A cascade approach for complex-type classification

Lauren Romeo[1], Sara Mendes[1,2], Núria Bel[1]

[1]Universitat Pompeu Fabra, Roc Boronat, 138, Barcelona (Spain)

[2]Centro de Linguística da Universidade de Lisboa, Av. Prof. Gama Pinto, 2, Lisboa (Portugal)

{lauren.romeo, sara.mendes, nuria.bel} @upf.edu

Abstract

The work detailed in this paper describes a 2-step cascade approach for the classification of complex-type nominals. We describe an experiment that demonstrates how a cascade approach performs when the task consists in distinguishing nominals from a given complex-type from any other noun in the language. Overall, our classifier successfully identifies very specific and not highly frequent lexical items such as complex-types with high accuracy, and distinguishes them from those instances that are not complex types by using lexico-syntactic patterns indicative of the semantic classes corresponding to each of the individual sense components of the complex type. Although there is still room for improvement with regard to the coverage of the classifiers developed, the cascade approach increases the precision of classification of the complex-type nouns that are covered in the experiment presented.

Keywords: complex-types, nominal classification, lexical-semantic classes

1. Introduction

The automatic identification of complex-type nominals (Pustejovsky, 1995; 2005) using distributional information extracted from corpus data, besides contributing to a more accurate modelling of the lexicon by providing a method towards a cost-effective inclusion of complex-type information in Language Resources (LRs), can also provide useful and often crucial information to Natural Language Processing (NLP) applications by making available this type of semantic information in LRs.

Differing from simple-type nouns, complex types are composed of more than one constituent sense that can be recovered both individually and simultaneously in context.

(1) a. The church discussed its role in society at the gathering. (ORGANIZATION)
b. The choir rehearses on Saturdays at the church. (LOCATION)
c. There is a collection organized (ORGANIZATION) by the church on Mulberry Street (LOCATION) this Sunday.

In this example, the noun *church* in (1a) denotes an ORGANIZATION, in (1b) a LOCATION and in (1c) the context requires the same single occurrence of the noun to denote both an ORGANIZATION and a LOCATION. The complexity of the selectional behavior of complex types in context makes it difficult to apply the standard notion of word sense, as used in automatic text processing tasks, to them. Word Sense Disambiguation systems, for instance, might be able to correctly identify the senses in both (1a) and (1b), however in (1c) a decision for a single sense would have to be made, despite the fact that in this case both senses are simultaneously activated by the context.

In fact, information on the sense composition of complex types can be crucial in NLP, as it allows for the reduction of the amount of lexical semantic processing (Buitelaar, 2000) in tasks such as Information Retrieval, semantic role annotation, high-quality Machine Translation and Summarization, as well as Question Answering.

Having demonstrated the feasibility of automatically classifying complex-type nominals using distributional information in previous work (Romeo *et al.*, 2013), in this paper we show that our approach to this problem is robust enough to be used at production level, and thus has the potential to be incorporated in a method for cost-effective inclusion of information on sense composition of lexical expressions in LRs.

In the following sections, we review the motivation and theoretical background of this work (Section 2); discuss the data preparation, present our classification experiments (Section 3) and discuss the results obtained (Section 4); we then conclude our paper with some final remarks and directions for future research (Section 5).

2. Background

Most approaches in lexical semantic classification do not distinguish among related senses of the same word, considering it either as part of a class or not (Hindle, 1990; Bullinaria, 2008; Bel *et al.*, 2012). In previous research, we outlined a strategy in which we used distributional evidence for automatically identifying complex types, i.e. nouns that simultaneously belong to multiple classes (Romeo *et al.*, 2013).

We worked with two complex types in English – LOCATION•ORGANIZATION (LOC•ORG) and EVENT•INFORMATION (EVT•INF) – making apparent that complex-type nouns demonstrate characteristic and indicative lexico-syntactic traits of more than one class, which allows for their automatic identification using distributional evidence.

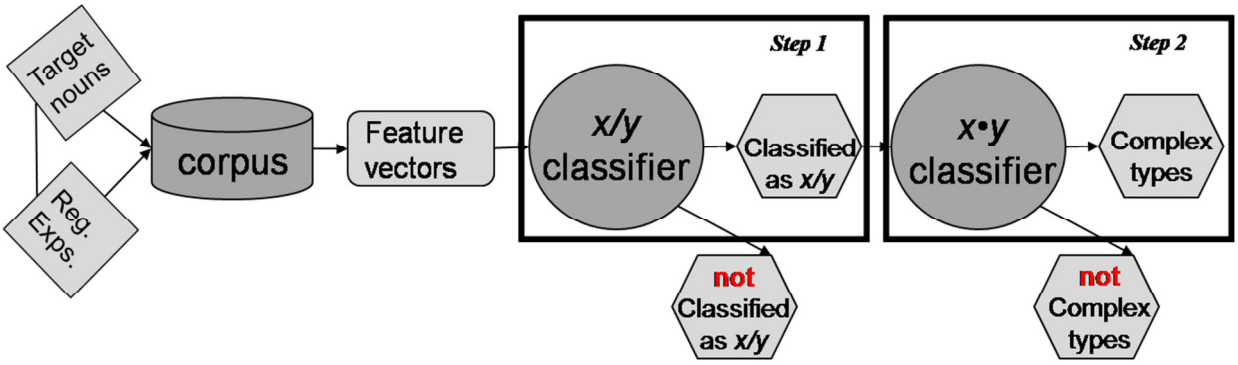

Figure 1: Workflow of the cascade experiment conducted for complex-type classification when no prior knowledge on the semantics of the nouns to be classified is available

Specifically, a cue-based lexical semantic classification methodology was applied to complex-type nominals obtaining an average performance of over 70% accuracy in distinguishing complex types from simple-type nouns belonging to semantic classes that correspond to any of the sense components of the former. However, this approach presupposes prior knowledge of whether nouns belong or not to what we will call the *x/y* group[1] – simple-type nouns from classes *x* and *y* and *x•y* complex-type nouns –, information which is not available in LRs. Overcoming this limitation is crucial for taking these classifiers from a purely experimental setup and making them usable in real scenarios.

In this paper, we aim at showing the feasibility of accomplishing this. To do so, we present an experiment showing how Romeo *et al.* (2013)'s approach can be used at production level, i.e. how it can be used when the task consists in identifying complex-type nominals from a given class by distinguishing them from any other noun in the language without any prior knowledge regarding the lexical semantic classes to which they belong. Accomplishing this requires extending the original approach (see Romeo *et al.* (2013) for a detailed description) to be able to not only separate complex-type nominals from simple-type nouns belonging to one of the classes corresponding to one of the sense components of the former, but to distinguish nouns belonging to a given complex-type class from any noun in the language, independently of the class to which they belong

As in Romeo *et al.* (2013), here we focus on two particular complex types[2] representative of the general characteristics of this type of nominal (Pustejovsky, 1995; 2005; Rumshisky *et al.*, 2007; Melloni and Jezek, 2009; Copestake and Herbelot, 2012):

ORGANIZATION•LOCATION ($\lambda x \bullet y \; \exists R \; [\alpha(\text{ORG}(x) \bullet \text{LOC}(y) \wedge R(x,y)]$): "the *church* prays during mass" vs. "the *church* is a large building"
EVENT•INFORMATION ($\lambda x \bullet y \; \exists R \; [\alpha(\text{EVT}(x) \bullet \text{INF}(y) \wedge R(x,y)]$): "the *interview* lasted for two hours" vs. "the *interview* was interesting"

Having verified the limitations of using an *n*-way classifier to accomplish our goal in the context of preliminary research leading up to the work presented here, we designed a cascade approach to our problem, by dividing it in 2 steps: (i) distinguishing *x/y* group nouns from any other noun in the language; and (ii) taking the nouns classified as belonging to the *x/y* group in Step 1 and distinguishing simple-type nouns from complex-type nouns.

This second step of the experiment corresponds to the task successfully performed in Romeo *et al.* (2013), with the sole difference that the classification is now performed on a potentially noisier dataset, as we are not classifying a controlled group of nouns, but rather the output of the first step of our cascade experiment. See Figure 1 for the workflow followed in the cascade experiments conducted, which is detailed in the following sections.

3. Experiments

In the context of the cascade experiment conducted to empirically demonstrate the feasibility of using the complex-type classifiers presented in Romeo *et al.* (2013) to identify complex types when no prior knowledge on the semantics of the nouns to be classified is available, we used distributional data gathered using lexico-syntactic patterns indicatory of the lexical semantic classes corresponding to the different sense components of the complex-type classes considered.

These lexico-syntactic patterns include information

[1] For the sake of simplicity, in this work, we will use the designation of *x/y* group to refer to the set of nouns consisting of simple-type nouns from classes *x* and *y* and complex-type nouns from the *x•y* complex type, e.g. LOC nouns, ORG nouns, and LOC•ORG nouns.

[2] The selection of these two complex-type classes was due to their wide inclusion in literature as representative of the general characteristics of complex-type nouns. Although the work presented in this paper regards the aforementioned two complex types, the approach can be extended to any other complex-type class, as the methodology we follow requires only the identification of lexico-syntactic patterns that are indicative of specific lexical semantic classes. Along this line, the methodology is class and domain-independent, relying only on the availability of the aforementioned lexico-syntactic patterns.

such as prepositions, selectional preferences, grammatical functions and morphological information (see Bel *et al.* (2012) for a full description), as illustrated by the examples in Table 1, which correspond to patterns considered per lexical semantic class.

The patterns are represented by regular expressions that indicate the entire lexico-syntactic context considered. For instance, as illustrated in Table 1, a nominal slot preceding a "creation" type verb in the past tense (x-VBD) is typically filled in by an ORG noun; a nominal slot preceded by a locative preposition (e.g. *inside*-IN) tends to be filled by a LOC noun; a target noun (x-NN) in the object position of a "transcribe" or "say" type verb (*submit*-V or *publish*-V, for instance) is generally an INF noun; while a target noun (x-NN) preceded by the preposition *during*-IN is typically an EVT noun.

In this work, we use these types of indicative contexts as cues for classification. As previously mentioned, the full description of all the cues considered per each of the simple-type classes we are working with can be found in Bel *et al.* (2012).

Class	Examples of lexico-syntactic patterns
ORG	**x-NN** (found|establish|organize)-VBD
LOC	(inside|outside)-IN (the|a|an)-(DT|Z) **x-NN**
INF	(submit|publish|report)-V* (the|a|an)-(DT|Z) **x-NN**
EVT	during-IN (the|a|an)-(DT|Z) **x-NN**

Table 1: Examples of lexico-syntactic patterns indicative of the lexico-semantic classes considered in this work.

The distributional information regarding these lexico-syntactic patterns, or cues, was extracted from a 60-million token PoS-tagged excerpt of the UkWaC corpus (Baroni *et al.*, 2009). To extract distributional information indicative of each *x/y* group, we combined the features indicative of class *x* with the features indicative of class *y*, i.e. we combined class-indicative features of LOC and ORG, in the case of the LOC/ORG classifier, and indicative cues for the EVT and INF classes in the case of the EVT/INF classifier.

The relative frequency of occurrence of each noun in each cue was stored in an n-dimensional vector, where *n* is the total number of cues used for each class. To classify, we used a Logistic Model Tree (LMT) (Landwehr *et al.*, 2005) Decision Tree classifier in the WEKA (Witten and Frank, 2005) implementation.

3.1 Data

For the experiments depicted in this paper we combined the gold standards used in Bel *et al.* (2012) for nominal classification and the gold standards specifically created in the context of our previous research on complex-type classification (Romeo *et al.*, 2013).

In order to train and evaluate the performance of the classifiers developed to identify complex-type nouns, information on the potential of a noun to be systematically interpreted in more than one sense (corresponding to the sense components of a complex type) was required. As this information is usually not included in LRs (see Boleda *et al.*, 2012), human annotations were used to create complex-type gold standards (see Romeo *et al.* (2013) for a detailed description on the construction of the complex-type gold standards). The constitution of the gold standards considered for our experiments is detailed in Table 2.

	Complex types	Simple types
ORG•LOC	79	184
EVT•INF	99	381

Table 2: Number of complex-type and simple-type nouns in the gold standards obtained by human annotation. This dataset was balanced with nouns annotated as neither LOC nor ORG, and EVT nor INF in the context of our experiments (see Section 3.2 for more details).

3.2 Defining training and test datasets

For the purpose of training a classifier and testing it with unseen data, thus emulating a real-life scenario, in which classification is to be performed on new data for which no prior semantic information is available, we divided our full dataset into training and test sets (70% for training and 30% for test) to evaluate the classifiers developed.

Having been built by human annotation, the datasets used in previous experiments (Romeo *et al.*, 2013) mirrored the unbalanced amount of complex and simple-type nominals in language[3]. Considering this, in the work presented here, we started by performing preliminary tests to verify which distribution of training data conveyed the best performing classifier, both in terms of discriminative power and accuracy on unseen data.

For training the *x/y* group classifiers, we experimented with two different data splits of the full dataset: one in which we used an equal amount of members of each of the classes considered for training; and another in which we used the same proportion of nouns from each class found in the original gold standard. In both cases, datasets were balanced with elements not belonging to the class, as mentioned earlier.

Considering both data splits, in this preliminary experiment the best classifiers were obtained when

[3] Moreover, as described and discussed in detail in Romeo *et al.* (2013), in the process of developing the complex-type gold standards, important asymmetries were observed in the ratio of nouns from each individual class corresponding to the sense components of a complex type that were annotated as having more than one potential sense. For instance, on the one hand, there were 9 EVT nouns that were tagged to have the potential to be interpreted as an INF noun, while on the other hand, there were 90 INF nouns that were considered to have the potential to be interpreted as an EVT noun.

Facts such as these point towards the possibility that complex-type sense components might not be equally prominent for a given complex-type noun or even complex-type class, which is naturally bound to be mirrored in distributional data. Although this is essentially an hypothesis which calls for further research, in Section 5 we will provide additional data supporting it, as we discuss the impact in our results of observed asymmetries in terms of the representativity, or frequency of occurrence, of distributional data indicative of specific sense components of complex types.

training with a balanced training set, i.e. when the classifiers were trained with an equal amount of nouns belonging to each of the classes to be identified by the classifier. Thus, the experimental results reported and discussed in the following sections are based on the results obtained when considering balanced datasets for training, whose constitution is presented in Table 3.

Balanced Datasets						
	LOC•ORG			EVT•INF		
type	S	C	not LOC/ORG	S	C	not EVT/INF
training	56	56	112	68	68	136
test	128	23	211	315	31	356

Table 3: Distribution of nouns in training and test datasets for the complex-type classes considered in this experiment. **C** corresponds to complex-type nouns, **S** to simple-type nouns either of class *x* or *y*, while the **not *x/y*** corresponds to nouns not belonging to the *x/y* group considered (see footnote 1).

3.3 The cascade experiment: complex-type noun classification in 2 steps

Having previously experimented with a single-step classification system, the results obtained made apparent that the nuanced distinctions a complex-type nominal classifier has to perform require a different approach. Complex types correspond to a very specific and complex linguistic phenomenon, with a strong impact in terms of semantic behavior in context, although observed in a limited amount of lexical items, facts which typically cause automatic systems to be unable to accurately model them. Specifically, the characteristic properties of this type of nouns causes their distributional data to be more disperse, besides partially overlapping with that of simple-type nouns corresponding to any of its sense components, which clearly raises problems to any automatic classification system.

All these observations led us to search for an alternative approach to the problem of complex-type nominal classification, namely the definition of a cascade approach with the potential of providing a partially filtered input to dedicated complex-type classifiers, thus allowing for better results.

Our expectations were confirmed by the results obtained and reported in this paper: on the one hand, the overall precision of a 2-step classification system significantly improves when compared with that of a single-step approach to this problem; on the other hand, the automatic filtering of information which is inherent to the cascade approach, although with a negative impact on recall, which is nonetheless not significant, as discussed in detail in the final sections of this paper, crucially contributes to an important noise reduction, and therefore to the reliability of the complex-type classification performed by the system.

This way, in order to consider the use of complex-type classifiers on a production level, we designed a 2-step cascade classification experiment, which is presented in detail below.

3.3.1. Distinguishing nouns in the *x/y* group from any other noun

The first step of the cascade experiment we propose consisted in training a classifier to distinguish nouns in the *x/y* group from nouns from any other class. Along this line, we consider all *x•y* complex-type nouns (either LOC•ORG or EVT•INF, in the case of the classifiers discussed in this paper), as well as simple-type nouns corresponding to the sense components of the complex-type class at stake (i.e. each of the components of the complex-type classes considered in this experiment: LOC and ORG; or EVT and INF), as members of the class. Thus, the goal of this step consists in coarsely distinguishing nouns belonging to the *x/y* group from nouns belonging to any other lexical semantic class.

To achieve this, we trained two classifiers, one for the LOC/ORG group and one for the EVT/INF group, with the LMT DT using 70% of our original dataset in a balanced selection of data, as detailed in the previous section (see Table 3). Each *x/y* group classifier model was then tested on unseen data (the remaining portion of the original dataset – cf. Table 3). Results obtained are detailed in Section 4.

3.3.2. Identifying complex-type nouns

The goal of the second step of the cascade experiment consisted in distinguishing *x•y* complex-type nouns from simple-type nouns, either from classes *x* or *y*, replicating the experiment presented in Romeo *et al.* (2013) on unseen and potentially noisier data. In this step, the output of the classification of the test set performed in Step 1, more precisely the nouns predicted to be members of the *x/y* group by the LOC/ORG and EVT/INF group classifiers, were then classified using a trained complex-type classifier, as proposed by Romeo *et al.* (2013)[4].

Thus, as the test set for this step, we used those nouns that were classified as members of the *x/y* class in Step 1, either correctly or not. Testing our complex-type classifiers with this information allows us evaluate their robustness in identifying complex-type nouns (i.e. in identifying, on the one hand, LOC•ORG nouns and, on the other hand, EVT•INF nouns), as they have to deal with a potentially noisier input consisting of *x/y* group nouns as identified by an automatic system whose average accuracy scores are in the mid-70% (see Table 4).

4. Results

Table 4 presents the results regarding the performance of the classifiers used in the cascade experiment, both with training and test data, and for the two complex-type classes considered.

The results obtained with the LOC/ORG group classifier, as well as the EVT/INF, i.e. the results obtained in the first step of the cascade workflow proposed, are

[4] To avoid the risk of overfitting, Romeo *et al.* (2013)'s classifiers were retrained, guaranteeing that there was no overlap between the test set used to evaluate the performance of the *x/y* group classifiers in the experiments depicted in this paper and the dataset used for training the complex-type classifiers.

consistent and promising: precision and recall are generally above 70% and there are no statistically significant differences[5] between the performance of the classifier in a 10-fold cross validation training setting and when the classification models are confronted with an input of unseen data. With regard to the results obtained in the second step of this experiment, further discussion is required. Section 5 focuses on this analysis of the results obtained, leading up to some final remarks and conclusions, which are presented in Section 6.

Step 1 of the cascade experiment: *x/y* group classification				
	LOC/ORG group classifier			
	accuracy	precision	recall	F-measure
training set	74.55%	0.751	0.746	0.744
test set	75.69%	0.755	0.757	0.754
	EVT/INF group classifier			
	accuracy	precision	recall	F-measure
training set	72.79%	0.729	0.728	0.728
test set	69.81%	0.711	0.698	0.697
Step 2 of the cascade experiment: complex-type classification				
	LOC•ORG complex-type classifier			
	accuracy	precision	recall	F-measure
training set	60.71%	0.607	0.607	0.607
test set	57.14%	0.877	0.571	0.667
	EVT•INF complex-type classifier			
	accuracy	precision	recall	F-measure
training set	59.56%	0.597	0.596	0.594
test set	56.69%	0.905	0.567	0.667

Table 4: Performance of classifiers in Step 1 and Step 2 of the cascade experiment on the classification of complex-type nouns with training and test datasets.

5. Discussion

As was to be expected, the performance of the complex-type classifiers in the training setting is consistent with the results reported in Romeo *et al.* (2013). Though slightly lower on the test setting, there is no statistically significant difference in the overall performance of the complex-type classifiers in the training and test settings, i.e. in a 10-fold cross validation setting and when used to classify the output of either the LOC/ORG or the EVT/INF group classifiers.

The considerably higher precision (statistically significant) of the complex-type classifier in the test setting, i.e. when used to classify unseen data, when compared with the results obtained in the training setting has, nonetheless, to be underlined and commented upon.

Overall, this seems to indicate that, not only can the complex-type classifier successfully handle instances corresponding to noise proceeding from the first step of the cascade experiment, ubiquitous in any production-level scenario, but also that, although one of the concerns with using a cascade approach was the possibility of error accumulation, the results obtained, and the significant increase in precision in the classification of complex-type nouns in particular, point towards the opposite. More precisely, these results indicate that the information provided to the *x•y* classifiers is somehow "cleaner".

Our hypothesis for explaining these results is that problematic cases are being filtered out in the first step of the cascade experiment, thus providing a cleaner input to the *x•y* complex-type classifier in Step 2, a fact that we will try to confirm via a detailed analysis of the results obtained, which is presented below. As the classifiers in Step 2 take as input only those nouns classified to be members of the *x/y* group in Step 1, if potentially problematic cases are filtered out by not being classified as members of the *x/y* group class, a natural consequence is the observed increase in precision (see Section 5.1 for a detailed discussion on the impact on precision of a cascade approach to complex-type classification such as the one proposed in this paper).

Moreover, it is interesting to note that this increase in precision does not have a relevant impact on recall: although the scores are slightly lower in the test setting, the difference between the recall scores in the training and test settings is not statistically significant. In Section 5.2 we provide more details regarding the performance of our classifiers in terms of recall and discuss some aspects that might be addressed in future work to improve it.

5.1 Complex-type classification of automatically identified *x/y* group nouns: impact on precision

In order to evaluate to which extent the first step of our cascade workflow is actually filtering out potentially problematic cases, we re-ran our complex-type classification with the full test set, i.e. as if the first step of the cascade workflow was performing with an accuracy of 100%, and compared the results obtained.

By doing this, we aim at identifying which types of nouns are being eliminated in the first step of our cascade experiment so as to verify whether the candidates that we are losing would be correctly dealt with by our complex-type classifiers. In this context, we observed that an important part of the nouns being eliminated in Step 1 are nouns that occur in corpus data with a low frequency. On the one hand, in the case of the EVT/INF classifier, 7 of the 11 nouns misclassified as not belonging to the EVT/INF group, and thus not included in the set of candidates provided as input to the EVT•INF complex-type classifier in Step 2, occurred with an absolute frequency of less than 200 times in the corpus. In fact, of those 7 nouns, 5 occurred with an absolute frequency of less than 20.

On the other hand, in the case of the LOC/ORG class, the absolute frequency of 6 of the 12 misclassified nouns not considered to belong to the LOC/ORG group was lower than 200 occurrences in the corpus, the absolute frequency of 3 of those 6 nouns being lower than 20. Thus, a large part of the misclassifications observed in the first step of our cascade experiment is due to their low

[5] In this work, statistical significance was calculated using Student's t-test with a 95-percent confidence interval.

frequency of occurrence in corpus data, which is bound to also affect the classification of these nouns as complex types.

In order to confirm this, as mentioned above, we submitted all the nouns misclassified in the first step of the cascade workflow to the complex-type classifiers in the second step to verify to which extent these are able to successfully classify such candidates. We obtained the following results: in the case of the EVT•INF class, 9 of the 11 nouns eliminated in the first step of our cascade experiment would be misclassified in Step 2 if they were to arrive to this step of the experiment, the 7 low-frequency nouns mentioned above being among these 9. In the case of the LOC•ORG class, the same would happen to 6 of the 12 nouns lost in Step 1, the overlap between the set of low-frequency nouns and that of misclassified complex-type nouns by the LOC•ORG classifier being perfect. These data make apparent that the increase in precision in Step 2 is directly explained by the fact that low-frequency complex-type nouns, which are problematic to any classifier due to the sparseness of the distributional information provided for classification, are being filtered out in Step 1, thus providing a cleaner amount of candidate nouns to be considered by the *x•y* classifier in Step 2, which therefore directly impacts its performance in terms of precision.

Naturally, not all of the nouns removed in Step 1 are necessarily problematic. For instance, we observed 2 cases of EVT•INF nouns and 6 cases of LOC•ORG nouns that are incorrectly classified in Step 1, and therefore not considered in Step 2, although they would be correctly classified at this stage, thus reducing the coverage of our classifiers, even if not at a statistically significant level (see Section 5.2).

But this is not the only aspect determining the scores of our classifiers in terms of recall, which is clearly the less strong aspect of the classifiers developed. In order to further understand what is impacting the recall scores obtained in complex-type classification we conducted an error analysis, which we discuss in detail in Section 5.2, focusing on those nouns that were misclassified as not belonging to a complex type by our classifiers.

5.2 Analyzing the recall of *x•y* classifiers

In the case of the EVT•INF classifier, the final results obtained in Step 2 demonstrate 5 incorrectly classified complex-type nouns, which are considered to be non-members of the EVT•INF class by our classifier. Of these 5 cases, one, the noun *newsflash*, is caused by insufficiency of distributional information (15 occurrences in total of this noun in corpus data). Due to its low frequency in our data, this noun only co-occurs 3 times with our class-indicative lexico-syntactic patterns, not providing sufficient information for the complex-type classifier to arrive at an accurate class membership decision for that particular noun. We have to underline that the complex-type classifiers in Step 2 must make more nuanced decisions, distinguishing between complex-type nouns, that should display characteristic features of both class *x* and class *y*, and simple-type nouns belonging to either class *x* or class *y*, which makes the availability of sufficient class-indicative distributional information all the more important.

As to the 4 remaining cases of incorrectly classified complex-type nouns, their misclassification cannot be attributed to low frequency, as these nouns have an absolute frequency of 190, 3881, 1779 and 538 times in the corpus. However, when looking into their individual feature vectors, we observed that the information being provided to the classifiers demonstrates a considerable asymmetry in terms of the frequency of use in language data of the different sense components of the complex type. When considering the distributional data as represented in the feature vectors of each of these complex-type nominals, we observed, for instance, that in the case of the complex-type noun *notice* 613 of its 697 co-occurrences with class-indicative lexico-syntactic patterns corresponded to features that are indicative of the INF class, while only 42 were indicative of the EVT class, and another 42 occurrences corresponded to information that was meant to be negative, i.e negative cues[6]. This same trend was also observed with the EVT•INF noun *quote* for which 123 of its 130 occurrences were in INF-indicative patterns, only 5 being in EVT-indicative patterns, and 2 in negative cues.

This way, we attribute misclassification in these cases to a lack of homogeneity in the representativity in corpora data of the features indicative of the different sense components of these particular lemmas. This point is further verified by the fact that these lemmas are correctly classified in Step 1 as members of the EVT/INF group, as this classifier is trained to identify nouns from each of the individual simple-type classes corresponding to the different sense components of a given complex type. Thus, even though there is an asymmetry in the frequency of use in language data of the distributional information represented in the feature vector of a complex-type noun provided to our classifiers, which causes its misclassification as a non-member of the *x•y* complex type, these nouns are not filtered out in the first step of our experiment as they have a significant number of features in common with nouns of one of the simple-type classes being considered by the classifiers in this step, and are therefore correctly classified as members of the *x/y* group.

In the case of LOC•ORG, the final results obtained in Step 2 demonstrate 4 incorrectly classified complex-type nouns, which were considered not to belong to the LOC•ORG class by our classifier. Of these 4 cases, one is caused by insufficiency of distributional information (23

[6] Following Bel *et al.* (2012), cues that were expected to be negative for the classes considered in this work were included in the set of lexico-syntactic patterns considered and provided to our classifiers. These are typically positive and very marked cues for other lexical semantic classes which are included as an attempt to capture correlations with other marks that separate class members from the non-members, and this way expected to contribute to a better partition of the classification space. As Bel *et al.* (2012) we will designate this type of distributional information as negative cues.

occurrences in total of this noun in corpus data) while the remaining three cases also displayed an asymmetry of occurrences in class-indicative lexico-syntactic patterns of the different sense components of the LOC•ORG complex type.

In the case of the LOC•ORG noun *borough*, 37 of its 54 occurrences in class-indicative lexico-syntactic patterns are indicative of the LOC class, while only 14 of its occurrences are class-indicative features for the ORG class, and 3 correspond to negative cues. This same trend was also observed with the LOC•ORG noun *unit*, for which 339 of its 387 occurrences corresponded to features considered indicative of the LOC class, while only 38 were features considered indicative of the ORG class, and 10 amounted to negative cues. The same was also true for the LOC•ORG noun *agency*, which has 189 of its 286 occurrences in features indicative of the LOC class and only 33 in features of the ORG class, while 14 occurrences corresponded to negative cues.

These examples serve to demonstrate the impact that asymmetry in the frequency of use of different sense components of a complex-type noun can have on results. In fact, although theses nouns are considered to be complex-type nominals in our gold standard, their distributional data is heavily biased towards one of the two sense components of the complex type. Considering this, future work should evaluate the possibility of devising strategies which attempt to smooth this bias in the distributional information extracted for this type of noun, in order to improve the accuracy of classifiers developed, specifically in terms of recall in classification results, by reducing the amount of false negatives.

6. Final Remarks

The cascade classification experiment depicted in this paper demonstrates that we can obtain state-of-the-art results (Bel *et al.*, 2012; Romeo *et al.*, 2013) when running a complex-type classification on a dataset of unseen nouns made up of nominals belonging to any lexical semantic class. Overall, our classifier successfully identifies very specific and not highly frequent lexical items such as complex-type nominals with high accuracy, and distinguishes them from those instances that are not complex types using lexico-syntactic patterns indicative of each of the individual classes corresponding to the different sense components of a complex type. Although there is still room for improvement with regard to the coverage of the classifiers developed, as discussed in detail in the paper, when compared to previous work (Romeo *et al.*, 2013), the cascade approach increases the precision of the complex-type nominal classification, even when no information on the lexical semantic class of the candidate nouns to be classified is available.

As our classifier is able to distinguish complex-type nouns from any other noun in the language without requiring any prior knowledge on its semantic properties, specifically on it belonging to any specific *x/y* group, as was the case in the approach depicted in Romeo *et al.* (2013), it has the potential to be used as a tool for production level setups and, therefore, be useful for a wide range of NLP systems and applications.

Results show that the classifiers built are able to accurately distinguish non-members of the complex types considered from elements belonging to these classes, although there are still important issues regarding the coverage of the complex-type classifiers, specifically due to the challenge of handling asymmetries in terms of the frequency of use in language data of different sense components of complex-type nouns. Far from being a trivial issue, future steps should include research on strategies to minimize the impact of this phenomenon in classification results.

7. Acknowledgements

This work was funded with the support of the SUR of the DEC of the Generalitat de Catalunya and the European Social Fund, by SKATER TIN2012-38584-C06-05, and by Fundação para a Ciência e a Tecnologia (FCT) post-doctoral fellowship SFRH/BPD/79900/2011.

8. References

M. Baroni, S. Bernardini, A. Ferraresi & E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3): 209-226.

N. Bel, L. Romeo & M. Padró. 2012. Automatic Lexical Semantic Classification of Nouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey: 1448-1455.

P. Buitelaar. 2000. Reducing lexical semantic complexity with systematic polysemous classes and underspecification. In *Proceedings of NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in NLP Systems*: 14-29.

J. A. Bullinaria. 2008. Semantic Categorization Using Simple Word Co-occurrence Statistics. In M. Baroni, S. Evert & A. Lenci (eds.) *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics,* Hamburg, Germany: 1-8.

A. Copestake & A. Herbelot. 2012. *Lexicalised compositionality*. Unpublished draft.

D. Hindle. 1990. Noun classification from predicate argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*: 268-275.

N. Landwehr, M. Hall & E. Frank. 2005. Logistic Model Trees. *Machine Learning*, 95(1-2): 161-205.

C. Melloni & E. Ježek. 2009. Inherent Polysemy of Action Nominals, presented at *Journées Sémantique et Modélisation (JSM 2009)*, Paris, France.

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

J. Pustejovsky. 2005. *A survey of dot objects*. Unpublished manuscript, Brandeis University, Waltham.

L. Romeo, S. Mendes & N. Bel. 2013. Towards the automatic classification of complex-type nominals. in *Proceedings of GL 2013 – 6th International*

Conference on Generative Approaches to the Lexicon, Pisa, Italy.

A. Rumshisky, V. Grinberg & J. Pustejovsky. 2007. Detecting Selectional Behavior of Complex Types in Text. In *Proceedings of GL 2007 – 4th International Workshop on Generative Approaches to the Lexicon*, Paris, France.

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