A Compositional Distributional Inclusion Hypothesis

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh*

School of Electronic Engineering and Computer Science
Queen Mary University of London
{d.kartsaklis;mehrnoosh.sadrzadeh}@qmul.ac.uk

Abstract. The distributional inclusion hypothesis provides a pragmatic way of evaluating entailment between word vectors as represented in a distributional model of meaning. In this paper, we extend this hypothesis to the realm of compositional distributional semantics, where meanings of phrases and sentences are computed by composing their word vectors. We present a theoretical analysis for how feature inclusion is interpreted under each composition operator, and propose a measure for evaluating entailment at the phrase/sentence level. We perform experiments on four entailment datasets, showing that intersective composition in conjunction with our proposed measure achieves the highest performance.

Key words: Computational Linguistics · Artificial Intelligence · Natural language processing · Textual entailment · Inclusion hypothesis · Compositionality · Distributional models

1 Introduction

Distributional models of meaning, where words are represented by vectors of co-occurrence frequencies gathered from corpora of text, provide a successful model for representing meanings of words and measuring the semantic similarity between them [22]. A pragmatic way for applying these models to entailment tasks is developed via the distributional inclusion hypothesis [8, 9, 11], which states that a word u entails a word v if whenever u is used so can be v. In distributional semantics terms, this means that contexts of u are included in contexts of v. For example, whenever 'boy' is used, e.g. in the sentence 'a boy runs', so can be 'person'; thus *boy* $\vdash$ *person*. By projecting this hypothesis onto a truth theoretical model, one may say that u and v stand in an entailment relation if by replacing u with v in a sentence presumed to be true, we produce a new sentence preserving that truth. For example, if the sentence 'a boy runs' is presumed to be true, so is the sentence 'a person runs', obtained by replacing 'boy' by 'person'.

One problem with distributional models of meaning is that they do not scale up to larger text constituents, such as phrases or sentences. The reason is that these do not frequently occur in corpora of text, thus the process of collecting reliable statistics to represent them as vectors does not witness the distributional hypothesis. This problem is usually addressed with the provision of a composition operator, the purpose of which is to produce vectors for phrases and sentences by combining their word vectors. Compositional distributional models of this form generally fall into three categories:

* Support by EPSRC for Career Acceleration Fellowship EP/J002607/1 and AFOSR International Scientific Collaboration Grant FA9550-14-1-0079 is gratefully acknowledged.

models based on simple element-wise operations between vectors, such as addition and multiplication [19]; tensor-based models in which relational words such as verbs and adjectives are multi-linear maps acting on noun (and noun-phrase) vectors [7, 10, 3]; and models in which the compositional operator is implemented as part of some neural network architecture [21, 12].

The purpose of this paper is to investigate, both theoretically and experimentally, the application of the distributional inclusion hypothesis on phrase and sentence vectors produced in a variety of compositional distributional models. We provide interpretations for the features of these vectors and analyse the effect of each compositional operator on the inclusion properties that hold for them. We further discuss a number of measures that have been used in the past for evaluating entailment at the lexical level. Based on the specificities introduced by the use of a compositional operator on word vectors, we propose an adaptation of the *balAPinc* measure [14]—which is currently considered a state-of-the-art in measuring entailment at the lexical level—for compositional distributional models.

The theoretical discussion is supported by experimental work. We evaluate entailment relationships between simple intransitive sentences, verb phrases, and transitive sentences, on datasets specifically created for the purposes of this work. We also present results on the $AN \vdash N$ task of [2], where the goal was to evaluate the extent to which an adjective-noun compound entails its noun. Our findings suggest that the combination of our newly proposed measure with intersective compositional models achieves the highest discriminating power when evaluating entailment at the phrase/sentence level.

Outline Sections 2 and 3 provide an introduction to compositional distributional semantics and to distributional inclusion hypothesis, respectively; Section 4 studies the inclusion properties of features in a variety of compositional distributional models, while Section 5 discusses the adaptation of the *balAPinc* measure to a compositional setting; Sections 6 and 7 deal with the experimental part; and finally, in Section 8 we briefly discuss our findings.

2 Compositional Distributional Semantics

Compositional distributional semantics represents meanings of phrases and sentences by combining the vectors of their words. In the simplest case, this is done by element-wise operations on the vectors of the words [19]. Specifically, the vector representation of a sequence of words $w_1, \ldots, w_n$ is defined to be:

$$\sum_i \overrightarrow{w}_i \qquad \text{or} \qquad \bigodot_i \overrightarrow{w}_i \tag{1}$$

where $\odot$ denotes element-wise multiplication.

A second line of research follows a more linguistically motivated approach and treats relational words as linear or multi-linear maps. These are then applied to the vectors of their arguments by following the rules of the grammar [7, 10, 3]. For example, an adjective is treated as a map $N \to N$, for N a basic noun space of the model. Equivalently, this map can be represented as a matrix living in the space $N \otimes N$. Similarly, a transitive verb is a map $N \times N \to S$, or equivalently, a "cube" or a tensor of order

3 in the space $N \otimes N \otimes S$, for S a basic sentence space of the model. Composition takes place by tensor contraction, which is a generalization of matrix multiplication to higher order tensors. For the case of an adjective-noun compound, this simplifies to matrix multiplication between the adjective matrix and the vector of its noun, while for a transitive sentence it takes the form:

$$\overrightarrow{svo} = (\overline{\text{verb}} \times \overrightarrow{\text{obj}}) \times \overrightarrow{\text{subj}} \tag{2}$$

where $\overline{\text{verb}}$ is a tensor of order 3. Compared to element-wise vector operations, note that tensor-based models adhere to a much stricter notion of composition, where the transition from grammar to semantics takes place via a structure-preserving map [7].

Finally, deep learning architectures have been applied to the production of phrase and sentence vectors, tailored for use in specific tasks. These methods have been very effective and their resulting vectors have shown state-of-the-art performances in many tasks. The main architectures usually employed are that of recursive or recurrent neural networks [21, 5] and convolutional neural networks [12]. Neural models are "opaque" for our purposes, in the sense that their non-linear multi-layer nature does not lend itself to be reasoned about in terms of the feature inclusion properties of the distributional inclusion hypothesis, and for this reason we do not deal with them in this paper.

3 The Distributional Inclusion Hypothesis

The distributional inclusion hypothesis (DIH) [8, 9, 11] is based on the fact that whenever a word u entails a word v, then it makes sense to replace instances of u with v. For example, 'cat' entails 'animal', hence in the sentence 'a cat is asleep', it makes sense to replace 'cat' with 'animal' and obtain 'an animal is asleep'. On the other hand, 'cat' does not entail 'butterfly', and indeed it does not make sense to do a similar substitution and obtain the sentence 'a butterfly is asleep'.

This hypothesis has inherent limitations, the main one being that it only makes sense in contexts that contain no logical words. For instance, the substitution of u for v would not work for sentences that have negations or quantifiers such as 'all' and 'none'. As a result, one cannot replace 'cat' with 'animal' in sentences such as 'all cats are asleep' or 'a cat is not asleep'. Despite this, the DIH has been subject to a good amount of study in the distributional semantics community and its predictions have been empirically validated to a good extent [9, 14].

Formally, if word u entails word v, then the set of features of u are included in the set of features of v. In the context of a distributional model of meaning, the term *feature* refers to a non-zero dimension of the distributional vector of a word. This makes sense since, according to DIH, word v subsumes the meaning of word u. Throughout this paper, we denote the features of a distributional vector $\overrightarrow{v}$ by $\mathcal{F}(\overrightarrow{v})$, hence we have:

$$u \vdash v \quad \text{whenever} \quad \mathcal{F}(\overrightarrow{u}) \subseteq \mathcal{F}(\overrightarrow{v}) \tag{3}$$

The research on the DIH can be categorised into two classes. In the first class, the degree of entailment between two words is based on the distance between the vector representations of the words. This distance must be measured by asymmetric means, since entailment is directional. Examples of measures used here are entropy-based measures

such as KL-divergence [4]. Abusing the notation and taking $\overrightarrow{u}$ and $\overrightarrow{v}$ to also denote their underling probability distributions, this is defined as follows:

$$D_{\mathrm{KL}}(\overrightarrow{v} \| \overrightarrow{u}) = \sum_i v_i (\ln v_i - \ln u_i) \tag{4}$$

KL-divergence is only defined when the support of $\overrightarrow{v}$ is included in the support of $\overrightarrow{u}$. In order to overcome this restriction, a variant referred to by α-skew [15] has been proposed. This is defined in the following way:

$$s_\alpha(\overrightarrow{u}, \overrightarrow{v}) = D_{\mathrm{KL}}(\overrightarrow{v} \| \alpha \overrightarrow{u} + (1-\alpha)\overrightarrow{v}) \tag{5}$$

where $\alpha \in (0, 1]$ serves as a smoothing parameter. ***Representativeness*** is another way of normalising KL-divergence; it is defined as follows:

$$R_{\mathrm{D}}(\overrightarrow{v} \| \overrightarrow{u}) = \frac{1}{1 + D_{\mathrm{KL}}(\overrightarrow{v} || \overrightarrow{u})} \tag{6}$$

Representativeness turns KL-divergence into a number in the unit interval $[0, 1]$. As a result we obtain $0 \leq R_{\mathrm{D}}(\overrightarrow{v} \| \overrightarrow{u}) \leq 1$, with $R_{\mathrm{D}}(\overrightarrow{v} \| \overrightarrow{u}) = 0$ when the support of $\overrightarrow{v}$ is not included in the support of $\overrightarrow{u}$ and $R_{\mathrm{D}}(\overrightarrow{v} \| \overrightarrow{u}) = 1$, when $\overrightarrow{u}$ and $\overrightarrow{v}$ represent the same distribution.

The research done in the second class attempts a more direct measurement of the inclusion of features, with the simplest possible case returning a binary value for inclusion or lack thereof. Measures developed by [23] and [6] advance this simple methods by arguing that not all features play an equal role in representing words and hence they should not be treated equally when it comes to measuring entailment. Some features are more "pertinent" than others and these features have to be given a higher weight when computing inclusion. For example, 'cat' can have a non-zero coordinate on all of the features 'mammal, miaow, eat, drink, sleep'. But the amount of these coordinates differ, and one can say that, for example, the higher the coordinate the more pertinent the feature. Pertinence is computed by various different measures, the most recent of which is *balAPinc* [14], defined as follows:

$$\mathit{balAPinc}(u, v) = \sqrt{\mathit{LIN}(u, v) \cdot \mathit{APinc}(u, v)} \tag{7}$$

where *LIN* is Lin's similarity [16] and *APinc* is an asymmetric measure defined as below:

$$\mathit{APinc}(u, v) = \frac{\sum_r \left[P(r) \cdot \mathit{rel}'(f_r) \right]}{|\mathcal{F}(\overrightarrow{u})|} \tag{8}$$

APinc applies the DIH via the idea that features with high values in $\mathcal{F}(\overrightarrow{u})$ must also have high values in $\mathcal{F}(\overrightarrow{v})$. In the above formula, f_r is the feature in $\mathcal{F}(\overrightarrow{u})$ with rank r; $P(r)$ is the precision at rank r; and $\mathit{rel}'(f_r)$ is a weight computed as follows:

$$rel'(f) = \begin{cases} 1 - \frac{\mathit{rank}(f, \mathcal{F}(\overrightarrow{v}))}{|\mathcal{F}(\overrightarrow{v})|+1} & f \in \mathcal{F}(\overrightarrow{v}) \\ 0 & o.w. \end{cases} \tag{9}$$

where *rank*$(f, \mathcal{F}(\overrightarrow{v}))$ shows the rank of feature f within the entailed vector. In general, *APinc* can be seen as a version of average precision that reflects lexical inclusion.

We will return to the topic of entailment measures in Section 5, where we propose variations on *APinc* and *balAPinc* that are more appropriate for entailment in compositional distributional models.

4 A Compositional Distributional Inclusion Hypothesis

In the presence of a compositional operator, features of a phrase/sentence adhere to some set-theoretic properties. In what follows, we present these properties for a number of operators in various compositional distributional models.

4.1 Element-wise Composition

For simple additive and multiplicative models, the set of features of the phrase/sentence are easily derived from the set of features of their words using the set-theoretic operations of union and intersection:

$$\mathcal{F}(\overrightarrow{v_1} + \cdots + \overrightarrow{v_n}) = \mathcal{F}(\overrightarrow{v_1}) \cup \cdots \cup \mathcal{F}(\overrightarrow{v_n}) \tag{10}$$

$$\mathcal{F}(\overrightarrow{v_1} \odot \cdots \odot \overrightarrow{v_n}) = \mathcal{F}(\overrightarrow{v_1}) \cap \cdots \cap \mathcal{F}(\overrightarrow{v_n}) \tag{11}$$

The features of a tensor product of vectors consists of tuples of same-indexed features, taken from their cartesian product:

$$\mathcal{F}(\overrightarrow{v_1} \otimes \cdots \otimes \overrightarrow{v_n}) = \{(v_i^1, \cdots, v_i^n) \mid v_i^j \in \mathcal{F}(\overrightarrow{v_j})\} \tag{12}$$

where v_i^j refers to the ith element of the jth vector. Point-wise minimum and maximum of vectors act inline with intersection and union respectively, providing a feature inclusion behaviour identical to addition and point-wise multiplication.

$$\mathcal{F}(\max(\overrightarrow{v_1}, \cdots, \overrightarrow{v_n})) = \mathcal{F}(\overrightarrow{v_1}) \cup \cdots \cup \mathcal{F}(\overrightarrow{v_n}) \tag{13}$$

$$\mathcal{F}(\min(\overrightarrow{v_1}, \cdots, \overrightarrow{v_n})) = \mathcal{F}(\overrightarrow{v_1}) \cap \cdots \cap \mathcal{F}(\overrightarrow{v_n}) \tag{14}$$

In order to see this, let us consider the $\max$ case. In the linear expansion notation, we have:

$$\max(\overrightarrow{v_1}, \cdots, \overrightarrow{v_n}) = \sum_i \max(v_i^1, v_i^2, \cdots, v_i^n)\overrightarrow{a}_i$$

where $\{\overrightarrow{a_i}\}_i$ is an orthonormal basis of space V where vectors $\overrightarrow{v_i}$ live. For any arbitrary dimension $\overrightarrow{a_j}$, it is the case that $\overrightarrow{a_j} \in \mathcal{F}(\max(\overrightarrow{v_1}, \cdots, \overrightarrow{v_n}))$ iff $\max(v_j^1, v_j^2, \cdots, v_j^n) \neq 0$. For this to happen, it suffices that one of the v_j^i's is nonzero, that is $v_j^1 \neq 0$ *or* $v_j^2 \neq 0$ *or* $\cdots$ *or* $v_j^n \neq 0$, which is equivalent to saying that $\overrightarrow{a_j} \in \mathcal{F}(\overrightarrow{v_1}) \cup \cdots \cup \mathcal{F}(\overrightarrow{v_n})$. The case for $\min$ is similar, with the difference that *or* is replaced with *and*, hence the set theoretic operation $\cup$ with $\cap$.

Element-wise composition has certain desirable properties in relation to the DIH. Firstly, it lifts naturally from the word level to phrase/sentence level; specifically, for two sentences $s_1 = u_1 \dots u_n$ and $s_2 = v_1 \dots v_n$ for which $u_i \vdash v_i,\ \forall i \in [1, n]$, it is always the case that $s_1 \vdash s_2$. This is a special case of a theorem proved in [1] for general tensor-based models. As an example, consider two intransitive sentences "*subj*$_1$ *verb*$_1$" and "*subj*$_2$ *verb*$_2$", for which we have $\mathcal{F}(\overrightarrow{subj}_1) \subseteq \mathcal{F}(\overrightarrow{subj}_2)$ and $\mathcal{F}(\overrightarrow{verb}_1) \subseteq \mathcal{F}(\overrightarrow{verb}_2)$; then, it is the case that:

$$\mathcal{F}(\overrightarrow{subj}_1) \cap \mathcal{F}(\overrightarrow{verb}_1) \subseteq \mathcal{F}(\overrightarrow{subj}_2) \quad \text{and} \quad \mathcal{F}(\overrightarrow{subj}_1) \cap \mathcal{F}(\overrightarrow{verb}_1) \subseteq \mathcal{F}(\overrightarrow{verb}_2)$$

and consequently:

$$\mathcal{F}(\overrightarrow{subj}_1) \cap \mathcal{F}(\overrightarrow{verb}_1) \subseteq \mathcal{F}(\overrightarrow{subj}_2) \cap \mathcal{F}(\overrightarrow{verb}_2)$$

A similar reasoning holds for the union-based case, since we have:

$$\mathcal{F}(\overrightarrow{subj}_1) \subseteq \mathcal{F}(\overrightarrow{subj}_2) \cup \mathcal{F}(\overrightarrow{verb}_2) \quad \text{and} \quad \mathcal{F}(\overrightarrow{verb}_1) \subseteq \mathcal{F}(\overrightarrow{subj}_2) \cup \mathcal{F}(\overrightarrow{verb}_2)$$

thus $\mathcal{F}(\overrightarrow{subj}_1) \cup \mathcal{F}(\overrightarrow{verb}_1) \subseteq \mathcal{F}(\overrightarrow{subj}_2) \cup \mathcal{F}(\overrightarrow{verb}_2)$. For the case of intersective composition, the above makes clear another DIH property that holds in contexts without logical words; that a phrase can be replaced with each one of its words, i.e. *red car* can be replaced with *car* and with *red*. Note, however, that in this case the same is not true for union-based composition, since the inclusion order becomes reversed, which is clearly unwanted.

4.2 Holistic Phrase/Sentence Vectors

In the ideal (but not so feasible) presence of a text corpus sufficiently large to provide co-occurrence statistics for phrases or even sentences, one could directly create vectors for larger text segments using the same methods as if they were words. This idea has been investigated in the context of entailment by [2], who present promising results for short adjective-noun compounds. *Holistic* vectors of this sort are interesting since they can be seen as representing (at least for short text segments) some form of idealistic distributional behaviour for text segments above the word level. For this reason, we briefly examine the relationship of these models with the compositional models of Section 4.1, with regard to their feature inclusion properties.

We consider the case of intersective composition. For a two-word phrase $w_1 w_2$ with a holistic vector $\overrightarrow{w_1 w_2}$, we start by noticing that $\mathcal{F}(\overrightarrow{w_1 w_2})$ is always a subset of $\mathcal{F}(\overrightarrow{w_1}) \cap \mathcal{F}(\overrightarrow{w_2})$ and specifically the subset referring to cases where w_1 and w_2 occur *together* in the same context, that is:

$$\mathcal{F}(\overrightarrow{w_1 w_2}) = [\mathcal{F}(\overrightarrow{w_1}) \cap \mathcal{F}(\overrightarrow{w_2})]_{|w_1, w_2} \subseteq \mathcal{F}(\overrightarrow{w_1}) \cap \mathcal{F}(\overrightarrow{w_2})$$

with the set equality to hold only when w_1 and w_2 occur exclusively in the same contexts, i.e. the presence of w_1 always signifies that w_2 is around and vice versa. The

relationship between holistic vectors and intersective composition can be leveraged to the phrase/sentence level. Recall the intransitive sentence example of Section 4.1; denoting the holistic vectors of the two sentences as $\overrightarrow{s_1 v_1}$ and $\overrightarrow{s_2 v_2}$, it is the case that:

$$\mathcal{F}(\overrightarrow{s_1 v_1}) \subseteq \mathcal{F}(\overrightarrow{s_2 v_2}) \subseteq \mathcal{F}(\overrightarrow{s_2}) \cap \mathcal{F}(\overrightarrow{v_2})$$

In other words, intersective composition preserves any entailment relation that holds at the holistic vector level, providing a faithful approximation of the holistic distributional behaviour. Note that for the case of union-based composition this approximation will be much more relaxed, and thus less useful in practice.

4.3 Tensor-based Models

For tensor-based models, one needs a different analysis. These models lie somewhere between intersective and union-based models. Consider the simple case of a matrix multiplication between a $m \times n$ matrix $\mathcal{M}$ and a $n \times 1$ vector $\overrightarrow{v}$, given below:

$$\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ w_{21} & \cdots & w_{2n} \\ \vdots & & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix} \times \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

The matrix $\mathcal{M}$ can be seen as a list of column vectors $(\overrightarrow{w_1}, \overrightarrow{w_2}, \cdots, \overrightarrow{w_n})$, where $\overrightarrow{w_i} = (w_{1i}, \cdots, w_{mi})^{\mathrm{T}}$. Then the result of the matrix multiplication becomes a combination of scalar multiplications of element v_i of the vector $\overrightarrow{v}$ with its corresponding vectors $\overrightarrow{w_i}$ of the matrix $\mathcal{M}$, as follows:

$$v_1 \overrightarrow{w_1} + v_2 \overrightarrow{w_2} + \cdots + v_n \overrightarrow{w_n}$$

By looking at matrix multiplication $\mathcal{M} \times \overrightarrow{v}$ in this way, we are able to describe the features of $\mathcal{F}(\mathcal{M} \times \overrightarrow{v})$ in terms of the features of $\overrightarrow{v}$ and the features of the $\overrightarrow{w_i}$'s of $\mathcal{M}$. This is as follows:

$$\mathcal{F}(\overline{w} \times \overrightarrow{v}) = \bigcup_{v_i \neq 0} \mathcal{F}(\overrightarrow{w_i}) \tag{15}$$

Generalizing slightly and calling v_i a feature whenever it is non-zero, the above can be written down in the following equivalent form:

$$\bigcup_i \mathcal{F}(\overrightarrow{w_i}) \mid_{\mathcal{F}(v_i)} \tag{16}$$

which means we collect features of each $\overrightarrow{w_i}$ vector but only up to "featureness" of v_i, that is up to v_i being non-zero.

The above procedure can be extended to tensors of higher order; a tensor of order 3, for example, can be seen as a list of matrices, a tensor of order 4 as a list of "cubes" and so on. For the case of this paper, we will not go beyond matrix multiplication and cube contraction. The concrete constructions of these matrices and cubes, presented in the next section, will make the above analysis more clear.

Concrete Tensor-based Constructions While the feature inclusion properties of a tensor-based model follow the generic analysis above, their exact form depends on the concrete constructions of their underlying tensors. In this section, we go over a few different methods of tensor construction and derive their feature inclusion properties.

We start by the construction presented in [10], which builds a tensor from the properties of the vectors of its arguments. For example, an intransitive verb gets assigned the vector $\sum_i \overrightarrow{Sbj}_i$, a verb phrase the vector $\sum_i \overrightarrow{Obj}_i$, and a transitive verb the matrix $\sum_i \overrightarrow{Sbj}_i \otimes \overrightarrow{Obj}_i$. Here, Sbj_i/Obj_i are the subjects/objects of the verb across the corpus. The features of the phrases vo and sentences sv, svo (where s/o are the subject/object of the phrase/sentence) are as follows:

$$\mathcal{F}(\overrightarrow{sv}) = \bigcup_i \mathcal{F}(\overrightarrow{Sbj}_i) \cap \mathcal{F}(\overrightarrow{s}) \qquad \mathcal{F}(\overrightarrow{vo}) = \bigcup_i \mathcal{F}(\overrightarrow{Obj}_i) \cap \mathcal{F}(\overrightarrow{o})$$
$$\mathcal{F}(\overrightarrow{svo}) = \bigcup_i \mathcal{F}(\overrightarrow{Sbj}_i \otimes \overrightarrow{Obj}_i) \cap \mathcal{F}(\overrightarrow{s}) \otimes \mathcal{F}(\overrightarrow{o})$$

The disadvantage of this model and a number of other models based on this methodology, e.g. [13, 18], is that their resulting representations of verbs have one dimension less than what their types dictate. According to the type assignments, an intransitive verb has to be a matrix and a transitive verb a cube, where as in the above we have a vector and a matrix. We remedy this problem by arguing that the sentence/phrase space should be spanned by the vectors of the arguments of the verb across the corpus. In order to achieve this, we create verb matrices for intransitive sentences and verb phrases by taking the outer product of the argument vectors with themselves, hence obtaining:

$$\overline{v}_{itv} := \sum_i \overrightarrow{Sbj}_i \otimes \overrightarrow{Sbj}_i \qquad \overline{v}_{vp} := \sum_i \overrightarrow{Obj}_i \otimes \overrightarrow{Obj}_i \tag{17}$$

When these verbs are composed with some subject/object to form a phrase/sentence, each vector in the spanning space is weighted by its similarity (assuming normalized vectors) with the vector of that subject/object, that is:

$$\overrightarrow{sv} = \overrightarrow{s} \times \overline{v}_{itv} = \sum_i \langle \overrightarrow{Sbj}_i | \overrightarrow{s} \rangle \overrightarrow{Sbj}_i \tag{18}$$

$$\overrightarrow{vo} = \overline{v}_{vp} \times \overrightarrow{o} = \sum_i \langle \overrightarrow{Obj}_i | \overrightarrow{o} \rangle \overrightarrow{Obj}_i \tag{19}$$

We call this model *projective*. For the case of a transitive verb (a function of two arguments), we define the sentence space to be spanned by the average of the argument vectors, obtaining:

$$\overline{v}_{trv} := \sum_i \overrightarrow{Sbj}_i \otimes \left(\frac{\overrightarrow{Sbj}_i + \overrightarrow{Obj}_i}{2} \right) \otimes \overrightarrow{Obj}_i \tag{20}$$
$$\overrightarrow{svo} = \sum_i \langle \overrightarrow{s} \mid \overrightarrow{Sbj}_i \rangle \left(\frac{\overrightarrow{Sbj}_i + \overrightarrow{Obj}_i}{2} \right) \langle \overrightarrow{Obj}_i \mid \overrightarrow{o} \rangle$$

Feature-wise, the above translate to the following:

$$\mathcal{F}(\overrightarrow{sv}) = \bigcup_i \mathcal{F}(\overrightarrow{Sbj}_i) \mid_{\mathcal{F}\left(\langle \overrightarrow{Sbj}_i | \overrightarrow{v} \rangle\right)} \qquad \mathcal{F}(\overrightarrow{vo}) = \bigcup_i \mathcal{F}(\overrightarrow{Obj}_i) \mid_{\mathcal{F}\left(\langle \overrightarrow{Obj}_i | \overrightarrow{v} \rangle\right)}$$

$$\mathcal{F}(\overrightarrow{svo}) = \bigcup_i \left(\mathcal{F}(\overrightarrow{Sbj}_i) \cup \mathcal{F}(\overrightarrow{Obj}_i)\right) \mid_{\mathcal{F}(\langle \overrightarrow{s} | \overrightarrow{Sbj}_i \rangle)\mathcal{F}(\langle \overrightarrow{Obj}_i | \overrightarrow{v} \rangle)}$$

Informally, we can think of the terms following the | symbol as defining a restriction on feature inclusion based on how well the arguments of the phrase/sentence fit to the arguments of the verb. We close this section by noting that in Section 6.2 we briefly present a statistical approach for creating the verb matrices based on holistic phrase vectors, along the lines of [3].

5 Measuring the CDIH

When computing entailment at the lexical level, *balAPinc* (Eq. 7) has been found to be one of the most successful measures [14]. However, the transition from words to phrases or sentences introduces extra complications, which we need to take into account. Firstly, in a compositional distributional model, the practice of considering only non-zero elements of the vectors as features becomes too restrictive and thus suboptimal for evaluating entailment; indeed, depending on the form of the vector space and the applied compositional operator (especially in intersective models), an element can get very low values without however ever reaching zero. This blurring of the notion of "featureness"— in which zero can be seen as a lower bound in a range of possible values— is in line with the quantitative nature of these models. In this paper we exploit this to the limit by letting $\mathcal{F}(\overrightarrow{w})$ to include all the dimensions of $\overrightarrow{w}$.

Secondly, we further exploit the continuous nature of distributional models by providing a stronger realization of the idea that $u \vdash v$ whenever v occurs in all the contexts of u. Let $f_r^{(u)}$ be a feature in $\mathcal{F}(\overrightarrow{u})$ with rank r and $f_r^{(v)}$ the corresponding feature in $\mathcal{F}(\overrightarrow{v})$, we remind that Kotlerman et al. consider that feature inclusion holds at rank r whenever $f_r^{(u)} > 0$ and $f_r^{(v)} > 0$; we strengthen this assumption by requiring that $f_r^{(u)} \leq f_r^{(v)}$. Incorporating these modifications in the *APinc* measure, we redefine $P(r)$ and $rel'(f_r)$ in Equation 8 as:

$$P(r) = \frac{\left|\{f_r^{(u)} | f_r^{(u)} \leq f_r^{(v)}, 0 < r \leq |\overrightarrow{u}|\}\right|}{r} \tag{21}$$

$$rel'(f_r) = \begin{cases} 1 & f_r^{(u)} \leq f_r^{(v)} \\ 0 & o.w. \end{cases} \tag{22}$$

Note that the new relevance function essentially subsumes the old one (Equation 9), since by definition high-valued features in $\mathcal{F}(\overrightarrow{u})$ must be even higher in $\mathcal{F}(\overrightarrow{v})$. We now re-define *APinc* at the phrase/sentence level to be the following:

$$\textit{SAPinc}(u, v) = \frac{\sum_r \left[P(r) \cdot \textit{rel}'(f_r)\right]}{|\overrightarrow{u}|} \tag{23}$$

where $P(r)$ and $rel'(f_r)$ are as defined in Equations 21 and 22, respectively, and $|\overrightarrow{u}|$ is the number of dimensions of $\overrightarrow{u}$. We further notice that when using *SAPinc*, a zero vector vacuously entails every other vector in the vector space, and it is entailed only by itself, as is the case for logical entailment.

We now proceed to examine the balanced *APinc* version, to which Kotlerman et al. refer as *balAPinc* (Equation 7). This is the geometric average of an asymmetric measure (*APinc*) with a symmetric one (Lin's similarity). The rationale of including a symmetric measure in the computation was that *APinc* tends to return unjustifyingly high scores when the entailing word is infrequent, that is, when the feature vector of the entailing word is very short; the purpose of the symmetric measure was to penalize the result, since in this case the similarity of the narrower term with the broader one is usually low. However, now that all feature vectors have the same length, such a balancing action is unnecessary; even more importantly, it introduces a strong element of symmetry in a measure that is intended to be strongly asymmetric. To cope with these issues, we propose to replace Lin's similarity with representativeness on KL-divergence (Equation 6), and define a sentence-level version of *balAPinc* between two word vectors $\overrightarrow{u}$ and $\overrightarrow{v}$ as follows:

$$\textit{SBalAPinc}(u, v) = \sqrt{R_{\mathrm{D}}(\overrightarrow{u} \| \overrightarrow{v}) \cdot \textit{SAPinc}(\overrightarrow{u}, \overrightarrow{v})} \tag{24}$$

Recall that $R_{\mathrm{D}}(p\|q)$ is asymmetric, measuring the extent to which q represents (i.e. is similar to) p. So the term $R_{\mathrm{D}}(\overrightarrow{u} \| \overrightarrow{v})$ in the above formula measures how well the *broader* term v represents the narrower one u; as an example, we can think that the term 'animal' is representative of 'cat', while the reverse is not true. The new measure aims at: (i) retaining a strongly asymmetric nature; and (ii) providing a more fine-grained element of evaluating entailment.

6 Experimental Setting

We evaluate the compositional models and the entailment measures presented above in four different tasks. Specifically, we measure upward-monotone entailment between (a) intransitive sentences; (b) verb phrases; (c) transitive sentences; and (d) adjective-noun compounds and nouns. The first three evaluations are based on datasets specifically created by us for the purposes of this paper, while for the adjective-noun task we use the dataset of [2]. In all cases, we first apply a compositional model to the phrases/sentences of each pair in order to create vectors representing their meaning, and then we evaluate the entailment relation between the phrases/sentences by using these composite vectors as input to a number of entailment measures. The goal is to see which combination of compositional model/entailment measure is capable of better recognizing strictly directional entailment relationships between phrases and sentences.

In all the experiments, we used a 300-dimensional PPMI vector space trained on the concatenation of UKWAC and Wikipedia corpora. The context was defined as a 5-word window around the target word.

6.1 Datasets

In this section we briefly describe the process we followed in order to create datasets for deciding entailment between subject-verb, verb-object, and subject-verb-object phrases

and sentences. Our goal was to produce pairs of phrases/sentences that stand in an upward-monotone entailment relationship to each other. When entailing and entailed phrases have exactly the same structure, as is in our case, one way to achieve that is to ensure that every word in the entailed phrase is a hypernym of the corresponding word in the entailing phrase. We achieved this by using hyponym-hypernym relationships taken by WordNet as follows.

Firstly, we extracted from the concatenation of UKWAC and Wikipedia corpora all verbs occurring at most 2.5 million times and at least 5000 times. Then, each verb was paired with a hypernym of its main synset, creating a list of 4800 pairs of verbs that stand in a hyponym-hypernym relation. Each verb was associated with a list of argument nouns; for the intransitive task this list contained nouns occurring in the corpus as subjects of the verbs, for the verb phrase nouns in an object relationship, and for the transitive task subject/object pairs. Starting from the most frequent cases, each argument of an entailing verb was paired with an argument of the corresponding entailed verb based a number of constraints (for example, each noun could occur at most 3 times as part of an entailing phrase, and a specific phrase can only occur once as entailing phrase).[1] We went through the phrase/sentence pairs manually and discarded any instance where we judged to be nonsensical. This process resulted in 135 subject-verb pairs, 218 verb-object pairs, and 70 subject-verb-object pairs, the phrases/sentences of which stand in a fairly clear entailment relationship. Each dataset was extended with the reverse direction of the entailments as negative examples, creating three strictly directional entailment datasets of 270 (subject-verb), 436 (verb-object) and 140 (subject-verb-object) entries. Table 1 presents a sample of positive entailments from each dataset.[2]

6.2 Compositional Models

We tested the additive and multiplicative compositional operators, as defined in Equation 1, a point-wise minimum model as discussed in Section 4.1, and a variation on the tensor-based model introduced via Equations 17-20. In relation to this latter model, informal experimentation showed that by taking into account directly the features of the distributional vector of the verb, the results improve. Let the distributional vector of the verb be $\overrightarrow{v}$ and the verb tensor be $\overline{v}_x$, as computed in Equations 17-20, for $x \in \{itv, vp, trv\}$. Then a new tensor is computed via the formula $\widetilde{v}_x := \overrightarrow{v} \odot \overline{v}_x$, the feature inclusion behaviour of which is derivable as follows:

$$\mathcal{F}(\widetilde{v}_x) = \mathcal{F}(\overrightarrow{v}) \cap \mathcal{F}(\overline{v}_x)$$

For the experiments on the intransitive and the verb-phrase datasets, we also use a least-squares fitting model for approximating the distributional behaviour of holistic vectors (see discussion in Section 4.2), along the lines of [3]. For each verb, we compute analytically an estimator for predicting the ith element of the resulting vector as follows:

$$\overrightarrow{w_i} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\overrightarrow{y_i}$$

[1] These constraints were much more relaxed for the transitive task, because of data sparsity problems.

[2] The datasets will become available at `http://compling.eecs.qmul.ac.uk/resources/`.

Subject-verb	Verb-object
evidence suggest ⊢ information express	develop skill ⊢ create ability
people believe ⊢ group think	solve problem ⊢ understand difficulty
paper present ⊢ material show	sign contract ⊢ write agreement
station serve ⊢ facility meet	reduce number ⊢ decrease amount
survey reveal ⊢ work show	publish book ⊢ produce publication
student develop ⊢ person create	sing song ⊢ perform music
company operate ⊢ organization manage	rejoin army ⊢ join force
player play ⊢ contestant compete	gain experience ⊢ obtain education
study demonstrate ⊢ examination show	serve purpose ⊢ meet goal
news come ⊢ message travel	identify area ⊢ determine location
summer finish ⊢ season end	promote development ⊢ support event
report note ⊢ document state	suffer injury ⊢ experience condition
book offer ⊢ product supply	undertake research ⊢ initiate investigation
tree mature ⊢ plant grow	drive car ⊢ handle vehicle

Subject-verb-object
report describe result ⊢ document explain process
report outline progress ⊢ document describe change
value suit budget ⊢ number meet standard
book present account ⊢ work show evidence
woman marry man ⊢ female join male
author retain house ⊢ person hold property
report highlight lack ⊢ document stress need
public trust reference ⊢ people accept message
study demonstrate importance ⊢ work show value
police fight crime ⊢ force compete activity
experiment test hypothesis ⊢ research evaluate proposal
university publish paper ⊢ body produce research
brochure outline feature ⊢ booklet explain concept
widow sell estate ⊢ woman exchange property

Table 1. Positive entailments from the three tasks at phrase and sentence level.

Here, the rows of matrix $\mathbf{X}$ are the vectors of the subjects (or objects) that occur with our verb, and $\overrightarrow{y_i}$ is a vector containing the ith elements of the holistic phrase vectors across all training instances; the resulting $\overrightarrow{w_i}$'s form the rows of our verb matrix. Finally, a non-compositional baseline, where the phrase is represented by the vector (or tensor) of its head verb, is also evaluated where appropriate.

6.3 Measures and Evaluation

We present results for a variety of entailment measures, including *SAPinc* and *SBal-APinc* as introduced in Section 5. KL-divergence is applied on smoothed vectors, as suggested by [4]. For α-skew, we use $\alpha = 0.99$ which in the past has showed the best reporting results [14]. *WeedsPrec* refers to the precision measure introduced by [23], while *ClarkeDE* denotes the degree of entailment measure of [6]. We also use strict feature inclusion as a baseline; in this case, entailment holds only when $\mathcal{F}(\overrightarrow{phrase_1}) \subseteq \mathcal{F}(\overrightarrow{phrase_2})$. After composition, all phrase/sentence vectors are normalized to unit length.

Regarding evaluation, since the tasks follow a binary classification objective and our models return a continuous value, we report area under curve (AUC). This reflects the generic discriminating power of a binary classifier by evaluating the task at every possible threshold.

7 Results

7.1 Phrase and Sentence Entailment

Table 2 presents the results for the phrase and sentence entailment experiments. As the numbers show, in all three tasks the highest performance is delivered by a combination of *SBalAPinc* or *SAPinc* with element-wise vector multiplication. Furthermore, it is interesting to note that *SBalAPinc* clearly outperforms *balAPinc* in every compositional model and every task. The ability of the proposed measure to better discriminate between positive and negative entailments is further demonstrated in Figure 1, where we examine the distributions of the two classes when using *balAPinc* (left) and *SBalAPinc* (right) in conjunction with multiplicative composition for the verb-object task.

	Subject-verb						**Verb-object**						**Subject-verb-object**					
Measure	Verb	$\odot$	MIN	$+$	$\otimes$	LstSqr	Verb	$\odot$	MIN	$+$	$\otimes$	LstSqr	Verb	$\odot$	MIN	$+$	$\otimes$	LstSqr
Inclusion	0.59	0.54	0.54	0.63	0.59	0.50	0.58	0.52	0.52	0.64	0.58	0.50	0.61	0.55	0.55	0.58	0.64	–
KL-div.	0.59	0.66	0.68	0.57	0.59	0.59	0.62	0.64	0.66	0.61	0.60	0.58	0.61	0.65	0.71	0.54	0.60	–
αSkew	0.63	0.75	0.72	0.74	0.65	0.62	0.65	0.74	0.70	0.75	0.66	0.57	0.66	0.74	0.74	0.71	0.70	–
WeedsPrec	0.67	0.75	0.75	0.65	0.67	0.59	0.67	0.70	0.71	0.68	0.67	0.56	0.69	0.79	0.78	0.59	0.69	–
ClarkeDE	0.57	0.66	0.63	0.62	0.59	0.56	0.58	0.67	0.63	0.63	0.60	0.53	0.58	0.67	0.63	0.60	0.61	–
APinc	0.69	0.78	0.78	0.72	0.70	0.60	0.69	0.75	0.75	0.74	0.70	0.56	0.74	0.76	0.77	0.65	0.74	–
balAPinc	0.65	0.72	0.71	0.70	0.67	0.58	0.66	0.70	0.69	0.71	0.67	0.55	0.67	0.71	0.71	0.64	0.70	–
SAPinc	0.65	**0.81**	0.74	0.72	0.71	0.63	0.62	**0.82**	0.74	0.72	0.68	0.58	0.59	**0.80**	0.73	0.67	0.75	–
SBalAPinc	0.65	**0.81**	0.75	0.72	0.69	0.64	0.66	**0.79**	0.74	0.73	0.68	0.59	0.63	**0.80**	0.76	0.67	0.76	–

Table 2. AUC scores for the three phrase and sentence entailment tasks. *Verb* is a non-compositional baseline based on comparing only the verb vectors of the two phrases, $\odot$ is element-wise vector multiplication, $+$ vector addition, $\otimes$ tensor-based composition, and *LstSqr* a least-square fitting model approximating the holistic distributional behaviour of the phrases.

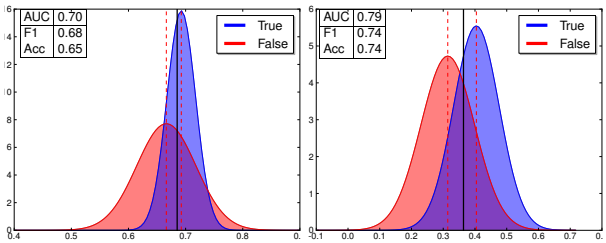

Fig. 1. The distributions of positive and negative entailments when using *balAPinc* (left) and *SBalAPinc* (right) in combination with multiplicative composition on the verb-object task. The dashed red lines indicate the means, while the thick black lines correspond to the thresholds that optimize *informedness*—equivalent to AUC subtended by the highest operating point [20].

7.2 Adjective-Noun Compounds

In this last experiment, we reproduce the $AN \vdash N$ task of [2], the goal of which is to assess the extent to which an adjective-noun compound (such as 'red car') entails the noun of the compound ('car'). The dataset contains 2450 pairs of $AN \vdash N$ entailments, half of which are negative examples that have been created by random permutation of the nouns at the right-hand side. We use this task as a proof of concept for the theory detailed in Section 4, since when using element-wise composition this sort of entailment always holds. The results, presented in Table 3, confirm the above in the most definite way. *SBalAPinc* achieves almost perfect classification when combined with multiplicative composition, while *SAPinc* shows top performance for union-based composition.

Measure	$\odot$	MIN	$+$
Inclusion	1.00	1.00	0.50
KL-divergence	1.00	1.00	0.87
αSkew	0.96	0.97	1.00
WeedsPrec	1.00	1.00	0.85
ClarkeDE	1.00	1.00	0.95
APinc	0.94	0.94	0.84
balAPinc	0.99	0.99	0.84
SAPinc	0.91	0.12	0.97
SBalAPinc	0.99	0.74	0.93

Table 3. AUC scores for the $AN \vdash N$ task.

8 Discussion

The experimental work presented in Sections 6 and 7 provides evidence that the measures introduced in this paper are appropriate for evaluating feature inclusion at the sentence level, especially in relation to element-wise vector multiplication as a compositional operator. This form of intersective composition seems to show a consistently high performance across all tested measures—an observation that is in line with the desired theoretical properties of these models as discussed in Section 4. This implies that the intersective composition is especially suitable for sentence entailment evaluation based on the CDIH. The reason may be the feature filtering methods applied by these models. The intersective filtering avoids generation of very dense vectors and thus facilitates entailment judgements based on the CDIH. On the other hand, union-based compositional models, such as vector addition, produce dense vectors for even very short sentences (Figure 2). In this case, entailment is better handled by information theoretic measures, and specifically the α-skew measure (Table 2), without however reaching the performance of intersective models and feature inclusion.

The tensor-based model presented in Section 4.3 can be seen as a combination of a union-based model (between the features of the arguments of the verb) and an intersective model (between the features of the distributional vector of the verb and the features

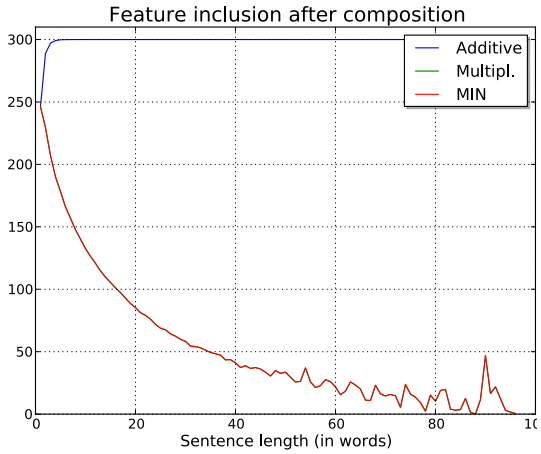

Fig. 2. Feature inclusion on the first million sentences of Wikipedia for three vector-based compositional models (using vectors of 300 dimensions). For sentence lengths greater than 5 words, additive composition produces dense vectors with all elements greater than zero. The feature inclusion behaviour of the two intersective models (vector multiplication and MIN) is identical, showing a polynomial decrease on the number of features for longer sentences.

of the vector of the verb phrase). While this idea does not seem to work very well in practice—as it returns results lower than those of the vector-based counterparts—the model outperforms the other full tensor model, that is the least-square fitting model. One reason is that tensor-based constructions similar to the ones in Equations 17-20 are more robust against data sparsity problems than statistical models based on holistic vectors of phrases and sentences.

In general, while intersective element-wise vector composition seems to be more aligned with a CDIH, tensor-based models, similar to the one presented in Section 4.3, provide an abundance of conceptual options, depending on how one creates the verb tensors. At the same time, the tensor-based models preserve the grammatical structure. Hence they can serve as an interesting test-bed for reasoning on entailment relations at the phrase or sentence level.

9 Conclusion and Future Work

In this paper we investigated the application of the distributional inclusion hypothesis on evaluating entailment between phrase and sentence vectors produced by compositional operators. We showed how the popular *balAPinc* measure for evaluating entailment at the lexical level can be lifted to a new measure *SBalAPinc* for use at the phrase/sentence level. Our results showed that intersective composition with *SBalAPinc* achieves the best performance. Experimenting with different versions of tensor models for entailment is an interesting topic that we plan to address in a future paper. Furthermore, the extension of word-level entailment to phrases and sentences provides connections with

natural logic [17], a topic that is worth a separate treatment and constitutes a future direction.

References

1. Balkır, E., Kartsaklis, D., Sadrzadeh, M.: Sentence Entailment in Compositional Distributional Semantics. In: Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM). Fort Lauderdale, FL (January 2016)
2. Baroni, M., Bernardi, R., Do, N.Q., Shan, C.c.: Entailment above the word level in distributional semantics. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 23–32. Association for Computational Linguistics, Avignon, France (April 2012), http://www.aclweb.org/anthology/E12-1004
3. Baroni, M., Zamparelli, R.: Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 1183–1193. Association for Computational Linguistics, Cambridge, MA (October 2010), http://www.aclweb.org/anthology/D10-1115
4. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. pp. 310–318. ACL '96, Association for Computational Linguistics, Stroudsburg, PA, USA (1996), http://dx.doi.org/10.3115/981863.981904
5. Cheng, J., Kartsaklis, D.: Syntax-aware multi-sense word embeddings for deep compositional models of meaning. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1531–1542. Association for Computational Linguistics, Lisbon, Portugal (September 2015), http://aclweb.org/anthology/D15-1177
6. Clarke, D.: Context-theoretic semantics for natural language: an overview. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics. pp. 112–119. Association for Computational Linguistics, Athens, Greece (March 2009), http://www.aclweb.org/anthology/W09-0215
7. Coecke, B., Sadrzadeh, M., Clark, S.: Mathematical foundations for a compositional distributional model of meaning. Linguistic Analysis 36 (2010)
8. Dagan, I., Lee, L., Pereira, F.C.N.: Similarity-based models of word cooccurrence probabilities. Mach. Learn. 34(1-3), 43–69 (1999), doi=10.1023/A:1007537716579
9. Geffet, M., Dagan, I.: The distributional inclusion hypotheses and lexical entailment. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05). pp. 107–114. Association for Computational Linguistics, Ann Arbor, Michigan (June 2005), http://www.aclweb.org/anthology/P05-1014
10. Grefenstette, E., Sadrzadeh, M.: Experimental support for a categorical compositional distributional model of meaning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1394–1404. Association for Computational Linguistics (2011)
11. Herbelot, A., Ganesalingam, M.: Measuring semantic content in distributional vectors. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. vol. 2, pp. 440–445. Association for Computational Linguistics (2013)
12. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 655–665. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/P14-1062
13. Kartsaklis, D., Sadrzadeh, M., Pulman, S.: A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In: COLING 2012, 24th

International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India. pp. 549–558 (2012)
14. Kotlerman, L., Dagan, I., Szpektor, I., Zhitomirsky-Geffet, M.: Directional distributional similarity for lexical inference. Natural Language Engineering 16(04), 359–389 (2010)
15. Lee, L.: Measures of distributional similarity. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 25–32 (1999)
16. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the International Conference on Machine Learning. pp. 296–304 (1998)
17. MacCartney, B., Manning, C.D.: Natural logic for textual inference. In: ACL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics (2007)
18. Milajevs, D., Kartsaklis, D., Sadrzadeh, M., Purver, M.: Evaluating neural word representations in tensor-based compositional settings. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 708–719. Association for Computational Linguistics, Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1079
19. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cognitive Science 34(8), 1388–1439 (2010)
20. Powers, D.M.: Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies 2(1), 37–63 (2011)
21. Socher, R., Huval, B., Manning, C., A., N.: Semantic compositionality through recursive matrix-vector spaces. In: Conference on Empirical Methods in Natural Language Processing 2012 (2012)
22. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research 37(1), 141–188 (2010)
23. Weeds, J., Weir, D., McCarthy, D.: Characterising measures of lexical distributional similarity. In: Proceedings of the 20th international conference on Computational Linguistics. No. 1015, Association for Computational Linguistics (2004)