Automatic acquisition of grammatical types for nouns

Núria Bel Sergio Espeja Montserrat Marimon

IULA
Universitat Pompeu Fabra
P. de la Mercè, 10-12
ES-08002 – Barcelona
{nuria.bel,sergio.espeja,montserrat.marimon}@upf.edu

Abstract

The work[1] we present here is concerned with the acquisition of deep grammatical information for nouns in Spanish. The aim is to build a learner that can handle noise, but, more interestingly, that is able to overcome the problem of sparse data, especially important in the case of nouns. We have based our work on two main points. Firstly, we have used distributional evidences as features. Secondly, we made the learner deal with all occurrences of a word as a single complex unit. The obtained results show that grammatical features of nouns is a level of generalization that can be successfully approached with a Decision Tree learner.

1 Introduction

Our work aims to the acquisition of deep grammatical information for nouns, because having information such as countability and complementation is necessary for different applications, especially for deep analysis grammars, but also for question answering, topic detection and tracking, etc.

Most successful systems of deep lexical acquisition are based on the idea that distributional features (i.e. the contexts where words occur) are associated to concrete lexical types. The difficulties are, on the one hand, that some filtering must be applied to get rid of noise, that is, contexts wrongly assessed as cues of a given type and, on the other hand, that for a pretty large number of words, their occurrences in a corpus of any length are very few, making statistical treatment very difficult.

The phenomenon of noise is related to the fact that one particular context can be a cue of different lexical types. The problem of sparse data is predicted by the Zipfian distribution of words in texts: there is a large number of words likely to occur a very reduced number of times in any corpus. Both of these typical problems are maximized in the case of nouns.

The aim of the work we present here is to build a learner that can handle noise, but, more interestingly, that is able to overcome the problem of sparse data. The learner must predict the correct type both when there is a large number of occurrences as well as when there are only few occurrences, by learning on features that maximize generalization capacities of the learner while controlling overfitting phenomena.

We have based our work on two main points. Firstly, we have used morphosyntactic information as features. Secondly, we made the learner deal with all occurrences of a word as a complex unit. In our system, linguistic cues of every occurrence are collected in the *signature* of the word (more technically a pair *lema* + *part of speech*) in a particular corpus. In the next sections we give further details about the features used, as well as about the use of signatures.

The rest of the paper is as follows. Section 2 presents an overview of the state of the art in deep lexical acquisition. In section 3, we introduce details about our selection of linguistically motivated

[1] This research was supported by the Spanish Ministerio de Educación y Ciencia: project AAILE, HUM2004-05111-C02-01/FILO, Ramón y Cajal, Juan de la Cierva Programs and PTA-CTE/1370/2003 with *Fondo Social Europeo*,.

Proceedings of NAACL HLT 2007, Companion Volume, pages 5–8,
Rochester, NY, April 2007.

cues to be used as features for training a Decision Tree (DT). Section 4 shortly introduces the methodology and data used in the experiments whose results are presented in section 5. And in section 6 we conclude by comparing with the published results for similar tasks and we sketch future research.

2 State of the art

Most of the work on deep lexical information acquisition has been devoted to verbs. The existing acquisition systems learn very specialized linguistic information such as verb subcategorization frame[2]. The results for verb subcategorization are mostly around the 0.8 of precision. Briscoe & Carroll (1997) reported a type precision of 0,76 and a type recall of 0.43. Their results were improved by the work of Korhonen (2002) with a type precision of 0.87 and a recall of 0.68 using external resources to filter noise. Shulte im Walde (2002) reports a precision of 0.65 and a recall of 0.58. Chesley & Salmon-Alt (2006) report a precision of 0.86 and a recall of 0.54 for verb subcategorization acquisition for French.

Lexical acquisition for nouns has been concerned mainly with ontological classes and has mainly worked on measuring semantic similarity on the basis of occurrence contexts. As for grammatical information, the work of Baldwin and Bond (2003) in acquisition of countability features for English nouns also tackles the very important problem of feature selection. Other work like Carroll and Fang's (2004) and Baldwin's (2005) have focused on grammatical information acquisition for HPSG based computational grammars. The latter is the most similar exercises to our work. Baldwin (2005) reports his better results in terms of type accuracy has been obtained by using syntactic information in a chunked and parsed corpus. The type F-scores for the different tested categories for English were: for verbs 0.47, for nouns 0.6 and for adjectives 0.832.

3 Feature selection

One of the most important tasks in developing machine learning applications is the selection of the features that leads to the smallest classification error. For our system, we have looked at distributional motivated features that can help in discriminating the different types that we ultimately use to classify words.

The lexical types used in deep analysis grammars are linguistic generalizations drawn from the distributional characteristics of particular sets of words. For the research we present here, we have taken the lexicon of a HPSG-based grammars developed in the LKB platform (Copestake, 2002) for Spanish, similarly to the work of Baldwin (2005). In the LKB grammatical framework, lexical types are defined as a combination of features. Lexical typology of nouns for Spanish, for instance, can be seen as a cross-classification of noun countability vs. mass distinctions, and subcategorization frame or valence, including prepositional selection. For example nouns as "temor" ('fear') and "adicción" ('adiction) belong to the type `n_ppde_pcomp_a_count` as they take two complements: one with *de* and the other with a bound preposition *a*, as in "El temor de la niña a los fantasmas" ('The girl's fear to ghosts') vs. "La adicción a la cocaína" ('The addiction to cocaine').

We decided to carry out the classification for each of the grammatical features that conform the cross-classified types as a better level of generalization than the type: *mass* and *countable*, on the one hand and, on the other hand, for subcategorization information three further basic features: *trans*, for nouns with thematic complements introduced by the preposition *de*, *intrans*, when the noun can appear with no complements and *pcomp* for nouns having complements introduced by a bound preposition. The complete type can be recomposed with the assigned features. "Temor" and "adicción" will be examples of *trans* and *pcomp_a*. They both have also to be assigned the feature *countable*. The combination of features assigned corresponds to the final type which is a definition of the complete behaviour of the noun with respect, for instance, optional complements.

We have used 23 linguistic cues, that is, the patterns of contexts that can be indicative of a particular feature. The most frequent cue that can be related to *countable* is for the noun to be found with plural morphology. A singular noun without determiner after a verb or a preposition is a cue of the noun being *mass*: "hay barro en el salón" ('there is mud in the living room") vs. "hay hombres en el

[2] Given the argument-adjunct distinction, subcategorization concerns the specification for a predicate of the number and type of arguments which it requires for well-formedness.

salón" ("there are men in the living room"). A further cue for *mass* is the presence of particular quantifiers, such as "más" ('more'), "menos" ('less'), etc. But these cues, based on a collection of lexical items, are less productive than other characteristics such as morphological number or presence of determiners, as they appear very scarcely in texts. Nevertheless, we should mention that most of mass nouns in Spanish can also appear in the contexts of countables, as in the case of "beer" when in constructions such as "three beers, please".

More difficult was to find cues for identifying the transitive nature of a noun. After some empirical work, we found a tendency of argumental complements to have a definite article: "temor de la niña" ('fear of the girl'), while modifiers tend to appear without determiners: "mesa de juegos" ('table of games'). Besides, we have taken as a cue the morphological characteristics of deverbal nouns. Suffixes such as "-ción", "-sión", and "-miento", are very much indicative of transitive nouns. Finally, to find the bound preposition of complements, we used a pattern for each possible preposition found after the noun in question.

We used Regular Expressions to implement the linguistic motivated patterns that check for the information just mentioned in a part of speech tagged corpus. The various patterns determine whether the linguistic cues that we have related to syntactic features are found in each occurrence of a particular word in a corpus. The positive or negative results of the *n* pattern checking are stored as binary values of a *n* dimensional vector, one for each occurrence. All vectors produced, one per occurrence of the word in question, are stored then in a kind of vector of vectors that we have called its *signature*. The term *signature* wants to capture the notion that the data it embodies is truly representative of a particular item, and that shows the details of its typical behavior. Particularly, we wanted linguistic cues appearing in different occurrences of the same word to be observed as related information. We have not dealt with ambiguity at all, however. One of the reasons was our focus on low frequency nouns.

4 Methodology and data

We have worked with the *Corpus Tècnic de l'IULA*, a multilingual part of speech tagged corpus which consists of domain specific texts. The section used for our evaluation was the Spanish with 1,091,314 words in the domain of economy and 4,301,096 for medicine. A dataset of 289 nouns, present in both subcorpora, was selected. It was important to compare the behavior of the same nouns in both corpus to check whether the learner was subject to unwanted overfitting.

We used the data for building a C4.5 DT classifier[3]. DT's are one well known and successful technique for this class of tasks when there is enough pre-annotated data available. DT's have the additional benefit that the results can be inspected. The signatures of the words in the Gold-Standard lists were extracted from the corpus of medicine and of the economy one. There was a further test set of 50 nouns with a single occurrence in the corpus of economy for testing purposes. The DT was trained with the signatures of the economy corpus, and the medicine ones as well as the singles set were used for testing.

5 Evaluation

The purpose of the evaluation was to validate our system with respect to the two problems mentioned: noise filtering and generalization capacity by measuring type precision and type recall. We understand type precision as a measure of the noise filtering success, and recall as a measure of the generalization capacity.

In the following tables we present the results of the different experiments. In Table 1, there is a view of the results of the experiment after training and testing with the signatures got in the smaller corpus. The results are for the assignment of the grammatical feature for the two values, *yes* and *no*. And the column named *global* refers to the total percentage of correctly classified instances.

		yes			*no*		
lt	***global***	***prec.***	***rec.***	***F***	***prec.***	***rec.***	***F***
MASS	0.67	0.4	0.26	0.31	0.73	0.83	0.78
COUNT	0.96	0.97	0.99	0.98	0	0	0
TRANS	0.85	0.73	0.45	0.55	0.86	0.95	0.91
INT	0.81	0.84	0.94	0.89	0.64	0.32	0.48
PCOMP	0.9	0.4	0.08	0.13	0.91	0.98	0.95

Table 1. DT results of economy signatures for training and test

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

The most difficult task for the learner is to identify nouns with bound prepositions. Note that there are only 20 nouns with prepositional complements of the 289 test nouns, and that the occurrence of the preposition is not mandatory, and hence the signatures are presented to the learner with very little information.

Table 2 shows the results for 50 nouns with only one occurrence in the corpus. The performance does not change significantly, showing that the generalization capacity of the learner can cope with low frequency words, and that noise in larger signatures has been adequately filtered.

		yes			*no*		
lt	***global***	***prec.***	***rec.***	***F***	***prec.***	***rec.***	***F***
MASS	0.71	0.5	0.16	0.25	0.73	0.93	0.82
COUNT	0.97	0.97	1	0.98	0	0	0
TRANS	0.85	0.75	0.46	0.57	0.87	0.96	0.91
INT	0.83	0.85	0.95	0.89	0.70	0.41	0.52
PCOMP	0.91	0	0	0	0.91	1	0.95

Table 2. DT results for training with signatures of the economy corpus and testing 50 unseen nouns with a single occurrence as test

Table 3 shows that there is little variation in the results of training with signatures of the economy corpus and testing with ones of the medicine corpus. As expected, no variation due to domain is relevant as the information learnt should be valid in all domains.

		yes			*no*		
lt	***global***	***prec.***	***rec.***	***F***	***prec.***	***rec.***	***F***
MASS	0.65	0.44	0.53	0.48	0.77	0.70	0.73
COUNT	0.97	0.97	1	0.98	0	0	0
TRANS	0.82	0.62	0.47	0.54	0.86	0.92	0.89
INT	0.78	0.82	0.92	0.86	0.58	0.35	0.43
PCOMP	0.81	0.31	0.28	0.29	0.92	0.93	0.93

Table 3. DT results for training with economy signatures and testing with medicine signatures

6 Conclusions

The obtained results show that the learning of grammatical features of nouns are learned successfully when using distributional linguistic information as learning features that allow the learner to generalize so as to maintain the performance in cases of nouns with just one occurrence.

There are however issues that should be further investigated. Grammatical features with low precision and recall results (*mass* and *pcomp*) show that some more research should be carried out for finding relevant linguistic cues to be used as learning features. In that respect, the local cues based on morphosyntactic tagging have proved to be useful, minimizing the text preprocessing requirements for getting usable results.

Acknowledgements

The authors would like to thank Jordi Porta, Daniel Chicharro and the anonymous reviewers for helpful comments and suggestions.

References

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. Annual Meeting of the ACL*.

Briscoe, T. and J. Carroll. 1997. "Automatic extraction of subcategorization from corpora". In *Proceedings of the Fifth Conference on Applied Natural Processing*, Washington.

Carroll, J. and A. Fang. 2004. "The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser". In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, Sanya City, China.

Chesley, P and S. Salmon-Alt. 2006. "Automatic extraction of subcategorization frames for French". In *Proc. of the LREC Conference*, Genoa.

Copestake, A.. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications.

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

Shulte im Walde, S. 2002. "Evaluating verb subcategorization frames learned by a German statistical grammar against manual definitions in the Duden Dictionary". In *Proceedings of the 10th EURALEX International Congress*, 187-197.

Witten, Ian H. and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco.