

# Algebraic compositional models for semantic similarity in ranking and clustering

Paolo Annesi, Valerio Storch, Danilo Croce and Roberto Basili

Dept. of Computer Science,  
University of Roma Tor Vergata, Roma, Italy  
{annesi,croce,basili}@info.uniroma2.it  
storch@uniroma2.it

**Abstract.** Although distributional models of word meaning have been widely used in Information Retrieval achieving an effective representation and generalization schema of words in isolation, the composition of words in phrases or sentences is still a challenging task. Different methods have been proposed to account on syntactic structures to combine words in term of algebraic operators (e.g. tensor product) among vectors that represent lexical constituents.

In this paper, a novel approach for semantic composition based on space projection techniques over the basic geometric lexical representations is proposed. In the geometric perspective here pursued, syntactic bi-grams are projected in the so called *Support Subspace*, aimed at emphasizing the semantic features shared by the compound words and better capturing phrase-specific aspects of the involved lexical meanings. State-of-the-art results are achieved in a well known benchmark for phrase similarity task and the generalization capability of the proposed operators is investigated in a cross-linguistic scenario, i.e. in the English and Italian Language.

## 1 Introduction

With the rapid development of the World Wide Web and the spread of human-generated contents, Information Retrieval (IR) has many challenges in discovering and exploiting those rich and huge information resources. Semantic search [3] improves search precision and recall by understanding user's intent and the contextual meaning of concepts in documents and queries. Semantic search extends the scope of traditional information retrieval paradigms from mere document retrieval to entity and knowledge retrieval, improving the conventional IR methods by looking at a different perspective, i.e. the meaning of words. However, the language richness and its intrinsic relation to the world and human activities make semantic search a very complex task. In a IR system, a user can express its specific user need with a natural language query like "... *buy a car* ...". This request can be satisfied by documents expressing the abstract concept of *buying something* and in particular the focus of the action is a car. This information

can be expressed inside a document collection in many different forms, e.g. the quasi-synonymic expression ”... *purchase an automobile ...*”. Accounting on lexical overlap with respect to the original query, a Bag-of-words based system would instead retrieve different documents, containing expressions such as ”... *buy a bag ...*” or ”... *drive a car ...*”. A proper semantic generalization is thus needed, in order to derive the correct *composition* of the target words, i.e. an action like *buy* and an object like *car*.

While compositional approaches to language understanding have been largely adopted, semantic tasks are still challenging for research in Natural Language Processing. Traditional logic-based approaches (as the Montague’s approach in [17] and [2]) rely on Frege’s principle for which the meaning of a sentence is a function of the meanings of its parts [10]. The resulting theory allows an algebra on the discrete propositional symbols to represent the meaning of arbitrarily complex expressions. Despite the fact that they are formally well defined, logic-based approaches have limitations in the treatment of ambiguity, vagueness and cognitive aspects intrinsically connected to natural language.

On the other hand, distributional models early introduced by Schütze [21] rely on the Word Space model. Here semantic uncertainty is managed through the statistical analysis of large scale corpora. Linguistic phenomena are then modeled according to a geometrical perspective, i.e. points in a high-dimensional space representing semantic concepts, such as words, and can be learned from corpora, in such a way that similar, or related, concepts are near each other in the space. Methods for constructing representations for phrases or sentences through vector composition has recently received a wide attention in literature (e.g. [15, 23]). However, vector-based models typically represent isolated words and ignore grammatical structure [23]. Such models have thus a limited capability to model compositional operations over phrases and sentences.

In order to overcome these limitations a so-called compositional distributional semantics (DCS) model is needed and its development is still object of on-going and controversial research (e.g. [5], [11]). A compositional model based on distributional analysis should provide semantic information consistent with the meaning assignment that is typical of human subjects. For example, it should support synonymy and similarity judgments on phrases, rather than only on single words. The objective should be a measure of similarity between quasi-synonymic complex expressions, such as ”... *buy a car ...*” vs. ”... *purchase an automobile ...*”. Another typical benefit should be a computational model for entailment, so that the representation for ”... *buying something ...*” should be implied by the expression ”... *buying a car ...*” but not by ”... *buying time ...*”. Distributional compositional semantics (DCS) need thus a method to define: (1) a way to represent lexical vectors  $\mathbf{u}$  and  $\mathbf{v}$ , for words  $u, v$  dependent on the phrase  $(r, u, v)$  (where  $r$  is a syntactic relation, such as verb-object), and (2) a metric for comparing different phrases according to the selected representations  $\mathbf{u}, \mathbf{v}$ . Existing models are still controversial and provide general algebraic operators (such as tensor products) over lexical vectors.

In this paper, we focus on the geometry of latent semantic spaces by proposing a novel distributional model for semantic composition. The aim is to model semantics of syntactic bigrams as projections in lexically-driven subspaces. Distances in such subspaces (called *Support Spaces*) emphasize the role of *common* features that constraint in "parallel" the interpretation of the involved lexical meanings and better capture phrase-specific aspects. In the following evaluations, operators will be employed to compose word pairs involved in specific syntactic structures. This resulting compositions will be evaluated according two different perspectives. First, similarity among compositions will be evaluated with respect to human annotators' judgments. Then, the operators generalization capability will be measured in order to prove their applicability in semantic search complex systems. Moreover the robustness of this Support Spaces based will be confirmed in a cross-linguistic scenario, i.e. in the English and Italian Language.

While Section 2 discusses existing methods of compositional distributional semantics, Section 3 presents our model based on support spaces. Experiments in Section 4 are used to show the beneficial impact of the proposed model and the contribution to semantic search systems. Finally, Section 5 derives the conclusions.

## 2 Related work

While compositional semantics allows to govern the recursive interpretation of sentences or phrases, traditional vector space models (as in IR [20]) and, mostly, semantic space models, such as LSA ([7, 13]), represent lexical information in metric spaces where individual words are represented according to the distributional analysis of their co-occurrences over a large corpus. Such models are based on the distributional hypothesis which assumes that words occurring within similar contexts are semantically similar (Harris in [12]).

Semantic spaces have been widely used for representing the meaning of words or other lexical entities (e.g. [23]), with successful applications in lexical disambiguation ([22]) or harvesting thesauri (as in Lin [14]). In this work we will refer to the so-called **word-based spaces**, in which words are represented by probabilistic information of their co-occurrences calculated in a fixed range window over all sentences. In such models, vector components correspond to the entries  $f$  of the vocabulary  $V$  (i.e. to features that are individual words). Weights are associated with each component, using different estimators of their correlation. In some works (e.g. [15]) pure co-occurrence counts are adopted as weighting functions  $f_i$ , where  $i = 1, \dots, N$  and  $N = |V|$ ; in other works (e.g. [18]), statistical functions like the pointwise mutual information between the target word  $w$  and the captured co-occurrences in the window are used, i.e.  $pmi(w, i) = \log_2 \frac{p(w, f_i)}{p(w) \cdot p(f_i)}$ .

A vector  $\mathbf{w} = (pmi_1, \dots, pm_i_N)$  models a word  $w$  and it is thus built over all the words  $f_i$  belonging to the dictionary. When  $w$  and  $f$  never co-occur in any window their  $pmi$  is by default set to 0. Weights of vector components depend on the size of the co-occurrence window and express the global statistics in the entire corpus. Larger values of the adopted window size aim to capture *topical*

*similarity* (as in the document based models of IR), while smaller sizes (usually between the  $\pm 1-3$  surrounding words) lead to representation better suited for *paradigmatic similarities* between word vectors  $\mathbf{w}$ . Cosine similarity between vectors  $\mathbf{w}_1$  and  $\mathbf{w}_2$  is modeled as the normalized scalar product, i.e.  $\frac{\langle \mathbf{w}_1, \mathbf{w}_2 \rangle}{\|\mathbf{w}_1\| \|\mathbf{w}_2\|}$  that expresses *topical* or *paradigmatic similarity* according to the different representations (e.g. window sizes). Notice that dimensionality reduction methods, such as LSA [7, 13] are also applied in some studies, to capture second order dependencies between features  $f$ , i.e. applying semantic smoothing to possibly sparse input data. Applications of an LSA-based representation to Frame Induction or Semantic Role Labeling are presented in [19] and [6], respectively.

The main limitation of distributional models of lexical semantic is their non-compositional nature: they are based on statistics related to the occurrences of the individual words in the corpus. In such models, the semantic of topological similarity functions is thus defined only for the comparison between individual words. That is the reason why distributional methods can not compute the meanings of phrases (and sentences) as effectively as they do indeed over individual words. Distributional methods have been recently extended to better account compositionality, in the so called distributional compositional semantics (DCS) approaches. Mitchell and Lapata in [15] follow Foltz [9] and assume that the contribution of the syntactic structure can be ignored, while the meaning of a phrase is simply the *commutative sum of the meanings of its constituent words*. More formally, [15] defines the composition  $\mathbf{p}^\circ = \mathbf{u} \circ \mathbf{v}$  of vectors  $\mathbf{u}$  and  $\mathbf{v}$  through an additive class of composition functions expressed by:

$$\mathbf{p}^+ = \mathbf{u} + \mathbf{v} \tag{1}$$

This perspective clearly leads to a variety of efficient yet shallow models of compositional semantics compared in [15]. For example pointwise multiplication is defined by the multiplicative function:

$$\mathbf{p}^\cdot = \mathbf{u} \odot \mathbf{v} \tag{2}$$

where the symbol  $\odot$  represents multiplication of the corresponding components, i.e.  $p_i = u_i \cdot v_i$ . Point-wise multiplication seems to best correspond with the intended effects of syntactic interaction, as experiments in [15] demonstrate. In [8], the concept of a *structured vector space* is introduced, where each word is associated with a set of vectors corresponding to different syntactic dependencies. Every word is thus expressed by a tensor, and tensor operations are imposed.

The main differences among these studies lies in (1) the lexical vector representation selected (e.g. some authors do not even commit to any representation, but generically refer to any lexical vector, as in [11]) as well as in (2) the adopted compositional algebra, i.e. the system of operators defined over such vectors. Generally, proposed operators do not depend on the involved lexical items, but a general purpose algebra is adopted. Since compositional structures are highly lexicalized, and the same syntactic relation triggers to very different semantic relations with respect to the different involved words, a proposal that makes the compositionality operators dependent on individual lexical vectors is hereafter discussed.

### 3 A quantitative model for compositionality

In order to determine the semantic analogies and differences between two phrases, such as "... *buy a car* ..." and "... *buy time* ...", a distributional compositional model is employed as follows. The involved lexicals are *buy*, *car* and *time*, while their corresponding vector representation will be denoted by  $\mathbf{w}_{buy}$ ,  $\mathbf{w}_{car}$  and  $\mathbf{w}_{time}$ . The major result of most studies on DCS is the definition of the function  $\circ$  that associates with  $\mathbf{w}_{buy}$  and  $\mathbf{w}_{car}$  a new vector  $\mathbf{w}_{buy\_car} = \mathbf{w}_{buy} \circ \mathbf{w}_{car}$ .

We consider this approach misleading since vector components in the word space are tied to the syntactic nature of the composed words and the new vector  $\mathbf{w}_{buy\_car}$  should not have the same type of the original vectors. Notice also that the components of  $\mathbf{w}_{buy}$  and  $\mathbf{w}_{car}$  express all their contexts, i.e. interpretations, and thus senses, of *buy* and *car* in the corpus. Algebraic operations are thus open to misleading contributions, brought by not-null feature scores of  $buy_i$  vs.  $car_j$  ( $i \neq j$ ) that may correspond to senses of *buy* and *car* that are not related to the specific phrase "*buy a car*". On the contrary, in a composition, such as the verb-object pair (*buy*, *car*), the word *car* influences the interpretation of the verb *buy* and viceversa. The model here proposed is based on the assumption that this influence can be expressed via the operation of projection into a subspace, i.e. a subset of original features  $f_i$ . A projection is a mapping (a selection function) over the set of all features. A subspace generated by a projection function  $\Pi$  local to the (*buy*, *car*) phrase can be found such that only the features specific to the phrase meaning are selected and the irrelevant ones are neglected. The resulting subspace has to preserve the compositional semantics of the phrase and it is called **support subspace** of the underlying word pair.

Consider the bigram composed of the words *buy* and *car* and their vectorial representation in a co-occurrence  $N$ -dimensional Word Space. Table 1 reports the  $k = 10$  features with the highest contributions of the point wise product of the pairs (*buy*, *car*) and (*buy*, *time*). The support space thus selects the most important features for both words, e.g. *buy.V* and *car.N*. Notice that this captures the conjunctive nature of the scalar product to which contributions come from feature with non zero scores in both vectors. It is clear that the two pairs give rise to different support subspaces: the main components related with *buy car* refer mostly to the automobile commerce area unlike the ones related with *buy time* mostly referring to the time wasting or saving. Similarity judgments about a pair can be thus better computed within its support subspace.

More formally  $k$ -dimensional support subspace for a word pair  $(u, v)$  (with  $k \ll N$ ) is the subspace spanned by the subset of  $n \leq k$  indexes  $\mathbf{I}^k(\mathbf{u}, \mathbf{v}) = \{i_1, \dots, i_n\}$  for which  $\sum_{t=1}^n u_{i_t} \cdot v_{i_t}$  is maximal. Given two pairs the similarity

Buy-Car	Buy-Time
<i>cheap::Adj</i>	<i>consume::V</i>
<i>insurance::N</i>	<i>enough::Adj</i>
<i>rent::V</i>	<i>waste::V</i>
<i>lease::V</i>	<i>save::In</i>
<i>dealer::N</i>	<i>permit::N</i>
<i>motorcycle::N</i>	<i>stressful::Adj</i>
<i>hire::V</i>	<i>spare::Adj</i>
<i>auto::N</i>	<i>save::V</i>
<i>california::Adj</i>	<i>warners::N</i>
<i>tesco::N</i>	<i>expensive::Adj</i>

**Table 1.** Features corresponding to dimensions in the  $k=10$  dimensional support space of bigrams *buy car* and *buy time*

between syntactic equivalent words (e.g. nouns with nouns, verbs with verbs) is measured in the support subspace derived by applying a specific projection function. Compositional similarity between *buy car* and the latter pairs (e.g. *buy time*) is thus estimated by (1) immersing  $w_{buy}$  and  $w_{time}$  in the selected "... *buy car* ..." support subspace and (2) estimating similarity between corresponding arguments of the pairs locally in that subspace. Therefore the similarity between syntactic equivalent words (e.g. *car* with *time*) within these new subspace is measured.

Therefore given a pair  $(u, v)$ , a unique matrix  $\mathbf{M}_{uv}^k = (m_{uv}^k)_{ij}$  is defined for a given projection  $\Pi^k(u, v)$  into the  $k$ -dimensional support space of any pair  $(u, v)$  according to the following definition:

$$(m_{uv}^k)_{ij} = \begin{cases} 1 & \text{iff } i = j \in \mathbf{I}^k(\mathbf{u}, \mathbf{v}) \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The vector  $\tilde{\mathbf{u}}$  projected in the support subspace can be thus estimated through the following matrix operation:

$$\tilde{\mathbf{u}} = \Pi^k(u, v) \quad \tilde{\mathbf{u}} = \mathbf{M}_{uv}^k \mathbf{u} \quad (4)$$

A special case of the projection matrix is given when no  $k$  limitation is imposed to the dimension and all the positive addends in the scalar product are taken. Notice also that two pairs  $p_1 = (u, v)$  and  $p_2 = (u', v')$  give rise to two different projections denoted by  $\mathbf{M}_1^k$  and  $\mathbf{M}_2^k$  and defined as:

$$(Left \text{ projection}) \Pi_1^k = \Pi^k(\mathbf{u}, \mathbf{v}) \quad (Right \text{ projection}) \Pi_2^k = \Pi^k(\mathbf{u}', \mathbf{v}') \quad (5)$$

It is also possible to define a unique symmetric projection  $\Pi_{12}^k$  corresponding to the combined matrix  $\mathbf{M}_{12}^k$  as follows:

$$\mathbf{M}_{12}^k = (\mathbf{M}_1^k + \mathbf{M}_2^k) - (\mathbf{M}_1^k \mathbf{M}_2^k) \quad (6)$$

where the mutual components that satisfy Eq. 3 are employed as  $\mathbf{M}_{12}^k$ .

As  $\Pi_1$  is the projection in the support subspace for the pair  $p_1$ , it is possible to immerse the latter pair  $p_2$  by applying Eq. 4. **This results in the two vectors  $\mathbf{M}_1^k \mathbf{u}'$  and the  $\mathbf{M}_1^k \mathbf{v}'$ .** It follows that a compositional similarity judgment between two phrase over the first pair support subspace can be expressed as:

$$\Phi_{p_1}^{(\circ)}(p_1, p_2) = \Phi_1^{(\circ)}(p_1, p_2) = \frac{\langle \mathbf{M}_1^k \mathbf{u}, \mathbf{M}_1^k \mathbf{u}' \rangle}{\|\mathbf{M}_1^k \mathbf{u}\| \|\mathbf{M}_1^k \mathbf{u}'\|} \circ \frac{\langle \mathbf{M}_1^k \mathbf{v}, \mathbf{M}_1^k \mathbf{v}' \rangle}{\|\mathbf{M}_1^k \mathbf{v}\| \|\mathbf{M}_1^k \mathbf{v}'\|} \quad (7)$$

where first cosine similarity between syntactically correlated vectors in the selected support subspaces are computed and then a composition function  $\circ$ , such as the sum or the product, is applied. Compositional function over the latter support subspace evoked by the pair  $p_2$  can be correspondingly denoted by  $\Phi_2^{(\circ)}(p_1, p_2)$ . A symmetric composition function can thus be obtained as a combination of  $\Phi_1^{(\circ)}(p_1, p_2)$  and  $\Phi_2^{(\circ)}(p_1, p_2)$  as:

$$\Phi_{12}^{(\diamond)}(p_1, p_2) = \Phi_1^{(\circ)}(p_1, p_2) \diamond \Phi_2^{(\circ)}(p_1, p_2) \quad (8)$$

where the composition function  $\diamond$  (again the sum or the product) between the similarities over the left and right support subspaces is applied. Notice how the left and right composition operators ( $\circ$ ) may differ from the overall composition operator  $\diamond$ . More details are discussed in [1].

## 4 Experimental Evaluation

This experimental evaluation aims to estimate the effectiveness of the proposed class of projection based methods in capturing similarity judgments over phrases and syntactic structures. In particular, a first evaluation is carried out to measure the correlation of the operator outcomes with judgments provided by human annotators. The generalization capability of the operators is measured in the second evaluation in order to prove their applicability in semantic search complex systems. Moreover the latter experiments are carried out in a cross-language setting, i.e. for english and italian datasets.

Type	First Pair	Second Pair	Rate
VO	<i>support offer</i>	<i>provide help</i>	7
	<i>use knowledge</i>	<i>exercise influence</i>	5
	<i>achieve end</i>	<i>close eye</i>	1
AdjN	<i>old person</i>	<i>right hand</i>	1
	<i>vast amount</i>	<i>large quantity</i>	7
	<i>economic problem</i>	<i>practical difficulty</i>	3
NN	<i>tax charge</i>	<i>interest rate</i>	7
	<i>tax credit</i>	<i>wage increase</i>	5
	<i>bedroom window</i>	<i>education officer</i>	1

**Table 2.** Example of Mitchell and Lapata dataset for the three syntactic relations verb-object (VO), adjective-noun (AdjN) and noun-noun (NN)

applied. Part-of-speech tagged words have been collected from the corpus to reduce data sparseness. Then all target words *tws* occurring more than 200 times are selected, i.e. more that 50,000 candidate features. Each column  $i$  of  $M$  represents a word  $w$  in the corpus. Rows model the target words  $tw$ , i.e. contain the  $p_{mi}$  values for the individual features  $f_i$ , as captured in a window of size  $\pm 3$  around  $tw$ . The most frequent 20,000 left and right features  $f_i$  are selected, so that  $M$  expresses 40,000 contexts. SVD is here applied to limit dimensionality to  $N = 100$ .

### 4.1 Experiment I

The first evaluation is carried out over the dataset proposed by [16], which is part of the *GEMS 2011 Shared Evaluation*. It consists of a list of 5,833 adjective-noun (AdjN), verb-object (VO) or noun-noun (NN) pairs, rated with scores ranging from

<sup>1</sup> The corpus is developed by the WaCky community and it is available in the Wacky project web page at <http://medialab.di.unipi.it/Project/QA/wikiCoNLL.bz2>

Two different word space are derived for the different languages. For English, the word space is derived from the ukWak [4], a web-based corpus consisting of about 2 billion tokens. For Italian, the Italian Wikipedia corpus<sup>1</sup> has been employed. It consists of about 200 million tokens and more than 10 million sentences. The space construction proceeds from an adjacency matrix  $M$  on which Singular Values decomposition ([7]) is then

1 to 7. In Table 2, examples of pairs and scores are shown. The correlation of the similarity judgements outputted by a DCS model against the human judgements is computed using Spearman’s  $\rho$ , a non-parametric measure of statistical dependence between two variables proposed by [15].

Model		AdjN	NN	VO
Mitchell&Lapata Word Space SVD	Additive	.69	.70	<b>.64</b>
	Multiplicative	.38	.43	.42
Support Subspace[1]	$\Phi^{(+)}, \Pi_{12}^k (k=30)$	<b>.70</b>	<b>.71</b>	.63
	$\Phi_{12}^{(+)}, \Phi_i^{(+)}, \Pi_i^k (k=40)$	.68	.68	<b>.64</b>
Agreement among Human Subjects	Max	.88	.92	.88
	Avg	.72	.72	.71

**Table 3.** Spearman’s  $\rho$  correlation coefficients across Mitchell and Lapata models and the projection-based models proposed in Section 3. Word space refers to the source spaces used as input to the LSA decomposition model.

Table 3 reports M&L performances in the first row, while in the last row the max and the average interannotator agreement scores for the three categories derived through a leave one-out resampling method are shown. Row 2 shows Spearman’s correlation for support subspace models discussed in [1] that better perform the distributional compositional task. Notice that different configurations according to the models described in Section 3 are used. For example, the system denoted as  $\Phi_{12}^{(+)}, \Phi_i^{(+)}, \Pi_i^k (k=40)$ , corresponds to a multiplicative symmetric composition function  $\Phi_{12}^{(+)}$  (as for Eq. 8) based on left and right additive compositions  $\Phi_i^{(+)}$  ( $i = 1, 2$  as in Eq. 7), derived through a projection  $\Pi_i^k$  in the support space limited to the first  $k = 40$  components for each pair (as for Eq. 5). The specific operator denoted by  $\Phi^{(+)}, \Pi_{12}^k (k=30)$  achieves the best performance over two out of three syntactic patterns (i.e. AdjN and NN) and is close to the best figures for VO. Experimental evaluation shows that the best performances are achieved by the projection based operators proposed. Notice that the distributional composition between verbs and objects is a very tricky task and results are in line with the additive model. Globally the results of our models are close to the average agreement among human subjects, this latter representing a sort of upper bound for the underlying task. It seems that latent topics (as extracted through SVD from sentence and word spaces) as well as the projections operators defined by support subspaces, provide a suitable comprehensive paradigm for compositionality. They seem to capture compositional similarity judgements that are significantly close to human ones. Notice that different settings of the projection operations can influence the performances. A more exhaustive study of the possible settings is presented in [1].

## 4.2 Experiment II

In this second evaluation, the generalization capability of the employed operators will be investigated. A verb (e.g. *perform*) can be more or less semantically

close to another verb (e.g. other verbs like *solve*, or *produce*) depending on the context in which it appears. The verb-object (VO) composition specifies the verb’s meaning by expressing one of its selectional preferences, i.e. its object. In this scenario, we expect that a pair such as *perform task* will be more similar to *solve issue*, as they both reflect an abstract cognitive action, with respect to a pair like *produce car*, i.e. a concrete production. This kind of generalization capability is crucial to effectively use this class of operators in a QA scenario by enabling to rank results according to the complex representations of the question. Moreover, both English and Italian languages can be considered to demonstrate the impact in a cross language setting. Figure 4 shows a manually developed dataset. It consists of 24 VO word pairs in English and Italian, divided into 3 different semantic classes: **Cognitive**, **Ingest Liquid** and **Fabricate**.

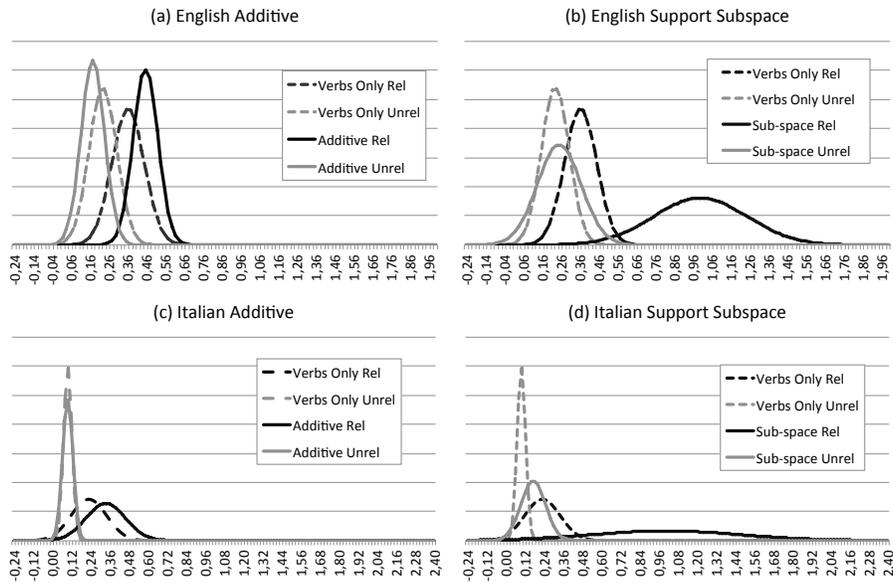
Semantic Class	English	Italian
Cognitive	<i>perform task</i> <i>solve issue</i> <i>handle problem</i> <i>use method</i> <i>suggest idea</i> <i>determine solution</i> <i>spread knowledge</i> <i>start argument</i>	<i>svolgere compito</i> <i>risolvere questione</i> <i>gestire problema</i> <i>applicare metodo</i> <i>suggerire idea</i> <i>trovare soluzione</i> <i>divulgare conoscenza</i> <i>iniziare ragionamento</i>
Ingest Liquid	<i>drink water</i> <i>ingest syrup</i> <i>pour beer</i> <i>swallow saliva</i> <i>assume alcohol</i> <i>taste wine</i> <i>sip liquor</i> <i>take coffee</i>	<i>bere acqua</i> <i>ingerire sciroppo</i> <i>versare birra</i> <i>inghiottire saliva</i> <i>assumere alcool</i> <i>assaggiare vino</i> <i>assaporare liquore</i> <i>prendere caff</i>
Fabricate	<i>produce car</i> <i>complete construction</i> <i>fabricate toy</i> <i>build tower</i> <i>assemble device</i> <i>construct building</i> <i>manufacture product</i> <i>create artwork</i>	<i>produrre auto</i> <i>completare costruzione</i> <i>fabbricare giocattolo</i> <i>edificare torre</i> <i>assemblare dispositivo</i> <i>costruire edificio</i> <i>realizzare prodotto</i> <i>creare opera</i>

**Table 4.** Cross-linguistic dataset

This evaluation aims to measure how the proposed compositional operators group together semantically related word pairs, i.e. those belonging to the same class, and separate the unrelated pairs. Figure 1 shows the application of two models, the Additive (eq. 1) and Support Subspace (Eq. 8) ones that achieve the best results in the previous experiment. The two languages are reported in different rows. Similarity distribution between the geometric representation of verb pair, with no composition, has been investigated as a baseline. For each language, the similarity distribution among the possible 552 verb pairs is estimated and two distributions of the **infra** and **intra-class** pairs are independently plotted. In order to summarize them, a Normal Distribution  $N(\mu, \sigma^2)$  of mean  $\mu$  and variance  $\sigma^2$  are employed. Each point represents the percentage  $p(x)$  of pairs in a group that have a given similarity value equal to  $x$ . In a given class, the VO-VO pairs of a DCS operator are expected to increase this probability with respect to the baseline pairs V-V of the same set. Viceversa, for pairs belonging to different classes, i.e. **intra-class** pairs. The distributions for the baseline

control set (i.e. **Verbs Only**, V-V) are always depicted by dotted lines, while DCS operators are expressed in continuous line.

Notice that the overlap between the curves of the **infra** and **intra-class** pairs corresponds to the amount of **ambiguity** in deciding if a pair is in the same class. It is the *error probability*, i.e. the percentage of cases of one group that by chance appears to have more probability in the other group. Although the actions described by different classes are very different, e.g. **Ingest Liquid** vs. **Fabricate**, most verbs are ambiguous: contextual information is expected to enable the correct decision. For example, although the class **Ingest Liquid** is clearly separated with respect to the others, a verb like *assume* could well be classified in the **Cognitive** class, as in *assume a position*.



**Fig. 1.** Cross-linguistic Gaussian distribution of infra (red) and inter (green) clusters of the proposed operators (continuous line) with respect to verbs only operator (dashed line)

The outcome of the experiment is that DCS operators are always able to increase the gap in the average similarity of the **infra** vs. **intra-class** pairs. It seems that the geometrical representation of the verb is consistently changed as most similarity distributions suggest. The compositional operators seem able to decrease the overlap between different distributions, i.e. reduce the ambiguity.

Figure 1 (a) and (c) report the distribution of the ML additive operator, that achieves an impressive ambiguity reduction, i.e. the overlap between curves is drastically reduced. This phenomenon is further increased when the Support

Subspace operator is employed as shown in Figure 1 (b) and (d): notice how the mean value of the distribution of semantically related word is significantly increased for both languages.

The probability of error reduction can be computed against the control groups. It is the decrease of the error probability of a DCS relative to the same estimate for the control (i.e. V-V) group. It is a natural estimator of the generalization capability of the involved operators. In Table 5 the intersection area for all the models and the decrement of the relative probability of error are shown. For English, the ambiguity reduction of the Support Subspace operator is of 91% with respect to the control set. This is comparable with the additive operator results, i.e. 92.3%. It confirms the findings of the previous experiment where the difference between these operators is negligible. For Italian, the generalization capability of support subspace operator is more stable, as its error reduction is of 62.9% with respect to the additive model, i.e. 54.2%.

Model	English		Italian	
	Probability of Error	Ambiguity Decrease	Probability of Error	Ambiguity Decrease
VerbOnly	.401	-	.222	-
Additive	.030	92.3%	.101	54.2%
SupportSubspace	.036	91.0%	.082	62.9%

**Table 5.** Ambiguity reduction analysis

## 5 Conclusions

In this paper, a distributional compositional semantic model based on space projection guided by syntagmatically related lexical pairs is defined. Syntactic bi-grams are here projected in the so called *Support Subspace* and compositional similarity scores are correspondingly derived. This represents a novel perspective on compositional models over vector representations with respect to shallow vector operators (e.g. additive or multiplicative operations) as proposed in literature, e.g. in [16]. The presented approach focuses on selecting the most important components for a specific word pair involved in a syntactic relation in order to have a more accurate estimator of their similarity.

The proposed method have been evaluated over the well known dataset in [16] achieving results close to the average human interannotator agreement scores. A first applicability study of such compositional models in typical IR systems was carried out. The operators' generalization capability was measured proving that compositional operators can effectively separate phrase structure in different semantic clusters. The robustness of such operators has been also confirmed in a cross-linguistic scenario, i.e. in the English and Italian Language. Future work on other compositional prediction tasks (e.g. selectional preference modeling) and over different datasets will be carried out to better assess and generalize the presented results.

## References

1. Annesi, P., Storch, V., Basili, R.: Space projections as distributional models for semantic composition (2012), submitted for publication

2. B. Coecke, M.S., Clark, S.: Mathematical foundations for a compositional distributed model of meaning. *Lambek Festschrift, Linguistic Analysis*, vol. 36 36 (2010), <http://arxiv.org/submit/10256/preview>
3. Baeza-Yates, R., Ciaramita, M., Mika, P., Zaragoza, H.: Towards semantic search. *Natural Language and Information Systems* pp. 4–11 (2008)
4. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources And Evaluation* 43(3), 209–226 (2009)
5. Baroni, M., Zamparelli, R.: Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In: *Proceedings of EMNLP 2010*. pp. 1183–1193. EMNLP '10, Stroudsburg, PA, USA (2010)
6. Croce, D., Giannone, C., Annesi, P., Basili, R.: Towards open-domain semantic role labeling. In: *ACL*. pp. 237–246 (2010)
7. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. *JASIS* 41(6), 391–407 (1990)
8. Erk, K., Pad, S.: A structured vector space model for word meaning in context (2008)
9. Foltz, P.W., Kintsch, W., Landauer, T.K., L, T.K.: The measurement of textual coherence with latent semantic analysis (1998)
10. Frege, G.: Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100, 25–50, translated, as ‘On Sense and Reference’, by Max Black
11. Grefenstette, E., Sadrzadeh, M.: Experimental support for a categorical compositional distributional model of meaning. *CoRR* abs/1106.4058 (2011)
12. Harris, Z.S.: *Mathematical Structures of Language*. Wiley, New York, NY, USA (1968)
13. Landauer, T.K., Dutnais, S.T.: A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* pp. 211–240 (1997)
14. Lin, D.: Automatic retrieval and clustering of similar word. In: *Proceedings of COLING-ACL*. Montreal, Canada (1998)
15. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: *Proceedings of ACL-08: HLT*. pp. 236–244 (2008)
16. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. *Cognitive Science* 34(8), 1388–1429 (2010)
17. Montague, R.: *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press (1974)
18. Pantel, P., Lin, D.: Document clustering with committees. In: *SIGIR-02*. pp. 199–206 (2002)
19. Pennacchiotti, M., Cao, D.D., Basili, R., Croce, D., Roth, M.: Automatic induction of framenet lexical units. In: *EMNLP*. pp. 457–465 (2008)
20. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. *Communications of the ACM* 18, 613–620 (1975)
21. Schütze, H.: Word space. In: Hanson, S.J., Cowan, J.D., Giles, C.L. (eds.) *NIPS* 5, pp. 895–902. Morgan Kaufmann Publishers, San Mateo CA (1993)
22. Schütze, H.: Automatic Word Sense Discrimination. *Computational Linguistics* 24, 97–124 (1998)
23. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37, 141 (2010), doi:10.1613/jair.2934