Multidirectional Associative Optimization of Function-Specific Word Representations

Daniela Gerz[♠◇] **Ivan Vulić**[♠] **Marek Rei**[♣] **Roi Reichart**[♡] **Anna Korhonen**[♠]
[♠]Language Technology Lab, University of Cambridge
[◇]PolyAI Limited, London
[♣]Department of Computing, Imperial College London
[♡]Faculty of Industrial Engineering and Management, Technion, IIT
{dan,ivan}@poly-ai.com, marek.rei@imperial.ac.uk
roiri@ie.technion.ac.il, alk23@cam.ac.uk

Abstract

We present a neural framework for learning associations between interrelated groups of words such as the ones found in Subject-Verb-Object (SVO) structures. Our model induces a joint function-specific word vector space, where vectors of e.g. plausible SVO compositions lie close together. The model retains information about word group membership even in the joint space, and can thereby effectively be applied to a number of tasks reasoning over the SVO structure. We show the robustness and versatility of the proposed framework by reporting state-of-the-art results on the tasks of estimating selectional preference and event similarity. The results indicate that the combinations of representations learned with our task-independent model outperform task-specific architectures from prior work, while reducing the number of parameters by up to 95%.

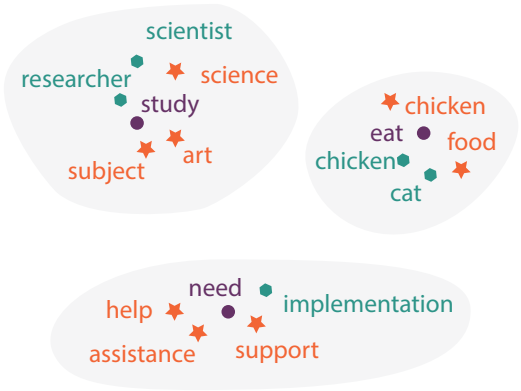


Figure 1: Illustration of three neighbourhoods in a function-specific space trained for the SVO structure (marked ○(S), ⬡(V), ☆(O)). The space is optimised such that vectors for plausible SVO compositions will be close. Note that one word can have several vectors, for example *chicken* can occur both as S and O.

1 Introduction

Word representations are in ubiquitous usage across all areas of natural language processing (NLP) (Collobert et al., 2011; Chen and Manning, 2014; Melamud et al., 2016). Standard approaches rely on the distributional hypothesis (Harris, 1954; Schütze, 1993) and learn a *single* word vector space based on word co-occurrences in large text corpora (Mikolov et al., 2013b; Pennington et al., 2014; Bojanowski et al., 2017). This purely context-based training produces general word representations that capture the broad notion of semantic relatedness and conflate a variety of possible semantic relations into a single space (Hill et al., 2015; Schwartz et al., 2015). However, this mono-faceted view of meaning is a well-known deficiency in NLP applications (Faruqui, 2016; Mrkšić et al., 2017) as it fails to distinguish between fine-grained word associations.

In this work we propose to learn a joint *function-specific* word vector space that accounts for the different roles and functions a word can take in text. The space can be trained for a specific structure, such as SVO, and each word in a particular role will have a separate representation. Vectors for plausible SVO compositions will then be optimized to lie close together, as illustrated by Figure 1. For example, the verb vector *study* will be close to plausible subject vectors *researcher* or *scientist* and object vectors *subject* or *art*. For words that can occur as either subject or object, such as *chicken*, we obtain separate vectors for each role: one for *chicken* as *subject* and another for *chicken* as *object*. The resulting representations capture more detailed associations in addition to basic distributional similarity and can be used to construct representations for the whole SVO structure.

To validate the effectiveness of our representation framework in language applications, we focus on modeling a prominent linguistic phenomenon: a general model of *who does what to whom* (Gell-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2872–2882
July 5 - 10, 2020.

Word	Nearest Neighbours
Subject	
memory	dream, feeling, shadow, sense, moment, consciousness
country	state, nation, britain, china, uk, europe, government
student	pupil, participant, learner, candidate, trainee, child
Verb	
see	saw, view, expect, watch, notice, witness
eat	drink, consume, smoke, lick, swallow, cook, ingest
avoid	eliminate, minimise, anticipate, overcome, escape
Object	
virus	bacteria, infection, disease, worm, mutation, antibody
beer	ale, drink, pint, coffee, tea, wine, soup, champagne
Joint SVO	
study (V)	researcher (S), scientist (S), subject (O), art (O)
eat (V)	food (O), cat (S), dog (S)
need (V)	help (O), implementation (S), support (O)

Table 1: Nearest neighbours in a function-specific space trained for the SVO structure. In the *Joint SVO* space (bottom) we show nearest neighbors for verbs (V) from the two other subspaces (O and S).

Mann and Ruhlen, 2011). In language, this event understanding information is typically captured by the SVO structures and, according to the cognitive science literature, is well aligned with how humans process sentences (McRae et al., 1997, 1998; Grefenstette and Sadrzadeh, 2011a; Kartsaklis and Sadrzadeh, 2014); it reflects the likely distinct storage and processing of objects (typically nouns) and actions (typically verbs) in the brain (Caramazza and Hillis, 1991; Damasio and Tranel, 1993).

The quantitative results are reported on two established test sets for compositional event similarity (Grefenstette and Sadrzadeh, 2011a; Kartsaklis and Sadrzadeh, 2014). This task requires reasoning over SVO structures and quantifies the plausibility of the SVO combinations by scoring them against human judgments. We report consistent gains over established word representation methods, as well as over two recent tensor-based architectures (Tilk et al., 2016; Weber et al., 2018) which are designed specifically for solving the event similarity task.

Furthermore, we investigate the generality of our approach by also applying it to other types of structures. We conduct additional experiments in a 4-role setting, where indirect objects are also modeled, along with a *selectional preference* evaluation of 2-role SV and VO relationships (Chambers and Jurafsky, 2010; Van de Cruys, 2014), yielding the highest scores on several established benchmarks.

2 Background and Motivation

Representation Learning. Standard word representation models such as skip-gram negative sampling (SGNS) (Mikolov et al., 2013b,a), Glove (Pennington et al., 2014), or FastText (Bojanowski et al., 2017) induce a single word embedding space capturing broad semantic relatedness (Hill et al., 2015). For instance, SGNS makes use of two vector spaces for this purpose, which are referred to as A_w and A_c. SGNS has been shown to approximately correspond to factorising a matrix $M = A_w A_c^T$, where elements in M represent the co-occurrence strengths between *words* and their *context* words (Levy and Goldberg, 2014b). Both matrices represent the same vocabulary: therefore, only one of them is needed in practice to represent each word. Typically only A_w is used while A_c is discarded, or the two vector spaces are averaged to produce the final space.

Levy and Goldberg (2014a) used dependency-based contexts, resulting in two separate vector spaces; however, the relation types were embedded into the vocabulary and the model was trained only in one direction. Camacho-Collados et al. (2019) proposed to learn separate sets of relation vectors in addition to standard word vectors and showed that such relation vectors encode knowledge that is often complementary to what is coded in word vectors. Rei et al. (2018) and Vulić and Mrkšić (2018) described related task-dependent neural nets for mapping word embeddings into relation-specific spaces for scoring lexical entailment. In this work, we propose a *task-independent* approach and extend it to work with a variable number of relations.

Neuroscience. Theories from cognitive linguistics and neuroscience reveal that single-space representation models fail to adequately reflect the organisation of semantic concepts in the human brain (i.e., *semantic memory*): there seems to be no single semantic system indifferent to modalities or categories in the brain (Riddoch et al., 1988). Recent fMRI studies strongly support this proposition and suggest that semantic memory is in fact a widely distributed neural network (Davies et al., 2009; Huth et al., 2012; Pascual et al., 2015; Rice et al., 2015; de Heer et al., 2017), where sub-networks might activate selectively or more strongly for a particular function such as modality-specific or category-specific semantics (such as objects/actions, abstract/concrete, animate/inanimate, animals, fruits/vegetables, colours, body parts, countries, flowers, etc.) (Warrington, 1975; Warrington and McCarthy, 1987; McCarthy and Warrington, 1988). This indicates a *function-specific*

division of lower-level semantic processing. Single-space distributional word models have been found to partially correlate to these distributed brain activity patterns (Mitchell et al., 2008; Huth et al., 2012, 2016; Anderson et al., 2017), but fail to explain the full spectrum of fine-grained word associations humans are able to make. Our work has been partly inspired by this literature.

Compositional Distributional Semantics. Partially motivated by similar observations, prior work frequently employs tensor-based methods for composing separate tensor spaces (Coecke et al., 2010): there, syntactic categories are often represented by tensors of different orders based on assumptions on their relations. One fundamental difference is made between atomic types (e.g., nouns) versus compositional types (e.g., verbs). Atomic types are seen as standalone: their meaning is independent from other types. On the other hand, verbs are compositional as they rely on their subjects and objects for their exact meaning. Due to this added complexity, the compositional types are often represented with more parameters than the atomic types, e.g., with a matrix instead of a vector. The goal is then to compose constituents into a semantic representation which is independent of the underlying grammatical structure. Therefore, a large body of prior work is concerned with finding appropriate composition functions (Grefenstette and Sadrzadeh, 2011a,b; Kartsaklis et al., 2012; Milajevs et al., 2014) to be applied on top of word representations. Since this approach represents different syntactic structures with tensors of varying dimensions, comparing syntactic constructs is not straightforward. This compositional approach thus struggles with transferring the learned knowledge to downstream tasks.

State-of-the-art compositional models (Tilk et al., 2016; Weber et al., 2018) combine similar tensor-based approaches with neural training, leading to task-specific compositional solutions. While effective for a task at hand, the resulting models rely on a large number of parameters and are not robust: we observe deteriorated performance on other related compositional tasks, as shown in Section 6.

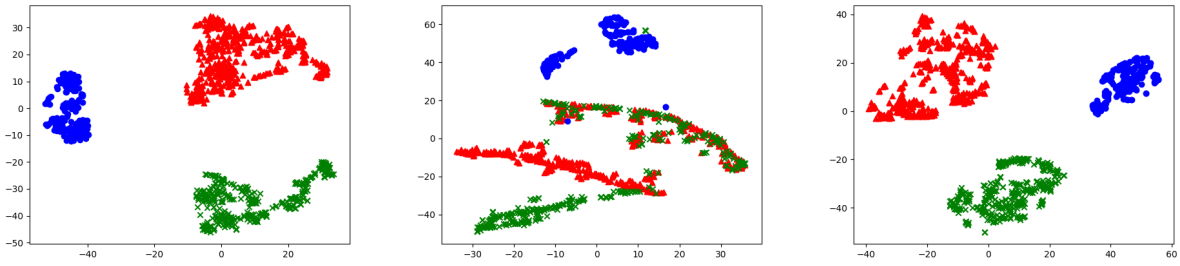
Multivariable (SVO) Structures in NLP. Modeling SVO-s is important for tasks such as compositional *event similarity* using all three variables, and *thematic fit* modeling based on *SV* and *VO* associations separately. Traditional solutions are typically based on clustering of word co-occurrence counts from a large corpus (Baroni and Lenci, 2010; Greenberg et al., 2015a,b; Sayeed et al., 2016; Emerson and Copestake, 2016). More recent solutions combine neural networks with tensor-based methods. Van de Cruys (2014) present a feed-forward neural net trained to score compositions of both two and three groups with a max-margin loss. Grefenstette and Sadrzadeh (2011a,b); Kartsaklis and Sadrzadeh (2014); Milajevs et al. (2014); Edelstein and Reichart (2016) employ tensor compositions on standard single-space word vectors. Hashimoto and Tsuruoka (2016) discern compositional and non-compositional phrase embeddings starting from HPSG-parsed data.

Objectives. We propose to induce function-specific vector spaces which enable a better model of associations between concepts and consequently improved event representations by encoding the relevant information directly into the parameters for each word during training. Word vectors offer several advantages over tensors: a large reduction in parameters and fixed dimensionality across concepts. This facilitates their reuse and transfer across different tasks. For this reason, we find our multidirectional training to deliver good performance: the same function-specific vector space achieves state-of-the-art scores across multiple related tasks, previously held by task-specific models.

3 Function-specific Representation Space

Our goal is to model the mutual associations (co-occurrences) between N groups of words, where each group represents a particular role, such as *subject* or *object* in an SVO structure. We induce an embedding matrix $\mathbb{R}^{|V_i| \times d}$ for every group $i = 1, \ldots, N$, where $|V_i|$ corresponds to the vocabulary size of the i-th group and the group vocabularies can partially overlap. For consistency, the vector dimensionality d is kept equal across all variables.

Multiple Groups. Without loss of generality we present a model which creates a function-specific vector space for $N = 3$ groups, referring to those groups as A, B, and C. Note that the model is not limited to this setup, as we show later in Section 6. A, B and C might be interrelated phenomena, and we aim for a model which can reliably score the plausibility of combining three vectors ($\vec{A}$,$\vec{B}$,$\vec{C}$) taken from this space. In addition to the full joint prediction, we aim for any two vector combinations



(a) Predicting $n \rightarrow 1$ (b) Predicting $1 \rightarrow n$ (c) Our multidirectional approach

Figure 2: The directionality of prediction in neural models is important. Representations can be of varying quality depending on whether they are induced at the input or output side of the model. Our multidirectional approach resolves this problem by training on shared representations in all directions.

($\vec{AB}$, $\vec{BC}$, $\vec{CA}$) to have plausible scores of their own. Observing relations between words inside single-group subspaces (A, B, or C) is another desirable feature.

Directionality. To design a solution with the necessary properties, we first need to consider the influence of *prediction directionality* in representation learning. A representation model such as SGNS (Mikolov et al., 2013a,b) learns two vectors for each word in one large vocabulary: one vector on the input side (word vector), another on the output side (context vector), with only the input word vectors being commonly used (Levy and Goldberg, 2014b). Here, we require several distinct vocabularies (i.e., three, one each for group A, B, and C). Instead of context vectors, we train the model to predict words from another group, hence directionality is an important consideration.

We find that prediction directionality has a strong impact on the quality of the induced representations, and illustrate this effect on an example that is skewed extremely to one side: an *n:1* assignment case. Let us assume data of two groups, where each word of group A_1 is assigned to exactly one of three clusters in group B_3. We expect a function-specific word vector space customised for this purpose to show three clearly separated clusters. Figure 2 visualises obtained representations.[1] Figure 2a plots the vector spaces when we use words on the input side of the model and predict the cluster: $A_1 \rightarrow B_3$; this can be seen as *n:1* assignment. In the opposite direction ($B_3 \rightarrow A_1$, *1:n* assignment) we do not observe the same trends (Figure 2b).

Representations for other and more complex phenomena suffer from the same issue. For example, the verb *eat* can take many arguments corresponding to various food items such as *pizza*, *beans*, or *kimchi*. A more specific verb such as *embark* might take only a few arguments such as *journey*, whereas *journey* might be fairly general and can co-occur with many other verbs themselves. We thus effectively deal with an *n:m* assignment case, which might be inclined towards *1:n* or *n:1* entirely depending on the words in question. Therefore, it is unclear whether one should rather construct a model predicting *verb* $\rightarrow$ *object* or *object* $\rightarrow$ *verb*. We resolve this fundamental design question by training representations in a *multidirectional* way with a *joint loss* function. Figure 2c shows how this method learns accurately clustered representations without having to make directionality assumptions.

4 Multidirectional Synchronous Representation Learning

The multidirectional neural representation learning model takes a list of N groups of words $(G_1, G_2, \ldots, G_N)$, factorises it into all possible "group-to-group" sub-models, and trains them jointly by combining objectives based on skip-gram negative sampling (Mikolov et al., 2013a,b). We learn a joint function-specific word vector space by using sub-networks that each consume one group G_i on the input side and predict words from a second group G_j on the output side, $i, j = 1, 2 \ldots, N; i \neq j$. All sub-network losses are tied into a single joint loss and all groups $G_1, \ldots, G_n$

[1] We train on 10K randomly selected German nouns (A_1) and their corresponding noun gender (B_3) from a German-English dictionary obtained from `dict.cc`, and train a 25-dim model for 24 epochs. Points in the figures show 1K words which were randomly selected from the 10K training vocabulary. The embedding spaces have been mapped to 2D with tSNE (van der Maaten and Hinton, 2012).

are shared between the sub-networks.

Sub-Network Architecture. We first factorise groups into sub-networks, representing all possible directions of prediction. Two groups would lead to two sub-networks $A \rightarrow B$ and $B \rightarrow A$; three groups lead to six sub-networks.

Similar to (Mikolov et al., 2013a,b), we calculate the dot-product between two word vectors to quantify their association. For instance, the sub-network $A \rightarrow B$ computes its prediction:

$$P_{A \rightarrow B} = \sigma(\vec{a} \cdot B_e^T + \vec{b}_{ab}) \quad (1)$$

where $\vec{a}$ is a word vector from the input group A, B_e is the word embedding matrix for the target group B, $\vec{b}_{ab}$ is a bias vector, and σ is the sigmoid function. The loss of each sub-network is computed using cross-entropy between this prediction and the correct labels:

$$\mathcal{L}_{A \rightarrow B} = cross\_entropy(P_{A \rightarrow B}, L_{A \rightarrow B}). \quad (2)$$

$L_{A \rightarrow B}$ are one-hot vectors corresponding to the correct predictions. We leave experiments with more sophisticated sub-networks for future work.

Synchronous Joint Training. We integrate all sub-networks into one joint model via two following mechanisms:

(1) Shared Parameters. The three embedding matrices referring to groups A, B and C are shared across all sub-networks. That is, we train one matrix per group, regardless of whether it is being employed at the input or the output side of any sub-network. This leads to a substantial reduction in the model size. For example, with a vocabulary of $50,000$ words and 25-dimensional vectors we work only with 1.35M parameters. Comparable models for the same tasks are trained with much larger sets of parameters: 26M or even up to 179M when not factorised (Tilk et al., 2016). Our modeling approach thus can achieve more that 95% reduction in the number of parameters.

(2) Joint Loss. We also train all sub-networks with a single joint loss and a single backward pass. We refer to this manner of joining the losses as *synchronous*: it synchronises the backward pass of all sub-networks. This could also be seen as a form of multi-task learning, where each sub-network optimises the shared parameters for a different task (Ruder, 2017). In practice, we perform a forward pass in each direction separately, then join all sub-network cross-entropy losses and backpropagate this joint loss through all sub-networks in order to update the parameters. The different losses are combined using addition:

$$\mathcal{L} = \sum_{\mu} \mathcal{L}_{\mu} \quad (3)$$

where μ iterates over all the possible sub-networks, $\mathcal{L}_{\mu}$ is the corresponding loss from one network, and $\mathcal{L}$ the overall joint loss.

When focusing on the SVO structures, the model will learn one joint space for the three groups of embeddings (one for S, V and O). The 6 sub-networks all share parameters and optimization is performed using the joint loss:

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{S \rightarrow V} + \mathcal{L}_{V \rightarrow S} + \mathcal{L}_{V \rightarrow O} \\ & + \mathcal{L}_{O \rightarrow V} + \mathcal{L}_{S \rightarrow O} + \mathcal{L}_{O \rightarrow S} \end{aligned} \quad (4)$$

The vectors from the induced function-specific space can then be composed by standard composition functions (Milajevs et al., 2014) to yield *event representations* (Weber et al., 2018), that is, representations for the full SVO structure.

5 Evaluation

Preliminary Task: Pseudo-Disambiguation. In the first evaluation, we adopt a standard *pseudo-disambiguation* task from the selectional preference literature (Rooth et al., 1999; Bergsma et al., 2008; Erk et al., 2010; Chambers and Jurafsky, 2010; Van de Cruys, 2014). For the three-group (S-V-O) case, the task is to score a *true* triplet (i.e., the (S-V-O) structure attested in the corpus) above all *corrupted* triplets (S-V'-O), (S'-V-O), (S-V-O'), where S', V' and O' denote subjects and objects randomly drawn from their respective vocabularies. Similarly, for the two-group setting, the task is to express a higher preference towards the attested pairs (V-O) or (S-V) over corrupted pairs (V-O') or (S'-V). We report accuracy scores, i.e., we count all items where *score(true)* $>$ *score(corrupted)*.

This simple pseudo-disambiguation task serves as a preliminary sanity check: it can be easily applied to a variety of training conditions with different variables. However, as pointed out by Chambers and Jurafsky (2010), the performance on this task is strongly influenced by a number of factors

such as vocabulary size and the procedure for constructing corrupted examples. Therefore, we additionally evaluate our models on a number of other established datasets (Sayeed et al., 2016).

Event Similarity (3 Variables: SVO). A standard task to measure the plausibility of SVO structures (i.e., *events*) is *event similarity* (Grefenstette and Sadrzadeh, 2011a; Weber et al., 2018): the goal is to score similarity between SVO triplet pairs and correlate the similarity scores to human-elicited similarity judgements. Robust and flexible event representations are important to many core areas in language understanding such as script learning, narrative generation, and discourse understanding (Chambers and Jurafsky, 2009; Pichotta and Mooney, 2016; Modi, 2016; Weber et al., 2018). We evaluate event similarity on two benchmarking data sets: **GS199** (Grefenstette and Sadrzadeh, 2011a) and **KS108** (Kartsaklis and Sadrzadeh, 2014). GS199 contains 199 pairs of SVO triplets/events. In the GS199 data set only the V is varied, while S and O are fixed in the pair: this evaluation prevents the model from relying only on simple lexical overlap for similarity computation.[2] KS108 contains 108 event pairs for the same task, but is specifically constructed without any lexical overlap between the events in each pair.

For this task function-specific representations are composed into a single *event representation/vector*. Following prior work, we compare cosine similarity of event vectors to averaged human scores and report Spearman's ρ correlation with human scores. We compose the function-specific word vectors into event vectors using simple addition and multiplication, as well as more sophisticated compositions from prior work (Milajevs et al., 2014, *inter alia*). The summary is provided in Table 4.

Thematic-Fit Evaluation (2 Variables: SV and VO). Similarly to the 3-group setup, we also evaluate the plausibility of SV and VO pairs separately in the 2-group setup. The selectional preference evaluation (Sayeed et al., 2016), also referred to as *thematic-fit*, quantifies the extent to which a noun fulfils the selectional preference of a verb given a role (i.e., agent:S, or patient:O) (McRae et al., 1997). We evaluate our 2-group function-specific spaces on two standard benchmarks: **1) MST1444** (McRae et al., 1998) contains 1,444 word pairs where humans provided thematic fit ratings on a scale from 1 to 7 for each noun to score the plausibility of the noun taking the agent role, and also taking the patient role.[3] **2) PADO414** (Padó, 2007) is similar to MST1444, containing 414 pairs with human thematic fit ratings, where role-filling nouns were selected to reflect a wide distribution of scores for each verb. We compute plausibility by simply taking the cosine similarity between the verb vector (from the V space) and the noun vector from the appropriate function-specific space (S space for agents; O space for patients). We again report Spearman's ρ correlation scores.

[2]For instance, the phrases *'people run company'* and *'people operate company'* have a high similarity score of 6.53, whereas *'river meet sea'* and *'river satisfy sea'* have been given a low score of 1.84.

Data set	Train	Test
SVO+iO	187K	15K
SVO	22M	214K
	Vocab size	**Freq.**
S	22K	people,one,company,student
V	5K	have,take,include,provide
O	15K	place,information,way,number
SV	69M	232K
	Vocab size	**Freq.**
S	45K	people,what,one,these
V	19K	be,have,say,take,go
VO	84M	240K
	Vocab size	**Freq.**
V	9K	have,take,use,make,provide
O	32K	information,time,service

Table 2: Training data statistics.

Model	Accuracy
4 Variables	
SVO+iO	0.950
3 Variables: SVO	
Van de Cruys (2009)	0.874
Van de Cruys (2014)	0.889
Tilk et al. (2016) ◇	0.937
Ours	0.943
2 Variables	
Rooth et al. (1999)	0.720
Erk et al. (2010)	0.887
Van de Cruys (2014)	0.880
Ours: SV	0.960
Ours: VO	0.972

Table 3: Accuracy scores on the pseudo disambiguation task. ◇ indicates our reimplementation.

[3]Using an example from Sayeed et al. (2016), the human participants were asked "how common is it for a {snake, monster, baby, cat} to frighten someone/something" (agent role) as opposed to "how common is it for a {snake, monster, baby, cat} to be frightened by someone/something" (patient role).

Training Data. We parse the ukWaC corpus (Baroni et al., 2009) and the British National Corpus (BNC) (Leech, 1992) using the Stanford Parser with Universal Dependencies v1.4 (Chen and Manning, 2014; Nivre et al., 2016) and extract co-occurring subjects, verbs and objects. All words are lowercased and lemmatised, and tuples containing non-alphanumeric characters are excluded. We also remove tuples with (highly frequent) pronouns as subjects, and filter out training examples containing words with frequency lower than 50. After preprocessing, the final training corpus comprises 22M SVO triplets in total. Table 2 additionally shows training data statistics when training in the 2-group setup (SV and VO) and in the 4-group setup (when adding indirect objects: SVO+iO). We report the number of examples in training and test sets, as well as vocabulary sizes and most frequent words across different categories.

Hyperparameters. We train with batch size 128, and use Adam for optimisation (Kingma and Ba, 2015) with a learning rate 0.001. All gradients are clipped to a maximum norm of 5.0. All models were trained with the same fixed random seed. We train 25-dimensional vectors for all setups (2/3/4 groups), and we additionally train 100-dimensional vectors for the 3-group (SVO) setup.

6 Results and Analysis

Pseudo-Disambiguation. Accuracy scores on the pseudo-disambiguation task in the 2/3/4-group setups are summarised in Table 3.[4] We find consistently high pseudo-disambiguation scores (>0.94) across all setups. In a more detailed analysis, we find especially the prediction accuracy of verbs to be high: we report accuracy of 96.9% for the 3-group SVO model. The vocabulary size for verbs is typically lowest (see Table 2), which presumably makes predictions into this direction easier. In summary, as mentioned in Section 5, this initial evaluation already suggests that our model is able to capture associations between interrelated groups which are instrumental to modeling SVO structures and composing event representations.

Event Similarity. We now test correlations of SVO-based event representations composed from a function-specific vector space (see Table 4) to human scores in the event similarity task. A summary of the main results is provided in Table 5. We also report best baseline scores from prior work. The main finding is that our model based on function-specific word vectors outperforms previous state-of-the-art scores on both datasets. It is crucial to note that different modeling approaches and configurations from prior work held previous peak scores on the two evaluation sets.[5] Interestingly, by relying only on the representations from the V subspace (i.e., by completely discarding the knowledge stored in S and O vectors), we can already obtain reasonable correlation scores. This is an indicator that the verb vectors indeed stores some selectional preference information as designed, i.e., the information is successfully encoded into the verb vectors themselves.

Composition	Reference	Formula
Verb only	Milajevs et al. (2014)	$\vec{V}$
Addition	Mitchell and Lapata (2008)	$\vec{S} + \vec{V} + \vec{O}$
Copy Object	Kartsaklis et al. (2012)	$\vec{S} \odot (\vec{V} \times \vec{O})$
Concat	Edelstein and Reichart (2016)	$[\vec{S},\vec{V},\vec{O}]$
Concat Addition	Edelstein and Reichart (2016)	$[\vec{S},\vec{V}] + [\vec{V},\vec{O}]$
Network	Ours	$\vec{S}\vec{V}^T + \vec{V}\vec{O}^T + \vec{S}\vec{O}^T$

Table 4: Composition functions used to obtain event vectors from function-specific vector spaces. $+$: addition, $\odot$: element-wise multiplication, $\times$: dot product. $[\cdot, \cdot]$: concatenation.

Model	Reference	Spearman's ρ GS199	Spearman's ρ KS108
Copy Object W2V	Milajevs et al. (2014)	<u>0.46</u>	0.66
Addition KS14	Milajevs et al. (2014)	0.28	<u>0.73</u>
	Tilk et al. (2016)	0.34	-
	Weber et al. (2018)	-	0.71
Ours: SVO d100			
Verb only	Ours	0.34	0.63
Addition	Ours	0.27	**0.76**
Concat	Ours	0.26	**0.75**
Concat Addition	Ours	0.32	**0.77**
Copy Object	Ours	0.40	0.52
Network	Ours	**0.53**	-

Table 5: Results on the event similarity task. Best baseline score is <u>underlined</u>, and the best overall result is provided in **bold**.

Thematic-Fit Evaluation. Correlation scores on two thematic-fit evaluation data sets are summarised in Table 6. We also report results with

[4] We also provide baseline scores taken from prior work, but the reader should be aware that the scores may not be directly comparable due to the dependence of this evaluation on factors such as vocabulary size and sampling of corrupted examples (Chambers and Jurafsky, 2010; Sayeed et al., 2016).

[5] Note the two tasks are inherently different. KS108 requires similarity between plausible triplets. Using the network score directly (which is a scalar, see Table 4) is not suitable for KS108 as all KS108 triplets are plausible and scored highly. This is reflected in the results in Table 5.

representative baseline models for the task: 1) a TypeDM-based model (Baroni and Lenci, 2010), further improved by Greenberg et al. (2015a,b) (**G15**), and 2) current state-of-the-art tensor-based neural model by Tilk et al. (2016) (**TK16**). We find that vectors taken from the model trained in the joint 3-group SVO setup perform on a par with state-of-the-art models also in the 2-group evaluation on SV and VO subsets. Vectors trained explicitly in the 2-group setup using three times more data lead to substantial improvements on PADO414. As a general finding, our function-specific approach leads to peak performance on both data sets. The results are similar with 25-dim SVO vectors.

Our model is also more light-weight than the baselines: we do not require a full (tensor-based) neural model, but simply function-specific word vectors to reason over thematic fit. To further verify the importance of joint multidirectional training, we have also compared our function-specific vectors against standard single-space word vectors (Mikolov et al., 2013b). The results indicate the superiority of function-specific spaces: respective correlation scores on MST1444 and PADO414 are 0.28 and 0.41 (vs 0.34 and 0.58 with our model). It is interesting to note that we obtain state-of-the-art scores calculating cosine similarity of vectors taken from *two groups* found in the *joint space*. This finding verifies that the model does indeed learn a joint space where co-occurring words from different groups lie close to each other.

Qualitative Analysis. We retrieve nearest neighbours from the function-specific (S, V, O) space, shown in Figure 1. We find that the nearest neighbours indeed reflect the relations required to model the SVO structure. For instance, the closest subjects/agents to the verb *eat* are *cat* and *dog*. The closest objects to *need* are three plausible nouns: *help*, *support*, and *assistance*. As the model has information about group membership, we can also filter and compare nearest neighbours in single-group subspaces. For example, we find subjects similar to the subject *memory* are *dream* and *feeling*, and objects similar to *beer* are *ale* and *pint*.

Model Variants. We also conduct an ablation study that compares different model variants. The variants are constructed by varying 1) the training regime: asynchronous (*async*) vs synchronous (*sync*), and 2) the type of parameter sharing: training on separate parameters for each sub-network (*sep*)[6] or training on shared variables (*shared*). In the asynchronous setup we update the shared parameters per sub-network directly based on their own loss, instead of relying on the joint synchronous loss as in Section 3.

Setup		Baselines		Ours	
Dataset	Eval	G15	TK16	SVO (d=100)	SV-VO (d=25)
MST1444	SV	0.36	-	**0.37**	0.31
	VO	0.34	-	**0.35**	0.35
	full	0.33	**0.38**	0.36	0.34
PADO414	SV	0.54	-	0.38	**0.55**
	VO	0.53	-	0.54	**0.61**
	full	0.53	0.52	0.45	**0.58**

Table 6: Results on the 2-variable thematic-fit evaluation. Spearman's ρ correlation.

	async		sync	
	sep	shared	sep	shared
3 Variables				
KS108 Verb only	0.56	0.48	0.58	**0.60**
KS108 Addition	0.51	0.66	0.73	**0.78**
GS199 Verb only	0.24	0.26	0.26	**0.34**
GS199 Network	0.10	0.40	0.28	**0.52**
2 Variables				
MST1444	0.17	0.10	0.30	**0.39**
PADO414	0.41	0.21	**0.44**	**0.44**

Table 7: Evaluation of different model variants, by training regime and parameter sharing.

Table 7 shows the results with the model variants, demonstrating that both aspects (i.e., shared parameters and synchronous training) are important to reach improved overall performance. We reach the peak scores on all evaluation sets using the *sync+shared* variant. We suspect that asynchronous training deteriorates performance because each sub-network overwrites the updates of other sub-networks as their training is not tied through a joint loss function. On the other hand, the synchronous training regime guides the model towards making updates that can benefit all sub-networks.

7 Conclusion and Future Work

We presented a novel multidirectional neural framework for learning function-specific word representations, which can be easily composed into multi-word representations to reason over event similarity and thematic fit. We induced a joint vector space

[6] With separate parameters we merge vectors from "duplicate" vector spaces by non-weighted averaging.

in which several groups of words (e.g., S, V, and O words forming the SVO structures) are represented while taking into account the mutual associations between the groups. We found that resulting function-specific vectors yield state-of-the-art results on established benchmarks for the tasks of estimating event similarity and evaluating thematic fit, previously held by task-specific methods.

In future work we will investigate more sophisticated neural (sub-)networks within the proposed framework. We will also apply the idea of function-specific training to other interrelated linguistic phenomena and other languages, probe the usefulness of function-specific vectors in other language tasks, and explore how to integrate the methodology with sequential models. The pre-trained word vectors used in this work are available online at: https://github.com/cambridgeltl/fs-wrep.

Acknowledgments

This work is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909) awarded to Anna Korhonen.

References

Andrew Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the ACL*, 5:17–30.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of EMNLP*, pages 59–68.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

José Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2019. Relational word embeddings. In *Proceedings of ACL*, pages 3286–3296.

Alfonso Caramazza and Argye E. Hillis. 1991. Lexical organization of nouns and verbs in the brain. *Nature*, 349(6312):788–790.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL*, pages 602–610.

Nathanael Chambers and Dan Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of ACL*, pages 445–453.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, pages 740–750.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1-4):345–384.

Ronan Collobert, Jason Weston, Lon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Tim Van de Cruys. 2009. A non-negative tensor factorization model for selectional preference induction. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 83–90.

Tim Van de Cruys. 2014. A neural network approach to selectional preference acquisition. In *Proceedings of EMNLP*, pages 26–35.

Antonio R. Damasio and Daniel Tranel. 1993. Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences of the United States of America*, 90(11):4957–60.

R. Rhys Davies, Glenda M. Halliday, John H. Xuereb, Jillian J. Kril, and John R. Hodges. 2009. The neural basis of semantic memory: Evidence from semantic dementia. *Neurobiology of Aging*, 30(12):2043–2052.

Lilach Edelstein and Roi Reichart. 2016. A factorized model for transitive verbs in compositional distributional semantics. *CoRR*, abs/1609.07756.

Guy Emerson and Ann A. Copestake. 2016. Functional distributional semantics. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 40–52.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Manaal Faruqui. 2016. *Diverse Context for Learning Word Representations*. Ph.D. thesis, Carnegie Mellon University.

Murray Gell-Mann and Merritt Ruhlen. 2011. The origin and evolution of word order. *Proceedings of the National Academy of Sciences*, 108(42):17290–17295.

Clayton Greenberg, Vera Demberg, and Asad Sayeed. 2015a. Verb polysemy and frequency effects in thematic fit modeling. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–57.

Clayton Greenberg, Asad Sayeed, and Vera Demberg. 2015b. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of NAACL-HLT*, pages 21–31.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011a. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of EMNLP*, pages 1394–1404.

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 Language Semantics*, pages 62–66.

Zellig S. Harris. 1954. Distributional Structure. *Word*, 10(2-3):146–162.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of ACL*, pages 205–215.

Wendy A. de Heer, Alexander G. Huth, Thomas L. Griffiths, Jack L. Gallant, and Frédéric E. Theunissen. 2017. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of natural language. In *Proceedings of QPL*, pages 249–261.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR (Conference Track)*.

Geoffrey Neil Leech. 1992. 100 million words of English: The British National Corpus (BNC).

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS*, pages 2177–2185.

Laurens van der Maaten and Geoffrey E. Hinton. 2012. Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55.

Rosaleen A. McCarthy and E. K. Warrington. 1988. Evidence for modality-specific meaning systems in the brain. *Nature*, 334(6181):428–430.

Ken McRae, Todd Ferretti, and Liane Amyote. 1997. Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12(2):137–176.

Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL-HLT*, pages 1030–1040.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR (Workshop Papers)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of EMNLP*, pages 708–719.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. *Proceedings of ACL*, pages 236–244.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Ashutosh Modi. 2016. Event embeddings for semantic script modeling. In *Proceedings of CoNLL*, pages 75–83.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL*, 5:309–324.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, pages 1659–1666.

Ulrike Padó. 2007. The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing.

Belen Pascual, Joseph C. Masdeu, Mark Hollenbeck, Nikos Makris, Ricardo Insausti, Song-Lin Ding, and Bradford C. Dickerson. 2015. Large-scale brain networks of the human left temporal pole: A functional connectivity MRI study. *Cerebral Cortex*, 25(3):680–702.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Procedings of EMNLP*, pages 1532–1543.

Karl Pichotta and Raymond J. Mooney. 2016. Learning statistical scripts with LSTM recurrent neural networks. In *Proceedings of AAAI*, pages 2800–2806.

Marek Rei, Daniela Gerz, and Ivan Vulić. 2018. Scoring lexical entailment with a supervised directional similarity network. In *Proceedings of ACL*, pages 638–643.

Grace E. Rice, Paul Hoffman, and Matthew A. Lambon Ralph. 2015. Graded specialization within and between the anterior temporal lobes. *Annals of the New York Academy of Sciences*, 1359(1):84–97.

M. Jane Riddoch, Glyn W. Humphreys, Max Coltheart, and Elaine Funnell. 1988. Semantic systems or system? Neuropsychological evidence re-examined. *Cognitive Neuropsychology*, 5(1):3–25.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of ACL*, pages 104–111.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic fit evaluation: An aspect of selectional preferences. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 99–105.

Hinrich Schütze. 1993. Word space. In *Proceedings of NIPS*, pages 895–902.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event participant modelling with neural networks. In *Proceedings of EMNLP*, pages 171–182.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. *Proceedings of NAACL-HLT*.

Elizabeth K. Warrington. 1975. The Selective Impairment of Semantic Memory. *Quarterly Journal of Experimental Psychology*, 27(4):635–657.

Elizabeth K. Warrington and Rosaleen A. McCarthy. 1987. Categories of knowledge. *Brain*, 110(5):1273–1296.

Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. 2018. Event representations with tensor-based compositions. In *Proceedings of AAAI*, pages 4946–4953.