

Distributional Inclusion Hypothesis for Tensor-based Composition

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh

Queen Mary University of London

School of Electronic Engineering and Computer Science

Mile End Road, London, UK

{d.kartsaklis;mehrnoosh.sadrzadeh}@qmul.ac.uk

Abstract

According to the distributional inclusion hypothesis, entailment between words can be measured via the feature inclusions of their distributional vectors. In recent work, we showed how this hypothesis can be extended from words to phrases and sentences in the setting of compositional distributional semantics. This paper focuses on inclusion properties of tensors; its main contribution is a theoretical and experimental analysis of how feature inclusion works in different concrete models of verb tensors. We present results for relational, Frobenius, projective, and holistic methods and compare them to the simple vector addition, multiplication, min, and max models. The degrees of entailment thus obtained are evaluated via a variety of existing word-based measures, such as Weed’s and Clarke’s, KL-divergence, APinc, balAPinc, and two of our previously proposed metrics at the phrase/sentence level. We perform experiments on three entailment datasets, investigating which version of tensor-based composition achieves the highest performance when combined with the sentence-level measures.

1 Introduction

Distributional hypothesis asserts that words that often occur in the same contexts have similar meanings (Firth, 1957). Naturally these models are used extensively to measure the semantic similarity of words (Turney and Pantel, 2010). Similarity is an a-directional relationship and computational linguists are also interested in measuring degrees of directional relationships between words. Distributional inclusion hypothesis is exactly about such relationships, and particularly, about entailment (Dagan et al., 1999; Geffet and Dagan, 2005; Herbelot and Ganesalingam, 2013). It states that a word u entails a word v if in any context that word u is used so can be word v . For example, in a corpus of sentences ‘a boy runs’, ‘a person runs’, ‘a person sleeps’, according to this hypothesis, $boy \vdash person$, since wherever ‘boy’ is used, so is ‘person’. Formally, we have that u entails v if features of u are included in features of v , where features are non-zero contexts. In this example, $boy \vdash person$, since $\{run\} \subset \{run, sleep\}$.

For the same reasons that the distributional hypothesis is not directly applicable to phrases and sentences, the distributional inclusion hypothesis does not scale up from words to larger language constituents. In a nutshell, this is because the majority of phrases and sentences of language do not frequently occur in corpora of text, thus reliable statistics cannot be collected for them. In distributional models, this problem is addressed with the provision of composition operators, the purpose of which is to combine the statistics of words and obtain statistics for phrases and sentences. In recent work, we have applied the same compositionality principles to lift the entailment relation from word level to phrase/sentence level (Kartsaklis and Sadrzadeh, 2016; Balkır et al., 2016a; Balkır et al., 2016b; Balkır, 2014). The work in (Balkır, 2014; Balkır et al., 2016b) was focused on the use of entropic measures on density matrices and compositional operators thereof, but no experimental results were considered; similarly, Bankova et al. (2016) use a specific form of density matrices to represent words for entailment purposes, focusing only on theory. In (Balkır et al., 2016a), we showed how entropic and other measures can be used on vectors as well as on density matrices and supported this claim with experimental results. In (Kartsaklis and Sadrzadeh, 2016), we focused on making the distributional inclusion hypothesis compositional, worked

out how feature inclusion lifts from words to compositional operators on them, and based on experimental results showed that intersective composition operators result in more reliable entailments. This paper takes a more concrete perspective and focuses on the feature inclusion properties of tensors constructed in different ways and composition operators applied to the tensors and vectors.

One can broadly classify the compositional distributional models to three categories: ones that are based on simple element-wise operations between vectors, such as addition and multiplication (Mitchell and Lapata, 2010); tensor-based models in which relational words such as verbs and adjectives are tensors and matrices contracting and multiplying with noun (and noun-phrase) vectors (Coecke et al., 2010; Grefenstette and Sadrzadeh, 2011; Baroni and Zamparelli, 2010); and models in which the compositional operator is part of a neural network (Socher et al., 2012; Kalchbrenner et al., 2014) and is usually optimized against a specific objective.

Tensor-based models stand in between the two extremes of element-wise vector mixing and neural net-based methods, offering a sufficiently powerful alternative that allows for theoretical reasoning at a level deeper than it is usually possible with black-box statistical approaches. Models of this form have been used in the past with success in a number of NLP tasks, such as head-verb disambiguation (Grefenstette and Sadrzadeh, 2011), term-definition classification (Kartsaklis et al., 2012), and generic sentence similarity (Kartsaklis and Sadrzadeh, 2014). In this paper we extend this work to entailment, by investigating (theoretically and experimentally) the properties of feature inclusion on the phrase and sentence vectors produced in four different tensor-based compositional distributional models: the relational models of (Grefenstette and Sadrzadeh, 2011), the Frobenius models of (Kartsaklis et al., 2012; Milajevs et al., 2014), the projective models of (Kartsaklis and Sadrzadeh, 2016), and the holistic linear regression model of (Baroni and Zamparelli, 2010).

Contrary to formal semantic models, and customary to distributional models, our entailments are non-boolean and come equipped with degrees. We review a number of measures that have been developed for evaluating degrees of entailment at the lexical level, such as the *APinc* measure and its newer version, *balAPinc* of Kotlerman et al. (2010), which is considered as state-of-the-art for word-level entailment. A newly proposed adaptation of these metrics, recently introduced by the authors in (Kartsaklis and Sadrzadeh, 2016), is also detailed. This measure takes into account the specificities introduced by the use of a compositional operator and lifts the measures from words to phrase/sentence level.

We experiment with these models and evaluate them on entailment relations between simple intransitive sentences, verb phrases, and transitive sentences on the datasets of (Kartsaklis and Sadrzadeh, 2016). Our findings suggest that the Frobenius models provide the highest performance, especially when combined with our sentence-level measures. On a more general note, the experimental results of this paper support that of previous work (Kartsaklis and Sadrzadeh, 2016) and strongly indicate that compositional models employing some form of intersective feature selection, i.e. point-wise vector multiplication or tensor-based models with an element of element-wise mixing (such as the Frobenius constructions), are more appropriate for entailment tasks in distributional settings.

2 Compositional distributional semantics

The purpose of a compositional distributional model is to produce a vector representing the meaning of a phrase or a sentence by combining the vectors of its words. In the simplest case, this is done by element-wise operations on the vectors of the words (Mitchell and Lapata, 2010). Specifically, the vector representation \vec{w} of a sequence of words w_1, \dots, w_n is defined to be:

$$\vec{w} := \vec{w}_1 + \vec{w}_2 + \dots + \vec{w}_n \qquad \vec{w} := \vec{w}_1 \odot \vec{w}_2 \odot \dots \odot \vec{w}_n$$

In a more linguistically motivated approach, relational words such as verbs and adjectives are treated as linear or multi-linear maps. These are then applied to the vectors of their arguments by following the rules of the grammar (Coecke et al., 2010; Grefenstette and Sadrzadeh, 2011; Baroni and Zamparelli, 2010). For example, an adjective is a map $N \rightarrow N$, for N a basic noun space of the model. Equivalently, this map can be represented as a matrix living in the space $N \otimes N$. In a similar way, a transitive verb is a map $N \times N \rightarrow S$, 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 following form, where $\overline{\text{verb}}$ is a tensor of order 3 and \times is tensor contraction:

$$\overline{\text{svdo}} = (\overline{\text{verb}} \times \overline{\text{obj}}) \times \overline{\text{subj}}$$

Finally, phrase and sentence vectors have been also produced by the application of neural architectures, such as recursive or recurrent neural networks (Socher et al., 2012; Cheng and Kartsaklis, 2015) and convolutional neural networks (Kalchbrenner et al., 2014). These models have been shown to perform well on large scale entailment tasks such as the ones introduced by the RTE challenge.

3 Distributional inclusion hypothesis

The distributional inclusion hypothesis (DIH) (Dagan et al., 1999; Geffet and Dagan, 2005; Herbelot and Ganesalingam, 2013) 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 limitations, the main one being that it only makes sense in upward monotone contexts. 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 and other limitations, the DIH has been subject to a good amount of study in the distributional semantics community and its predictions have been validated (Geffet and Dagan, 2005; Kotlerman et al., 2010).

Formally, the DIH says that 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. By denoting the features of a distributional vector \vec{v} by $\mathcal{F}(\vec{v})$, one can symbolically express the DIH as follows:

$$u \vdash v \quad \text{whenever} \quad \mathcal{F}(\vec{u}) \subseteq \mathcal{F}(\vec{v}) \tag{1}$$

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 (Chen and Goodman, 1996). KL-divergence is only defined when the support of \vec{v} is included in the support of \vec{u} . To overcome this restriction, a variant referred to by α -skew (Lee, 1999) has been proposed (for $\alpha \in (0, 1]$ a smoothing parameter). *Representativeness* provides another way of smoothing the KL-divergence. The formulae for these are as follows, where abusing the notation we take \vec{u} and \vec{v} to also denote the probability distributions of u and v :

$$D_{\text{KL}}(\vec{v} \parallel \vec{u}) = \sum_i v_i (\ln v_i - \ln u_i) \quad s_\alpha(\vec{u}, \vec{v}) = D_{\text{KL}}(\vec{v} \parallel \alpha \vec{u} + (1 - \alpha) \vec{v})$$

$$R_{\text{D}}(\vec{v} \parallel \vec{u}) = \frac{1}{1 + D_{\text{KL}}(\vec{v} \parallel \vec{u})}$$

Representativeness turns KL-divergence into a number in the unit interval $[0, 1]$. As a result we obtain $0 \leq R_{\text{D}}(\vec{v} \parallel \vec{u}) \leq 1$, with $R_{\text{D}}(\vec{v} \parallel \vec{u}) = 0$ when the support of \vec{v} is not included in the support of \vec{u} and $R_{\text{D}}(\vec{v} \parallel \vec{u}) = 1$, when \vec{u} and \vec{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 Weeds et al. (2004) and Clarke (2009) 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* (Kotlerman et al., 2010), where *LIN* is Lin’s similarity (Lin, 1998) and *APinc* is an asymmetric measure:

$$balAPinc(u, v) = \sqrt{LIN(u, v) \cdot APinc(u, v)} \quad APinc(u, v) = \frac{\sum_r [P(r) \cdot rel'(f_r)]}{|\mathcal{F}(\vec{u})|}$$

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

$$rel'(f) = \begin{cases} 1 - \frac{rank(f, \mathcal{F}(\vec{v}))}{|\mathcal{F}(\vec{v})|+1} & f \in \mathcal{F}(\vec{v}) \\ 0 & o.w. \end{cases} \quad (2)$$

where $rank(f, \mathcal{F}(\vec{v}))$ 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.

4 Measuring feature inclusion at the phrase/sentence level

In recent work, the authors of this paper introduced a variation of the *APinc* and *balAPinc* measures aiming to address the extra complications imposed when evaluating entailment at the phrase/sentence level (Kartsaklis and Sadrzadeh, 2016). The modified measures differ from the original ones in two aspects. 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, see Sections 5 and 6), an element can get very low values without however ever reaching zero. The new measures exploit this blurring of the notion of “featureness” to the limit, by letting $\mathcal{F}(\vec{w})$ to include all the dimensions of \vec{w} .

Secondly, the continuous nature of distributional models is further exploited 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 $\mathcal{F}(\vec{u})$ with rank r and $f_r^{(v)}$ the corresponding feature in $\mathcal{F}(\vec{v})$, we remind that Kotlerman et al. consider that feature inclusion holds at rank r whenever $f_r^{(u)} > 0$ and $f_r^{(v)} > 0$; the new measures strengthen this assumption by requiring that $f_r^{(u)} \leq f_r^{(v)}$. Incorporating these modifications in the *APinc* measure, $P(r)$ and $rel'(f_r)$ are redefined as follows:

$$P(r) = \frac{|\{f_r^{(u)} | f_r^{(u)} \leq f_r^{(v)}, 0 < r \leq |\vec{u}|\}|}{r} \quad rel'(f_r) = \begin{cases} 1 & f_r^{(u)} \leq f_r^{(v)} \\ 0 & o.w. \end{cases} \quad (3)$$

The new relevance function subsumes the old one, as by definition high-valued features in $\mathcal{F}(\vec{u})$ must be even higher in $\mathcal{F}(\vec{v})$. The new *APinc* at the phrase/sentence level thus becomes as follows:

$$SAPinc(u, v) = \frac{\sum_r [P(r) \cdot rel'(f_r)]}{|\vec{u}|} \quad (4)$$

where $|\vec{u}|$ is the number of dimensions of \vec{u} . Further, we notice that *balAPinc* is the geometric average of *APinc* with Lin’s similarity measure, which is symmetric. According to (Kotlerman et al., 2010), the rationale of including a symmetric measure was that *APinc* tends to return unjustifiably 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; more importantly, it introduces a strong element of symmetry in a measure that is intended to be strongly asymmetric. We cope with this issue by replacing Lin’s measure with representativeness on KL-divergence,¹ obtaining the following new version of *balAPinc*:

¹Using other asymmetric measures is also possible; the choice of representativeness on KL-divergence was based on informal experimentation which showed that this combination works better than other options in practice.

$$SBalAPinc(u, v) = \sqrt{R_D(\vec{u} \parallel \vec{v}) \cdot SAPinc(\vec{u}, \vec{v})} \quad (5)$$

Recall that $R_D(p \parallel q)$ is asymmetric, measuring the extent to which q represents (i.e. is similar to) p . So the term $R_D(\vec{u} \parallel \vec{v})$ 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 entailment evaluation.

5 Generic feature inclusion in compositional models

In the presence of a compositional operator, features of phrases or sentences adhere to set-theoretic properties. For simple additive and multiplicative models, the set of features of a phrase/sentence is derived from the set of features of their words using union and intersection. It is slightly less apparent (and for reasons of space we will not give details here) that the features of point-wise minimum and maximum of vectors are also derived from the intersection and union of their features, respectively. That is:

$$\begin{aligned} \mathcal{F}(\vec{v}_1 + \dots + \vec{v}_n) &= \mathcal{F}(\max(\vec{v}_1, \dots, \vec{v}_n)) = \mathcal{F}(\vec{v}_1) \cup \dots \cup \mathcal{F}(\vec{v}_n) \\ \mathcal{F}(\vec{v}_1 \odot \dots \odot \vec{v}_n) &= \mathcal{F}(\min(\vec{v}_1, \dots, \vec{v}_n)) = \mathcal{F}(\vec{v}_1) \cap \dots \cap \mathcal{F}(\vec{v}_n) \end{aligned}$$

As shown in (Kartsaklis and Sadrzadeh, 2016), element-wise composition of this form 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 kind of lifting of the entailment relationship from words to the phrase/sentence level also holds for tensor-based models (Balkır et al., 2016a).

In general, feature inclusion is a more complicated process for tensor-based settings, since in this case the composition operation is matrix multiplication and tensor contraction. As an example, consider the simple case of a matrix multiplication between a $m \times n$ matrix \mathbf{M} and a $n \times 1$ vector \vec{v} . Matrix \mathbf{M} can be seen as a list of column vectors $(\vec{w}_1, \vec{w}_2, \dots, \vec{w}_n)$, where $\vec{w}_i = (w_{1i}, \dots, w_{mi})^\top$. The result of the matrix multiplication is a combination of scalar multiplications of each element v_i of the vector \vec{v} with the corresponding column vector \vec{w}_i of the matrix \mathbf{M} . That is, we have:

$$\begin{pmatrix} w_{11} & \dots & w_{1n} \\ w_{21} & \dots & w_{2n} \\ \vdots & & \vdots \\ w_{m1} & \dots & w_{mn} \end{pmatrix} \times \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = v_1 \vec{w}_1 + v_2 \vec{w}_2 + \dots + v_n \vec{w}_n$$

Looking at $\mathbf{M} \times \vec{v}$ in this way enables us to describe $\mathcal{F}(\mathbf{M} \times \vec{v})$ in terms of the union of $\mathcal{F}(\vec{w}_i)$ ’s where v_i is non zero, that is, we have:

$$\mathcal{F}(\mathbf{M} \times \vec{v}) = \bigcup_{v_i \neq 0} \mathcal{F}(\vec{w}_i)$$

By denoting v_i a feature whenever it is non-zero, we obtain an equivalent form as follows:

$$\bigcup_i \mathcal{F}(\vec{w}_i) \mid_{\mathcal{F}(v_i)} \quad (6)$$

The above notation says that we collect features of each \vec{w}_i vector but only up to “featureness” of v_i , that is up to v_i being non-zero. This 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.

6 Feature inclusion in concrete constructions of tensor-based models

While the previous section provided a generic analysis of the feature inclusion behaviour of tensor-based models, the exact feature inclusion properties of these models depend on the specific concrete constructions, and in principle get a form more refined than that of simple intersective or union-based composition. In this section we investigate a number of tensor-based models with regard to feature inclusion, and derive their properties.

6.1 Relational model

As a starting point we will use the model of Grefenstette and Sadrzadeh (2011), which adopts an extensional approach and builds the tensor of a relational word from the vectors of its arguments. More specifically, the tensors for adjectives, intransitive verbs, and transitive verbs are defined as below, respectively:

$$\overrightarrow{adj} = \sum_i \overrightarrow{Noun}_i \quad \overrightarrow{verb}_{IN} = \sum_i \overrightarrow{Sbj}_i \quad \overrightarrow{verb}_{TR} = \sum_i \overrightarrow{Sbj}_i \otimes \overrightarrow{Obj}_i \quad (7)$$

where \overrightarrow{Noun}_i , \overrightarrow{Sbj}_i and \overrightarrow{Obj}_i refer to the distributional vectors of the nouns, subjects and objects that occurred as arguments for the adjective and the verb across the training corpus. For the case of a subject-verb sentence and a verb-object phrase, composition reduces to element-wise multiplication of the two vectors, and the features of the resulting sentence/phrase vector get the following form (with \vec{s} and \vec{o} to denote the vectors of the subject/verb of the phrase/sentence):

$$\mathcal{F}(\overrightarrow{s\vec{o}}) = \bigcup_i \mathcal{F}(\overrightarrow{Sbj}_i) \cap \mathcal{F}(\vec{s}) \quad \mathcal{F}(\overrightarrow{v\vec{o}}) = \bigcup_i \mathcal{F}(\overrightarrow{Obj}_i) \cap \mathcal{F}(\vec{o}) \quad (8)$$

For a transitive sentence, the model of Grefenstette and Sadrzadeh (2011) returns a matrix, computed in the following way:

$$\overrightarrow{s\vec{v}\vec{o}}_{Rel} = \overrightarrow{verb} \odot (\vec{s} \otimes \vec{o})$$

where \overrightarrow{verb} is defined as in Equation 7. By noticing that $\mathcal{F}(\vec{u} \otimes \vec{v}) = \mathcal{F}(\vec{u}) \times \mathcal{F}(\vec{v})$, with symbol \times to denote in this case the *cartesian product* of the two feature sets, we define the feature set of a transitive sentence as follows:

$$\mathcal{F}(\overrightarrow{s\vec{v}\vec{o}}_{Rel}) = \bigcup_i \mathcal{F}(\overrightarrow{Sbj}_i) \times \mathcal{F}(\overrightarrow{Obj}_i) \cap \mathcal{F}(\vec{s}) \times \mathcal{F}(\vec{o}) \quad (9)$$

Equation 9 shows that the features of this model are pairs (f_{sbj}, f_{obj}) , with f_{sbj} a subject-related feature and f_{obj} an object-related feature, providing a fine-grained representation of the sentence. Throughout this paper, we refer to this model as *relational*.

6.2 Frobenius models

As pointed out in (Kartsaklis et al., 2012), the disadvantage of the relational model 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. A solution presented in (Kartsaklis et al., 2012) suggested the use of *Frobenius* operators in order to expand vectors and matrices into higher order tensors by embedding them into the the diagonals. For example, a vector is embedded into a matrix by putting it in the diagonal of a matrix and padding the off-diagonal elements with zeros. Similarly, one can embed a matrix into a cube by putting it into the main diagonal and pad the rest with zeros. Using this method, for example, one could transform a simple intersective model in tensor form by embedding the context vector of a verb \vec{v} first into a matrix and then into a cube. For a transitive sentence, one could use the matrix defined in Equation 7 and derive a vector for the meaning of the sentence in two ways, each one corresponding to a different embedding of the matrix into a tensor of order 3:

$$\overrightarrow{s\vec{v}\vec{o}}_{CpSbj} = \vec{s} \odot (\overrightarrow{verb} \times \vec{o}) \quad \overrightarrow{s\vec{v}\vec{o}}_{CpObj} = (\vec{s}^T \times \overrightarrow{verb}) \odot \vec{o} \quad (10)$$

We refer to these models as Copy-Subject and Copy-Object, correspondingly. In order to derive their feature inclusion properties, we first examine the form of the sentence vector produced when the verb is composed with a new subject/object pair:

$$\begin{aligned}\vec{s}\vec{v}\vec{o}_{\text{CpSbj}} &= \vec{s} \odot (\overline{verb} \times \vec{o}) = \vec{s} \odot \sum_i \overrightarrow{Sbj}_i \langle \overrightarrow{Obj}_i | \vec{o} \rangle \\ \vec{s}\vec{v}\vec{o}_{\text{CpObj}} &= (\vec{s}^\top \times \overline{verb}) \odot \vec{o} = \vec{o} \odot \sum_i \overrightarrow{Obj}_i \langle \vec{s} | \overrightarrow{Sbj}_i \rangle\end{aligned}$$

We can now define the feature sets of the two models using notation similar to that of Equation 6:

$$\begin{aligned}\mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpSbj}}) &= \mathcal{F}(\vec{s}) \cap \bigcup_i \mathcal{F}(\overrightarrow{Sbj}_i) \Big|_{\mathcal{F}(\langle \overrightarrow{Obj}_i | \vec{o} \rangle)} \\ \mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpObj}}) &= \mathcal{F}(\vec{o}) \cap \bigcup_i \mathcal{F}(\overrightarrow{Obj}_i) \Big|_{\mathcal{F}(\langle \vec{s} | \overrightarrow{Sbj}_i \rangle)}\end{aligned}\tag{11}$$

The symbol $|$ defines a restriction on feature inclusion based on how well the arguments of the sentence fit to the arguments of the verb. For a subject-object pair (Sbj, Obj) that has occurred with the verb in the corpus, this translates to the following:

- *Copy-Subject*: Include the features of Sbj up to similarity of Obj with the sentence object
- *Copy-Object*: Include the features of Obj up to similarity of Sbj with the sentence subject

Note that each of the Frobenius models puts emphasis on a different argument of a sentence; the Copy-Subject model collects features of the subjects that occurred with the verb, while the Copy-Object model collects features from the verb objects. It is reasonable then to further combine the two models in order to get a more complete representation of the sentence meaning, and hence its feature inclusion properties. Below we define the feature sets of two variations, where this combination is achieved via addition (we refer to this model as *Frobenius additive*) and element-wise multiplication (*Frobenius multiplicative*) of the vectors produced by the individual models (Kartsaklis and Sadrzadeh, 2014):

$$\begin{aligned}\mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{FrAdd}}) &= \mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpSbj}}) \cup \mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpObj}}) \\ \mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{FrMul}}) &= \mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpSbj}}) \cap \mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpObj}})\end{aligned}\tag{12}$$

where $\mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpSbj}})$ and $\mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{CpObj}})$ are defined as in Equation 11.

6.3 Projective models

In this section we provide an alternative solution and remedy the problem of having lower dimensions than the required by arguing that the sentence/phrase space should be spanned by the vectors of the arguments of the verb *across* the corpus. Thus we create verb matrices for intransitive sentences and verb phrases by summing up *projectors* of the argument vectors, in the following way:

$$\vec{v}_{itv} := \sum_i \overrightarrow{Sbj}_i \otimes \overrightarrow{Sbj}_i \quad \vec{v}_{vp} := \sum_i \overrightarrow{Obj}_i \otimes \overrightarrow{Obj}_i\tag{13}$$

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:

$$\vec{s}\vec{v}_{\text{Prj}} = \vec{s}^\top \times \vec{v}_{itv} = \sum_i \langle \overrightarrow{Sbj}_i | \vec{s} \rangle \overrightarrow{Sbj}_i \quad \vec{v}\vec{o}_{\text{Prj}} = \vec{v}_{vp} \times \vec{o} = \sum_i \langle \overrightarrow{Obj}_i | \vec{o} \rangle \overrightarrow{Obj}_i\tag{14}$$

Translating the above equations to feature inclusion representations will give:

$$\mathcal{F}(\vec{s}\vec{v}_{\text{Prj}}) = \bigcup_i \mathcal{F}(\vec{S}b_{j_i}) \mid_{\mathcal{F}(\langle \vec{S}b_{j_i} \mid \vec{s} \rangle)} \quad \mathcal{F}(\vec{v}\vec{o}_{\text{Prj}}) = \bigcup_i \mathcal{F}(\vec{O}b_{j_i}) \mid_{\mathcal{F}(\langle \vec{O}b_{j_i} \mid \vec{o} \rangle)} \quad (15)$$

with symbol \mid to define again a restriction on feature inclusion based on the similarity of the arguments with the subject or object of the sentence/phrase. For the subject-verb case, this reads: “include a subject that occurred with the verb, up to its similarity with the subject of the sentence”. 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:

$$\vec{v}_{trv} := \sum_i \vec{S}b_{j_i} \otimes \left(\frac{\vec{S}b_{j_i} + \vec{O}b_{j_i}}{2} \right) \otimes \vec{O}b_{j_i}$$

The meaning of a transitive sentence then is computed as:

$$\vec{s}\vec{v}\vec{o}_{\text{Prj}} = \vec{s}^\top \times \vec{v}_{trv} \times \vec{o} = \sum_i \left[\langle \vec{s} \mid \vec{S}b_{j_i} \rangle \langle \vec{O}b_{j_i} \mid \vec{o} \rangle \left(\frac{\vec{S}b_{j_i} + \vec{O}b_{j_i}}{2} \right) \right] \quad (16)$$

Feature-wise, the above translates to the following:

$$\mathcal{F}(\vec{s}\vec{v}\vec{o}_{\text{Prj}}) = \bigcup_i \left(\mathcal{F}(\vec{S}b_{j_i}) \cup \mathcal{F}(\vec{O}b_{j_i}) \right) \mid_{\mathcal{F}(\langle \vec{s} \mid \vec{S}b_{j_i} \rangle) \mathcal{F}(\langle \vec{O}b_{j_i} \mid \vec{o} \rangle)} \quad (17)$$

Note that in contrast with the relational and Frobenius models, which all employ an element of inter-sective feature selection, the projective models presented in this section are purely union-based.

6.4 Inclusion of verb vectors

The models of the previous sections provide a variety of options for representing the meaning of a verb from its arguments. However, none of these constructions takes into account the distributional vector of the verb itself, which includes valuable information that could further help in entailment tasks. We remedy this problem by embedding the missing information into the existing tensors; for example, we can amend the tensors of the projective model as follows:

$$\vec{v}_{itv} = \sum_i \vec{S}b_{j_i} \otimes (\vec{S}b_{j_i} \odot \vec{v}) \quad \vec{v}_{vp} = \sum_i (\vec{O}b_{j_i} \odot \vec{v}) \otimes \vec{O}b_{j_i} \quad (18)$$

with \vec{v} denoting the distributional vector of the verb. In the context of an intransitive sentence, now we have the following interaction:

$$\vec{s}\vec{v} = \vec{s}^\top \times \vec{v}_{itv} = \vec{s}^\top \times \sum_i \vec{S}b_{j_i} \otimes (\vec{S}b_{j_i} \odot \vec{v}) = \vec{v} \odot \sum_i \langle \vec{s} \mid \vec{S}b_{j_i} \rangle \vec{S}b_{j_i} \quad (19)$$

We see that the result of the standard projective model (Equation 14) is now enhanced with an additional step of inter-sective feature selection. In feature inclusion terms, we get:

$$\mathcal{F}(\vec{s}\vec{v}) = \mathcal{F}(\vec{v}) \cap \mathcal{F}(\vec{s}\vec{v}_{\text{Prj}}) \quad \mathcal{F}(\vec{v}\vec{o}) = \mathcal{F}(\vec{v}) \cap \mathcal{F}(\vec{v}\vec{o}_{\text{Prj}}) \quad (20)$$

It is easy to show that similar formulae hold for the relational and Frobenius models.

7 Experimental setting

We evaluate the feature inclusion behaviour of the tensor-based models of Section 6 in three different tasks; specifically, we measure upward-monotone entailment between (a) intransitive sentences; (b) verb phrases; and (c) transitive sentences. We use the entailment datasets introduced in (Kartsaklis and Sadrzadeh, 2016), which consist of 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 has been created using hypernym-hyponym relationships from WordNet, and it 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. Some examples of positive entailments from all three categories include:²

Subject-verb	Verb-object	Subject-verb-object
evidence suggests ⊢ information expresses survey reveals ⊢ work shows player plays ⊢ contestant compete study demonstrates ⊢ examination shows summer finishes ⊢ season ends	sign contract ⊢ write agreement publish book ⊢ produce publication sing song ⊢ perform music reduce number ⊢ decrease amount promote development ⊢ support event	book presents account ⊢ work shows evidence woman marries man ⊢ female joins male author retains house ⊢ person holds property study demonstrates importance ⊢ work shows value experiment tests hypothesis ⊢ research evaluates proposal

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.

We experimented with a variety of entailment measures, including *SAPinc* and *SBalAPinc* as in (Kartaklis and Sadrzadeh, 2016), their word-level counterparts (Kotlerman et al., 2010), KL-divergence (applied to smoothed vectors as in Chen and Goodman (1996)), α -skew with $\alpha = 0.99$ as in Kotlerman et al. (2010), *WeedsPrec* as in Weeds et al. (2004), and *ClarkeDE* as in Clarke (2009). We use strict feature inclusion as a baseline; in this case, entailment holds only when $\mathcal{F}(\vec{phrase}_1) \subseteq \mathcal{F}(\vec{phrase}_2)$. For compositional operators, we experimented with element-wise vector multiplication and MIN, vector addition and MAX, and the tensor models presented in Section 6. Informal experimentation showed that directly embedding distributional information from verb vectors in the tensors (Section 6.4) works considerably better than the simple versions, so the results we report here are based on this approach. We also present results for a least squares fitting model, which approximates the distributional behaviour of holistic phrase/sentence vectors along the lines of (Baroni and Zamparelli, 2010). Specifically, for each verb, we compute an estimator that predicts the i th element of the resulting vector as follows:

$$\vec{w}_i = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{y}_i$$

Here, the rows of matrix \mathbf{X} are the vectors of the subjects (or objects) that occur with our verb, and \vec{y}_i is a vector containing the i th elements of the holistic phrase vectors across all training instances; the resulting \vec{w}_i 's form the rows of our verb matrix. This model could be only implemented for verb-object and subject-verb phrases due to data sparsity problems. As our focus is on analytic properties of features, we did not experiment with any neural model.

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. 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.

8 Results and discussion

The results are presented in Table 1 (subject-verb and verb-object task) and Table 2 (subject-verb-object task). In all cases, a combination of a Frobenius tensor model with one of the sentence-level measures (*SAPinc*) gives the highest performance. In general, *SAPinc* and *SBalAPinc* work very well with all the tested compositional models, achieving a cross-model performance higher than that of any other metric, for all three tasks. From a feature inclusion perspective, we see that models employing an element of intereseective composition (vector multiplication, MIN, relational and Frobenius tensor models) have consistent high performances across all the tested measures. The reason may be that the intereseective filtering avoids generation of very dense vectors and thus facilitates entailment judgements based on the DIH. On the other hand, union-based compositional models, such as vector addition, MAX, and the projective tensor models, produce dense vectors for even very short sentences, which affects negatively the evaluation of entailment. The non-compositional verb-only baseline was worse than any compositional model other than the least-squares model, which is the only tensor model that did not perform well; this indicates that our algebraic tensor-based constructions are more robust against data sparsity problems than statistical models based on holistic vectors of phrases and sentences.

²The datasets are available at <http://compling.eecs.qmul.ac.uk/resources/>.

Subject-verb task									
Model	Inclusion	KL-div	α Skew	WeedsPrec	ClarkeDE	APinc	balAPinc	SAPinc	SBalAPinc
Verb	0.59	0.59	0.63	0.67	0.57	0.69	0.65	0.65	0.65
\odot	0.54	0.66	0.75	0.75	0.66	0.78	0.72	0.81	0.81
min	0.54	0.68	0.72	0.75	0.63	0.78	0.71	0.74	0.75
+	0.63	0.57	0.74	0.65	0.62	0.72	0.70	0.72	0.72
max	0.63	0.57	0.70	0.65	0.60	0.71	0.65	0.71	0.71
Least-Sqr	0.50	0.59	0.62	0.59	0.56	0.60	0.58	0.63	0.64
\otimes_{proj}	0.59	0.59	0.65	0.67	0.59	0.70	0.67	0.71	0.69
$\otimes_{\text{rel/frob}}$	0.54	0.64	0.77	0.74	0.68	0.78	0.73	0.84	0.83
Verb-object task									
Model	Inclusion	KL-div	α Skew	WeedsPrec	ClarkeDE	APinc	balAPinc	SAPinc	SBalAPinc
Verb	0.58	0.62	0.65	0.67	0.58	0.69	0.66	0.62	0.66
\odot	0.52	0.64	0.74	0.70	0.67	0.75	0.70	0.82	0.79
min	0.52	0.66	0.70	0.71	0.63	0.75	0.69	0.74	0.74
+	0.64	0.61	0.75	0.68	0.63	0.74	0.71	0.72	0.73
max	0.64	0.62	0.73	0.68	0.62	0.72	0.68	0.62	0.66
Least-Sqr	0.50	0.58	0.57	0.56	0.53	0.56	0.55	0.58	0.59
\otimes_{proj}	0.58	0.60	0.66	0.67	0.60	0.70	0.67	0.68	0.68
$\otimes_{\text{rel/frob}}$	0.52	0.63	0.75	0.71	0.67	0.75	0.70	0.82	0.79

Table 1: AUC results for the subject-verb and verb-object tasks. ‘Verb’ refers to a non-compositional baseline, where the vector/tensor of the phrase is taken to be the vector/tensor of the head verb. \odot , + refer to vector element-wise multiplication and addition, respectively, \otimes_{proj} to the projective tensor models of Section 6.3, and $\otimes_{\text{rel/frob}}$ to the construction of Section 6.1. The tensor models (except the least squares one) are further enhanced with information from the distributional vector of the verb, as detailed in Section 6.4.

Model	Inclusion	KL-div	α Skew	WeedsPrec	ClarkeDE	APinc	balAPinc	SAPinc	SBalAPinc
Verb	0.61	0.61	0.66	0.69	0.58	0.74	0.67	0.59	0.63
\odot	0.55	0.65	0.74	0.79	0.67	0.76	0.71	0.80	0.80
min	0.55	0.71	0.74	0.78	0.63	0.77	0.71	0.73	0.76
+	0.58	0.54	0.71	0.59	0.60	0.65	0.64	0.67	0.67
max	0.58	0.55	0.68	0.58	0.58	0.63	0.61	0.60	0.61
Least-Sqr	-	-	-	-	-	-	-	-	-
\otimes_{rel}	0.51	0.64	0.78	0.79	0.69	0.79	0.72	0.84	0.83
\otimes_{proj}	0.64	0.60	0.70	0.69	0.61	0.74	0.70	0.75	0.76
\otimes_{CpSbj}	0.57	0.65	0.73	0.77	0.63	0.73	0.68	0.79	0.78
\otimes_{CpObj}	0.54	0.62	0.73	0.72	0.64	0.76	0.71	0.81	0.79
\otimes_{FrAdd}	0.60	0.60	0.75	0.72	0.67	0.77	0.75	0.84	0.82
\otimes_{FrMul}	0.55	0.62	0.76	0.81	0.68	0.78	0.73	0.86	0.83

Table 2: AUC results for the subject-verb-object task. \otimes_{Rel} refers to the relational tensor model of Section 6.1, while \otimes_{CpSbj} , \otimes_{CpObj} , \otimes_{FrAdd} , and \otimes_{FrMul} to the Frobenius models of Section 6.2. As in the other tasks, the distributional vector of the verb has been taken into account in all tensor-based models except Least-Sqr.

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 with a focus on tensor-based models. Our results showed that intersective composition in general, and the Frobenius tensor models in particular, achieve the best performance when evaluating upward monotone entailment, especially when combined with the sentence-level measures of (Kartsaklis and Sadrzadeh, 2016). Experimenting with different versions of tensor models for entailment is an interesting topic that we plan to pursue further in a future paper. Furthermore, the extension of word-level entailment to phrases and sentences provides connections with natural logic (MacCartney and Manning, 2007), a topic that is worth a separate treatment and constitutes a future direction.

Acknowledgments

The authors gratefully acknowledge support by EPSRC for Career Acceleration Fellowship EP/J002-607/1 and AFOSR International Scientific Collaboration Grant FA9550-14-1-0079.

References

- E. Balkır, D. Kartsaklis, and M. Sadrzadeh. 2016a. Sentence Entailment in Compositional Distributional Semantics. In *Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, Fort Lauderdale, FL, January.
- Esma Balkır, Mehrnoosh Sadrzadeh, and Bob Coecke, 2016b. *Topics in Theoretical Computer Science: The First IFIP WG 1.8 International Conference, TTCS 2015, Tehran, Iran, August 26-28, 2015, Revised Selected Papers*, chapter Distributional Sentence Entailment Using Density Matrices, pages 1–22. Springer International Publishing, Cham.
- Esma Balkır. 2014. Using density matrices in a compositional distributional model of meaning. Master’s thesis, University of Oxford.
- Desislava Bankova, Bob Coecke, Martha Lewis, and Daniel Marsden. 2016. Graded entailment for compositional distributional semantics. *arXiv preprint arXiv:1601.04908*.
- M. Baroni and R. Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL ’96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jianpeng Cheng and Dimitri Kartsaklis. 2015. 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*, pages 1531–1542, Lisbon, Portugal, September. Association for Computational Linguistics.
- Daoud Clarke. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 112–119. Association for Computational Linguistics.
- Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36.
- Ido Dagan, Lillian Lee, and Fernando C. N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Mach. Learn.*, 34(1-3):43–69.
- John R. Firth. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, pages 1–32.
- Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 107–114, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computational Linguistics.
- Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 440–445. Association for Computational Linguistics.
- Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland, June. Association for Computational Linguistics.
- Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of natural language. In B. Coecke, I. Hasuo, and P. Panangaden, editors, *Quantum Physics and Logic 2014 (QPL 2014)*. *EPTSC 172*, pages 249–261.
- Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2016. A Compositional Distributional Inclusion Hypothesis. In *Proceedings of the 9th Conference on Logical Aspects of Computational Linguistics (LACL)*, Lecture Notes in Computer Science. Springer. To appear.
- Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. 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*, pages 549–558.

- Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(04):359–389.
- Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32.
- Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, pages 296–304.
- Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *ACL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics.
- Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 708–719, Doha, Qatar, October. Association for Computational Linguistics.
- Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.
- R. Socher, B. Huval, C. Manning, and Ng. A. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Conference on Empirical Methods in Natural Language Processing 2012*.
- Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.
- Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, number 1015. Association for Computational Linguistics.