A Semi-supervised Type-based Classification of Adjectives: Distinguishing Properties and Relations

Matthias Hartung, Anette Frank

Computational Linguistics Department
Heidelberg University
{hartung, frank}@cl.uni-heidelberg.de

Abstract

We present a semi-supervised machine-learning approach for the classification of adjectives into property- vs. relation-denoting adjectives, a distinction that is highly relevant for ontology learning. The feasibility of this classification task is evaluated in a human annotation experiment. We observe that token-level annotation of these classes is expensive and difficult. Yet, a careful corpus analysis reveals that adjective classes tend to be stable, with few occurrences of class shifts observed at the token level. As a consequence, we opt for a type-based semi-supervised classification approach. The class labels obtained from manual annotation are projected to large amounts of unannotated token samples. Training on heuristically labeled data yields high classification performance on our own data and on a data set compiled from WordNet. Our results suggest that it is feasible to automatically distinguish adjectives denoting properties and relations, using small amounts of annotated data.

1. Introduction

One important function of adjectives used as modifiers in natural language is that they elicit certain properties of nouns. Therefore, adjectives have been examined for the task of learning attributes for ontology induction (Almuhareb and Poesio, 2004). In line with Almuhareb (2006), we aim at the corpus-based induction of conceptual knowledge, with a focus on learning attributes of concept classes based on the meaning of noun-modifying adjectives. For example, from the cooccurrence of a noun and a property-denoting adjective, as in *red car*, we want to be able to infer that cars have an attribute COLOR. Relation learning is another problem relevant in ontology induction. Here, we are concerned with relation-denoting adjectives, as in *environmental science*, that are to be represented as relational concepts ('SCIENCE about/addressing the ENVIRONMENT').

In this paper, we examine whether the task of automatically distinguishing property- and relation-denoting adjectives is feasible in principle. For this purpose, we adopt an adjective classification scheme that separates adjectives into subtypes relevant for ontology learning. This classification scheme is evaluated in two tasks. First, we assess the validity of the scheme in a human annotation task. In a second step, we present a machine-learning approach for the automatic classification of adjectives into property- and relation-denoting lexical types.

In our annotation experiment we observe that token-level annotation for adjective types is difficult and expensive. At the same time, careful analysis of the annotated corpus reveals that adjective types tend to be stable, with only few occurrences of class shifts observed at the token level. This ability of an adjective to change its class on the token level will be denoted as *class volatility* throughout the paper.

A second observation is that features that may be used to separate the two classes in a machine-learning approach are essentially type-based, focusing on grammatical properties that are not exhibited by all instances in particular contexts. These insights suggest a type-based classification approach, similar to recent work in semantic verb classification by Miyao and Tsujii (2009). Based on the observed low class volatility, we use the token-level annotations from our annotated corpus as seeds for the acquisition of a large training set by semi-supervised instance generation. The classifiers trained on this heuristically annotated set identify property- and relation-denoting adjectives with high precision well above the baseline.

2. Related Work

Using adjectives for attribute learning has first been proposed by Almuhareb and Poesio (2004) and Cimiano (2006). Cimiano's work on this particular task is based on the investigation of adjective-noun phrases from corpora. For every adjective modifying a noun, its possible attributes are extracted from WordNet (Fellbaum, 1998) and associated with the respective noun. As this approach depends on an external lexical resource, it is obviously limited in coverage. Almuhareb (2006) aims at learning this information on a larger scale by means of a pattern-based approach

that operates on large web-based corpora. The outcome of his work on this task, however, is considerably affected by the lack of a separation between property-denoting and relational adjectives, such that a large number of adjectives is erroneously identified as attribute-denoting by his system.

Classification schemes similar to the one we envisage here have been presented by Torrent (2006) for Catalan and Raskin and Nirenburg (1998) for English. Their goal was the creation of a large-scale adjective lexicon for NLP tasks. The most fundamental difference between the work of Raskin and Nirenburg and ours is that they created their resource manually. In contrast, we aim at automatic classification, as effective automatic methods have the advantage that they can be applied to novel, specialized domains. Torrent (2006) made use of clustering techniques to automatically establish adjective classes in Catalan. She obtained various sets of clusters that were evaluated against a human-annotated gold standard, yielding up to 73% accuracy.[1] Since our aim is the targeted acquisition and classification of adjectives for the purpose of ontology learning, we opt for a classification approach that allows us to pre-specify (and possibly refine and extend) appropriate target classes for concept learning – which is not an option with clustering.

Finally, Amoia and Gardent (2008) present a (manual) classification of adjectives that relies on logical properties of adjectives in the tradition of Montague (1974). While this perspective is orthogonal to our work, their work might be useful to supplement our approach, by providing further adjective classes that can be sorted out as being neither property-denoting nor relational.

Methodologically, our approach is related to a great body of work in automatic verb classification (most recently, Miyao and Tsujii (2009)), going back to the empirical work of Levin (1993). Although in this field the number of target classes is by far greater and aimed at a conceptual semantic classification, the common denominator between verb semantic classes and the adjective classes considered here is that certain properties on the type level are constitutive for class membership, while on the token level only a single property is observable at a time. In line with this strand of work on Levin-style verb classification, our classification approach will operate on the type level.

[1] A strict comparison of the two approaches will not be possible due to the different languages considered and divergences regarding the selected target classes.

3. Corpus Annotation

As a basis for distinguishing adjective classes relevant for ontology learning, we adopt the three-way classification that has been proposed for Catalan adjectives by Torrent (2006). According to the class labels (**b**asic, **e**vent-related and **o**bject-related), we name this classification scheme *BEO classification*. We give a brief overview of the properties exhibited by the BEO classes, paying special attention to their relevance for ontology learning.

3.1. Classification Scheme

Basic Adjectives. Basic adjectives denote values of an attribute of an entity. They denote either discrete values of an attribute, such as 'oval' for an object's attribute SHAPE, or sets of possible discrete values, as for 'young' for the attribute AGE.

(1) oval table ↔ SHAPE(table)=oval
young girl ↔ AGE(girl)=young

Event-related Adjectives. These adjectives modify an associated event the referent of the noun takes part in, as illustrated by the following paraphrases:

(2) eloquent person ↔ person that *speaks* eloquently

(3) interesting book ↔ book that is interesting to *read*

Object-related Adjectives. This class comprises adjectives that are morphologically derived from a noun, denoted as $A_{/N}$ and N_b, respectively, as in (4). In these cases, N_b refers to an entity that acts as a semantic dependent of the head noun N.

(4) economic$_{[A_{/N}]}$ crisis$_{[N]}$ ↔ crisis of the *economy*$_{[N_b]}$
political$_{[A_{/N}]}$ debate$_{[N]}$ ↔ debate on *politics*$_{[N_b]}$

BEO classes in Ontology Learning. As seen above, the BEO classes distinguish properties (basic and event-related adjectives) from relational meanings (object-related adjectives). This distinction can be utilized in ontology learning for the acquisition of property-based concept descriptions and semantic relations between concepts, respectively.

3.2. Annotation Process

Methodology. To validate the BEO classification scheme, we ran an annotation experiment with three human annotators. We compiled a list of 200 high-frequency adjectives from the British National Corpus[2] and for each of them randomly extracted five ex-

[2] We used version 3 of the BNC XML Edition, available from: `http://www.natcorp.ox.ac.uk/`

ample sentences. The annotators labelled each item as BASIC, EVENT, OBJECT or IMPOSSIBLE. The latter was supposed to be used in case the annotators were unable to provide a label due to erroneous examples, insufficient context, or instances belonging to alternative classes of adjectives not considered here.

Ambiguities between BEO Classes. The most notable ambiguity among BEO classes holds between basic and event-related adjectives. Consider the following competing analyses for *fast horse*:

(5) fast horse ↔ VELOCITY(horse)=fast
fast horse ↔ horse that *runs* fast

We argue that this ambiguity sheds light on the difference between *independent* and *founded* properties of an object (cf. Guarino (1992)). For disambiguation we propose the inference patterns in (6).

(6) ENT(ity)'s property of being ADJ(ective) is due to ENT's ability to EVENT.
If ENT was not able to EVENT, it would not be an ADJ ENT.

Applied to (5), these patterns indicate that, in the case of a horse, being fast should be formalized as a property that is founded on the horse's inherent ability to run (or, at least, to move). If this ability was absent, it would no longer be possible to qualify the horse as being fast. Hence, we prefer the event reading for (5).

4. Corpus Analysis

4.1. Agreement Figures

Table 1 displays agreement figures for our annotation experiment in terms of Fleiss' Kappa (Fleiss, 1971). Total agreement between all three annotators amounts to $\kappa = 0.404$. Note that we observe substantial agreement of $\kappa = 0.762$ between two of the annotators, which suggests that the upper bound is higher than the observed overall agreement.
Table 2 displays the overall agreement figures broken down into the four class labels. These results underline our intuition that the distinction between the classes BASIC and EVENT is very difficult even for human subjects.

4.2. Re-Analysis: Binary Classification Scheme

This observation led us to re-analyze our data using a binary classification that collapses basic and event-related adjectives into one class. This re-analysis is merely a shift in granularity: both basic and event-related adjectives denote properties, whereas object-related adjectives denote relations. Re-analyzing the data in this way improves overall agreement to $\kappa = 0.69$. See Table 3 for detailed agreement figures.
The remaining disagreements between annotators have been manually adjudicated by one of the authors. After adjudication, the data set[3] contains 689 adjective *tokens* that are unambiguously annotated, given the respective context, as denoting a property, while 138 tokens are labeled as relational. In total, 190 (out of 200) lexical adjective *types* are covered. The missing mass is due to items marked as IMPOSSIBLE by at least one annotator.

	Annot. 1	Annot. 2	Annot. 3
Annot. 1	—	0.762	0.235
Annot. 2	0.762	—	0.285
Annot. 3	0.235	0.285	—

Table 1: Agreement figures in terms of Fleiss' κ

	BASIC	EVENT	OBJECT	IMPOSS
κ	0.368	0.061	0.700	0.452

Table 2: Category-wise κ-values for all annotators

	BASIC+EVENT	OBJECT	IMPOSS
κ	0.696	0.701	-0.003

Table 3: Category-wise κ-values, bi-partite classification

4.3. Class Volatility

In order to judge the possibility of a *type-based* automatic adjective classification, we need to quantify the degree of class volatility we observe in the annotated corpus, i.e. the proportion of lexical types that are assigned alternating class labels at the token level.
We identified 12 adjectives that are volatile in the sense that they can undergo a type shift between basic and event-related vs. object-related adjectives[4] on the token level. Thus, the proportion of volatile types in the data set amounts to 6.3%.
In a further adjudication step, the number of volatile types could be reduced to 5 by evaluating fine-grained interpretation differences. Table 4 displays the full list of adjectives considered before and after adjudication, including their frequency distribution over the two classes. The subset of adjectives established as "true volatiles" after adjudication is given in boldface.
In the following, we discuss some typical cases of shifts between property-denoting and relational inter-

[3] The annotated corpus will be made freely available on request.

[4] Henceforth, we will refer to these binary classes as ATTR(ibutive) and REL(ational).

Type	after adjudication			before adjud.	
	#ATTR	#REL	#ambig.	#ATTR	#REL
black	2	2	0	2	2
male	4	1	0	4	1
personal	2	2	1	2	3
political	2	2	1	1	4
white	3	1	0	3	1
detailed	5	0	0	4	1
mental	0	5	0	2	3
military	0	5	0	1	4
nuclear	0	5	0	1	4
professional	0	5	0	3	2
regional	0	5	0	1	4
technical	0	4	0	1	3

Table 4: Overview of volatile adjectives in the data set

pretations of adjectives.

4.3.1. Shifts from ATTR to REL

(7) Certain stations in BLACK rural areas or town locations were expected to be used exclusively by Africans.

(8) The suburban commuter station was emphatically a MALE preserve at certain times of day.

Both *black* in (7) and *male* in (8) have to be assigned a relational interpretation even though the basic meaning of these adjectives is property-denoting. This shift can be analyzed as a metonymic process where the adjective is re-interpreted as referring to an entity to which the respective property applies (concretely: *black people*). This entity, in turn, acts as an argument in a relation with the head noun. Thus, *black rural areas* in (7) and *male preserve* in (8) can be paraphrased as *rural areas inhabited by black people* and *a preserve occupied by male people*, respectively.

4.3.2. Shifts from REL to ATTR

In the following example, we observe a shift from a relational to a property-based adjective reading:

(9) But then aren't you taking a POLITICAL stance, rather than an aesthetic one?

We argue that a *political stance*, as in (9), does not denote a particular *stance on politics* (which would be the obvious relational interpretation), but a property: a stance that is *politically motivated* or *held for political reasons*. The given context crucially elicits the class-delimiting function of the adjective, in that different subtypes of stances are contrasted.

5. Semi-supervised Type-based Classification of Adjectives

5.1. Features for Classification

Our classification approach is based on the observation that property- and relation-denoting adjectives systematically differ with regard to their behaviour in certain grammatical constructions. These differences can be captured in terms of lexico-syntactic patterns (Amoia and Gardent, 2008; Beesley, 1982; Raskin and Nirenburg, 1998; Torrent, 2006). We can cluster these patterns into groups (see Table 5): `I` (features encoding comparability), `II` (gradability), `III` (predicative use), `IV` and `V` (particular constructions). All these feature groups encode grammatical properties that can be found with property-denoting adjectives only, while relational adjectives do not license them. As a positive feature for relational adjectives, we consider morphological derivation from nouns (group `VI`), e.g. *criminal – crime*, *economic – economy*). This information was extracted from CELEX2 (Baayen et al., 1996).

5.2. Semi-supervised Instance Generation

A major problem we encounter with the features presented above is their severe sparsity. Applied to our annotated corpus of 1000 sentences, the complete feature set yields only 10 hits.

Given the results of our corpus analysis in Section 4, however, we can raise the classification task to the type level, under the proviso that class volatility is limited to only a small number of adjective types and contextual occurrences. Under this assumption, we can use our annotated data set as seed material for the automatic acquisition of a large annotated corpus by semi-supervised instance generation. In this process, the unanimous class labels gathered from the manually annotated corpus are projected to the unannotated data. This means that potential class changes on the token level are disregarded.

5.3. Data Set Construction

Using the heuristic annotation projection technique described above, we created two data sets for our classification experiments. These provide the training data for two classifiers: a decision tree (ADTree) and a meta classifier that makes use of boosting. For the experiments reported here, we relied on classifier implementations of Weka (Witten and Frank, 2005).

Data Set 1. The first data set we created is based on the manually annotated corpus described above. We identified all adjective types in the corpus that exhibit perfect agreement across all annotators and are

Group	Feature	Pattern	Example
I	as	as **JJ** as	*as cheap as possible*
	comparative-1	**JJR** NN	*halogen produces a brighter light*
	comparative-2	RBR **JJ** than	*more famous than your enemies*
	superlative-1	**JJS** NN	*this is the broadest question*
	superlative-2	the RBS **JJ** NN	*one of the most beautiful buildings in Europe*
II	extremely	an extremely **JJ** NN	*an extremely nice marriage*
	incredibly	an incredibly **JJ** NN	*an incredibly low downturn*
	really	a really **JJ** NN	*a really simple solution*
	reasonably	a reasonably **JJ** NN	*a reasonably clear impression*
	remarkably	a remarkably **JJ** NN	*a remarkably short amount of time*
	very	DT very **JJ**	*gets onto a very dangerous territory*
III	predicative-use	NN (WP|WDT)? is|was|are|were RB? **JJ**	*my digital camera is nice*
	static-dynamic-1	NN is|was|are|were being **JJ**	*the current joint unit was being successful*
	static-dynamic-2	be RB? **JJ** .	*Be absolutely certain:*
IV	one-proform	a/an RB? **JJ** one	*a hard one*
V	see-catch-find	see|catch|find DT NN **JJ**	*90% found the events relevant and useful*
VI	morph	adjective is morphologically derived from noun	*culture → cultural*

Table 5: Set of features used for classification

not found to be volatile. This yields 164 property-denoting and 18 relational types, which we use as seeds for heuristic token-level annotation. For each lexical adjective type, we acquired a corpus of 5000 sentences from a subsection of the ukWaC corpus (Ferraresi et al., 2008) to which the labels from the annotated corpus were projected. We refer to this data set as DS1.

Data Set 2. In order to assess the soundness of our features on a larger sample and to evaluate whether our method of heuristic annotation projection can be generalized to different data sets, we also compiled a gold standard of property-denoting and relational adjectives from WordNet[5] (Fellbaum, 1998).

Like any other PoS category, adjectives in WordNet are organized in *synsets*, i.e. sets of (nearly) synonymous types. Every synset constitutes a *word sense*, thus reflecting fine-grained meaning differences. All lexical knowledge in WordNet is encoded in terms of semantic relations between word senses. The information of interest for our task is captured by the relations *attribute* and *pertainymy*. Presence of an *attribute* relation between an adjective and a noun sense indicates that the noun denotes a property and the adjective specifies a possible value of this property. A *pertainymy* relation linking an adjective and a noun sense indicates a relational adjective meaning. If neither an *attribute* nor a *pertainymy* relation is specified for a given adjective, nothing can be inferred regarding the binary classification considered here.

For the construction of our gold standard, we collected all adjectives from WordNet that are unambiguously property-denoting or relational, meaning that *all* of their senses are marked with either the *attribute* or the *pertainymy* relation. This yields 3727 property-denoting and 3655 relational types (i.e., roughly one third of the overall 21486 adjective types in WordNet). We only considered adjectives with more than 2000 occurrences in the same subsection of the ukWaC corpus we had used for the construction of DS1. The final data set comprises 246 property-denoting and 140 relational adjective types. Again, we extracted up to 5000 sentences from ukWaC for each of these adjectives, and assigned them the class labels ATTR and REL, respectively. The resulting data set is referred to as DS2.

6. Evaluation

As our classification is intended to be used in ontology learning tasks, we evaluate the performance of the classifiers in separating property-denoting vs. relational adjectives in terms of precision and recall. Depending on whether attribute or relation learning is in focus, it is important to achieve high performance for the respective target category of adjectives rather than good overall accuracy for both classes.

In the following, we report on the classification performance on both data sets, based on different feature combinations: `all-feat` comprises all features individually, while in `all-grp` we collapsed them into groups (see Table 5). As a morphological lexicon might not be available in all domains and languages, we also experimented with a feature combination `no-morph` that incorporates all the collapsed features from `all-grp` except the morphological derivation feature from group `VI`.

All results reported are statistically significant over the respective baseline (McNemar's test; $p < 0.05$).

[5]For these experiments we used WordNet 3.0.

	ATTR			REL			
	P	R	F	P	R	F	Acc
all-feat	0.95	**0.98**	0.96	**0.71**	0.56	0.63	0.93
all-grp	0.95	**0.98**	0.96	**0.71**	0.56	0.63	0.93
no-morph	**0.96**	0.96	0.96	0.67	**0.67**	**0.67**	0.93
Baseline	*0.90*	*1.00*	*0.95*	*0.00*	*0.00*	*0.00*	*0.00*

Table 6: Class-based precision and recall scores for the ADTree (DS1, cross-validation, unbalanced)

	ATTR			REL			
	P	R	F	P	R	F	Acc
all-feat	1.00	**0.93**	**0.97**	**0.94**	**1.00**	**0.97**	**0.97**
all-grp	1.00	**0.93**	**0.97**	**0.94**	**1.00**	**0.97**	**0.97**
no-morph	1.00	0.92	0.95	0.93	0.99	0.96	0.95
Baseline	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*

Table 7: Class-based precision and recall scores for the ADTree (DS1, cross-validation, oversampled)

	ATTR			REL			
	P	R	F	P	R	F	Acc
all-feat	**0.96**	**0.99**	**0.97**	0.79	**0.61**	0.69	**0.95**
all-grp	**0.96**	**0.99**	**0.97**	**0.85**	**0.61**	**0.71**	**0.95**
no-morph	0.95	0.96	0.95	0.56	0.50	0.53	0.91
Baseline	*0.90*	*1.00*	*0.95*	*0.00*	*0.00*	*0.00*	*0.00*

Table 8: Class-based precision and recall scores for the Boosted Learner (DS1, cross-validation, unbalanced)

6.1. Results on Data Set 1

We ran a first experiment on the heuristically annotated data set, using 10-fold cross validation. As the data is highly skewed towards the property-denoting class, we also created a balanced data set by random oversampling (Batista et al., 2004). The results on the balanced and the unbalanced data set are compared against a baseline classifier that always votes for the majority class.

Precision and Recall. Precision and recall figures achieved by the decision tree for both classes of adjectives are summarized in Tables 6 and 7. We observe very high precision values for the ATTR class on both the unbalanced and the balanced data set, while precision for REL adjectives is lower in both cases. By contrast, recall shows inverted performance, yielding a drop for the ATTR class and a strong increase for the REL class when switching from the unbalanced to the balanced data set, as is to be expected.

We obtain very high precision values, well above the baseline, for both classes when an equal number of training instances is provided. This indicates that our classification approach can be applied equally well for attribute and relation learning.

As our classification might also be useful for tasks different from ontology learning, we also report accuracy scores, as shown in the rightmost column of the tables. Comparing the performance on balanced and unbalanced data, we observe a slight increase of 0.03 points on average due to oversampling.[6]

In a comparison between the decision tree and the boosted learner (see Tables 6 and 8), we observe slight improvements for the ATTR class, but – more importantly – a considerable increase on the REL class when the `all-grp` combination is used with boosting. Apparently, this classifier benefits from collapsing individual features into groups, thus merging the values of sparse features. For this classifier, at least, the morphological feature provides valuable information, while the decision tree performs surprisingly well on the unbalanced set when this feature is omitted. Interestingly, this affects both classes, even though morphological derivation is the only positive feature we provided for the REL class.

In sum, our results indicate that automatically distinguishing property- and relation-denoting adjectives at the type level is possible with high accuracy, even on the basis of small training sets.

Type	**ATTR Tokens**	**REL Tokens**	**IMPOSS Tokens**
beautiful (ATTR)	50	0	0
black (ATTR)	35	7	8
bright (ATTR)	45	1	4
heavy (ATTR)	42	0	8
new (ATTR)	50	0	0
civil (REL)	0	49	1
commercial (ATTR)	5	44	1
cultural (REL)	2	48	0
environmental (REL)	0	48	2
financial (REL)	0	46	4

Table 9: Volatility of prototypical class members

Class Volatility. Yet, as discussed in Section 4.3, a type-based classification approach runs the risk of being affected by class shifts on the token level. This is not reflected by the evaluation carried out on the heuristically acquired corpus. In order to investigate the strength of this effect, we selected five adjective types of each class and inspected a random sample of 50 tokens for each type. As example cases, we chose types that were automatically classified with high confidence scores, since, at this point, we were particularly interested in the class change potential of prototypical class members.

The results of this investigation are shown in Table 9. The columns labelled with ATTR and REL display counts of tokens that matched one of our target categories, whereas the rightmost column subsumes all

[6] We obtained comparable results for balanced data sets created using random undersampling.

	ATTR			REL			
	P	R	F	P	R	F	Acc
all-feat	0.85	**0.82**	0.83	0.70	0.75	0.72	0.79
all-grp	**0.91**	0.80	**0.85**	**0.71**	**0.86**	**0.77**	**0.82**
no-morph	0.87	0.80	0.83	0.69	0.79	0.73	0.79
Baseline	*0.64*	*1.00*	*0.53*	*0.00*	*0.00*	*0.00*	*0.00*

Table 10: Class-based precision and recall scores for the Boosted Learner (DS2, test set)

	ATTR			REL			
	P	R	F	P	R	F	Acc
all-feat	**0.87**	0.76	**0.81**	**0.79**	**0.89**	**0.84**	**0.82**
all-grp	0.83	**0.77**	0.80	**0.79**	0.84	0.81	0.81
no-morph	0.80	0.71	0.75	0.74	0.82	0.78	0.77
Baseline	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*

Table 11: Class-based precision and recall scores for the Boosted Learner (DS2, training set, cross-validation, oversampled)

tokens that could not be assigned to the ATTR or REL class. The main reason that accounts for the majority of these cases are contexts where the adjective is part of a multi-word expression that does not elicit either a property or a relation, e.g. *black hole*, *look bright* or *heavy metal band*.

The average class volatility on the token level amounts to 8.6%. These figures can be considered as rough estimates for the average error that is introduced by raising our classification task to the type level. Still, our findings suggest that class volatility is not an issue that affects entire classes on a large scale, but seems to be limited to individual contexts.

6.2. Results on Data Set 2

With 246 property-denoting vs. 140 relational adjective types, the class distribution on DS2 is less skewed as compared to DS1. Further, DS2 offers sufficient training data for both classes. DS2 was therefore separated into training (80%) and test data (20%). The test set contains 49 property-denoting and 28 relational adjectives. In order to compare the results achieved on the test set against a balanced evaluation set, we also performed cross-validation on the training set after random oversampling.

On DS2, the boosted classifier yields the best results. Detailed figures are displayed in Tables 10 and 11. While all feature combinations perform well above the baseline, the `all-grp` combination achieves the best results for both classes. Considering all features without collapsing them into groups yields lower performance in general, except for recall on the ATTR class. Omitting the derivation feature leads to a slight decrease in performance.

Comparing the performance on the test set against cross-validation on the oversampled training set (see Table 11) shows consistent results. Even though the effect of oversampling is less prominent than on DS1, the impact of balanced class distributions in the training data is clearly observable.

The results on DS2 underline that property-denoting adjectives can be identified with high precision and decent recall. With regard to relational adjectives, we also observe highly satisfactory recall scores, while precision is lower, but still acceptable.

6.3. Discussion

Our experiments show good and consistent results on both DS1 and DS2. The pattern-based features we use for classification on the type level achieve high performance on the identification of property-denoting adjectives. Due to semi-supervised heuristic instance generation, the approach involves a moderate annotation effort. It should also be applicable to attribute learning in specialized domains, where no linguistic resources are available.

For the identification of relational adjectives, both classifiers perform robustly. Contrary to the attribute learning task, the applicability of our approach on relational adjectives benefits from external morphological resources and a sufficient amount of training data for this class.

Our type-based classification obviously runs the risk of being affected by type shifts on the token level. However, our findings on DS1, as well as empirical investigation of the annotated corpus suggest that class volatility is not an issue that affects entire classes on a large scale, but seems to be limited to individual contexts. This result is corroborated by examining WordNet. Analyzing the distribution of property-denoting and relational readings over the different word senses of adjectives in WordNet we found that 13.9% of all types exhibit volatile word senses that cannot be uniformly assigned a property-denoting or a relational reading. Even though this proportion is higher than the one we observed in our corpus, it is still tractable. This holds all the more as, in a random sample, an average of only 8.6% of the tokens of prototypical members of the ATTR and REL class were found to switch classes. By further investigation of classified data on the token level, we hope to obtain useful contextual features that are indicative for class shifts. This has to be left for future work, however.

7. Conclusions and Outlook

In this paper, we presented a semi-supervised machine learning approach for classifying property-denoting

vs. relational adjectives. This classification is a prerequisite for the task of learning attributes together with their values from text corpora. The results of our annotation experiment show that we can distinguish properties and relations in the denotation of adjectives at high performance levels, as long as the more fine-grained distinction between independent and founded properties is abstracted from. To compensate for sparse training data on the token level, we generated additional training instances in a semi-supervised manner, relying on observed low class volatility at the token level. Further performance improvements can be expected from contextual features that detect class changes on the token level. This issue needs to be addressed in future research. Another open issue concerns the feasibility of separating a third class of adjectives that are neither property- nor relation-denoting. As adjectives of this kind are too sparse in our annotated data and since they do not constitute a homogeneous class in WordNet, we could not investigate the problem in this paper. In future work, we will explore whether we can extend our approach towards a 3-way classification, using linguistic class descriptions offered by Amoia and Gardent (2008).

In summary, we consider our semi-supervised type-based adjective classification as an attractive method for supporting ontology learning in different languages or specialized domains, where appropriate lexical resources are not yet available.

8. References

Abdulrahman Almuhareb and Massimo Poesio. 2004. Attribute-based and Value-based Clustering. An Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Abdulrahman Almuhareb. 2006. *Attributes in Lexical Acquisition*. Ph.D. Dissertation, Department of Computer Science, University of Essex.

Marilisa Amoia and Claire Gardent. 2008. A Test Suite for Inference Involving Adjectives. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

R.H. Baayen, R. Piepenbrock, and L. Gulikers. 1996. *CELEX2*. Linguistic Data Consortium, Philadelphia.

Gustavo Batista, Ronaldo Prati, and Maria Carolina Monard. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations*, 6:20–29.

Kenneth R. Beesley. 1982. Evaluative Adjectives as One-Place Predicates in Montague Grammar. *Journal of Semantics*, 1(3):195–249.

Philipp Cimiano. 2006. *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*. Springer.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and Evaluating ukWaC, a Very Large Web-derived Corpus of English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Nicola Guarino. 1992. Concepts, Attributes and Arbitrary Relations. *Data & Knowledge Engineering*, 8:249–261.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, Ill.

Yusuke Miyao and Jun'ichi Tsujii. 2009. Supervised Learning of a Probabilistic Lexicon of Verb Semantic Classes. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, pages 1328–1337.

Richard Montague. 1974. English as a Formal Language. pages 247–270.

Victor Raskin and Sergei Nirenburg. 1998. An Applied Ontological Semantic Microtheory of Adjective Meaning for Natural Language Processing. *Machine Translation*, 13:135–227.

Gemma Boleda Torrent. 2006. *Automatic Acquisition of Semantic Classes for Adjectives*. Ph.D. Dissertation, Pompeu Fabra University, Barcelona.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, Cal., 2nd edition.