The Structure of the Lexicon in the Task of Automatic Lexical Acquisition

Núria Bel
Sergio Espeja
Montserrat Marimon
Universitat Pompeu Fabra

In the task of automatic lexical acquisition, i.e. the induction of lexical information from texts, there have been no attempts to exploit theoretically-based models of the structure of the lexicon. Works like those of Bybee (1988) and Langacker (1987) propose a highly structured lexicon where words are related paradigmatically by phonological similarity and where lexical features are an emergent characteristic of the resulting structure. If so, a machine learning algorithm such as a Decision Tree (DT, Quinlan, 1945) should be able to learn the correlation between particular lexical features and the formal characteristics of words. In our experiment, the machine learner should be able to find a correlation between characters that form the words used for training it and the nominal feature /mass/. The ability of the trained learner to predict correctly whether nouns that it has not been shown in the training phase are mass nouns or not is proof that such a correlation exists and that it can be considered an emergent feature of the paradigmatic relations that relate words in the lexicon. The obtained results prove that a structured lexicon can provide information on lexical features.

1. Introduction

In the task of automatic lexical acquisition, that is the induction of lexical information from texts, there have been no attempts to exploit theoretically based models of the structure of the lexicon. Works like those of Bybee (1988) and Langacker (1987) propose a highly structured lexicon where words are related paradigmatically by phonological similarity and where lexical features are emergent characteristics of the resulting structure. If so, a machine learning algorithm such as a Decision Tree (DT, Quinlan, 1945) should be able to learn the correlation that these authors claim to exist between particular lexical features and the sounds of words. What we present here is the experimentation done in order to test this hypothesis.

For our experiment, the chosen feature is the feature *mass* for nouns. We have trained a DT with a set of correctly classified examples of nouns (i.e. *mass=yes* for nouns such as *aluminium*, and *mass=no* for *table*) to check whether only the formal similarity of words helps to predict this feature, as stated in the works mentioned before. The machine learner should be able to find a correlation between characters that form the words used for training it and the feature *mass*. The ability of the trained learner to predict correctly whether nouns that it has not been shown in the training phase are mass nouns or not will be the proof that such a correlation exists and that it can be considered an emergent feature of the paradigmatic relations that relate words in the lexicon.

The results of our experiments show that the feature *mass* can be derived from a lexicon and used to classify automatically words with an accuracy that ranges from 75% for English nouns to 86% for Spanish nouns. The differences in the figures are mainly due to the number of lexical entries used for training: the English sample was of 7,576 nouns, while the Spanish one was of about 23,000. For both languages, and compared with a frequency based baseline of 65% and 77.7% respectively[1], our results are an evidence that a structured lexicon can provide information on lexical features.

[1] The frequency based baseline means that a learner using it will classify as mass=no all the nouns of the test set, and a 65% and a 77,7% will be right, for English and Spanish respectively, because most of the nouns are mass=no. Note that such a classifier will be unable to identify any mass noun.

2. Mass-count distinctions and lexical acquisition

Lexical semantics has addressed the distinctions between mass and countable nouns at length. In several languages, Spanish and English among them, the distinction between mass and count nouns is grounded on morphosyntactic criteria (basically, the possibility to take plural form and the selection of specific determiners, among others) as a reflection of a lexical characterization based in the denotation of the word. However, despite of the original characterization, nouns can change this lexical feature when being in particular syntactic contexts. The works by Leech (1981) and his lexical rules, Ostler and Atkins (1991) with their lexical inference rules, and Pustejovsky (1995) with the concept of type coertion, have tried to formalize the systematic relations that hold between mass and countable readings of nouns and to describe the contexts that trigger the changes of a feature initially lexical. Gillon (1992) offers a summary of cases of conversion for English. A mass noun converts into a count noun in order to denote units of it, or kinds of what it denotes. Usually, conversion is made by pluralizing the noun: "two coffees". Less evident examples are those nouns denoting emotions and mental states which can (at least some of them) also become countable by pluralizing them: "my sister has two *anxieties*: her health and her children". Finally, another group of nouns can denote, when in plural, instances of the denotation of the mass noun, etc.: "All the efforts will be needed" vs. "Much effort". Conversion from count nouns to mass nouns is also possible, provided that certain conditions exist: a clear context for it, and a possible denotation of the "largest aggregate" (Ostler and Atkins, 1991). This last condition is specially productive with animals and plants for human consumption: "There is lamb for dinner today", but also "This is a house made of cedar". The most typical context for triggering this mass reading is an undetermined noun phrase. Compare the examples above with: "There is a lamb for dinner today" and "This is a house made of one cedar".

In the task of lexical acquisition, the mass feature has been until now induced by taking into account the syntactic contexts where the word occurs, which, as we have seen, can influence the reading. Baldwin and Bond (2003) is a well known attempt to induce mass/count information from parsed corpus, using parallel supervised classifiers that take into account different syntactic cues: head number, modifier number, subject-verb agreement, the occurrence in N of N constructions, etc. The accuracy of their system was of 94.6% in classifying English nouns into four classes related to countability, for a gold-standard test set that, however, accepted a double classification, i.e. a noun could be both mass and count. Besides, and more crucially, their method was said to obtain the preferred countability class, because the value was got by assessing the relative token occurrence. That is, the preferred class will be the one more frequent for each word. Bel et al. (2007) also dealt with the mass feature for Spanish nouns with an accuracy of 67.0% for the mass feature, and only using a part of speech tagged corpus and local contexts. With their method, also the most frequent value is computed as the preferred one.

Our current experiment has tried to assess whether the lexical value for the mass feature can be induced from a lexica before the action of the context where the noun is inserted. By knowing the base reading of a word, we could apply contextual rules to obtain the derived readings which are found in texts.

3. Experimental setup, methodology and data

In order to account for the contribution of information coming from a structured view of the lexicon, we have carried an experiment with two different lexica. In the experiments we used existing English and Spanish dictionaries of the HPSG-based grammars developed in the LKB system (Copestake 2002) mostly for parsing purposes: the English Resource Grammar (Copestake and Flickinger 2000) and the Spanish Resource Grammar (Marimon et al. 2007). These lexica have been taken as the gold-standard test set.

For the experiment, every noun is represented as a fixed-length vector based on the orthographical characters as components, ordered in the inverse direction and preprocessed as to

group some of the most productive suffixes.[2] In (1) we see vectors for the English words *alarm*, *aluminium*, *ambition*, *annex* and *antenna*.

(1) Examples of vectors used for building the DT classifier.

```
m,r,a,l,a,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,MS
m,u,n,i,m,u,l,a,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,MS
tion,i,b,m,a,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,MS
x,e,n,n,a,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,CNT
```

We used the lexical data for building specific C4.5 DT classifiers.3 DT's are one well known and successful technique for classification, although delivering better results when there is a very large set of pre-annotated data available. DT's have the additional benefit that the results can be inspected. Figure 1 is a partial view of the DT resulting from the English experiment where the importance of the derivational suffixes is evident. In Figure 2, we show a partial view of the DT for Spanish.

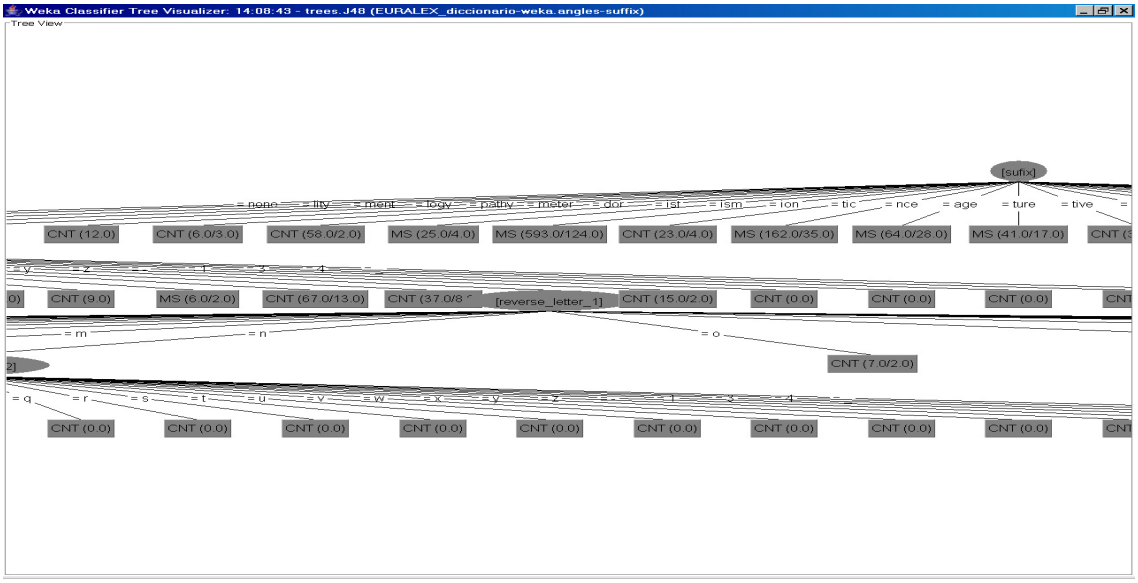

Figure 1. Partial view of the DT classifier results for English nouns.

[2] Suffixes used for English were: *-lity*, *-ment*, *-logy*, *-pathy*, *-meter*, *-dor*, *-ist*, *-ism*, *-tion*, *-tic*, *-nce*, *-age*, *-ture*, *-tive*.

[3] We have used WEKA J48 decision tree classifier (Witten and Frank 2005).

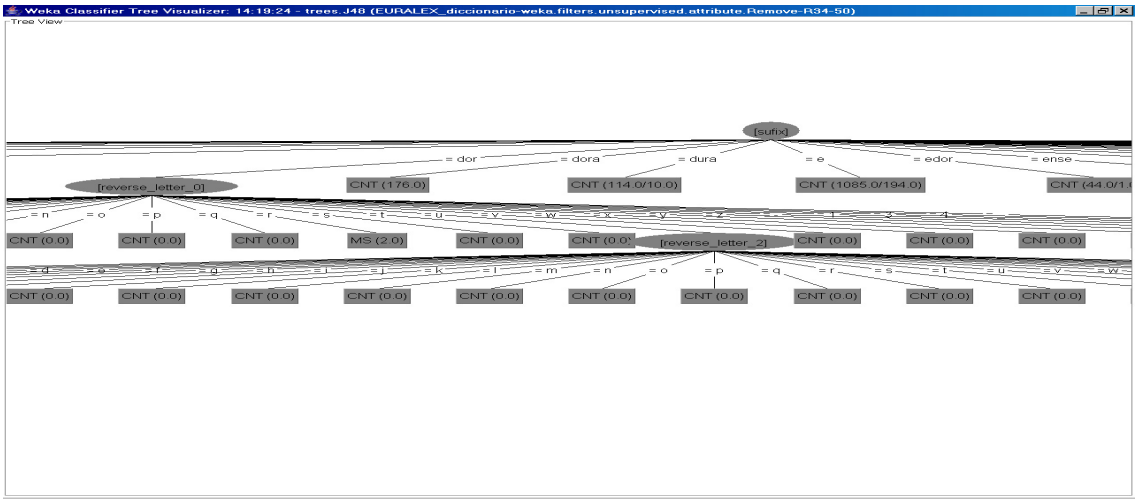

Figure 2. Partial view of the DT classifier for Spanish nouns.

The usual accuracy measures as *type precision* (percentage of feature values correctly assigned to all values assigned) and *type recall* (percentage of correct feature values as found in the gold-standard) have been used. F1 is the usual score combining precision and recall. Note that ambiguity has not been treated at all, being the evaluation against a unique correct type which is considered the primary type because, as said before, we are not taking context into account. Thus "translation", for instance, is encoded as mass=yes, despite of the fact that in certain contexts it takes a countable reading.

The results for our experiment with 7,576 English nouns to be classified as mass=yes or mass=no achieved a global accuracy of 75.6%. For Spanish, we could use a gold-standard file made of 23,507 nouns, and the DT classifier achieved an accuracy of 86.2%. As expected, with a significant larger noun set, the DT achieved better results. These figures have to be compared with the baseline that was of 65% for English and 77.7% for Spanish (see footnote 1). Thus, for both languages significant results were achieved.

The statistical significance of the results was checked by using the Kappa coefficient, which was of 0.40 for English and 0.51 for Spanish. Kappa coefficient indicates that the agreement between the learner and the human encoded gold-standard is more than the expected agreement only by chance, and it is a measure of the statistical significance of the results. The values ranging from 0.4 to 0.7 are considered acceptable. In Table 1, we supply more details on the evaluation of the results.

LEARNER TASK		Precision	Recall	F1
English Mass=	no	0.75	0.92	0.83
English Mass=	yes	0.75	0.44	0.56
Spanish Mass=	no	0.86	0.97	0.91
Spanish Mass=	yes	0.85	0.46	0.60

Table 1. Results of the experiment

Also interesting to the hypothesis of the structured lexicon was to check the significance of the contribution of suffixes compared to stems. We produced a version of the experiment training the learner only with stems and another version only with suffixes. Results are in Table 2 in terms of *accuracy* (percentage of correct feature values as found in the gold-standard), where we see that suffixes bring the largest part of the information, at least, concerning the feature studied.

	Accuracy	Kappa coefficient
Stem + suffix	86.03 %	0.53
Only stem	79.19 %	0.1
Only suffix	84.61 %	0.46

Table 2. Showing the influence of suffix information for the Spanish test set

Note that kappa coefficient for the only stem version of the experiment decreases dramatically, indicating that the results are very close to those that could be achieved by chance. As can be seen in Fig. 1 and 2, the importance of suffixes is evident also in the accuracy results. Although, for instance, the case of suffix *-dor* in Fig. 2 is not conclusive and the DT looks for further characters in the word to take a decision. This is what the results of Table 2 seem to reflect. The results signal the relevant nature of suffixes, but nevertheless counting with all the information of the word delivers better results.

4. Conclusions

The results of our experiments show that information about mass feature for nouns can be induced from the lexicon as an emergent feature of the structure created by formal, in our case orthographical similarities. Hence, our results support Bybee (1988) and Langacker (1987) hypothesis on the structure of the lexicon. However, for inducing features by means of a DT, it is necessary to count with a large number of entries. The comparison between our experiments with English and Spanish shows that results improve when a large lexica is available.

Our experiment contributes to the area of lexical acquisition where there have been other attempts to automatically learn lexical information like mass/count distinction, being the best known the works by Bond and Vatikiotis-Bateson (2002), O'Hara et al. (2003), Baldwin and Bond (2003), and Bel et al. (2007). Direct comparison of the results cannot be done because of the differences in the quantity of information used in the experiments, preprocessing of the materials, strategies, etc., but we expect in a near future to find out whether the combination of the predictions made by our classifier with that of classifiers using syntactic information can deliver significant hints about the mass/count distinction.

References

Baldwin, T.; Bond, F. (2003). Learning the countability of English nouns from corpus data. In *Proceedings of the 41st Annual Meeting of the ACL*. Sapporo, Japan. 463-470.

Bel, N.; Espeja, S.; Marimon, M. (2007). "Automatic Acquisition of Grammatical Types for Nouns". In *Human Language Technologies 2007: Proceedings of the North American Chapter of the ACL Companion Volume, Short Papers*. Rochester. 5-8.

Bold, F.; Vatikiotis-Bateson, C. (2002). "Using an ontology to determine English Countability". In *Proceedings of the 19th Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan. 99-105.

Bybee, J. (1988). "Morphology as Lexical Organization". In Hammond, M; Noonan, M. (eds.). *Theoretical Morphology. Approaches in Modern Linguistics*. San Diego: Academic Press. 119-141.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications.

Copestake, A.; Flickinger, D. (2000). "An open-source grammar development environment and broad-coverage English grammar using HPSG". In *Proceedings of the Second conference on Language Resources and Evaluation.* Athens. 1124-1132.

Gillon, B. (1992). "Towards a common semantics for English count and mass nouns". *Linguistics and Philosophy* 15. 597-639.

Langacker, R. (1987). *Foundations of Cognitive Grammar*. *Theoretical Prerequisites*. Vol. I. Stanford: Stanford University Press.

Leech, G. N. (1981). *Semantics*. Harmondsworth: Penguin Books.

Marimon, M.; Bel, N.; Espeja, S.; Seghezzi, N. (2007). "The Spanish Resource Grammar: Pre-processing Strategy and Lexical Acquisition". In Baldwin, T. et al. (eds.). *Proceedings of the ACL2007 Workshop on Deep Linguistic Processing*. Prague Czech Republic. 105-111

Ostler, N.; Atkins, B. T. S. (1991). "Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules". In Pustejovsky, J.; Bergler, S. (eds.). *Lexical Semantics and Knowledge Representation.* Berlin: Springer. 87-100.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: The MIT Press.