What a neural language model tells us about spatial relations

Mehdi Ghanimifard **Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden
{mehdi.ghanimifard,simon.dobnik}@gu.se

Abstract

Understanding and generating spatial descriptions requires knowledge about *what* objects are related, their functional interactions, and *where* the objects are geometrically located. Different spatial relations have different functional and geometric bias. The wide usage of neural language models in different areas including generation of image description motivates the study of what kind of knowledge is encoded in neural language models about individual spatial relations. With the premise that the functional bias of relations is expressed in their word distributions, we construct multi-word distributional vector representations and show that these representations perform well on intrinsic semantic reasoning tasks, thus confirming our premise. A comparison of our vector representations to human semantic judgments indicates that different bias (functional or geometric) is captured in different data collection tasks which suggests that the contribution of the two meaning modalities is dynamic, related to the context of the task.

1 Introduction

Spatial descriptions such as "the chair is to the left of the table" contain spatial relations "to the left of" the semantic representations of which must be grounded in visual representations in terms of geometry (Harnad, 1990). The apprehension of spatial relations in terms of scene geometry has been investigated through acceptability scores of human judges over possible locations of objects (Logan and Sadler, 1996). In addition, other research has pointed out that there is an interplay between geometry and object-specific function in the apprehension of spatial relations (Coventry et al., 2001). Therefore, spatial descriptions must be grounded in two kinds of knowledge (Landau and Jackendoff, 1993; Coventry et al., 2001; Coventry and Garrod, 2004; Landau, 2016). One kind of knowledge is referential meaning, expressed in the geometry of scenes (geometric knowledge or *where* objects are) while the other kind of knowledge is higher- level conceptual world knowledge about interactions between objects which is not directly grounded in perceivable situations but is learned through our experience of situations in the world (functional knowledge or *what* objects are related). Furthermore, Coventry et al. (2001) argue that individual relations have a particular geometric and functional bias and "*under*" and "*over*" are more functionally-biased than "*below*" and "*above*". For instance, when describing the relation between a person and an umbrella in a scene with a textual context such as "*an umbrella ___ a person*", "*above*" is associated with stricter geometric properties compared to "*over*" which covers a more object-specific extra-geometric sense between the target and the landmark (i.e. *covering* or *protecting* in this case). Of course, there will be several configurations of objects that could be described either with "*over*" or "*above*" which indicates that the choice of a description is determined by the speaker, in particular what aspect of meaning they want to emphasise. Coventry et al. (2001) consider this bias for prepositions that are geometrically similar and therefore the functional knowledge is reflected in different preferences for objects that are related. However, such functional differences also exist between geometrically different relations.

This poses two interesting research questions for computational modelling of spatial language. The first one is how both kinds of knowledge interact with individual spatial relations and how models of spatial language can be constructed and learned within end-to-end deep learning paradigm. Ramisa et al. (2015) compare the performance of classifiers using different multi-modal features (visual, geometric and textual) to predict a spatial preposition. Schwering (2007) applies semantic similarity metrics of spatial relations on geo-

Proceedings of SpLU-RoboNLP, pages 71–81
Minneapolis, MN, June 6, 2019.

graphical data retrieval. Collell et al. (2018) show that word embeddings can be used as predictive features for common sense knowledge about location of objects in 2D images. The second question is related to the extraction of functional knowledge for applications such as generation of spatial descriptions in a robot scenario. Typically, a robot will not be able to observe all object interactions as in (Coventry et al., 2004) to learn about the interaction of objects and choose the appropriate relation. Following the intuition that the functional bias of spatial relations is reflected in a greater selectivity for their target and landmark objects, Dobnik and Kelleher (2013, 2014) propose that the degree of association between relations and objects in the corpus of image descriptions can be used as filters for selecting the most applicable relation for a pair of objects. They also demonstrate that entropy-based analysis of the targets and landmarks can identify the functional and geometric bias of spatial relations. They use descriptions from a corpus of image descriptions because here the prepositions in spatial relations are used mainly in the spatial sense. The same investigation of textual corpora such as BNC (Consortium et al., 2007) does not yield such results as there prepositions are used mainly in their non-spatial sense.[1] Similarly, Dobnik et al. (2018) inspect the perplexity of recurrent language models for different descriptions containing spatial relations in the Visual Genome dataset of image captions (Krishna et al., 2017) in order to investigate their bias for objects.

In this paper, we follow this line of work and (i) further investigate what semantics about spatial relations are captured from descriptions of images by generative recurrent neural language models, and (ii) whether such knowledge can be extracted, for example as vector representations, and evaluated in tests. The neural embeddings are opaque to interpretations per se. The benefit of using recurrent language models is that they allow us to (i) deal with spatial relations as multi-word expressions and (ii) they learn their representations within their contexts:

(a) *a cat on a mat*
(b) *a cat on the top of a mat*
(c) *a mat under a cat*

[1]We may call this metaphoric or highly functional usage which is completely absent of the geometric dimension.

In (a) and (b), the textual contexts are the same "*a cat ___ a mat*" but the meaning of the spatial relations, one of which is a multi-word expression, are slightly different. In (c) the context is made different through word order.

The question of what knowledge (functional or geometric) should be represented in the models can be explained in information-theoretic terms. The low surprisal of a textual language model on a new text corpora is an indication that the model has encoded the same information content as the text. In the absence of the geometric knowledge during the training of the model, this means that a language model encodes the relevant functional knowledge. We will show that the degree to which each spatial description containing a spatial relation encodes functional knowledge in different contexts can be used as source for building distributional representations. We evaluate these representations intrinsically in reasoning tests and extrinsically against human performance and human judgment.

The contributions of this paper are:

1. It is an investigation of the semantic knowledge about spatial relations learned from textual features in recurrent language models with intrinsic and extrinsic methods of evaluation on internal representations.
2. It proposes a method of inspecting contextual performance of generative neural language models over a wide categories of contexts.

This paper is organised as follows: in Section 2 we describe how we create distributional representations with recurrent neural language models, in Section 3 we describe our computational implementations that build these representations, and in Section 4 we provide their evaluation. In Section 5 we give our final remarks.

2 Neural representations of spatial relations

Distributional semantic models produce vector representations which capture latent meanings hidden in association of words in documents (Church and Hanks, 1990; Turney and Pantel, 2010). The neural word embeddings were initially introduced as a component of neural language models (Bengio et al., 2003). However, subsequently neural language models such as word2vec (Mikolov et al., 2013) and GloVe (Pennington

et al., 2014) have become used to specifically learn word embeddings from large corpora. The word embeddings trained by these models capture world-knowledge regularities expressed in language by learning from the distribution of context words which can be used for analogical reasoning[2]. Moreover, sense embeddings (Neelakantan et al., 2014) and contextual embeddings (Peters et al., 2018) have shown to provide fine-grained representation which can discriminate between different word senses or contexts, for example in substituting synonym words and multi-words in sentences (McCarthy and Navigli, 2007).

However, meaning is also captured by generative recurrent neural language models used to generate text rather than predict word similarity. The focus of our work is to investigate what semantics about spatial relations is captured by these models. Generative language models use the chain rule of probability for step-by-step prediction of the next word in a sequence. In these models, the probability of a sequence of words (or sometimes characters) is defined as the multiplication of conditional probabilities of each word given the previous context in a sequence:

$$P(w_{1:T}) = \prod_{t=1}^{T-1} P(w_{t+1}|w_{1:t}) \quad (1)$$

where T is the length of the word sequence. The language model estimates the probability of a sequence in Equation (1) by optimising parameters of a neural network trained over sufficient data. The internal learned parameters includes embeddings for each word token which can be used as word level representations directly.

An alternative way of extracting semantic prediction from a generative neural language model which we are going to explore in this paper is to measure the fidelity of the model's output predictions against a new ground truth sequence of words. This is expressed in the measure of *Perplexity* as follows:

$$PP(S) = (\prod_{s \in S} P(w_{1:t} = s))^{\frac{-1}{|S|}} \quad (2)$$

where S is a collection of ground truth sentences. Perplexity is a measure of the difficulty of a generation task which is based on the information theoretic concept of entropy (Bahl et al., 1983). It is based on *cross-entropy* which takes into account the probability of a sequence of words in ground truth sentences and the probability of a language model generating that sequence. It is often used for intrinsic evaluation of word- error rates in NLP tasks (Chen et al., 1998). However, in this paper we use perplexity as a measure of fit of a pre-trained generative neural language model to a collection of sentences.

Our proposal is as follows. We start with the hypothesis that in spatial descriptions some spatial relations (those that we call functional) are more predictable from the associated word contexts of targets and landmarks than their grounding in the visual features. Hence, this will be reflected in a perplexity of a (text-based) generative language model trained on spatial descriptions. Descriptions with functionally-biased spatial relations will be easier to predict by this language model than geometrically-biased spatial descriptions and will therefore have lower perplexity. If two sequences of words where only the spatial relations differ (but target and landmark contexts as well as other words are the same) have similar perplexity, it means that such spatial relations have similar selectional requirements and are therefore similar in terms of functional and geometric bias. We can exploit this to create vector representations for spatial relations as follows. Using a dictionary of spatial relations, we extract collections of sentences containing a particular spatial relation from a held-out dataset not used in training of the language model. The collection of sentences with a particular spatial relation are our context templates. More specifically, for our list of spatial relations $\{r_1, r_2, ..., r_k\}$, we replace the original relation r_i with a target relation r_j in its collection of sentences, e.g. we replace *to the right of* $_i$ with *in front of* $_j$. The outcome is a collection of artificial sentences $S_{i \to j}$ that are identical to the human-generated sentences except that they contain a substituted spatial relation. The perplexity of the language model on these sentences represents the association between the original spatial relation and the context in which this has been projected:

$$PP(S_{i \to j}) = PP_{i,j} = P(rel_i, c_{rel_j})^{\frac{1}{-N'}} \quad (3)$$

where c_{rel_j} is the context of rel_i, and $PP_{i,j}$ is the perplexity of the neural language model on the

[2]For example, "a is to a^* as b is to b^*" can be queried with simple vector arithmetic $king - man + woman \approx queen$. More specifically, with a search over vocabulary with cosine similarity: $\underset{b^* \in V/\{a^*,b,a\}}{arg\,max} \; cos(b^*, a^* - a + b)$

sentence collection where relation rel_i is artificially placed in the contexts of relation rel_j. If rel_i and rel_j are associated with similar contexts, then we expect low perplexity for $S_{i \to j}$, otherwise the perplexity will be high. Finally, the perplexity of rel_i against each collection c_{rel_j} is computed and normalised within each collection (Equation 4) and the resulting vector per rel_i over all contexts is represented as a unit vector (Equation 5).

$$m_{i,j} = \frac{PP_{i,j}}{\sum_{i'=1}^{k} PP_{i',j}} \quad (4)$$

$$\hat{v_i} = \frac{v_i}{||v_i||} \qquad v_i = (m_{i,1}, ..., m_{i,k})^T \quad (5)$$

where $\hat{v_i}$ is the vector representation of the relation rel_i. These vectors create a matrix. In a particular cell of some row and some column, high perplexity means that the spatial relation in that row is less swappable with the context in the column, while a low perplexity means that the spatial relation is highly swappable with that context. This provides a measure similar to mutual information (PPMI) in traditional distributional vectors (Church and Hanks, 1990).

In conclusion, representing multi-word spatial relations in a perplexity matrix of different contexts allows us to capture their semantics based on the predictions and the discriminatory power of the language model. If all spatial relations are equally predictable from the language model such vector representations will be identical and vector space norms will not be able to discriminate between different spatial relations. In the following sections we report on the practical details how we build the matrix (Section 3) and evaluate it on some typical semantic tasks (Section 4). The implementation and evaluation code: `https://github.com/GU-CLASP/what_nlm_srels`

3 Dataset and models

3.1 Corpus and pre-processing

We use Visual Genome region description corpus (Krishna et al., 2017). This corpus contains 5.4 million descriptions of 108 thousand images, collected from different annotators who described specific regions of each image. As stated earlier, the reason why we use a dataset of image descriptions is because we want to have spatial usages of prepositions. Other image captioning datasets such as MSCOCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015) could also be used. However, our investigation has shown that since the task in these datasets in not to describe directly the relation between selected regions, common geometric spatial relations are almost missing in them: there are less then 30 examples for "*left of*" and "*right of*" in these datasets.

After word tokenisation with the space operator, we apply pre-processing which removes repeated descriptions per-image and also descriptions that include uncommon words with frequency less than 100[3] Then we split the sentences into 90%-10% portions. The 90% is used for training the language model (Section 3.2), and 10% is used for generating the perplexity vectors by extracting sentences with spatial relations that represent our context bins (Section 3.3). The context bins are used for generating artificial descriptions $S_{i \to j}$ on which the language model is evaluated for perplexity.

3.2 Language model and GloVe embeddings

We train a generative neural language model on the 90% of the extracted corpus (Section 3.1) which amounts to 4,537,836 descriptions of maximum length of 29 and 4,985 words in the vocabulary. We implement a recurrent language model with LSTM (Hochreiter and Schmidhuber, 1997) and a word embeddings layer similar to Gal and Ghahramani (2016) in Keras (Chollet et al., 2015) with TensorFlow (Abadi et al., 2015) as back-end. The Adam optimiser (Kingma and Ba, 2014) is used for fitting the parameters. The model is set up with 300 dimensions both for the embedding- and the LSTM units. It is trained for 20 epochs with a batch size of 1024.

In addition to the generative LSTM language model, we also train on the same corpus GloVe (VG) embeddings with 300 dimensions and a context-window of 5 words. Finally, we also use pre-trained GloVe embeddings on the Common Crawl (CC) dataset with 42B tokens[4].

[3] The pre-processing leaves 5,042,039 descriptions in the corpus with maximum 31 tokens per sentence. The relatively high threshold of 100 tokens is chosen to insure sufficient support in the 10% of held-out data for bucketing. We did not use OOV tokens because the goal of the evaluation is to capture object-specific properties about spatial relations and OOV tokens would interfere with this.

[4] `http://nlp.stanford.edu/data/glove.42B.300d.zip`

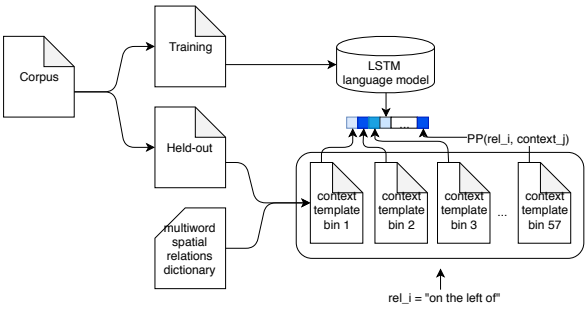


Figure 1: Generating perplexity-based vectors for each spatial relation.

3.3 Perplexity vectors

Based on the lists of spatial prepositions in (Landau, 1996) and (Herskovits, 1986), we have created a dictionary of spatial relations which include single word relations as well as all of their possible multi-word variants. This dictionary was applied on the 10% held-out dataset where we found 67 single- and multi-word spatial relation types in total. As their frequency may have fallen below 100 words due to the dataset split, we further remove all relations below this threshold which gives us 57 relations. We also create another list of relations where composite variants such as "to the left of" and "on the left of" are grouped together as "left" which contains 44 broad relations. We group the sentences by the relation they are containing to our context bins using simple pattern matching on strings. Table 1 contains some examples of our context bins. The bins are used for artificial sentence generation as explained in the previous section.

Relation (rel_i)	Context bin (c_{rel_i})
above	scissors _____ the pen
	tall building _____ the bridge
	...
below	pen is _____ scissors
	bench _____ the green trees
	...
next to	a ball-pen _____ the scissors
	car _____ the water
	...

Table 1: Examples of context bins based on extracted descriptions from Visual Genome. The images that belong to these descriptions are shown in Appendix B.

For each of the 67 spatial relations extracted from the larger corpus, there are 57 collections of sentences (=the number of relations in the smaller corpus). Hence, there are $3,819 (= 67 \times 57)$ possible projections $S_{i \to j}$, where a relation i is placed in the context j, including the case where there is no swapping of relations when $j = i$. The process is shown in Figure 1. The vector of resulting perplexities in different contexts is normalised according to Equation 5 which gives us perplexity vectors (P-vectors) as shown in Figure 2.

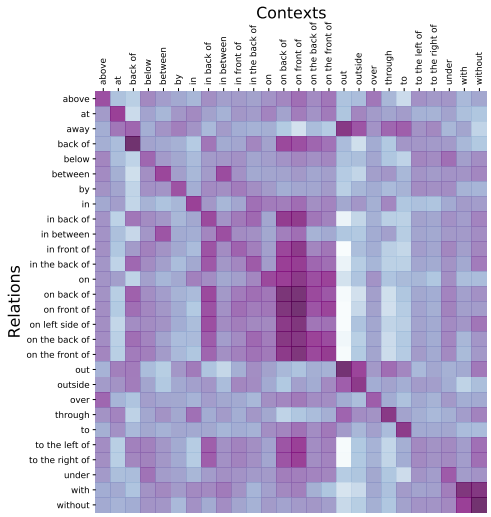


Figure 2: A matrix of perplexity vectors for 28 spatial relations and 26 contexts. For the full 67×57 matrix see Appendix C. The rows represent spatial relations and columns represent the normalised average perplexity of a language model when this relation is swapped in that context.

In addition to the P-vectors we also create representations learned by the word embedding layer in the generative language model that we train. For each of the 44 broad single-word spatial relations we extract a 300-dimensional embedding vector from the pre-trained recurrent language model (LM-vectors). In order to produce LM-vectors for the multi-word spatial relations, we simply sum the embeddings of the individual words. For example the embedding vector for "to the left of" is $v_{to} + v_{the} + v_{left} + v_{of}$. The same method is also used for the GloVe embeddings.

3.4 Human judgments

In order to evaluate our word representations we compare them to three sources of human judgments. The first one are judgments about the the fit of each spatial relation over different geometric locations of a target object in relation to a landmark which can be represented as spatial templates (Logan and Sadler, 1996). The second

are 88,000 word association judgments by English speakers from (De Deyne et al., 2018). In each instance participants were presented a stimulus word and were asked to provide 3 other words. The dataset contains 4 million responses on 12,000 cues. Based on the collective performance of annotators, the dataset provides association strengths between words (which contain any kind of words, not just spatial words) as a measure of their semantic relatedness. Finally, we collected a new dataset of word similarity judgments using Amazon Mechanical Turk. Here, the participants were presented with a pair of spatial relations at a time. Their task was to use a slider bar with a numerical indicator to express how similar the pair of words are. The experiment is similar to the one described in (Logan and Sadler, 1996) except that in our case participants only saw one pair of relations at a time rather than the entire list. The shared vocabulary between these three datasets covers *left*, *right*, *above*, *over*, *below*, *under*, *near*, *next*, *away*.

4 Evaluation

As stated in Section 2 the P-vectors we have built are intended to capture the discriminatory power of a generative language model to encode and discriminate different spatial relations, their functional bias. In this section we evaluate the P-vectors on several common intrinsic and extrinsic tests for vectors. If successful, this demonstrates that such knowledge has indeed been captured by the language model. We evaluate both single- and multi-word relations.

4.1 Clustering

Method Figure 2 and its complete version in Appendix C show that different spatial relations have different context fingerprints. To find similar relations in this matrix we can use *K-means clustering*. K-mean is a non-convex problem: different random initialisation may lead to different local minima. We apply the clustering on 67 P-vectors for multi-word spatial relations and qualitatively examine them for various sizes k. The optimal number of clusters is not so relevant here, only that for each k we get reasonable associations that follow our semantic intuitions.

Results As shown in Table 2, with $k = 30$, the clustering of perplexity vectors shows acceptable semantics of each cluster. There are clusters with synonymous terms such as (15. *above*, *over*) or

1. to	18. up; down; off
2. on	19. with; without
3. away	20. together; out
4. here	21. outside; inside
5. into	22. near; beside; by
6. from	23. top; front; bottom
7. during	24. in between; between
8. back of	25. along; at; across; around
9. through	26. beneath; below; under; behind
10. alongside	27. right; back; left; side; there
11. along side	28. to the left of; to the right of; next to
12. underneath	29. in back of; in the back of; on the back of; at the top of
13. in; against	30. on the top of; on side of; on the bottom of; on left side of; on top of; on the front of; on back of; on the side of; on front of; on bottom of
14. in front of	
15. above; over	
16. to the side	
17. onto; toward	

Table 2: K-means clusters of spatial relations based on their P-vectors.

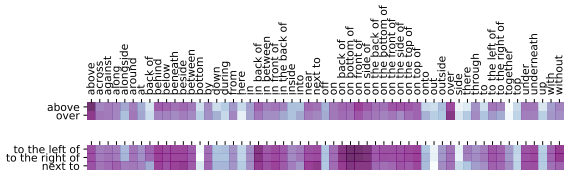


Figure 3: The P-vectors of two clusters.

(26. *below*, *under*). Some clusters have variants of multi-word antonymous such as (30. *on the top of*, *on the bottom of*). Other clusters have a mixture of such relations, e.g. (27. *right*, *back*, *left*, *side*, and *there*).

Discussion The inspection of the perplexities of two of these clusters in Figure 3 shows that the language model has learned different selectional properties of spatial relations: *above* and *over* are generally more selective of their own contexts, while *to the left of* and *to the right of* show a higher degree of confusion with a variety of the P-vector contexts. High degree of confusion in *left* and *right* is consistent with the observation in (Dobnik and Kelleher, 2013) that these relations are less dependent on the functional relation between particular objects and therefore have a higher geometric bias. On the other hand, *above* and *over* seem to be more selective of their contexts. The functional distinction between *above* and *over* is mildly visible: the shades of blue in *above* are slightly darker than *over*.

4.2 Analogical reasoning with relations

The intrinsic properties of vector representations (the degree to which they capture functional associations between relations and their objects) can be tested with their performance in analogical reasoning tasks. We compare the performance of

	Single word	Multi-words
GloVe (CC)	0.56	0.36
GloVe (VG)	0.43	0.29
LM	0.86	0.45
P-vectors	0.62	0.47
Random	0.11	0.05

Table 3: The accuracies of different representations on the word analogy test.

the P-vectors (Section 3.3), the embeddings of the language model used to create the P-vectors and GloVe embeddings (Section 3.2) in two analogical tasks which require both geometric and functional reasoning.

4.2.1 Predicting analogical words

Method The task is similar to the analogy test (Mikolov et al., 2013; Levy et al., 2015) where two pairs of words are compared in terms of some relation "a is to a' as b is to b'". We manually grouped spatial relations that are opposite in one geometric dimension to 6 groups. These are: Group 1: left, right; Group 2: above, below; Group 3: front, back; Group 4: with, without; Group 5: in, out; and Group 6: up, down. We generate all possible permutations of these words for the analogical reasoning task which gives us 120 permutations. We expand these combinations to include multi-word variants. This dataset has 85,744 possible analogical questions such as (*above* :: *below*, *to the left of* :: ?). We accept all variants of a particular relation (e.g. *to the right side of* and *to the right of*) as the correct answer.

Results As shown in in Table 3, on the single-word test suite, the LM-embeddings perform better than other models. On multi-word test suite the P-vectors perform slightly better. On both test suites, GloVe trained on Common Crawl performs better than GloVe trained on Visual Genome. However, its performance on multi-word relations is considerably lower. We simulated random answers as a baseline to estimate the difficulty of the task. Although the multi-word test suite has $\sim$700 times more questions than the test suite with single-word relations, it is only approximately 2-times more difficult to predict the correct answer in the multi-word dataset compared to the single-word dataset.

Discussion The perplexity of the language model on complete context phrases (Multi-words) is as good indicator of semantic relatedness as the word embeddings of the underlying language model and much better than GloVe embeddings. The good performance of the P-vectors explains the errors of the language model in generating spatial descriptions. The confusion between *in front of* and *on the back of* is similar to the confusion between *to the left of* and *to the right of* in terms of their distribution over functional contexts. Hence, a similar lack of strong functional associations allows the vectors to make inference about geometrically related word-pairs. This indicates that functional and geometric bias of words are complementary. There are two possible explanations why P-vectors perform better than LM-embeddings on multi-word vectors: (i) low-dimensions of P-vectors (57D) intensify the contribution of spatial contexts for analogical reasoning compared to high-dimensional LM-embeddings (300D); (ii) summing the vectors of the LM-embeddings for multi-words reduces their discriminatory effect.

4.2.2 Odd-one-out

Method Based on the semantic relatedness of words, the goal of this task is to find the odd member of the three. The ground truth for this test are the following five categories of spatial relations, again primarily based on geometric criteria: X-axis: left, right; Y-axis: above, over, under, below; Z-axis: front, back; Containment: in, out; and Proximity: near, away. Only the Y-axis contains words that are geometrically similar but functionally different, e.g. *above*/*over*. In total there are 528 possible instances with 3,456 multi-word variations. The difficulty of the task is the same for both single- and multi-word expressions as the choice is always between three words. Hence, the random baseline is 0.33.

Results Table 4 shows the accuracy in predicting the odd relation out of the three. We also add a comparison to fully geometric representations captured by spatial templates (Logan and Sadler, 1996). Ghanimifard and Dobnik (2017) show that spatial templates can be compared with Spearman's rank correlation coefficient $\rho_{X,Y}$ and therefore we also include this similarity measure. Since our groups of relations contain those that are geometric opposites in each dimension, we take the absolute value of $|\rho_{X,Y}|$. Spatial templates are not able to recognise relatedness without the right distance measure, $|\rho_{X,Y}|$. LM-embeddings perform better than other vectors in both tests, but

	Single word		Multi-words	
	$1-cos$	$\|\rho\|$	$1-cos$	$\|\rho\|$
GloVe (CC)	0.62	0.68	0.52	0.58
GloVe (VG)	0.61	0.61	0.58	0.59
LM	0.87	0.90	0.82	0.88
P-vectors	0.72	0.70	0.64	0.52
Sp Templates	0.22	1.0	-	-

Table 4: The accuracies in odd-one-out tests.

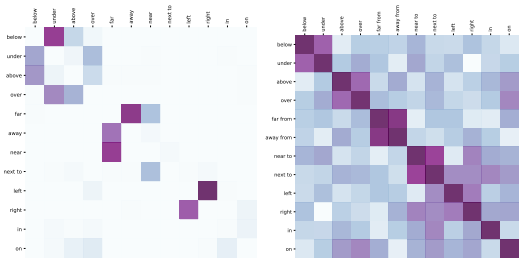
P-vectors follow closely. All models have a low performance on the multi-word test suite. When using $|\rho_{X,Y}|$ all vectors other than P-vectors produce better results. While we do not have an explanation for this, it is interesting to observe that $|\rho_{X,Y}|$ is a better measure of similarity than cosine.

Discussion The results demonstrate that using functional representations based on associations of words can predict considerable information about geometric distinctions between relations, e.g. distinguishing *to the right of* and *above*, and this is also true for P-vectors. As stated earlier, our explanation for this is that functional and geometric knowledge is in complementary distribution. This has positive and negative implications for joint vision and language models used in generating spatial descriptions. In the absence of geometric information, language models provide strong discriminative power in terms of functional contexts, but even if geometric latent information is expressed in them, an image captioning system still needs to ground each description in the scene geometry.

4.3 Similarity with human judgments

We compare the cosine similarity between words in LM- and P-vector spaces with similarities from (i) word association judgments (De Deyne et al., 2018), (ii) our word similarity judgments from AMT, and (iii) spatial templates (Section 3.4). We take the maximum subset of shared vocabulary between them, including *on*, *in* only shared between (i) and (ii). Since (i) is an association test, unrelated relations do not have association strengths. There are 55 total possible pairs of 11 words, while only 28 pairs are present in (i) as shown in Figure 4.

Method We take the average of the two way association strengths if the association exists and for (i) we assign a zero association for unrelated pairs such as *left* and *above*. Spearman's rank correlation coefficient $\rho_{X,Y}$ is used to compare the calculated similarities.



(a) (b)

Figure 4: (i) Word association judgments and (ii) word similarity judgments

Results Table 5 shows ranked correlations of different similarity measures. Spatial templates do not correlate with (WA) word associations and (WS) word similarities. On 28 pairs there is a weak negative correlation between spatial templates and WS. The correlation of similarities of two different human judgments is positive but weak ($\rho = 0.33$). The similarities predicted by LM-vectors and P-vectors correlate better with WA than WS.

	55 pairs		28 pairs	
	WA	WS	WA	WS
SpTemp	-0.02	-0.08	0.06	-0.35
LM	0.48^{***}	0.15	0.59^{***}	0.08
P	0.48^{***}	0.19	0.40^{**}	-0.08

p-values: $* < 0.01$, $** < 0.01$, $*** < 0.001$

Table 5: Spearman's ρ between pairwise lists of similarities. WA are similarities based on word associations and WS are direct word similarities from human judgments.

Discussion The low correlation between the two similarities from human judgments is surprising. Our explanation is that this is because of different priming to functional and geometric dimension of meaning in the data collection task. In the WA task participants are not primed with the spatial domain but they are providing general word associations, hence functional associations. On the other hand, in the WS task participants are presented with two spatial relations, e.g. *left of* and *right of*, and therefore the geometric dimension of meaning is more explicitly attended. We also notice that judgments are not always unison, the same pair may be judged as similar and dissimilar which further confirms that participants are selecting between two different dimensions of meaning. This observation is consistent with our argument that LM-vectors and P-vectors encode functional knowledge. Both representations correlate

better with WA than with WS. Finally, (Logan and Sadler, 1996) demonstrate that WS judgments can be decomposed to dimensions that correlate with the dimensions of the spatial templates. We leave this investigation for our future work.

5 Conclusion and future work

In the preceding discussion, we have examined what semantic knowledge about spatial relations is captured in representations of a generative neural language model. In particular, we are interested if the language model is able to encode a distinction between functional and geometric bias of spatial relations and how the two dimensions of meaning interact. The idea is based on earlier work that demonstrates that this bias can be recovered from the selectivity of spatial relations for target and landmark objects. In particular, (i) we test the difference between multi-word spatial relations at two levels: the word embeddings which are a form of internal semantic representations in a language model and the perplexity-based P-vectors which are external semantic representations based on the language model performance; (ii) we project spatial relations in the contexts of other relations and we measure the fit of the language model to these contexts using perplexity (P-vectors); (iii) we use these contexts to build a distributional model of multi-word spatial relations; (iv) in the evaluation on standard semantic similarity tasks, we demonstrate that these vectors capture fine semantic distinctions between spatial relations; (v) we also demonstrate that these representations based on word-context associations latently capture geometric knowledge that allows analogical reasoning about space; this suggests that functional and geometric components of meaning are complementary: (vi) doing so we also demonstrated that generation of spatial descriptions is also dependent on textual features, even if the system has no access to the visual features of the scene. This has implications for baselines for image captioning and how we evaluate visual grounding of spatial relations.

Our work could be extended in several ways, including by (i) using the knowledge about the bias of spatial relations to evaluate captioning tasks with spatial word substitutions (Shekhar et al., 2017a,b); (ii) examining how functional knowledge is complemented with visual knowledge in language generation (Christie et al., 2016; Delecraz et al., 2017) (iii) using different contextual embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) for the embedding layer of the generative language model rather than our specifically-trained word embeddings; note that P-vectors are representations of collections of context based on the performance of the decoder language model while ELMo and BERT are representations of specific context based on the encoder language model; (iv) comparing language models for spatial descriptions from different pragmatic tasks. As the focus of image captioning is to best describe the image and not for example, spatially locate a particular object, the pragmatic context of image descriptions is biased towards the functional sense of spatial relations. Our analysis should be extended to different kinds of corpora, for example those for visual question answering, human-robot interaction, and navigation instructions where we expect that precise geometric locating of objects receives more focus. Therefore, we expect to find a stronger geometric bias across all descriptions and a lower performance of our representations on analogical reasoning.

Acknowledgements

We are grateful to the anonymous reviewers for their helpful comments. The research of the authors was supported by a grant from the Swedish Research Council (VR project 2014-39) to the Centre for Linguistic Theory and Studies in Probability (CLASP) at Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. 1983. A maximum likelihood approach to continu-

ous speech recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2:179–190.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Stanley F Chen, Douglas Beeferman, and Ronald Rosenfeld. 1998. Evaluation metrics for language models. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 275–280. Citeseer.

François Chollet et al. 2015. Keras. https://github.com/keras-team/keras.

Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. *arXiv preprint arXiv:1604.02125*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

BNC Consortium et al. 2007. The british national corpus, version 3 (bnc xml edition). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium.*

Kenny R Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V Richards. 2004. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In *International Conference on Spatial Cognition*, pages 98–110. Springer.

Kenny R Coventry and Simon C Garrod. 2004. *Saying, seeing, and acting: the psychological semantics of spatial prepositions*. Psychology Press, Hove, East Sussex.

Kenny R Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of memory and language*, 44(3):376–398.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, pages 1–20.

Sebastien Delecraz, Alexis Nasr, Frédéric Béchet, and Benoit Favre. 2017. Correcting prepositional phrase attachments using multimodal corpora. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 72–77.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, pages 1–12, Montpellier, France. Association for Computational Linguistics.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. *Language and space*, pages 317–363.

Barbara Landau. 2016. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive Science*, 41(2):321–350.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

G.D. Logan and D.D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In M. Bloom, P.and Peterson, L. Nadell, and M. Garrett, editors, *Language and Space*, pages 493–529. MIT Press.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2641–2649. IEEE.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220.

Angela Schwering. 2007. Evaluation of a semantic similarity measure for natural language spatial relations. In *International Conference on Spatial Information Theory*, pages 116–132. Springer.

Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. Vision and language integration: moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

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