Computing Semantic Compositionality in Distributional Semantics

Emiliano Guevara
Tekstlab, ILN, University of Oslo, e.r.guevara@iln.uio.no

Abstract

This article introduces and evaluates an approach to semantic compositionality in computational linguistics based on the combination of Distributional Semantics and supervised Machine Learning. In brief, distributional semantic spaces containing representations for complex constructions such as Adjective-Noun and Verb-Noun pairs, as well as for their constituent parts, are built. These representations are then used as feature vectors in a supervised learning model using multivariate multiple regression. In particular, the distributional semantic representations of the constituents are used to predict those of the complex structures. This approach outperforms the rivals in a series of experiments with Adjective-Noun pairs extracted from the BNC. In a second experimental setting based on Verb-Noun pairs, a comparatively much lower performance was obtained by all the models; however, the proposed approach gives the best results in combination with a Random Indexing semantic space.

1 Introduction

Probably the most important missing ingredient from the current NLP state-of-the-art is the ability to compute the meaning of complex structures, i.e. semantically compositional structures. In this paper, I propose a methodological approach and a series of experiments designed to teach computers the ability to compute the compositionality of (relatively simple) complex linguistic structures. This work uses a combination of Distributional Semantics and Machine Learning techniques. The starting data in the experiments reported below are multidimensional vectorial semantic representations extracted from electronic corpora. This work extends the basic methodology presented in Guevara (2010) with new data collection techniques, improved evaluation metrics and new case studies.

Compositionality is probably one of the defining properties of human language and, perhaps, a nearly uncontroversial notion among linguists. One of the best-known formulations of compositionality is:

(1) *The Principle of Compositionality:*
The meaning of a complex expression is a function of the meaning of its parts and of the syntactic rules by which they are combined. (Partee, ter Meulen and Wall, 1990: 318)

The Principle of Compositionality is a standard notion in many different fields of research, notably in logic, in philosophy of language, in linguistics and in computer science; this intrinsic multi-disciplinarity makes tracing back its recent history somewhat difficult.

The recent years have witnessed an ever-increasing interest in techniques that enable computers to automatically extract semantic information from linguistic corpora. In this paper I will refer to this new field in general as Distributional Semantics. Distributional Semantics, in short, extracts spatial representations of meaning from electronic corpora by using distributional (i.e. statistical) patterns of word usage. The main hypothesis in Distributional Semantics is the so-called *distributional hypothesis of meaning*, expressing the fact that "words that occur in the same contexts tend to have similar meanings" (Pantel, 2005). The distributional hypothesis of meaning is ascribed to Zellig Harris, who proposed a general distributional methodology for linguistics.

Since representations in Distributional Semantics are spatial in nature (e.g. vectors representing points in a multidimensional space), differences in meaning are captured through differences in location:

in the multidimensional space, two semantically (i.e. *distributionally*) similar words are *closer* than two words that are dissimilar. See Sahlgren (2006) and Turney and Pantel (2010) for detailed overviews of the methodology and applications of Distributional Semantics.

2 Compositionality in distributional semantics: state-of-the-art

I stressed above that computers are still not able to deal with the compositionality of meaning. However basically true, this statement should be qualified somewhat. Previous work in the field has produced a small number of operations to approximate the composition of vectorial representations of word meaning. In particular, given two independent vectors $v1$ and $v2$, the semantically compositional result $v3$ is modelled by one of the following four basic operations: vector addition, vector pointwise-multiplication, tensor product or linear regression.

In the literature on Information Retrieval, *vector addition* is the standard approach to model the composed meaning of a group of words (or a document) as the sum of their vectors (see, among many others, Widdows, 2004: ch. 5). More schematically:

(2) *Vector addition:* $v1_i + v2_i = v3_i$
Given two independent vectors $v1$ and $v2$, the compositional meaning of $v3$ consists of the sum of the corresponding components of the original vectors.

Mitchell and Lapata (2008) introduce a whole family of models of compositionality based on vector addition and pointwise-multiplication (and a weighted combination of both), evaluated on a sentence similarity task inspired by Kintsch (2001). While the additive model captures the compositionality of meaning by considering all available components, multiplicative models only operate on a subset of them, i.e. non-zero components. They claim that when we pointwise-multiply the vectors representing two words, we obtain an output that captures their composition; actually, this operation is keeping in the output only the components which had corresponding non-zero values: whether this operation has any relation with semantics is still unclear. However, in their experiments, Mitchell and Lapata prove that the pointwise-multiplicative model and the weighted combination of the additive and the multiplicative models perform equally well. Of these, only the simple multiplicative model will be tested in the experiments I present in the following section.

(3) *Vector pointwise multiplication:* $v1_i \times v2_i = v3_i$
Each corresponding pair of components of $v1$ and $v2$ is multiplied to obtain the corresponding component of $v3$.

Widdows (2008) proposes to apply a number of more complex vector operations imported from quantum mechanics to model composition in semantic spaces, in particular *tensor product* and the related operation of *convolution product*. Widdows (2008) obtains results indicating that both the tensor product and the convolution product perform better than the simple additive model in two small experiments (relation extraction and phrasal composition). Giesbrecht (2009) presents a more complex task, singling out non-compositional multiword expressions. Her results clearly show that tensor product outperforms vector addition, multiplication and convolution.

(4) *Tensor product:* $v1 \otimes v2 = v3$
where $v3$ is a matrix whose *ij*-th entry is equal to $v1_i \times v2_j$

However, since the tensor product (also called outer product) of two vectors produces a result with higher dimensionality (a matrix), it cannot be directly compared against the other methods, which instead generate compositional representations in the same original space. In the experiments reported in the following section, we will use the circular convolution composition method (Plate, 1991): in brief, circular convolution is a mathematical operation that effectively compresses the tensor product of two vectors onto the original space, thus allowing us to compare its outcome with that of the other methods here reviewed.

(5) *Circular convolution:* $v1 \otimes v2 = v3$
where $v3 = \sum_{j=0}^{n-1} v1_j v2_{i-j}$

It is interesting to note that a great deal of attention has recently been devoted to the tensor product as the basic operation for modelling compositionality, even at the sentential level (e.g. Grefenstette *et al*. 2010), through a combination of mathematical operations and symbolic models of logic (inspired by Clark and Pulman, 2007). Although extremely motivating and thought provoking, these proposals have not been tested on empirical grounds yet.

A common thread ties all the approaches briefly outlined above: all information that is present in the systems is conveyed by the vectors $v1$ and $v2$, e.g. the independent word representations, while completely disregarding $v3$ (the composed vector). Furthermore, all of these approaches are based on the application of a single geometric operation on the independent vectors $v1$ and $v2$. It seems highly unlikely that just one geometric operation could reliably represent *all* the semantic transformations introduced by *all* syntactic relations in *every* language.

Guevara (2010) and Baroni and Zamparelli (2010) introduce a different approach to model semantic compositionality in distributional spaces by extracting context vectors from the corpus also for the composed vector $v3$. For example, Guevara collects vector representations for *nice* and *house*, but also for the observed pair *nice_house*. With these data, a model of Adjective-Noun (AN) compositionality is built by using a supervised machine learning approach: multivariate multiple linear regression analysis by partial least squares. This method is able to learn the transformation function that best approximates $v3$ on the basis of both $v1$ and $v2$. Baroni and Zamparelli (2010) use a slightly different methodology: assuming that each adjective is a linear transformation function (i.e. the function to be learnt by the algorithm), they model AN compositionality by approximating $v3$ only on the basis of $v2$ (the noun) but running a different regression analysis for each adjective in their data.

The approach proposed by Guevara (2010) is really only an extension of the full additive model of Mitchell and Lapata (2008), the only difference being that adopting a supervised learning methodology ensures that the weight parameters in the function are estimated optimally by linear regression. In the following section, I present a new series of experiments that refine, extend and improve this approach to model the compositionality of adjacent AN and VN pairs by linear regression.

(6) *Compositionality by regression:* $Av1 + Bv2 = v3$
where A and B are weight matrices estimated by the supervised learning algorithm using multivariate multiple linear regression.

3 Compositionality by regression

Let us reconsider the highly underspecified definition of the Principle of Compositionality. Let us start by setting the *syntactic relation* that we want to focus on for the purposes of this study: following Guevara (2010) and Baroni and Zamparelli (2010), I model the semantic composition of adjacent Adjective-Noun pairs expressing attributive modification of a nominal head. In a second analogous experiment, I also model the *syntactic relation* between adjacent Verb-Noun expressing object selection by the verbal head.

The *complex expression* and its *parts* are, respectively, adjacent Adjective-Noun and Verb-Noun[1] pairs and their corresponding constituents (respectively, adjectives and nouns, verbs and nouns) extracted from the British National Corpus. Furthermore, the *meaning* of both complex expressions and their constituents is assumed to be the multidimensional context vectors obtained by building semantic spaces.

What remains to be done, therefore, is to model the *function* combining meanings of the constituent parts to yield the meaning of the resulting complex expression. This is precisely the main assumption made in this article. Since we are dealing with multidimensional vector representations of meaning, we suggest that compositionality can be interpreted as a *linear transformation function* mapping two

[1]Actually, the extracted Verb-Noun pairs are not always strictly adjacent, an optional determiner was allowed to occur between verb and noun. Thus, phrases such as "raise money" and "visit a client" were both included.

independent vectors in a multidimensional space into a composed vector in the same space. Moreover, considering that each component in the independent vectors $v1$ and $v2$ is a candidate predictor, and that each component in the composed vector $v3$ is a dependent variable, it is proposed to formulate compositionality of meaning in Distributional Semantics as a problem of multivariate multiple regression. Such a formulation allows us to model compositionality by applying well-known standard machine learning techniques such as the Multilayer Perceptron or Support Vector Machines.

However, since word sequences in corpora tend to have low frequency distributions (usually lower than the frequency of their constituents) and very sparse vectorial representations, it is very difficult to build datasets where the number of observations (the size of the dataset) is greater than the number of variables considered (the dimensions of the vector in the dataset). This issue is known as the *curse of dimensionality*, and specific mathematical techniques have been developed to deal with it. In our experiments, we use one such regression technique, Partial Least Squares.

3.1 Partial least squares regression

Partial Least Squares Regression (PLS) is a multivariate regression technique that has been designed specifically to treat cases where the curse of dimensionality is a serious issue. PLS has been successfully applied in a wide range of different scientific fields such as spectroscopy, chemistry, brain imaging and marketing (Mevik and Wehrens, 2007).

PLS predicts the output matrix Y from information found in both the input matrix X and in Y. It does so by looking for a set of *latent variables* in the data that perform a simultaneous decomposition of both matrices while trying to explain as much as possible of the covariance between X and Y. Next, PLS carries out regression using the decomposition of X to predict Y. Thus, PLS performs the prediction by extracting the latent variables with the best predictive power. PLS is a robust regression technique that is particularly efficient in situations with a high number of predictors and few observations (Abdi, 2007, Hastie *et al.*, 2009). Standard linear regression will fail in such cases.

3.2 Experimental setup

3.2.1 Corpus and construction of the dataset

Using a lemmatised and POS tagged version of the BNC, a list of adjacent AN pair candidates was extracted with simple regex-based queries targeting sequences composed of [Det/Art–A–N] (i.e. pairs expressing attributive modification of a nominal head like *'that little house'*). In order to ensure the computational attainability of the successive steps, the candidate list was filtered by frequency (> 400) obtaining 1,367 different AN pairs.

A new version of the BNC was then prepared to represent the selected AN lemma pairs as a single token; for example, while in the original BNC the phrase [*nice houses*] consists in two separate POS-tagged lemmas, *nice_AJ* and *house_NN*, in the processed corpus it appears as a single entry *nice_AJ_house_NN*). The corpus was also processed by stop-word removal (very high frequency items, mainly functional morphemes). The re-tokenization process of the BNC enables us to extract independent context vectors for each AN pair in our list ($v3$) and their corresponding constituents (A and N, respectively $v1$ and $v2$), while ensuring that the extracted vectors do not contain overlapping information.

The same preprocessing steps were carried out to extract VN pair candidates. Sequences composed of [V-(Det/Art)–N] with an optional determiner were targeted and filtered by frequency (> 400), resulting in a first list of 545 VN pairs. This list contained a large amount of noise due to lemmatisation and POS-tagging problems (e.g. *housing association*), and it also contained many very frequent lexicalized items (e.g. *thank goodness*). The list was manually cleaned, resulting in 193 different VN pairs.

3.2.2 Building semantic spaces and composition models

For each syntactic relation (AN and VN), two different semantic spaces were built with the S-Space package (Jurgen and Stevens, 2010): a *Hyperspace Analogue to Language* space (HAL, Burgess and

Lund, 1997) and a *Random Indexing* space (RI, Sahlgren, 2006). The spaces were built using the same vocabulary, the 23,222 elements in the corpus with a frequency $\geq$ 100 (comprising both individual lemmas and all the selected AN pairs) and the same contextual window of 5 words to the left and to the right of the target (either a word or a AN/VN pair).

HAL is a co-occurrence based semantic space that corresponds very closely to the well-known *term-by-term* matrix collection method. However, given the size of our vocabulary, the resulting matrix is extremely large (23,222 $\times$ 23,222). HAL reduces the dimensionality of the space by computing the variances of the row and column vectors for each word, and discarding the elements with lowest variance. The dimensionality of this space was reduced to the 500 most informative dimensions, thus ending with a size of 23,222 $\times$ 500. The vectors in this space were normalised before the successive steps.

RI avoids the problem of dimensionality of semantic spaces by applying a different strategy to collect the context vectors. Each word in the corpus is assigned an initial unique and randomly generated *index vector* of a fixed size. As the algorithm scans the corpus one token at a time, the vector of the target word is incrementally updated by combining it with the index vector of the context. In order to keep the comparability of the built spaces, the RI space was built with 500-dimensional index vectors, thus obtaining a space of 23,222 $\times$ 500 dimensions. The vectors in this space were also normalised.

With the AN/VN pair vectors and their corresponding constituents (respectively $v3$, $v1$ and $v2$), four different models of compositionality were built from each semantic space (HAL and RI) in each of the considered syntactic relations:

- an additive model (ADD) $v1_i + v2_i = v3_i$
- a simple multiplicative model (MUL) $v1_i \times v2_i = v3_i$
- a circular convolution model (CON) $v1 \otimes v2 = v3$
- a partial least squares regression model (PLS) $Av1 + Bv2 = v3$

In addition, two baseline models were introduced in the evaluation process. The baseline models were built by simply extracting the context vectors for the constituents in each pair from each space (A and N, V and N, respectively $v1$ and $v2$).

Of all the considered models, only PLS requires a stage of parameter estimation, i.e. training. In order to accomplish this, the data were randomly divided into a training set (1,000 AN pairs – 73%) and a test set (the remaining 367 AN pairs – 27%). In the much smaller VN dataset, the training set was built with 133 pairs (69%) and the test set with 60 pairs (31%). These parameters for the regression models were estimated by performing a 10-fold cross-validation in the training phase. All the models were built and evaluated using the R statistical computing environment and simple Python scripts. In particular, the regression analysis was carried out with the **pls** package (Mevik and Wehrens, 2007). After various preliminary trials, the PLS model's predictions were computed by using the first 50 latent variables.

3.3 Evaluation

The evaluation of models of compositionality is still a very uncertain and problematic issue. Previous work has relied mainly on "external" tasks such as rating sentence similarity or detection idioms. These evaluation strategies are "external" in the sense that each compared model produces a set of predictions which are then used in order to reproduce human annotation of datasets that do not have a representation in the semantic space under consideration. For example, Mitchell and Lapata (2008) use their models to approximate the human ratings in their sentence similarity dataset. Giesbrecht (2009) also uses human annotated data (manually classified collocations, compositional and non-compositional) in her evaluation task. However, any evaluation task requiring hand-annotated datasets will have a considerable cost in resource building. At present time, there are no suitable datasets in the public domain.

I propose instead to take a radically different point of view, developing "internal" evaluation tasks that try to measure how well the proposed models approximate the distributional patterns of corpus-extracted composed vectors. That is to say, I want to compare the predicted output of every model (i.e. a predicted context vector for $v3$) with the real observation of $v3$ that was collected from the corpus. The

following subsections present a few experimental evaluation methods based on neighbour analysis and on the Euclidean measure of distance.

The evaluation strategies here presented rests on the sensible assumption that if a model of AN compositionality is reliable, its predicted output for any AN pair, e.g. *weird_banana*, should be in principle usable as a substitute for the corresponding corpus-attested AN vector. Moreover, if such a model performs acceptably, it could even be used predict the compositionality of unattested candidates like *shadowy_banana*: this kind of operations is the key to attaining human-like semantic performance.

3.3.1 Correlation analysis between modelled predictions and observations

Let us start the comparative evaluation of the modelled predictions by considering the results of a series of Mantel correlation tests. First, distance matrices were computed for the observations in the test sets and then the same was done for each of the prediction models. Then, each of the models' distance matrices was compared against the distance matrix of the observations trying to determine their degree of correlation. The null hypothesis in each Mantel test is that the distance matrices being compared are unrelated. The aim of this task is similar to the evaluation method used by Mitchell and Lapata (2008): we try to find out which model has the strongest correlation with the original data, with the difference that in our case no "external" human ratings are used.

	HAL		RI	
Model	*Correlation*	*Simul. p-value*	*Correlation*	*Simul. p-value*
PLS	**0.5438081**	0.001	**0.4341146**	0.001
ADD	0.5344057	0.001	0.3223733	0.001
MUL	0.3297359	0.001	0.1811038	0.002
CON	-0.05557023	0.956	-0.02584655	0.727

Table 1: Adjective-Noun pairs. Mantel tests of correlation (max. correlation = 1)

Considering the results for the AN dataset in Table 1, with the PLS and ADD models we can reject the null hypothesis that the two matrices (distance matrix between the observed AN pairs and distance matrix between each model's predictions) are unrelated with p-value = 0.001 in both the semantic spaces (HAL and RI). MUL also allows the null hypothesis to be rejected, but with a lower correlation (and with a greater p-value = 0.002 in RI). Having obtained the highest observed correlation in both settings, the PLS model is highly positively associated with the observed data. Also ADD and MUL have produced predictions that are positively correlated with the observed AN vectors. CON is not correlated with the original data. In other words, PLS and ADD seem to be much better that the remaining models in reproducing unseen AN pairs; overall, however, PLS produces the closest approximation of the corpus-based test set. Finally, although both semantic spaces (HAL and RI) produce the same ordering among the models, it seems that the predictions using the HAL space are relatively closer to the observed data.

	HAL		RI	
Model	*Correlation*	*Simul. p-value*	*Correlation*	*Simul. p-value*
PLS	0.2186435	0.003	0.1113741	0.116
ADD	**0.4094653**	0.001	**0.1290508**	0.124
MUL	0.1375934	0.042	-0.08865458	0.8
CON	0.05153776	0.174	-0.08186146	0.807

Table 2: Verb-Noun pairs. Mantel tests of correlation (max. correlation = 1)

Turning to the VN dataset, the obtained results are much less promising (see Table 2). As a general observation, the correlations between each of the models and the observations are very low, except for ADD in the HAL semantic space. In addition, ADD obtains the best correlation also in the RI space. PLS comes in second place. Given that PLS is based on the estimation of parameters from training data, its low performance can be attributed to the size of dataset (only 133 VN examples used for training). On

the contrary, ADD, MUL and CON do not have this excuse and their extremely low performance must be due to other factors. Finally, it is very clear that HAL produces better correlations for all the models.

3.3.2 Observation-based neighbour analysis

For this and for the remaining evaluation protocols, a preliminary step was taken. Since our intention is to base the evaluation on the analysis of nearest neighbours, we extracted an identical subset of the built semantic spaces (HAL and RI, which originally had a vocabulary of 23,222 items) in order to compute a distance matrix of a manageable size.

In the Adjective-Noun dataset, the extracted subset comprises vectors for all the observed AN vectors in both the training and test sets (1,367 items), all the corresponding predictions, the NOUN- and ADJ-baseline models, the 2,500 most frequent nouns (not included in the baseline) and the 2,500 most frequent adjectives (not included in the baseline). The distance matrix for the selected sub-space was then created by using the Euclidean measure of distance, resulting in a 8,666 $\times$ 8,666 matrix.

The Verb-Noun dataset was treated in the same way, extracting vectors for all the VN observations, the corresponding predictions from each model, the VERB- and NOUN-baseline models and the 1,000 most frequent nouns and verbs in the space (not overlapping with the baselines); this resulted in a 2,420 $\times$ 2,420 distance matrix.

Following Guevara's (2010) neighbour analysis, for each observed AN pair in the test datasets, the list of n-top neighbours were extracted from the distance matrix (n=10 and n=20). Then, the resulting neighbour lists were analysed to see if any of the modelled predictions was to be found in the n-top list. The ADJ- and NOUN-baselines were introduced in the evaluation to further compare the appropriateness of each model. Below we only report the results obtained with n=20, but very similar results were obtained in the 10-top neighbour setting.

As can be observed from Table 3, in the HAL space, PLS obtains the highest score, followed by the NOUN-baseline at a short distance and then by the ADJ-baseline at a greater distance. The performance of the remaining models is negligible. A different situation can be seen for the RI space, where the winner is the NOUN-baseline followed by PLS and ADJ.

	HAL	RI
Model	*Predictions found*	*Predictions found*
ADD	0	0
CON	0	0
MUL	3	0
PLS	**152**	112
ADJ	32	53
NOUN	123	**144**

Table 3: AN pairs. Observation-based neighbour analysis (max. score = 367)

It is interesting to see that PLS is actually competing against the NOUN-baseline alone, being the rival models almost insensible to the evaluation task. This same pattern will be seen in the other evaluation tasks. Furthermore, the score differences obtained by PLS and the NOUN-baseline are significant (HAL p-value = 0.03275, RI p-value = 0.01635, 2-sample test for equality of proportions).

The VN dataset gave much poorer results, once more. In fact, it is almost pointless to comment anything except that only MUL was able to rank its predictions in top-20 neighbours six times (only in the HAL space) and that PLS managed to do the same 9 times (only in the RI space). The maximum score in this setting was 60.

3.3.3 Prediction-based neighbour analysis

Building on the previous neighbour analysis, a new task was set up by changing the starting point for neighbour extraction. In this case, for each modelled AN pair in the test dataset in each composition

model, the list of *n*-top neighbours were extracted from the distance matrix (*n*=10 and *n*=20). Then, the resulting neighbour lists were analysed to see if the originally observed corresponding AN pair was to be found in the *n*-top list. The same procedure was used with the VN dataset.

Below we only report the results obtained with *n*=20, but very similar results were obtained in the 10-top neighbour setting. This task at first did not seem to be particularly difficult, but the obtained results were very poor.

	HAL	RI
Model	*Predictions found*	*Predictions found*
ADD	2	0
CON	0	0
MUL	0	0
PLS	**32**	**25**
ADJ	6	2
NOUN	26	16

Table 4: AN pairs. Prediction-based neighbour analysis (max. score = 367)

The winner in this experiment was PLS, once again followed by the NOUN-baseline. However, the score differences obtained by PLS and the NOUN-baseline are not significant (HAL p-value = 0.4939, RI p-value = 0.1985, 2-sample test for equality of proportions). The main observation to be made is that the obtained scores are surprisingly low if compared with the previous evaluation task. The reason for this difference is to be found in the homogeneity and specialization that characterizes each of the models' neighbour sets: each model produces predictions that are relatively very close to each other. This has the consequence that the nearest neighbour lists for each model's predictions are, by and large, populated by items generated in the same model, as shown in Table 5. In conclusion, although PLS obtained the highest score in this task, we cannot be sure that it performed better than the NOUN-baseline. In any case, the remaining composition models did not reach the performance of PLS.

Model	*Same-model items*
ADD	3626 (98,8 %)
CON	3670 (100 %)
MUL	3670 (100 %)
PLS	2767 (75,4 %)
NOUN	1524 (41,5 %)
ADJ	1382 (36,1 %)

Table 5: AN pairs. Same-model neighbours in each models' top-10 list of neighbours extracted from HAL semantic space (total items in each list = 3670)

The VN dataset once again did not produce interesting results. As a brief note, ADD won in the HAL space (but managing to score only two observations in its predictions' top-20 neighbours) while PLS won in the RI space as before, scoring 5 observations in its predictions' top-20 neighbours (max. score 60).

3.3.4 Gold-standard comparison of shared neighbours

Our previous evaluation methods targeted the distance between predictions and observations, i.e. the ability of each model to reproduce unseen AN/VN pairs. Changing perspective, it would be desirable to test if the models' predictions show a similar distributional behaviour with respect to the corresponding observed vector and to other words in the semantic space.

To test this idea, the *n*-top neighbour-lists (*n*=10 and *n*=20) for the observed AN/VN pairs were extracted and taken to be the gold-standard. Then, each prediction's *n*-top list of neighbours was analysed looking for shared neighbours with respect to the corresponding gold-standard list. Each time a shared neighbour was found, 1 point was awarded to the model.

Table 6 summarises the results obtained with n=20 (similar figures obtained with n=10) in the AN dataset. Although by a small margin, the winner in this task is PLS. Even if the obtained scores are still rather low (in the best cases, about 17% of all the available points were obtained), this experiment represents a significant improvement over Guevara's (2010) reported results, which reached only about 10% of the maximum score. Here again the same ordering of models can be observed: after PLS we find the NOUN- and ADJ-baselines, leaving the performance of the remaining models at a extremely modest level. Additionally, the score differences obtained by PLS and the NOUN-baseline are highly significant (HAL p-value = 2.363e-08, RI p-value = 0.0003983, 2-sample test for equality of proportions).

	HAL	RI
Model	*Shared neighbours*	*Shared neighbours*
ADD	28	0
CON	0	0
MUL	5	0
PLS	**1299**	**1267**
ADJ	259	534
NOU	1050	1108
Total shared:	2641	2909

Table 6: AN pairs. Gold-standard comparison of shared neighbours (max. score = 7340)

Table 7 summarises the results obtained in the VN dataset, which show a considerable improvement over the preceding evaluation methods. Here we have to clear winners, ADD in the HAL space and PLS in the RI space. Interestingly, although the numbers are still on the low side, ADD obtained 8.6% of the total points, with shared neighbours for 35 out of 60 VN pairs; PLS obtained 21% of the total, with shared neighbours for 40 out of 60 VN pairs. In particular this last score is (21%) is the highest one ever obtained with gold-standard comparison of shared neighbours (also considering Guevara's 2010 results).

	HAL	RI
Model	*Shared neighbours*	*Shared neighbours*
ADD	**103**	0
CON	0	0
MUL	31	0
PLS	0	**253**
VERB	0	0
NOUN	0	0
Total shared:	134	253

Table 7: VN pairs. Gold-standard comparison of shared neighbours (max. score = 1200)

4 Conclusions

This paper proposes an improved framework to model the compositionality of meaning in Distributional Semantics. The method, Partial Least Squares Regression, is well known in other data-intensive fields of research, but to our knowledge had never been put to work in computational semantics.

PLS outperformed all the competing models in the reported experiments with AN pairs. In particular, the PLS model generates compositional predictions that are closer to the observed composed vectors than those of its rivals. This is an extremely promising result, indicating that it is possible to generalize linear transformation functions beyond single lexical items in Distributional Semantics' spaces.

It is remarkable that PLS did not actually have to compete against any of the previously proposed approaches to compositionality, but only against the NOUN- and ADJ-baselines, and in particular against the former. This fact is expected from a theoretical point of view: since the Noun is the head of the AN pair, it is likely that the complex expression and its head share much of their distributional properties. PLS nearly always outperformed the NOUN-baseline, but only by small margins, which indicates that

there is a still plenty of space for improvement. Our experiments also show that AN compositionality by regression performs nearly equally well in semantic spaces of very different nature (HAL and RI).

The second dataset used in this paper contained VN pairs. Generally, this dataset did not produce good results with any of the considered approaches to model compositionality. This rather negative result may be due to its relatively smaller size, but this excuse may only be applied to PLS, the only model that relies on parameter estimation. Surprisingly, though, the gold-standard comparison of shared neighbours gave much better results, with ADD performing well in the HAL space and PLS performing very well in the RI space. Even if the VN dataset did not produce excellent results, it highlights some interesting issues. First, not all syntactic relations may be equally "easy" to model. Second, different evaluation methods may favor competing approaches. Finally, some approaches may be particularly successful with a specific distributional space architecture (like PLS and RI, and ADD and HAL).

This work has intentionally left the data as raw as possible, in order to keep the noise present in the models at a realistic level. The combination of Machine Learning and Distributional Semantics here advocated suggests a very promising perspective: transformation functions corresponding to different syntactic relations could be learned from suitably processed corpora and then combined to model larger, more complex structures, probably also recursive phenomena. It remains to prove if this approach is able to model the symbolic, logic-inspired kind of compositionality that is common in Formal Semantics; being inherently based on functional items, it is at present time very difficult and computationally intensive to attain, but hopefully this will change in the near future.

References

Abdi, H. (2007). Partial least squares regression. In N. Salkind (Ed.) *Encyclopaedia of Measurement and Statistics*. Thousand Oaks (CA), Sage.

Baroni, M. and A. Lenci. (2009). One semantic memory, many semantic tasks. In *Proc. of the EACL Workshop on Geometrical Models of Natural Language Semantics*, 3–11. Athens, ACL.

Baroni, M. and R. Zamparelli. (2010, to appear). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP 2010*.

Burgess, C. and K. Lund. (1997). Modeling parsing constraints with high-dimensional context space. *"Language and Cognitive Processes"*, 12, 177-210.

Clark, S. and S. Pulman. (2007). Combining symbolic and distributional models of meaning. In *Proceedings of the AAAI Spring Symposium on Quantum Interaction,* 52–55. Stanford (CA).

Giesbrecht, E. (2009). In Search of Semantic Compositionality in Vector Spaces. In *Proceedings of ICCS 2009*, Moscow, Russia, 173–184. Berlin, Springer.

Grefenstette, E. Bob Coecke, S Pulman, S. Clark and M. Sadrzadeh. (2010). Concrete Compositional Sentence Spaces. Talk presented at ESSLLI 2010, August 16-20, 2010, University of Copenhagen.

Guevara, E. (2010). A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. In *Proc. of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, 33-37. Uppsala, ACL.

Jurgens, D. and K. Stevens. (2010). The S-Space Package: An Open Source Package for Word Space Models. In *Proceedings of the ACL 2010 System Demonstrations,* 30-35. Uppsala, ACL.

Kintsch, W. 2001. Predication. *"Cognitive Science"*, 25 (2), 173–202.

Mevik, B.-H. and R. Wehrens. (2007). The pls package: principal component and partial least squares regression in R. *"Journal of Statistical Software"*, 18 (2), 1–24.

Mitchell, J. and M. Lapata. (2008). Vector- based Models of Semantic Composition. In *Proceedings of the 46th Annual Meeting of the ACL*, 236–244. Columbus, OH, ACL.

Pantel, P. (2005). Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Conference of the ACL*, 125–132. Morristown, ACL.

Partee, B.H., A. ter Meulen and R.E. Wall. (1990). *Mathematical methods in linguistics*. Dordrecht, Kluwer.

Plate, T.A. (1991). Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, 30–35. Sydney.

Sahlgren, M. (2006). *The Word Space Model*. Ph.D. dissertation. Stockholm University.

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

Widdows, D. (2004). *Geometry and Meaning*. Stanford, CSLI publications.

Widdows, D. (2008). Semantic Vector Products: Some Initial Investigations. Paper presented at the *Second AAAI Symposium on Quantum Interaction*. Oxford, 26th–28th March 2008.