Modeling Semantics and Pragmatics of Spatial Prepositions via Hierarchical Common-Sense Primitives

Georgiy Platonov Yifei Yang Haoyu Wu Jonathan Waxman

Marcus Hill Lenhart K. Schubert

Department of Computer Science, University of Rochester
gplatono@cs.rochester.edu
{yyang99, hwu36, jwaxman2, mhill24}@u.rochester.edu
schubert@cs.rochester.edu

Abstract

Understanding spatial expressions and using them appropriately is necessary for seamless and natural human-machine interaction. However, capturing the semantics and appropriate usage of spatial prepositions is notoriously difficult, because of their vagueness and polysemy. Although modern data-driven approaches are good at capturing statistical regularities in the usage, they usually require substantial sample sizes, often do not generalize well to unseen instances and, most importantly, their structure is essentially opaque to analysis, which makes diagnosing problems and understanding their reasoning process difficult. In this work, we discuss our attempt at modeling spatial senses of prepositions in English using a combination of rule-based and statistical learning approaches. Each preposition model is implemented as a tree where each node computes certain intuitive relations associated with the preposition, with the root computing the final value of the prepositional relation itself. The models operate on a set of artificial 3D "room world" environments, designed in Blender, taking the scene itself as an input. We also discuss our annotation framework used to collect human judgments employed in the model training. Both our factored models and black-box baseline models perform quite well, but the factored models will enable reasoned explanations of spatial relation judgements.

1 Introduction

Prepositions in general and spatial prepositions in particular form a notoriously difficult lexical class because of their inherent vagueness and polysemy. Pragmatics plays crucial role in determining both which prepositions are licensed for usage in a given situation and the range of configurations (i.e., locations of the arguments) of which the licensed preposition holds true. Spatial senses of prepositions are sensitive to miscellaneous factors such as shapes and salience of the argument objects, presence of meronymy (part-of) relations, typicality, etc. *On* provides a good example of such a semantically rich preposition. When we say that one object is on another one, we strongly imply the relation of physical support between them. But support relation comes in many forms and occurs in diverse physical configurations:

a) an apple on the table
b) a book on the shelf
c) a picture on the wall
d) a fly on the ceiling
e) a shirt on the person
f) a lamp on the post
g) a fish on a hook
h) a sail on a ship

Such variety makes capturing the meaning in a computational model difficult. Yet, locative expressions involving prepositions are pervasive in natural languages and, therefore, interpretation and understanding of their meaning is important for AI, especially in use cases involving grounded human-machine interactions. Another important requirement for modern AI systems is interpretability and explainability. While neural networks can efficiently learn complex statistical distributions from large datasets, they are predominantly opaque from the common-sense analysis perspective.

Our approach to computational models for spatial prepositions is based on the following considerations. To begin with, even though the range of senses of spatial relations together with the heavy dependence on pragmatic considerations make capturing their meaning with simple mathematical criteria difficult, it is still possible to account for many of the above aspects in a principled way. People's judgments about whether a particular relation holds

in a given case can be quite variable; therefore it should suffice to provide models that estimate the probability that arbitrary judges would consider the relation to hold. This approach is aligned with a view of predicate vagueness as variability in applicability judgments (Kyburg, 2000; Lassiter and Goodman, 2017), enabling Bayesian interpretation. Next, since the usage of locative expressions is pragmatic, the ultimate success criterion in assessing models of prepositional predicates should also be pragmatic; i.e, in physical settings we often use such predicates to identify a referent (*the blue book in front of the laptop*) or to specify a goal (*put the laptop on the table*), so our models should allow a natural language system to interpret such usages as a human would.

Last, but not least, our approach facilitates explainability. Each relation is built from a combination of simpler relations, whose value can be retrieved and used to provide a justification for a particular judgement. For example, in order for one object to be next to another, they need to be close to each other and at about the same elevation. Thus, the latter criteria are included as factors in determining the value of the *next-to* relation, and their values could be used to generate meaningful explanations for any particular judgement made by the model.

In the following sections, we discuss related work, and then outline our modeling framework by examining the primitive concepts that are used as building blocks, and showing how these concepts come together in modeling a specific preposition. We then evaluate our approach in a "room world" domain, making use of Blender graphics software. We discuss two different sets of models, one purely neural network-based, implemented as a collection of multi-layer perceptrons, and another where models are implemented as trees, where each node computes a probabilistic rule. We describe our annotation framework for collecting human spatial judgments and evaluate our models. We summarize our contributions, and directions for future work, in the concluding section.

2 Related Work

In what follows, the first and second arguments of a preposition are referred to as *figure* and *ground*, respectively, when used in locative settings (Talmy, 1975).

The 3D approach to modeling spatial relations, as opposed to modeling based on 2D images, is informed by the cognitive science perspective. It is likely that people conceptualize their immediate surroundings as a 3D space defined by the three principal orientation axes of the body (Tversky et al., 1999). Moreover, 2D map-like space representations employed in navigation can be easily computed from a 3D "mental image" of the environment. It seems reasonable to assume that a potential embodied agent, such as robot, would also benefit from constructing such 3D "mental images" of its surroundings. Indoor scenarios for spatial modeling are particularly conducive to such approach (Bower and Morrow, 1990).

Developing computational models for spatial prepositions is a long-standing problem in the field of computational linguistics and NLP, and the attempts date back to the late 1960s. Early work followed mainly geometric intuitions, relying on the concepts of contiguity, surface, etc. (Cooper, 1968). A very good review of the semantic and pragmatic issues involved in spatial expressions is contained in Herskovits (1985). Herskovits' analysis identified a variety of important factors that influence correctness judgments in the application of spatial prepositions, illustrating these factors with many striking examples (e.g., the role of object types and typicality in contrasts such as *the house on the lake* vs. **the truck on the lake*, or the role of the figure/ground distinction and object size and type in contrasts like *The bicycle is near Mary's house* vs. *?Mary's house is near the bicycle*). Herskovits also proposed various abstract principles constraining the meaning and use of spatial prepositions. Our work borrows many of the elements of Herskovits' analysis, but is more narrowly focused on application to a particular setting (the room world), and is distinguished by our emphasis on developing computational models capable of actually evaluating the truth of prepositional relations in the chosen domain.

A number of methodologies rooted in application of topological notions to defining semantics of spatial prepositions arose aiming at spatial reasoning using abstract qualitative primitives to encode relations between objects (Cohn and Renz, 2008; Cohn, 1997). One example of such an approach is the Region Connection Calculus (RCC) and its modifications (Chen et al., 2015; Li and Ying, 2004). At the heart of RCC lies the notion of connectedness. Two nonempty regions are con-

nected if and only if their topological closures have a nonempty intersection. Starting with this primitive, one may proceed to define more useful spatial relations such as part-of (*x* is a part of *y* if every object that is connected to *x* is also connected to *y*) and overlapping (*x* and *y* overlap if there is a *z* that is a part of both *x* and *y*). Continuing in the same fashion one can define several other topological notions and then use them to describe spatial configurations of objects. While mathematically appealing and facilitating rigorous inference, these qualitative methods are too strict and unable to capture the semantic richness of natural language descriptions of spatial configurations of objects, since they neglect aspects such as orientation, size, shape, and argument types.

Conceptually, the way we define the spatial relations in our model is similar to the *spatial template* approach, discussed in Logan and Sadler (1996). This approach is based on the idea of defining a region of acceptability around the reference object that captures the typical locations of the relatum for this relation and determining how well the actual relatum fits this region. Our work is also similar in spirit and goals to the work by Bigelow et al. (2015), which combined the imagistic space representations with spatial templates and applied it to a story understanding task. In their approach, the authors used explicit Blender graphics modeling of a scene to represent the objects in question and their relative configurations. In their model, each region of acceptability is a three dimensional rectangular region (more precisely, a prism with a rectangular base) representing the set of points for which the given spatial relation holds. For example if one has a pair of two objects, A and B, and wants to determine whether A is on top of B, A is checked to determine whether it is in the region of acceptability located directly above B. Probabilistic reasoning is supported by using values from 0 to 1 to represent the portion of the relatum that falls into a particular region of acceptability.

In recent years, attempts have been made to use statistical learning models, especially deep neural networks, to learn spatial relations. The work by Bisk et al. (2018) is concerned with learning to transduce verbal instructions, e.g., *"Move the McDonald's block so it's just to the right (not touching) the Twitter block"* into block displacements in a simulated environment. This system, unlike ours, relies on deep learning and does not use high-level cognitively-motivated spatial relation models. The CLEVR dataset (Johnson et al., 2017) and its modified versions, such as (Liu et al., 2019), lays out an explicit spatial question answering challenge that has inspired a flurry of visual reasoning works, e.g., (Kottur et al., 2019) and (Mao et al., 2019), which achieves near-perfect scores on the CLEVR questions. Common shortcomings of these approaches are reliance on synthetic data of limited variety (only a few simple geometric shapes are present), two-dimensional image-based model of the world, very limited ground-truth models of spatial relations (e.g., *left* means any amount laterally to the left, regardless of depth or intervening objects, etc.), and use of domain-specific procedural formalisms for linguistic semantics.

Other noteworthy recent examples of dataset-driven work are (Chang et al., 2014) and (Yu and Siskind, 2017). The former inverts the learning problem, in a sense; the task was not to learn how to describe object relationships, but rather to automatically generate a scene based on a textual description. The latter employed models of spatial relations to locate and identify similar objects in several video streams.

We should separately mention the spatial modelling studies by Malinowski and Fritz (2014) and, especially, Collell et al. (2017), which apply deep neural networks to learning spatial templates for triplets of form (relatum, relation, referent). The latter work does this in an implicit setting, that is, it uses relations that indirectly suggest certain spatial configurations, e.g., *(person, rides, horse)*. Their model is capable not only of learning a spatial template for specific arguments but also of generalizing that template to previously unseen objects; e.g., it can infer the template for *(person, rides, elephant)*. These approaches, however, rely on the analysis of 2D images rather than attempting to model relations in an explicitly represented 3D world.

Our approach can be seen as an attempt at quantitative implementations inspired by the criteria that have been discussed in psychologically and linguistically oriented studies (Garrod et al., 1999; Herskovits, 1985; Tyler and Evans, 2003). Studies of human judgements of spatial relations show that overly formal qualitative models with sharp boundaries generally cannot do justice to the usage of locative expressions in natural settings. We previously mentioned a study (Bigelow et al., 2015) that applied 3D graphics scene modeling to a story

understanding task, allowing reasoning about the relative configuration and visibility of objects in the scene. Another example of an imagistic reasoning system was implemented as part of the planning system for the robot Ripley (Roy et al., 2004). Ripley used three-dimensional representation of its body, operator and workspace, reconstructed from two-dimensional view coming from Ripley's cameras.

Our work is very similar in spirit and execution to (Platonov and Schubert, 2018) and (Richard-Bollans et al., 2020b,a). All these studies model prepositions using specially designed 3D environments in Blender or Unity and employ similar sets of metrics to define the meaning of the prepositions. The studies by Platonov & Schubert differ from the present work in that the rules were less flexible (fewer parameters) and parameter values were hand-adjusted, while in our work we use gradient descent-based optimization to learn optimal values. The studies by Richard-Bollans *ete al.* relied on the prototype and exemplar approaches, using learning from data to estimate the prototype parameters or the exemplar configuration. Our work is, by contrast, rule-based (although one might argue that the parameters in our rules implicitly encode prototype properties). None of the prior studies explore generation of justifications for the spatial judgements.

3 Task Description

We explore spatial prepositions as applied to the so-called "room worlds" - 3D scenes depicting room interiors filled with common everyday items such as furniture, appliances, food items, etc.

The objects in the scene are designed in a particular way, so that their meronomy corresponds to that of the real objects. That is, the mesh consists of parts that are usually distinguished by people (e.g., for a chair, its seat, legs, back, ets., are separate objects that can be accessed by our system). This is useful for part-based inferences, e.g., a book is on a bookshelf when it is on one of the shelves. The objects are also annotated with other additional tags such as frontal vectors that indicate where the "front" of an object is, object type, etc. We have designed 52 scenes containing about 10-30 objects admissible for annotation as figure objects. Since our annotation task involves describing the location of a figure object in relation to other objects (grounds), objects that form the environment (walls, ceiling, floor) are not admissible as figures (however, they can be used as grounds as in *the poster is on the north wall*).

This serves as a realistic domain for evaluating spatial relations. We designed the annotation task so as to achieve a balance between obtaining a significant number of annotations and collecting some information about human preferences. In each annotation instance the annotator is presented with a screenshot of a room world scene and is asked to describe the location of a highlighted figure object. First, the annotator is to pick a single best-fitting preposition and a corresponding ground object. After that, they are to indicate all other relations that they believe to hold between the figure and the ground (if the most appropriate relation chosen was *between*, they are to indicate which relations hold between the figure and the first ground object). They are then asked to repeat the same procedure for up to two more times. The reason for such an approach is that while the second part of the annotation (choosing all the relations holding between the given and the selected objects) produces coverage of pairwise relations, many such judgements feel forced and unnatural to human annotators (during earlier explorations it was noted that vagueness of locative expressions leads to annotators overthinking when making judgements). The laws of conversational implicature predict that, in everyday usage, various locatives will not occur with uniform frequencies. When several possible prepositions are applicable, people tend to choose those prepositions that disambiguate better or imply stronger relations, e.g., *on* is preferable to *touching* or *near* even though these relations often co-occur. Hence, the first part of the annotation process allows annotators to freely choose the most natural or "obvious" options.

At the moment, because of the scarcity of data (see Table 2 for the number of collected annotations), we don't distinguish between the two annotation types when training and testing our models. In principle, one can assign different weights to different annotations to skew the model towards relying on the best-choice annotations more.

4 Model Details

We have developed two kinds of models. The first one is a series of simple multi-layer perceptrons (one per each relation), and the second is our main rule-based model, which is implemented as a net-

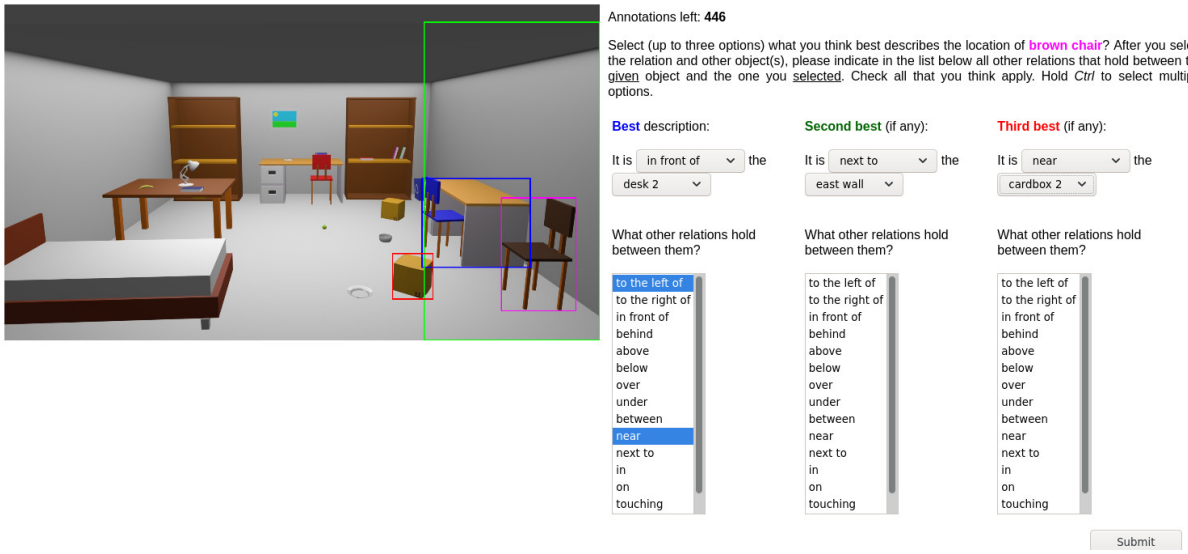

Figure 1: An example of a room world scene and the accompanying annotation controls. Best viewed in color.

work (more precisely, an arborescence) of nodes that compute meaningful hand-crafted relations used for determining the values of the prepositions. Each node realizes one or more differentiable operations which allows us to train the model using standard gradient descent-based optimization. The main reasons for developing the pure NN-based solution are to provide the baseline performance metric against which we compare our main models. Each model is essentially a binary classifier used to predict the likelihood that a particular relation holds between given objects.

4.1 Neural Baseline

Our baseline model consists of a number of independent binary classifiers (one for each spatial relation) and employs a 2-hidden layer architecture for each network. The baseline models take figure and ground objects' centroids, bounding boxes, and frontal vectors as input features. For each relation we iteratively tested different hidden layer structures in the 15-36 units range and selected one that performs the best (on average, across 5 randomized re-runs). We chose SELU activations (Klambauer et al., 2017) in the hidden layers and the logistic sigmoid function as an output non-linearity, which was the best combination based on our empirical exploration. We used binary cross-entropy as the standard binary classification loss. The model was trained using the PyTorch stochastic gradient descent optimizer with learning rate $\eta = 0.003$ and momentum $\alpha = 0.9$. We experimented with different regularization terms, but didn't notice any consistent performance gains (probably due to the small size of our networks and dataset). Main reason for the simplicity of the neural baseline is the small size of the dataset of annotations (under 7000 in total).

4.2 Rule-Based Model

We rely on a soft rule-based approach and imagistic scene representation for computing spatial relations. Each spatial preposition is implemented as a binary or ternary probabilistic predicate computed hierarchically as a combination of more primitive relations that we call *factors*. These factors encode typical more basic relations that affect whether a particular spatial preposition holds. They are usually either different senses of the same preposition or they co-occur with the preposition in most/all configurations that license the usage of that preposition. The set of factors ranges from those computing geometric properties (e.g., locations, sizes, and distances) to ones computing non-geometric, or functional ones (e.g., physical properties of the relata, such as part structure, or the location of the "front" of an object). There are several combinatory rules that determine how the factors are combined to produce a composite value. Typically, the factor values are linearly combined, multiplied together, or the maximum among them is taken, depending on the relation. For example, when one object is "on" another, it is often higher than the second object, and typically supported by it. The factors that we compute represent such primitive relations that often accompany higher-level relations of "on-

ness", "above-ness", etc. A list of example factors is presented in the Table. 1 below.

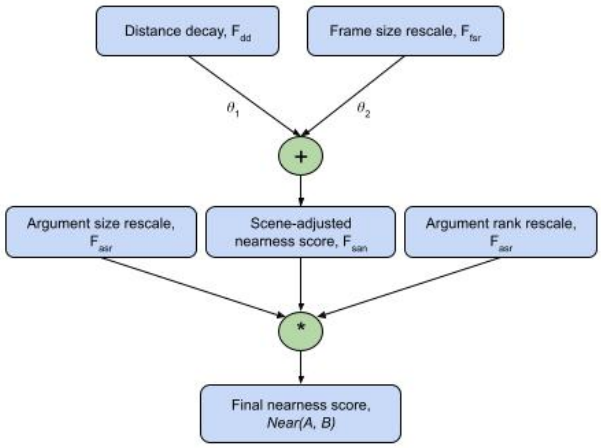

Figure 2: Structure of the factor network for *near*.

The factor tree for each relation is different, however, the general underlying principles can be understood by considering an example. One such example factor network is presented in Fig. 2. When computing *Near*(A, B), we start by computing the absolute distance, $d(A, B)$, between A and B. How this distance is computed depends on the geometry of the arguments. In the default case, assuming that both objects are roughly compact, $d(A, B)$ is simply the Euclidean distance between the centroids of A and B since, in this case, the centroid is a good approximation of the "general location" of an object. On the other hand, if, say, A or B is planar (extended in any two dimensions compared to the third, e.g., a wall, a book, a TV, etc.), linear (extended in one dimension, e.g., a pen), or generally concave (e.g., a table), then $d(A, B)$ is the minimum between the centroid distance and the distance between two closest points of A and B. We then compute scaled distance $d_{sc}(A, B)$ by dividing the absolute distance by the sum of the argument sizes, which are approximated by the radius of the circumscribed sphere. Intuitively, scaled distance provides a "size invariant" measure of the closeness of the two objects. Its value should be close to 1 when the objects are adjacent to each other, regardless of the their sizes. Next, we compute the *distance decay factor*, F_{dd}, as

$$F_{dd}(A, B) = \sigma(\theta_{dd} d_{sc}(A, B)),$$

where σ is logistic sigmoid and θ_{dd} is a learned parameter. The value of this factor gives a context-independent measure of nearness, which is then sequentially modified by a rescaling that takes into account context information. We compute the *scene-adjusted nearness*, F_{san}, as a linear combination

$$F_{san}(A, B) = \theta_1 F_{dd}(A, B) + \theta_2 F_{fsr}(A, B),$$

where $\theta_1, \theta_2 \geq 0, \theta_1 + \theta_2 = 1$, and

$$F_{fsr} = 1 - \frac{d(A, B)}{frame\_size}$$

is the *frame-size rescale factor*. The latter gives an estimate of nearness by considering the absolute distance between the objects relative to the size of the *frame*, i.e., psychologically salient part of the world. Currently, frame size is taken to be the size of the entire scene. However, in principle this can be extended to be chosen depending on argument locations, e.g., if two small objects are on top of a table, we can make the frame be the area of the table top. The final nearness score is computed as

$$Near(A, B) = F_{san}(A, B) F_{asr}(A, B) F_{arr}(A, B).$$

Here, F_{asr} is the *argument size-rescaling factor*,

$$F_{asr}(A, B) = 0.9 + 0.1 \cdot \sigma(\theta_{asr}(B.size - A.size)),$$

if $A.size > B.size$, and $F_{asr}(A, B) = 1$ otherwise. This factor encodes the intuition that, when using *near* to locate objects, the ground object is typically chosen to be bigger and fixed. Compare *?the house is near the car* vs. *the car is near the house*. Thus, when the figure is bigger than the ground we reduce the nearness score a bit, so that $Near(Bookshelf, Banana)$ returns a lower value than $Near(Banana, Bookshelf)$ (other things being equal). However, as should be clear from the formula for F_{asr}, we only allow the size difference adjustment to vary in the interval $[0.9, 1.0]$. In this way, the system would prefer to use the correct order of the arguments when making a nearness judgement on its own, while still recognizing that the relation might hold for the reverse order of the arguments.

The F_{arr} is the *argument ranking rescaling factor*. This factor lowers the nearness score if there are other objects that have a higher value of F_{san}. That is, it lowers the score in proportion to how far the current figure object is from being the best candidate figure object for a given selection of the ground and the relation. It is computed as

$$F_{arr} = e^{-\theta_{arr}(rank-1)},$$

Factor	Description
$to\_the\_right\_of\_deictic(a, b, o)$	Represents the deictic (here - viewer-specific) sense of the *to the right of* with respect to the observer o
$in\_front\_of\_intrinsic(a, b)$	Represents the intrinsic (object-centered) sense of *in front of*
$frame\_size\_rescale(a, b)$	Relative distance between a and b based on the size of the current perceptual frame
$supporting(a, b)$	Direct support relation, i.e., whether a supports b
$indirectly\_supporting(a, b)$	Indirect support relation, i.e., whether a supports some c which, in turn, supports b
$touching(a, b)$	Whether a and b are in contact with each other
$in\_direction(a, b, v)$	Computes whether b is in the general direction defined by a vector v with respect to a
$higher\_than\_centroidwise(a, b)$	Determines whether a is higher than b in terms of their centroid locations
$at\_same\_height(a, b)$	Computes whether a and b are roughly at the same elevation (in terms of centroids or their base level)

Table 1: Some of the factors used in computing spatial relations. In our system, we use the term observer to refer to the properties of the viewer, i.e., viewer location and gaze direction.

where $rank$ is the number of other objects C such that $F_{san}(C, B) > F_{san}(A, B)$.

Regarding sense ambiguity, different relations can be evaluated with respect to different coordinate frames. For example, for several projective relations, e.g., *to the right of*, we consider three cases, deictic, extrinsic and intrinsic. The so-called *deictic to the right of* is computed based on viewer's perspective. Here, one object is considered to be to the right of another, if its projection onto the viewer's visual plane is to the right of that of the latter. The *extrinsic to the right of* is based on the global coordinate system imposed by the world, i.e., front-right sides of the room. Finally, the *intrinsic to the right of* is determined based on the intrinsic coordinate system of the ground object, i.e., A is intrinsically to the right of B if it is on the right side of B. Note that not all objects have intrinsic orientations, and in these cases this sense of the relation is assigned 0. These different senses are evaluated based on the known observer properties (location and gaze direction), global orientation vectors of the world (fixed and always known), and frontal vector of an object (when applicable, i.e., the object has inherent orientation), respectively. When dealing with multiple senses, the model selects the one with the maximal value as an output.

The rule-based models are implemented as custom computational graphs using the PyTorch framework. We use binary cross-entropy loss and Adam as an optimizer, with the learning rate $\eta = 0.01$ and L2 regularization coefficient 0.1. The models are trained using back-propagation of error. Each object (3D mesh) in the scene is encapsulated in a separate Python object. It should be noted that we use these Python objects as input features, and not the numerical vectors as is common in the ML work.

5 Evaluation and Discussion

Evaluation data for both types of models are presented in Table. 2. Overall, both models performed reasonably well, apart from the cases such as *in front of, behind* and *touching* where the rule-based model performed better thanks to additional available information. The results clearly show that it is possible to produce reasonable judgments for most spatial relations even with purely geometric information. However, our main goal was to demonstrate that even when they fall short, our rule-based models still compare reasonably well with pure neural network-based approaches, with the added benefit of being interpretable thanks to their formulation in terms of meaningful decision criteria that correspond to human intuitions about spatial relations. Another important aim of our exploration was to evaluate whether the factors we selected are appropriate and sufficient for modeling the semantics of the locative senses of prepositions. While it is difficult to extrapolate our performance results to novel settings, we believe that our room worlds are representative of a significant subset of everyday settings where locative expressions are apt to be used. The annotation process is still ongoing and we are working on an additional set of scenes depicting outdoor environments. As such, the numbers in the table are subject to change, as the breadth of configurations covered and annotation data is increased. Scale differences between the two domains might affect the boundaries of applicability of the prepositions as well as their relative psychological preference ordering. Whenever possible we rely on approximations to the real 3D meshes of objects, using centroids and bounding boxes. This allows us to focus on the most salient features of objects' shapes and maintain relatively high performance. The system generates responses in real time which is relevant to the responsive-

		Pure NN model				Rule-based model			
relation	total instances	accuracy	precision	recall	F1	accuracy	precision	recall	F1
to the right of	214	0.94	1.00	0.89	0.94	0.94	0.97	0.92	0.94
to the left of	152	0.89	0.85	1.00	0.92	0.95	1.00	0.90	0.95
in front of	127	0.73	0.66	0.90	0.76	0.85	0.81	0.93	0.87
behind	97	0.76	0.68	0.91	0.78	0.86	0.80	0.91	0.85
above	74	1.00	1.00	1.00	1.00	0.90	1.00	0.85	0.92
below	86	0.82	0.92	0.80	0.85	0.87	0.97	0.78	0.87
between	220	0.96	1.00	0.93	0.96	0.95	1.00	0.87	0.93
next to	331	0.97	0.97	1.00	0.98	0.95	0.94	1.00	0.97
touching	82	0.76	0.74	0.83	0.78	0.99	1.00	0.97	0.98
near	296	0.90	0.91	0.95	0.93	0.93	0.95	0.93	0.94
on	346	0.8	0.81	0.89	0.85	0.89	0.94	0.88	0.91

Table 2: Performance statistics for the rule-based (RB) and pure MLP (NN) models. We excluded the data for *under, over* and *in*, as the number of collected annotations was insufficient. The total instances column refers to the test set instances, which constitute between 20% and 30% of all collected annotations, depending on the relation.

ness during a dialogue with the user (see the next subsection).

5.1 On Explainability

The main reason for using the rule-based approach is its interpretability. Specifically, our tree-of-factors implementation of spatial models allows backwards-generated justification of the final judgment. Each factor represents some higher-level semantic concept which can be readily translated into natural language. The tree of factors computed during the forward computation phase is preserved and is traversed in the backward direction starting from the root (representing the final output, i.e., the result of the evaluation of the preposition model). The mechanism for selecting the relevant factors for each node of the tree is as follows. If the combination rule for the current node (the way the factors of its immediate children are combined) is a product, then if the node value ≥ 0.5, return all the child nodes; otherwise, return the child node with the smallest value. If the combination rule for the children is a weighted linear combination of factor values, then if the current node value is ≥ 0.5, return the highest contributing factor node or nodes (total contribution includes their value and weight); otherwise, return the value of the node with the largest weight. Finally, if the combination rule is the max operation, then if the current node value is ≥ 0.5, return the child node with maximum value; otherwise return all the child nodes. One exception is the *touching* relation, for which the explanation procedure returns a particular part of the ground object as a justification (if the relation holds, that is). For example, the relation *Touching(Green Book, Bookshelf)* holds because the relation *Touching(Green Book, shelf 2)* holds, where *shelf 2* is part of the *Bookshelf*. In this case, relation between parts is considered a primitive, i.e., non-decomposable into more primitive relations, and so the justification process ends there. We are currently working on incorporating our models into an existing dialogue system that, given the returned factor(s), will generate an output in English. The interpretability of our models is to be evaluated in a dialogue-based setting.

As an example of the operation of the explanation procedure, consider simplified factor network for *to the right of* in Fig. 3.

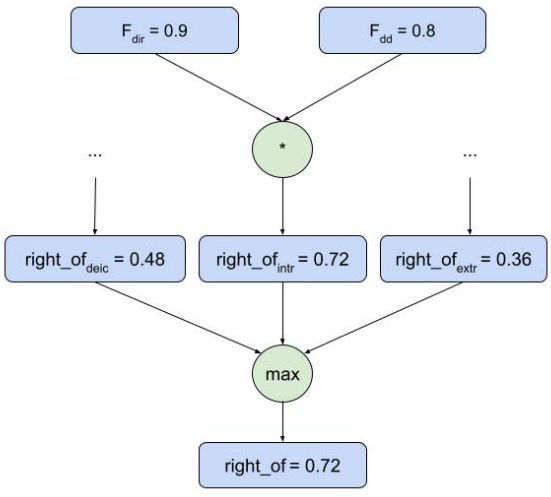

Figure 3: An example of an explanation procedure.

The numbers in the nodes are the respective values of the factors that the node computes. Assume that the system is being asked whether A is to the right of B. Assume further that the final output value is *right of = 0.72*, which corresponds to "yes". Now, if the user inquires why the system arrived at that conclusion, the following process unfolds. The node for the final score for *to the right of* takes the maximum over three values: deictic *right_of*$_{deic}$, intrinsic *right_of*$_{intr}$ and extrinsic *right_of*$_{extr}$. Since the maximum is taken, one of those nodes must be equal to the final value. Hence, the explanatory routine returns the corresponding node and its value (*right_of*$_{intr}$, 0.72). The corresponding interpretation will be (when bridging with the dialogue system is completed) something like "A is to the right of B because A is located on the right side of B". If asked further as to why the intrinsic relation holds, the system will analyze the intrinsic score's contributing factors, namely F_{dir} (directional factor that defines the "right-side" region for an object) and F_{dd} (distance decay, measuring how far apart the objects are). Since the combination rule used is multiplication and the value of the current node (intrinsic right) is 0.72 (i.e., relation holds), it follows that both factors must hold as well. The system will return the list of the nodes and their values, i.e., [(F_{dir}, 0.9), (F_{dd}, 0.8)] as a result. The straightforward interpretation of the latter would be "A is on the right side of B, because it is located in the general rightward direction w.r.t. to B and it is close enough". This process can continue until leaf nodes are reached, which do not admit further decomposition and are treated as primitives. Alternatively, assume that the value F_{dd} is only 0.4 (A is too far from B). This low value will propagate downstream and affect the intrinsic *right_of*$_{intr}$ and the final *right of* score. In this case, the system will supply a negative answer to the original question. When asked why A is not to the right of B, it will return the list of all senses [(right_of$_{deic}$, 0.48), ...] which has a straightforward interpretation of "A is not to the right of B because none of the senses apply". If queried why, say, the intrinsic sense does not apply, the system returns the lowest-value node contributing to the intrinsic sense node, i.e., [(F_{dd}, 0.4)], which translates into "A is too far from B to be on its right side".

Note the contrast with standard approaches to explainability in deep neural networks (e.g., modular neural networks), where the model can usually only answer "what" questions about its decisions (i.e., we know what kind of thing a module computes), but not the "why" or "how" questions about the reasons a given module arrived at a particular output.

6 Conclusion

We considered the problem of designing intuitive computational models of spatial prepositions that combine geometrical information as well as some pieces of commonsense knowledge and contextual information about the arguments. Our main aim was to develop spatial semantic models that rely on psychologically plausible criteria and facilitate justification of spatial judgements produced by the models, and to compare such an approach against a more mainstream black box statistical learning architecture acting as a baseline. We believe that combining the power of data-driven methods and interpretable, algorithmic models is the way forward in AI in general and, in particular, is necessary in order to incorporate context and background knowledge information needed to model spatial expressions properly. This work is one step in that direction.

References

Eric Bigelow, Daniel Scarafoni, Lenhart Schubert, and Alex Wilson. 2015. On the need for imagistic modeling in story understanding. *Biologically Inspired Cognitive Architectures*, 11:22–28.

Yonatan Bisk, Kevin J Shih, Yejin Choi, and Daniel Marcu. 2018. Learning interpretable spatial operations in a rich 3d blocks world. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gordon H Bower and Daniel G Morrow. 1990. Mental models in narrative comprehension. *Science*, 247(4938):44–48.

Angel Chang, Manolis Savva, and Christopher D Manning. 2014. Learning spatial knowledge for text to 3d scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2028–2038.

Juan Chen, Anthony G Cohn, Dayou Liu, Shengsheng Wang, Jihong Ouyang, and Qiangyuan Yu. 2015. A survey of qualitative spatial representations. *The Knowledge Engineering Review*, 30(01):106–136.

Anthony G Cohn. 1997. Qualitative spatial representation and reasoning techniques. In *KI-97: Advances in Artificial Intelligence*, pages 1–30. Springer.

Anthony G Cohn and Jochen Renz. 2008. Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, page 551.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2017. Acquiring common sense spatial knowledge through implicit spatial templates. *arXiv preprint arXiv:1711.06821*.

Gloria S Cooper. 1968. A semantic analysis of english locative prepositions. Technical report, BOLT BERANEK AND NEWMAN INC CAMBRIDGE MA.

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

Annette Herskovits. 1985. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9(3):341–378.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. *arXiv preprint arXiv:1706.02515*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

Alice Kyburg. 2000. When vague sentences inform: a model of assertability. *Synthese*, 124:175–191.

Daniel Lassiter and Noah D. Goodman. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194:3801–3836.

Sanjiang Li and Mingsheng Ying. 2004. Generalized region connection calculus. *Artificial Intelligence*, 160(1):1–34.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4185–4194.

Gordon D Logan and Daniel D Sadler. 1996. A computational analysis of the apprehension of spatial relations. *Language and space*.

Mateusz Malinowski and Mario Fritz. 2014. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv preprint arXiv:1411.5190*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

Georgiy Platonov and Lenhart Schubert. 2018. Computational models for spatial prepositions. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 21–30.

Adam Richard-Bollans, Lucía Gómez Álvarez, and Anthony G Cohn. 2020a. Modelling the polysemy of spatial prepositions in referring expressions. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 703–712.

Adam Richard-Bollans, Brandon Bennett, and A Cohn. 2020b. Automatic generation of typicality measures for spatial language in grounded settings. In *European Conference on Artificial Intelligence*. Leeds.

Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. 2004. Mental imagery for a conversational robot. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(3):1374–1383.

Leonard Talmy. 1975. Figure and ground in complex sentences. In *Annual Meeting of the Berkeley Linguistics Society*, volume 1, pages 419–430.

Barbara Tversky, Julie Bauer Morrison, Nancy Franklin, and David J Bryant. 1999. Three spaces of spatial cognition. *The Professional Geographer*, 51(4):516–524.

Andrea Tyler and Vyvyan Evans. 2003. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press.

Haonan Yu and Jeffrey Mark Siskind. 2017. Sentence directed video object codiscovery. *International Journal of Computer Vision*, 124(3):312–334.