

CCG Categories for Distributional Semantic Models

Paramita Mirza

University of Trento

paramita.paramita@unitn.it

Raffaella Bernardi

University of Trento

bernardi@disi.unitn.it

Abstract

For the last decade, distributional semantics has been an active area of research to address the problem of understanding the semantics of words in natural language. The core principal of the distributional semantic approach is that the linguistic context surrounding a given word, which is represented as a vector, provides important information about its meaning. In this paper we investigate the possibility to exploit Combinatory Categorical Grammar (CCG) categories as syntactic features to be relevant for characterizing the context vector and hence the meaning of words. We find that the CCG categories can enhance the representation of verb meaning.

1 Introduction

The distributional semantic approach is based on the idea that the meaning of a word relies heavily on its context. Hence, the meaning of a word can be represented as a vector of its co-occurrence frequency with the neighbouring words. There have been several works that explore ways to improve the representation of word meaning by incorporating syntactic information in the context vector, dependency relations between words being the commonly used syntactic features. Dependency-based Distributional Semantic Models (DSMs) have been tested against several tasks and shown to be among the best performing word space models (Erk and Pado, 2008; Cruys, 2008; Baroni and Lenci, 2009; Baroni and Lenci, 2010).

In this paper we investigate an alternative view on the syntactic features that can be used to enrich the context vector, namely Combinatory Categorical Grammar (CCG) categories, which provide a transparent relation between syntactic category and semantic type of a linguistic expression.

Hence, we propose to build a CCG-based DSM using a corpus annotated by a CCG parser.

We test the model on word categorization tasks, in particular concrete noun and verb categorization. We are interested in investigating how the role of context changes in capturing lexical meaning among the different word categories (nouns vs. verbs). Furthermore, we explore the performance of the model in capturing the different categories of verbs, based on several verb classifications studied in the literature.

By comparing the model based on CCG categories with an analogous one based on Part of Speech (PoS), we study the role of richer syntactic information in the task of word categorization. Finally, we include also function words (grammatical words) in the context vector instead of assuming that only content words (i.e. nouns, verbs, adjectives, adverbs) are relevant in capturing the word meaning. We find that for some cases function words are useful to distinguish different classes of verbs.

2 “Supertags” for Distributional Semantic Models

We propose to investigate the role of constituent structures and features encoding tense information, by building a distributional model with dimensions tagged by “supertags”, namely by Combinatory Categorical Grammar (CCG) categories.

CCG is the categorial grammar version studied by the Edinburgh research group led by Steedman (2000), which has been used to theoretically analyse several linguistic phenomena. It has been used for building a CCGbank (Hockenmaier, 2003) and has been implemented into an efficient and wide coverage parser (Clark and Curran, 2007)¹. Below we will briefly describe the CCG categories, without going into details about the grammar.

¹However, for our experiments we used the revised version presented in Honnibal et al. (2007)

CCG language consists of atomic and complex categories where the latter are built out of the former by means of the directional implication operators *Output\Input* and *Output/Input*. For instance, an intransitive verb is assigned the category $S\backslash NP$, which means that it wants an *NP*-argument on its left.

The atomic categories considered are *S*, *NP*, *N*, and *PP*, but they are also enriched with features that further specify sub-categorization information. Bare nouns are distinguished from non-bare nouns by enriching the *N*-category: $N[nb]$ (non bare) and N (bare). Sentences and verb phrases are distinguished by means of the features enrichment of the *S* category. Sentences are distinguished into: $S[dcl]$ (declarative sentences), $S[wq]$ (wh-questions), $S[q]$ (yes-no questions), $S[qem]$ (embedded questions), $S[em]$ (embedded declaratives), $S[frg]$ (sentence fragments), $S[for]$ (small clauses headed by *for*), $S[intj]$ (interjections) and $S[inv]$ (elliptical inversion). Verbs carry tense features such as: $S[b]\backslash NP$ (bare infinitives, subjunctives and imperatives), $S[to]\backslash NP$ (to-infinitives), $S[ps]\backslash NP$ (past participles in passive mode), $S[pt]\backslash NP$ (past participles used in active mode) and $S[ng]\backslash NP$ (present participles).

3 Data Sets

In the following we describe the data sets that are used to carry out the experiments presented in Section 4. We will start with the classification of concrete nouns and then move to several verb classifications.

Concrete nouns: We take the data set developed for the shared task at the ESSLLI 2008 Workshop on Lexical Semantics². The data set consists of 44 concrete nouns extracted from McRae et al. (2005). The nouns are grouped into 6 semantic categories, which are 4 categories of natural objects (*bird*, *groundAnimal*, *fruitTree*, and *green*) and 2 categories of man-made artifacts (*tool* and *vehicle*).

Furthermore, the nouns can also be classified into 3 classes: *bird* and *groundAnimal* are grouped together into *animal* class; *fruitTree* and *green* into *vegetable*; *tool* and *vehicle* into *artifact*. This hierarchical structure of the data set makes it possible to perform several tasks of categorization on one

data set.

Verbs (classification based on Levin’s criteria): Inspired by the classification originally proposed in Levin (1993) and further revised in Vinson and Vigliocco (2008), the organizers of the ESSLLI 2008 Workshop have proposed a data set of 45 verbs classified into 9 semantic classes (*communication*, *mentalState*, *motionManner*, *motionDirection*, *changeLocation*, *bodySense*, *bodyAction*, *exchange*, and *changeState*), further grouped into 5 classes: *communication* and *mentalState* into *cognition*; *motionManner*, *motionDirection*, and *changeLocation* into *motion*; *bodySense* and *bodyAction* into *body*; *exchange*; and *changeState*.

Verbs (argument structure distinctions): Merlo and Stevenson (2001) consider thematic relations to be crucial for verb classification, and hence propose a classification of verbs that is coarser than the one proposed by Levin and considered to be appropriate for numerous language engineering tasks. In particular, the relevant features to be considered are causativity, animacy, the passive vs. active voice, and the use of past-participle vs. simple past. They consider the argument-structure, which is the thematic roles assigned by the verbs, to be the discriminative main property. To this end, three classes of verbs are defined: *unergative*, *unaccusative*, and *object-drop*.

The *unergative* are intransitive *activity* verbs whose transitive form can be the causative counterpart of the intransitive form. The subject of an intransitive activity verb is specified by an *agent*, while the subject of the transitive form is indicated by the *agent of causation* (e.g. “The horse *raced* past the barn” and “The jockey *raced* the horse past the barn”).

The *unaccusative* verbs are intransitive *change-of-state* verbs. The transitive counterpart of these verbs exhibits the causative/inchoative alternation. The subject of the transitive unaccusative verb is marked by the *agent of causation*, but the alternating argument becomes a *theme* (e.g. “The butter *melted* in the pan” and “The cook *melted* the butter in the pan”).

The *object-drop* verbs are again *activity* verbs that exhibit a non-causative diathesis alternation in which the object is simply optional. The thematic assignment is *agent* for the subject and *theme* for the optional object (e.g. “The boy *played*” and “The boy *played* soccer”).

²<http://wordspace.collocations.de/doku.php/data:esslli2008:start>

The data set comprises 18 unergative, 19 unaccusative, and 20 object-drop verbs.

Verbs (positive and negative): The distinction between these two classes of words is studied in sentiment analysis and used in opinion mining. To obtain the negative verbs, we started from the list provided by Hu and Liu (2004)³. We removed those verbs that were not among the target words of our models (see Section 4), and finally kept only the most frequent 500 verbs. The positive verbs were extracted by choosing the 500 most frequent verbs in the corpus that are not in the negative class.

Verbs (upward and downward monotonic): If we take a logical view on the verb classification issue, verbs can be divided into upward and downward monotonic. For instance, let us consider two sentences (1) “We *know* the epidemic spread quickly” and (2) “We *doubt* the epidemic spread quickly”. From (1) we can infer the relaxed version “We know the epidemic spread” but we cannot infer the restricted one “We know the epidemic spread quickly via fleas”. Whereas the reverse happens for (2), from which we cannot infer “We doubt the epidemic spread” but we can infer “We doubt the epidemic spread quickly via fleas”.

In formal semantics, *doubt* is called a downward-entailing operator (it reverses the order of the arguments it takes) and *know* is called an upward-entailing operator (it preserves the order.) We take a data set of 29 downward-entailing verbs identified in Danescu-Niculescu-Mizil et al. (2009) based on Ladusaw (1980) (e.g. *avoid*, *block*, *decline*), and compare it to a set of non-downward-entailing verbs obtained by extracting verbs that do not belong to downward class and have similar frequencies in our corpus.

4 Experiments

We consider the unsupervised approach (i.e. clustering) to perform the word categorization tasks. Our goal is to bring to light the different role of syntactic information in capturing the meaning of nouns and verbs, and to investigate the role of function and content words, as well as tense features, in the different verb classifications previously discussed.

³<https://github.com/williamgunn/SciSentiment/blob/master/negative-words.txt>

4.1 Distributional Semantics Models

Our two models, CCG-DSM and PoS-DSM, are harvested from two large corpora, Wikipedia and ukWaC. The former contains approximately 820 million words put together into 43.7 million sentences. While ukWaC (Ferraresi et al., 2008) is a very large (>2 billion words) corpus of British English built by web crawling, limited to the .uk Internet domain. For both models, we consider as target words the 10K most frequent nouns (excluding proper nouns and nouns containing numbers), the 5K most frequent verbs, and the 5K most frequent adjectives. The two models differ with respect to their dimensions as specified below.

PoS-based model (PoS-DSM): For building the PoS-DSM, the corpora have been tokenized and annotated with TreeTagger⁴. As dimensions we took 20K most frequent PoS tagged words (e.g. *fruit_NN*, *use_VBG*) considering both content and function words. There are 376 PoS tagged function words, 204 words of them are unique lemmas.

CCG-based model (CCG-DSM): For building the CCG-DSM, the corpora have been analyzed by the CCG parser (Honnibal et al., 2007), and the dimensions are 20K most frequent CCG tagged words (e.g. *fruit_N*, *use_(S[ng]\NP)/NP*). There are 1,499 CCG tagged function words and 18,501 content words. Among the function words, there are many words with more than one CCG category, only 196 of them are unique lemmas; among the content words there are 8,812 unique lemmas. For example, *be* is associated with 61 different CCG categories which differ either in terms of features (e.g. *(S[b]\NP)/NP* vs. *(S[dcl]\NP)/NP*) or in terms of the arguments (e.g. *(S[dcl]\NP)/PP* vs. *(S[dcl]\NP)/S*).

For each model, we evaluate both the *complete model* (compl), which is the model containing both content and function words as the dimension, and the model built using only *content words* (cont). The latter model is obtained from the complete version by leaving in only nouns, verbs, adjective and adverbs as the dimension, based on their PoS tags.⁵ Moreover, we observe two different context windows: *2 size context window* (2win) in which only 2 words before and

⁴<http://www.ims.uni-stuttgart.de/projekte/complex/TreeTagger/>

⁵As a consequence of this filtering method, the model based on only content words includes also negation, e.g. *not*, since it is tagged as an adverb.

Model	6-way		3-way		2-way		Average Entropy	Average Purity
	Entropy	Purity	Entropy	Purity	Entropy	Purity		
Van de Cruys (dependency)	0.173	0.841	0.000	1.000	0.000	1.000	0.058	0.947
CCG-DSM-cont-senwin-raw	0.243	0.773	0.067	0.977	0.755	0.682	0.355	0.811
PoS-DSM-cont-senwin-raw	0.243	0.773	0.067	0.977	0.755	0.682	0.355	0.811
Van de Cruys (BoW)	0.334	0.682	0.539	0.705	0.983	0.545	0.619	0.644

Table 1: Concrete nouns clustering result

2 words after a given target word are considered to co-occur together with the target word and hence determine the context words; and *sentence size context window* (*senwin*) in which we assume that all words within the same sentence of a given target word are the context words. Finally, we consider the following different weighting schemas: Positive Point-wise Mutual Information (PPMI), Exponential Point-wise Mutual Information (EPMI), Positive Local Mutual Information (PLMI), and Positive Log Weighting (PLOG), besides the raw co-occurrence frequencies.

For the data sets developed by the organizer of the ESSLLI 2008 Workshop, which are the concrete noun categorization and the verb categorization based on Levin’s classes, we report also the results of the dependency based model of Cruys (2008) – that resulted to be the one best performing at the workshop. Cruys (2008) compare the dependency based model with a Bag-of-Words model (BoW) to study the effects of syntactic information.

4.2 Clustering Algorithm

We follow the instructions given in the ESSLLI 2008 Workshop for all our experiments, using CLUTO toolkit (Karypis, 2003) for clustering. We use the k-means algorithm of CLUTO using the *rbr* parameter with global optimization, which repeatedly bisects the objects until the desired number of clusters is reached. As for the other parameters we use the default values.

4.3 Evaluation Measures

To evaluate the cluster quality, we use the two standard measures available in CLUTO: *entropy* measures the degree of “disorder” in a cluster (i.e. how many objects from different classes grouped into one cluster), the best result is obtained with value 0; while *purity* (Zhao and Karypis, 2001) measures the degree to which a cluster contains words from one class only (i.e. the proportion of

the most frequent class in the cluster), the best result is obtained with value 1.

CLUTO also provides tools for analysing the discovered clusters, which can be used to gain a better understanding of the set of objects assigned to each cluster and to provide brief summaries about the cluster’s contents. The set of *descriptive features* is determined by selecting the features that contribute the most to the average similarity between the objects of each cluster. For each descriptive feature, a certain number is given, which denotes the percentage of the within cluster similarity that this particular feature can explain.

5 Results and Analysis

We will present the experiment results and analysis for the various data sets explained in Section 3.

5.1 Concrete Nouns

As previously discussed, since the data set is organized hierarchically, it is possible to do several tasks of clustering, namely 6-way, 3-way, and 2-way clustering. Table 1 reports the detailed results for the three clustering tasks separately. For the CCG-DSM, we report the results only of the best performing version, which is the model with only content words as dimensions, the sentence as context window, and using raw frequency values (CCG-DSM-cont-senwin-raw).

For comparison, Table 1 shows also the results achieved at the ESSLLI 2008 Workshop by the other models previously described. CCG-DSM achieves better results than the BoW models (Cruys, 2008), but it is outperformed by the model based on the dependency relation – even though in 3-way clustering task the purity and entropy values of both models are comparable. Finally, in this particular experiment setup, PoS-DSM achieves exactly the same result, showing that the CCG categories are not really helpful in this particular task. Below we will present the qualitative analysis of the CCG-DMS and PoS-DSM for this task.

The CCG-DSM model successfully discriminates nouns of *vegetable* (cluster 0), *animal* (cluster 1), and *artifact* (cluster 2) classes. However, one noun from the *animal* class (“chicken”) is grouped together with nouns of the *vegetable* class. Table 2 reports the top 10 descriptive features for each cluster obtained by the CCG-DSM in the 3-way clustering experiment, while Table 3 reports the PoS-DSM ones.

Cluster	Descriptive features
0	other_N/N 5.2%, fruit_N 4.1%, apple_N 3.0%, not_(S\NP)\(S\NP) 2.3%, tomato_N 2.3%, potato_N 2.2%, crop_N 2.2%, tree_N 2.2%, onion_N 1.7%, also_(S\NP)\(S\NP) 1.7%
1	other_N/N 6.0%, not_(S\NP)\(S\NP) 5.2%, bird_N 2.8%, year_N 2.3%, animal_N 2.2%, also_(S\NP)\(S\NP) 2.2%, dog_N 2.0%, large_N/N 1.9%, many_N/N 1.8%, include_(S[dcI]\NP)/NP 1.8%
2	not_(S\NP)\(S\NP) 5.9%, other_N/N 3.9%, small_N/N 2.4%, first_N/N 2.4%, use_(S[ngI]\NP)/NP 2.3%, time_N 2.0%, new_N/N 1.9%, water_N 1.6%, also_(S\NP)\(S\NP) 1.6%, year_N 1.6%

Table 2: Descriptive features for 3-way concrete nouns clustering by CCG-DSM-cont-senwin-raw

Cluster	Descriptive features
0	also_RB 4.4%, other_JJ 4.1%, not_RB 4.0%, fruit_NN 3.0%, then_RB 1.7%, small_JJ 1.4%, fresh_JJ 1.3%, vegetable_NNS 1.2%, apple_NNS 1.2%, large_JJ 1.2%
1	not_RB 7.1%, also_RB 5.9%, other_JJ 4.4%, species_NNS 3.5%, bird_NNS 2.0%, large_JJ 1.4%, many_JJ 1.4%, sea_NN 1.4%, animal_NNS 1.3%, first_JJ 1.3%
2	not_RB 7.6%, also_RB 4.4%, other_JJ 2.8%, then_RB 2.8%, use_VBN 2.8%, small_JJ 2.0%, use_VBG 2.0%, water_NN 1.8%, first_JJ 1.8%, time_NN 1.7%

Table 3: Descriptive features for 3-way concrete nouns clustering by PoS-DSM-cont-senwin-raw

We can see that the descriptive features used by CCG-DSM and PoS-DSM are similar, most of them are nouns and adjectives. Thus, in this case CCG categories do not give more information than PoS tags.

5.2 Levin Inspired Verb Classification

Table 4 reports the comparison of models’ performance on clustering the 45 verbs of the ESLLI 2008 Workshop. The best performance of CCG-DSM is achieved using the following experiment

setup: dimensions include both function and content words, context window is of size 2, and the weighting scheme is EPMI. Although the overall performance is lower than the one for the concrete nouns, with the average purity of 0.678 vs. 0.811, it is higher than one obtained by the best performing model at the Workshop, namely the model based on dependency relations (0.678 vs. 0.612). Moreover, the average entropy is reduced significantly: 0.323 vs. 0.436 for the CCG-based model and the dependency-based model respectively.

The confusion matrix for the 5-way verb clustering (Table 5) shows that the model obtains high purity and low entropy for the cluster 0 (10 verbs of the *motion* class out of 15), cluster 1 (7 verbs of the *cognition* out of 10) and cluster 3 (8 verbs of the *body* class out of 10). However, it confuses the verbs of the *motion* class and the verbs of the *changeState* class. Several verbs from the *motion* class, such as “fall”, “pull”, “push”, and “rise” are considered as the verbs of the *changeState* class instead. The same confusion also happens between the *exchange* and *cognition* class: “evaluate”, “request”, and “suggest” are categorized as *exchange* verbs instead of *cognition*. The descriptive features for each cluster are described in Table 6.

Cluster	Classes					Entropy	Purity
	ex ¹	mo ²	cs ³	bo ⁴	co ⁵		
0	0	10	1	0	0	0.189	0.909
1	0	0	0	2	7	0.329	0.778
2	0	4	4	0	0	0.431	0.500
3	0	0	0	8	0	0.000	1.000
4	5	1	0	0	3	0.582	0.556

¹ exchange ² motion ³ changeState ⁴ body
⁵ cognition

Table 5: Confusion matrix for 5-way verbs clustering (ESLLI Workshop 2008)

It is interesting to notice the change of the context-window parameter in the best performing model: while the meaning of concrete nouns are better captured by looking at the sentence window, verbs are more influenced by the surrounding words. From Table 6 we could see that the features which are found to be more descriptive of the classes mostly are not nouns and adjectives as before, but adverbs and auxiliary verbs.

5.3 Argument Structure Distinction

As what we have done so far, we report the results of the best performing CCG-DSM, which again is

Model	9-way		5-way		Average	Average
	Entropy	Purity	Entropy	Purity	Entropy	Purity
Van de Cruys (dependency)	0.408	0.556	0.464	0.667	0.436	0.612
CCG-DSM-compl-2win-epmi	0.340	0.600	0.305	0.756	0.323	0.678
PoS-DSM-compl-2win-epmi	0.351	0.622	0.364	0.733	0.358	0.678
Van de Cruys (BoW)	0.442	0.556	0.463	0.600	0.453	0.578

Table 4: Verbs clustering result (ESLLI 2008 Workshop classification)

Cluster	Descriptive features
0	upon .(S/S)/(S[ng]\NP) 1.1%, smoothly _(S\NP)\(S\NP) 1.0%, bicycle _N 0.9%, past ._(S\NP)\(S\NP)/NP 0.9%, see _(S[pss]\NP)/(S[ng]\NP) 0.9%,
1	and _(S\NP)/(S\NP) 3.4%, password _N/PP 2.1%, worth _(S[adj]\NP)/(S[ng]\NP) 2.1%, have _(S[dc]\NP)/(S[pt]\NP)/NP 1.5%, openly _(S\NP)\(S\NP) 1.3%,
2	damaged _N/N 3.1%, trigger _N 2.9%, wound _(S[pss]\NP) 2.9%, sharply _(S\NP)\(S\NP) 2.7%, apart _(S[adj]\NP) 2.7%,
3	eat _(S[b]\NP) 5.7%, make _(S[b]\NP)/S[dc]\NP 3.3%, deeply _(S\NP)\(S\NP) 2.8%, make _(S[dc]\NP)/S[dc]\NP 2.8%, like _(S[dc]\NP)/S[dc]\NP 2.5%,
4	that _(S[bem]/S[b]) 3.2%, evidence _N/(S[to]\NP) 2.4%, Right _N 2.1%, effectiveness _N/PP 1.7%, tribute _N 1.6%,

Table 6: Descriptive features for 5-way verbs clustering (ESLLI Workshop 2008)

the model presented above: dimensions are both function and content words, the window context of size 2, with PPMI weighting schema (CCG-DSM-compl-2win-ppmi). The model obtains 0.544 entropy and 0.772 purity and it outperforms the PoS-DSM. Using the same experiment setup, PoS-DSM is able to cluster the verbs with 0.719 purity and 0.658 entropy.

Table 7 and Table 8 provide an error analysis of this task. The most common mistake is that verbs of the *object-drop* class, such as “carve”, “clean”, “knit”, “pack”, “swallow”, and “wash” are considered to be of *unaccusative* class by the model. While “divide” and “open”, which belong to *unaccusative* class, are clustered together into the *object-drop* class.

Interestingly, the descriptive features relevant for this classification task carry several tense features. Recall, the feature abbreviations are: $S[dc]$ (declarative sentences), $S[b]$ (bare infinitives, subjunctives and imperatives), $S[to]$ (to-infinitives)

Cluster	Classes			Entropy	Purity
	unacc ¹	objdrop ²	unerg ³		
0	0	3	16	0.397	0.842
1	2	11	0	0.391	0.846
2	17	6	2	0.734	0.680

¹ unaccusative ² object-drop ³ unergative

Table 7: Confusion matrix for argument structure distinction (Merlo & Stevenson)

Cluster	Descriptive features
0	around _PR 0.7%, see _(S[pss]\NP)/(S[ng]\NP) 0.7%, around _(S\NP)\(S\NP) 0.6%, around ._(S\NP)\(S\NP)/(S[ng]\NP) 0.5%, around _PP/PP 0.5%, along _(S\NP)\(S\NP) 0.4%, off _PR 0.4%, past ._(S\NP)\(S\NP)/NP 0.4%, see ._(S[dc]\NP)/(S[ng]\NP)/NP 0.3%, backward _N 0.3%
1	begin _(S[b]\NP)/(S[ng]\NP) 0.2%, start _(S[dc]\NP)/(S[ng]\NP) 0.1%, begin _(S[dc]\NP)/(S[ng]\NP) 0.1%, eligible _(S[adj]\NP)/(S[to]\NP) 0.1%, continue _(S[b]\NP)/(S[ng]\NP) 0.1%, start _(S[pt]\NP)/(S[ng]\NP) 0.1%, start _(S[b]\NP)/(S[ng]\NP) 0.1%, continue _(S[dc]\NP)/(S[ng]\NP) 0.1%, try _(S[b]\NP)/(S[ng]\NP) 0.1%, Manor _N 0.1%
2	partially _(S\NP)/(S\NP) 0.3%, gently _(S\NP)\(S\NP) 0.3%, slowly _(S\NP)\(S\NP) 0.2%, completely _(S\NP)\(S\NP) 0.2%, once _(S/S)/(S[pss]\NP) 0.2%, liquid _N 0.2%, begin _(S[dc]\NP)/(S[to]\NP) 0.2%, start _(S[dc]\NP)/(S[to]\NP) 0.2%, start _(S[pt]\NP)/(S[to]\NP) 0.2%, gently _(S\NP)\(S\NP) 0.2%

Table 8: Descriptive features for argument structure distinction (Merlo & Stevenson)

$[pss]$ (past participles in passive mode), $S[ng]$ (present participles), $S[pt]$ (past participles used in active mode). Merlo and Stevenson (2001) theory indeed has foreseen the relevance of the distinction between passive vs. active voice, as well as the usage of past-participle vs. simple past.

From the descriptive features of each cluster we could infer that *unergative* verbs are verbs which tend to occur together with “around”, “along”, or

”past”; the verbs of *object-drop* class tend to co-occur with ”begin” or ”start” in the form of gerund (e.g. ”begin playing”, ”start studying”); whereas the verbs of *unaccusative* class usually occur together with ”begin” or ”start” in to-infinitive form (e.g. ”start to melt”, ”begin to boil”).

5.4 Positive and Negative Verbs

We report the results obtained by the best performing CCG-DSM, namely the one with both function and content words as dimensions, context window of size 2, and PLOG weighting scheme (CCG-DSM-compl-2win-plog). The model achieves 0.946 purity and 0.255 entropy. However, the same results are obtained also by the PoS-DSM using the same parameters.

Cluster	Classes		Entropy	Purity
	positive	negative		
0	500	54	0.461	0.903
1	0	446	0.000	1.000

Table 9: Confusion matrix for positive vs. negative verb clustering

Looking at the confusion matrix shown in Table 9, it can be seen that the model assign 54 negative verbs to the cluster of positive verbs (cluster 0). Some of the negative verbs that are failed to be clustered as negative verbs are not strictly negative, for instance, ”blow”, ”hang”, ”issue”, and ”knock”. However, there are also other verbs that the model fails to recognize as negative which obviously have negative nuance, such as ”break”, ”die”, ”kill”, and ”reject”.

Cluster	Descriptive features
0	the _NP/N 0.2%, to _(S[to]\NP)/(S[b]\NP) 0.2%, and _conj 0.2%, a _NP/N 0.2%, be _(S[dcl]\NP)/(S[pss]\NP) 0.2%, have _(S[dcl]\NP)/(S[pt]\NP) 0.2%, it _NP 0.2%, in _(S\NP)\(S\NP)/NP 0.2%, of _PP/NP 0.1%, they _NP 0.1%
1	the _NP/N 2.4%, and _conj 2.0%, to _(S[to]\NP)/(S[b]\NP) 1.9%, be _(S[dcl]\NP)/(S[pss]\NP) 1.6%, a _NP/N 1.5%, have _(S[dcl]\NP)/(S[pt]\NP) 1.3%, by _(S\NP)\(S\NP)/NP 1.3%, he _NP 1.2%, it _NP 1.2%, they _NP 1.2%

Table 10: Descriptive features for positive vs. negative verbs clustering

The descriptive features behind this clustering are reported in Table 10. Quite impressively, the descriptive features are dramatically changed with respect to the ones seen so far. They are all function words, we see for the first time an important

role to be played by pronouns, prepositions, coordination and determiners.

5.5 Downward Monotonic Verbs

We report the results obtained by the best performing CCG-DSM (CCG-DSM-compl-2win-empi) with 0.732 entropy and 0.786 purity, and the confusion matrix is shown in Table 11. Out of the 28 downward monotonic verbs, CCG-DSM misses to consider only three verbs as such, namely ”doubt”, ”luck”, and ”withstand”. The verbs that are wrongly considered downward monotonic are: ”acknowledge”, ”address”, ”convince”, ”cooperate”, ”demand”, ”halt”, ”merge”, ”outline”, and ”reconstruct”.

Cluster	Classes		Entropy	Purity
	non-DM	DM		
0	9	25	0.834	0.735
1	19	3	0.575	0.864

Table 11: Confusion matrix for non-DW vs. DW monotonic verb clustering

Cluster	Descriptive features
0	rom _PP/(S[ng]\NP) 1.3%, sug- gestion _N/S[em] 1.3%, tempor- arity _(S\NP)\(S\NP) 1.1%, possibility _N/S[em] 0.8%, strictly _(S\NP)\(S\NP) 0.7%, until _(S\NP)\(S\NP)/PP 0.7%, re- lations _N 0.7%, act _S[ng]\NP 0.6%, strongly _(S\NP)/(S\NP) 0.5%, no- tion _N/S[em] 0.5%
1	seriously _(S\NP)/(S\NP) 3.9%, heav- ily _(S\NP)\(S\NP) 2.0%, knee _N/PP 2.0%, yourself _NP 2.0%, needle _N 1.8%, time _N/(S[ng]\NP) 1.8%, reward _N/PP 1.5%, siege _N 1.4%, reason _N/(S[to]\NP) 1.4%, duty _(N/PP)/PP 1.3%

Table 12: Descriptive features for non-DW vs. DW monotonic verbs clustering

Interestingly, the downward entailing verbs are recognized mostly by means of preposition and adverbs as specified in Table 12. Using the same experimental set up, the PoS-DSM performs worse with 0.966 entropy and 0.607 purity. The model fails to recognize the downward monotonic verbs, assigning 12 downward entailing verbs in one cluster and 16 in the other.

6 Conclusions

We have shown that while the richer CCG tags encoding both constituent structures and some other information, such as verb tense features and bare

vs. not-bare noun distinctions, they are not so relevant for noun classification. However, they indeed play an important role in distinguishing some classes of verbs. Thus, embedding CCG categories in the semantic space might be useful to give better representation of the meaning of verbs.

On the one hand, the CCG-DSM obtains equal results with PoS-DSM in distinguishing positive vs. negative verbs and concrete nouns, and on the later task the dependency model obtains better results. On the other hand, the CCG-DSM outperforms the dependency based one for the verb classification inspired by Levin's classes, with the average purity of 0.678 vs. 0.612 and the average entropy of 0.323 vs. 0.436. It outperforms PoS-DSM in the argument structure based distinction proposed by Merlo and Stevenson (2001) as well as in detecting downward entailing verbs.

Moreover, the experiments show that the size of context window have different impacts in the different classification tasks. The sentence context window is more informative for representing the meaning of nouns, whereas for verbs the more relevant information for distinguishing their classes is found within the context window of size 2.

Finally, while content words are the dimensions required by the semantic space of nouns to better picture them, verbs require to also consider function words. In particular, to distinguish negative from positive verbs (in the sense of sentiment analysis) a major role is played by grammatical words like coordination, pronouns, and prepositions; whereas adverbs seems to be more relevant for recognizing downward entailing verbs.

References

- Marco Baroni and Alessandro Lenci. 2009. One distributional memory, many semantic spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, October.
- Stephen Clark and James Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.
- Tim Van de Cruys. 2008. A comparison of bag of words and syntax-based approaches for word categorization. In M. Baroni, S. Evert, and A. Lenci, editors, *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, pages 47–54.
- Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Richard Duce. 2009. Without a 'doubt'? Unsupervised discovery of downward-entailing operators. In *Proceedings of NAACL HLT*, pages 137–145.
- Katrin Erk and Sebastian Pado. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP 2008*.
- Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop*.
- Julia Hockenmaier. 2003. *Data and Models for Statistical Parsing with Combinatory Categorical Grammar*. Ph.D. thesis, University of Edinburgh.
- M. Honnibal, J. R. Curran, and J. Bos. 2007. Rebanking CCGBank for improved interpretation. In *Proceedings of the 48th annual meeting of ACL*, pages 207–215.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04*, pages 168–177, New York, NY, USA. ACM.
- George Karypis, 2003. *CLUTO: A Clustering Toolkit*.
- William A. Ladusaw. 1980. *Polarity Sensitivity as Inherent Scope Relations*. Garland Press, New York. Ph.D. thesis date 1979.
- Beth Levin. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago, Ill.: University of Chicago Press.
- Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, November.
- Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Comput. Linguist.*, 27(3):373–408, September.
- Mark Steedman. 2000. *The Syntactic Process*. Cambridge, MA: MIT Press.
- P. Vinson and G. Vigliocco. 2008. Feature norms for a large set of object and event concepts. *Behavior Research Methods*, 40(1):183–190.
- Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis.