# An Exploration of Discourse-Based Sentence Spaces for Compositional Distributional Semantics

**Tamara Polajnar, Laura Rimell, and Stephen Clark**
Computer Laboratory
University of Cambridge
Cambridge, UK
{tamara.polajnar,laura.rimell,stephen.clark}@cl.cam.ac.uk

## Abstract

This paper investigates whether the wider context in which a sentence is located can contribute to a distributional representation of sentence meaning. We compare a vector space for sentences in which the features are words occurring within the sentence, with two new vector spaces that only make use of surrounding context. Experiments on simple subject-verb-object similarity tasks show that all sentence spaces produce results that are comparable with previous work. However, qualitative analysis and user experiments indicate that extra-sentential contexts capture more diverse, yet topically coherent information.

## 1 Introduction

Distributional word representations (Turney and Pantel, 2010) have proven useful for a wide variety of tasks, including lexical similarity, sentiment analysis, and machine translation. By far the most typical method of building distributional word vectors is based on co-occurrences in a small context window around the word. In contrast, there has been little investigation of different distributional representations for sentences, though the current hypothesis is that the wider discourse in which the sentence is situated may provide relevant information (Baroni et al., 2014; Clark, 2013, 2015). If word representations could be composed into sentence vectors that reflect typical discourse contexts, this might be of great use in sentence-level tasks such as sentence similarity, automatic summarisation, and textual entailment.

Previous work in compositional distributional semantics largely defines the sentence vector space to be the same as the noun space (Kartsaklis et al., 2012; Socher et al., 2011b, 2012), and produces sentence vectors in that space by a sequence of operations on word representations. However, embedding a sentence into a vector space whose dimensions are based on lexical semantics may fail to capture important aspects of sentential meaning. We believe there are two reasons behind the rather surprising lack of attention to sentence spaces. The first is doubt as to whether the distributional hypothesis applies to sentences, i.e. whether sentence meaning is contextual. The second is a question of data sparsity in obtaining contextual sentence representations.

In this paper we explore the idea that contextual sentence representations are viable, and that the surrounding discourse, in the form of adjacent sentences, provides useful information for modelling sentence meaning. We introduce two sentence spaces based on extra-sentential context, one consisting of a variety of context words and the other only of the surrounding verbs, and compare them with an intra-sentential contextual sentence space similar to that proposed in Grefenstette et al. (2013).

We situate our work within the Categorial framework (Coecke et al., 2010; Baroni et al., 2014; Clark, 2013, 2015) where nouns and sentences are considered atomic types, represented as vectors, and other words as functions, represented as tensors. This framework provides a natural setting in which the sentence space can differ from the spaces of sentence constituents, since argument-taking words such as verbs are maps from argument space into sentence space. Following Grefenstette and Sadrzadeh (2011a,b) and Kartsaklis et al. (2012) we focus on simplified sentences consisting of a subject, transitive verb, and object (SVO). We train transitive verb tensors using a single-step multilinear regression algorithm.

We evaluate our composed representations on two standard SVO sentence similarity tasks. The results show that the discourse-based sentence spaces perform competitively, both with the intra-



*Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 1–11,
Lisboa, Portugal, 18 September 2015. 

sentential contextual space and with previous work on SVO composition, although not beating the state of the art on these tasks. We then provide a qualitative analysis of the topics resulting from Singular Value Decomposition in each sentence space, showing that both intra- and extra-sentential spaces contain highly coherent topics, but that the extra-sentential spaces are able to group together SVO triples with greater lexical diversity. We evaluate topic coherence with a novel SVO triple intrusion task.

## 2 Background and Related Work

The majority of previous work producing vector representations for sentences uses the same space for sentences as for words. Within the Categorial framework, several previous experiments (Grefenstette and Sadrzadeh, 2011a,b; Kartsaklis et al., 2012; Kartsaklis and Sadrzadeh, 2014) have defined the sentence space to be the same as the noun space. The noun space is based on co-occurrences with frequent words in the corpus in a small window, which may not be the ideal space to represent sentences, which have distinct semantics involving propositional meaning and links to surrounding discourse (see Section 2.1 for more detail).

Neural language modelling approaches such as Socher et al. (2011a, 2013) recursively build sentence representations from constituent word vectors, which themselves are embeddings based on local context, such that the phrase space after each composition step remains the same, including the space for sentences at the root of a derivation. In these models the features are less interpretable, but since the original word embeddings are based on local co-occurrences, sentences are effectively being represented in a lexical semantic space.

Grefenstette et al. (2013) use a dedicated sentence space for SVO sentences, in which the features are intra-sentential co-occurrences of VO pairs and SVO triples with the 10,000 most frequent words in the corpus. They learn tensors for transitive verbs by multi-stage linear regression, incorporating objects and subjects in two separate steps. Fried et al. (2015) also use an intra-sentential sentence space when learning low-rank approximations for verb tensors (see Section 2.1). We experiment with a similar intra-sentential space alongside our extra-sentential spaces. Another intra-sentential sentence space is described in Le and Mikolov (2014), who learn embeddings for larger text segments, including sentences, based on n-grams internal to the text segments. A variant of this approach was also adopted by Fried et al. (2015) to learn verb tensors mapping to an intra-sentential sentence space for the Categorial framework using single-step linear regression.

Sentence spaces need not be contextual, but may also represent other aspects of meaning relevant to propositions, such as plausibility or feature norms (McRae et al., 1997). A non-contextual option that has been previously implemented is a two-dimensional "plausibility space", in which the sentence vector represents a plausibility judgement. This type of space was explored in theory in Clark (2013, 2015) and implemented with multilinear regression training for verb tensors by Polajnar et al. (2014a).

Contemporaneously with our work, Kiros et al. (2015) have used an encoder-decoder recurrent neural network architecture to encode a sentence vector conditioned on the previous and following sentences, providing further support for the utility of extra-sentential context for sentence meaning.

### 2.1 Categorial Framework Background

In the Categorial framework, nouns are represented as vectors, while argument-taking words such as verbs and adjectives are represented as functions. Specifically, they are tensors that perform multilinear transformations of lower-dimensional tensors, e.g. noun vectors. The Categorial Grammar derivation of a sentence guides the combination of vector and tensor objects representing the words in the sentence to ultimately produce a single sentence vector.

For example, a transitive verb in Combinatory Categorial Grammar (CCG) has the syntactic type $(S\backslash NP)/NP$, which defines it as a function that takes a noun phrase as an input from the right, and then another noun phrase from the left, to produce a sentence. Interpreting such categories under the Categorial framework is straightforward. First, for each atomic category there is a corresponding vector space; in this case the sentence space $\mathbf{S}$ and the noun space $\mathbf{N}$.[1] Hence the meaning of a noun or noun phrase, for example *people*, will be a vector in the noun space: $\overrightarrow{people} \in \mathbf{N}$. In order to obtain the meaning of a transitive verb, each slash is re-

[1] In practice, for example using the CCG parser of Clark and Curran (2007), there will be additional atomic categories, such as $PP$, but not many more.

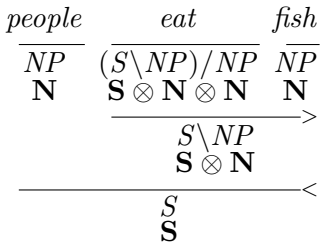


Figure 1: Syntactic reduction and tensor-based semantic types for a transitive verb sentence.

placed with a tensor product operator, so that the meaning of *eat*, for example, is a 3rd-order tensor: $\overline{eat} \in \mathbf{S} \otimes \mathbf{N} \otimes \mathbf{N}$. Just as in the syntactic case, the meaning of a transitive verb is a function (a multi-linear map) which takes two noun vectors as arguments and returns a sentence vector.

Meanings combine using *tensor contraction*, which can be thought of as a multi-linear generalisation of matrix multiplication (Grefenstette, 2013). Consider first the adjective-noun case, for example *black cat*. The syntactic type of *black* is $N/N$; hence its meaning is a 2nd-order tensor (matrix): $\overline{black} \in \mathbf{N} \otimes \mathbf{N}$. In the syntax, $N/N$ combines with $N$ using the rule of forward application ($N/N \ N \Rightarrow N$), which is an instance of function application. Function application is also used in the tensor-based semantics, which, for a matrix and vector argument, corresponds to matrix multiplication.

Figure 1 shows how the syntactic types combine with a transitive verb, and the corresponding tensor-based semantic types. Note that, after the verb has combined with its object $NP$, the type of the verb phrase is $S\backslash NP$, with a corresponding meaning tensor (matrix) in $S \otimes N$. This matrix then combines with the subject vector, through matrix multiplication, to give a sentence vector.

Some previous work in the Categorial framework has taken the sentence space to be the same as the noun space (Grefenstette and Sadrzadeh, 2011a,b; Kartsaklis et al., 2012; Kartsaklis and Sadrzadeh, 2014). The verb is defined, not as an $\mathbf{S} \otimes \mathbf{N} \otimes \mathbf{N}$ tensor, but an $\mathbf{N} \otimes \mathbf{N}$ matrix summing the outer products of its observed subjects and objects. Because this results in a type mismatch when presented with two noun vector arguments, tensor contraction cannot be used directly to produce a sentence vector. Instead, various combinations of matrix multiplication, pointwise multiplication, and addition are employed. As a result, the sentence representation is a purely compositional function of the context vectors of its component words; the observed contexts of SVO triples are not part of the representation.

One effect of reducing a verb tensor to a matrix is to reduce the number of parameters required to learn the verb. However, recent work in the Categorial framework offers other ways to reduce the number of parameters while retaining the higher type of the tensor. Fried et al. (2015) introduce low-rank approximations for verb tensors, which provide a large reduction in the number of parameters while increasing the speed of training, without substantial loss in accuracy on standard SVO tasks.

## 3 Sentence Spaces

We focus on SVO triples, which we also refer to as transitive sentences or simply sentences. Although real-world sentences are more complex, SVO is currently the standard grammatical construction for sentence composition within the Categorial framework, because it is manageable for current learning methods.

This section describes our three contextual sentence spaces. The first follows Grefenstette et al. (2013) and Fried et al. (2015) in using intra-sentential word co-occurrenes with SVO triples. The others use extra-sentential co-occurrences. We consider the extra-sentential spaces to be a primitive way of incorporating the surrounding discourse into distributional representations. We know that individual sentences are linked to their neighbours in a coherent discourse, and investigate whether that linkage can be leveraged for natural language understanding. We make the assumption that Wikipedia articles, the source of our vectors, are a good source of coherent sequences of sentences.

### 3.1 Intra-Sentential Context

Following previous work, our first sentence space is the intra-sentential context of the SVO triple. We call this the Internal Distributional (IDist) space. We first select the top $N = 10,000$ most frequent words from the corpus (excluding stopwords) as contexts. Any of these words appearing inside the same sentence as the SVO triple are counted as features for that triple. Figure 2 shows an example of IDist.

When the V, S, or O itself is frequent enough

| |
|---|
| $\mathbf{S}_{t-2}$**:** M. Atget captured the old Paris in his pictures. $\mathbf{S}_{t-1}$**:** His photographs show the city in its various facets. $\mathbf{S}_t$**:** He photographed stairwells and architectural details. $\mathbf{S}_{t+1}$**:** His interests also extended to the environs of Paris. $\mathbf{S}_{t+2}$**:** He also photographed street-hawkers and small tradesmen, as well as popular amusements. |
| **IDist**: stairwell, architectural, detail |
| **DDist**: capture, old, paris, picture, photograph, show, city, various, interest, extend, popular, amusement |
| **DVerb**: capture, show, extend, photograph |

Figure 2: Example features in sentence spaces for a target sentence $\mathbf{S}_t$.

to be one of the context words, we had to decide whether to retain or discard it as context for the triple. We chose to discard the verb, because it is the verb tensor itself that is being learned. On the other hand, if either the S or O is one of the context words, we retained it as context for the triple. The reasoning is related to a somewhat strange aspect of using intra-sentential context for a sentence: as composition methods become more sophisticated, and more of the sentence is included in the composition, there would eventually be no intra-sentential context left to use if all composed words were removed.

### 3.2 Extra-Sentential Contexts

Our other sentence spaces use the surrounding discourse as context for a sentence. There are many ways one could create a discourse context for an SVO triple, with the size of the context ranging from the surrounding sentences to the full document, and the context features ranging from the same words as in IDist, to specific parts of speech, phrase types, or discourse markers deemed more representative of sentence meaning.

We define two extra-sentential sentence spaces. Both use a window of two sentences on either side of the target sentence $\mathbf{S}_t$. The first space is the Discourse Distributional (DDist) space, which takes as context features any of the top 10,000 words from the corpus occurring in the two sentences either side of $\mathbf{S}_t$ (but not in $\mathbf{S}_t$ itself). This sentence space is analogous to IDist, but using an extra- rather than intra-sentential window.

The second space is the Discourse Verb (DVerb) space, which takes as context features any verbs occurring in the two sentences either side of $\mathbf{S}_t$. This space was loosely inspired by work on unsupervised learning of narrative event chains (Chambers and Jurafsky, 2008, 2009), in which sequences of events such as *accuse – claim – argue – dismiss* or *appoint – work – oversee – retire* are extracted from text. That work links event types which share a protagonist in a connected discourse; in contrast, we do not check whether neighbouring verbs share arguments, but simply hypothesise that verbs near the target verb represent related events and are therefore particularly suited to be context features. Figure 2 shows examples of DVerb and DDist.

We expect DVerb to suffer from a certain amount of data sparsity since the number of verbs in a window of two sentences on either side of the target can be expected to be low, despite the fact that we do not restrict the context features to the main verbs of those sentences. DVerb is therefore the most speculative of our sentence spaces.

### 3.3 Combined Spaces

In order to examine the interaction between the intra- and extra-sentential contexts, we also create two combined spaces: ID.DD, a concatenation of IDist and DDist, and ID.DV, a concatenation of IDist and DVerb. To create the combined spaces, we use the vector spaces as defined above which are created separately and reduced to 20-dimensions (Section 4). Then for each triple we concatenate the vector from each of the spaces we are combining to create a 40-dimensional vector.

## 4 Training

To train the noun vectors and verb tensors we used an October 2013 download of Wikipedia articles, which was tokenised using the Stanford NLP tools,[2] lemmatised with the Morpha lemmatiser (Minnen et al., 2001), and parsed with the C&C parser (Clark and Curran, 2007).

We selected a total of 345 verbs, which include the verbs in our test datasets, along with some additional high-frequency verbs included to produce more representative sentence spaces. To train the verbs, we required high-quality SVO triples that occurred enough times in the corpus to provide us with distributional representations of their contexts. For each verb we therefore selected up to

[2]http://nlp.stanford.edu/software/index.shtml

600 triples which occurred more than once and contained subject and object nouns that occurred at least 100 times. This resulted in $M \approx 150,000$ triples overall.

We first generated distributional vectors for all the nouns contained in the training triples and the test datasets. We used Wikipedia as the source corpus, with sentences as the context window and the top $N = 10,000$ most frequent words (excluding stopwords) as the context words. Following the procedure outlined in Polajnar and Clark (2014), we employed t-test weighting (Curran, 2004) and context selection, and reduced our noun vectors ($\mathbf{n}$) to $K = 100$ dimensions using Singular Value Decomposition (SVD).

For each verb $V$ we have a set of $M_V$ training instances, where each instance $i \in M_V$ consists of subject and object noun vectors $\mathbf{n}^{(s)}{}_i$, $\mathbf{n}^{(o)}{}_i$ and a true sentence space representation vector $\mathbf{t}_i$. The vector $\mathbf{t}_i$ is the SVD-reduced version of the Wikipedia context vector for the triple $\mathbf{n}^{(s)}{}_i \; \mathbf{V} \; \mathbf{n}^{(o)}{}_i$.

The true IDist and DDist vectors were generated using the same $N = 10,000$ context words as for the nouns, weighted by t-test. The entire $M \times N$ matrix was reduced to $S = 20$ or $S = 40$ dimensions.[3]

The DVerb context words consist of $N = 2,641$ verbs that occurred at least 10 times within the two sentences surrounding our triples. DVerb was also weighted using t-test and the matrix encoding the co-occurrence of triples with verb contexts was reduced with SVD to produce an $M \times S$ matrix.

**Regression (reg)** We learn the values of the $S \times K \times K$ tensor representing the verb as parameters ($\mathbf{V}$) of a regression algorithm. To train the tensor we minimise the sum of the mean squared errors ($MSQE$) between each of the training sentence space vectors $\mathbf{t}_i$ and classifier predictions $\mathbf{s}_i$ using the following regularised objective:

$$O(\mathbf{V}) = -\frac{1}{M_V}\left[\sum_{i=1}^{M_v} MSQE(\mathbf{t}_i, \mathbf{s}_i) + \frac{\lambda}{2}||\mathbf{V}||\right]$$

where the $l$-th index of the predicted sentence vector is produced by tensor contraction

$$s_l = \sum_{jk} V_{ljk} n_j^{(s)} n_k^{(o)} \quad (1)$$

between the tensor and the subject and object noun vectors $\mathbf{n}^{(s)}$ and $\mathbf{n}^{(o)}$. The training was performed through gradient descent with ADADELTA (Zeiler, 2012), with minibatches, and with 10% of the training triples reserved as a validation set for early stopping. The regularisation parameter was set to $\lambda = 0.05$ without tuning.

**Distributional Tensors (dist)** As an alternative to learning the verb function, we produce a verb tensor using a procedure inspired by Grefenstette and Sadrzadeh (2011a). The intuition behind this method is that the tensor should encode higher values for topics that frequently co-occur within the subject, object, and sentence vectors in the triples used to train a particular verb. Specifically, we generate an $S \times K \times K$ tensor $\mathbf{V}$ for each verb as the average of the tensor products ($\otimes$) of $K$-dimensional subject and object vectors and the $S$-dimensional sentence space vector ($\mathbf{s}$) from the training triples:

$$\mathbf{V} = \frac{1}{M_V}\left[\sum_{i=1}^{M_V} \mathbf{s}_i \otimes \mathbf{n}^{(s)}{}_i \otimes \mathbf{n}^{(o)}{}_i\right]$$

where $M_V$ is the number of training triples for the verb $V$. Our procedure differs from Grefenstette and Sadrzadeh (2011a) because it generates a tensor, while they treated verbs as matrices and effectively disregarded the sentence space.

## 5 Quantitative Experiments

We perform two experiments using composed sentence vectors. The first involves disambiguation of a polysemous verb in the context of its subject and object, and the second involves measurement of sentence similarity, without disambiguation. We make use of two existing SVO datasets.

### 5.1 Datasets

**GS11** The first dataset is from Grefenstette and Sadrzadeh (2011a) (GS11), and consists of 200 sentence pairs (400 sentences total). Each sentence pair shares a subject and an object. The first member of the pair has an ambiguous verb, while the second has a 'landmark' disambiguating verb. Gold standard annotation provides similarity ratings for each pair on a scale of 1 (low) to 7 (high). For example, *people try door* and *people test door* have high similarity ratings, while *people try door* and *people judge door* have low ratings.

[3] We examined other configurations of noun and sentence space dimensions. Larger tensors learned by regression or distributionally did not consistently lead to increased scores. Although the dimensionality of the sentence space is small, the $K \times K \times S$ tensors are sufficiently large that we believe they are already capturing a significant amount of information from the interaction of the noun and sentence spaces.

| GS11 | Distributional | | Regression | |
|---|---|---|---|---|
| | S=20 | S=40 | S=20 | S=40 |
| **IDist** | 0.18 | 0.15 | 0.31 | **0.33** |
| **DDist** | 0.18 | 0.20 | 0.26 | 0.27 |
| **DVerb** | **0.21** | **0.21** | **0.32** | 0.32 |
| **ID.DV** | - | **0.22** | - | **0.33** |
| **ID.DD** | - | 0.19 | - | 0.29 |

| KS14 | Distributional | | Regression | |
|---|---|---|---|---|
| | S=20 | S=40 | S=20 | S=40 |
| **IDist** | 0.15 | 0.07 | **0.42** | **0.43** |
| **DDist** | 0.17 | **0.17** | 0.34 | 0.37 |
| **DVerb** | **0.18** | 0.14 | 0.33 | 0.37 |
| **ID.DV** | - | **0.22** | - | **0.40** |
| **ID.DD** | - | 0.16 | - | 0.38 |

Table 1: Spearman-$\rho$ results for the GS11 dataset (left) and KS14 dataset (right).

**KS14** The second dataset (Kartsaklis and Sadrzadeh, 2014) (KS14), consists of 72 sentences arranged into 108 sentence pairs. The sentences in each pair do not share verbs, subjects, or objects. Gold standard annotation provides similarity ratings for each pair on a scale of 1 (low) to 7 (high). For example, *medication achieve result* and *drug produce effect* have high similarity ratings, while *author write book* and *delegate buy land* have low ratings. Sentence pairs with mid-similarity ratings tend to have high relatedness but are not mutually substitutable, e.g. *team win match* and *people play game*.

Both tasks are formulated as ranking tasks. Each SVO triple is composed as in Equation 1 and the resulting vectors are compared using cosine to give a similarity value. Sentence pairs are ordered according to similarity and Spearman's $\rho$ is used to compare the automatically-obtained similarity ranking with that obtained from the gold standard judgements.

### 5.2 Results

Table 1 shows the results for the two tasks. Each task is evaluated with both the distributionally-built tensors and regression trained tensors and with 20 and 40 dimensional sentence spaces. This led to eight separate experiments. Overall, different conditions favour different sentence spaces. DVerb achieves the highest or near-highest score for all the GS11 experiments, which is interesting given that DVerb is the sparsest sentence space of the three. Although it is well known that the arguments of an ambiguous verb are important for disambiguation, this result suggests that extra-sentential verb co-occurrences may also reflect different verb senses. On the other hand, IDist achieves some of the highest overall scores with regression training, on both GS11 and KS14. DDist and DVerb lag somewhat behind IDist on KS14. We also note that the results on all experiments are higher with regression-trained tensors.

In the combined space experiments, we find that IDist and DVerb provide mutually complementary information and high scores that are close to or outperform single space models.

To put these results in context, our regression results for IDist, DVerb, and ID.DV are comparable with the highest distributional results in Fried et al. (2015) ($\rho = 0.34$ on GS11 and $\rho = 0.42$ on KS14), which were obtained with a sentence-internal space with 100-dimensional vectors, much higher dimensionality than ours. Kartsaklis and Sadrzadeh (2014) obtain $\rho = 0.42$ on GS11, using a distributional matrix with a composition method which effectively disregards the sentence space, and a 300-dimensional noun space. The state-of-the art for KS14 is $\rho = 0.58$ with vector addition and 100- or 300-dimensional vectors (Polajnar et al., 2014b; Kartsaklis and Sadrzadeh, 2014), demonstrating that so far, no sophisticated composition method has been able to beat vector addition on this dataset.[4] Although our contextual sentence spaces do not reach the state of the art, their performance is good enough to show that the method is viable and merits continued development.

## 6 Qualitative Analysis

In this section we provide a qualitative analysis of how the sentence spaces represent meaning. We contrast the space that has been used in previous literature (IDist) with the extra-sentential spaces (DDist, DVerb) to highlight the differences encoded by different contextual information.

### 6.1 Topic Comparison

In a word-context matrix, it is common to perform qualitative analyses of dimensionally reduced spaces by looking at the top-weighted words per topic, where the topics are induced by a dimensionality reduction technique. In our case,

[4]The results of Milajevs et al. (2014) and Hashimoto and Tsuruoka (2015) are not comparable, as they average across annotators for each SVO pair. The standard treatment of these datasets considers each annotator judgement as a separate test point, which leads to lower results overall.

| | IDist | DDist | DVerb |
|---|---|---|---|
| | **Topic 5** | **Topic 9** | **Topic 2** |
| 1 | fire **destroy** building - *fire building downtown rebuild disastrous main* | man **start** business - *business company businessman work shop* | wind **cause** damage - *damage flood cause dissipate report total destroy* |
| 2 | fire **damage** building - *building fire severely rebuild badly disastrous* | man **become** partner - *firm partner law solicitor company business* | tornado **cause** damage - *touch damage destroy dissipate cause rate* |
| 3 | building **suffer** fire - *fire building rebuild restore severe porch remodel* | man **join** business - *business father businessman firm educate company* | tornado **destroy** home - *damage touch injure destroy spawn cause* |
| 4 | Fire **destroy** building - *fire building great salem rebuilt displacement* | company **change** name - *company product inc. acquire subsidiary* | tornado **kill** people - *strike confirm touch destroy damage kill dissipate* |
| 5 | building **replace** building - *building consulate construct current* | man **become** owner - *owner business purchase businessman serve* | storm **kill** people - *dissipate cause flood strike estimate destroy* |
| 20 | fire **destroy** Building - *building fire disastrous syndicate richardson* | man **marry** widow - *daughter firstly marry die son widow sir marriage* | tornado **strike** town - *touch strike damage injure destroy rate sweep* |
| 21 | building **replace** one - *building brick wooden one demolish* | company **offer** product - *product insurance products customer company* | storm **destroy** house - *flood damage destroy neighbor dissipate affect* |
| 22 | building **cover** area - *building area meter floor square storey* | company **announce** plan - *company million announce merger* | flooding **damage** home - *cause impact amount isolate collapse* |
| 23 | man **enter** building - *building petrol thor suspected printing randall* | man **marry** Elizabeth - *son daughter elizabeth die tudor eldest bury* | wind **destroy** house - *weaken dissipate damage estimate evacuate* |
| 24 | people **destroy** building - *building machinery explosive withdrawal* | man **join** firm - *firm law counsel attorney partner practice serve clerk* | storm **drop** rainfall - *dissipate weaken cause flood total damage* |

Table 2: Top triples (in roman type, with verb in bold) for two sample topics per space with S=20. The top distributional terms for each triple are listed in italics.

the sentence space was trained by using, not the co-occurrences of *words* and contexts, but of *SVO triples* and contexts. Therefore, we can look at the highest-weighted triples from the training data for each topic.

Table 2 shows sample topics from IDist, DDist, and DVerb. Every triple has a weighting in every dimension; here we show the five top-weighted triples from the chosen topics, as well as five triples at ranks 20-24. We also show the top-weighted context words for that triple from the original unreduced space.

All three spaces show strong topical coherence; however, lexical coherence seems a greater factor in the clustering of triples for IDist. The IDist topic seems to rely heavily on the word *building*, which occurs as either subject or object (or both) in all of the triples shown here, and in fact in 23 of the top 30 triples. It can also be seen as a top context for many of the triples. Although the overall topic appears to be mostly about damaged buildings, there are several instances of triples that have to do with buildings, but not with damage, for example *building replace one* and *man enter building*.[5] Since the arguments can serve as contexts, it is very likely that triples containing similar arguments will be clustered together.

The DDist topic exhibits moderate coherence, but also more lexical variety in the argument slots than IDist. This topic is also an example of the interleaving that occurs when there are over 150,000 triples grouped into only 20 topics, with marriage triples interspersed with business-related ones. The top highest ranked context words for DDist triples often contain the subject or the object from the triple. Since the subject and object were not counted as co-occurrences for DDist (as they are intra-sentential, and DDist contexts are explicitly extra-sentential), this would seem to indicate that DDist does indeed incorporate some discourse continuity, as the entities are mentioned in surrounding sentences.

The DVerb topic appears quite coherent, and also exhibits more lexical diversity in the subject and object slots than IDist. Some top triples include light verbs such as *cause*, with the subjects and objects making it clear that they are relevant to the topic: *wind cause damage, tornado cause damage*. This is particularly exciting because the subjects and objects were not encoded in the feature space that was used to produce the topics. This space only contains the surrounding verbs, so the topical grouping of *wind* and *tornado* with *storm* and *flooding* is produced by their co-occurrence with the highly-frequent context verbs such as *destroy, damage*, and *injure*.

[5]To avoid sparsity, all instances of masculine pronouns were replaced with "man" and feminine pronouns with "woman" during preprocessing.

### 6.2 Coherence Analysis

To further explore the coherence of the topics in each sentence space, we introduce a *triple intrusion task*. This task is based on the word intrusion task for evaluation of topic models (Chang et al., 2009). In the word intrusion task, the top five words from a topic are grouped together with a sixth word which ranks low in that topic, but high in other topics. Human annotators must pick the "intruder" from a randomly-ordered list of these six words. The more coherent the topic, the easier it is for humans to identify the intruder.

Analogously, we ask human annotators to identify an intruding SVO triple. In the first version of this experiment (top5), we carry the word intrusion method over directly to triples. The top five SVO triples from each topic are chosen. The intruder is chosen as the lowest-ranking SVO triple from a topic that is also ranked in the top 1% of triples in at least one other topic. This ensures that the intruder is semantically plausible in its own right.

In the second version of the triple intrusion task (lexdiv), we explore the interaction of topic coherence with lexical coherence, by choosing from each topic the five highest-ranked SVO triples having no lexical overlap with one another. In this way we seek to test the intuition that arose from direct examination of the topics, namely that some sentence spaces have topics exhibiting semantic coherence along with greater lexical diversity.

To obtain the lexdiv triples for a topic, we begin with the top-ranked triple. We then add the next highest ranked triple which shares no lexical items (subject, verb, or object) with the first triple. We proceed to add triples in this way until we have a set of five triples. In some cases it is necessary to go fairly far down the topic rankings to find such a set; the average rank of the lowest-ranked triple for IDist is 186.5, 64.3 for DDist, and 63.9 for DVerb. These rankings themselves indicate that IDist is less lexically diverse than DDist and DVerb. The intruder is obtained as in the top5 setting, except that we also require it not to have any lexical overlap with any of the five high-ranked triples, to ensure that it blends in. Sample triple sets are shown in Figure 3. The triples were randomised and the rank was not displayed to the annotators.

We created sets of six triples for all topics from our three sentence spaces and two experiment settings, yielding 120 sets in all. We randomised the order of sets and distributed them among four annotators such that each set of triples was annotated by two annotators. The annotators were PhD students and postdoctoral researchers in computer science or linguistics. They were given no background on the source of the triples, and were instructed to pick the odd one out from each set.

| **IDist (top5)** | **DDist (lexdiv)** |
|---|---|
| man join force | man play character |
| people kill man | woman join cast |
| force take part | station air program |
| man send force | executive produce series |
| force cross river | show win award |
| program provide student | region become part |

Figure 3: Intrusion examples before randomisation. The intruder is shown as the last item in each set.

We report model accuracy and Fleiss' kappa ($\kappa$) for each sentence space and setting. Model accuracy is the proportion of examples for which the annotator chose the correct intruder. For model accuracy, we report the average accuracy over two annotators. Since no single annotator saw all the sets of triples, we arbitrarily assigned annotators to be the first or second annotator on a given division of the data. Higher model accuracy corresponds to greater topic coherence.

Higher human accuracy on the lexdiv setting would imply that a topic exhibits greater lexical diversity at higher ranks, or else that it maintains greater semantic coherence further down the ranks. Either way, the property of semantic coherence with greater lexical diversity is an interesting one from the perspective of utility for tasks such as paraphrasing and automatic summarisation.

Fleiss' $\kappa$ provides a slightly different perspective on topic coherence, as a measurement of how often the annotators agreed on their choice of intruder, serving also as a check on model accuracy since it rules out random success on intruder identification. Again, the higher the inter-annotator agreement, the more coherent the topic.

The results are given in Table 3. We observe that accuracy was consistenty lower for lexdiv than for top, which is unsurprising, since the task is much harder: in many cases for top5, all five top triples share at least one lexical item and sometimes more, while the intruder is often lexically distinct. For top5, IDist shows the highest accuracy (0.85), indicating that its topics are most coherent, or possibly that because they are the most *lexically* coherent, the intruder is easiest to iden-

| | Accuracy | | Fleiss' $\kappa$ | |
|---|---|---|---|---|
| **Space** | **top5** | **lexdiv** | **top5** | **lexdiv** |
| **IDist** | 0.85 | 0.45 | 0.88 | 0.54 |
| **DDist** | 0.78 | 0.58 | 0.55 | 0.75 |
| **DVerb** | 0.75 | 0.58 | 0.82 | 0.63 |

Table 3: Triple intrusion task: model accuracy average over two (amalgamated) annotators and Fleiss' $\kappa$.

tify. However, IDist shows the lowest accuracy for lexdiv (0.45), as well as the greatest dropoff in accuracy from top5 to lexdiv, a drop of 0.40, compared to 0.20 for DDist and 0.27 for DVerb. It appears that when triples are restricted to be lexically diverse, DDist and DVerb are more semantically coherent, with an accuracy of 0.58. We note that DVerb results would likely improve with more data and more stringent triple selection. Since we allow triples that occur two or more times, there are some triples in DVerb that are extremely sparse, and occur with only one verb., e.g. *saint pray temple* which only co-occurs with *use*, or *man plug setup* which only co-occurs with *play*, and also appears to be a result of parser error. There is at least one topic where many such triples have been grouped together, making DVerb evaluation more difficult for annotators.

We observe similar effects for Fleiss' $\kappa$. Annotaters generally achieved much higher agreement on top5 than lexdiv. The exception is DDist-top5, where agreement was much lower than for lexdiv; since model accuracy was high, it appears that each annotator had trouble with different examples, a fact for which we find no obvious explanation. IDist-top5 again achieves the highest agreement (0.88), indicating the task is fairly easy, but there is a steep dropoff to lexdiv, whereas DVerb shows a much smaller dropoff, and DDist and DVerb both show higher agreement for lexdiv than IDist does

## 7 Conclusions

We have introduced and evaluated two distributional vector spaces based on extra-sentential contexts. Results on two standard similarity tasks demonstrate that these spaces are effective in modelling sentence meaning for SVO sentences. Furthermore, a qualitative analysis indicates that extra-sentential spaces differ from the standard intra-sentential space in ways that may not be captured by the similarity tasks. The next step, therefore, is to experiment on tasks where discourse plays a larger role, such as script induction or automatic summarisation.

We have also explored only a small fraction of the many possible contextual sentence spaces. At a minimum, the role of the size and symmetry of the extra-sentential context in the quality of sentence vectors should be investigated. Future work could also investigate other, more sophisticated models that go beyond simple sentence adjacency; for example, making use of the Penn Discourse Treebank (Prasad et al., 2008, 2014).

## Acknowledgments

Tamara Polajnar and Stephen Clark are supported by ERC Starting Grant DisCoTex (306920). Laura Rimell and Stephen Clark are supported by EPSRC grant EP/I037512/1.

## References

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9:5–110.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-HLT*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL-IJCNLP*.

J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 21 (NIPS-09)*, page 288296, Vancouver, Canada.

Stephen Clark. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics second edition (to appear)*. Wiley-Blackwell, 2015.

Stephen Clark. Type-driven syntax and semantics for composing meaning vectors. In Chris Heunen, Mehrnoosh Sadrzadeh, and Edward Grefenstette, editors, *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*, pages 359–377. Oxford University Press, 2013.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG

and log-linear models. *Computational Linguistics*, 33(4):493–552.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. In J. van Bentham, M. Moortgat, and W. Buszkowski, editors, *Linguistic Analysis (Lambek Festschrift)*, volume 36, pages 345–384. 2010.

James R. Curran. *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh, 2004.

Daniel Fried, Tamara Polajnar, and Stephen Clark. 2015. Low-rank tensors for verbs in compositional distributional semantics. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, Bejing, China.

Edward Grefenstette. *Category-Theoretic Quantitative Compositional Distributional Models of Natural Language Semantics*. PhD thesis, University of Oxford, 2013.

Edward Grefenstette and Mehrnoosh Sadrzadeh. July 2011a. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011b. Experimenting with transitive verbs in a discocat. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Languge*, Edinburgh, Scotland, UK.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. 2013. Multi-step regression learning for compositional distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2015. Learning embeddings for transitive verb disambiguation by implicit tensor factorization. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, Beijing, China.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. June 2014. A study of entanglement in a categorical framework of natural language. In *Proceedings of the 11th Workshop on Quantum Physics and Logic (QPL)*, Kyoto, Japan.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In *Proceedings of COLING*, pages 549–558.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *CoRR*, abs/1506.06726. URL `http://arxiv.org/abs/1506.06726`.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of ICML*.

K. McRae, V. R. de Sa, and M. S. Seidenberg. Jun 1997. On the nature and scope of featural representations of word meaning. *Journal of experimental psychology. General*, 126(2):99–130. ISSN 0096-3445.

Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of EMNLP*, Doha Qatar.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Tamara Polajnar and Stephen Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *14th Conference of the European Chapter of the Association for Computational Linguistics, EACL'14*, Gothenburg, Sweden.

Tamara Polajnar, Luana Fagarasan, and Stephen Clark. 2014a. Reducing dimensions of tensors in type-driven distributional semantics. In *Proceedings of EMNLP 2014*, Doha, Qatar.

Tamara Polajnar, Laura Rimell, and Stephen Clark. 2014b. Using sentence plausibility to learn the semantics of transitive verbs. *CoRR*, abs/1411.7942. URL `http://arxiv.org/abs/1411.7942`.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B.Webber. 2008. The Penn discourse TreeBank 2.0. In *Proceedings of LREC*, pages 2,9612,968.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn discourse TreeBank, comparable corpora and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. 2011a. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of NIPS*.

Richard Socher, Cliff Lin, Andrew Y. Ng, and Christopher D. Manning. 2011b. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, Bellevue, Washington.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1201–1211, Jeju, Korea.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of ACL*.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.