Towards the automatic classification of complex-type nominals

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 presented here depicts experiments toward the automatic classification of complex-type nominals using distributional information. We conducted two experiments: classifying complex-type nominals as members of multiple individual lexical classes, and building a dedicated classifier for complex-type nominals, distinguishing them from simple types. We discuss the promising results obtained, with a focus on asymmetries observed and on lines to be explored in the future.

1 Introduction

In this article we evaluate the possibility to automatically identify dot-type nominals using distributional information extracted from corpus data. This work has a two-fold motivation. First, to contribute to a more accurate modeling of the lexicon, by providing a method towards a cost-effective inclusion of dot-type information in Language Resources (LRs), which will thus mirror a complex, systematic and productive linguistic phenomenon. Second, to make this type of semantic information available in LRs to provide useful and often crucial information to Natural Language Processing (NLP) applications.

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, as illustrated below.

(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 dot-object selectional behavior in context, as illustrated in (1), makes it difficult to apply to complex types the standard notion of word sense, as used in automatic text processing tasks. Traditional word sense disambiguation (WSD) 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 both senses are simultaneously activated by the context.

Having rich information available on complex types not only can reduce the search space in disambiguation tasks, and thus the number of decisions needed, but can also provide grounds to opt for the non-disambiguation of instances when relevant, for example in co-predication contexts like (1c). Moreover, knowledge of the entire sense potential of a given word is sometimes required for specific tasks (see for instance Rumshisky et al. (2007) and Lenci et al. (2010)).

Thus, 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.

In this paper we evaluate the possibility to employ information from actual language use as encoded in corpus data to acquire information on the sense composition of complex types. In line with approaches that explore corpus-based definitions of fine-grained distinctions that emerge as abstractions over the combinatorial patterns of lexical items (Ježek and Lenci, 2007), we use a classification approach based strictly on distributional evidence available in a corpus to automatically identify complex types.

As 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), our goal is to outline a strategy which automatically accounts for those nouns that belong to multiple classes, specifically to pinpoint complex-type nouns using distributional evidence. In this context we discuss an experiment involving two complex types in English: LOCATION•ORGANIZATION (LOC•ORG) and EVENT•INFORMATION (EVT•INF). Our hypothesis is that complex-type nouns demonstrate characteristic and indicative lexico-syntactic traits of more than one class, which allow us to use lexico-syntactic patterns over corpus data to automatically identify nouns for which there is distributional evidence of their membership to more than one class.

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

2 Motivation and theoretical background

2.1 Complex types

Dot objects, or nouns with complex types, are composed by more than one constituent type, each representative of a distinct sense, between which holds a regular and predictable relation. As thoroughly discussed in the literature (Pustejovsky, 1995; 2005), there is strong linguistic motivation for considering the existence of such objects. First, the knowledge we have of concepts associated with *books* and *doors*, for instance, is not characterizable as a conjunction of simple types. Second, the notion of complex types captures a type of inherent logical polysemy, occurring in regular, predictable patterns, i.e. systematically recurrent, namely cross-linguistically.

Building on arguments that show traditional sense-enumerating lexicons are not only uneconomical, but also present instances of systematic phenomena as arbitrary and idiosyncratic features of single words, which do not account for the productive nature of their potential underlying regularities (Pethrö, 2000), and thus render unfeasible the task of listing all possible meanings of a word (Kilgariff, 1992), the Generative Lexicon Theory (GL) (Pustejovsky, 1995) explores and formalizes the shifts of meaning of these objects in context. This represents an important step towards implementing systems that can assign meaning to words dynamically depending on the context in which they occur (Cooper, 2005).

Here we assume Pustejovsky's (1995) definition of dot types as a Cartesian product of types with a particularly restricted interpretation. This means that the product $\tau_1 \times \tau_2$, of types τ_1 and τ_2, each denoting sets, alone does not adequately determine the semantics of the dot object. The relation *R*, which structures the component types, must also be seen as part of the definition of the semantics of the lexical conceptual paradigm of the complex type. Thus, for the dot object $\tau_1 \bullet \tau_2$ to be well-formed, there must be a relation R that structures the elements τ_1 and τ_2, a concept that is formalized in GL (Pustejovsky, 1995: 149) as:

$$\begin{bmatrix} \alpha \\ ARGSTR = \begin{bmatrix} ARG1 = \mathbf{x}: \tau_1 \\ ARG2 = \mathbf{y}: \tau 2 \end{bmatrix} \\ QUALIA = \begin{bmatrix} \tau_1 \tau_2 \_\mathbf{lcp} \\ FORMAL = R(x,y) \end{bmatrix} \end{bmatrix} \quad (2)$$

This formalization accounts for one of the properties that makes complex types unique and distinguishes them, for instance, from cases of homonymy[1]: the possibility for their distinctive senses to be active at the same time (Pustejovsky, 1995: 223), illustrated in (1c). The levels of representation and generative mechanisms in GL predict a noun like *church*, represented below, occurs not only in contexts typical of class x: ORG (see (1a)) and of class y: LOC (see (1b)), but also in contexts which activate the relation $R_1(x,y)$, i.e. contexts where both ORG and LOC senses are simultaneously activated (see (1c)).

$$\begin{bmatrix} \mathbf{church} \\ ARGSTR = \begin{bmatrix} ARG1 = \mathbf{x}: \mathbf{organization} \\ ARG2 = \mathbf{y}: \mathbf{location} \end{bmatrix} \\ QUALIA = \begin{bmatrix} \mathbf{org} \bullet \mathbf{loc\_lcp} \\ FORMAL = R_1(x,y) \end{bmatrix} \end{bmatrix} \quad (3)$$

These properties distinguish dot objects from simple types, unified types or standard generalization on types (cf. Pustejovsky, 1995: 141 and ff.). Moreover, the possibility to have word

[1] Utt and Padó (2011) consider the importance of this distinction, proposing an automatic polysemy classifier. Boleda et al. (2012) also put forth an approach for predicting regular sense alternations in corpus data. However, both methods are based on external rich language resources, which besides only being available for a very restricted set of languages, do not necessarily mirror language use, as noted in the latter work.

senses that semantically compose these words either individually or simultaneously activated, depending on the selectional environment, presents a challenge to NLP systems that deal with identifying word senses in context. In fact, these follow a one-word, one-sense approach, designed to identify a single sense in each decision. Thus, as argued in Section 1, including information on the semantics of dot objects in LRs can contribute to an overall improvement in performance of NLP systems.

2.2 Exploring the Distributional Hypothesis to identify complex-type nouns

Considering the above characterization of dot types, we assume them to be members of more than one lexical class, more precisely members of each class corresponding to the senses they are composed of. As members of more than one class, complex types are expected to occur in indicatory contexts of more than one individual class. With this in mind, we evaluate the possibility to automatically identify complex types using a cue-based classification methodology.

Based on the Distributional Hypothesis (Harris, 1954), cue-based lexical semantic classification (Merlo and Stevenson, 2001) builds on the assumption that lexical semantic classes are emergent properties of a number of words that recurrently co-occur in a number of particular contexts. Thereby, as proposed by Bybee and Hopper (2001) and Bybee (2010), we understand lexical semantic classes as generalizations that come about when there is a systematic co-distribution for a number of words in a number of contexts. Different contexts where a number of words tend to occur thus become linguistic *cues* of a particular semantic property that a set of words has in common. Using these cues to gather indicatory distributional information provides evidence that discriminates members of a class from other lexical items.

We hypothesize that the classification of a noun as a member of the different individual classes that correspond to the senses that compose a complex type indicate its potential to belong to a given dot type. Parting from the cue-based nominal lexical semantic classification work reported in Bel et al. (2012), we apply this methodology to complex-type nominals. This allows us to analyze the distributional behavior of nouns belonging to more than one class and to which extent binary classifiers can accurately deal with such items.

As members of more than one class, we expect complex-type nouns to disperse their occurrences between indicatory contexts of different classes. Thereby, one of our goals consists in evaluating to which extent this can be problematic to binary classifiers. Specifically, we will verify whether the available distributional information indicatory of each individual class is strong enough for an automatic cue-based classification for this type of noun to work.

3 Data preparation

The sense composition of complex types discussed in previous sections forms the basis of our hypothesis in which we claim that these nominals should exhibit linguistic behavior characteristic of each simple-type class that makes up their sense composition. To verify this hypothesis and thus provide empirical evidence of multiple class membership for complex-type nouns, we implemented the cue-based lexical semantic classification experiment described below.

3.1 Classes considered

In line with the argument presented above, we focus on two complex types representative of the general characteristics of dot objects (Pustejovsky, 1995; 2005; Rumshisky et al., 2007; Melloni and Ježek, 2009; Copestake and Herbelot, 2012):

ORGANIZATION·LOCATION (λx•y ∃R [α (ORG(x)•LOC(y)∧R(x,y)]): "the *church* prays during mass" vs. "the *church* is a large building"

EVENT·INFORMATION (λx•y ∃R [α (EVT(x) •INF(y)∧R(x,y)]): "the *interview* lasted for two hours" vs. "the *interview* was interesting"

3.2 Description of the gold standard

In their nominal classification experiments, Bel et al. (2012) used gold standards created by extracting nouns from WordNet (Miller et al., 1990) which contained a sense corresponding to each of the lexical classes they studied. As our aim in this work is to automatically identify which nouns are complex-type nominals, we needed gold standards composed of nouns with the potential to be systematically interpreted in more than one sense to evaluate the results obtained in our experiments. As this information is usually not included in LRs, and specifically in Wordnet (see Boleda et al., 2012), we resorted to human annotation to create the gold standards.

Three experts, either native or highly proficient English speakers, annotated each noun from the original Bel et al. (2012) lists for their

potential to contain another known sense. The annotators were given the automatically extracted list of nouns from each class and were asked to annotate whether those nouns could have a specific sense, different from the one encoded in the original gold standard.

Being simply provided with the original gold standard lists and a general definition of a target sense, annotators were asked to mark with *yes* or *no* whether they thought each individual noun in the list could be interpreted as a member of the target class, besides potentially having any other sense. With this annotated information, we used a voting scheme to build the gold standard, including in it the nouns considered to be members of more than one class by at least two annotators.

3.2.1 Asymmetry of sense components

Previous work has reported asymmetries regarding the prominence of senses that compose complex types (see, for example, Rumshisky et al. (2007) and Ježek and Melloni (2011)), as one sense is more generally used or constitutes a preferred interpretation[2]. Confirming this observation, evidence from psycholinguistic studies (Frisson and Pickering, 2001) demonstrated that although more than one sense interpretation is available for a given word, the vast majority of speakers tend to consistently choose one interpretation over the other.

Several authors established relations between this type of asymmetry and complex types, particularly with regard to the nature of the relations holding between their sense components. An important part of the work developed on this matter has focused on classes whose sense components are ontologically related, in particular on the PROCESS•RESULT complex-type.

Ježek and Melloni (2011) characterize the properties of the polysemy involved in this case arguing it arises from the fact that a RESULT object type is temporally and causally dependant on a PROCESS type as an event is the pre-condition for the (coming into) existence of the object (RESULT). Thus, PROCESS readings can be considered more prominent as they are also reflected when the RESULT sense is active while the reverse does not hold true. The EVT•INF complex type, can be considered a sub-case of the former. Formalized in (4), the aforementioned unique properties of this dot type are represented in the AGENTIVE role.

$$\begin{bmatrix} \textbf{illustration} \\ ARGSTR = \begin{bmatrix} ARG1 = \textbf{x : event} \\ ARG2 = \textbf{y : information} \end{bmatrix} \\ QUALIA = \begin{bmatrix} \tau_1 \bullet \tau_2 \_ \textbf{lcp} \\ FORMAL = R\,(x,y) \\ AGENTIVE = x(z,y) \end{bmatrix} \end{bmatrix} \quad (4)$$

Just as is the case for PROCESS•RESULT nominals, we expect the prominence of senses for this complex type to be asymmetric. The data obtained in our annotation task are consistent with this expectation (see Table 1), as 90 of the 149 INFORMATION nouns in Bel et al.'s (2012) gold standard are considered to also have an EVENT sense, whereas only 9 of the 273 EVENT nouns are annotated as also having an INFORMATION sense. Moreover, these human annotation results constitute a source of quantitative information providing evidence that support the existence of asymmetries of prominence of the different sense components of complex types.

	# of complex types per class	ratio of complex types per class
ORG as LOC	38	0.28
LOC as ORG	46	0.37
INFO as EVT	90	0.60
EVT as INFO	9	0.03

Table 1. Distribution of dot types per lexical class

Regarding the LOC•ORG complex type, there is neither an ontological relation between its meaning components nor such a clear asymmetry in the prominence of its sense components. Yet, differences observed can be attributed to relations generally holding between objects in the world. For instance, an ORGANIZATION, as a more abstract concept, is typically associated to a physical reality, namely the LOCATION which hosts this abstract object and makes it "perceivable". Reversely, LOCATION, as a physical point in space, is often independent of any other reality. Thus, in the lexicon, we observe words primarily denoting an ORGANIZATION that also refer to the LOCATION that hosts it, whereas the reverse is observed only in considerably stricter conditions, as illustrated by *congress* and *schoolyard* in (5).

(5) a. The *congress* (ORG) decided to vote the new rule into power after the recess.
b. The new rule was voted to power in the *congress* (ORG or LOC).
c. #The *schoolyard* decided to vote the new law into power after the recess.
d. The new rule was voted to power in the *schoolyard* (LOC).

[2] As often discussed in the literature (e.g. Bybee, 2010), these two aspects are not independent from each other: frequency of use tends to impact preferred interpretations. This is nonetheless a debate outside the scope of this work.

Asymmetry in the prominence of complex-type sense components is thus related to the nature of the systematic relation holding between them, which is different for each complex-type paradigm. Moreover, the ratio of nouns in each individual class annotated as having more than one potential sense, makes apparent the representativity of this phenomenon for each class (see Table 1). This provides crucial insight when analyzing our results, particularly to evaluate whether the asymmetries reported in this section have an overall impact in the automatic identification of complex types.

4 Experiments

In our experiments, we considered English nouns from the LOCATION, ORGANIZATION, INFORMATION, and EVENT classes. We used a part of the UkWaC corpus (Baroni et al., 2009) consisting of 60 million PoS-tagged tokens. To gather distributional evidence, we employed lexico-syntactic patterns indicatory of each individual class including prepositions, selectional preferences, grammatical functions and morphological information (see Bel et al. (2012) for a detailed description of patterns used). Each pattern was translated into a regular expression used over the corpus to identify occurrences of nouns in marked contexts. 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 Trees (LMT) (Landwehr et al., 2005) Decision Tree classifier in the WEKA (Witten and Frank, 2005) implementation.

As detailed in Section 3.2, our gold standards are derived from the lists used by Bel et al. (2012) to reflect the phenomenon of multiple class membership of complex types. As there is a larger ratio of simple types in language, which is mirrored in our gold standards (see Table 2), a baseline based on the majority class would not allow us to assess the quality of the results depicted here. Thereby, to evaluate our results, we compare them against the performance of state-of-the-art classifiers for simple types, reported in Bel et al. (2012).

4.1 Complex types as members of individual simple-type classes

As mentioned earlier, the basic hypothesis for our experiments is that complex-type nominals, as members of more than one lexical class (see Section 2 for more details), demonstrate characteristic lexico-syntactic traits of multiple classes, and thus occur in indicatory contexts of the different classes that correspond to their sense components. However, as members of more than one class, the distributional behavior of complex-type nouns is expected to be more disperse, as occurrences are divided between indicative contexts of different classes. Given this, the experiment reported in this section aims to provide evidence as to whether this distributional information, though disperse, is strong enough to allow for an automatic identification of the different sense components of complex types in a classification task.

To accomplish this, we used the binary classifiers described in Bel et al. (2012), which were developed to automatically classify nouns into previously known lexical semantic classes, not taking into consideration polysemy. Based on word occurrences in specific contexts in a corpus, these classifiers simply consider a given noun either as a member of a class or not.

In the experiment reported in this section, we used a binary classifier for each sense component (organization, location, event and informational object) of the complex types considered. We started by verifying the binary classifiers capacity to identify complex-type nouns as members of the class corresponding to their most prominent sense, indicated in bold in Table 2.

	complex types correctly classified as members of the class (%)	ratio of classified complex types per members of the class
ORG•LOC as ORG	58.69	0.22
ORG•**LOC** as LOC	89.47	0.25
EVT•**INFO** as INFO	71.11	0.43
EVT•INFO as EVT	77.78	0.03

Table 2. Complex types correctly identified as members of the class corresponding to their prominent sense

The results reported in Table 2 make apparent that dot-type nominals provide enough distributional evidence indicatory of their most prominent sense so that their automatic classification as members of the class it corresponds to is possible. The results obtained are actually in line with the performance of the same classifiers with simple-type nominals reported by Bel et al. (2012), where a 66.21% and a 73.05% accuracy are obtained respectively for the LOCATION and the EVENT nouns classifiers.

With this in mind, we proceeded to verify whether this is also observed when considering less prominent sense components by performing a cross-classification of the nouns in our study using the binary classifiers mentioned above, essentially emulating the human annotation task described in Section 3.2. More precisely, we used trained bi-

nary classifiers for each class to classify the human-annotated lists of nouns, i.e. each classifier trained for simple-type classification of nouns of semantic type τ_1 was provided with a list of nouns with τ_2 as their prominent sense.

To illustrate this, a noun like *church*, defined as a LOCATION (τ_1) in Bel et al.'s (2012) gold standards, was checked for its occurrence in lexico-syntactic patterns indicatory of ORGANIZATION (τ_2) nouns, i.e. whether it shows distributional evidence indicatory of another class. Our claim is that having τ_1 nouns that occur in contexts indicatory of τ_2 allowing them to be classified as members of τ_2 provides evidence toward our hypothesis: given the sense composition of complex types, they should be considered members of more than one lexical semantic class, a fact that automatic classifiers should account for.

Table 3 presents the results of precision and recall of the cross-classification of complex-type nouns as members of the class corresponding to non-prominent sense components, in bold.

	Precision	Recall	Ratio
ORG•**LOC** as LOC	77.78	15.21	0.06
ORG•LOC as ORG	57.14	21.05	0.06
EVT•INFO as EVT	64.44	32.22	0.19
EVT•**INFO** as INFO	6.67	66.67	0.03

Table 3. Results of cross-classification (in %)

With our cross-classification, we replicate the annotation task automatically (see Section 3.2). The results in Table 3 allow us to make three main observations. First, the performance of cross-classification is in line with that of the classifiers used when dealing with simple-type nominals and when classifying complex types as members of the class corresponding to its most prominent sense component[3]. This indicates that complex types do occur in contexts typical of the different classes corresponding to their sense components, i.e. they belong to more than one class and behave as such.

The second aspect made apparent from the results in Table 3 is the overall low recall. These results are consistent with the work of Rumshisky et al. (2007) and the discussion in Section 3.2.1, specifically the asymmetries in terms of prominence of the different meaning components of complex types. This is reflected in the frequency of occurrences in contexts indicatory of a given class, which represents the information provided to our classifiers.

[3] The low precision reported for EVT•**INFO** as INFO is not independent of the reduced amount of nouns (9) of this type in the gold standard (see Table 1).

The noun *church*, for instance, occurred in contexts typical of LOCATION nouns with a relative frequency of 0.015 and of 0.030 in contexts typical of ORGANIZATION nouns. This is also the case of the noun *jurisdiction,* which occurred with a relative frequency of 0.039 in contexts typical of ORGANIZATION nouns and just 0.014 in contexts typical of LOCATION nouns. This provides evidence that more distributional information is available toward one sense over another, which is bound to affect classification results, particularly when the asymmetry is large.

Thus, the representation of senses in distributional data has an impact on our classification results, being responsible, in particular, for insufficient distributional evidence towards class membership for an important part of nouns in our list, which explains the low recall observed.

Thirdly, although the absolute numbers are lower due to the aforementioned recall, the ratio of complex types per class shows similar tendencies to the human annotation results. In fact, the ratios of complex types for the ORGANIZATION and LOCATION classes are balanced, along the lines of the human annotation results (see Table 1), whereas a big asymmetry is observed for the INFORMATION and EVENT classes, again mirroring the human annotation results (see Section 3.2.1).

Given our objective to verify whether complex-type nominals provide distributional evidence concurrent with more than one semantic class, our cross-classification experiment shows that the distributional information available generally indicates that complex types demonstrate a distributional behavior typical of members of more than one class, though the information available is not enough to correctly classify a part of the nouns studied, as indicated by the low recall observed in Table 3.

However, in this experiment we only consider a part of the distributional data for each complex type at a time. Having demonstrated that complex types show distributional behavior typical of members of more than one class and being clear that more information has to be considered for classifiers to achieve a better performance, we propose to include indicatory contexts of each of the classes composing the complex type in the same classifier, this way accounting for its full sense potential in the classification task.

4.2 Distinguishing complex types from simple-type nouns

The experiments described in Section 4.1 show that complex-type distributional evidence

is indicatory of class membership to more than one class, but also that individually this information is often not sufficient for automatic systems to perform accurately and robustly. Thus, we put forth a new experiment to classify complex types built upon these observations. Our cross-classification experiments considered distributional information available for each word in contexts indicative of each class corresponding to one of its senses individually. In this section we depict an experiment where we combine contextual cues indicatory of each individual class that corresponds to the different sense components of a complex type to train a classifier.

The goal of this experiment is to automatically distinguish complex types from simple types by training a dedicated classifier. This approach combines the distributional information characteristic of each individual sense component of the complex type in a single classifier, providing it with more information at a time, which we expect to raise both precision and recall. Along this line, we collected distributional evidence of nouns by simultaneously using the cues for each class corresponding to the different sense components of the complex types considered in this work. We provided this information to the classifier as well as the human annotated gold standard for training. As in the previous experiment, we used LMTs (Landwehr et al., 2005), this time in a 10 fold cross-validation setting. Table 4 presents the results of the classification of ORG•LOC and EVT•INF complex types.

	Accuracy	Precision	Recall	F-Measure
ORG•LOC	67.68%	0.62	0.67	0.62
EVT•INFO	78.75%	0.72	0.78	0.72

Table 4. Results of complex-type classifiers

The results above demonstrate that by combining cues indicatory of different individual semantic classes and thus providing distributional evidence of the entire sense potential of a complex-type to the classifier we are able to automatically classify complex types, distinguishing them from simple-type nominals. As in the previous experiment, in order to be distinguished from simple-type nominals, complex types must demonstrate sufficient distributional evidence in contexts indicatory of classes corresponding to their different sense components.

By combining the distributional information indicatory of two classes and providing it simultaneously to the classifier, we improve the results previously obtained and attain accuracy in line with state-of-the-art simple-type classifiers (see Bel et al.'s (2012) results regarding nominal lexical semantic classification in English). Moreover, this approach overcomes the main issue in the results depicted in Section 4.1, which was low recall.

A final observation on the results attained regards the difference of more than 10% of accuracy between the classifiers for both complex types considered. Previously discussed work by Ježek and Melloni (2011) (see Section 3.2.1) help us identify possible causes for these contrasts, such as an ontological dependence between component types of dot types like EVT•INF, whose occurrences have both sense components of the dot object generally simultaneously present. However, the same is not true for complex types such as ORG•LOC nouns, which results in a more disperse distributional behavior between indicatory contexts of each sense component of the dot object, constituting a challenge for classifiers, which naturally impacts performance.

5 Final Remarks

The classifiers developed in this work consider contexts indicatory of each nominal class that corresponds to a sense component of a complex-type. As shown, our classifiers are able to automatically identify nouns that display characteristic properties of different simple types, namely LOCATION and ORGANIZATION, and EVENT and INFORMATION. By achieving this, we demonstrate the validity of our hypothesis that dot-object nouns simultaneously display distributional characteristics of the different classes that correspond to their sense components.

Although, we obtain results in line with state-of-the-art performance of simple-type classifiers by combining contextual information for the different sense components of complex types, we still do not capture those contexts where only dot-type nouns can occur (i.e. contexts that are unique to these nouns and clearly separate them from simple types and homonyms). Given the specific properties of EVT•INF nouns, the weight of this type of contexts can be hinted by the different performance of the classifiers developed, as discussed in the previous section.

In future work we will evaluate to which extent using the contexts specific to complex types, i.e. contexts which "convoke" different sense components simultaneously (see, for instance, Šimon and Huang (2009), Pustejovsky (2007) and Cruse (2000)), can result in a still more reliable classifier, with the potential to contribute to cost-effectively create more accurate LRs for NLP.

Acknowledgments

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.

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.

G. Boleda, S. Padó & J. Utt. 2012. Regular Polysemy: A Distributional Model. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM),* Montreal, Canada: 151-160.

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.

J. Bybee. 2010. *Language, usage and cognition*. Cambridge University Press, Cambridge.

J. Bybee & P. Hopper. 2001. *Frequency and the emergence of language structure*. John Benjamins, Amsterdam.

R. Cooper. 2005. Do delicious lunches take a long time?, *GSLT internal conference*.

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

A. Cruse. 2000. Aspects of the micro-structure of word meanings. In Y. Ravin & C. Leacock (eds.), *Polysemy: Theoretical and Computational Approaches*. Oxford University Press.

S. Frisson & M. J. Pickering. 2009. Semantic Underspecification in Language Production. *Language and Linguistics Compass*, 3(1). Blackwell Publishing, Ltd.

Z. Harris. 1954. Distributional structure. *Word*, 10(23): 146-162.

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

E. Ježek & A. Lenci. 2007. When GL meets the corpus. A data driven investigation of semantic types and coercion phenomena. In P. Bouillon, L. Danlos & K. Kanzaky (eds.), *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon*, Paris, France.

E. Ježek & C. Melloni. 2011. Nominals, polysemy and co-predication. *Journal of cognitive science*, 12.

A. Kilgariff. 1992. *Polysemy*. PhD Thesis, University of Sussex, UK.

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

A. Lenci, M. Johnson & G. Lapesa. 2010, Building an Italian FrameNet through semi-automatic corpus analysis. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*: 12-19.

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

P. Merlo & S. Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3): 373-408.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross & K. J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): 235-244.

J. Utt & S. Padó. 2011. Ontology-based distinction between polysemy and homonymy. *Proceedings of the Ninth International Conference on Computational Semantics (IWCS'11)*, Stroudsburg, PA: 265-274.

G. Pethö. 2001. What is polysemy? A survey of current research and results. In E. T. Ne´meth & K. Bibok (eds.), *Pragmatics and Flexibility of Word Meaning*. Elsevier, Amsterdam: 175-224.

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

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

J. Pustejovsky. 2007. Type Theory and Lexical Decomposition. In P. Bouillon & C. Lee (eds.), *Trends in Generative Lexicon Theory*. Kluwer Publishers.

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

P. Šimon & C. Huang. 2010. Cross-sortal Predication and Polysemy. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 2010)*: 853-861.

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