Computational Models for Spatial Prepositions

Georgiy Platonov
Department of Computer Science
University of Rochester
gplatono@cs.rochester.edu

Lenhart Schubert
Department of Computer Science
University of Rochester
schubert@cs.rochester.edu

Abstract

Developing computational models of spatial prepositions (such as *on, in, above, etc.*) is crucial for such tasks as human-machine collaboration, story understanding, and 3D model generation from descriptions. However, these prepositions are notoriously vague and ambiguous, with meanings depending on the types, shapes and sizes of entities in the argument positions, the physical and task context, and other factors. As a result truth value judgments for prepositional relations are often uncertain and variable. In this paper we treat the modeling task as calling for assignment of probabilities to such relations as a function of multiple factors, where such probabilities can be viewed as estimates of whether humans would judge the relations to hold in given circumstances. We implemented our models in a 3D blocks world and a room world in a computer graphics setting, and found that true/false judgments based on these models do not differ much more from human judgments that the latter differ from one another. However, what really matters pragmatically is not the accuracy of truth value judgments but whether, for instance, the computer models suffice for identifying objects described in terms of prepositional relations, (e.g., *the box to the left of the table*, where there are multiple boxes). For such tasks, our models achieved accuracies above 90% for most relations.

1 Introduction

Spatial prepositions are pervasive in natural languages and, therefore, interpretation and understanding of their meaning is critical to tasks involving NLP. The computational challenges are aggravated by the versatility and vagueness of these prepositions, and their sensitivity to miscellaneous factors such as shapes, sizes and salience of the relata, part-of relations, typicality, etc. *On* provides a good example of such semantically rich prepositions. When we say that one object is on another one, we strongly imply the relation of physical support between them. But support relations can be quite subtle, and can occur in diverse physical configurations:

Example 1

a. a book on a shelf,
b. a picture on a wall,
c. a shirt on a person,
d. a lamp on a post,
e. a paragraph on a printed page,
f. a fish on a hook,
g. a sail on a ship,
h. a fly on the ceiling.

In and *over* provide additional examples of semantically subtle, versatile prepositions. While it is conceivable that the diverse meanings of these prepositions are unrelated and arose from disparate communicative and historical pressures, there are strong arguments that this is not the case (Tyler and Evans, 2003). In fact, it is very likely that all or most of the different meanings associated with a preposition are based on some underlying primary meaning from which they all originated. In the case of *on*, it seems plausible that the initial meaning was essentially support by a more or less horizontal surface, which was then extended to further support relations, and metaphorized to nonspatial relations during the evolution of language. Because of this richness, it seems that no single criterion can capture all the instances where relations such as *on*, *in*, *over*, etc., hold.

However, there are three considerations that prompted us to proceed with the design of intuitive computational models of some of the most prevalent spatial prepositions: First, while no simple mathematical criterion can characterize any

Proceedings of the First International Workshop on Spatial Language Understanding (SpLU-2018), pages 21–30
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

one of these relations, we can identify prototypical cases where the relations hold, and by considering such cases one by one, we can also zero in on non-geometrical factors that affect "truth" judgments in these cases. Second, people's judgments about whether a prepositional 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. And third, the ultimate success criterion in assessing models of prepositional predicates should 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. Our results for referent identification suggest that our current models are nearly good enough for such purposes in various "blocks world" and "room world" configurations.

In developing a conceptual framework for modeling several common prepositional relations, we tried to achieve a trade-off: On one hand, we tried to avoid overcomplicating the model, keeping the number of primitive concepts used in the framework to a minimum. On the other, we strove to make the framework general enough to cover a wide range of objects and configurations.

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 two test domains, a blocks world and a "room world", making use of Blender graphics software. We show that our computational models judge the chosen prepositional relations accurately enough in both worlds to enable rather good referent identification in relation to independent human judgments. We summarize our contributions, and directions for future work, in the concluding section.

2 Related work

Understanding the essence of the spatial prepositions is a major, long-standing task from NLP, linguistic and cognitive science perspectives. Attempts to develop a computational model for spatial prepositions date back to the late 60s. The earliest attempts followed mainly geometric intuitions, relying on the concepts of contiguity, surface, etc. (Cooper, 1968). However, an impressively thorough study emerged in the 80s (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. Compared to her study, our work is more narrowly focused on a few prepositions and two kinds of "worlds", but is distinguished by our emphasis on developing a computational model capable of actually evaluating the truth of prepositional relations in the chosen worlds.

A quite distinctive approach based on topology arose in 90s. A number of methodologies rooted in this idea were aimed 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 connected 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 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.

It is no surprise that a significant amount of research on locative expressions and spatial relations has been conducted in modern robotics. Using natural language is the most efficient way to issue a command to robots, and since they have to operate in the physical world, understanding the way humans describe space is crucial. Current state-of-the-art approaches to grounding natural language commands in general, and spatial commands in particular, are based on probabilistic graphical models (PGM) such as *Generalized Grounding Graphs* (G^3) (Tellex et al., 2011) and *Distributed Correspondence Graphs* (DCG) (Howard et al., 2014) and their modifications (Broad et al., 2016; Paul et al., 2016; Boteanu et al., 2016; Chung et al., 2015).

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 graphics modeling of a scene using Blender 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. Noteworthy examples are Bisk et al. (2017) and Chang et al. (2014). The first study was dedicated to learning spatial prepositions from images with accompanying textual annotation data within a blocks world domain. The experimental task was based on a series of images showing step-by-step construction of various structures on a table. Any two consecutive images differed in one block movement, and each image was paired with a textual description of that change. A deep neural architecture was used to pair the spatial descriptions with movements and positions of blocks in the images. The second study in a sense inverted the learning problem; the task was not to learn how to describe object relationships, but rather to automatically generate a scene based on a textual description. As such the work revisits well-studied terrain (Coyne and Sproat, 2001). Another recent study in this area is Yu and Siskind (2017), wherein spatial relation models are used 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 suggests 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.

3 Proposed Model[1]

Here we describe an example of our models for spatial prepositions as well as some of the underlying concepts and intuitions. The factors that contribute to the semantics of the prepositions can be divided into geometric and non-geometric ones. Geometric factors are relatively straightforward; they include locations, sizes and distances. Non-geometric factors include background knowledge about the relata—their physical properties, roles, the way we interact with them—as well as the perceived "frame" and the presence and characteristics of other objects within that frame.

We use a 3D modeling approach in our work. Thus geometric factors can be directly inferred from the coordinates of the polygonal meshes

[1] The implementation and all the accompanying data can be found at https://github.com/gplatono/SRP/tree/master/blender_project

comprising the object's model. We add additional geometric and non-geometric knowledge about the objects by manually attaching labels or tags to the meshes. Our approach is a rule-based one. Each spatial relation takes two (or three, in case of *between*) arguments and applies a sequence of metrics evaluating various criteria, such as distance, whether the objects are in contact, whether they possess certain properties, etc. Each metric returns a real number from $[0, 1]$. Where these metrics represent contributing factors to a relation, they are usually combined linearly into a normalized compound metric, with weights representing the importance of the factors. In some cases two factors are multiplied together, so that each scales the other. For relations with multiple prototypes, the final metric is just the maximum, i.e., we pick the best match.

Whenever possible we rely on approximations to the real 3D meshes of objects, using centroids and bounding boxes (smallest rectangular regions encompassing the objects). There are two main reasons for that. First, we are trying to achieve near real-time performance. Second, in many circumstances, given the object shapes and distances between them, the approximations yield acceptable results. Among the basic geometric primitives used in our models are various distances, scaled by object dimensions, e.g., scaled centroid distance (SCD):

$$SCD(A, B) = \frac{d(Centroid(A), Centroid(B))}{Radius(A) + Radius(B)}.$$

Here d is just the Euclidean distance and $Radius$ gives the radius of the sphere, circumscribed around the object. Given two ideally-shaped objects (cubes or spheres) the scaled distance between them will be equal to 1 exactly when they are touching each other. This is a useful measure if the objects are convex or located relatively far apart.

We also introduce similar metrics for certain types of objects that are not compact, i.e., poorly approximated by a sphere. For example, "the chair is near the wall" doesn't mean that the chair is close to the geometric center of the wall. In this case it makes more sense to measure the distance between the center of the chair and the plane of the wall. We use the labels "planar" and "rod" to mark regularly shaped non-compact objects such as walls and pencils, and introduce special distance metrics for these categories. In certain cases, when an object is very irregular or if high precision is required (e.g., when determining if two objects are touching each other) we compute pairwise vertex-to-vertex distances between two meshes.

Another important geometric primitive is an infinite conic region, defined at a vertex by an orientation vector and the angular width of the cone. This primitive is used in computing so-called projective prepositions, such as *above*, etc. This is similar to the idea of an acceptance area in (Bigelow et al., 2015). Also, for prepositions like *to the left/right of*, whose value depends on the observer's vantage point, we project the arguments' meshes onto the observer's visual plane (orthogonal to its frontal or "view" vector) and then work with 2D data, either bounding boxes or entire mesh projections.

One example of non-geometric knowledge that we use is meronymy (part-whole relationships). This knowledge is crucial for dealing with synechdoche, as in "the book is on a bookshelf". In such a case we don't usually mean that the book is directly on the bookshelf (however, this might be the case in certain contexts), but rather that it is located on one the shelves of the bookshelf. Also, knowing about parts is not enough since many real-life objects have multiple parts but we usually interact with just some of them. For example "a magnet on the frigde" will probably be used to designate a situation where the magnet is attached to the fridge's door rather than stuck on the fridge's top surface. Thus, typical interactions affect the salience of different parts and aspects of objects. In our models we mark such salient parts of an object with a special tag.

As noted earlier, the semantics of spatial prepositions does not just depend on their arguments; the perceived frame or scale and the statistics of objects in the vicinity are additional important factors. For some prepositions we first compute the raw value (between 0 and 1) representing the context-independent value of that preposition's metric. That metric is then modified by scaling it up or down depending on the values of this same metric for other objects in the scene. For example, suppose that the raw nearness metric $near\_raw(A, B)$ for two objects A and B is 0.55 out of 1.0. This reflects the fact that without further context, this is an ambiguous situation. However, if B is the closest object to A, i.e.,

$near\_raw(C, A) < 0.55, \forall C(C \neq B)$, we can say that B is *relatively* near A. In this case the final score $near(A, B)$ will be boosted by a small amount (depending on the distribution of the objects in the scene), which will make a more definite judgment possible.

Finally, let's consider the relation *on* as an example, where multiple simple metrics come together. As noted in Example 1, there are many possible configurations that can be described using *on*. Based on these configurations we can discern several stereotypical scenarios, or prototypes, and introduce special rules, each covering one such prototype. For *on* such prototypes include cases where one object is in contact with the upper surface of another; where it is attached to the salient surface of another; where it is part of a group of objects (i.e., stack), such that this whole group is on the second object; etc. We can describe *on* as (partially) depicted in algorithm 1 below.

Algorithm 1 On (The notation <3D-vector>.z refers to the vertical component)

```
procedure On(A,B)
    $on \leftarrow 0.5 * ((Above(A, B) + Touching(A, B))$
    if $Planar(B)$ and $Larger(B, A)$ and $centroid(A).z > 0.5 * dimensions(A).z$ then
        $on \leftarrow max(on, Touching(A, B))$
    ...
    for $C$ in $B$ do
        if $WorkingPart(C)$ then
            $on \leftarrow max(on, On(A, C))$
    for $C$ in $Scene \setminus \{A, B\}$ do
        if $On(C, B) > 0.5$ and $\neg Salient(C)$ then
            $on \leftarrow max(on, 0.95 * On(A, C) * On(C, B))$
```

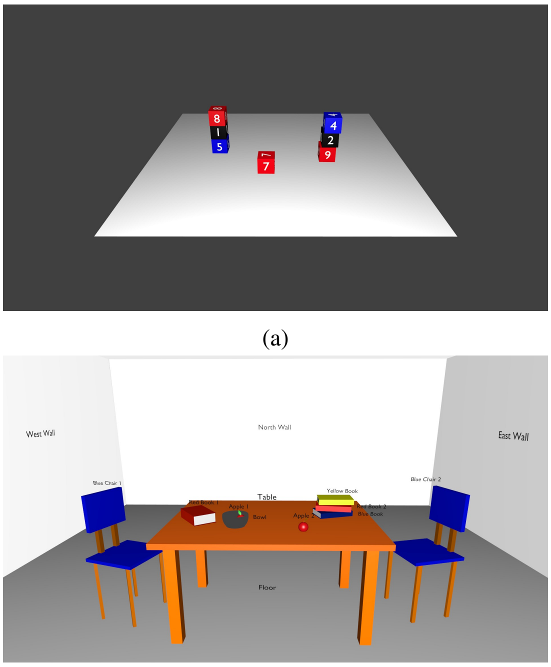
As can be seen, we compute *on* by consecutively applying different rules, corresponding to the aforementioned prototypes, and taking the best fit, i.e., the one whose metric has the maximum value. The first rule captures the canonical scenario where an object is directly above another and in contact with it. The next rule applies to situations where an object is in contact with another, bigger, planar object, such as a wall. In addition, the object should be well above the ground, so we require its centroid to be located higher than half of the object's height ($centroid(A).z > 0.5 * dimensions(A).z$). We next apply a few more rules covering such standard scenarios. We also check for the possibility of synechdoche by iterating through an object's interactive parts and checking if the relatum can be said to be on one of them. Finally, we check for transitivity: if A is on B and B is on C, then A is likely to be on C. However, the transitivity of *on* is limited. Salient objects break transitivity; e.g., if a book is on the table and the table on the floor, the book cannot be said to be on the floor. (Salience, as used here, is a static, context-independent property of an object.) Also, if there are too many intermediaries between two objects (a book on top of the stack of books, which is, in turn, on the table), the applicability of *on* decreases. This is probably due to the fact that a pile of objects becomes an increasingly salient, composite object the bigger it grows.

4 Testing domains and the annotation effort for spatial prepositions

We now describe the domains in which we tested our models as well as the experimental setup for annotating spatial configurations of objects. The annotated data serve two purposes. First, in order to measure the performance of our rule-based system, in terms of how well it captures the range of meanings of several spatial prepositions, we need to collect actual instances of human spatial judgments. Second, the collected dataset can be used in the future to teach a machine learning model the spatial relations.[2] We chose to study the spatial relations in two domains: a blocks world and a "room world". The first domain consists of a square plane with multiple colored cubical blocks on it, while the second domain represents a typical room interior, containing various everyday items, e.g., furniture, books, food, appliances, etc. The relatively simple blocks world allows us to isolate and investigate the geometric components of the meaning of a particular preposition, while the more complex room domain adds pragmatic considerations to the mix. Both domains are represented as a set of 3D scenes modeled in Blender (Blender Online Community, 2018). 3D models for the scenes were mostly created ad hoc, directly in Blender, using its standard visual modeling tools. The reason behind this is that most pub-

[2] However, while our dataset suffices for evaluating our rule-based model, it will require expansion, perhaps via crowdsourcing, for ML purposes

licly available models are designed with different purposes in mind and their part structure in incompatible with our needs. However, several models were borrowed from the public collection of models on Blend Swap (BlendSwap.com, 2018), available under the Creative Commons licence.



(a)

(b)

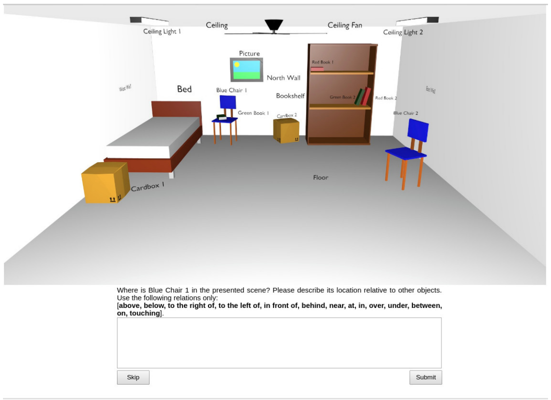
Figure 1: An example configuration for the the blocks world domain (a) and the room world domain (b)

We set up two different annotation tasks – a truth-judgment task and a description task. In the truth-judgment task, the annotator is presented with a screenshot of a scene from either domain and asked whether the particular relation holds between the given objects ("Is block 1 to the right of block 2?"). The possible qualitative response options form a Likert scale, with five items: "YES", "RATHER YES", "UNCERTAIN", "RATHER NO", "NO". In the description task, the annotator is given a screenshot of a scene from either domain and an object from that scene. The annotator is then asked to describe the object's location, in terms of a single prepositional relation to another object in the scene (or two objects, in exceptional cases like *between* or straddling two objects), so as to identify it uniquely. The annotator is encouraged to provide multiple descriptions, if there are several natural ways to pick out the object uniquely. The list of acceptable prepositions includes the following: *above, below, to the right, to the left, in front of, behind, near, at, in, over, under, between, on*, and *touching*.[3]

The objects present in the scenes were selected so as to allow for sufficiently varied configurations, combining large immovable items of furniture with multiple portable items. There was no specific plan behind the object placement in any scene, except to ensure that the target object can be uniquely described, and the overall configuration does not look unnatural or anomalous. To make unique descriptive identification of objects nontrivial, some of the objects, such as chairs or books, were presented in the scenes in several identical copies. Annotators were not allowed to directly refer to the objects by their name (every object in the scene was accompanied by a unique identifier to make it easier for the participant to locate it), but, instead, the participants were asked to use only the type and/or color of the objects when referring to them. Examples of acceptable descriptions include "to the left of another black block", "between a table and a bookshelf", "at the bed", and "on two blue blocks". In order to automate the process and facilitate gathering of the dataset, an annotation tool was deployed online and about a dozen volunteers (grad and undergrad students from the computer science department) were asked to participate in the preliminary data collection (Fig. 2). Since the task is straightforward they received minimal training; only the restrictions on the response format (only one preposition, unique identifiability, etc.) were made clear to them.

A number of scenes were created for the proposed annotation tasks. For the description task, 151 scenes were created. For the blocks world, each scene was designed to allow three questions (identification of three different blocks), while the more context-rich room world scenes supported 7-8 questions on average. For the truth-judgment task, 192 scenes were designed, with one question per scene. These 192 scenes are comprised of four variations of 48 basic scenes. The variations are: basic scene (with just the relation argument objects present in the scene), basic scene with bigger frame size (zoomed out), basic scene

[3] These particular prepositions were chosen in part because of their naturalness for describing configurations of objects in the original domain (the blocks world) – unlike *across, around, throughout, with, etc.*, and in part by the practical need to limit the number of prepositions to be modeled while still including the most widely used ones.



Where is Blue Chair 1 in the presented scene? Please describe its location relative to other objects. Use the following relations only:
[**above, below, to the right of, to the left of, in front of, behind, near, at, in, over, under, between, on, touching**].

Figure 2: The annotation website. The instructions say ”Where is Blue Chair 1 in the presented scene? Please describe its location relative to other objects.” The instructions are followed by the list of the fourteen admissible prepositions.

with context (additional objects added), and basic scene with context and bigger frame size. The collected dataset contains approximately 3500 annotations in total, with about 1500 annotations for the truth-judgment task and 2000 for the description task. It was split into a parameter tuning part and a disjoint test set with the latter containing about 800 annotations, split approximately equally between the description and truth-judgment tasks.

5 Evaluation

The model was evaluated as follows. For the truth-judgment task, the model was used to evaluate the given relation and its arguments. Both the numerical answer provided by the model and the annotator's answer were then transformed to the ordinal scale to compute the agreement coefficient. The human responses were converted from the Likert scale “YES”, “RATHER YES”, “UNCERTAIN”, “RATHER NO”, “NO” into integers 5 to 1, respectively. The metric value generated by the model was transformed as follows: Values in $[0, 0.2)$ correspond to 1, those in $[0.2, 0.4)$ to 2, ..., those in $[0.8, 1]$ to 5. For the description task, given a human description of a target object in relation to a reference object, the model was given the reference object and relation, and was required to identify the object being described.

We used both standard and weighted versions of Cohen's Kappa as an inter-annotator agreement metric with the weighting penalty $w(i, j) = \|i - j\|$, where i and j are the ordinal conversions of the responses of human annotators and our system. The agreement values were computed as follows. First, all pairwise agreement values between annotators and between each annotator and the system were computed. Next, the corresponding averages (of human-human and human-system pairs, respectively) were found.

For the initial data set (the part used to some extent to tune the model parameters), the accuracy breakdown was as follows. For weighed Kappa, the average pairwise human-human inter-annotator agreement value was 0.717, whereas the average pairwise system-human agreement metric was 0.682. For standard Kappa, the respective values were 0.536 and 0.479.

For an independent data set used for final evaluation, the values were: human-human agreement, weighted Kappa - 0.76, human-system agreement, weighted Kappa - 0.71, human-human agreement, standard Kappa - 0.52, human-system agreement, standard Kappa - 0.49. Again, all these numbers are pairwise averages. As expected, inter-annotator agreement was not very high.[4] The somewhat lower system-human agreement is still close enough to human-human agreement to indicate the plausibility of our models. Since humans manage to identify referents perfectly well using spatial relations, despite the vagueness of these relations, the key question then was how well our models would do for such usages.

relation	total occurrences	accuracy
to the right of	210	89%
to the left of	212	94%
in front of	118	92%
behind	104	96%
above	81	99%
below	43	98%
over	29	96%
under	135	95%
between	168	93%
at	17	94%
touching	71	93%
near	196	82%
in	31	100%
on	166	90%

Table 1: Fourteen relations, together with the total occurrences within the dataset used for tuning (different annotation) and accuracy per relation.

[4] This is not a flaw to be remedied, but simply a reflection of the vagueness of the prepositional relations.

For the description task we computed the accuracy in terms of the percentage of tests with correctly identified objects. The overall system accuracy on the testing data was about 93%; while imperfect, this is an encouraging result. The detailed breakdown for separate relations is provided in Table 1.

relation	total occurrences	accuracy
to the right of	33	88%
to the left of	30	87%
in front of	24	96%
behind	25	92%
above	12	100%
below	11	100%
over	0	0%
under	33	97%
between	37	86%
at	4	100%
touching	30	93%
near	55	93%
in	7	100%
on	75	89%

Table 2: Fourteen relations, together with the total occurrences within the dataset used for final testing (different annotation) and accuracy per relation.

6 Discussion and 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. In our experiments in a blocks world and a room world, we achieved reasonable agreement with human "truth" judgments and quite good agreement in a referential description task. We are not aware of other models that achieved this level of success in comparably diverse environments.

All of the existing methods we mentioned have significant limitations; typically they deal adequately with some aspects but fall short on others. The lexical semantics models in linguistics provide the most comprehensive theory of spatial relations as they are used in language. As such they are particularly relevant to natural language processing applications. However, their biggest drawbacks (at least when they attempt to address the polysemy of the prepositions) is that they are hardly formalizable and make reference to large amounts of background knowledge about how people interact with the world. Neither handcrafting that background knowledge nor learning it automatically from data seems feasible at present. On the other hand, research aimed at precise qualitative spatial models typically puts the emphasis on providing formal frameworks that enable rigorous inference, rather than on approximating human spatial representations and judgments. Unsurprisingly, this bias results in models that are suitable for certain applications, such as navigation and autonomous problem solving, but not for human-machine interaction. A separate problem is that of reconciling qualitative and quantitative spatial models.

Computational approaches popular today mostly rely on learning the meaning of prepositions from data. While they are closer to capturing their natural usage patterns, such models are trained on limited datasets in toy tasks. The generalization capabilities of such models are questionable. In our opinion the path towards comprehensive models of spatial prepositions lies at the intersection of these two major paradigms. The core meanings can be captured by meticulous analysis of the behaviour of the prepositions, while machine learning methods can be applied to adjust the weights of various *a priori* significant factors and ultimately to learn diverse additional pragmatic factors that influence human judgments in context, but are very hard to describe explicitly.

A couple of further insights we gained are worth noting. First, as indicated by the disparity we observed between judgments of truth and identification of referents, experimental design is of utmost importance in this area. Special attention needs to be paid to ensure that the experimental task is natural and sufficiently varied; at the same time, the task should enable isolating the specific meaning aspects of particular prepositions, so that they can be modeled individually. These desiderata are not easily achieved.

Second, physics plays an important role in our understanding of spatial relations. For example, as noted at the outset, *on* is closely connected with the *support* relation; thus, a cable or a rope hanging from the ceiling and touching the table under it will probably not be considered to be on the table. This example breaks the rule-based definition of *on* that we presented above. We did not address the physical aspects of the meaning of spatial prepositions in our work. This deficiency will have to be rectified if our models of spatial prepositions are to correspond more fully to our everyday intuition.

Acknowledgments

This work was supported by DARPA grant W911NF-15-1-0542. We thank our annotation team for their contributions, and the anonymous referees for their very useful comments.

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. 2017. Learning interpretable spatial operations in a rich 3d blocks world. *arXiv preprint arXiv:1712.03463*.

Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam.

BlendSwap.com. 2018. *Blendswap.com*.

Adrian Boteanu, Thomas Howard, Jacob Arkin, and Hadas Kress-Gazit. 2016. A model for verifiable grounding and execution of complex natural language instructions. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 2649–2654. IEEE.

Alexander Broad, Jacob Arkin, Nathan Ratliff, Thomas Howard, Brenna Argall, and Distributed Correspondence Graph. 2016. Towards real-time natural language corrections for assistive robots. In *RSS Workshop on Model Learning for Human-Robot Communication*.

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.

Istvan Chung, Oron Propp, Matthew R Walter, and Thomas M Howard. 2015. On the performance of hierarchical distributed correspondence graphs for efficient symbol grounding of robot instructions. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 5247–5252. IEEE.

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. *Handbook of knowledge representation*, 3:551–596.

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, DTIC Document.

Bob Coyne and Richard Sproat. 2001. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 487–496. ACM.

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

Thomas M Howard, Stefanie Tellex, and Nicholas Roy. 2014. A natural language planner interface for mobile manipulators. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 6652–6659. IEEE.

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.

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*.

R Paul, J Arkin, N Roy, and TM Howard. 2016. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. *Proceedings of Robotics: Science and Systems (RSS), Ann Arbor, Michigan, USA*.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, volume 1, page 2.

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.