Space projections as distributional models for semantic composition

Paolo Annesi, Valerio Storch, Roberto Basili

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

Abstract. Empirical distributional methods account for the meaning of syntactic structures by combining words according to algebraic operators (e.g. tensor product) acting over the corresponding 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 line with Frege's context principle, the meaning of a phrase is modeled in terms of the subset of properties shared by the co-occurring words. 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 phrase similarity task, used as a benchmark for this class of methods.

1 Introduction

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 [1] and [2]) rely on Frege's principle for which the meaning of a sentence is a function of the meanings of its parts [3]. The resulting theory allows an algebra on the discrete propositional symbols to represent the meaning of arbitrarily complex expressions.
More recently, distributional models of lexical semantics have been proposed (e.g. Firth [4] or Schuetze [5]) based on Wittgenstein's later philosophy, whereby the lexical meanings is determined by their context of use [6]. This seems to be a completely orthogonal *view* on meaning representation with respect to logical models.

Computational models of semantics based on symbolic logic representations can account naturally for the meaning of sentences, through the notion of compositionality for which the meaning of complex expressions can be determined by using the meanings of their constituent and the rules to combine them. 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. For instance, the sentence *Meanwhile, the bank closed* could either refer to the closing time of an office, or to the "*cease to operate*" sense of a bankrupt. Logic-based approaches present strict limitation towards these tasks and bring often inadequate tool to model and overcome the uncertainty of phenomena like select the proper interpretation of a specific verb-object pair.

Distributional models early introduced by Schütze [7] and recently surveyed in [8] rely on the Word Space model, inspired by Information Retrieval. They manage semantic uncertainty through mathematical functions grounded in probability or vector spaces. Points in the space represent semantic concepts, such as words, and can be learned from corpora, in such a way that similar, or related, concepts are near to one another in the space. The distance between two points (via angular or Euclidean metrics) represents semantic dissimilarity between concepts. Methods for constructing representations for phrases or sentences through vector composition has recently received a wide attention in literature (e.g. [9]). However, vector-based models typically represent isolated words and ignore grammatical structure [8]. They 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. [10], [11]) A compositional model based on distributional analysis should provide semantic information consistent with the meaning assignment 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) needs thus a method to define: (1) a way to represent lexical vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, 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 $\boldsymbol{u}$, $\boldsymbol{v}$.

Existing models are still controversial and provide general algebraic operators (such as tensor products) over lexical vectors. By focusing on the geometry of latent semantic spaces a novel distributional model for semantic composition is proposed . 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 the involved lexical meanings and better capture phrase-specific aspects. 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.

2 Related work

While compositional semantics allows to govern the recursive interpretation of sentences or phrases, vector space models (as in IR [12]) and, mostly, semantic space models, such as LSA ([13, 14]), represent lexical information in metric spaces where individual words are represented according to the distributional analysis of their co-occurrences over a large corpus.

Distributional models are based on the theory that words occurring within similar contexts are semantically similar (Harris in [15]). Words are represented as vectors and their meaning is distributed across many dimensions. Word meaning is obtained empirically by examining the contexts in which a word appears. The meaning of a word w corresponds strictly to the distributional context in which w occurs, i.e. depends on the contexts distribution it shares with other words. Vector components reflect the corresponding contexts so that two words close in the space are systematically found in similar contexts. This suggests that they are related by some type of generic semantic relation, either paradigmatic (e.g. synonymy, hyperonymy, antonymy) or syntagmatic (e.g. meronymy, conceptual and phrasal association), as observed in Sahlgren [16].

Semantic spaces have been widely used for representing the meaning of words or other lexical entities (e.g. [8]), with successful applications in lexical disambiguation ([5]) or harvesting thesauri (as in Lin [17]). In this work we will refer to the so-called **word-based spaces**, in which target words are represented by gathering probabilistic information of their co-occurences calculated in a fixed range window over all sentences. In such that models vectors components correspond to the entries f of the vocabulary V (i.e. to features that are individual words). In some works (e.g. [9]) pure co-occurrence counts are adopted as weights for individual features f_i, where $i = 1, ..., N$ and $N = |V|$; in other works (e.g. [18]), weights are the pointwise mutual information scores between the target word w and the captured co-occurences in the window,

$$pmi(w, i) = log_2 \frac{p(w, f_i)}{p(w) \cdot p(f_i)} \qquad i = 1, ..., N$$

A vector $\boldsymbol{w} = (pmi_1, ..., pmi_N)$ for a word w 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 $\boldsymbol{w}$. Cosine similarity between vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ is modeled as the normalized scalar product, i.e.

$$\frac{\langle \boldsymbol{w}_1, \boldsymbol{w}_2 \rangle}{\|\boldsymbol{w}_1\| \, \|\boldsymbol{w}_2\|}$$

that expresses *topical* or *paradigmatic similarity* according to the different representations (e.g. window sizes). Notice that dimensionality reduction methods,

such as LSA [13, 14] 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 ([20]), respectively.

The main drawback of the above models is their non-compositional nature: they ignore the grammatical structure underlying phrases, such as "*... buy a car ...*" that are thus not clearly connected to the base vectors $\boldsymbol{w}_{buy}$ and $\boldsymbol{w}_{car}$. Distributional methods hence can not compute the meanings of phrases (and sentences) as efficiently as they do indeed over words.

2.1 Distributional Compositional Semantic Models

Distributional methods have been thus recently extended to better account compositionality, in the so called distributional compositional semantics (DCS) approaches. Mitchell and Lapata in [9] follow Foltz [21] and assume that the contribution of 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, [9] defines the composition $\boldsymbol{p}^{\circ} = \boldsymbol{u} \circ \boldsymbol{v}$ of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ through an additive class of composition functions expressed by:

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

This perspective clearly leads to a variety of efficient yet shallow models of compositional semantics compared in [9]. For example pointwise multiplication is defined by the multiplicative function:

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

where the symbol $\odot$ represents multiplication of the corresponding components, i.e. $p_i = u_i \cdot v_i$. Since the cosine similarity function is insensitive to the vectors magnitude, in [9] a more complex asymmetric type of function called *dilation* is introduced. It consists in multiplying vectors $\boldsymbol{v}$ by the quadratic factor $\boldsymbol{u} \cdot \boldsymbol{u}$ and $\boldsymbol{v}$ by a stretching factor λ as follows: $\boldsymbol{p}^{d} = (\boldsymbol{u} \cdot \boldsymbol{u})\boldsymbol{v} + (\lambda - 1)(\boldsymbol{u} \cdot \boldsymbol{v})\boldsymbol{u}$ Notice that either $\boldsymbol{u}$ can be used to dilate $\boldsymbol{v}$, or $\boldsymbol{v}$ can be used to dilate $\boldsymbol{u}$. The best dilation factor λ for the dilation models is studied and tuned in [9]. Dilation and point-wise multiplication seem to best correspond with the intended effects of syntactic interaction, as experiments in [9] demonstrate.

In [22], the concept of a *structured vector space* is introduced, where each word is associated to 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. In most work, 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 gives rise to very different operators 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

Let's start to discuss the above compositional model over an example, where we want to model the semantic analogies and differences between "*... buy a car ...*" and "*... buy time ...*". The involved lexicals are *buy*, *car* and *time*, while their corresponding vector representation will be denoted by $\boldsymbol{w}_{buy}$ $\boldsymbol{w}_{car}$ and $\boldsymbol{w}_{time}$. The major result of most studies on DCS is the definition of the function $\circ$ that associates to $\boldsymbol{w}_{buy}$ and $\boldsymbol{w}_{car}$ a new vector $\boldsymbol{w}_{buy\_car} = \boldsymbol{w}_{buy} \circ \boldsymbol{w}_{car}$.

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 $\boldsymbol{w}_{buy\_car}$ should not have the same type of the original vectors. Mathematical operations between the two input vectors (e.g. point wise multiplication $\odot$ as in Eq. 2) produce a vector for a structure (i.e. a new type) that possess the same topological nature of the original vectors. As these latter are dedicated to express arguments, i.e. a verb and its object in the initial space, the syntactic information (e.g. the relation and the involved POS) carried independently by them is neglected in the result. For example, the structure "*... buy a car ...*" combines syntactic roles that are different and the antisymmetric relationship between the head verb and the modifier noun is relevant. The vectorial composition between $\boldsymbol{w}_{buy}$ and $\boldsymbol{w}_{car}$, as proposed in Eq. 2 [23], even if mathematically correct, results in a vector $\boldsymbol{w}_{buy\_car}$ that does not exploit this syntactic constraint and may fail to express the underlying specific semantics.

Notice also that the components of $\boldsymbol{w}_{buy}$ and $\boldsymbol{w}_{car}$ express all their contexts, i.e. interpretations, and thus senses, of *buy* and *car* in the corpus. Some mathematical operations, e.g. the tensor product between these vectors, 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 local to the (buy, car) phrase can be found such that only the features specific to its meaning are selected. It seems a necessary condition that any correct interpretation of the phrase has to be retrieved and represented on the subspace of the properties shared by the proper sense of individual co-occurring words. In order to separate these word senses and neglect irrelevant ones, a projection function Π must identify these common semantic features. 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 by the words *buy* and *car* and their vectorial representation in a co-occurrence $N-$dimensional Word Space. Notice that different vectors are usually derived for different POS tags, so that the verbal and nominal use of *buy* are expressed by two different vectors, i.e. *buy*.V and *buy*.N. Every component of the vectors in a word space expresses the co-occurrence strength (in terms of frequency or *pmi*) of *buy*.V with respect to one feature, i.e. a co-occurring POS tagged word such as *cost*.N, *pay*.V or *cheaply*.Adv. The support space 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. Moreover, the feature score is a weight, i.e. a function of the relevance of a feature for the represented word.

As an example, let us consider the phrase "*... buy time ...*". Although the verb *buy* is the same of "*... buy a car ...*", its meaning (i.e. to do something in order to achieve more time) is clearly different. Since vector $\boldsymbol{w}_{buy}$ expresses at least both possible meanings of the verb *buy*, different subspaces must be evoked in a distributional model for *buy car* vs. *buy time*.

Ranking features from the most important to the least important for a given phrase (i.e. pair $\boldsymbol{u}$ and $\boldsymbol{v}$) can be done by sorting in decreasing order the components $p_i = u_i \cdot v_i$, i.e. the addends in the scalar product. This leads to the following useful:

Buy-Car	Buy-Time
cheap::Adj	*consume::V*
insurance::N	*enough::Adj*
rent::V	*waste::V*
lease::V	*save::In*
dealer::N	*permit::N*
motorcycle::N	*stressful::Adj*
hire::V	*spare::Adj*
auto::N	*save::V*
california::Adj	*warner::N*
tesco::N	*expensive::Adj*

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

Definition (k-dimensional support space). A $k-$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 $\mathbf{I}^k(\boldsymbol{u}, \boldsymbol{v}) = \{i_1, ..., i_n\}$ for which $\sum_{t=1}^{n} u_{i_t} \cdot v_{i_t}$ is maximal. We will hereafter denote the set of indexes characterizing the support subspace of order k as $\mathbf{I}^k(\boldsymbol{u}, \boldsymbol{v})$.

Table 1 reports the $k = 10$ features with the highest contributions of the point wise product of the pairs *(buy,car)* and *(buy,time)*. 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 computed within its support subspace. Given two pairs the similarity 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. In the above example, the meaning representation of *buy* and *car* is obtained by projecting both vectors in their own subspace in order to capture the (possibly multiple) senses supported by the pair. Then, compositional similarity between *buy car* and the latter pairs (e.g. *buy time*) is estimated by (1) immersing w_{buy} and w_{time} in the selected "*... buy car ...*" support subspace and (2) estimating similarity between cor-

responding arguments of the pairs locally in that subspace. As exemplified in Table 1, two pairs give rise to two different support spaces, so that there are two ways of projecting the two pairs. In order to provide precise definitions for these notions, formal definitions will be hereafter provided through linear algebra operators.

Space projections and compositionality. Support spaces (of dimension k) are isomorphic to projections in the original space. A projection $\Pi^k(u, v)$ can be used and provides a computationally simple model for expressing the intrinsic meaning of any underlying phrase (u, v). Given a pair (u, v), a unique matrix $\mathbf{M}^k_{uv} = (m^k_{uv})_{ij}$ 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:

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

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

$$\tilde{\boldsymbol{u}} = \Pi^k(u, v) \qquad \tilde{\boldsymbol{u}} = \mathbf{M}^k_{uv}\boldsymbol{u} \tag{4}$$

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. This maximal support subspace, denoted by removing the superscript k, i.e. as $\mathbf{M}_{uv} = (m_{uv})_{ij}$, is defined as follows:

$$(m_{uv})_{ij} = \begin{cases} 0 & iff \quad i \neq j \quad or \quad u_i \cdot v_i \leq 0, \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

From Eq. 5 it follows that the support subspace components are those with positive product.

Definition. (*Left and Right Projections*). Two phrases (u, v) and (u', v') give rise to two different projections, defined as follows

$$(\textit{Left} \text{ projection}) \ \Pi^k_1 = \Pi^k(\boldsymbol{u}, \boldsymbol{v}) \qquad (\textit{Right} \text{ projection}) \ \Pi^k_2 = \Pi^k(\boldsymbol{u}^{'}, \boldsymbol{v}^{'}) \tag{6}$$

We will denote the two projection matrices as $\mathbf{M}^k_1$ and $\mathbf{M}^k_2$, correspondingly. In order to achieve a unique symmetric projection Π^k_{12}, it is possible to define the corresponding combined matrix $\mathbf{M}^k_{12}$ as follows:

$$\mathbf{M}^k_{12} = (\mathbf{M}^k_1 + \mathbf{M}^k_2) - (\mathbf{M}^k_1\mathbf{M}^k_2) \tag{7}$$

where the mutual components that satisfy Eq. 3 (or Eq 5) are employed as $\mathbf{M}^k_{12}$ (or $\mathbf{M}_{12}$ respectively).

Compositional Similarity Judgments. The projection function that locates the support subspace of a word pair (v, o), whose syntactic type is *verb-object*, i.e. `VO`, will be hereafter denoted by $\Pi_{vo}(\boldsymbol{v}, \boldsymbol{o})$. Given two word pairs

$p_1 = (v, o)$ and $p_2 = (v', o')$, we define here a compositional similarity function $\Phi(p_1, p_2)$ as a model of the similarity between the underlying phrases. As the support subspace for the pair p_1 is defined by the projection Π_1, it is possible to immerse the latter pair p_2 by applying Eq. 4. **This results in the two vectors $\mathbf{M}_1 \boldsymbol{v}'$ and the $\mathbf{M}_1 \boldsymbol{o}'$.** It follows that a compositional similarity judgment between two verbal phrase over the left support subspace can be expressed as:

$$\Phi^{(\circ)}_{p_1}(p_1, p_2) = \Phi^{(\circ)}_1(p_1, p_2) = \frac{\langle \mathbf{M}^k_1 \boldsymbol{v}, \mathbf{M}^k_1 \boldsymbol{v}' \rangle}{\left\| \mathbf{M}^k_1 \boldsymbol{v} \right\| \left\| \mathbf{M}^k_1 \mathbf{v}' \right\|} \circ \frac{\langle \mathbf{M}^k_1 \boldsymbol{o}, \mathbf{M}^k_1 \boldsymbol{o}' \rangle}{\left\| \mathbf{M}^k_1 \boldsymbol{o} \right\| \left\| \mathbf{M}^k_1 \boldsymbol{o}' \right\|} \tag{8}$$

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 . Notice how the compositional function over the right support subspace evoked by the pair p_2 can be correspondingly denoted by $\Phi^{(\circ)}_2(p_1, p_2)$. A symmetric composition function can thus be obtained as a combination of $\Phi^{(\circ)}_1(p_1, p_2)$ and $\Phi^{(\circ)}_2(p_1, p_2)$ as:

$$\Phi^{(\diamond)}_{12}(p_1, p_2) = \Phi^{(\circ)}_1(p_1, p_2) \diamond \Phi^{(\circ)}_2(p_1, p_2) \tag{9}$$

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 $\diamond$, as we will see in experiments. The above definitions in fact characterize several projection functions Π^k, local composition function $\Phi^{(\circ)}_1$ as well as global composition function $\Phi^{(\diamond)}_{12}$. It is thus possible to define variants of the models presented above according to four main parameters:

Support Selection. Two different projection functions Π have been defined in Eq. 3 and Eq. 5, respectively. The MAXIMAL SUPPORT Π denotes the support space defined in Eq. 5. The k-DIMENSIONAL SUPPORT defined in Eq. 3 is always denoted by the superscript k in Π^k instead.

Symmetry of the similarity judgment. A SYMMETRIC judgment (denoted by simple Φ_{12}) involves Eq. 9 in which compositionality depends on both left and right support subspaces. In an ASYMMETRIC projection the support subspace belonging to a single (left Φ_1, or right Φ_2) pair is chosen. In all the experiments we applied Eq. 8, by only considering the left support subspace, i.e. Φ_1.

Symmetry of the support subspace. A support subspace can be build as:

- an INDEPENDENT SPACE, where different, i.e. left and right, support subspaces are built, through different projection functions $\mathbf{M}_1$ and $\mathbf{M}_2$ independently
- a UNIFIED SPACE, where a common subspace is built according to Eq. 7, and denoted by the projection matrix $\mathbf{M}_{12}$

Composition function. The composition function Φ° in Eq. 8 and 9 can be the product or the sum as well. We will denote Φ^{+}_i or $\Phi^{\cdot}_i$ as well as Φ^{+} and $\Phi^{\cdot}$ to emphasize the use of sum or product in Eq. 8 and 9. The only case in which no

combination is needed is when the unified support space (as in Eq. 7) is used, and thus no left or right Π_i is applied, but just Π_{12}.

4 Experimental Evaluation

The aim of this evaluation is to estimate if the proposed class of projection based methods for distributional compositional semantics is effective in capturing similarity judgments over phrases and syntactic structures. We tested our method over binary phrase structures represented by verb-object, noun-noun ad adjective-noun. Evaluation is carried out over the dataset proposed by [23], 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 1 to 7. In Table 2, examples of pairs and scores are shown: notice how the similarity between the (`VO`) *offer support* and *provide help* is higher than the one between *achieve end* and *close eye*. The correlation of the similarity judgements output by a DCS model against the human judgements is computed using Spearman's ρ, a non-parametric measure of statistical dependence between two variables proposed by [9].

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

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

We employed two different word spaces derived from a corpus, i.e. ukWak [24], including about 2 billion tokens. Each space construction proceeds from an adjacency matrix M on which Singular Values decomposition ([13]) is then applied. Part-of-speech tagged words have been collected from the corpus to reduce data sparseness. Then all target words *tw*s occurring more than 200 times are selected, i.e. more that 50,000 candidate features. A first space, called *sentence-based space*, is derived by applying SVD to a M=term×sentence adjacency matrix. Each column of M represents thus a sentence of the corpus, with about 1,500,000 sentences and *tf-idf* scores for words w in each row. The dimensions of the resulting SVD matrix in the sentence-based space is $N = 250$.

The second space employed is a *word space* built from the ukWak co-occurrences where left contexts are treated differently from the right ones for each target word tw. Each column in M represents here a word w in the corpus and in rows we found the *pmi* values for the individual features f_i, as captured in a window of size ± 3 around w. 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 $N = 100$.

Comparative analysis with results previously published in [23] has been carried out. We also recomputed the performance measures of operators in [23] (e.g. M&L multiplicative or additive models of Eq. 1 and 2) over all the word spaces specifically employed in the rest of our experiments.

Table 3 reports M&L performances in first three rows. In the last row of the Table the max and the average interannotator agreement scores for the three categories derived through a leave one-out resampling method, are shown. For each category with a set of subjects responses of size m, a set of $m-1$ (i.e., the response data of all but one subject) and a set of size one (i.e., the response data of the single remaining subject) are derived. The average rating of the set of $m-1$ subjects is first calculated and then Spearman's ρ correlation coefficient with respect to the singleton set is computed. Repeating this process m times results in an average and maximum score among the results (as reported in row 6). The distributional compositional models discussed in this paper are shown in rows 4 and 5, where different configurations are used according to the models described in Section 3. For example, the system denoted in Table 3 as $\Phi_{12}^{(+)}$, $\Phi_{i}^{(+)}$, Π_{i}^{k} (k=40), corresponds to an additive symmetric composition function $\Phi_{12}^{(+)}$ (as for Eq. 9) based on left and right additive compositions $\Phi_{i}^{(+)}$ ($i = 1, 2$ as in Eq. 8), derived through a projection Π_{i}^{k} in the support space limited to the first $k = 40$ components for each pair (as for Eq. 6).

First, Mitchell and Lapata operators applied onto our sentence and word space models over perform results previously presented in [23] (i.e. row 2 and 3 vs. row 1). This is mainly due to the benefits of the SVD modeling adopted here. The use of *pmi* scores in word spaces or *tf-idf* values in sentence spaces, then subject to the SVD factorization, is beneficial for the multiplicative and additive models proposed in the past.

The best performances are achieved by the projection based operators proposed in this paper. The word space version (denoted by $\Phi^{(+)}$, Π_{12}^{k} (k=30)) gets the best performance over two out of three syntactic patterns (i.e. `AdjN` and `NN`) and is close to the best figures for `VO`. Notice how parameters of the projection operations influence the performance, so that different settings provide quite different results. This is in agreement with the expected property for which different syntactic compositions require different vector operations.

If compared to the sentence space, a word space, based on a small window size, seems better capture the lexical meaning useful for modeling the syntactic composition of a pair. The subset of features, as derived through SVD, in a resulting support space is very effective as it is in good agreement with human judgements (ρ=0.71) A sentence space leads in general to a more topically-oriented lexical representations and this seems slightly less effective. In synthesis it seems that specific support subspaces are needed: a unified additive model based on a Word Space is better for adjective-noun and compound nouns while the additive symmetric model based on a sentence space is much better for verb-object pairs.

A general property is that 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 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.

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, tensorial algebraic operations) as proposed elsewhere, e.g. in [23]. The approach presented here focuses on first selecting the most important components for a specific word pair in a relation and then modeling their similarity. This captures their meanings locally relevant to the specific context evoked by the pair. The proposed *projection-based* method of

Model		AdjN	NN	VO
Mitchell&Lapata, [23]	Additive	.36	.39	.30
	Multiplicative	.46	.49	.37
	Dilation	.44	.41	.38
Mitchell&Lapata Topical SVD	Additive	.53	.67	.63
	Multiplicative	.29	.35	.40
	Dilation	.44	.49	.50
Mitchell&Lapata Word Space SVD	Additive	.69	.70	.64
	Multiplicative	.38	.43	.42
	Dilation	.60	.57	.61
Sentence Space	$\Phi_1^{(+)}$, Π_1^k (k=20)	.58	.62	.64
	$\Phi_{12}^{(+)}$, $\Phi_i^{(+)}$, Π_i^k (k=40)	.55	**.71**	.65
	$\Phi_{12}^{(+)}$, $\Phi_i^{(+)}$, Π_i^k (k=10)	.49	.65	**.66**
Word Space	$\Phi^{(+)}$, Π_{12}^k (k=30)	**.70**	**.71**	.63
	$\Phi_{12}^{(\cdot)}$, $\Phi_i^{(+)}$, Π_i^k (k=40)	.68	.68	**.64**
	$\Phi_{12}^{(\cdot)}$, $\Phi_i^{(\cdot)}$, Π_i	**.70**	.65	.61
Agreement among Human Subjects	Max	.88	.92	.88
	Avg	.72	.72	.71

Table 3. Spearman's ρ correlation coefficients across Mitchell and Lapata models and the projection-based models proposed in Section 3. Topical Space and Word space refer to the source spaces. is used as input to the LSA decomposition model.

DCS, evaluated over a well known dataset ([23]), is very effective for the syntactic structures of VO, NN and AdjN. It achieves the same results than the average human interannotator agreement, by outperforming most previous results ([23]). Future work on other compositional prediction tasks (e.g. selectional preference modeling or the ranking of short texts) and over different datasets will be carried out to better assess and generalize the presented results.

References

1. Montague, R.: Formal Philosophy: Selected Papers of Richard Montague. Yale University Press (1974)
2. B. Coecke, M.S., Clark, S.: Mathematical foundations for a compositional distributed model of meaning. Lambek Festschirft, Linguistic Analysis, vol. 36 **36** (2010)
3. Frege, G.: Über sinn und bedeutung. Zeitschrift für Philosophie und philosophische Kritik **100** (1892/1960) 25–50 Translated, as 'On Sense and Reference', by Max Black.
4. Firth, J.: A synopsis of linguistic theory 1930-1955. In: Studies in Linguistic Analysis. Philological Society, Oxford (1957) reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.
5. Schütze, H.: Automatic Word Sense Discrimination. Computational Linguistics **24** (1998) 97–124
6. Wittgenstein, L.: Philosophical Investigations. Blackwells, Oxford (1953)
7. Schütze, H.: Word space. In Hanson, S.J., Cowan, J.D., Giles, C.L., eds.: NIPS 5. Morgan Kaufmann Publishers, San Mateo CA (1993) 895–902
8. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research **37** (2010) 141
9. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: In Proceedings of ACL-08: HLT. (2008) 236–244
10. Baroni, M., Zamparelli, R.: Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. EMNLP '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 1183–1193
11. Grefenstette, E., Sadrzadeh, M.: Experimental support for a categorical compositional distributional model of meaning. CoRR **abs/1106.4058** (2011)
12. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM **18** (1975) 613–620
13. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JASIS **41** (1990) 391–407
14. 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 (1997) 211–240
15. Harris, Z.S.: Mathematical Structures of Language. Wiley, New York, NY, USA (1968)
16. Sahlgren, M.: The Word-Space Model. PhD thesis, Stockholm University (2006)
17. Lin, D.: Automatic retrieval and clustering of similar word. In: Proceedings of COLING-ACL, Montreal, Canada (1998)
18. Pantel, P., Lin, D.: Document clustering with committees. In: SIGIR-02, Montreal, Canada (2002) 199–206
19. Pennacchiotti, M., Cao, D.D., Basili, R., Croce, D., Roth, M.: Automatic induction of framenet lexical units. In: EMNLP. (2008) 457–465
20. Croce, D., Giannone, C., Annesi, P., Basili, R.: Towards open-domain semantic role labeling. In: ACL. (2010) 237–246
21. Foltz, P.W., Kintsch, W., Landauer, T.K., L, T.K.: The measurement of textual coherence with latent semantic analysis (1998)
22. Erk, K., Pad, S.: A structured vector space model for word meaning in context (2008)

23. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cognitive Science **34** (2010) 1388–1429
24. 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** (2009) 209–226