# Sentence Entailment in Compositional Distributional Semantics

**Esma Balkır, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh** [*]
{*e.balkir, d.kartsaklis, m.sadrzadeh*}*@qmul.ac.uk*
School of Electronic Engineering and Computer Science
Queen Mary University of London

## Abstract

Distributional semantic models provide vector representations for words by gathering co-occurrence frequencies from corpora of text. Compositional distributional models extend these representations from words to phrases and sentences. In categorical compositional distributional semantics these representations are built in such a manner that meanings of phrases and sentences are functions of their grammatical structure and the meanings of the words therein. These models have been applied to reasoning about phrase and sentence level similarity. In this paper, we argue for and prove that these models can also be used to reason about phrase and sentence level entailment. We provide preliminary experimental results on a toy entailment dataset.

The distributional hypothesis (Firth 1957) provides a model in which meanings of words are represented by vectors. These vectors are built from frequencies of co-occurrences of words within contexts. Compositional distributional models (Mitchell and Lapata 2010) extend these vector representations from words to phrases/sentences. They work alongside the principle of compositionality, employing the fact that the meaning of a string of words is a function of the meanings of the words therein.

The vectorial word and phrase/sentence representations have been applied to similarity-based language tasks such as disambiguation and semantic similarity (Schütze 1998; Turney 2006). In order to apply the distributional representations to entailment tasks, distributional semanticists adhere to a distributional inclusion hypothesis: if word $v$ entails word $w$, then the contexts of word $v$ are included in the contexts of word $w$. This means that whenever word $v$ is used, word $w$ can be used retaining a valid meaning. Whereas there has been an extensive amount of work on this hypothesis at the word level, e.g. see (Dagan, Lee, and Pereira 1999; Weeds, Weir, and McCarthy 2004; Kotlerman et al. 2010), not much has been done when it comes to phrases/sentences. The work on entailment between quantified noun phrases (Baroni et al. 2012) is an exception, but it does not take into account composition. Compositionality is what is needed for a modular approach to the textual entailment challenge (Dagan, Glickman, and Magnini 2006), where entailment is to be decided for complex sentences of language.

Categorical compositional distributional semantics (CCDS) is a compositional distributional model where vectorial meanings of phrases/sentences are built from the vectors of the words therein and grammatical structures of the phrases/sentences (Coecke, Sadrzadeh, and Clark 2010). These models offer a general mathematical setting where the meaning of any phrase/sentence of language can in principle be assigned a vectorial representation. Fragments of them have been instantiated to concrete data and have been applied to word and phrase/sentence similarity-based tasks, outperforming the models where grammar was not taken into account (Grefenstette and Sadrzadeh 2011; Kartsaklis and Sadrzadeh 2013).

In this paper we show how CCDS can be used to reason about entailment in a compositional fashion. In particular, we prove how the general compositional procedure of this model extends the entailment relation from words to strings of words whose grammatical structure is obtained from their syntactic parses. Previous work on word level entailment shows how entropy-based notions such as KL-divergence can be used to formalise the distributional inclusion hypothesis (Dagan, Lee, and Pereira 1999; Herbelot and Ganesalingam 2013). In the current paper we prove that in CCDS this notion soundly extends from word vectors to sentence vectors and provides a notion of sentence entailment similar to that of Natural Logic (MacCartney and Manning 2007).

In the presence of correlations between contexts, the notion of KL-divergence naturally lifts from vectors to density matrices via von Neumann entropy. The results of this paper build on the developments of (Balkır 2014; Balkır, Sadrzadeh, and Coecke 2015); we generalise the instantiations of CCDS from vectors to density matrices and argue that the notion of relative entropy on density matrices gives rise to a richer notion of word and sentence level entailment. Density matrices have been previously used in compositional distributional semantics to represent parsing information (Blacoe, Kashefi, and Lapata 2013) and ambiguity of meaning (Piedeleu et al. 2015).

We conclude by providing a small scale experiment on data obtained from British National Corpus (BNC) applied to a toy short-sentence entailment task. This involves implementing a concrete way of building vectors and density matrices for words and composing them to obtain sentences.

[*] Author order is alphabetical.

## Categorical Compositional Distributional Semantics (CCDS)

These models rely on the theory of compact closed categories. For general definitions of these categories see (Coecke, Sadrzadeh, and Clark 2010). In its most abstract form, a CCDS is denoted by

$$(\mathcal{C}_{\text{Syn}}, \mathcal{C}_{\text{Sem}}, F\colon \mathcal{C}_{\text{Syn}} \to \mathcal{C}_{\text{Sem}})$$

It consists of a compact closed category for syntax, a compact closed category for semantics, a strongly monoidal functor, and a principle of lexical substitution:

$$[\![w_1 w_2 \cdots w_n]\!] := F(\alpha)([\![w_1]\!] \otimes [\![w_2]\!] \otimes \cdots [\![w_n]\!]) \qquad (1)$$

for $w_1 w_2 \cdots w_n$ a string of words, $\alpha$ its grammatical structure, $F(\alpha)$ the translation of $\alpha$ to a distributional setting, and $[\![x]\!]$ the distributional meaning of a word or a string of words.

In practice, these abstract models are instantiated as concrete models. Below we describe the cases of $(\text{PRG}, \text{FVect}_{\mathbb{R}}, F)$ for vectors as distributional representations, and $(\text{PRG}, \mathcal{CPM}(\text{FHilb}_{\mathbb{R}}), F)$ for density matrices.

In favour of a formal description of the underlying mathematical setting, we adhere to the technical language of category theory; the reader less familiar with this language may skip this section.

### Instantiation to $(\text{PRG}, \text{FVect}_{\mathbb{R}}, F)$

On the syntactic side, we work with a pregroup grammatic model of syntax due to Lambek (2001). A pregroup algebra is a compact closed category. It consists of a partially ordered monoid where each element has a left and a right adjoint $\text{PRG} = (P, \leq, \cdot, 1, (-)^r, (-)^l)$. The notion of adjunction here means that for each $p \in P$, we have a $p^r$ and a $p^l$ in $P$ such that:

$$p \cdot p^r \leq 1 \leq p^r \cdot p \qquad p^l \cdot p \leq 1 \leq p \cdot p^l$$

A pregroup grammar is a pregroup algebra $T(B)$ generated over the set of basic grammatical types of a language, e.g. the set $B = \{n, s\}$ for $n$ denoting the type of a noun phrase and $s$ the type of a sentence. It comes equipped with a relation $R \subseteq T(B) \times \Sigma$ that assigns to the vocabulary $\Sigma$ of a language grammatical types from $T(B)$. For example, adjectives of English have type $n \cdot n^l$, intransitive verbs have type $n^r \cdot s$ and transitive verbs have type $n^r \cdot s \cdot n^l$.

In a pregroup algebra, a string of words of the vocabulary $w_1 w_2 \cdots w_n$ has grammatical structure $\alpha$ when for $t_i \in R[w_i]$, there is a morphism $t_1 \cdot t_2 \cdot \cdots \cdot t_n \overset{\alpha}{\to} t$ in the pregroup algebra seen as a compact closed category.

On the semantic side, we work with $\text{FVect}_{\mathbb{R}}$: the compact closed category of finite dimensional vector spaces over reals $\mathbb{R}$ and the linear maps between the spaces. For each vector space $V$, its dual space $V^*$ is its left and right adjoint $V^l = V^r = V^*$. In the presence of a fixed basis we have $V^* \cong V$. This category has the following morphisms for $\otimes$ the tensor product between vector spaces:

$$\epsilon_V \colon V \otimes V \to \mathbb{R} \qquad \eta_V \colon \mathbb{R} \to V \otimes V$$

Given $\sum_{ij} C_{ij} \overrightarrow{v_i} \otimes \overrightarrow{v_j} \in V \otimes V$ and a basis $\{\overrightarrow{v}_i\}_i$ for $V$, these maps are concretely defined as follows:

$$\epsilon_V(\sum_{ij} C_{ij} \overrightarrow{v_i} \otimes \overrightarrow{v_j}) := \sum_{ij} C_{ij} \langle \overrightarrow{v_i} | \overrightarrow{v_j} \rangle \quad \eta(1) := \sum_i \overrightarrow{v_i} \otimes \overrightarrow{v_i}$$

The syntax-semantics map is the functor

$$F\colon \text{PRG} \to \text{FVect}_{\mathbb{R}}$$

given on basic types by $F(n) := N$ and $F(s) = S$ for $N$ and $S$ two vector spaces in $\text{FVect}_{\mathbb{R}}$. This functor is strongly monoidal, resulting in equalities on elements such as

$$F(p \cdot q) = F(p) \otimes F(q) \quad F(1) = \mathbb{R} \quad F(p^r) = F(p^l) = F(p)^*$$

and on morphisms such as

$$F(p \leq q) = F(p) \to F(q)$$

and

$$F(p \cdot p^r \leq 1) = \epsilon_{F(p)} \quad F(1 \leq p^r \cdot p) = \eta_{F(p)}$$

plus two similar ones for the left adjoints.

In this setting, the distributional meanings of words are vectors, hence the principle of lexical substitution instantiates as follows:

$$\overrightarrow{w_1 w_2 \cdots w_n} := F(\alpha)(\overrightarrow{w}_1 \otimes \overrightarrow{w}_2 \otimes \cdots \otimes \overrightarrow{w}_n) \qquad (2)$$

for $\overrightarrow{w_i}$ the vector representation of word $w_i$.

### Instantiation to $(\text{PRG}, \mathcal{CPM}(\text{FHilb}_{\mathbb{R}}), F)$

The syntactic side is as in the previous case. On the semantic side, we work in the (dagger) compact closed category $\mathcal{CPM}(\text{FHilb}_{\mathbb{R}})$ over finite dimensional Hilbert spaces and completely positive maps (Selinger 2007). Here, objects are of the form $V \otimes V^*$, elements of which represent density operators, that is, they are self-adjoint, semi-definite positive, and have trace 1. A completely positive map between two density matrices preserves this structure. Formally, this is a morphism $f\colon V \otimes V^* \to W \otimes W^*$ for which there exists a vector space $X$ and a linear map $g\colon V \to X \otimes W$ such that $f = (g \otimes g) \circ (1_{W \otimes W} \otimes \eta_X)$ in $\text{FHilb}_{\mathbb{R}}$. The $\epsilon$ and $\eta$ maps of this category are obtained by the images of the respective maps in $\text{FHilb}_{\mathbb{R}}$.

The categorical compositional distributional semantics works along the following functor:

$$F\colon \text{PRG} \to \text{FHilb}_{\mathbb{R}} \to \mathcal{CPM}(\text{FHilb}_{\mathbb{R}})$$

Here, the distributional meanings of words are density matrices, hence the principle of lexical substitution instantiates as follows:

$$\widehat{w_1 \cdots w_n} := F(\alpha)(\hat{w}_1 \otimes \cdots \otimes \hat{w}_n) \qquad (3)$$

for $\hat{w}_i$ the density matrix representation of word $w_i$ and $\otimes$ the tensor product in $\mathcal{CPM}(\text{FHilb}_{\mathbb{R}})$.

## KL-Divergence and Relative Entropy

For a vector space $V$ with a chosen orthonormal basis $\{\overrightarrow{v_i}\}_i$, a normalized vector $\overrightarrow{v} = \sum_i p_i \overrightarrow{v_i}$ can be seen as a probability distribution over the basis. In this case one can define a notion of entropy for $\overrightarrow{v}$ as follows:

$$S(\overrightarrow{v}) = -\sum_i p_i \ln p_i$$

which is the same as the entropy of the probability distribution $P = \sum_i p_i$ over the basis.

For two vectors $\overrightarrow{v}, \overrightarrow{w}$ with probability distributions $P$ and $Q$, the distance between their entropies, referred to by Kullback-Leibler divergence, is defined as:

$$KL(\overrightarrow{v}||\overrightarrow{w}) = \sum_j p_j (\ln p_j - \ln q_j)$$

This is a measure of distinguishability. One can define a degree of representativeness based on this measure:

$$R_{KL}(\overrightarrow{v}, \overrightarrow{w}) = \frac{1}{1 + KL(\overrightarrow{v}||\overrightarrow{w})}$$

This is a real number in the unit interval. When there are non zero weights on the basis elements of $\overrightarrow{v}$ that are zero in $\overrightarrow{w}$, then $\ln 0 = \infty$ (by convention $0 \ln 0 = 0$) and so $R_{KL}(\overrightarrow{v}, \overrightarrow{w}) = 0$. So when the support of $P$ is not included in the support of $Q$ then $R_{KL} = 0$, and when $P = Q$ then $R_{KL} = 1$.

Both KL-divergence and representativeness are asymmetric measures. The following measure, referred to by Jensen-Shannon divergence, provides a symmetric version:

$$JS(\overrightarrow{v}, \overrightarrow{w}) = \frac{1}{2}\left[S(P||\frac{P+Q}{2}) \; + \; S(Q||\frac{P+Q}{2})\right]$$

If there are correlations between the basis of $V$, these can be represented by a positive semi-definite symmetric matrix. Suppose we write this matrix in the chosen orthonormal basis as $\hat{v} = \sum_{ij} p_{ij} \overrightarrow{v_i} \otimes \overrightarrow{v_j}$. The diagonal entries of $\hat{v}$ are probabilities over the basis, so we have:

$$\sum_{ii} p_{ii} = 1$$

The non-diagonal entries denote the correlations between the basis. The correlation between $\overrightarrow{v_i}$ and $\overrightarrow{v_j}$ is the same as the correlation between $\overrightarrow{v_j}$ and $\overrightarrow{v_i}$. $\bar{v}$ given in the form above is the matrix form of a density operator in the chosen basis $\{\overrightarrow{v_i}\}_i$.

Density matrices have a notion of entropy called von Neumann entropy, defined as follows:

$$N(\hat{v}) = -\text{Tr}(\hat{v} \ln \hat{v})$$

They also have a notion of KL-divergence:

$$N(\hat{v}||\hat{w}) = \text{Tr } \hat{v}(\ln \hat{v} - \ln \hat{w})$$

The representativeness between two density matrices is defined in a similar way as for vectors. It is a real number in the unit interval, with 0 and 1 values as described before:

$$R_N(\hat{v}, \hat{w}) = \frac{1}{1 + N(P||Q)}$$

The density matrix version of the Jensen-Shannon divergence is obtained by replacing $S$ with $N$.

A vector can be represented as a diagonal density matrix on the chosen basis $\{ovv_i\}_i$. In this case, entropy and von Neumann entropy are the same, since the density matrix has no information on its non-diagonal elements, denoting a zero correlation between the chosen basis.

## Distributional Inclusion Hypothesis for Vectors and Density Matrices

According to the distributional inclusion hypothesis (DIH) if word $v$ entails word $w$ then the set of contexts of $v$ are included in the set of contexts of $w$. This makes sense since it means that whenever word $v$ is used in a context, it can be replaced with word $w$, preserving the meaning. In other words, in such cases, meaning of $w$ subsumes meaning of $v$. For example, ‘cat’ entails ‘animal’, hence in the sentence ‘A cat is drinking milk’, one can replace ‘cat’ with ‘animal’ and the meaning of the sentence stays valid. On the other hand, ‘cat’ does not entail ‘goldfish’, evident from the fact that the sentence ‘A goldfish is drinking milk’ is very unlikely to appear in a real corpus.

Different asymmetric measures on probability distributions have been used to model and empirically evaluate the DIH. Entropy-based measures such as KL-divergence is among successful such measures. Take the orthonormal basis of a distributional space to be the context lemmas of a corpus and this measure becomes zero if there are contexts with zero weights in $\overrightarrow{v}$ that do not have zero weights in $\overrightarrow{w}$. In other words, $R_{KL}(\overrightarrow{v}, \overrightarrow{w}) = 0$ when $v$ does not entail $w$. The contrapositive of this provides a degree of entailment:

$$\overrightarrow{v} \vdash \overrightarrow{w} \quad \Rightarrow \quad R_{KL}(\overrightarrow{v}, \overrightarrow{w}) \neq 0 \tag{4}$$

The $\alpha$-skew divergence of Lee (Lee 1999) and a symmetric version of it based on $JS$ (Dagan, Lee, and Pereira 1999) are variations on the above.

Similarly, for density matrices one can use the degree of representativeness of two density matrices $R_N$ to check for inclusion of contexts.

$$\hat{v} \vdash \hat{w} \quad \Rightarrow \quad R_N(\hat{v}, \hat{w}) \neq 0 \tag{5}$$

Here contexts can be single context lemmas for the diagonal elements where the basis are reflexive pairs $(p_i, p_i)$; contexts can also be pairs of two context lemmas for the non-diagonal elements where the basis are pairs $(p_i, q_j)$ with $p_i \neq q_j$. So not only we are checking inclusion over single contexts, but also over correlated contexts. The following example shows why this notion leads to a richer notion of entailment.

For the sake of simplicity suppose we do not care about the frequencies per se, but whether the bases occurred with the target word at all. So the entries are always either 1 or 0. Consider a distributional space with basis {aquarium, pet, fish} and two target words: ‘cat’ and ‘goldfish’ therein. Assume that we have seen ‘cat’ in the context of ‘fish’, and also independently, in the context of ‘pet’. Assume further that we have seen the word ‘goldfish’ in the context of ‘aquarium’, and also in the contexts of ‘pet’ and ‘fish’, but whenever it was in the context of ‘pet’, ‘fish’ was also around: for

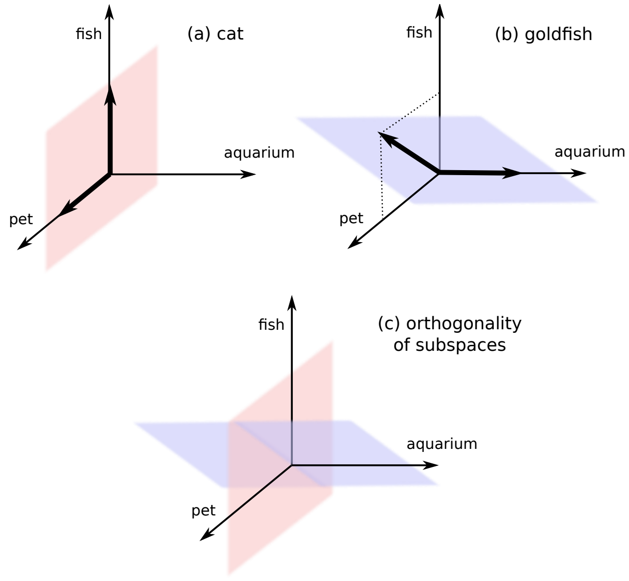

Figure 1: Inclusion of subspaces in the ‘goldfish’ example.

example they always occurred in the same sentence. Hence, we have never seen ‘goldfish’ with ‘pet’ or ‘fish’ separately. This signifies a correlation between ‘pet’ and ‘fish’ for the target word ‘goldfish’.

This correlation is not representable in the vector case and as a result, whereas ‘cat’ does not normally entail ‘goldfish’, its vector representation does, as the set of contexts of ‘cat’ is included in the set of contexts of ‘goldfish’:

| | aquarium | pet | fish |
|---|---|---|---|
| goldfish | 1 | 1 | 1 |
| cat | 0 | 1 | 1 |

By moving to a matrix setting, we are able to represent this correlation and get the correct entailment relation between the two words. In this case, the basis are pairs of the original basis elements. Abbreviating them to their first letters, the matrix representations of ‘cat’ and ‘goldfish’ become:

| goldfish | a | p | f |
|---|---|---|---|
| a | 1 | 0 | 0 |
| p | 0 | 1 | 1 |
| f | 0 | 1 | 1 |

| cat | a | p | f |
|---|---|---|---|
| a | 0 | 0 | 0 |
| p | 0 | 1 | 0 |
| f | 0 | 0 | 1 |

It is not immediately apparent from the matrix representations that the contexts of ‘goldfish’ do not include the contexts of ‘cat’. However to assess the inclusions of contexts, one needs to compare the spans of their eigenvectors with non-zero eigenvalues. For ‘cat’, these eigenvectors are $[(1,0,0),(0,1,0)]$ and for ‘goldfish’ they are $[(1,0,0),(0,1,1)]$. The spans of each are depicted in Figure 1, and one can clearly see that neither one is a subspace of the other.

Without taking correlations of the basis into account, DIH has been strengthened from another perspective and by the realization that contexts should not be all treated equally. Various measures were introduced to weight the contexts based on their *prominence*, for example by taking into account their rank (Weeds, Weir, and McCarthy 2004; Clarke 2009; Kotlerman et al. 2010). From the machine learning side, classifiers have been trained to learn the entailment relation at the word level (Baroni et al. 2012). All of these improvements are applicable to the above density matrix setting.

## Categorical Compositional Distributional Entailment

Montague style semantics of natural language comes equipped with a notation of entailment, where for two sentences $s_1$ and $s_2$, we have “$s_1$ entails $s_2$” iff for $\phi_1$ and $\phi_2$ their logical translations, we have that $\phi_1 \vdash \phi_2$ in a logical system such as first order logic. So entailment becomes the question of derivability in a logic.

In distributional semantics, similar to the case of co-occurrence distributions for phrases/sentences, the inclusion hypothesis does not naturally extend from words to phrases/sentences. One cannot say that a sentence $s_1$ entails the sentence $s_2$ since the contexts of $s_1$ are included in contexts of $s_2$. Similar to the case of co-occurrence counts, it is not clear what the contexts of a sentence are, they cannot be counted directly, they are not the sum or multiplication of the contexts of the words therein. So like for similarity, entailment should be computed compositionally.

In a CCDS, in either of its instantiations to vectors and density matrices, the $F$ functor provides a translation of sentences of natural language to a distributional setting, which comes inherently with a compositional notion of entailment, as defined below:

**Definition.** Categorical compositional distributional entailment (CCDE). *For two strings $v_1 v_2 \cdots v_n$ and $w_1 w_2 \cdots w_n$, and $X$ either $KL$ or $N$, we have $v_1 v_2 \cdots v_n \vdash w_1 w_2 \cdots w_n$ whenever $R_X(\llbracket v_1 \cdots v_n \rrbracket, \llbracket w_1 \cdots w_n \rrbracket) \neq 0$*

We show that this entailment can be made compositional; that is, we derive a phrase/sentence-level entailment from the entailments between the words thereof. What makes this possible is the concept of ‘upward monotonicity’ from Natural Logic (MacCartney and Manning 2007). Roughly put, this expresses that phrases/sentences of an upward monotone vocabulary entail each other.

**Proposition.** *For all $i, 1 \leq i \leq n$ and $v_i, w_i$ upwardly monotone words, we have*

$$v_i \vdash w_i \quad \Rightarrow \quad v_1 v_2 \cdots v_n \vdash w_1 w_2 \cdots w_n$$

**Outline of proof.** First consider the case of density matrices. By Eq. 5 and CCDE, it suffices to show:

$$R_N(\hat{v}_i, \hat{w}_i) \neq 0 \;\Rightarrow\; R_N(\widehat{v_1 \cdots v_n}, \widehat{w_1 \cdots w_n}) \neq 0$$

By definition of positive operators $\hat{R}(\hat{v}_i, \hat{w}_i) \neq 0$ is equivalent to the existence of $r_i \in \mathbb{R}$ and a positive operator $\hat{v'}_i$ such that $\hat{w}_i = r_i \hat{v}_i + \hat{v'_i}$. Assuming these, it suffices to prove that there exists $q \in \mathbb{R}$ and $\hat{\pi'}$ a positive operator such that

$$F(\beta)(\hat{w}_1 \otimes \cdots \otimes \hat{w}_n) = qF(\alpha)(\hat{v}_1 \otimes \cdots \otimes \hat{v}_n) + \hat{\pi'}$$

according to the principle of density lexical substitution (Eq. 3) and for $\alpha$ and $\beta$ grammatical structures of the $w$ and $v$ sentences respectively.

Second, by Lambek's *switching lemma* (Lambek 2001), a sequence of epsilon and identity maps suffice for the representation of the grammatical structure of any sentence in a pregroup grammar. Applying this to $F(\beta)$ makes the left hand side of the above equality equivalent to

$$(\hat{w}_1 \otimes \cdots \otimes \hat{w}_k) \circ \cdots \circ (\hat{w}_l \otimes \cdots \otimes \hat{w}_s)$$

where each $\hat{w}_j$ is either a density matrix $\hat{w}_i$ for $1 \leq i \leq n$ or an identity map over the types of these density matrices.

Finally, if each $\hat{w}_i$ is substituted with its corresponding assumption $r_i\hat{v}_i + \hat{v_i'}$ then by bi-linearity of the $\otimes$ the above will become equivalent to an expression as follows:

$$(r_1 r_2 \cdots r_n)(\hat{v_1} \otimes \cdots \otimes \hat{v_n}) + \Pi$$

where clearly $r_1 r_2 \cdots r_n \in \mathbb{R}$ and it is straightforward to see that $\Pi$ is a sum of expressions of the form $r_1(\hat{v_1} \otimes \cdots \otimes \hat{v'}_k) + r_2(\hat{v_2'} \otimes \cdots \otimes \hat{v_n'}) + \cdots + (\hat{v_1'} \otimes \cdots \hat{v_n'})$ and hence a positive operator. So we have found the $q$ and $\hat{\pi}'$ that we were after.

For the case of vectors, the proof proceeds as above, as vectors are density matrices that only have diagonal elements. □

The above proposition means if $w_1$ represents $v_1$ and $w_2$ represents $v_2$ and so on until $w_n$ and $v_n$, then the string $w_1 w_2 \cdots w_n$ represents the string $v_1 v_2 \cdots v_n$ compositionally, from meanings of phrases/sentences. The degree of representativeness of words – either based on KL-divergence or von Neumann entropy – extends to the degree of representativeness of phrases and sentences.

## Working with Real Data

In this section we present an application of the proposed model in a phrase entailment task based on data collected from a corpus.

**Dataset.** In order to create our dataset we first randomly selected 300 verbs from the most frequent 5000 words in the British National Corpus, and randomly picked either a hyponym or a hyponym from WordNet, provided that it occurred more than 500 times. Next, each entailing verb was paired with one of its common subject or object nouns, while the corresponding entailed verb was paired with an appropriate hypernym of this noun, where

*hyponym* $\vdash$ *hypernym*

This created 300 phrase entailments of the form

*subject*$_1$ *verb*$_1$ $\vdash$ *subject*$_2$ *verb*$_2$

and *verb*$_1$ *object*$_1$ $\vdash$ *verb*$_2$ *object*$_2$.

From these, we selected 23 phrase pairs to reflect a range of entailment degrees.

The degree of entailment between the produced phrases were evaluated by 16 humans, who provided their scores in a scale from 1 (no entailment) to 7 (entailment), following the common practice in the relevant literature—see, for example, (Mitchell and Lapata 2010). Each entailment was scored by the average across all annotators.

**Basic vector space.** The distributional space where the vectors of the words live is a 300-dimensional space produced by non-negative matrix factorization (NMF). The original vectors were 10,000-dimensional vectors weighted by pointwise mutual information (PMI), for which the contexts counts had been collected from a 5-word window around each target word.

**Entailment via KL-divergence in** FVect$_\mathbb{R}$**.** For degrees of entailment obtained via KL-divergence, we work on the instantiation of CCDS to FVect$_\mathbb{R}$. The vector representation of a verb-noun phrase is obtained by applying Equation 1:

$$\overrightarrow{\text{verb noun}} = F(\alpha)(\overline{v} \otimes \overrightarrow{n}) = (1_S \otimes \epsilon_N)(\overline{v} \otimes \overrightarrow{n})$$

This simplifies to the matrix multiplication between the matrix of the verb and the vectors of the noun:

$$\overline{v} \times \overrightarrow{n} \tag{6}$$

The vector of a noun-verb phrase is computed similarly, where in this case $\alpha$ will be $\epsilon_n \otimes 1_s$ and the final matrix multiplication becomes $\overrightarrow{n}^{\mathrm{T}} \times \overline{v}$, for $\overrightarrow{n}^{\mathrm{T}}$ the transpose of the vector of the noun. For details of these computations, we refer the reader to our previous work (Coecke, Sadrzadeh, and Clark 2010; Grefenstette and Sadrzadeh 2011; Kartsaklis, Sadrzadeh, and Pulman 2012), where these have been worked out for a variety of different examples.

Vectors of nouns $\overrightarrow{n}$ are created using the usual distributional method. For producing the verb matrices, we work with a variation of the method suggested in (Grefenstette and Sadrzadeh 2011), referred to by *relational*. We build matrices as follows:

$$\overline{v} = \sum_i \overrightarrow{n}_i \otimes (\overrightarrow{v} \odot \overrightarrow{n}_i)$$

where $\overrightarrow{n_i}$ enumerates the nouns that the verb has modified across the corpus and $\overrightarrow{v}$ is the distributional vector of the verb built in the same way as the nouns. The original relational method computed the matrix of the verb by encoding in it the information about the noun arguments of the verb across the corpus. The above formulation enriches this encoding by also taking into account the context vector of the verb, hence also encoding direct information about the verb itself.

By substituting this in the matrix multiplication of Equation 6 and simplifying it, we obtain the following vector representation for each phrase (verb-noun or noun-verb):

$$\overrightarrow{phrase} = \overrightarrow{v} \odot \sum_i \langle \overrightarrow{n} \mid \overrightarrow{n}_i \rangle \overrightarrow{n}_i$$

Roughly speaking, the above says that the vector meaning of any such phrase represents the contextual properties of the verb of the phrase together with the common contextual properties of the noun of the phrase and the nouns that the verb has modified across the corpus.

**Entailment via relative entropy in** $\mathcal{CPM}$(FHilb$_\mathbb{R}$)**.** In the case of degrees of entailment using relative entropy, we work with the instantiation of CCDS to $\mathcal{CPM}(\text{FHilb}_{\mathbb{R}})$, where Equation 1 results in a density matrix, computed as follows for a verb-noun phrase:

$$\widehat{\text{verb noun}} = F(\alpha)(\hat{v} \otimes \hat{n}) = (1_S \otimes \epsilon_N)(\hat{v} \otimes \hat{n})$$

where $\hat{v}$ and $\hat{n}$ are the density matrices of the verb and the noun, respectively, and $\otimes$ the tensor product in $\mathcal{CPM}(\text{FHilb}_{\mathbb{R}})$. This simplifies to the following formula:

$$\text{Tr}_N(\hat{v} \circ (\hat{n} \otimes 1_S)) \qquad (7)$$

For details, see Piedeleu et al. (2015). The density matrix of a noun-verb phrase is computed similarly by swapping the corresponding identity $1_S$ and epsilon maps $\epsilon_N$ in $\alpha$.

The density matrix for a word $w$, regardless of its grammatical type, is created as:

$$\hat{w} = \sum_i p_i \overrightarrow{c_i} \otimes \overrightarrow{c_i}$$

where $i$ iterates through all contexts of $w$ and $\overrightarrow{c_i}$ is a context vector computed as the average of the vectors of all other words in the same context with $w$. The correlations between the contexts in this case is taken to be the joint probability of their basis elements, which correspond to words annotated with POS tags.

Substituting these in Equation 7 and simplifying it results in the following density matrix representation for each phrase:

$$\widehat{phrase} = \hat{v}^{\text{T}} \times \hat{n} \times \hat{v}$$

Again, the formulation is the same for a verb-noun or a noun-verb phrase. In simple terms, this results in a density matrix that takes into account the contextual properties of the verb, the noun, and the nouns that the verb has modified across the corpus, with the added value that these contextual properties are now enriched: they also reflect the correlations between the contexts.

**Entailment for simple vector composition.** Finally, as a comparison, we also work with degrees of entailment obtained by computing KL-divergence on a simple compositional model achieved via element-wise addition and element-wise multiplication of the vectors of the words in the phrase:

$$\overrightarrow{phrase_+} = \overrightarrow{v} + \overrightarrow{n} \qquad \overrightarrow{phrase_\odot} = \overrightarrow{v} \odot \overrightarrow{n}$$

where $\overrightarrow{v}$ and $\overrightarrow{n}$ denote the distributional vectors of the verb and the noun, respectively.

The experiment proceeds as follows: We firstly produce phrase vectors (or density matrices) by composing the vectors or the density matrices of the individual words in each phrase, and then we compute an entailment value for each pair of phrases; in the case of vectors, this value is given by the representativeness on the KL-divergence between the phrase vectors, while for the density matrix case it is the representativeness on the von Neumann entropy between the density matrices of the phrases. The performance of each model is expressed as the Spearman's correlation of the model predictions with the human judgements.

The results are presented in Table 1, where a non-compositional baseline is also included: we computed $R_{KL}$ for the lexical vectors of the heads of the sentences, that is their verbs. The upper bound is the inter-annotator agreement.

| **Model** | $\rho$ | **Inf** | **F1** | **Acc** |
|---|---|---|---|---|
| Baseline (vector of verb) | 0.24 | 0.37 | 0.57 | 0.74 |
| Categorical | | | | |
| $R_{KL}$ (vectors) | 0.66 | 0.56 | 0.74 | 0.78 |
| $R_N$ (density matrices) | 0.48 | 0.60 | 0.76 | 0.78 |
| Simple | | | | |
| $R^{+}_{KL}$ (e.w. addition) | 0.52 | 0.52 | 0.71 | 0.78 |
| $R^{\odot}_{KL}$ (e.w. multipl.) | 0.41 | 0.32 | 0.64 | 0.61 |
| Upper bound | 0.66 | | | |

Table 1: Results for a phrase entailment experiment.

We also present informedness, F1-score and accuracy for a binarized variation of the task, in which a phrase pair is classified as "entailment" or "non-entailment" depending on whether its average human score was above or below the mean of the annotation range. Note that informedness is an information-theoretic measure that takes into account the true negatives count (something that is not the case for F1-score, for example) and thus it is more appropriate for small and relatively balanced datasets such as ours. The numbers we present for the binary task are based on selecting an appropriate threshold for each model, above of which entailment scores are classified as positive. This threshold was selected in order to optimize informedness.

All the compositional models (for both vectors and density matrices) outperformed the non-compositional baseline. In the correlation task, the categorical vector model $R_{KL}$ was better, achieving a score that matches the inter-annotator agreement; in the classification task, the categorical density matrix model $R_N$ is ahead in every measure.

A snapshot of the results including the highest and lowest pairs according to human judgements are shown in Table 2. We see that although each model returns values in a slightly different range, all of them follow to some extent the general pattern of human annotations. From all three models, the predictions of the model based on element-wise multiplication of vectors are quite marginal. The categorical models and addition of vectors return more balanced results, without avoiding small mistakes.

## Conclusion and Future Directions

We reviewed the categorical compositional distributional model of meaning. This model extends the distributional hypothesis from words to strings of words. We showed how the model can also extend the distributional inclusion hypothesis (DIH) from words to strings. In this case, one is able to derive entailment results over strings of words, from the entailments that hold between their constituent words.

We also reviewed the existing notion of KL-divergence and its application to word level entailment on vector representations of words. We then argued for and showed how

| **Entailment** | **Humans** | **Categorical** | | **Simple** | |
|---|---|---|---|---|---|
| | | $R_{KL}(0.12)$ | $R_N(0.17)$ | $R^{+}_{KL}$ (0.13) | $R^{\odot}_{KL}$ (0.08) |
| arrange task ⊢ organize work | 5.50 (0.785) - T | 0.164 - T | 0.371 - T | 0.192 - T | 0.142 - T |
| recommend development ⊢ suggest improvement | 5.38 (0.768) - T | 0.146 - T | 0.250 - T | 0.182 - T | 0.084 - T |
| advertise notice ⊢ announce sign | 5.38 (0.768) - T | 0.114 - F | 0.187 - T | 0.100 - F | 0.090 - T |
| confirm number ⊢ approve performance | 1.81 (0.258) - F | 0.111 - F | 0.140 - F | 0.087 - F | 0.084 - T |
| recall time ⊢ cancel term | 1.63 (0.232) - F | 0.070 - F | 0.169 - F | 0.126 - F | 0.072 - F |
| editor threathen ⊢ application predict | 1.13 (0.161) - F | 0.082 - F | 0.184 - T | 0.092 - F | 0.080 - F |

Table 2: A snapshot of a phrase entailment experiment. The human judgements are between 1 and 7, with their values normalised between 0 and 1 in brackets. The model predictions are between 0 and 1. T and F indicate classification of each phrase pair as entailment or non-entailment according to each model. The numbers that appear in brackets at the headers are the classification thresholds optimizing informedness for the various models.

moving from vectors to density matrices strengthens the DIH. Finally, we presented a preliminary toy experiment on an entailment task for short subject-verb sentences and verb-object phrases and compared the correlation between the degrees of entailment as predicted by the model and as judged by humans. In this task both vector-based and categorical compositions performed above the baseline.

On the theoretical side, we proved that strings of words whose grammatical structures are their syntactic parses admit a compositional notion of entailment. Extending this result to strings with meaning postulates, such as Frobenius algebras for relative pronouns, constitutes future work. Regarding the experimental side, the theoretical argument of the paper favours categorical composition over simple element-wise operators between vectors, and the results presented here are supportive to this. The density matrices formulation, which in theory is the most powerful, worked better on the classification task; furthermore, informal experimentation showed that a non-compositional model based solely on the relative entropy of the density matrices of the verbs scores a $\rho$ much higher than the corresponding vector-based baseline, providing additional evidence about the richness of the proposed representation. A large scale experiment to verify the predictions properly is under way and constitutes work in progress.

KL-divergence and relative entropy give rise to an ordering on vectors and density matrices respectively. They represent the difference in the information contents of the underlying vectors and density matrices. Exploring this order and the notion of logic that may arise from it is work in progress. The work of Widdows (2004) and Preller (2011) might be relevant to this task.

## Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments. Sadrzadeh is supported by EPSRC CAF grant EP/J002607/1, Kartsaklis by AFOSR grant FA9550-14-1-0079 and Balkır by a Queen Mary Vice Principal scholarship.

## References

Balkır, E.; Sadrzadeh, M.; and Coecke, B. 2015. Distributional sentence entailment using density matrices. *CoRR* abs/1506.06534.

Balkır, E. 2014. Using density matrices in a compositional distributional model of meaning. Master's thesis, University of Oxford.

Baroni, M.; Bernardi, R.; Do, N.-Q.; and Shan, C.-c. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 23–32. Association for Computational Linguistics.

Blacoe, W.; Kashefi, E.; and Lapata, M. 2013. A quantum-theoretic approach to distributional semantics. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 847–857.

Clarke, D. 2009. Context-theoretic semantics for natural language: an overview. In *Proceedings of the workshop on geometrical models of natural language semantics*, 112–119. Association for Computational Linguistics.

Coecke, B.; Sadrzadeh, M.; and Clark, S. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis* 36.

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. Springer. 177–190.

Dagan, I.; Lee, L.; and Pereira, F. C. N. 1999. Similarity-based models of word cooccurrence probabilities. *Mach. Learn.* 34(1-3):43–69.

Firth, J. R. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis* 1–32.

Grefenstette, E., and Sadrzadeh, M. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1394–1404. Association for Computational Linguistics.

Herbelot, A., and Ganesalingam, M. 2013. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, 440–445. Association for Computational Linguistics.

Kartsaklis, D., and Sadrzadeh, M. 2013. Prior disambiguation of word tensors for constructing sentence vectors. In

*Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNL)*, 1590–1601.

Kartsaklis, D.; Sadrzadeh, M.; and Pulman, S. 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*, 549–558.

Kotlerman, L.; Dagan, I.; Szpektor, I.; and Zhitomirsky-Geffet, M. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering* 16(04):359–389.

Lambek, J. 2001. Type grammars as pregroups. *Grammars* 4(1):21–39.

Lee, L. 1999. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 25–32.

MacCartney, B., and Manning, C. D. 2007. Natural logic for textual inference. In *ACL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics.

Mitchell, J., and Lapata, M. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8):1388–1439.

Piedeleu, R.; Kartsaklis, D.; Coecke, B.; and Sadrzadeh, M. 2015. Open system categorical quantum semantics in natural language processing. In *Proceedings of the 6th Conference on Algebra and Coalgebra in Computer Science*.

Preller, A. 2011. From Sentence to Concept, a Linguistic Quantum Logic. Technical Report RR-11019, LIRMM.

Schütze, H. 1998. Automatic Word Sense Discrimination. *Computational Linguistics* 24:97–123.

Selinger, P. 2007. Dagger compact closed categories and completely positive maps. *Electronic Notes in Theoretical Computer Science* 170:139–163.

Turney, P. D. 2006. Similarity of semantic relations. *Computational Linguistics* 32(3):379–416.

Weeds, J.; Weir, D.; and McCarthy, D. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, number 1015. Association for Computational Linguistics.

Widdows, D. 2004. *Geometry and Meaning*. Center for the Study of Language and Information/SRI.