Modelling Polysemy in Adjective Classes by Multi-Label Classification

Gemma Boleda
GLiCom
Universitat Pompeu Fabra
08003 Barcelona
gemma.boleda@upf.edu

Sabine Schulte im Walde
IMS
University of Stuttgart
70174 Stuttgart
schulte@ims.uni-stuttgart.de

Toni Badia
GLiCom
Universitat Pompeu Fabra
08003 Barcelona
toni.badia@upf.edu

Abstract

This paper assesses the role of multi-label classification in modelling polysemy for language acquisition tasks. We focus on the acquisition of semantic classes for Catalan adjectives, and show that polysemy acquisition naturally suits architectures used for multi-label classification. Furthermore, we explore the performance of information drawn from different levels of linguistic description, using feature sets based on morphology, syntax, semantics, and n-gram distribution. Finally, we demonstrate that ensemble classifiers are a powerful and adequate way to combine different types of linguistic evidence: a simple, majority voting ensemble classifier improves the accuracy from 62.5% (best single classifier) to 84%.

1 Introduction

This paper reports on a series of experiments to explore the automatic acquisition of semantic classes for Catalan adjectives. The most important challenge of the classification task is to model the assignment of polysemous lexical instances to multiple semantic classes, combining a) a state-of-the-art Machine Learning architecture for *Multi-label Classification* (Schapire and Singer, 2000; Ghamrawi and McCallum, 2005) and an *Ensemble Classifier* (Dietterich, 2002) with b) the definition of features at various levels of linguistic description.

A proper treatment of polysemy is essential in the area of lexical acquisition, since polysemy represents a pervasive phenomenon in natural language. However, previous approaches to the automatic acquisition of semantic classes have mostly disregarded the problem (cf. Merlo and Stevenson, 2001 and Stevenson and Joanis, 2003 for English semantic verb classes, or Schulte im Walde, 2006 for German semantic verb classes). There are a few exceptions to this tradition, such as Pereira et al. (1993), Rooth et al. (1999), Korhonen et al. (2003), who used soft clustering methods for multiple assignment to verb semantic classes.

Our work addresses the lack of methodology in modelling a polysemous classification. We implement a multi-label classification architecture to handle polysemy. This paper concentrates on the classification of Catalan adjectives, but the general nature of the architecture should allow related tasks to profit from our insights.

As target classification for the experiments, a set of 210 Catalan adjectives was manually classified by experts into three simple and three polysemous semantic classes. We deliberately decided in favour of a small-scale, broad classification. So far, there is little work on the semantic classification of adjectives, as opposed to verbal semantic classification. The semantic classification we propose is a first step in characterising adjectival meaning, and can be refined and extended in subsequent work.

The experiments also provide a thorough comparison of feature sets based on different levels of linguistic description (morphology, syntax, semantics). A set of features is defined for each level of description, and its performance is assessed within the series of experiments. An ensemble classifier comple-

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 171–180, Prague, June 2007.

ments the classification architecture, by optimising the combination of these different types of linguistic evidence.

Our task is motivated by the fact that adjectives play an important role in sentential semantics: they are crucial in determining the reference of NPs, and in defining properties of entities. Even using only three different classes, the information acquired could be applied to, e.g., identify referents in a given context in Dialog or Question Answering systems, and to induce properties of objects within Information Extraction tasks. Furthermore, with the semantic classes corresponding to broad sense representations, they can be exploited for Word Sense Disambiguation.

The remainder of this paper is organised as follows. Section 2 provides background on Catalan adjectives, and Section 3 presents the Gold Standard classification. Section 4 introduces the methodology of the multi-label classification experiments, Section 5 discusses the results, and the improved ensemble classifier is presented in Section 6.

2 Catalan adjective classes

The definition and characterisation of our target semantic classification follows the proposal by Raskin and Nirenburg (1998) within the framework of Ontological Semantics(Nirenburg and Raskin, 2004). In Ontological Semantics, an ontology of concepts modelling the world is explicitly defined, and the semantics of words are mapped onto elements of the ontology. The classification pursued in this paper is drawn up based on the ontological sort of adjectival denotation: all adjectives denote properties, but these properties can be instantiated as simple attributes (*basic adjectives*), relationships to objects (*object-related adjectives*), or relationships to events (*event-related adjectives*).

Basic adjectives are the prototypical adjectives which denote attributes or properties and cannot be decomposed further (such as *bonic* 'beautiful', *gran* 'big'). In Ontological Semantics, these adjectives are mapped to concepts of type *attribute*. For instance, the semantics of the adjective *gran* specifies a mapping to the *size-attribute* element in the ontology. As for event-related adjectives, they have an event component in their meaning and are therefore mapped onto *event* concepts in the ontology. For instance, the semantics of *tangible* ('tangible') includes a pointer to the event element *touch* in the ontology. Similarly, object-related adjectives are mapped onto object concepts in the ontology: *deformació nasal* ('nasal deformity') can be paraphrased as *deformity that affects the nose*, so *nasal* evokes the object *nose*.

The semantic distinctions are mirrored at several levels of linguistic description, such as morphology, syntax, and semantics. For instance, there is a clear relationship between morphological type and semantic class: basic adjectives are typically non-derived, object adjectives tend to be denominal, and event adjectives are usually deverbal. This is the default mapping that one expects from the morphology-semantics interface. As an example for syntactic evidence, basic adjectives in Catalan can be used non-restrictively (in a pre-nominal position) and also predicatively, while object adjectives typically cannot.

However, the correspondences between the linguistic properties and the semantic classes are not one-to-one mappings. Taking the morphological level as an example, some denominal adjectives are basic (such as *vergonyós* 'shy', from *vergonya* 'shyness'). Conversely, some object adjectives are not synchronically denominal (such as *botànic* 'botanical'), and some deverbal adjectives are not event-related, such as *amable* (lit. 'suitable to be loved'; has evolved to 'kind, friendly'). In such cases, the semantic class can be better traced in the distributional properties, not the morphological properties of the adjective.

The proposed classification accounts for some cases of adjectival polysemy. For instance, *familiar* has an object reading (related to the Catalan noun for 'family'), and a basic reading (corresponding to the English adjective 'familiar'):

(1) reunió familiar / cara familiar
meeting familiar / face familiar

'family meeting / familiar face'

Similarly, the participial adjective *sabut* ('known') has an event-related sense, corresponding to the verb *saber* ('know'), and a basic sense equivalent to 'wise':

(2) conseqüència sabuda / home sabut
consequence known / man wise
'known consequence / wise man'

The polysemy between our proposed classes, as exemplified in (1) and (2), is the kind of polysemy we aim to model in the acquisition experiments reported in this paper.

3 Gold Standard classes

As a Gold Standard for the experiments to follow, 210 Catalan adjectives were classified by three experts. The adjectives were randomly sampled from an adjective database (Sanromà, 2003), balancing three factors of variability: frequency, morphological type, and suffix. An equal number of adjectives was chosen from three frequency bands (low, medium, high), from four derivational types (denominal, deverbal, non-derived, participle), and from a series of suffixes within each type. The derivational type and suffix of each adjective were available in the adjective database, and had been manually encoded.

Three experts assigned the 210 lemmata to one out of six classes: each adjective was tagged as basic (B), event (E), object (O), or as polysemous between basic and event (BE), between basic and object (BO), or between event and object (EO). The decisions were reached by consensus. The distribution of the Gold Standard material across classes is shown in the last column of Table 6 (Section 5.2).

In the acquisition experiments, our aim is to automatically assign a class to each adjective that can be simple (B, E, O) or complex (BE, BO, EO), in case of polysemy.

4 Classification method

Adjective classification was performed within a two-level architecture for multi-label classification: first, make a binary decision on each of the classes, and then combine the classifications to achieve a final, multi-label classification. We therefore decomposed the global decision on the (possibly polysemous) class of an adjective into three binary decisions: Is it basic or not? Is it event-related or not? Is it object-related or not? The individual decisions were then combined into an overall classification that included polysemy. For example, if a lemma was classified both as basic and as object in each of the binary decisions, it was deemed polysemous (BO). The motivation behind this approach was that polysemous adjectives should exhibit properties of all the classes involved. As a result, positive decisions on each binary classification can be made by the algorithm, which can be viewed as implicit polysemous assignments.

This classification architecture is very popular in Machine Learning for multi-label problems, cf. (Schapire and Singer, 2000; Ghamrawi and McCallum, 2005), and has also been applied to NLP problems such as entity extraction and noun-phrase chunking (McDonald et al., 2005). The remainder of this section describes other methodological aspects of our experiments.

4.1 Classifier: Decision Trees

As classifier for the binary decisions we chose Decision Trees, one of the most widely used Machine Learning techniques for supervised experiments (Witten and Frank, 2005). Decision Trees provide a transparent representation of the decisions made by the algorithm, and thus facilitate the inspection of results and the error analysis. The experiments were carried out with the freely available Weka software package. The particular algorithm chosen, Weka's J48, is the latest open source version of C4.5 (Quinlan, 1993). For an explanation of decision tree induction and C4.5, see Quinlan (1993) and Witten and Frank (2005, Sections 4.3 and 6.1).

4.2 Feature definition

Five levels of linguistic description, formalised as different feature sets, were chosen for our task. They included evidence from morphology (*morph*), syntax (*func*, *uni*, *bi*), semantics (*sem*), plus a combination of the five levels (*all*). Table 1 lists the linguistic levels, their explanations, and the number of features used on each level.[1] Morphological features (*morph*) encode the derivational type (denominal, deverbal, participial, non-derived) and the suffix (in case the adjective is derived) of each adjective, and correspond to the manually encoded informa-

[1] In level *all*, different features were used for each of the three classes. Table 1 reports the mean number of features across the three classes.

Level	Explanation	# Features
morph	morphological (derivational) properties	2
func	syntactic function	4
uni	uni-gram distribution	24
bi	bi-gram distribution	50
sem	distributional cues of semantic properties	18
all	combination of the 5 linguistic levels	10.3

Table 1: Linguistic levels as feature sets.

tion from the adjective database. Syntactic and semantic features encode distributional properties of adjectives. Syntactic features comprise three subtypes: (i) the syntactic function (level *func*) of the adjective, as assigned by a shallow Constraint Grammar (Alsina et al., 2002), distinguishing the modifier (pre-nominal or post-nominal) and predicative functions; (ii) a unigram distribution (level *uni*), independently encoding the parts of speech (POS) of the words preceding and following the adjective, respectively; and (iii) a bigram distribution (level *bi*), the POS bigram around the target adjective, considering only the 50 most frequent bigrams to avoid sparse features. Semantic features (level *sem*) expand syntactic features with heterogeneous shallow cues of semantic properties. Table 2 lists the semantic properties encoded in the features, as well as the number of heuristic cues defined for each property. As an example, one of the shallow cues used for gradability was the presence of degree adverbs (*més* 'more', *menys* 'less') to the left of the target adjectives. The last set of features, *all*, combines features from all levels of description. However, it does not contain all features, but a selection of the most relevant ones (further details in Section 4.3).

property	#
non-restrictivity	1
predicativity	4
gradability	4
syntactic function of head noun	3
distance to the head noun	1
binaryhood (adjectives with two arguments)	1
agreement properties	2

Table 2: Semantic features.

4.3 Feature selection

Irrelevant features typically decrease performance by 5 to 10% when using Decision Trees (Witten and Frank, 2005, p. 288). We therefore applied a feature selection algorithm. We chose a feature selection method available in Weka (*WrapperSubsetEval*) that selects a subset of the features according to its performance within the Machine Learning algorithm used for classification. Accuracy for a given subset of features is estimated by cross-validation over the training data. Because the number of subsets increases exponentially with the number of features, this method is computationally very expensive, and we used a best-first search strategy to alleviate this problem.

We additionally used the feature selection procedure to select the features for level *all*: for each class, we used only those features that were selected by the feature selection algorithm in at least 30% of the experiments.

4.4 Differences across linguistic levels

One of our goals was to test the strengths and weaknesses of each level of linguistic description for the task of adjective classification. This was done by comparing the accuracy results obtained with each of the feature sets in the Machine Learning experiments. Following a standard procedure in Machine Learning, we created several partitions of the data to obtain different estimates of the accuracy of each of the levels, so as to be able to perform a significance test on the differences in accuracy. We performed 10 experiments with 10-fold cross-validation (*10x10 cv* for short), so that for each class 100 different binary decisions were made for each adjective. For the comparison of accuracies, a standard paired t-test could not be used, because of the inflated Type I error probability when reusing data (Dietterich, 1998). Instead, we used the *corrected resampled t-test* as proposed by Nadeau and Bengio (2003).[2]

5 Classification results

5.1 Accuracy results

The accuracy results for each of the binary decisions (basic/non-basic, event/non-event, object/non-object) are depicted in Table 3.[3] Level *bl* corresponds to the baseline: the baseline accuracy was determined by assigning all lemmata to the most frequent class. The remaining levels follow the nomenclature in Table 1 above. Each column contains the mean and the standard deviation (marked by ±) of the accuracy for the relevant level of information over the 100 results obtained with 10x10 cv.

	Basic	**Event**	**Object**
bl	65.2 ±11.1	76.2 ±9.9	71.9 ±9.6
morph	72.5 ±7.9	89.1 ±6.0	84.2 ±7.5
func	73.6 ±9.3	76.0 ±9.3	81.7 ±7.4
uni	66.1 ±9.4	75.1 ±10.6	82.2 ±7.5
bi	67.4 ±10.6	72.3 ±10.2	83.0 ±8.3
sem	72.8 ±9.0	73.8 ±9.6	82.3 ±8.0
all	**75.3** ±7.6	**89.4** ±5.7	**85.4** ±8.7

Table 3: Accuracy results for binary decisions.

As one might have expected, the best results were obtained with the *all* level (bold faced in Table 3), which is the combination of all feature types. This level achieved a mean improvement of 12.3% over the baseline. The differences in accuracy results between most levels of information were, however, rather small. For the object class, all levels except for *func* and *uni* achieved a significant improvement over the baseline. For the basic class, no improvement over the baseline was significant according to the corrected resampled t-test. And for the event class, only levels *morph* and *all* offered a significant improvement in accuracy; the remaining levels even obtained a slightly lower accuracy score.

These results concern the three individual binary decisions. However, our goal was not to obtain three separate decisions, but a single classification including polysemy. Table 4 shows the accuracy results for the classification obtained by combining the three individual decisions for each adjective. We report two accuracy measures, full and partial: full accuracy required the class assignments to be identical; partial accuracy only required some overlap in the classification of the Machine Learning algorithm and the Gold Standard for a given class assignment. The motivation for calculating partial overlap was that a class assignment with some overlap with the Gold Standard (even if they were not identical) is generally more useful than a class assignment with no overlap.

	Full	**Partial**
bl	51.0 ±0.0	65.2 ±0.0
morph	60.6 ±1.3	87.8 ±0.4
func	53.5 ±1.8	79.8 ±1.3
uni	52.3 ±1.7	76.7 ±1.0
bi	52.9 ±1.9	76.9 ±1.8
sem	52.0 ±1.3	78.7 ±1.7
all	**62.3** ±2.3	**90.7** ±1.6

Table 4: Accuracy results for combined decisions.

Again, the best results were obtained with the *all* level. The second best results were obtained with level *morph*. These results could have been expected from the results obtained by the individual decisions (Table 3); however, note that the differences between the various levels are much clearer in the combined classification than in the individual binary decisions.

Table 5 shows the two-by-two comparisons of the accuracy scores. Each cell contains the difference in accuracy means between two levels of description, as well as the level of significance of the difference. The significance is marked as follows: * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. If no asterisk is shown, the difference was not significant.

Under the strictest evaluation condition (full accu-

[2]Note that the corrected resampled t-test can only compare accuracies obtained under two conditions (algorithms or, as is our case, feature sets); ANOVA would be more adequate. In the field of Machine Learning, there is no established correction for ANOVA for the purposes of testing differences in accuracy (Bouckaert, 2004). Therefore, we used multiple t-tests instead, which increases the overall error probability of the results for the significance tests.

[3]The accuracy for each decision was computed independently. For instance, a *BE* adjective was judged correct within the basic class iff the decision was *basic*; correct within the event class iff the decision was *event*; and correct within the object class iff the decision was *non-object*.

agreement	level	bl	morph	func	uni	bi	sem
full	morph	9.7***					
	func	2.5*	-7.1***				
	uni	1.4	-8.3***	-1.1			
	bi	2.0	-7.7***	-0.6	0.6		
	sem	1.0	-8.7***	-1.5	-0.4	1.0	
	all	11.4***	1.7	8.9***	10.0***	9.4***	10.4***
partial	morph	-22.6***					
	func	14.6***	-8.0***				
	uni	11.4***	-11.1***	-3.1**			
	bi	11.7***	-10.9***	-2.9**	0.2		
	sem	13.4***	-9.1***	-1.1	2.0	1.8	
	all	25.4***	2.9*	10.9***	14.0***	13.8***	12.0***

Table 5: Comparison of accuracy scores across linguistic levels.

racy), only levels *morph*, *func*, and *all* significantly improved upon the baseline. Levels *morph* and *all* are better than the remaining levels, to a similar extent. In the partial evaluation condition, all levels achieved a highly significant improvement over the baseline ($p < 0.001$). Therefore, the classifications obtained with any of the feature levels are more useful than the baseline, in the sense that they present more overlap with the Gold Standard.

The best result obtained for the full classification of adjectives with our methodology achieved a mean of 62.3% (full accuracy) or 90.7% (partial accuracy), which represents an improvement of 11.3% and 25.5% over the baselines, respectively. Levels including morphological information were clearly superior to levels using only distributional information.

These results suggest that morphology is the best single source of evidence for our task. However, recall from Section 3 that the sampling procedure for the Gold Standard explicitly balanced for morphological factors. As a result, denominal and participial adjectives are underrepresented in the Gold Standard, while non-derived and deverbal adjectives are overrepresented. Moreover, previous experiments on different datasets (Boleda et al., 2004; Boleda et al., 2005) provided some evidence that distributional information outperforms morphological information for our task. Therefore, we cannot conclude from the experiments that morphological features are the most important information for the classification of Catalan adjectives in general.

5.2 Error analysis

The error analysis focuses on the two best feature sets, *morph* and *all*. Table 6 compares the errors made by the experiment classifications (based on the two sets of features) against the Gold Standard classification. To obtain a unique experiment classification for each feature level in this comparison, we applied majority voting across the 10 different classifications obtained in the 10 experiment runs for each of the linguistic levels. The table rows correspond to the Gold Standard classification and the columns correspond to the experiment classifications with the feature levels *all* and *morph*, respectively. The matches (the diagonal elements) are in italics, and off-diagonal cells representing the largest numbers of mismatches are boldfaced. The overall number of mistakes made by both levels with majority voting is almost the same: 86 (*morph*) vs. 89 (*all*). However, the mismatches are qualitatively quite different.

Level *morph* uniformly mapped denominal adjectives to both basic and object (BO). Because of this overgeneration of BOs, 31 lemmata that were tagged as either basic or object in the Gold Standard were assigned to BO. In contrast, level *all* was overly discriminative: most of the BO cases (16 out of 23), as well as 16 object adjectives, were assigned to basic. This type of confusion could be explained by the fact that some non-prototypical basic adjectives were as-

		all						**morph**						
		B	BE	BO	E	EO	O	B	BE	BO	E	EO	O	*Total*
GS	B	*94*	**12**	0	0	1	0	*82*	2	**10**	**11**	2	0	*107*
	BE	1	*6*	0	0	0	0	0	*1*	0	6	0	0	*7*
	BO	**16**	1	*5*	1	0	0	5	0	*16*	2	0	0	*23*
	E	5	**23**	1	*7*	1	0	4	7	0	*25*	1	0	*37*
	EO	0	2	0	0	*4*	0	0	0	0	6	*0*	0	*6*
	O	**16**	1	6	2	0	*5*	6	0	**21**	3	0	*0*	*30*
	Total	*132*	*45*	*12*	*10*	*6*	*5*	*97*	*10*	*47*	*53*	*3*	0	*210*

Table 6: Levels *all* and *morph* against the Gold Standard.

signed to the basic class in the Gold Standard, because they did not fit the narrower definitions of the event and object classes, but these adjectives do not behave like typical basic adjectives.

As for event adjectives, the *morph* level assigned almost all deverbal adjectives to the event class, which worked well in most cases (26). However, this mapping cannot distinguish deverbal adjectives with a basic meaning (11 basic and 6 BE adjectives in the Gold Standard). Level *all*, including morphological and distributional cues, also shows difficulties with the event class, but of a different nature. Feature examination showed that the distributional differences between basic and event adjectives are not robust. For instance, according to t-tests performed on the Gold Standard ($\alpha = 0.05$), only three of the 18 semantic features exhibit significant mean differences for classes basic and event. In contrast, ANOVA across the 6 classes ($\alpha = 0.05$) yields significant differences for 16 out of the 18 features, which indicates that most features serve to distinguish object adjectives from basic and event adjectives. As a result of the lack of robust distributional differences between basic and event adjectives, 35 basic or event adjectives were classified as BE when using the *all* level as feature set.

Further 23 event adjectives were incorrectly classified as BE by the *all* level, but correctly classified by the *morph* level, because they are deverbal adjectives. These cases involved adjectives derived from stative verbs, such as *abundant* ('abundant') or *preferible* ('preferable'). Feature analysis revealed that deverbal adjectives derived from stative verbs are more similar to basic adjectives than those derived from process-denoting verbs.

To sum up, the default morphological mapping mentioned in Section 2 works well in most cases but has a clear ceiling, as it cannot account for deviations from the expected mapping. Distributional cues are more sensitive to these deviations, but fail mostly in the distinction between basic and event, because the differences in syntactic distribution between these classes are not robust.

6 An improved classifier

The error analysis in the previous section has shown that, although the number of mistakes made with level *morph* and *all* is comparable, the kinds of mistakes are qualitatively very different. This suggests that mixing features for the construction of a single Decision Tree, as is done in level *all*, is not the optimal way to combine the strengths of each level of description. An alternative combination can be achieved with an *ensemble classifier*, a type of classifier that has received much attention in the Machine Learning community in the last decade (Dietterich, 2002). When building an ensemble classifier, several class proposals for each item are obtained, and one of them is chosen on the basis of majority voting, weighted voting, or more sophisticated decision methods. It has been shown that in most cases, the accuracy of the ensemble classifier is higher than the best individual classifier (Freund and Schapire, 1996; Dietterich, 2000; Breiman, 2001). Within NLP, ensemble classifiers have been applied, for instance, to genus term disambiguation in machine-readable dictionaries (Rigau et al., 1997), using a majority voting scheme upon several heuristics, and to part of speech tagging, by combining the class predictions of different algorithms (van Halteren et

Levels	Full Ac.	Part. Ac.
morph+func+uni+bi+sem+all	84.0 ±0.06	95.7 ±0.02
func+uni+bi+sem	81.5 ±0.04	95.9 ±0.01
morph+func+sem+all	72.4 ±0.03	89.3 ±0.02
bl	51.0 ±0.0	65.2 ±0.0
all	62.3 ±2.3	90.7 ±1.6

Table 7: Results for ensemble classifier.

al., 1998). The main reason for the general success of ensemble classifiers is that they gloss over the biases introduced by the individual systems.

We implemented an ensemble classifier by using the different levels of description as different subsets of features, and applying majority voting across the class proposals from each level. Intuitively, this architecture is analogous to having a team of linguists and NLP engineers, each contributing their knowledge on morphology, n-gram distribution, syntactic properties, etc., and have them reach a consensus classification. We thus established a different classification for each of the 10 cross-validation runs by assigning each adjective to the class that received most votes. To enable a majority vote, at least three levels have to be combined. Table 7 contains a representative selection of the combinations, together with their accuracies. Also, the accuracies obtained with the baseline (*bl*) and the best single level (*all*) are included for comparison.

In any of the combinations tested, accuracy improved over 10% with respect to the *all* level. The best result, a mean of 84% (full accuracy), was obtained by combining all levels of description. These results represent a raw improvement over the baseline of 33%, and 21.7% over the best single classifier. Also note that with this procedure 95.7% of the classifications obtained with the ensemble classifier present partial overlap with the class assignments in the Gold Standard.

These results show that the combination of different sources of linguistic evidence is more important than the type of information used. As an example, consider the second ensemble classifier in Table 7: this classifier excludes the two levels that contain morphological information (*morph* and *all*), which represents the most successful individual source of information for our dataset. Nevertheless, the combination achieved 19.2/20.9% more accuracy than levels *all* and *morph*, respectively.

7 Related work

Adjectives have received less attention than verbs and nouns within Lexical Acquisition. Work by Hatzivassiloglou and colleagues (Hatzivassiloglou and McKeown, 1993; Hatzivassiloglou and McKeown, 1997; Hatzivassiloglou and Wiebe, 2000) used clustering methods to automatically identify adjectival scales from corpora.

Coordination information was used in Bohnet et al. (2002) for a classification task similar to the task we pursue, using a bootstrapping approach. The authors, however, pursued a classification that is not purely semantic, between quantitative adjectives (similar to determiners, like *viele* 'many'), referential adjectives (*heutige*, 'of today'), qualitative adjectives (equivalent to basic adjectives), classificatory adjectives (equivalent to object adjectives), and adjectives of origin (*Stuttgarter*, 'from Stuttgart').

In a recent paper, Yallop et al. (2005) reported experiments on the acquisition of syntactic subcategorisation patterns for English adjectives.

Apart from the above research with a classificatory flavour, other lines of research exploited lexical relations among adjectives for Word Sense Disambiguation (Justeson and Katz, 1995; Chao and Dyer, 2000). Work by Lapata (2001), contrary to the studies mentioned so far, focused on the meaning of adjective-noun combinations, not on that of adjectives alone.

8 Conclusion

This paper has presented an architecture for the semantic classification of Catalan adjectives that explicitly includes polysemous classes. The focus of the architecture was on two issues: *(i) finding an appropriate set of linguistic features,* and *(ii) defining an adequate architecture for the task.* The investigation and comparison of features at various linguistic levels has shown that morphology plays a major role for the target classification, despite the caveats raised in the discussion. Morphological features related to derivational processes are among the simplest types of features to extract, so that the approach can be straightforwardly extended to languages similar to Catalan with no extensive need of resources.

Furthermore, we have argued that polysemy acquisition naturally suits multi-label classification architectures. We have implemented a standard architecture for this class of problems, and demonstrated its applicability and success. The general nature of the architecture should be useful for related tasks that involve polysemy within the area of automatic lexical acquisition.

Our work has focused on a broad classification of the adjectives, similarly to Merlo and Stevenson (2001), who classified transitive English verbs into three semantic classes. The small number of classes might be considered as an over-simplification of adjective semantics, but the simplified setup facilitates a detailed qualitative evaluation. In addition, as there has been virtually no work on the acquisition of semantic classes for adjectives, it seems sensible to start with a small number of classes and incrementally build upon that. Previous work has demonstrated that multi-label classification is applicable also to a large number of classes as used in, e.g., document categorisation (Schapire and Singer, 2000). This potential can be exploited in future work, addressing a finer-grained adjective classification.

Finally, we have demonstrated that the combination of different types of linguistic evidence boosts the performance of the system beyond the best single type of information: ensemble classifiers are a more adequate way to combine the linguistic levels of description than simply merging all features for tree construction. Using a simple, majority voting ensemble classifier, the accuracy jumped from 62.5% (best single classifier) to 84%. This result is impressive by itself, and also in comparison to similar work such as (Rigau et al., 1997), who achieved a 9% improvement on a similar task. Our insights are therefore useful in related work which involves the selection of linguistic features in Machine Learning experiments.

Future work involves three main lines of research. First, the refinement of the classification itself, based on the results of the experiments presented. Second, the use of additional linguistic evidence that contributes to the semantic class distinctions (e.g., selectional restrictions). Third, the application of the acquired information to broader NLP tasks. For example, given that each semantic class exhibits a particular syntactic behaviour, information on the semantic class should improve POS-tagging for adjective-noun and adjective-participle ambiguities, probably the most difficult distinctions both for humans and computers (Marcus et al., 1993; Brants, 2000). Also, semantic classes might be useful in terminology extraction, where, presumably, object adjectives participate in terms more often than basic adjectives.[4]

Acknowledgements

We thank Roser Sanromà for providing us with the manually annotated database of adjectives, and for being part of the Gold Standard annotation committee. Special thanks also to the Institut d'Estudis Catalans for proding us with the research corpus. The comments of three anonymous reviewers helped improve the paper. Finally, the financial support of the Universitat Pompeu Fabra and its Translation and Philology Department and the Fundación Caja Madrid is gratefully acknowledged.

References

À. Alsina, T. Badia, G. Boleda, S. Bott, À. Gil, M. Quixal, and O. Valentín. 2002. CATCG: a general purpose parsing tool applied. In *Proceedings of Third International Conference on Language Resources and Evaluation (LREC-02)*, Las Palmas, Spain.

B. Bohnet, S. Klatt, and L. Wanner. 2002. An approach to automatic annotation of functional information to adjectives with an application to German. In *Proceedings of the LREC Workshop on Linguistic Knowledge Acquisition and Representation*, Las Palmas, Spain.

G. Boleda, T. Badia, and E. Batlle. 2004. Acquisition of semantic classes for adjectives from distributional evidence. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1119–1125, Geneva, Switzerland.

G. Boleda, T. Badia, and S. Schulte im Walde. 2005. Morphology vs. syntax in adjective class acquisition. In *Proceedings of the ACL-SIGLEX 2005 Workshop on Deep Lexical Acquisition*, pages 1119–1125, Ann Arbor, USA.

R. Bouckaert. 2004. Estimating replicability of classifier learning experiments. In *Proceedings of ICML.*

T. Brants. 2000. Inter-annotator agreement for a german newspaper corpus. In *Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

[4]Horacio Rodríguez, p. c., April 2007.

L. Breiman. 2001. Random forests. *Mach. Learn.*, 45:5–23.

G. Chao and M. G. Dyer. 2000. Word sense disambiguation of adjectives using probabilistic networks. In *Proceedings of COLING*, pages 152–158.

T.G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

T.G. Dietterich. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.*, 40:5–23.

T.G. Dietterich. 2002. Ensemble learning. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. The MIT Press.

Y. Freund and R.E. Schapire. 1996. Experiments with a new boosting algorithm. In *Proceedings of ICML*, pages 148–156.

N. Ghamrawi and A. McCallum. 2005. Collective multi-label classification. In *Proceedings of 14th Conf. on Information and Knowledge Management*.

V. Hatzivassiloglou and K. R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of ACL*, pages 172–182.

V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL/EACL*, pages 174–181.

V. Hatzivassiloglou and J. M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING*, pages 299–305, Morristown, NJ, USA. Association for Computational Linguistics.

J. S. Justeson and S. M. Katz. 1995. Principled disambiguation: Discriminating adjective senses with modified nouns. *Computational Linguistics*, 21(1):1–27.

A. Korhonen, Y. Krymolowski, and Z. Marx. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of ACL*, pages 64–71.

M. Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of NAACL*.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

R. McDonald, K. Crammer, and F. Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *Proceedings of HLT-EMNLP*, pages 987–994.

P. Merlo and S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Comp. Ling.*, 27(3):373–408.

C. Nadeau and Y. Bengio. 2003. Inference for the generalization error. *Mach. Learn.*, 52(3):239–281.

S. Nirenburg and V. Raskin. 2004. *Ontological Semantics*. MIT Press.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional Clustering of English Words. In *Proceedings of ACL*, pages 183–190.

R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

V. Raskin and S. Nirenburg. 1998. An applied ontological semantic microtheory of adjective meaning for natural language processing. *Mach. Trans.*, 13(2-3):135–227.

G. Rigau, J. Atserias, and E. Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of EACL*, pages 48–55.

M. Rooth, S. Riezler, D. Prescher, G. Carroll, and F. Beil. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of ACL*.

R. Sanromà. 2003. Aspectes morfològics i sintàctics dels adjectius en català. Master's thesis, Universitat Pompeu Fabra.

R.E. Schapire and Y. Singer. 2000. Boostexter: A boosting-based system for text categorization. *Mach. Learn.*, 39(2-3):135–168.

S. Stevenson and E. Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of CoNLL*.

H. van Halteren, J. Zavrel, and W. Daelemans. 1998. Improving data driven wordclass tagging by system combination. In *Proceedings of ACL*, pages 491–497.

I.H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

J. Yallop, A. Korhonen, and T. Briscoe. 2005. Automatic acquisition of adjectival subcategorization from corpora. In *Proceedings of ACL*, Ann Arbor, Michigan.