Choosing which to use? A study of distributional models for nominal lexical semantic classification

Lauren Romeo*, Gianluca E. Lebani†, Núria Bel*, Alessandro Lenci†

*Universitat Pompeu Fabra
Roc Boronat, 138, Barcelona (Spain)
lauren.romeo@upf.edu, nuria.bel@upf.edu

† Computational Linguistics Laboratory, University of Pisa
via Santa Maria, 36, Pisa (Italy)
gianluca.lebani@for.unipi.it, alessandro.lenci@ling.unipi.it

Abstract

This paper empirically evaluates the performances of different state-of-the-art distributional models in a nominal lexical semantic classification task. We consider models that exploit various types of distributional features, which thereby provide different representations of nominal behavior in context. The experiments presented in this work demonstrate the advantages and disadvantages of each model considered. This analysis also considers a combined strategy that we found to be capable of leveraging the bottlenecks of each model, especially when large robust data is not available.

Keywords: Distributional Models, Lexical Semantic Classes, Nominal Classification

1. Introduction

Lexical semantic classes are generalizations that aim to capture similarities among meanings of different lexical entities. Nominal lexical semantic class information represents crucial input for many Natural Language Processing (NLP) applications, such as Information Extraction, the building and extending of semantic ontologies, Machine Translation and Question Answering, as well as for tasks such as the acquisition of selectional preferences of predicative expressions. Yet, the lexical semantic tagging of nouns in large lexica is still mostly done manually, thus the high cost of this task can hinder the production of rich lexica for different languages and/or domains.

Distributional models that infer the meaning of words from similarity of contexts can highly reduce the cost of this task by capturing the indicative properties of lexical semantic classes. Recent work has demonstrated the usefulness in applying the Distributional Hypothesis (Harris, 1954) to create predictive models that are representative of the behavior of a lexical semantic class (cf. for instance: (Stevenson et al., 1999; Lin and Pantel, 2001; Joanis et al., 2007; Baroni and Lenci, 2010; Bel et al., 2012; Boleda et al., 2012; Romeo et al., 2012; Lenci, in press). In this way, the distributional behavior of some class members is utilized to make further predictions on class membership. However, distributional spaces can vary greatly with different models, and thus far, there has been little consensus on the conditions that prefer one model over another.

This paper empirically evaluates the performances of different distributional models in a nominal lexical semantic classification task. Our goal is to provide empirical evidence to determine what features produce a more accurate classification, which can thus increase the reliability of automatically constructed resources and simultaneously reduce the problems of the high costs to build large, domain-tuned lexica.

We consider models that exploit different types of distributional features, which thereby produce different representations of nouns, as found in context. Thus, the aim of this paper is to explore different distributional models used for nominal lexical semantic classification. In the next sections, we describe each of the models considered (Section 2) as well as the methodology followed (Section 3). We present the classification results obtained (Section 4) and explore how the performance of the classifier varies between distributional models (Section 5). We conclude with final remarks and future research regarding the use of distributional models in classification tasks (Section 6).

2. Related Work

Nominal lexical semantic classes gather together properties that appear to be linguistically significant for a number of linguistic phenomena. According to the linguistic tradition, words that can be inserted in the same contexts can be said to belong to the same class (Harris, 1954). Thus, we define lexical classes to be linguistic generalizations drawn from the characteristics of the contexts where a number of words tend to appear.

Classification approaches propose training a classifier with distributional information about their occurrences in selected contexts where words belonging to a class can occur (e.g. certain quantifiers, such as numerals, can directly modify count nouns but not mass nouns (Gillon, 1992)). The whole set of occurrences of a word are then considered to be features which define its class membership, either because the word is observed in a number of particular con-

	DM	LING	LINE
	sub-int-happen-V	x-NN when-WRB	car
accident	sub-int-occur-v	until-IN the-DT x-NN	injury
	obj-cause-v	since-IN the-DT x-NN	road

Table 1: Example of features used for each model for the event noun accident. In DM, features represent the syntactic position of a target noun as a combination of dependency ("sub-int") or its dependent head ("happen-V"); in LINE, features represent linguistically-motivated class indicatory lexico-syntactic contexts, such as a target noun ("x-NN") preceding a specific adverb ("when-WRB"); in LINE, features represent simple co-occurring words in a 5-word context window.

texts or because it is not. However, distributional representations can vary greatly, depending on the specific aspects of meaning they are designed to model. Thus, selecting the most useful and/or indicative features is one of the most important tasks in nominal lexical semantic classification, as the features considered in a distributional model directly affect the classification predictions produced.

Along this line, distributional models can be considered either *structured*, if they collect corpus derived information in the form of word pairs and dependency relations (selected references include: Grefenstette (1994), Padó and Lapata (2010), and Baroni and Lenci (2010)), *unstructured*, if they consider simple word co-occurrence statistics (selected references include: Lund and Burgess (1996), Landauer and Dumais (1997) and Bullinaria and Levy (2007; 2012)) or *linguistic*, if they collect distributional information with hand-written pattern based features (selected references include: Bel et al. (2007; 2012)).

As previously mentioned, in the context of this paper, we consider distributional models that exploit different representations of features. Along this line, we consider a structured distributional semantic resource, an unstructured linear model and a linguistic model. We acknowledge that there is a variety of distributional models currently in the state of the art (for a general survey see Turney and Pantel (2010)). However, we consider the aforementioned models because they are representative of the main state-of-the-art models currently used for noun classification, which is the focus of the work presented here. In this way, we do not consider these models to be the only approach to this task, although, as mentioned, we do consider them the most representative. The following subsections detail each of the models considered.

2.1. Distributional Memory

Distributional Memory (DM: Baroni and Lenci (2010)) is representative of a structured distributional model. DM is proposed as a general purpose resource for semantic modelling. It consists of work-link-word tuples, which are extracted with different levels of lexicalization using different types of pre-defined lexico-syntactic patterns. The framework of DM was designed to exploit corpus data to its full extent for any type of semantic task. Thus, the information considered in DM attempts to overcome the limitations of ad-hoc or manually constructed patterns, especially in the wake of exponential growth of availability of corpus data. In this way, by collecting one set of statistics from the corpus, this model can exploit different views of extracted data and different algorithms to tackle various tasks.

Extensive and systematic studies have been conducted with DM, including but not limited to similarity judgments, synonynym detection, noun categorization, detection of selectional preferences, etc., which demonstrate that it is both versatile and comprehensive enough to address a variety of semantic tasks. Overall, in the large battery of experiments considered in their seminal work regarding DM, Baroni and Lenci (2010) report that in nearly all of the considered test sets, their best implementation is at least as good as other algorithms reported in the state of the art, or among the top of the state-of-the-art ranking.

It is also noted that although DM performs on par with other models, it is a general model, thus its parameters are not modified for any specific task which can result in a less accurate performance when considering specifically tailored models for a given task. The discussion in Section 5. contains a detailed analysis regarding how the features considered in DM provide distributional information that successfully overcomes bottlenecks, such as noise and sparsity, which are characteristic of other distributional models used in nominal lexical semantic classification tasks. In the work presented in this paper, we used the TYPEDM instance of DM, which is readily available for download and use.[1]

2.2. Linear Models

Bullinaria and Levy (2007; 2012) and Bullinaria (2008). have extensively explored unstructured distributional models. Considering that aspects of word-meaning can be induced using simple word co-occurrence counts from corpus data, they studied semantic word categorization as a function of window type and size, semantic vector distribution, as well as corpus size.

Bullinaria and Levy (2012) report the best performance using their models for semantic categorization at approximately 80%. However, they noted that their model becomes compromised with smaller corpora, as it demonstrated sensitivity to a reduction in data when used for semantic categorization tasks. We, too, observed such a sparsity effect on the performance of our LINE model in our classification experiments on smaller corpus data (see section 5. for details). However, it is not exclusive to unstructured models, as structured models can also be subject to problems concerning sparse data (Turney and Pantel, 2010).

[1] http://clic.cimec.unitn.it/dm/

2.3. Linguistic models

Linguistic models consider manually identified lexico-syntactic patterns that represent generalizations upon different contexts where a number of words belonging to a class tend to occur. These patterns cue a semantic property that a set of words, or class, may have in common, which is then used as an indicator for members of that class. Besides the inclusion of lexical information (e.g. a set of prototypical verbal lemmas of which a target noun is recurrently a subject), linguistic models take into account the crucial role that syntax can have in defining the distributional properties of classes by specifying patterns made of a combination of lemmas and PoS.

This approach has been used for the lexical semantic classification of verbs (Merlo and Stevenson, 2001), which selected very specific ad-hoc linguistic cues for classifying verbs undergoing different types of diathesis alternations, while Joanis et al. (2007) considered general linguistic information, such the frequency of filled syntactic positions or slots, tense and voice of occurring verbs, etc., to classify English verbs into a number of Levin (1993) classes. Although much more work has been done regarding the classification of verbs (see Korhonen (2010) and Schulte im Walde (2009) for a survey of the state of the art), this approach has also had success with the classification of nouns. For instance, Baldwin and Bond (2003) and Baldwin (2005) considered different syntactic cues for nominal classification, such as the PoS tags of neighboring words that take into account head number, modifier number, subject-verb agreement, the occurrence in N of N constructions, etc. In their experiments with nominal classification, Bel et al. (2007) considered local contexts in a PoS-tagged corpus as features.

More recently, Bel et al. (2012) expanded on the previous work in nominal lexical semantic classification and considered classes such as EVENT, HUMAN, CONCRETE, SEMIOTIC, LOCATION and MATTER. They identified lexico-syntactic patterns indicative of semantic properties of a given class (such as: a prepositional phrase headed by "during" is indicative for eventive nouns). Although achieving an accuracy of around 80%, they concluded that the selection/identification of particular lexico-syntactic information can, on one hand, limit the amount of data considered resulting in sparse vectors and nouns in the gold standard that were not found in any of the contexts that were taken as cues. On the other hand, they can introduce noise into the vectors as the cues may not be indicative for a single class (for a detailed analysis of the results obtained by this model see Section 4.1.).

2.4. Combining linear and linguistic information

In an attempt to further explore the potential of models that use surface information, we also considered the combination of features of the unstructured linear model and the linguistic model into a fourth model that uses both linguistically-motivated information as well as linear context as features.

In combining the features from these two models, we can determine whether the distributional information of one model can be compensated with the distributional information of the other, especially in the case that one of the models provides insufficient data for classification (see Section 5. for a detailed discussion of this combinatory strategy).

3. Methodology

The goal of this work is to empirically evaluate the performance of different distributional models in a nominal lexical semantic classification task. In this way, we considered models that exploit different types of distributional features, thereby providing different representations of nominal behavior in context.

As described in detail above, the first model considered is the structured Distributional Memory model (DM: (Baroni and Lenci, 2010)), which is a generalized framework for distributional semantics that uses word by link–word tuples from a dependency parse of a corpus as features. This is the only model considered that incorporates syntactic information provided by a dependency parser.

The second model considered is an unstructured linear model (henceforth: LINE), built by extracting tokens in context windows of a target noun. In this model, features consist of tokens extracted from a standard 5-word context window (Evert, 2008), to the right and to the left of each target word.

The third model considered is a linguistically-motivated model (henceforth: LING), which uses as features the lexico-syntactic information considered indicative of a given lexical semantic class. These features are manually selected and include lexico-syntactic information such as selectional preferences, grammatical marks, prepositions and suffixes (as detailed in Bel (2012)).

Finally, the combinatory strategy that we consider (henceforth: LINGLINE) uses the feature information from both the LING and the LINE models. Table 1 presents examples of features from each of the models considered.

3.1. Data Preparation

Each of the models was trained on a concatenation of the ukWaC corpus (Baroni et al., 2009), a mid-2009 dump of the English Wikipedia, and the British National Corpus (Burnard, 2007), following Baroni and Lenci (2010), for a total size of approximately 2.83 billion tokens. Although the same corpus was used to extract each of the models, DM considers the full syntactic annotation (tokens, PoS tags and syntactic dependency information). The features considered for the DM model were extracted from corpus data using the DM methodology to extract tuples, as defined by Baroni and Lenci (2010)

LINE considers only tokens as features. To extract the features for this model, all PoS and punctuation were removed from the corpus data and 5-word windows containing a token of at least 3 characters were extracted for each target noun, as defined in our gold standard.

LING considers only tokens and corresponding PoS tags. To extract the features for this model from corpus data,

each of the lexico-syntactic patterns identified were specified through regular expressions with PoS tags given after each token to identify occurrences of nouns in the indicated contexts. See Table 2 for the number of features considered per model and class.

	Targets	DM	LING	LINE
COM	208	775,747	16	27,095
EVT	211	687,019	20	27,086
HUM	208	656,023	17	27,078
LOC	114	572,191	22	27,073
ORG	111	535,675	16	27,042

Table 2: Number of target nouns per class and number of features per class for each model considered

3.2. Experiments

We use positive Local Mutual Information (pLMI: Evert (2008)) as our weighting scheme because it is an approximation of the log-likelihood ratio measure that has been shown to be a very effective weighting scheme, especially in the case of sparse frequency counts (see Baroni and Lenci (2010) for more details). The pLMI was calculated using the DISSECT toolkit (Dinu et al., 2013). Following standard practice (Bullinaria and Levy, 2007), negative weights are raised to 0. The information for each class and model was compiled into a sparse matrix consisting of four elements: target word, feature, weight and class membership information, which was provided to the classifier.

The work presented here consists of classification experiments conducted with the following lexical semantic classes in English: INFORMATIONAL OBJECT (COM), EVENT (EVT), HUMAN (HUM), LOCATION (LOC), ORGANIZATION (ORG). Each class was selected because of ongoing research on nominal regular polysemy (Pustejovsky, 1995), which also serves to provide insight to interpret results obtained.

Gold standards that consist of nouns containing a sense from WordNet (Fellbaum, 1998) that correspond to a class considered in our experiments (e.g. people in the case of HUM) were used for evaluation. The gold standards were balanced with respect to class members and elements not belonging to the class (see targets in Table 2, which presents the number of class members, each appearing *n* times in any corpus). The actual occurrences of target nouns in the corpus determined the final lists.

We performed each binary classification experiment using a CART Decision Tree (DT) classifier, a k Nearest Neighbor (kNN) classifier and a Support Vector Machine (SVM) classifier, as implemented in the Scikit-Learn toolkit (Pedregosa et al., 2011). For a baseline, we implemented a dummy-classifier which generates predictions uniformly at random. We selected this classifier to compare the success of our classifiers against a random decision.

All evaluations were conducted in a 10-fold cross validation setting. Due to space constraints, hereafter, we report on the results obtained using only DT. We focus on the results obtained using DT because they provide a balance between the characteristics of different machine learning classification algorithms. Also, DT tend to perform better when dealing with categorical features, which is important for binary classifications (Kotsiantis, 2007).

4. Results

The results in Table 3 show that overall the F1 score of each model demonstrates a statistically significant improvement ($p < 0.05$) over the baseline. In regards to the performance of the individual models, we observed that DM obtains the highest overall results, with its F1 score demonstrating a statistically significant improvement ($p < 0.05$) over the F1 score of both LING and LINE. We attribute this result to the inclusion of syntactic information provided by a dependency parse in the model, which is the main difference between the DM model and the LING and LINE models. We reflect on this point in more detail in Section 5.

Interestingly, the LING and the LINE models, which both consider more shallow features, achieve an F1 score of 0.76 and 0.77, respectively, which can already be considered for use in NLP tasks. However, there is no statistical significance between the F1 scores of theses models, although there is a slight difference in their recall and precision, especially when considering individual classes. This implies that each model has different advantages in regards to the lexical semantic classification of nouns (the Error Analysis in Section 4.1. and the resulting Discussion in Section 5. reflect upon this point in detail).

With respect to the individual classes, we observe that HUM and ORG obtain a better classification from LING while COM, EVT and LOC obtain better classification results with LINE, thus indicating that the distributional model selected for classification should consider the indicative properties of the class being classified (Bel et al., 2012). For instance, the LING model benefits classes, such as ORG and HUM, that have readily identifiable class-specific features, such as morphological or grammatical marks while the LINE model benefits classes in which the features considered to be indicative of a class in linguistically-motivated models may fail to handle the heterogeneity of members as they occur in actual language use. For these types of classes, the information provided by linguistic cues may be too disperse in feature vectors to be accurately captured by classifiers, thus the simple linear features that are used in the LINE model are more appropriate to capture the basic patterns of occurrences of lemmas of those more broadly defined classes (i.e. classes with a larger internal variation), as found in corpus data.

In Table 3, we also observed that the combined LINGLINE model demonstrates a statistically significant improvement ($p < 0.05$) over both the LING and the LINE models, respectively. As there is no statistical difference between the LING and the LINE models, individually, these results demonstrate the benefit of simultaneously considering the features of both models.

	DM			LING			LINE			LINGLINE			Baseline		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
COM	0.87	0.88	0.88	0.68	0.67	0.67	0.79	0.78	0.78	0.81	0.80	0.80	0.52	0.58	0.54
EVT	0.85	0.81	0.83	0.81	0.79	0.79	0.83	0.85	0.84	0.80	0.85	0.81	0.47	0.54	0.50
HUM	0.92	0.91	0.91	0.88	0.9	0.88	0.76	0.78	0.76	0.89	0.84	0.86	0.53	0.58	0.55
LOC	0.83	0.81	0.81	0.73	0.77	0.74	0.8	0.77	0.78	0.84	0.84	0.83	0.48	0.54	0.50
ORG	0.84	0.82	0.83	0.72	0.74	0.72	0.72	0.76	0.73	0.79	0.77	0.77	0.51	0.54	0.52
MacroAvg	0.86	0.84	0.84	0.76	0.77	0.76	0.78	0.78	0.77	0.82	0.82	0.81	0.52	0.50	0.55

Table 3: Precision (P), Recall (R), and F-Measure for classification using each model of each class when considering larger corpus data

This improvement underlines the compensatory effect of using information provided by the combination of features from LING and LINE. For instance, LING includes indicative yet potentially sparse and/or noisy features while LINE includes a simply large amount of co-occurrence information. In the scope of the work presented here, noise refers to an effect that is attributed to the fact that for many features there is not an 1-1 association with a specific class, which causes many of the surface patterns to be ambiguous.

In this way, where the distributional information provided by the features of the LING model is not sufficient for the classifier to make a decision regarding class membership, LINE can provide extra information to the classifier to arrive at a generally more reliable decision (see the Discussion regarding the confusion matrices in Section 5. for more details) and vice versa. However, we acknowledge that LINGLINE still does not outperform DM, which again emphasizes the potential of the added value provided by the richer syntactic information available in DM. This result also confirms the bias of structured DSMs to identify paradigmatically similar words, which is important to note, especially as paradigmatic similarity forms the basis for semantic classification.

The evaluation of our distributional models also considers the effect of corpus size. Hence, we also conducted classification experiments on a smaller amount of corpus data[2]. Along this line, we used a 95 million token excerpt of the corpus (approx. 20 times smaller than the large corpus) to train each model. Following the previously described procedure, we obtained the results depicted in Table 4.

Again, the F1 score of each of the models, when trained on smaller corpus data, demonstrates a statistically significant improvement ($p < 0.05$) over the baseline. In regards to the performance of the individual models, we can observe that LINGLINE, obtains the highest overall results, with its F1 score demonstrating a statistically significant improvement ($p < 0.05$) over the F1 measure of both LING and LINE.

However, when considering smaller corpus data, LING produces a significantly higher F1 score than LINE which implies that the reduction of corpus data effectively reduces the distributional data needed by LINE to accurately predict class membership. Thus, the performance of LINGLINE when using smaller corpus data, again affirms the compensatory benefit of combining the features of both models, especially in the case where one model lacks sufficient information to make an accurate classification decision, such as observed in the results obtained by the LINE model in Table 4.

4.1. Error Analysis

In order to better interpret the scores obtained by each model, we conducted an error analysis based on the confusion matrices that resulted from each classification experiment. In this way, we have been able to identify the bottleneck of each model as a function of its resulting False Positives (FPs: the items incorrectly classified as members of the target class) and False Negatives (FNs: the items incorrectly classified as not belonging to the target class). Roughly speaking, FPs can be interpreted as a consequence of "noisy" feature vectors, while FNs can be interpreted as a consequence of sparsity, or lack of evidence in the feature vectors. In what follows, we summarize the observations that can be drawn from the error patterns showed by each model in the different corpus settings.

Larger corpus data (2.83 billion tokens): We first look at the types of features that are being considered. As described in detail in Section 2.3., the LING model consists of manually-identified linguistic features that are considered indicative of semantic properties of a class. However, as explained in Bel et al. (2012), these features are not individually indicative of a given class, as their predictive power arises through correlations between a set of these features. Although these features provide specific information regarding the behavior of a given lemma as found in corpus data, the information provided by a single individual feature is not necessarily class-specific, which can introduce noisy information into the feature vector, hindering the ability of the classifier to make an accurate decision.

To further inspect this, we considered the semantic class to which a given FP belongs. We noticed that there was a large amount of EVT nouns that were classified as COM nouns and vice-versa, a large amount of COM nouns that were classified as EVT nouns. For example, the COM noun *reservation* was incorrectly classified as an EVT noun in

[2]Given that the DM tensor is distributed as a readily available semantic resource that was pre-trained on the full corpus, the analysis on smaller corpus data was not performed for this model.

	LING			LINE			LINGLINE			Baseline		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
COM	0.66	0.68	0.66	0.74	0.73	0.73	0.72	0.73	0.72	0.49	0.54	0.51
EVT	0.71	0.71	0.71	0.66	0.64	0.65	0.74	0.71	0.71	0.47	0.53	0.50
HUM	0.87	0.86	0.86	0.70	0.70	0.69	0.91	0.87	0.89	0.53	0.59	0.56
LOC	0.69	0.64	0.64	0.64	0.62	0.61	0.74	0.74	0.73	0.48	0.56	0.51
ORG	0.81	0.76	0.78	0.70	0.70	0.69	0.77	0.77	0.76	0.55	0.56	0.56
MacroAvg	0.74	0.73	0.73	0.68	0.67	0.67	0.77	0.76	0.76	0.50	0.55	0.52

Table 4: Precision (P), Recall (R), and F-Measure for classification using LING, LINE, and LINGLINE for each class when considering smaller corpus data

the LING model, while the EVT noun *discrepancy* was incorrectly classified as a COM noun. This trend was also observed with ORG and HUM nouns. For instance, we saw that a large part of the FPs of ORG are members of the HUM class (such as: *comedian* and *graduate*) and a large part of FPs of HUM are members of the ORG class (such as: *choir* and *regime*). A high amount of confusion between FPs of the LOC and EVT classes was also observed. The binary setting of the classification task did not allow for an analogous analysis to be conducted on FNs. Table 5 presents the overall results of this analysis.

	COM	EVT	HUM	LOC	ORG
COM	0	36	43	29	26
EVT	49	0	27	49	26
HUM	20	20	0	22	52
LOC	23	37	16	0	18
ORG	21	21	33	27	0

Table 5: Confusion matrix of true semantic classes of FPs with the larger corpus data

On the one hand, the FPs can be due to the fact that HUM nouns, for instance, are explicitly marked, either grammatically or morphologically (i.e. suffixes such as "-er", "or", "ir" or the subject of psychological-type verbs), while ORG nouns can be considered collective HUM nouns, or a subset of this class (see Romeo et al. (2012) for a detailed discussion of this phenomenon).

On the other hand, these mis-classifications are also related to very particular cases of lexical ambiguity. For instance, COM and EVT nouns, as well as LOC and EVT nouns, have been considered in literature as examples of regular polysemy (see Pustejovsky (1995)), in which a lemma can be selected for as more than one sense. Under this assumption, some misclassifications can be caused by the fact that the lemma is also a member of another (potentially related) semantic class. Nonetheless, a discussion of polysemy goes beyond the scope of this paper, although, it is important to note that there is a systematicity in the mis-classification of the nouns which can be attributed to the lexical ambiguity of certain nouns.

Smaller Corpus Data (90 million tokens): When training with smaller corpus data, we observed that the results of the LING model were quite similar to its results on the full corpus data. However, the most significant difference was observed in the results of the LINE model. With smaller corpus data, we can directly see where the unstructured information of the LINE model is compensated with the linguistic information of the LING model, highlighting the value of combining the distributional information from both models in LINGLINE (see differences between models in precision and recall in Table 4), especially in the case that a model can not obtain sufficient distributional data to make an accurate classification decision. In this way, our results show that the LINGLINE model effectively reduces the overall ratio of FPs and FNs, resulting in more accurate classifications, as well as reflects a broader coverage.

In regards to the semantic classes of the obtained FPs, we observed trends similar to those discussed with regard to the results obtained with the larger corpus data. Along this line, we can say that although the amount of distributional information is reduced, the tendencies of the behavior of the nouns as found in general corpus data remain consistent.

5. Discussion

The goal of the work presented here is to empirically evaluate the performances of different distributional models in a nominal lexical semantic classification task. Overall, the DM model consistently obtains the strongest performance. We attribute this, on one hand, to the inclusion of syntactic information provided by the dependency parse in the model, which provides structure to the lexical information considered as features. On the other hand, we consider that general structural information (e.g. syntactic parse patterns, copulative structures, position with relation to a verb link, attribute nouns, prepositional phrases, etc.), provided by the features contemplated in the DM model, become indicative of a given lexico-semantic class with the incorporation of specific lexical information, especially when a given feature occurs many times with several members of a given lexico-semantic class. Thus, our results indicate that the quality of classification tasks increases with the inclusion of syntactic annotation, as demonstrated by DM.

In regards to the other distributional models that we considered, the results demonstrate that the LING model is de-

pendent on the availability of specific lexico-syntactic information in corpus data and the accuracy of a classifier to correlate the relations between a set of individual features that together are indicative of a given semantic class. However, as the results in Table 4 indicate, unlike the LINE model, the LING model does not necessarily require a large amount of corpus data, thus indicating that when available, the linguistic cues of the LING model, do provide sufficient information to the classifier to make accurate classification decisions.

The LINE model, however, is dependent on the availability of a large amount of corpus data to ensure a sufficient amount of surface information. Consequently, when a large amount of data is not available, the LINE model looses its predictive capacity. Moreover, as the model considers the use of many shallow features, there is a risk that many of those features are uninformative, thus providing no useful indicative information to the classifier. Although we did not test the DM model on the smaller corpus data, our intuition is that it would behave like the LINE model with respect to data sparsity. Therefore, we consider sensitivity to data sparseness as a general problem of distributional models, independently of their being structured or not.

Moreover, we must also consider that not all lexical classes may be equally identifiable through surface features (see the differences in the F1 scores of each individual class in Table 3 and 4). In this way, the availability of contextual distributional information, such as that considered in LINE, can help to overcome limitations of manually identified class-indicative features, such as low frequency of target occurrences or simply a sheer lack of class-indicative marks. Along this line, we can consider the LING model to be ideal when trained on smaller corpora while the LINE model is more predictive when trained on large general corpus data.

However, we also consider the results obtained when combining information from the LING and the LINE models in the LINGLINE model. We observed that this combination of information produced a compensatory effect in which each of the models provides information that may be lacking when considering the distributional information provided by only one model, especially in the case of the LINE model when trained on smaller corpus data. Thus, we consider that the combined LINGLINE model obtains state-of-the-art results (Bel et al., 2012) and on different-sized corpora. Along this line, as a large, robust, syntactic dependency-parsed (large) corpus is not always available for all languages, domains and/or tasks being considered, the joint exploitation of linguistically-motivated cues and linear co-occurrence features, as demonstrated by LINGLINE, is a viable alternative.

6. Final Remarks

Overall, our study provides an empirical evaluation of classifications produced by distributional models that exploit different degrees of feature extraction criteria. The experiments presented in this work demonstrate the advantages and disadvantages of each model considered. Our results consistently indicate that the quality of classification increases with the complexity of syntactic information considered in the features of distributional models. Along this line, the analysis conducted in this work resulted in a strategy that, by combining distributional linguistic and linear features, is capable of leveraging the bottlenecks of each model, especially when large robust data is not available. The results obtained serve to increase the reliability of automatically constructed resources that require nominal lexical semantic class information.

A limitation of this work is the assumption that the lexical semantic classes considered are monosemous, which, as we demonstrated in the Error Analysis in Section 4.1., can have a negative effect on the results, especially when considering ambiguous entities as members of a given class. Future research will address the inclusion of polysemy in our models (Pustejovsky, 1995; Bullinaria and Levy, 2007; Bel et al., 2012), as we consider it a ubiquitous phenomenon that must be accounted for in any distributional semantic space as well as in any classification scenario.

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 the EBES-IULA/UPF mobility grant, as well as the PRIN grant 20105B3HE8 *"Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary"*, funded by the Italian Ministry of Education, University and Research.

8. References

T. Baldwin and F. Bond. 2003. Learning the countability of english nouns from corpus data. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 463–470.

T. Baldwin. 2005. General-purpose lexical acquisition: Procedures, questions and results. In *Proceedings of the Pacific Association for Computational Linguistics 2005*, pages 23–32.

M. Baroni and A. Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

M. Baroni, S. Bernardini, A. Ferraresi, and 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, S. Espeja, and M. Marimon. 2007. Automatic acquisition of grammatical types for nouns. In *Human Language Technologies 2007: Proceedings of the North American Chapter of the ACL Companion Volume, Short Papers*.

N. Bel, L. Romeo, and M. Padró. 2012. Automatic lexical semantic classification of nouns. In *Language Resources and Evaluation Conference (LREC 2012)*.

G. Boleda, S. Schulte im Walde, and T. Badia. 2012. Modeling regular polysemy: A study of the semantic classification of catalan adjectives. *Computational Linguistics*, 38(3):575–616.

J.A. Bullinaria and J. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

J.A. Bullinaria and J. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stoplists, stemming and svd. *Behavior Research Methods*, 44:890–907.

J. A. Bullinaria. 2008. Semantic categorization using simple word co-occurrence statistics.Â. In *the ESSLLI Workshop on Distributional Lexical Semantic*.

L. Burnard. 2007. Reference Guide for the British National Corpurs (XML Edition).

G. Dinu, N. Pham, and M. Baroni. 2013. Dissect: Distributional semantics composition toolkit. In *Proceedings of the System Demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, East Stroudsburg PA: ACL.

S. Evert, 2008. *Corpora and collections*. Corpurs Linguistics. An international Handbook. Mouton de Gruyter, Berlin, 58 edition.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

B. S. Gillon. 1992. Towards a common semantics for english count and mass nouns. *Linguistics and Philosophy*, 15:597–639.

G. Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer, Boston, MA.

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

S. Schutle im Walde, 2009. *The Induction of Verb Frames and Verb Classes from Corpora*, pages 952–972. Corpurs Linguistics. An international Handbook. Mouton de Gruyter, Berlin.

E. Joanis, S. Stevenson, and D. James. 2007. A general feature space for automatic verb classification. *Natural Language Engineering*, 14:337–367.

A. Korhonen. 2010. Automatic lexical classification: bridging research and practice. *Philosophical Transaction of the Royal Society*, 368:3621–3632.

S. B. Kotsiantis. 2007. Supervised machine learning: A review of classification techniques. *Informatica (Slovenia)*, 31(3):249–268.

A. Lenci, in press. *Carving Verb Class from Corpora.* Word Classes. John Benjamins, Amsterdam - Philadelphia.

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

D. Lin and P. Pantel. 2001. Induction of semantic classes from natural language text. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 317–322.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28:203–208.

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

S. Padó and M. Lapata. 2010. Dependency-based constructions of semantic space models. *Computational Linguistics*, 33(2):161–199.

F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

J. Pustejovsky. 1995. *The Generative Lexicon*. Oxford University Press, Oxford.

L. Romeo, S. Mendes, and N. Bel. 2012. Identifying lexical semantic class in an unsupervised clustering task. In *Proceedings of the 24th International Conference in Computational Linguistics (COLING2012)*.

S. Stevenson, P. Merlo, N. Kariaeva, and K. Whitehouse. 1999. Supervised leaning of lexical semantic verb classes using frequency distributions. In *Proceedings of SigLex-99 Standarizing Lexical Resources*, pages 15–22.

T.Landauer and S.Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psycological Review*, 104(2):211–240.

P.D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.