Exploring the Functional and Geometric Bias of Spatial Relations Using Neural Language Models

Simon Dobnik[*]   **Mehdi Ghanimifard**[*]   **John D. Kelleher**[†]

[*]CLASP and FLOV, University of Gothenburg, Sweden

[†]Dublin Institute of Technology, Ireland

[*]{simon.dobnik,mehdi.ghanimifard}@gu.se [†]john.d.kelleher@dit.ie

Abstract

The challenge for computational models of spatial descriptions for situated dialogue systems is the integration of information from different modalities. The semantics of spatial descriptions are grounded in at least two sources of information: (i) a geometric representation of space and (ii) the functional interaction of related objects that. We train several neural language models on descriptions of scenes from a dataset of image captions and examine whether the functional or geometric bias of spatial descriptions reported in the literature is reflected in the estimated perplexity of these models. The results of these experiments have implications for the creation of models of spatial lexical semantics for human-robot dialogue systems. Furthermore, they also provide an insight into the kinds of the semantic knowledge captured by neural language models trained on spatial descriptions, which has implications for image captioning systems.

1   Introduction

Spatial language understanding is fundamental requirement for human-robot interaction through dialogue. A natural task for a human to request a robot to fulfil is to retrieve or replace an object for them. Consequently, a particularly frequent form of spatial description within human-robot interaction is a *locative expression*. A locative expression is a noun phrase that describes the location of one object (the *target object*) relative to another object (the *landmark*). The relative location of the target object is specified through a prepositional phrase:

Bring me *the big red book* *on* *the table*.

Target   *Landmark*

Prepositional Phrase

Locative Expression

In order to understand these forms of spatial descriptions a robot must be equipped with computational models of the spatial semantics of prepositions that enable them to ground the semantics of the locative expression relative to the context of the situated dialogue.

A natural approach to developing these computational models is to define them in terms of scene *geometry*. And, indeed, there is a tradition of research that follows this path, see for example (Logan and Sadler, 1996; Kelleher and Costello, 2005, 2009). However, there is also a body of experimental and computational research that has highlighted that the semantics of spatial descriptions are dependent on several sources of information beyond scene geometry, including *functional semantics* (which encompasses a range of factors such as world knowledge about the typical interactions between objects, and object affordances) (Coventry and Garrod, 2004). We can illustrate this distinction between geometric and functionally defined semantics using a number of examples. To illustrate a geometric semantics: assuming a spatial meaning, anything can be described as *to left of* anything else so long the spatial configuration of the two objects is geometrically correct. However, as (Coventry et al., 2001) has shown the spatial description *the umbrella is over the man* is sensitive to the protective affordances of the umbrella to stop rain, and is appropriate in contexts where, the umbrella is not in a geometrically prototypical position above the man, so long as the umbrella is protecting the man from the rain.

A further complication with regard to modelling the semantics of spatial descriptions is that experimental results indicate that the contribution of geometrical and functional factors is not the same for every spatial relation (Garrod et al., 1999; Coventry et al., 2001). This experimental work shows that there is an interplay between function and ge-

ometry in the definition of spatial semantics and therefore the spatial meaning of given spatial relation is neither fully functional nor fully geometric. Rather, spatial terms can be ordered on a spectrum based on the sensitivity of their semantics to geometric or functional factors.

Given the distinction between geometric and functional factors in shaping spatial semantics, a useful analysis that would inform the design and creation of computational models of spatial semantics is *to identify the particular semantic bias (geometric/functional) that each spatial term evinces*. However, such an analysis is difficult. Native speakers do not have strong intuitions about the bias of prepositions and such bias had to be established experimentally (Coventry et al., 2001; Garrod et al., 1999) or through linguistic analysis (Herskovits, 1986, p.55).[1] Reviewing the literature on this experimental and analytic work reveals that prepositions such as *in*, *on*, *at*, *over*, *under* have been identified as being functionally biased, whereas *above*, *below*, *left of* and *right of* are geometrically biased. Other spatial relations may be somewhere in between. In this paper we will use these relations as ground-truth pointers against which our methods will be evaluated. If the method is successful, then we are able to make predictions about those relations that have not been verified for their bias experimentally. Knowing the bias of a spatial relation is useful both theoretically and practically. Theoretically, it informs us about the complexity of grounded semantics of spatial relations. In particular, it engages with the "what" and "where" debate where it has been argued that spatial relations are not only spatial (i.e. geometric) (Landau and Jackendoff, 1993; Coventry and Garrod, 2004; Landau, 2016). Practically, the procedure to estimate the bias is useful for natural language generation systems, for example in situated robotic applications that cannot be trained end-to-end. Given that a particular pair of objects can be described geometrically with several spatial relations, the knowledge of functional bias may be used as a filter, prioritising those relations that are more likely for a particular pair of objects, thereby incorporat-

ing functional knowledge. This approach to generation of spatial descriptions is therefore similar to the approach that introduces a cognitive load based hierarchy of spatial relations (Kelleher and Kruijff, 2006) or a classification-based approach that combines geometric (related to the bounding box), textual (word2vec embeddings) and visual features (final layer of a convolutional network) (Ramisa et al., 2015). The functional geometric bias of spatial relations could also be used to inform semantic parsing, for example in prepositional phrase attachment resolution (Christie et al., 2016; Delecraz et al., 2017).

Previous work has investigated metrics of the semantic bias of spatial prepositions, see (Dobnik and Kelleher, 2013, 2014). (Dobnik and Kelleher, 2013) uses (i) normalised entropy of target-landmark pairs to estimate variation of targets and landmarks per relation and (ii) log likelihood ratio to predict the strength of association of target-landmark pairs with a spatial relation and presents ranked lists of relations by the degree of argument variation or strength of the association respectively. The approach hypothesises that functionally biased relations are more selective in the kind of targets and landmarks they co-occur with. The reasoning behind this is that geometrically it is possible to relate a wider range of objects than in the case where additional functional constrains between objects are also applied. (Dobnik and Kelleher, 2014) generalises over landmarks and targets in WordNet hierarchy and estimates the generality of the types of landmark. Again, the work hypothesises that functional relations are more restricted in their choice of target and landmark objects and therefore are generally more specific in terms of the WordNet hierarchy. Both papers present results compatible with the hypotheses where the functional or geometric nature of prepositions is predicted in line with the experimental studies (Garrod et al., 1999; Coventry et al., 2001).

Sensitive to the fact that relations such as *in* and *on* not only have spatial usage but also usages that may be considered metaphoric (Steen et al., 2010), both (Dobnik and Kelleher, 2013) and (Dobnik and Kelleher, 2014) were based on an analysis of a corpus of image captions. The idea being that descriptions of images are more likely to contain spatial descriptions grounded in the image. For similar reasons, we also employ a corpus of image descriptions (larger than in the previous work).

---

[1]The discussion of Herskovits focuses on interaction of objects conceptualised as geometric shapes, for example *on:* contiguity with line or surface. The fact that the interacting objects can be conceptualised as different geometric shapes points and therefore related by a particular prepositions points to their functional nature as discussed here.

This paper adopts a similar research hypothesis to (Dobnik and Kelleher, 2014, 2013), namely that: it is possible to distinguish between functionally biased and geometrically biased spatial relations by examining the diversity of the contexts in which they occur. Defining the concept of context in terms of the *target* and *landmark* object pairs that a relation occurs within, the rationale of this hypothesis is that: geometrically biased relations are more likely to be observed in a more diverse set of contexts, compared to functionally biased relations, because the use of a geometrically biased relation only presupposes the appropriate geometric configuration whereas the use of a functionally biased relation is also constrained by object affordances or typical interactions.

However, the work presented in this paper provides a more general analytical technique based on a neural language model (Bengio et al., 2003; Mikolov et al., 2010) which is applied to a larger dataset of spatial descriptions. We use neural language models as the basic tool for our analysis because they are already commonly used to learn the syntax and semantics of words in an unsupervised way. The contribution of this paper in relation to (i) the previous analyses of geometric and functional aspects of spatial relations is that it examines whether similar predictions can be made using these more general tools of representing meaning of words and phrases; the contribution to (ii) deep learning of language and vision is that it examines to what extent highly specific world-knowledge can be extracted from a neural language model. The paper proceeds as follows: in Section 2 we describe the datasets and their processing, in Section 3 we describe the basics behind language models and the notion of perplexity, in Section 4 and 5 we present and discuss our results. We conclude in Section 6.

The code that was used to produce the datasets and results discussed in this paper can be found at: `https://github.com/GU-CLASP/functional-geometric-lm`.

2 Datasets

The Amsterdam Metaphor Corpus (Steen et al., 2010) which is based on a subsection of a BNC reveals that the spatial sense of prepositions are very rare in genres such as news, fiction and academic texts. For example, *below* only has two instances that are not labelled as a metaphor and more than 60% of fragments with *in*, *on*, and *over* are not used in their spatial sense. For this reason Dobnik and Kelleher (2013) use two image description corpora (IAPR TC-12 (Grubinger et al., 2006) and Flickr8k (Rashtchian et al., 2010)) where spatial uses of prepositions are common. They apply a dependency parser and a set of post-processing rules to extract spatial relations, target and landmark object triplets. The size of this extracted dataset is 96,749 instances and is relatively small for training a neural language model. (Kordjamshidi et al., 2017) released CLEF 2017 multimodal spatial role labelling dataset (mSpRL) which is a human annotated subset of the IAPR TC-12 Benchmark corpus for spatial relations, targets and landmarks (Kordjamshidi et al., 2011) containing 613 text files and 1,213 sentences. While this dataset could not be used to train a language model directly, a spatial role labelling classifier could be trained on it to identify spatial relations and arguments which would then be used to produce a bootstrapped dataset for training a neural language model.

Recently, Visual Genome (Krishna et al., 2017) has been released which is a crowd-source annotated corpus of 108K images which also includes annotations of *relationships* between (previously annotated) bounding boxes. Relationships are predicates that relate objects which include spatial relations (2404639, "cup on table"), verbs (2367163, "girl holding on to bear") as well as combinations of verbs and spatial relations (2317920, "woman standing on snow") and others. We use this dataset in the work reported here. Its advantage is that it contains a large number of annotated relationships but the disadvantage is that these are collected in a crowd-sourced setting and are therefore sometimes noisy but we assume these are still of better quality than those from a bootstrapped machine annotated dataset.

To extract spatial relations from the annotated relationships, we created a dictionary of their syntactic forms based on the lists of English spatial relations in Landau (1996) and Herskovits (1986). For the training data we preserve all items annotated as relationships as single tokens ("jumping_over") and we simplify some of the composite spatial relations based on our dictionary, e.g. "left of" and "to the left of" become "left" to increase the frequency of instances. This choice could have affected our results if done without careful consid-

eration. While compound variants of spatial relations have slightly different meanings, we only collapsed those relations for which we assumed this would not affect their geometric or functional bias. Furthermore, Dobnik and Kelleher (2013) show that compound relations cluster with their non-compound variants using normalised entropy of target-landmark pairs as a metric. Finally, some variation was due to the shorthand notation used by the annotators, e.g. "to left of". The reason behind keeping all relation(ships) in the training set is to train the language model on as many targets and landmarks as possible and to learn paradigmatic relations between them. We normalise all words to lowercase and remove the duplicate descriptions per image (created by different annotators). We also check for and remove instances where a spatial relation is used as an object, e.g. "chair on left". We remove instances where one of the words has fewer than 100 occurrences in the whole dataset which reduces the dataset size by 10%. We add start and end tokens to the triplets ($\langle s \rangle$ *target relation landmark* $\langle /s \rangle$) as required for training and testing a language model. The dataset is shuffled and split into 10 folds that are later used in cross-validation. In the evaluation, we take 20 samples per spatial relation from the held out data of those relations that are members of the dictionary created previously. This way the average perplexity is always calculated on the same number of samples per each relation.[2]

3 Language model and perplexity

3.1 Language model

Probabilistic language models capture the sequential properties of language or paradigmatic relations between sequences of words. Using the chain rules of probabilities they estimate the likelihood of a sequence of words:

$$P(w_{1:T}) = \sum_{t=1}^{T} P(w_{t+1}|w_{1:t}) \qquad (1)$$

Neural language models estimate probabilities by optimising parameters of a function represented in a neural architecture (Bengio et al., 2003):

$$\hat{P}(w_{t+1}|w_{1:t} = v_{k_{1:t}}) = f(v_{t-1};\Theta) = \hat{y}_t \qquad (2)$$

where Θ represents parameters of the model, f being the composition of functions within the neural network architecture, and $v_{k_{1:t}}$ the words up to time t in the sequence. The output of the function is $\hat{y}_t \in R^n$, a vector of probabilities, with each dimension representing the probability of a word in the vocabulary. The loss of a recurrent language model is the average surprisal for each batch of data (Graves et al., 2013; Mikolov et al., 2010):

$$loss(S) = -\sum_{s \in S} \sum_{t=0}^{|s|} \frac{\log(\hat{y}_t(v_{k_{t+1}}))}{|S| \times |s|} \qquad (3)$$

Note that our architecture is deliberately simple as we apply it in an experimental setting with constrained descriptions[3]. We use a Keras implementation (Chollet et al., 2015), and fit the model parameters with Adam (Kingma and Ba, 2014) with a batch size of 32 and iterations of 20 epochs. On each iteration the language model is optimised on the loss which is related to perplexity as described in the following section.

3.2 Perplexity

Instead of calculating the averages of likelihoods from Equation 1, which might get very low on long sequences of text, we use perplexity which is an exponential measure for average negative log likelihoods of the model. This solves the representation problem with floating points and large samples of data.

$$Perplexity(S,P) = 2^{E_S[-log_2(P(w_{1:T}))]} \qquad (4)$$

where $w_{1:T}$ is an instance in a sample collection S. Perplexity is often used for evaluating language models on test sets. Since language models are optimised for low perplexities[4], the perplexity of a trained model can be used as a measure of fit of the model with the samples.

4 Varying targets and landmarks

4.1 Hypotheses

As a language model encodes semantic relations between words in a sequence we therefore expect that the distinction between functional and geometric spatial relations will also be captured by

---

[2]The reason we use 20 sample is that this is also the size of the 10% test folds in the down-sampled dataset described later. In selecting 20 items for the test-set we also ensure that it contains the vocabulary in the down-sampled training folds.

[3]For more details on the architecture see Section A.1 in the supplementary material, in particular Figure 6 and Equation 5.

[4]Equation 4 is related to Equation 3 as perplexity is 2^{Loss} given a neural model as the likelihood model.

it. As functionally biased spatial relations are used in different situational contexts than geometrically biased spatial relations, we expect that a language model will capture this bias in different distributions of target and landmark objects in the forms of the perplexity of phrases. Our weak hypothesis is that the perplexity of phrases on the test set reflects the functional-geometric bias of a spatial relation (Hypothesis 1). We take the assumption that functionally-biased relations are more selective in terms of their target and landmark choice (Section 1) and consequently sequences such as `<s> target relation landmark </s>` with functional relations have a higher predictability in the dataset resulting in a lower perplexity in the language model (Hypothesis 2). Related to this hypothesis, there is a stronger hypothesis that target and landmark are predictable with a given functional spatial relation (Hypothesis 3).

4.2 Method

We train two language models as described in Section 3.1. For training and evaluation 10-fold cross-validation is used and average results are reported. We ensure that the evaluation sets contain no vocabulary not seen during the training. The language model 1 (LM1) is trained on unrestricted frequencies of instances. In training the language model 2 (LM2) we down-sample relations so that they are represented with equal frequencies. The dataset to train LM2 contains 200 instances of each possible relations while the evaluation set contains 20 instances for each spatial relation. Note that using this method some targeted spatial relations might disappear from the evaluation set as their frequency in the held-out data is too low. In addition to the requirement that the evaluation set contains no out-of-vocabulary items, the target and landmarks are included without restriction on their frequency, as they occur with these spatial relations.

4.3 Results

Figure 1 shows the estimated average perplexities of a subset of spatial relations, those that satisfy the sampling frequency requirement described in Section 4.2. Functionally and geometrically biased spatial relations as identified experimentally in the literature (Section 1) are represented with orange and blue bars respectively. There is a tendency that functionally biased relations lead to lower mean perplexity of phrases (Hypothesis
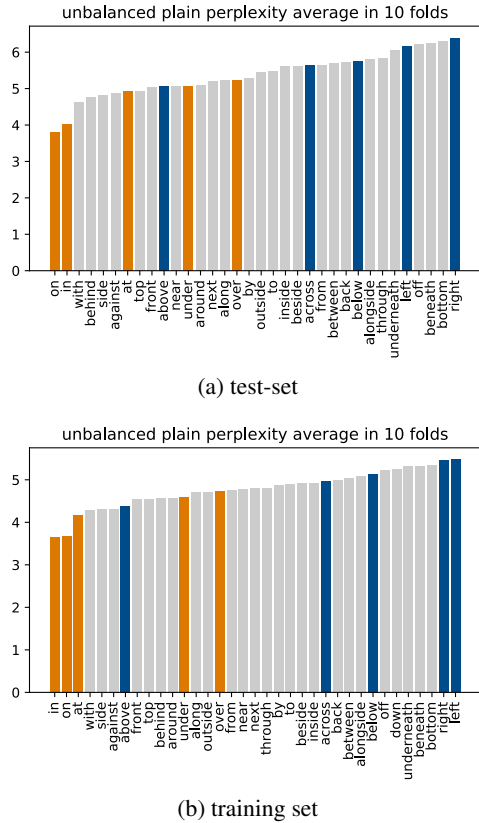

(a) test-set

(b) training set

Figure 1: Mean perplexities of spatial descriptions of LM1 (orange: functionally biased, blue: geometrically biased relations).

2 is confirmed) and also that there is a tendency that spatial relations of a particular bias cluster together (Hypothesis 1 is also confirmed). We report results both on the training set and the test set which show the same tendencies. This means that our model generalises well on the test set and that the latter is representative.

However, in the language model the perplexities are biased by the frequency of individual words: more frequent words are more likely and therefore they are associated with lower LM perplexity. The results show high Spearman's rank correlation coefficient $\rho = 0.90$ between frequencies of spatial relation in the dataset and the perplexity of the model on the test set: on (329,529) > in (108,880) > under (11,631) > above (8,952) > over (5,714) > at (4,890) > below (2,290) > across (1,230) > left (996) > right (891). For the purposes of our investigation in predictability of target-landmark pairs (Hypothesis 3) we should avoid the bias in the training set. In order to exclude the bias of frequencies of relations, we evaluate LM2 where spatial relations are presented with equal frequen-
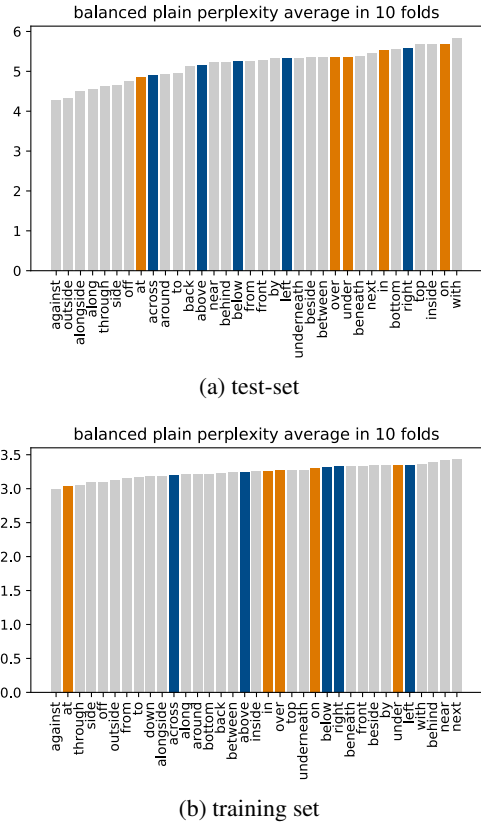
(a) test-set

(b) training set

Figure 2: Mean perplexities of LM2 by spatial relation (orange: functionally biased, blue: geometrically biased).

cies in training. Figure 2 shows the ranking of spatial relations by the perplexities when the language model was trained with balanced frequencies. The two kinds of spatial relations are less clearly separable as the colours overlap (Hypothesis 3 is not confirmed). In comparison to Figure 1 there is an observable trend that all instances lead to lower perplexities in the training set which is the effect of down-sampling on vocabulary size. Figure 2 also shows that phrases with geometrically biased spatial relations have a higher change towards lower perplexities.

Noting that the frequency of using functionally-biased spatial relations are higher in English, this bias and our strong hypothesis for predictability of target-landmark pairs can be expressed with simple joint probabilities which we are estimating with the language model:

$$P(target, relation, landmark) =$$
$$P(relation)P(target, landmark|relation)$$

It is possible that targets and landmarks that occur with these relations are very specific to these rela-

tions but infrequent with other relations. When we remove their frequency support provided by the frequency of relations these targets and landmarks become infrequent in the dataset and therefore less probable which on overall results in higher perplexities of phrases with functionally-biased relations. Specificity of targets and landmarks can be a source of these results.

To provide (some) evidence for this assumption, Figure 3 shows the average ratios of unique types over total types of targets and landmarks in the balanced dataset over 10-folds on which LM2 was trained. There is a very clear division between functionally and geometrically biased spatial relations in terms of the uniqueness of targets, functionally-biased relations are occurring with more unique ones which contributes to higher perplexity of LM2. There is less clear distinction between the two kinds of spatial relations in terms of uniqueness of landmarks. Some functional relations such as *on* occur with fewer unique landmarks than targets (from .11 to .06), some geometric relations such as *right* occur with more unique landmarks than targets (from .07 to .11). The asymmetry between targets and landmarks is expected since the choice of landmarks in the image description task is restricted by the choice of the targets (as well as other contextual factors such as visual salience). They have to be "good landmarks" to relate the targets to. A functional relation-landmark pair is more related to the target through the landmark's affordances whereas a geometric relation-landmark pair is more related to the target through geometry. This might explain for example, why *on* has fewer, but *right* has more unique landmarks than targets. On the other hand there are also relations where the ratio of unique targets and landmarks is very similar, for example *at* (.14 and .14). Overall, it appears that if uniqueness of objects is contributing to the perplexity of the language model of phrases which functionally-biased relations (which in this balanced dataset is the case) then this is more contributed by targets rather than the landmarks.

To further explore the idea of asymmetry between targets and landmarks we re-arranged the targets and landmarks in the descriptions from the balanced dataset that LM2 was trained to `<s> landmark relation target </s>` and trained LM2′. The average perplexities over 10-folds of cross-validation are shown in Figure 4.
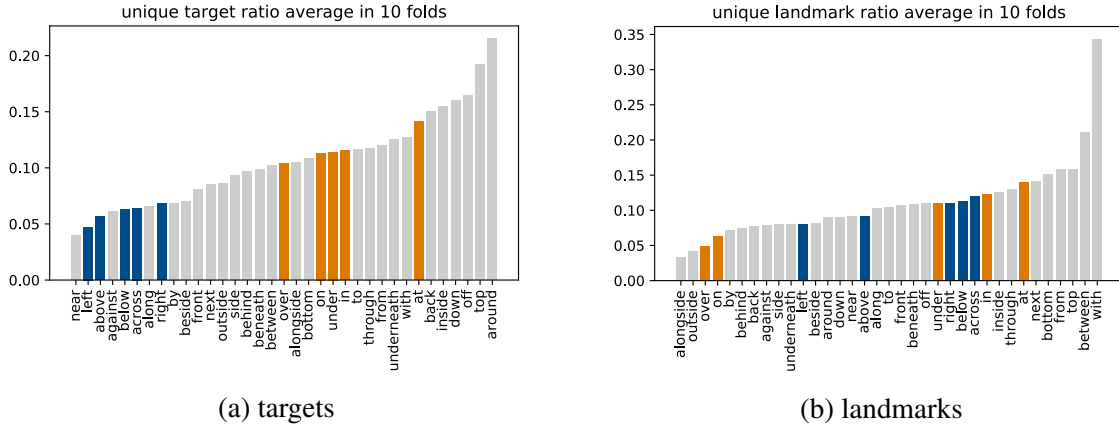
(a) targets   (b) landmarks

Figure 3: Ratio between unique types and all types per spatial relation in the balanced dataset for LM2.

Comparing Figure 4 with Figure 2 we first observe that the perplexity of LM2′ on the descriptions is overall several magnitudes lower than the perplexity of LM2 (max 0.06, max 140). Secondly, we observe that the perplexities of phrases containing different relations are very similar and that there is no separation of phrases by perplexity depending on the relation bias. The results are in line with our argument above. Knowing the landmark, it is much easier for the language model to predict the relation (of either kinds) and the target.
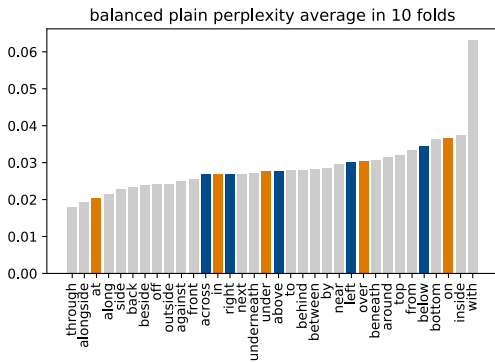

Figure 4: Mean perplexities of LM2′ by spatial relation (orange: functionally biased, blue: geometrically biased)

In conclusion, the explanation why descriptions with functionally-biased relations have a higher perplexity than descriptions with geometrically-biased descriptions appears to be twofold: (i) functionally-biased relations are more selective of their targets as expressed by the uniqueness counts, and (ii) functional relations are also more selective of their landmarks but this fact works against the performance of the language model.

As it is trained on the sequence left to right, it has to learn to predict relations only on the basis of targets which in the case of functionally-biased relations are represented by more unique tokens than geometrically-biased relations. The more informative words, the landmarks, that would enable the language model to predict a functional relation, comes last, after the relation has already been seen. The possible reason why geometrically-biased relations lead to lower perplexities of a language model on descriptions is because they have fewer unique targets. Hence, our Hypothesis 1 which linked selectivity of functionally-biased relations to low perplexity of phrases can be refuted. In spatial relations the order of the semantic interpretation of tokens (that we want to capture in these experiments) is different from the linear syntactic order of order which can be captured by the language model. When this order is changed as in LM2′ our predictions come closer to the hypothesis (Figure 4).[5]

By removing the frequency bias on spatial relations in LM2 we fix the distribution of spatial relations and examine the effect of distribution of targets and landmarks on perplexities of phrases (spatial relation as fixed context). In the following section, we fix the distributions of targets and landmarks of each spatial relation and examine the perplexity of phrases when another spatial relation is projected in this context (targets-landmarks as fixed context).

---

[5]Modulo that landmarks are, as discussed above, well-predictive of relations of both kinds.

5 Varying spatial relations

5.1 Hypotheses

Given a particular spatial relation, the distribution of targets and landmarks that occur with it creates a particular signature of targets and landmarks, the target-landmark context of a spatial relation. In this experiment, we investigate the effect on perplexity of phrases when another spatial relation is projected in such a target-landmark context. Given different selectivity of functionally- and geometrically-biased spatial relations, namely the functionally-based spatial relations are more selective of their targets and landmarks and therefore create more specific contexts, we should observe differences in perplexities of phrases when other spatial relations are projected in these contexts. In particular, we hypothesise that geometrically-biased spatial relations are more easily swappable than functionally-biased spatial relations as measured by the perplexity of a language model trained on the original, non-swapped phrases (Hypothesis 4).

5.2 Method

We use LM2 from Section 4 (trained on the balanced frequencies of spatial relations) with no additional training from the previous experiment. We group descriptions in the evaluation set by spatial relation. For each phrase containing a particular spatial relation, we replace it with every other spatial relation and estimate the perplexity of the resulting phrase using a language model. Finally, we calculate the mean of perplexities over all phrases. We use 10-fold cross-validation and report the final means across the 10 folds.

5.3 Results

Figure 5 shows a %-increase in mean complexities from those in Figure 2 when LM2 is applied on phrases with swapped relations in the contexts of the original relations. Hence, the column "at" shows the %-increase in perplexities of phrases that originally contained *at* in the validation dataset but this was replaced by all other spatial relations. Comparing with Figure 2 the estimated perplexities are higher across all cases which is predictable. There is a weak tendency that replacing functionally-biased relations with other relations leads to higher perplexities of spatial descriptions than replacing geometrically-biased relations, but the distinction is not clear cut
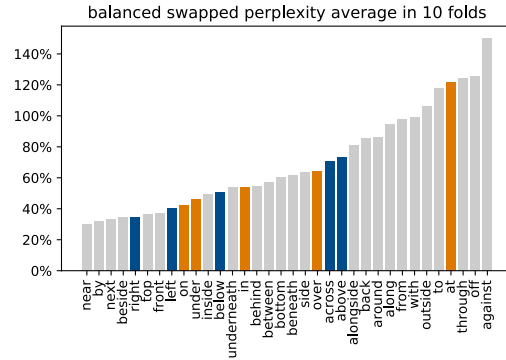

Figure 5: %-increase in perplexities of LM2 shown per context of the original preposition when swapped with another one.

(Hypothesis 4 partially confirmed). The lack of a clear distinction between two classes of descriptions confirms our previous observations about landmarks and targets: the LM has learned particular contexts for both kinds of descriptions.

6 Discussion and conclusion

We explored the degree that the functional and geometric character of spatial relations can be identified by a neural language model by focusing on spatial descriptions of controlled length and containing normalised relations. Our first question was about the implications of using a neural language model for this task. The previous research (Dobnik and Kelleher, 2013) used normalised entropy of target-landmarks per relation and log likelihood ratio between target-landmarks and relations to test this. These are focused measures that estimate the variation and the strength of association of words in a corpus. On the other hand, a language model provides a more general probabilistic representation of the entire description. As such it captures any kind of associations between words in a sequence. The other important observation is that it captures sequential relations in the direction left to right and as we have seen the sequential nature of the language model does not correspond precisely with the order in which semantic arguments of spatial relations are interpreted. However, nonetheless we can say that language models are able to capture a distinction between functional and geometric spatial relations (plus other semantic distinctions) to a similar degree of success as previously reported measures. Our initial hypothesis about the greater selectivity of spatial relations

for its arguments is correct but it is exemplified in a greater perplexity of a language model in the context of balanced spatial relations. We argued that this has to do with the fact that the targets are more unique to these relations (which is consequence of a greater specificity for arguments of functionally biased relations) and is also related to the way a sequential language model works. In the unbalanced dataset, the perplexity of the language model is reversed (it is lower with functionally biased relations) because the specificity of targets to relations is boosted with greater frequency of functionally-biased relations. The fact that functionally-biased relations are more frequent is probably related to the fact that such descriptions are more informative than purely geometric ones if available for a particular pair of objects.

We can only report tendencies based on the perplexities of our language models as our conclusions. This is because the functional-geometric bias is graded, because the predictions are highly dependent on the quality and the size of the dataset, and because other semantic relations might also be expressed by this measure. We chose a large contemporary dataset of image descriptions because we hope that it contains a high proportion of prepositions used as spatial relations. However, there is no guarantee that all prepositions in this dataset are used this way. We observe that there is considerable variation of obtained values across the 10-folds of cross-validation and we report the mean values over all folds. As an illustration, in the supplementary material (Section A.2) we give an example of graphs from two intermediary folds.

Using a language model in this task we have also learned new insights about the way language models encode spatial relations in image descriptions. It has been pointed out (cf. (Kelleher and Dobnik, 2017) among others) that convolutional neural networks with an attention model are designed to detect objects whereas spatial relations between objects are likely to be predicted by the language model. In this work we show that language models are not only predicting the relation (which is expected) but are able to distinguish between different classes of relations thus encoding finer semantic distinctions. This tells us that language models are able to encode a surprising amount of information about world knowledge with a usual caveat that it is difficult to separate

several strands of this knowledge.

The work can be extended in several ways. One way is to study dataset effects on the predicted results. Datasets with descriptions of robotic actions and instructions may be particularly promising as they focus on spatial uses. Different normalisations of spatial relations have a significant effect on the results. In this work composite spatial relations such *on the left side of* are normalised to simple spatial relations such as *left*. However, these could be treated as separate relations as difference between may exist. A more systematic examination of clusters of spatial relations would eventually tell us what other spatial relations not yet identified as functionally or geometrically biased have similar properties to those that have identified as such experimentally.

Acknowledgements

References

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

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*, 1604.02125 [cs.CV].

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 apprehension of Over, Under, Above and Below. *Journal of Memory and Language*, 44(3):376–398.

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

Simon Garrod, Gillian Ferrier, and Siobhan Campbell. 1999. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR benchmark: A new evaluation resource for visual information systems. In *Proceedings of OntoImage 2006: Workshop on language resources for content-based mage retrieval during LREC 2006*, Genoa, Italy. European Language Resources Association.

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

John D. Kelleher and Fintan J. Costello. 2005. Cognitive representations of project prepositions. In *In Proceedings of the Second ACL-Sigsem Workshop on The Linguistic Dimensions of Prepositions and their Used In Computational Linguistic Formalisems and Applications*.

John D. Kelleher and Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12 –13 June*, volume 1 of *CLASP Papers in Computational Linguistics*, pages 41–52, Gothenburg, Sweden.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 1041–1048. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017. CLEF 2017: Multimodal spatial role labeling (mSpRL) task overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 367–376, Cham. Springer International Publishing.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3):4:1–4:36.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 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(S2):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.

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.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

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, Lisbon, Portugal. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, CA. North American Chapter of the Association for Computational Linguistics (NAACL).

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

A Supplementary material
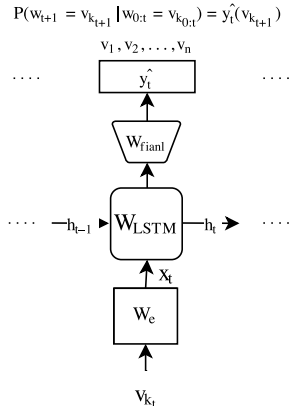
A.1 Language Model Architecture

Figure 6: The recurrent language model diagram with LSTM recurrent unit.

The neural language model architecture with the Long-Short Terms Memory (LSTM) function and its parameters, similar to tied weights in (Gal and Ghahramani, 2016):

- $W_e \in R^{n \times d}$ for word embeddings,

- $W_{LSTM} \in R^{2d \times 4d}$ for parameters of the Long-Short Term Memory function,

- $W_{Final} \in R^{d \times n}$ of the final dense layer with *softmax*.

where n is the vocabulary size for $V = \{v_1, v_2, ..., v_n\}$ and d is both the embeddings size and the memory size in LSTM. For mini-batches from training data, these parameters are being updated using a stochastic gradient descent to minimise the loss.

$$x_t = \delta_{v_{k_t}} \cdot W_e \tag{5}$$

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(\begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \cdot W_{LSTM} \right) \tag{6}$$
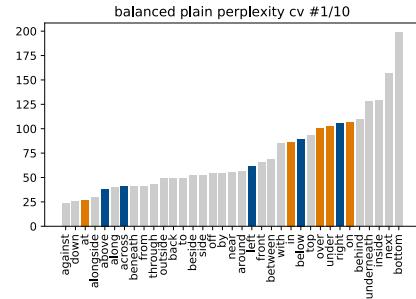
$$c_t = f \circ c_{t-1} + i \circ g \tag{7}$$
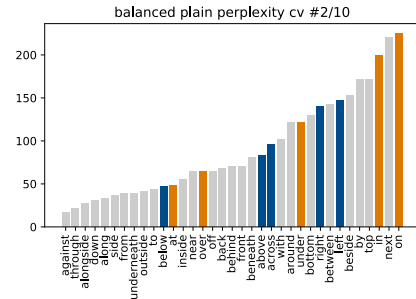
$$h_t = o \circ \tanh(c_t) \tag{8}$$

$$\hat{y}_t = \text{softmax}(h_t \cdot W_{final} + b) \tag{9}$$

where $\delta_{v_{k_t}}$ represents the one-hot encoding of the t-th word in the sequence. The x_t is the word embedding for this word, and two vectors c_t and h_t represent the states of the recurrent unit. Figure 6 illustrates the same equation.

A.2 Evaluation

(a)

(b)

Figure 7: Mean perplexities of LM2 by spatial relation for (a) folds 1 and (b) 2 (orange: functionally biased, blue: geometrically biased).