the gold standard.

The problems caused by sparse data, i.e. the lack of evidence, and therefore abundance of zero values, in cue-based lexical acquisition are addressed by Joanis et al. (2007), who reported that even using medium to high frequency words (for instance, their test set were verbs with more than 100 occurrences in the British National Corpus, BNC), they had to pre-process missing values by substituting zeros for a trimmed mean value of the observed values for this particular cue in the whole test set, as we will see in section 3.

For our experiments, we have addressed two tasks which specially suffer the problem of zero values even with medium to high frequency words: the classification of mass nouns, which we have tested for Spanish nouns; and the classification of concrete/abstract nouns, which we have addressed for English nouns. In the following sections we give details of the cues used to create both datasets and we motivate them.

[1] The context taken as a cue was the use of personal pronouns (not *it*) as subject.

[2] F-score is the harmonized mean value of precision and recall.

2.1 Cues for mass/countable distinction

In several languages, Spanish and English among them, the distinction between mass and non-mass nouns is grounded on morphosyntactic cues of a lexical class that is based on the denotation of the word (Gillon, 1992 for English; and Bosque, 1996 for Spanish)[3].

For our experiments with Spanish nouns, we have used Bel et al. (2007) cues and data. The cues devised for identifying Spanish mass/non-mass nouns are the following:
- Plural morphology: Spanish mass nouns tend to appear in singular more than in plural. We have registered both in different attributes.
- Singular undetermined noun phrases after a verb or a preposition are a clear cue of the head noun being a mass noun: "hay barro en el salón" ('there is mud in the living room") vs. "hay hombres/*hombre en el salón" ("there are men/*man in the living room"). We have used a cue for each possible context.
- The co-occurrence with particular quantifiers, such as "más" ('more'), "menos" ('less'), "poco" ('few'), etc. in singular is also a cue for mass nouns.
- Derivational suffixes of nouns such as "-ción" or "-dad" are considered cues, as well.
The cues based on a collection of lexical items are rarely observed, but their predictive nature is very high, becoming one of the typical examples of frequently zero-valued attribute.

2.2 Cues for abstract nouns in English

Concrete/abstract classification (Baroni et al. 2008) has a more severe problem of sparse data than mass class. The distinction between abstract/concrete nouns is more semantically based than the mass one, and contextual morphosyntactic cues are harder to find. Our final set of cues consists of nine cues for abstract nouns and five meant to identify non-abstract nouns, as follows:
- According to Light (1996), suffixation is the most powerful cue to identify English abstract nouns. The suffixes we have used as cues are: *-ness*, *-tion*, *-ity*, *-ism*, *-dom*, *-ment*, *-tude*, *-ence*. We have also included the *–ing* ending to capture most of the verbal nominalizations, which tend to be abstract, although this cue has introduced some noise (as in 'building', which is a concrete noun).
- For non-abstract noun detection, we have used the following suffixes: *-er, -or, -ist*, trying to capture nouns that refer to persons or instruments: *doctor, teacher, dentist*, *opener,* etc. For both types of suffixes we have also included as separate cues the possibility of being coordinated with a noun that has one of these suffixes, as coordination tends to link nouns belonging to the same class. While in the case of suffixes used to build abstract

[3] Despite their lexical characterization, nouns can change this lexical feature when being in particular syntactic contexts, which makes the lexical acquisition of this feature a very interesting scenario, as the number of cues that are meant to identify the members of the mass class is reduced.

nouns we indeed meet the problem of missing values, i.e. a pretty large number of abstract nouns do not have any of the suffixes that would be a cue; in the case of suffixes used to build concrete nouns, the problem is noise, i.e. there are many abstract nouns that have these endings, for instance 'answer'.

- Concrete nouns tend to co-occur with adjectives that refer to colour and size, for instance *big, small, huge, large, little, long, short, thin, tall, round, medium*. Thus, the co-occurrence with these words has been considered a cue.
- Abstract nouns have been found to co-occur more with particular determiners such as *much, little* or *that*.

3. Missing values

Missing or zero values are due to the optional nature and variety of the contexts of occurrence we are using for identifying lexical classes. For instance, in our previous example 'there is mud in the living room', the word 'mud' can also appear with other determiners that are less informative for our classification, i.e. "the mud". Moreover, as we have seen in the previous section, in order to identify a class it is necessary to use a number of different possible contexts in which a word cannot occur simultaneously. Note that it is impossible for low frequency words to occur in every of these contexts, and a number of the attributes of its vector representation will be necessarily zeros. Note that for 11 out of 17 cues in our English test set, less than 40% of the nouns were observed in these contexts once or more times. Supervised machine learning methods such as DT's can handle missing values by assigning a probability to each of the possible values, which is calculated based on the frequencies of the various values of an attribute in the examples. However, when the training set contains a majority of cases with a missing value, labelled both as positive and negative examples, the cue loses its predictive power. Although a word occurs in one of these informative contexts, it will not be taken into account because the abundance of other cases where it will be zero-valued will make the learner ignore it, being the numerous cases of zeros more salient than the few with values. In Figure 1, we can see the relation between the occurrences of a word (tokens) and the number of different cues observed, at least once, in a corpus of 3.3 Million words in our experiment of abstract-noun classification in English.

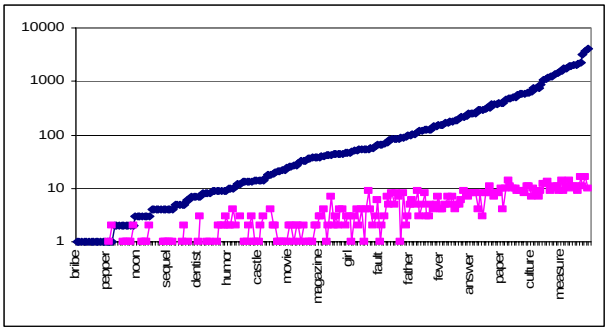

Figure 1: Relation between token frequency (dark line) and the number of different cues observed, at least once, for 213 English nouns in a corpus of 3.3 M words. The scale is logarithmic in order to see the details of the low frequency words. It is clear that the lower the frequency is, the fewer cues tend to be observed.

In other works on lexical information acquisition, the problems created by zero values could have remained hidden by the selection of testing datasets with high frequency words in large corpora. For instance, Merlo and Stevenson (2001) extracted the counts of only 20 verbs per class from a 65-million-word tagged ACL/DCI corpus (Brown and Wall Street Journal 1987-1989), and from the 29-million-word parsed Collins corpus (Collins, 1997), selecting those samples with a minimum of 10 occurrences per verb. But, according to Zipf (1935), there will be a large number of words, especially nouns, that will appear fewer times in any corpus of any length. For instance, Yallop et al. (2005) calculated that in the 100M-word BNC, from a total of 124,120 distinct adjectives, 70,246 occur only once, 106,464 less than ten times and 119,337 less than a hundred times. Similar figures would result from the analysis of noun occurrences. As already said, these low frequency words will not occur in every possible context, and hence the proliferation of missing values will make different cue-based methods for lexical acquisition inoperative. If the learner has been trained only with the more frequent words, during classification the lower frequency ones will not be displaying enough evidence, and hence will be misclassified. If the training material takes them into account, zero values will characterise both positive and negative examples, thus creating an uncertainty that the system will not be able to solve. The uncertainty is due to the fact that either a zero value is indeed a negative value, i.e. the cue is that it has not been observed; or a zero value indicates that the cue was just not observed in the examined corpus, by chance, because of its optional nature.

Missing values have been a topic for Machine Learning methods and, of course, for the application of ML to other NLP tasks. For instance, for Decision Trees, Quinlan (1986) proposed different methods to estimate the unknown values. These included using the modal value, using Bayesian probabilities, determining the unknown value using a DT, and distributing the unknown examples according to the known examples. According to Mingers (1989) experiments, the easiest and best method to deal with zero values is to assume that they can be substituted at a given point for the most common value observed among the other examples in the same class. For NLP tasks, Katz (1987) and Baayen and Sproat (1996), among others, acknowledged the importance of pre-processing in that way low frequency events for Markovian methods, for instance.

For lexical classification, Joanis et al. (2007) experimented with more than 100 occurrences per verb in the BNC, but still they had to substitute missing values for a mean value obtained from the test set, as we will see later. Korhonen (2002) and Korhonen and Krymolowski (2002), in a different learning scenario, smoothed zero values with back-off estimates, i.e. informative prior probabilities calculated on a previously available set of classified verbs (i.e. Wordnet). In all cases the treatment of missing values increased accuracy in the classification task, showing the impact of this phenomenon in language data.

Our experiments try to improve these results by addressing the uncertainty mentioned before with the objective of separating zero values into, on the one hand, negative values and, on the other hand, likely to be unobserved values. As we will show, we are proposing a back-off solution, computing the estimates that are to substitute the zero values from the correlations holding among the different cues that define a class. The estimation is done by applying Bayesian methods that take into account two factors: the fact that there are several cues to identify each lexical class and the fact that there are actual data coming from actual occurrences of the word, although there are few of them. If several cues are meant to describe one lexical class, then it is possible to use the knowledge that there is such a correlation in order to make predictions about the likelihood of the one that is missing, taking as a reference the related cue or cues in actual occurrences, as we explain in the next section.

4. Harmonization

Our aim was to distinguish between negative cues and unobserved cues for improving the material to be supplied to the learner, and our proposal was to make this distinction by using information on the correlations among the cues. According to Anderson (1991), categorization is justified by the observation that objects tend to cluster in terms of their attributes. Thus, if a number of attributes are used to identify a class, to know just a part of the attributes still helps to predict the values of the other, unobserved, attributes.

In the preceding section we have described the different cues that we have used to identify a particular lexical class, i.e. no plural morphology, absence of determiner, etc. The idea behind our approach is that when a cue for identifying class A, for instance, is matched, it is more likely to observe other cues that would also be significant for class A, than to observe the ones that are used for identifying other classes. Our hypothesis is that classification results would improve if we could distinguish real zero values, the negative case, from those cases where the fact of being unobserved is related to low frequency or to the non-obligatory character of the contexts used as cues. Thus, we approach this task by computing, for each of these zero-valued components, the probability of there being a positive value.

To sum up, we preprocess the data gathered from a corpus by smoothing missing values with an estimate of their probability of being a positive cue, based on the information supplied by the concept of lexical classes. We call the results "harmonized vector" because the values are harmonically re-evaluated. The calculation is made as follows.

4.1 Basic vector representation

We have used Regular Expressions to implement patterns that check for the cues in a part-of-speech tagged corpus. A vector representing a word consists of as many components as cues devised to identify members, and also non-members, of the class that our classifier is intended to learn and predict, as detailed in section 2. The positive or negative results of the *n*-pattern checking in all the occurrences of a word are stored as numeric values of an *n*-dimensional vector. Thus, a single vector summarizes all the occurrences of a word (the type) by encoding how many times each cue has been observed or whether it has not been observed (zero value). This vector, which we call Frequency, is the input for the different smoothing and harmonization techniques used in the experiments.

4.2 The probability of being a positive cue

As already said, the frequency vector can contain many zero-valued components. Our hypothesis is that classification results would improve if we could distinguish actual zero values, the negative case, from those cases likely to be related to low frequency or to the non-obligatory character of some of the contexts used as cues. Thus, we have approached the task by computing, for each of these zero-valued components, the probability of there being a positive value. We have used the knowledge we have about the particular cues that jointly identify the classes into which we want to classify words, and the frequency vector obtained as a representation of their occurrences in a corpus, as we have explained in section 4.1.

Following Anderson (1991), we want to obtain (1), the probability of a component's value given all the components of the vector where it appears.

$$(1)\quad P_i(j \mid v) = \sum_k P(k \mid v) P_i(j \mid k)$$

Where *v* is the vector representing the occurrences of the word, *k* is the class that is being used for classification and *j* is the value of the component number *i* of the vector *v*. (1) wants to make a prediction about unobserved attributes. In our case, we want to make a prediction about the probability of *j* being positive, given the actual vector *v* registering the matched cues from all the occurrences of a word. The summation is across all the possible values of class *k*, because we want to find the probability of *j=1* taking into account both values for *k*, *yes* and *no*.

We start with $P(k \mid v)$, that is, the probability of the class given the vector matched cues, because it involves the calculation of $P_i(j \mid k)$, as we will see below. The probability that a given vector belongs or not to a class is approximated as the probability of a class given a particular vector, $P(k \mid v)$, and can be calculated by Bayesian inversion as (2):

$$(2)\quad P(k \mid v) = \frac{P(k)P(v \mid k)}{\sum_k P(k)P(v \mid k)}$$

Where *P(k)* is the prior probability of class *k*, and *P(v|k)* is the likelihood of vector *v* given class *k*. The prior is estimated for the current experiments with a value 0.5 for each class, in order to prevent unbalanced classes from affecting the results of the harmonization experiment. We will see in the conclusions that it is a future task to use a more realistic prior.

To calculate the likelihood of the vector given the class *P(v|k)*, we consider the likelihood of the components, and again following Anderson (1991) we use (3):

(3) $P(v \mid k) = \prod_i P_i(j \mid k)$

The likelihood of each vector component given a class should had been computed from the training sample, but our hypothesis is that because our data is Zipfian-distributed and there are many missing values for *j*, the training sample cannot be used. The same reasons we have argued for preprocessing the data to be supplied to the learner apply here, as well: too many missing values will supply the assessment with uncertain information[4]. In order to solve that point we have probabilistically modeled the linguistic information contained in the definition of lexical classes. For instance, for the case of the abstract class that we have defined in section 2, we have used five cues, out of seventeen, to identify concrete nouns. It follows that the probability of $P_i(j|k)$ is 1 for each of these cues: as linguists, we expect that if a word belongs to the class *concrete*, it will appear with each of these cues. However, as we know that none of them is obligatory and that a word can have low frequency, we have heuristically tuned its likelihood to a probability of 0.5, because the cue can be seen or not.

	abstract	Concrete
Suffix=0	0.5	1.0
Suffix=1	0.5	0.0
SC_Adj=0	1.0	0.5
SC_Adj=1	0.0	0.5

Table 1. A sample of the likelihood of particular cues *j* given a class *k*, *P(j|k)*, for abstract/concrete.

In Table 1, we show a sample of the probabilities of the two types of cues for the abstract case: those meant to identify the abstract class (for instance, suffixes such as *–ness*, *-tion*, etc.) and those meant to identify the concrete class (for instance, size or colour adjectives). Table 1 reads as follows: the probability of occurring with a particular suffix (Suffix=1) given an *abstract* noun is 0.5, but if the word is *concrete* is 0. We cannot be certain about the positive case, but we are so about the negative case. This is a very simplistic approach to model the information about the lexical class. In Table 2, we show the probabilities estimated for the more complex *mass* class in Spanish, where we have been able to take into account more cues and, crucially, also the fact that a given cue can identify two classes in a three-class lexical model[5].

[4] We have compared $P_i(j|k)$ assessed with linguistic information, as just explained, with that obtained from the training data. As expected, the strongest differences are mainly in the assessment of the cues that are less frequent, thus confirming our initial suspicions.

5 The lexical model addressed mass, count and mass-count names. Merlo and Stevenson (2001) and Joanis et al. (2007) had identified an important aspect about the distribution of cues for lexical classes: there are classes that share cues.

	mass=no	**mass=yes**
Plu=no	0.50	0.54
Plu=yes	0.50	0.46
Undet=no	1.00	0.50
Undet=yes	0.00	0.50
Quant=no	1.00	0.50
Quant=yes	0.00	0.50
Morfo=no	0.63	0.66
Morfo=yes	0.37	0.34

Table 2. A sample of the likelihood of particular cues *j* given a class *k*, *P(j|k)*, for the mass class.

Note that the likelihood of the attributes obtained from the knowledge of which cues are under the same class, $P_i(j|k)$, is used in (3) and also in (1). Thus, the posterior probability of the class given the vector, and the likelihood of the cue given the same class are multiplied to take all the available information into account.

In practice, the results of harmonization can be summed up with the following explanation. When the initial frequency vector has one or more positive cues from those defined to identify a particular class, then all the components of cues related to the same class get some probability of being positive, i.e. the case of *acero* or *desabastecimiento* in Table 3 below. The cues related to the other class remain zero-valued. When the frequency vector has positive values for two or more cues, each one related to different classes (because of noise, for instance), then the value obtained by the initially zero-valued attributes after harmonization is kept as zero, as in the case of *agua*.

Note that frequency information is lost, however, as now all the vector components encode probabilities.

Harmonized	Frequency	Word
0,1,0,1,0,1,1,0,1, 0,0,1,1,0	0,3,0,1,0,1,1,0,1, 0,0,1,1,0	agua ('water')
1,1,0.5,0.5,0.5,1,1,1,1, 0,0,0,0,0	1,2,0,0,0,2,1,1,2, 0,0,0,0,0	acero ('steel')
0.5,0.5,0.5,0.5,0.5,0.5,1, 0.5,0.5,0,0,0,0,0	0,0,0,0,0,0,1,0,0, 0,0,0,0,0	desabastecimiento ('shortage')

Table 3: Comparative view of harmonized and frequency-based vectors for different occurrence patterns of positive mass nouns in Spanish. The last five components are cues for non-mass class

We have kept this simple scenario in order to clearly see the impact of harmonization. However, it is a future task to assess priors and conditional probabilities differently to obtain better predictions.

5. Experiments and Methodology

For our classification experiments we have used a C4.5 Decision Tree classifier (Quinlan, 1993), a Support Vector Machine (LibSVMs, Chang and Lin, 2001, with a Radial basis function as kernel), which has recently shown its adequacy to the task (Joanis et al. 2007). We have used the implementation of the Weka platform (Witten and Frank, 2005). For each experiment we have used a 10-fold cross-validation testing.

The baseline has been a simple majority-class classifier, as computed from an actual MT dictionary of 35.000 lemmas for the Spanish case, and from the gold-standard files for the English case, which had to be 50% but some nouns have been deleted because they did not occur in our corpus.

Our experiment had to confirm that the information supplied by the harmonized vectors improved the classification results, even when the frequencies of the words were low. Thus, the corpus used to gather data has been considerably small, compared to other works. We have used the IULA-UPF Multilingual Technical Corpus (Cabré et al., 2006), a part-of-speech tagged corpus which consists of domain-specific texts. The section used in our evaluation contains 1 million words for Spanish in the domain of economy; and 3.3 million words from different domains for English.

The Spanish dataset consists of the vectors for 250 different nouns, as used by Bel et al. (2007), where there are 102 nouns with the value *mass=yes*. The evaluation has been done by comparing with Bel et al. (2007) gold-standard files, which were manually encoded. The number of occurrences for each type varies, as shown in Table 4. Each word has been represented by a vector of 23 attributes encoding the positive and negative cues reported in section 2, gathered from the corpus just mentioned. The data contains noise and a high number of missing values, which reproduces a scenario that suffers from the sparse data problem: for the five best attributes ranked by a chi-squared filter, there are 89, 115, 139, 156 and 209 vectors, respectively, that have zero as value.

#of occurrences (tokens)	1	2-10	11-50	>50
Spanish types	44	101	58	47
English types	15	55	55	62

Table 4. Token frequency per type in the used corpora.

The English dataset for the task of concrete-abstract classification has initially consisted of 250 nouns selected by Altarriba et al. (1999), as we have used their ratings of concreteness as our Gold standard. However, there were 63 nouns that did not occur in our corpus and we have deleted them. The final test set has been of 187 types. Frequency information about them in our corpus is shown in Table 4, where we can see that the majority of nouns occur less than 50 times. The English dataset has suffered, more than the Spanish one, from the sparse data problem. For the first 5 best chi-squared ranked attributes there are 107, 111, 129, 185, 169 samples, respectively, that are zero-valued.

For the different experiments, the test sets we have used are:

- Zero-valued components, substituted by a simple mean based smoothing, which has been calculated on the members of each class in the dataset.
- Zero-valued components, substituted by a trimmed mean value, which has been obtained from the dataset and calculated by discarding 30% of the lowest and the highest scores.
- Frequency of each cue.
- The harmonized vectors.

6. Results and discussion

Our harmonization method wanted to discriminate real negative cues (i.e. the significant fact is that the cue has not been observed) from lack of evidence, as this uncertainty negatively affects classification already when learning from the data. For the learner, zero values for incompatible labelled examples make the cue lose its predictive or discriminative capacity, and then, although a sample shows the sought cue correctly, the learner will not take it into account. Thus, the objective of our experiment was to test to what extent the qualified information supplied by the harmonized vector contributes to the accuracy of the classification, independently of the datasets and of the methods used.

Results are presented in terms of accuracy, i.e. the percentage of correct classifications out of all the classifications, in Table 5. They confirm that harmonized vectors improve learning results, if compared with the results obtained by using smoothing methods such as mean and trimmed mean values computed from the available datasets, as proposed by Mingers (1989) and Joanis et al. (2007). Our method also slightly improves the results of just using the frequency of the cues in our test set, but differences in the Spanish set are not significant. As said, frequency information is meant to handle noise.

	Spanish Mass		**English Abstract**	
Experiment	**DT**	**SVM**	**DT**	**SVM**
Mean	74.2	63.8	57.8	61.0
Trimmed mead	77.5	67.4	55.6	61.0
Frequency	79.9	79.1	61.4	64.1
Harmonized	82.8	80.7	76.1	70.1
Baseline	74.8		61.5	

Table 5: Results of the experiments

DT and SVM results are similar, and there are some improvements of the classification results in both datasets and for the two methods. The best accuracy is achieved with the Spanish dataset working with DT's, which is in contradiction with Joanis et al. (2007) experimental results, where they reported to have gained by using SVM's. Some reasons for this could be the fact that their vectors have a higher dimensionality (224 attributes) than ours (23 attributes), and the size of the corpus we have already commented in section 5. However, and although a proper comparison of results is impossible due to the differences between both datasets, our results seem to be in line with their results: 77.5% of accuracy for English verbs when working with DT's. Our results show an improvement also compared to Bel et al. (2007), which report 67% of accuracy for the same tasks, using a vector-of-vectors representation for classifying Spanish mass nouns with a DT. The improvements when using an informed estimate support Korhonen (2002) results, when she used informed priors computed from WordNet. But the availability of previously classified resources can be a problem, while our proposal only makes use of the

knowledge of which cues are used together to identify every class.

An analysis of the errors shows that the classifiers using harmonized vectors cannot handle noise, if compared with vectors using frequency information. Very frequent words tend to occur in most of the proposed contexts, although in different proportions.

The improvement in the Spanish set is not statistically significant, however. Error analysis shows that it has to do with the loss of frequency information, which in the other experiments helps in noise filtering, but not in the harmonized vector. Our class-based probability information increases the chances of a zero value being a positive, but cannot correctly handle the actually observed data where noise exists. It remains for future work to introduce noise filtering.

Another remarkable aspect has to do with the size of the training corpus, which for the Spanish case is rather small. However, in comparison with the results of the English experiment, with a 3-times larger corpus, the size does not seem to be an important factor, as the bad results with frequency vectors and DT show for the English case. As expected, the phenomenon of missing values, the problem that really affects the DT's, is to a certain extent independent of the size of the corpus because of the Zipfian distribution of words.

7 Conclusions

The results of our experiments support our proposal to use harmonized vectors to overcome one of the problems caused by sparse data in lexical acquisition: abundance of uncertain zero values. By just using the knowledge of the correlations that hold among different cues used to identify a class, we can estimate the likelihood of each zero-valued component being positive. Our method opens the possibility of working with cues, which have a strong predictive power but very low occurrence. It is also a topic for future research to combine harmonization with frequency information in a smart way.

8. Acknowledgements

I want to thank Joan Banach and Liana Egiazarian for their valuable contribution with the English dataset.

9. References

Altarriba, J., Bauer, L. M., & Benvenuto, C. 1999. Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers, 31*.

Anderson, J.R. 1991. 'The adaptive nature of human categorization'. *Psychological Review*, 98(3).

Baayen, H. and Sproat, R. 1996. Estimating lexical priors for low-frequency morphologically ambiguous forms. Comput. Linguist. 22, 2.

Baldwin, T. 2005. 'Bootstrapping Deep Lexical Resources: Resources for Courses', ACL-SIGLEX 2005. Workshop on Deep Lexical Acquisition.

Baldwin, T. and F. Bond. 2003. "Learning the Countability of English Nouns from Corpus Data". Proceedings of the 41st. ACL Annual Meeting.

Baldwin, T.; Bender, E.M; Flickinger, D. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference LREC.*

Baroni, M.; Evert, S. and Lenci, A. (eds.) 2008. ESSLLI Workshop on Distributional Lexical Semantics.

Bel, N.; Espeja, S.; Marimon, M. 2007. Automatic Acquisition of Grammatical Types for Nouns. In H LT 2007: The Conference of the NAACL. Companion Volume, Short Papers.

Bosque, Ignacio (ed.). 1996. *El sustantivo sin determinación. La ausencia de determinante en la lengua española*, Madrid: Visor.

Brent, M. R. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics* 19.

Briscoe, E. & Carroll, J. 1993. Generalised probabilistic LR parsing for unification-based grammars. *Computational Linguistics* 19.1.

Cabré, M. T.; Bach, C.; Vivaldi, J. 2006. *10 anys del Corpus de l'IULA*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra

Collins, M. J. 1997. Three generative, lexicalised models for statistical parsing. In Proceedings of ACL-97.

Dorr, B. and Jones, D. 1996. Use of Syntactic and Semantic Filters for Lexical Acquisition: Using WordNet to Increase Precision. In Proceedings of the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases. Toulouse, France.

Chang; C-C and Lin, C-J. 2001. LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Gillon, B. 1992."Towards a common semantics for English count and mass nouns," Linguistics and Philosophy 15.

Harris, Z. 1951. *Structural Linguistics*. Chicago University Press.

Joanis, E. and Stevenson, S. 2003. A general feature space for automatic verb classification. Proceedings of the 10 Conference of the European Chapter of the Association for Computational Linguistics (EACL-03).

Joanis, E; Stevenson, S; and James, D. 2007. A General Feature Space for Automatic Verb Classification. *Natural Language Engineering*.

Katz., S. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing, March, 35.

Korhonen, A. 2002. Subcategorization acquisition. As Technical Report UCAM-CL-TR-530, University of Cambridge, UK.

Korhonen, A. and Krymolowski, Y. 2002. On the Robustness of Entropy-Based Similarity Measures in Evaluation of Subcategorization Acquisition Systems. In Proceedings of the Sixth Conference on Natural Language Learning.

Korhonen, A., Krymolowski, Y. and Collier, N. 2008. The Choice of Features for Classification of Verbs in Biomedical Texts. *Proceedings of Coling 2008*.

Light, M. 1996. Morphological cues for lexical semantics.

Proceedings of the 34th ACL.
Merlo P. and Stevenson S. 2001. Automatic Verb Classification based on Statistical Distribution of Argument Structure, *Computational Linguistics*, 27:3.
Mingers, J. 1989. An empirical comparison of pruning methods for decision tree induction. *Machine Learning, 4* (2), 227–243.
Quinlan, J.R. 1986. Induction of decision trees, *Machine Learning, 1*, 81–106.
Quinlan, R.J. 1993. C4.5: Programs for Machine Learning. Series in Machine Learning. Morgan Kaufman, San Mateo, CA.
Schülte im Walde, S. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*.
Yallop, J., Korhonen, A. and Briscoe, E.J. 2005. 'Automatic acquisition of adjectival subcategorisation from corpora', Proceedings of the 43rd Assoc. for Comp. Ling., Morgan Kaufmann.
Zipf, G. K., 1935. *The Psycho-Biology of Language*, Houghton Mifflin, Boston.
Witten, I. H. and Frank E. 2005. Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.