Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning

Roshanak Mirzaee
Michigan State University
mirzaee@msu.edu

Parisa Kordjamshidi
Michigan State University
kordjams@msu.edu

Abstract

Recent research shows synthetic data as a source of supervision helps pretrained language models (PLM) transfer learning to new target tasks/domains. However, this idea is less explored for spatial language. We provide two new data resources on multiple spatial language processing tasks. The first dataset is synthesized for transfer learning on spatial question answering (SQA) and spatial role labeling (SpRL). Compared to previous SQA datasets, we include a larger variety of spatial relation types and spatial expressions. Our data generation process is easily extendable with new spatial expression lexicons. The second one is a real-world SQA dataset with human-generated questions built on an existing corpus with SpRL annotations. This dataset can be used to evaluate spatial language processing models in realistic situations. We show pretraining with automatically generated data significantly improves the SOTA results on several SQA and SpRL benchmarks, particularly when the training data in the target domain is small.

1 Introduction

Understanding spatial language is important in many applications such as navigation (Zhang and Kordjamshidi, 2022; Zhang et al., 2021; Chen et al., 2019), medical domain (Datta et al., 2020; Kamel Boulos et al., 2019; Massa et al., 2015), and robotics (Venkatesh et al., 2021; Kennedy et al., 2007). However, few benchmarks have directly focused on comprehending the spatial semantics of the text. Moreover, the existing datasets are either synthetic (Mirzaee et al., 2021; Weston et al., 2015; Shi et al., 2022) or at small scale (Mirzaee et al., 2021; Kordjamshidi et al., 2017).

The *synthetic datasets* often focus on specific types of relations with a small coverage of spatial semantics needed for spatial language understanding in various domains. Figure 2 indicates the coverage of sixteen spatial relation types (in Table 1) collected from existing resources (Randell et al., 1992; Wolter, 2009; Renz and Nebel, 2007). The *human-generated datasets*, despite helping study the problem as evaluation benchmarks, are less helpful for training models that can reliably understand spatial language due to their small size (Mirzaee et al., 2021).

Three boxes called one, two and three exist in an image. Box one contains a big yellow melon and a small orange watermelon. **Box two has a small yellow apple**. A small orange apple is inside and touching this box. Box one is in box three. **Box two** is to the **south** of, **far from** and to the **west** of **box three**. A **small yellow watermelon** is **inside box three**.

Q: Is **the yellow apple** to the **west** of the **yellow watermelon**? Yes

Q: Where is **box two** relative to the **yellow watermelon**? Left, Below, Far

(a) SpARTUN - A synthetic large dataset provided as a source of supervision.

A **grey car** is parking in **front** of a **grey house with brown window frames** and **plants on** the **balcony**.

Q: Are the **plants in front of** the **car**? No

Q: Are the **plants in** the **house**? Yes

(b) ReSQ - A human-generated dataset for probing the models on realistic SQA

Figure 1: Two new datasets on SQA

In this work, we build a new synthetic dataset on SQA, called SpARTUN[1] (Fig 1a) to provide a source of supervision with broad coverage of spatial relation types and expressions[2].

[1] **Spa**tial **R**easoning and role labeling for **T**ext **UN**nderstanding

[2] We only consider explicit spatial semantics and the

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6148 - 6165
December 7-11, 2022

Formalism (General Type)	Specific value	Spatial type/Spatial value)	Expressions (e.g.)
Topological	RCC8	DC (disconnected) EC (Externally Connected) PO (Partially Overlapped) EQ (Equal) TPP (Tangential Proper Part) NTPP (Non-Tangential Proper Part) TPPI (Tangential Proper Part inverse) NTPPI (Non-Tangential Proper Part inverse)	disjoint touching overlapped equal covered by in, inside covers has
Directional	Relative	LEFT, RIGHT BELOW, ABOVE BEHIND, FRONT	left of, right of under, over behind, in front
Distance	Qualitative	Far, Near	far, close

Table 1: Spatial relation types and examples of spatial language expressions.

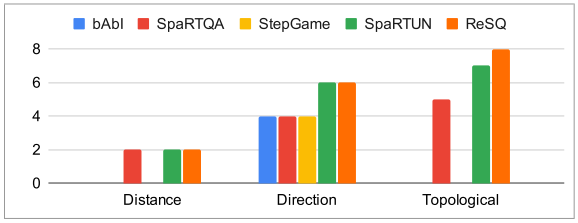

Figure 2: The comparative coverage of relation types based on Table 1 for SQA datasets.

To generate SpaRTUN, we follow the idea of SpartQA (Mirzaee et al., 2021) benchmark and generate scene graphs from a set of images. The edges in this graph yield a set of triplets such as ABOVE(blue circle, red triangle), which are used to generate a scene description (i.e., a story).

In SpaRTUN, we map the spatial relation types in triplets (e.g., ABOVE) to a variety of spatial language expressions (e.g., over, north, above) to enable the transfer learning for various data domains [3]. We also build a logical spatial reasoner to compute all possible direct and indirect spatial relations between graph nodes. Then, the questions of this dataset are selected from the indirect relations.

To evaluate the effectiveness of SpaRTUN in transfer learning, we created another dataset named ReSQ[4] (Fig 1b). This dataset is built on mSpRL (Kordjamshidi et al., 2017) corpus while we added human-generated spatial questions and answers to its real image descriptions. This dataset comparatively reflects more realistic challenges and complexities of the SQA problem.

We analyze the impact of SpaRTUN as source of extra supervision on several SQA and SpRL benchmarks. To the best of our knowledge, we are the first to use synthetic supervision for the SpRL task. Our results show that the auto-generated data successfully improves the SOTA results on mSpRL and SpartQA-Human, which are annotated for SpRL task. Moreover, further pretraining models with SpaRTUN for SQA task improves the result of previous models on ReSQ, StepGame, and SpartQA-Human benchmarks. Furthermore, studying the broad coverage of spatial relation expressions of SpaRTUN in realistic domains demonstrates that this feature is a key factor for transfer learning.

The contributions of this paper can be summarized as: **(1)** We build a new synthetic dataset to serve as a source of supervision and transfer learning for spatial language understanding tasks with broad coverage of spatial relation types and expressions (which is easily extendable); **(2)** We provide a human-generated dataset to evaluate the performance of transfer learning on real-world spatial question answering; **(3)** We evaluate the transferability of the models pretrained with SpaRTUN on multiple SQA and SpRL benchmarks and show significant improvements in SOTA results.

2 Related Research

Requiring large amounts of annotated data is a well-known issue in training complex deep neural mod-

Metaphoric usages and implicit meaning are not covered in this work.

[3]The full list of spatial expressions used in this dataset and the dataset generation code are provided in https://github.com/HLR/SpaRTUN.

[4]**Re**al-world **S**patial **Q**uestions

els (Zhu et al., 2016) that is extended to spatial language processing tasks. In our study, we noticed that all available large datasets on SQA task including bAbI (Weston et al., 2015), SPARTQA-AUTO (Mirzaee et al., 2021), and StepGame (Shi et al., 2022) are, all, synthetic.

bAbI is a simple dataset that covers a limited set of relation types, spatial rules, and vocabulary. StepGame focuses on a few relation types but with more relation expressions for each and considers multiple reasoning steps. SPARTQA-AUTO, comparatively, contains more relation types and needs complex multi-hop spatial reasoning. However, it contains a single linguistic spatial expression for each relation type. All of these datasets are created based on controlled toy settings and are not comparable with real-world spatial problems in the sense of realistic language complexity and coverage of all possible relation types. SPARTQA-HUMAN (Mirzaee et al., 2021) is a human-generated version of SPARTQA-AUTO with more spatial expressions. However, this dataset is provided for probing purposes and has a small training set that is not sufficient for effectively training deep models.

For the SPRL task, MSPRL and SpaceEval (SemEval-2015 task 8) (Pustejovsky et al., 2015) are two available datasets with spatial roles and relation annotations. These are small-scale datasets for studying the SPRL problem. From the previous works which tried transfer learning on SPRL task, (Moussa et al., 2021) only used it on word embedding of their SPRL model, and (Shin et al., 2020) used PLM without any specifically designed dataset for further pretraining. These issues motivated us to create SPARTUN for further pretraining and transfer learning for SQA and SPRL.

Transfer learning has been used effectively in different NLP tasks to further fine-tune the PLMs (Razeghi et al., 2022; Alrashdi and O'Keefe, 2020; Magge et al., 2018). Besides transfer learning, several other approaches are used to tackle the lack of training data in various NLP areas, such as providing techniques to label the unlabeled data (Enayati et al., 2021), using semi-supervised models (Van Krieken et al., 2019; Li et al., 2021) or data augmentation with synthetic data (Li et al., 2019; Min et al., 2020). However, transfer learning is a simple way of using synthetic data as an extra source of supervision at no annotation cost. Compared to the augmentation methods, the data in the transfer learning only needs to be close to the target task/domain (Ma et al., 2021) and not necessarily the same. Mirzaee et al. is the first work that considers transfer learning for SQA. It shows that training models on synthetic data and finetuning with small human-generated data results in a better performance of PLMs. However, their coverage of spatial relations and expressions is insufficient for effective transfer learning to realistic domains.

Using logical reasoning for building datasets that need complex reasoning for question answering is used before in building QA datasets (Clark et al., 2020; Saeed et al., 2021). More recent efforts even use the path of reasoning and train models to follow that (Tafjord et al., 2021). However, there are no previous works to model spatial reasoning as we do here with the broad coverage of spatial logic.

3 Transfer Learning for Spatial Language Understanding

To evaluate transfer learning on spatial language understanding, we select two main tasks, spatial question answering (SQA) and spatial role labeling (SPRL). Given the popularity of PLMs in transfer learning (Khashabi et al., 2020; Ma et al., 2021; Clark et al., 2020), we design PLM-based models for this evaluation. In the rest of this section, we describe each task and model in detail.

3.1 Spatial Question Answering

In spatial question answering, given a scene description, the task is to answer questions about the spatial relations between entities (e.g., Figure 1). Here, we focus on challenging questions that need multi-hop spatial reasoning over explicit relations. We consider two question types, YN (Yes/No) and FR(Find relations). The answer to YN is chosen from "Yes" or "No," and the answer to FR is chosen from a set of relation types.

We use a PLM with classification layers as a baseline for the SQA task. We use a binary classification layer for each label for questions with more than one valid answer and a multi-class classification layer for questions with a single valid answer. To predict the answer, we pass the concatenation of the question and story to the PLM (more detail in (Devlin et al., 2019).) The final output of $[CLS]$ token is passed to the classification layer and depending on the question type, a label or multiple labels with the highest probability are chosen as the final answer.

We train the models based on the summation of the cross-entropy losses of all binary classifiers in multi-label classification or the single cross-entropy for a single classifier in multi-classification. In the multi-label setting, we remove inconsistent answers by post-processing during the inference phase. For instance, LEFT and RIGHT relations cannot be valid answers simultaneously.

3.2 Spatial Role Labeling

Spatial role labeling (Kordjamshidi et al., 2010, 2011) is the task of identifying and classifying the spatial roles (Trajector, Landmark, and spatial indicator) and their relations. A relation is selected from the relation types in Table 1 and assigned to each triplet of (Trajector, Spatial indicator, Landmark) extracted from the sentence. We call the former **spatial role extraction** and the latter **spatial relation**[5] **extraction** (Figure 3).

Several neural models have been proposed to solve spatial role (Mazalov et al., 2015; Ludwig et al., 2016; Datta and Roberts, 2020). We take a similar approach to prior research (Shin et al., 2020) for the extraction of spatial roles (entities (Trajector/Landmark) and spatial indicators).

First, we separately tokenize each sentence in the context and use a PLM (which is BERT here) to compute the tokens representation. Next, we apply a BIO tagging layer on tokens representations using (O, B-entity, I-entity, B-indicator, I-indicator) tags. A Softmax layer on BIO tagger output is used to select the spatial entities and spatial indicators with the highest probability. For training, we use CrossEntropy loss given the spatial annotation.

For the spatial relation extraction model, similar to (Yao et al., 2019; Shin et al., 2020), we use BERT and a classification layer to extract correct triplets. Given the output of the spatial role extraction model, for each combination of $(\text{spatial entity}(tr), \text{spatial\_indicator}(sp), \text{spatial entity}(lm))$ in each sentence we create an input[6] and pass it to the BERT model. To indicate the position of each spatial role in the sentence, we use segment embeddings and add 1 if it is a role position and 0 otherwise.

The $[CLS]$ output of BERT will be passed to a one-layer MLP that provides the probability for the triplet (see Fig 3). Compared to the prior research,

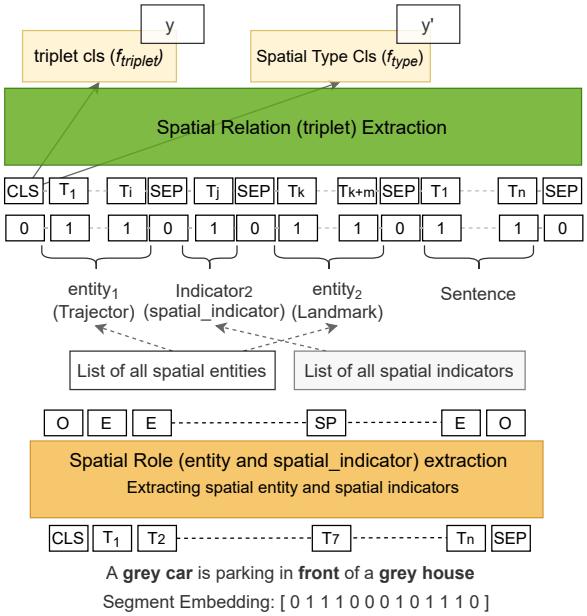

Figure 3: Spatial role labeling model includes two separately trained modules. E: entity, SP: spatial_indicators. As an example, triplet (a grey house, front , A grey car) is correct and the "spatial_type = FRONT", and (A grey car, front, a grey house) is incorrect, and the "spatial_type = NaN".

we predict the spatial type for each triplet as an auxiliary task for spatial relation extraction. To this aim, we apply another multi-class classification layer[7] on the same $[CLS]$ token. To train the model, we use a joint loss function for both relation and type modules (more detail in Appendix B).

4 SpaRTUN: Dataset Construction

To provide a source of supervision for spatial language understanding tasks, we generate a synthetic dataset with SQA format that contains SpRL annotation of sentences. We build this dataset by expanding SpartQA in multiple aspects. The following additional features are considered in creating SpaRTUN:

F1) A broad coverage of various types of spatial relations and including rules of reasoning over their combinations (e.g. $\text{NTPP}(a, b)$, $\text{LEFT}(b, c) \rightarrow \text{LEFT}(a, c)$) in various domains.

F2) A broad coverage of spatial language expressions and utterances used in various domains.

F3) Including extra annotations such as the supporting facts and number of reasoning steps for SQA to be used in complex modeling.

[5] In different works like (Kordjamshidi et al., 2010), the triplet and relation are used interchangeably.

[6] [CLS, tr, SEP, sp, SEP, lm, SEP, $sentence$, SEP]

[7] The classes are relation types in Table 1 alongside a NaN class for incorrect triplets.

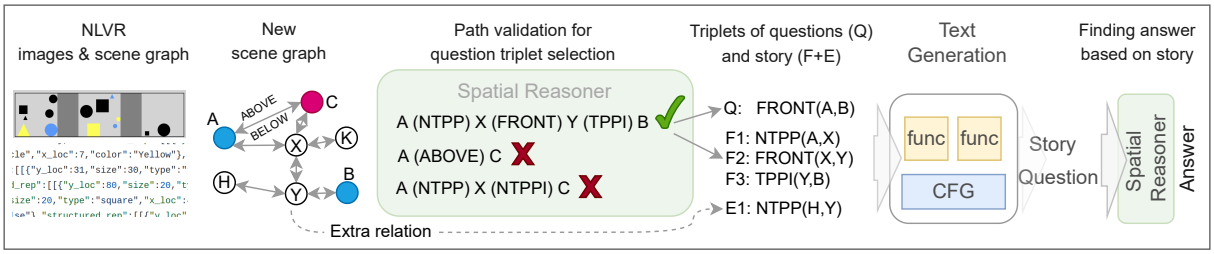

Figure 4: The data construction flow of SPARTUN. First, we generate scene graphs from NLVR images. Then a spatial reasoner validates each path between each pair of entities in this graph. All facts (F) in the selected $path$ and some extra facts (E) from the scene graph are selected as story triplets, and the start and end nodes of the $path$ are selected as question triplets. Finally, we pass all triplets to a text generation module and compute the final answer. We ignore paths with length one (e.g., $A(\ ABOVE)C$) and only keep questions that need multi-hop reasoning.

In the rest of this section, we describe the details of creating SPARTUN and the way we support the above mentioned features. Figure 4 depicts SPARTUN data construction flow.

Spatial Relation Computation. Following SPARTQA-AUTO, we use the NLVR scene graphs (Suhr et al., 2017) and compute relations between objects in each block based on their given coordinates. NLVR is limited to 2D relation types[8], therefore to add more dimensions (FRONT and BEHIND), we randomly change the LEFT and RIGHT to BEHIND and FRONT in a subset of examples. Moreover, there are no relations between blocks in NLVR descriptions.

To expand the types of relations, we extend this limitation and randomly assign relations[9] to the blocks while ensuring the spatial constraints are not violated. Then, we create a new scene graph with computed spatial relations. The nodes in this graph represent the entities (objects or blocks), and the directed edges are the spatial relations.

Question Selection. There are several paths between each pair of entities in the generated scene graph. We call a path valid if at least one relation can be inferred between its start and end nodes can be inferred. For example, in Figure 4, $\text{NTPP}(A, X),\ \text{FRONT}(X, Y),\ \text{TPPI}(Y, B)$ is valid since it results in $\text{FRONT}(A, B)$ while $\text{NTPP}(A, X),\ \text{NTPPI}(X, C)$ is not a valid path –there is no rules of reasoning that can be applied to infer new relations.

To verify the validity of each path, we pass its edges, represented as triplets in the predicate-arguments form to a logical spatial reasoner (implemented in Prolog) and query all possible relations between the pair. The number of triplets in each path represents the number of reasoning steps for inferring the relation.

We generate the question triplets from the paths with the most steps of reasoning (edges). This question will ask about the spatial relationship between the head and tail entity of the selected path. The triplets in this path are used to generate the story and are annotated as supporting facts. Additionally, the story will include additional information (extra triplets) unnecessary for answering the question to increase the complexity of the task.

Spatial Reasoner. We implement several rules (in the form of Horn clauses shown in Table 2) in Prolog, which express the logic between the relation types (described in Table 1) in various formalisms and model the logical spatial reasoning computation (see Appendix B.1). Compared to previous tools (Wolter, 2009), we are the first to include the spatial, logical computation between multiple formalisms. This reasoner validates the question/queries based on the given facts. For instance, by using the Combination rule in Table 2 over the set of facts $\{\text{NTPP}(A, X),\ \text{FRONT}(X, Y),\ \text{TPPI}(Y, B)\}$, the reasoner returns *True* for the query $\text{FRONT}(A, B)$ and *False* for $\text{FRONT}(B, A)$ or $\text{BEHIND}(A, B)$.

Text generation. The scene description is generated from the selected story triplets in question selection phase and using a publicly available context-free grammar (CFG) provided in SPARTQA-AUTO. However, we increase the variety of spatial expressions by using a vocabulary of various entity properties and relation expressions (e.g., above, over, or north for ABOVE relation type) taken from exist-

[8]The relations types included in NLVR are: DC, EC, *PP relations, LEFT, RIGHT, BELOW, and ABOVE.

[9]All relation in Table 1 except EQ

Not	$\forall(X,Y) \in Entities$	$R \in \{Dir \vee PP\}$	IF $R(X,Y)$	$\Rightarrow \text{NOT}(R\_reverse(X,Y))$
Inverse	$\forall(X,Y) \in Entities$	$R \in \{Dir \vee PP\}$	IF $R(Y,X)$	$\Rightarrow R\_reverse(X,Y)$
Symmetry	$\forall(X,Y) \in Entities$	$R \in \{Dis \vee (RCC-PP)\}$	IF $R(Y,X)$	$\Rightarrow R(X,Y)$
Transitivity	$\forall(X,Y,Z) \in Entities$	$R \in \{Dir \vee PP\}$	IF $R(X,Z), R(Z,Y)$	$\Rightarrow R(X,Y)$
Combination	$\forall(X,Y,Z,H) \in Entities$	$R \in Dir, *PP \in PP$	IF $*PP(X,Z), R(Z,H), *PPi(Z,Y)$	$\Rightarrow R(X,Y)$

Table 2: Designed spatial rules. Dir: Directional relations (e.g., LEFT), Dis: Distance relations (e.g., FAR), PP: all Proper parts relations (NTPP, NTPPI, TPPI, TPP), $RCC-PP$: All RCC8 relation except proper parts relations. $*PP$: one of TPP or NTPP. $*PPi$: one of NTPPi or TPPi.

ing resources (Freeman, 1975; Mark et al., 1989; Lockwood et al., 2006; Stock et al., 2022; Herskovits, 1986) We map the relation types and the entity properties to the lexical forms in our collected vocabulary.

For the question text, we generate the entity description and relation expression for each question triplet. The entity description is generated based on a subset of its properties in the story. For instance, an expression such as "a black object" can be generated to refer to both "a big black circle" and "a black rectangle". We generate two question types, YN (Yes/No) questions that ask whether a specific relation exists between two entities, and FR (Find Relations) questions that ask about all possible relations between them. To make YN questions more complex, we add quantifiers ("all" and "any") to the entities' descriptions.

Our text generation method can flexibly use an extended vocabulary to provide a richer corpus to supervise new target tasks when required.

Finding Answers. We search all entities in the story based on the entity descriptions (e.g., all circles, a black object) in each question and use the spatial reasoner to find the final answer.

SpRL Annotations. Along with generating the sentences for the story and questions, we automatically annotate the described spatial configurations with spatial roles and relations (trajector, landmark, spatial indicator, spatial type, triplet, entity ids). These annotations are based on a previously proposed annotation scheme of SpRL and provide free annotations for the SpRL task.

To generate SpaRTUN, we use 6.6k NLVR scene graphs for training and 1k for each dev and test set. We collect 20k training, 3k dev, and 3k test examples for each FR and YN question (see Table 3)[10]. On average, each story of SpaRTUN contains eight sentences and 91 tokens that describe on average 10 relations between different mentions of entities. More details about the dataset statistics can be seen in Appendix A.1.

5 Experimental Results

The focus of this paper is to provide a generic source of supervision for spatial language understanding tasks rather than proposing new techniques or architectures. Therefore, in the experiments, we analyze the impact of SpaRTUN on SQA and SpRL using the PLM-based models described in Section 3.

In all experiments, we compare the performance of models *fine-tuned with the target datasets* with and without *further pretraining on **syn**thetic **sup**ervision* (*SynSup*). All codes are publicly available[11]. The details of experimental settings and hyperparameters of datasets are provided in the Appendix.

5.1 Spatial Question Answering

Here, we evaluate the impact of SpaRTUN and compare it with the supervision received from other existing synthetic datasets. Since the datasets that we use contain different question types, we supervise the models based on the same question type as the target task[12].

The baselines for all experiments include a majority baseline (MB) which predicts the most repeated label as the answer to all questions, and a pretrained language model, that is, BERT here. We also report the human accuracy in answering the questions for the human-generated datasets[13]. For all experiments, to evaluate the models, we measure the accuracy which is the percentage of correct predictions in the test sets.

[10] All data are provided in the English language.: The corpus is in English.

[11] https://github.com/HLR/Spatial-QA-tasks

[12] StepGame only has FR question types. Hence, we use the model trained on FR questions for both FR and YN target tasks.

[13] All human results gathered by scoring the human answers over a subset of the test set.

5.1.1 SQA Evaluation Datasets

Dataset	Train	Dev	Test
bAbI	8992	992	992
SpartQA-Auto (YN)	26152	3860	3896
SpartQA-Auto (FR)	25744	3780	3797
SpartQA-Human (YN)	162	51	143
SpartQA-Human (FR)	149	28	77
ReSQ	1008	333	610
StepGame	50000	1000	10000
SpaRTUN (YN)	20334	3152	3193
SpaRTUN (FR)	18400	2818	2830

Table 3: Size of SQA benchmarks.

bAbI We use tasks 17 and 19 of bAbI. Task 17 is on spatial reasoning and contains binary Yes/No questions. Task 19 is on path finding and contains FR questions with answers in {LEFT, RIGHT, ABOVE, BELOW} set. The original dataset contains west, east, north, and south, which we mapped to their corresponding relative relation type.

SpartQA-Human is a small human-generated dataset containing YN and FR questions that need multi-hop spatial reasoning. The answer of YN questions is in {Yes, No,DK} where DK denotes *Do not Know* is used when the answer cannot be inferred from the context. The answer to FR questions is in {left, right, above, below, near to, far from, touching, DK}[14].

StepGame is a synthetic SQA dataset containing FR questions which need k reasoning steps to be answered ($k = 1$ to 10). The answer to each question is one relation in {left, right, below, above, lower-left, upper-right, lower-right, upper-left} set.

ReSQ We created this dataset to reflect the natural complexity of real-world spatial descriptions and questions. We asked three volunteers (English-speaking undergrad students) to generate Yes/No questions for mSpRL dataset that contains complex human-generated sentences. The questions require at least one step of reasoning. The advantage of ReSQ is that the human-generated spatial descriptions and their spatial annotations already exist in the original dataset. The statistics of this dataset are provided in Appendix A.2.

One of the challenges of the ReSQ, which is not addressed here, is that the questions require spatial commonsense knowledge in addition to capturing the spatial semantics. For example, by using commonsense knowledge from the sentence, "a lamp hanging on the ceiling", we can infer that the lamp is above all the objects in the room. To compute the human accuracy, we asked two volunteers to answer 100 questions from the test set of ReSQ and compute the accuracy.

[14]Since the relation types are not used in SpartQA, the answer is selected from a fixed set of relation expressions

Model	*SynSup*	17^{1k}	19^{500}
MB	-	51.9	10.6
BERT	-	87.39	34.53
BERT	SpartQA-A	90.42	**100**
BERT	StepGame	87.39	99.89
BERT	SpaRTUN-S	**92.43**	98.99
BERT	SpaRTUN	90.02	99.89

Table 4: Impact of using synthetic supervision on the bAbI tasks. All the models are further fine-tuned on the training set of task 17 (size = 1k) and 19 (size = 500), and test on bAbI test sets.

Model	*SynSup*	YN	FR
MB	-	53.60	24.52
BERT	-	**49.65**	18.18
BERT	SpartQA-A	39.86	48.05
BERT	StepGame	44.05	11.68
BERT	SpaRTUN-S	44.75	37.66
BERT	SpaRTUN	48.25	**50.64**
Human	-	90.69	95.23

Table 5: Transfer learning on SpartQA-Human. SpartQA-A stands for SpartQA-Auto.

5.1.2 Transfer Learning in SQA

The following experiments demonstrate the impact of transfer learning for SQA benchmarks considering different supervisions.

Due to the simplicity of **bAbI** dataset, PLM can solve this benchmark with 100% accuracy (Mirzaee et al., 2021). Hence we run our experiment on only 1k and 500 training examples of task 17 and task 19, respectively. Table 4 demonstrates the impact of synthetic supervision on both tasks of bAbI. The results with various synthetic data are fairly similar for these two tasks. However, pretraining the model with the simple version of SpaRTUN, named SpaRTUN-S, performs better than other synthetic datasets on task 17. This can be due to the fewer relation expressions in SpaRTUN-S, which follows the same structure as task 17.

In the next experiment, we investigate the impact of SpaRTUN on **SpartQA-Human** result. Comparing the results in Table 5, we find that even though the classification layer for SpartQA-

		k steps of reasoning									
Model	SynSup	1	2	3	4	5	6	7	8	9	10
TP-MANN	-	85.77	60.31	50.18	37.45	31.25	28.53	26.45	23.67	22.52	21.46
BERT	-	98.44	94.77	91.78	71.7	57.56	50.34	45.17	39.69	35.41	33.62
BERT	SPARTQA-A	98.63	94.95	91.94	77.74	68.37	61.67	57.95	50.82	46.86	44.03
BERT	SPARTUN-S	**98.70**	**95.21**	**92.46**	77.93	69.53	62.14	57.37	48.79	44.67	42.72
BERT	SPARTUN	98.55	95.02	92.04	**79.1**	**70.34**	**63.39**	**58.74**	**52.09**	**48.36**	**45.68**

Table 6: Result of models with and without extra synthetic supervision on StepGame.

Model	***SynSup***	**Accu**
MB	-	50.21
BERT	-	57.37
BERT	SPARTQA-AUTO	55.08
BERT	StepGame	60.14
BERT	SPARTUN-S	58.03
BERT	SPARTUN	**63.60**
Human	-	90.38

Table 7: Results with and without extra supervision on ReSQ. The Human accuracy is the performance of human on answering a subset of test set.

AUTO and SPARTQA-HUMAN are the same, the model trained on SPARTUN has a better transferability. It achieves 2.6% better accuracy on FR and 9% better accuracy on YN questions compared to SPARTQA-AUTO. YN is, yet, the most challenging question type in SPARTQA-HUMAN and none of the PLM-based models can reach even the simple majority baseline.

Table 6 demonstrates our experiments on **StepGame**. BERT without any extra supervision, significantly, outperforms the best reported model in Shi et al., TP-MANN, which is based on a neural memory network. As expected, all the PLM-based models almost solve the questions with one step of reasoning (i.e. where the answer directly exists in the text). However, with increasing the steps of reasoning, the performance of the models decreases. Comparing the impact of different synthetic supervision, SPARTUN achieves the best result on $k > 3$. For questions with $k <= 3$, SPARTUN-S achieves competitive similar results compared to SPARTUN. Overall, the performance gap in SPARTUN-S, SPARTQA-AUTO and SPARTUN shows that more coverage of relation expressions in SPARTUN is effective.

In the next experiment, we show the influence of SPARTUN on real-world examples, which contain more types of spatial relations and need more rules of reasoning to be solved. Table 7 shows the result of transfer learning on **RESQ**. This result shows that the limited coverage of spatial relations and expression in SPARTQA-AUTO impacts the performance of BERT negatively. However, further pretraining BERT on SPARTUN-S improves the result on RESQ. This can be due to the higher coverage of relation types in SPARTUN-S than SPARTQA-AUTO. Using SPARTUN for further pretraining BERT has the best performance and improves the result by 5.5%, indicating its advantage for transferring knowledge to solve real-world spatial challenges.

5.2 Spatial Role Labeling

Here, we analyze the influence of the extra synthetic supervision on SPRL task when evaluated on human-generated datasets. Table 8 shows the number of sentences in each SPRL benchmarks.

The pipeline model provided in Section 3, contains two main parts, a model for spatial role extraction (SRol) and a model for spatial relation extraction (SRel), which we analyze separately.

We further pretrain the BERT module in these models and then fine-tune it on the target domain. We use Macro F1-score (mean of F1 for all classes) to evaluate the performance of the SRol and SRel models.

5.2.1 SPRL Evaluation Datasets

Dataset	**Train**	**Dev**	**Test**
SPARTQA-AUTO (story)	25755	16214	16336
SPARTQA-AUTO (question)	23584	15092	15216
SPARTQA-HUMAN (story)	176	99	272
SPARTQA-HUMAN (question)	155	127	367
SPARTUN (story)	48368	7031	7191
SPARTUN (question)	38734	5970	6023
MSPRL	481	-	461

Table 8: Number of sentences of SPRL benchmarks. To train the SPARTQA-AUTO, we only use the 3k training examples (23 - 25k sentences).

Model	*SynSup*	mSpRL	SpartQA-H
R-Inf	-	80.92	-
SRol	-	**88.59**	55.8
SRol	SpartQA-A	88.41	57.28
SRol	SpaRTUN	88.03	**72.43**

Table 9: Evaluating spatial role extraction (SRol) on two mSpRL and SpartQA-Human(SpartQA-H) datasets with and without synthetic supervision.

Model	*SynSup*	mSpRL	SpartQA-H
R-Inf	-	68.78	-
SRel	-	69.12	S: 48.58 Q: 49.46
SRel	SpartQA-A	68.84	S: 58.32 Q: 55.17
SRel	SpaRTUN	**71.23**	S: **61:53** Q: **63.22**

Table 10: Spatial relation extraction (SRel) on mSpRL and SpartQA-Human(SpartQA-H) with and without synthetic supervision. Since the questions(Q) and stories(S) in SpartQA-Human have different annotations (questions have missing roles), we separately train and test this model on each.

mSpRL is a human-curated dataset provided on SpRL task. This dataset contains spatial description of real-world images and corresponding SpRL annotations (see Appendix A.6).

SpartQA-Human did not contain SpRL annotations. Hence, we asked two expert volunteers to annotate the story/questions of this dataset. Then another expert annotator checked the annotation and discarded the erroneous ones. As a result, half of this training data is annotated with SpRL tags.

5.2.2 Transfer learning in SpRL

Table 9 demonstrates the influence of synthetic supervision in spatial role extraction evaluated on mSpRL and SpartQA-Human.

We compare the result of SRol model with the previous SOTA, "R-Inf" (Manzoor and Kordjamshidi, 2018), on mSpRL dataset. R-Inf uses external multi-modal resources and global inference. All of the BERT-based SRol models outperform the R-Inf, which shows the power of PLMs for this task. However, since the accuracy of the SRol is already very high, using synthetic supervision shows no improvements compared to the model that only trained with mSpRL training set for the SRol. In contrast, on SpartQA-Human, using synthetic supervision helps the model perform better. Especially, using SpaRTUN increases the performance of the SRol model dramatically, by 15%.

In table 10, we show the result of SRel model (containing spatial relation extraction and spatial relation type classification) for spatial relation extraction, with and without extra supervision from synthetic data. Same as SRol model, extra supervision from SpaRTUN achieves the best result when tested on SpartQA-Human.

For mSpRL, we compared the SRel model with R-Inf on spatial relation extraction. As table 10 demonstrates we improve the SOTA by 2.6% on F1 measure using SpaRTUN as synthetic supervision. Also, model further pretrained on SpartQA-Auto gets lower result than model with no extra supervision due to the limited relation expressions used in this data.

In conclusion, our experiments show the efficiency of SpaRTUN in improving the performance of models on different benchmarks due to the flexible coverage of relation types and expressions.

6 Conclusion and Future Work

We created a new synthetic dataset as a source of supervision for transfer learning for spatial question answering (SQA) and spatial role labeling (SpRL) tasks. We show that expanding the coverage of relation types and combinations and spatial language expressions can provide a more robust source of supervision for pretraining and transfer learning. As a result, this data improves the models' performance in many experimental scenarios on both tasks when tested on various evaluation benchmarks. This data includes rules of spatial reasoning and the chain of logical reasoning for answering the questions that can be used for further research in the future.

Moreover, we provide a human-generated dataset on a realistic SQA task that can be used to evaluate the models and methods for spatial language understanding related tasks in real-world problems. This data is an extension of a previous benchmark on SpRL task with spatial semantic annotations. As a result, this dataset contains annotations for both SpRL and SQA tasks.

In future work, we plan to investigate explicit spatial reasoning over text by neuro-symbolic models. Moreover, using our methodology to generate synthetic spatial corpus in other languages or for other types of reasoning, such as temporal reasoning, is an exciting direction for future research.

Limitations

Though we aim for a broad coverage of relation types and relations, we collected this from our available resources and spatial lexicons but this is not by any means complete. There can be relations and expressions that are not covered. In particular, the relation expressions are limited to verbs and prepositions. The performance and reasoning ability of our models is improved with transfer learning but this is, certainly, far from the natural language understanding desiderata. Our models are based on large language models and need GPU resources to execute.

Acknowledgements

This project is supported by National Science Foundation (NSF) CAREER award 2028626 and partially supported by the Office of Naval Research (ONR) grant N00014-20-1-2005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation nor the Office of Naval Research. We thank all reviewers for their helpful comments and suggestions. We also thank Sania Sinha and Timothy Moran for their help in the human data generation and annotations.

References

Reem Alrashdi and Simon O'Keefe. 2020. *Automatic Labeling of Tweets for Crisis Response Using Distant Supervision*, page 418–425. Association for Computing Machinery, New York, NY, USA.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Surabhi Datta and Kirk Roberts. 2020. A hybrid deep learning approach for spatial trigger extraction from radiology reports. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 50. NIH Public Access.

Surabhi Datta, Yuqi Si, Laritza Rodriguez, Sonya E Shooshan, Dina Demner-Fushman, and Kirk Roberts. 2020. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of biomedical informatics*, 108:103473.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Saman Enayati, Ziyu Yang, Benjamin Lu, and Slobodan Vucetic. 2021. A visualization approach for rapid labeling of clinical notes for smoking status extraction. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 24–30, Online. Association for Computational Linguistics.

John Freeman. 1975. The modelling of spatial relations. *Computer graphics and image processing*, 4(2):156–171.

A. Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press.

Maged N Kamel Boulos, Guochao Peng, and Trang VoPham. 2019. An overview of geoai applications in health and healthcare. *International journal of health geographics*, 18(1):1–9.

William G Kennedy, Magdalena D Bugajska, Matthew Marge, William Adams, Benjamin R Fransen, Dennis Perzanowski, Alan C Schultz, and J Gregory Trafton. 2007. Spatial representation and reasoning for human-robot collaboration. In *AAAI*, volume 7, pages 1554–1559.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Parisa Kordjamshidi, Marie-Francine Moens, and Martijn van Otterlo. 2010. Spatial Role Labeling: Task definition and annotation scheme. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 413–420. European Language Resources Association (ELRA).

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 *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 367–376. Springer.

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 (TSLP)*, 8(3):1–36.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–554.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. *arXiv preprint arXiv:1909.00126*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Kate Lockwood, Ken Forbus, D Halstead, and Jeffrey Usher. 2006. Automatic categorization of spatial prepositions. In *Proceedings of the 28th annual conference of the cognitive science society*, pages 1705–1710.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Oswaldo Ludwig, Xiao Liu, Parisa Kordjamshidi, and Marie-Francine Moens. 2016. Deep embedding for spatial role labeling. *arXiv preprint arXiv:1603.08474*.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *35th AAAI Conference on Artificial Intelligence*.

Arjun Magge, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2018. Deep neural networks and distant supervision for geographic location mention extraction. *Bioinformatics*, 34(13):i565–i573.

Umar Manzoor and Parisa Kordjamshidi. 2018. Anaphora resolution for improving spatial relation extraction from text. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 53–62.

David M Mark et al. 1989. *Languages of spatial relations: Researchable questions & NCGIA research agenda*. National Center for Geographic Information and Analysis Santa Barbara ....

Wouter Massa, Parisa Kordjamshidi, Thomas Provoost, and Marie-Francine Moens. 2015. Machine reading of biological texts: bacteria-biotope extraction. In *Proceedings of the 6th international conference on bioinformatics models, methods and algorithms*, pages 55–64. SCITEPRESS.

Alexey Mazalov, Bruno Martins, and David Matos. 2015. Spatial role labeling with convolutional neural networks. In *Proceedings of the 9th Workshop on Geographic Information Retrieval*, pages 1–7.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Alaeddine Moussa, Sébastien Fournier, Khaoula Mahmoudi, Bernard Espinasse, and Sami Faiz. 2021. Spatial role labeling based on improved pre-trained word embeddings and transfer learning. *Procedia Computer Science*, 192:1218–1226.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.

David A Randell, Zhan Cui, and Anthony G Cohn. 1992. A spatial logic based on regions and connection. *KR*, 92:165–176.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*.

Jochen Renz and Bernhard Nebel. 2007. Qualitative spatial reasoning using constraint calculi. In *Handbook of spatial logics*, pages 161–215. Springer.

Mohammed Saeed, Naser Ahmadi, Preslav Nakov, and Paolo Papotti. 2021. RuleBERT: Teaching soft rules to pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1460–1476, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop

spatial reasoning in texts. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, AAAI '22.

Hyeong Jin Shin, Jeong Yeon Park, Dae Bum Yuk, and Jae Sung Lee. 2020. Bert-based spatial information extraction. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 10–17.

Kristin Stock, Christopher B Jones, Shaun Russell, Mansi Radke, Prarthana Das, and Niloofar Aflaki. 2022. Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science*, 36(3):547–584.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Emile Van Krieken, Erman Acar, and Frank Van Harmelen. 2019. Semi-supervised learning using differentiable reasoning. *arXiv preprint arXiv:1908.04700*.

Sagar Gubbi Venkatesh, Anirban Biswas, Raviteja Upadrashta, Vikram Srinivasan, Partha Talukdar, and Bharadwaj Amrutur. 2021. Spatial reasoning from natural language instructions for robot manipulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11196–11202. IEEE.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Diedrich Wolter. 2009. Sparq-a spatial reasoning toolbox. In *AAAI Spring Symposium: Benchmarking of Qualitative Spatial and Temporal Reasoning Systems*, page 53.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Yue Zhang, Quan Guo, and Parisa Kordjamshidi. 2021. Towards navigation by reasoning over spatial configurations. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 42–52, Online. Association for Computational Linguistics.

Yue Zhang and Parisa Kordjamshidi. 2022. Explicit object relation alignment for vision and language navigation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 322–331, Dublin, Ireland. Association for Computational Linguistics.

Xiangxin Zhu, Carl Vondrick, Charless C Fowlkes, and Deva Ramanan. 2016. Do we need more training data? *International Journal of Computer Vision*, 119(1):76–92.

Jordan Zlatev. 2008. Holistic spatial semantics of thai. In *Cognitive linguistics and non-Indo-European languages*, pages 305–336. De Gruyter Mouton.

A Datasets

A.1 SpaRTUN

As we described in Section 4 to cover more spatial expressions and spatial relation types, we provide an extendable vocabulary of these spatial phenomena. The entire vocabulary of supported relation expressions and entity properties are provided in Figure 10.

Statistic information: Each example in SpaRTUN contains a story that describe the spatial relation between entities and some questions which ask about indirect relations between entities. On average, each story contains eight sentences and 91 tokens, which describe ten relations on average.

We follow SpartQA for dataset split. The number of questions in each train, dev, and test sets is provided in Table 3. YN questions can have two answers "Yes," which is the answer to 54% of questions, and "No," which is the answer to 46% of questions.

FR is a question type with multiple answers. In below, you can see the percentage of existence of each relation in the whole data: { left : 10%, right:10%, above: 27%, below: 26%, behind: 19%, front: 10%, near: 2%, far: 15%, dc: 26%, ec: 7%, po: 0.2%, tpp: 2%, ntpp: 10%, tppi: 3%, and ntppi: 8% }

A.2 ReSQ

The ReSQ dataset generated over the context of mSpRL dataset. For each group of sentences (describing an image), we ask three volunteers (English-speaking undergraduate students) to generate at least four Yes/No questions. On average, they spent 20 minutes generating questions for each group of sentences which, in total, they spent 210 hours generating the whole data. After gathering the data, another undergrad student check the questions and remove the incorrect ones and keep the rest. The train set is provided on the train set of mSpRL, and since it does not have a dev set, we split the 32% of test data (equal to 20% of the training set) and keep it as the dev set. 50% of questions in this data are "Yes" and 50% are "No". The static information of this dataset comes in Table 3.

To compute the human accuracy we ask two undergraduate students, one from those who create the questions and one new volunteer to answer 100 questions from the test set of ReSQ. In the end a third students grade their answers.

A.3 bAbI

This dataset is automatically generated data including samples with two sentences describing relationships between three objects and Yes/No questions asking about the existence of a relation between two objects (Fig 5) focuses on multi-hop spatial reasoning question answering.

"The pink rectangle is below the red square. The red square is below the blue square."
1. Is the red square below the pink rectangle? No
2. Is the pink rectangle below the blue square? Yes

Figure 5: An example of bAbI

bAbI task 19, contain questions asking about the directed path from one room to another. More statistic information of this dataset comes in table 3.

A.4 SpartQA

SpartQA-Auto contains more complex textual context (story) and questions requiring complex multi-hop spatial reasoning (e.g. Fig 6). This datasets contains one large synthesized (SpartQA-Auto) and a small human-generated (SpartQA-Human) subsets.

One of the advantages of SpartQA is the SpRL annotation of whole data (Contexts and Questions) provided with the main dataset. In this work, we also recruited two experts annotator which spent 270 hours annotating 2k sentences in SpartQA-Human using WebAnno framework[15]. Then another expert annotator checks their annotation and discards the wrong ones. The statistic information of SpartQA comes in Table 3.

A.5 StepGame

StepGame is another synthesized datasets described in this paper. You can check a sample of this dataset in Figure 7.

A.6 mSpRL

SpRL is the task of identifying and classifying the spatial arguments of the spatial expressions mentioned in a sentence (Kordjamshidi et al., 2010). The mSpRL(Kordjamshidi et al., 2017) is a dataset provided on SpRL task.. The statistic data of this dataset comes in Table 11. A SpRL can have following spatial semantic component (Zlatev, 2008) on the static environment, **trajector** (the main

[15] https://webanno.github.io/webanno/

STORY:
We have three blocks, A, B and C. Block B is to the right of block C and it is below block A. Block A has two black medium squares. Medium black square number one is below medium black square number two and a medium blue square. It is touching the bottom edge of this block. The medium blue square is below medium black square number two. Block B contains one medium black square. Block C contains one medium blue square and one medium black square. The medium blue square is below the medium black square.

QUESTIONS:
FB: Which block(s) has a medium thing that is below a black square? A, B, C
FB: Which block(s) doesn't have any blue square that is to the left of a medium square? A, B
FR: What is the relation between the medium black square which is in block C and the medium square that is below a medium black square that is touching the bottom edge of a block? Left
CO: Which object is above a medium black square? the medium black square which is in block C or medium black square number two? medium black square number two
YN: Is there a square that is below medium square number two above all medium black squares that are touching the bottom edge of a block? Yes

Figure 6: An example of SPARTQA-AUTO

Story:
0:"B is south east of J."
1:"X is under E."
2:"K is to the left of Z and is on the **same horizontal plane**."
3:"If L is the **center of a clock face**, E is **located between 10 and 11**."
4:"S is positioned above Q."
5:"Q is d**iagonally to the bottom right** of L."
6:"C and S are horizontal and C is to the left of S."
7:"I is **above** B with a small gap between them."
8:"E is **above** N and **to the left** of N."
9:"Q is **below and to the right** of B."
10:"X is to the left of C with a small gap between them."

question:"What is the relation of the agent L to the agent J? "lower-right"

Figure 7: StepGame. An example of questions which need 10 steps of reasoning.

	Train	Test	All
Sentences	600	613	1213
Trajectors	716	874	1590
Landmarks	612	573	1185
Spatial Indicators	666	795	1461
Spatial Triplets	761	939	1700

Table 11: MSPRL statistics.

entity), **landmark**(the reference entity), and **spatial_indicator** (the spatial term describing the relationship between trajector and landmark.). The dynamic environment can also have *path*, *region*, *direction*, and *motion*. To understand MSPRL better you can take a look at Figure 8. In this figure the spatial value assigned to each spatial triplet can be chosen from Table 1.

The white car in the street, is in front of the blue building.

Trajector1 = The white car
Landmark1 = the street
Spatial_indicator = *in*
Genaral type = Topological (RCC8 (TPP))

TPP(the white car, the street)

Trajector1 = The white car
Landmark1 = the blue building
Spatial_indicator = *in front*
Genaral type = Directional (FRONT)

FRONT(the white car, the blue building)

Figure 8: Spatia Role Labeling (SpRL).

B Models and modules

We use the huggingFace[16] implementation of pre-trained BERT base which has 768 hidden dimensions. All models are trained on the training set, evaluated on the dev set, and reported the result on the test set. For training, we train the model until no changes happen on the dev set and then store and use the best model on the dev set. We use AdamW ((Loshchilov and Hutter, 2017)) optimizer on all models and modules.

For SQA tasks we use Focal Loss (Lin et al., 2017) with $\gamma = 2$. For spatial argument extraction,

[16] https://huggingface.co/transformers/v2.9.1/model_doc/bert.html

we use cross-entropy loss for BIO-tagging, and for spatial relation extraction, we use the summation of loss for each spatial relation and relation type classification part.

$$Loss = \sum \text{CrossEntropyLoss}(\text{p}', \text{y}') + \text{BCELoss}(\text{p}, \text{y}) \quad (1)$$

The rest of experimental setting such as number of epochs, batch size, and learning rate are provided in Table 13. This settings are chosen after trial and test on the dev set of the target task.

Dataset	YN	FR
SPARTUN	92.83	93.66
SPARTUN-Simple	90.30	93.66
SPARTUN-Clock	-	87.13
SPARTQA	82.05	94.17

Table 12: Result of BERT (SQA) model trained and test on two synthetic supervision data.

Besides, The result of BERT model trained on SPARTUN and SPARTUN and tested on the same dataset are provided in Table 12. SPARTUN-Simple only contains one spatial expresion for each relation types, and SPARTUN-Clock contains all relation expression plus clock expressions (Column 5 in Table 10a) for relation types.

B.1 Logic-based spatial reasoner

We consider the logic rules mentioned in Figure 2 and in the form of the Horn clauses. we collect the different combinations of spatial relations mentioned in Table 1 and implement the logic-based spatial reasoner. Figure 9a shows an example of some parts of our code on *LEFT* relation. In Figure 9b, on the left, some facts are given, and the query "$ntppi(room, X)$" ask about all objects that existed in the room. Below each query, there are all possible predictions for them.

Experiment	DS	epochs	batch size	learning rate	classifier type
PLM	SPARTQA YN	3	8	8e-06	boolean cls
PLM	SPARTQA FR	30	8	8e-06	boolean cls
PLM	StepGame	30	4	4e-06	multi-class
PLM	SPARTUN YN	4	8	8e-06	boolean cls
PLM	SPARTUN FR	10	8	8e-06	boolean cls
SQA experiments					
bAbI task 17		100	4	4e-06	boolean cls
bAbI task 17	SPARTUN	100	4	4e-06	boolean cls
bAbI task 17	SPARTQA	100	4	4e-06	boolean cls
bAbI task 17	StepGame	100	4	4e-06	boolean cls
bAbI task 19		60	4	4e-06	boolean cls
bAbI task 19	SPARTUN	30	4	4e-06	boolean cls
bAbI task 19	SPARTQA	30	4	4e-06	boolean cls
bAbI task 19	StepGame	30	4	4e-06	boolean cls
SPARTQA-HUMAN YN		60	4	4e-06	boolean cls
SPARTQA-HUMAN YN	SPARTUN	40	4	4e-06	boolean cls
SPARTQA-HUMAN YN	SPARTQA	50	4	4e-06	boolean cls
SPARTQA-HUMAN YN	StepGame	30	4	4e-06	boolean cls
SPARTQA-HUMAN FR		40	1	2e-06	boolean cls
SPARTQA-HUMAN FR	SPARTUN	40	4	4e-06	boolean cls
SPARTQA-HUMAN FR	SPARTQA	40	1	2e-06	boolean cls
SPARTQA-HUMAN FR	StepGame	40	4	4e-06	boolean cls
StepGame		30	4	4e-0	multi-class
StepGame	SPARTUN	30	4	4e-06	multi-class
StepGame	SPARTQA	30	4	4e-06	multi-class
RESQ		50	4	2e-06	boolean cls
RESQ	SPARTUN	50	4	4e-06	boolean cls
RESQ	SPARTQA	50	4	4e-06	boolean cls
RESQ	StepGame	50	4	4e-06	boolean cls
SPRL experiments					
SRol	SPARTQA	3	1	2e-06	-
SRel	SPARTQA	5	1	2e-07	-
SRol	SPARTUN	5	1	8e-06	-
SRel	SPARTUN	10	1	2e-07	-
SRol - SPARTQA-HUMAN		40	1	2e-05	-
SRol - SPARTQA-HUMAN	SPARTQA	50	1	2e-06	-
SRol - SPARTQA-HUMAN	SPARTUN	7	1	4e-07	-
SRol - MSPRL		50	1	2e-06	-
SRol - MSPRL	SPARTQA	50	1	2e-06	-
SRol - MSPRL	SPARTUN	50	1	2e-06	-
SRel - SPARTQA-HUMAN		50	1	2e-06	-
SRel - SPARTQA-HUMAN	SPARTQA	50	1	2e-06	-
SRel - SPARTQA-HUMAN	SPARTUN	50	1	8e-07	-
SRel - MSPRL		50	1	2e-05	-
SRel - MSPRL	SPARTQA	70	1	4e-06	-
SRel - MSPRL	SPARTUN	70	1	6e-06	-

Table 13: The hyperparameters and setups information for each experiment. The first three rows are related to further pretraining model on the synthetic data. These models are used in the other experiments as the further pretrained models.

```
left_2(X,Y) :-
     left_2(X,Y,[]).

left_2(X,Y, _) :-
     left_1(X,Y).

left_2(X,Z, Visited) :-
     left_1(X,Y),
     Y \= X,
     \+ member(Y, Visited),
     left_2(Y,Z, [Y|Visited]),
     Y \= Z,
     X \= Z.

left_2_check(X,Y) :- left_2(X,Y),

right_2(X,Z) :- left_2(Z,X).
```

```
left_3(X,Z) :-
     left_2_and_eq(X,Z);

     left_3__(X,Z).

left_2_and_eq(X,Z) :-
     left_2(X,Z);

     eq_2(X,Y),
     left_2_check(Y,Z);

     eq_2(Y,Z),
     left_2_check(X,Y).
```

```
left_3__(X,Z) :-

     has_pp_rel(X,Y),
     X \= Y,
     left_2_and_eq(Y,H),
     Y \= H,
     has_ppi_rel(H,Z),
     H \= Z;

     has_pp_rel(X,Y),
     X \= Y,
     left_2_and_eq(Y,Z),
     Y \= Z;

     left_2_and_eq(X,H),
     X \= H,
     has_ppi_rel(H,Z),
     H \= Z.
```

(a) Examaple of implemented rule clauses in Prolog.

```
%Facts:

left_(box1, box2).
above_(box1, box2).
below_(box3,box2).

ntpp_(apple,box1).
ntpp_(pear, box3).
ntppi_(box2,grapes).
ntppi_(room,box1).
ntppi_(room,box2).
dc_(box2,box1).
eq_(apple1,apple).
eq_(apple2,apple1).
```

```
%Query:

?- ntppi(room, X).
X = box1 ;
X = box2 ;
X = apple ;
X = grapes ;
X = apple1 ;
X = apple2 ;
X = apple1 ;
false.
```

```
?- left(X, grapes).
X = apple ;
X = apple1 ;
X = apple2 ;
X = apple1 ;
X = box1 ;
false.
```

(b) Example of Facts, Query, and answer of implemented model

Figure 9: Logic-bases spatial reasoner

formalism	Type		Cardinals	Clocks
Directional	left	"to the left of", "on the left side of", "to the left-hand side of"	"west of", "to the west of"	"at 9:00 position relative to", "at 9: 00 position regarding to", "at 9 o'clock position regarding to"
	right	"to the right of", "on the right side of", "to the right-hand side of"	"east of", "to the east of"	"at 3:00 position relative to", "at 3: 00 position regarding to", "at 3 o'clock position regarding to"
	below	"above", "over"	"north of", "to the north of"	"at 12:00 position relative to", "at 12:00 position regarding to", "at 12 o'clock position regarding to"
	above	"below", "under"	"south of", "to the south of"	"at 6:00 position relative to", "at 6: 00 position regarding to", "at 6 o'clock position regarding to"
	behind	"behind"		
	front	"in front of"		
Distances	far	"far from", "farther from", "away from"		
	near	"near to", "close to"		
Topological	DC	disconnected from		
	EC	"touch[es]"		
	PO	"overlap[s]"		
	EQ	-		
	TPP	"covered by", "inside and touching"		
	TPPI	"cover[s]"		
	NTPP	"in", "inside", "within"		
	NTPPI	"ha[s/ve]","contain[s]"		

(a) List of relation expression supported in SPARTUN.

properties	
block	Block, box
blocks	Blocks, boxes
object_general_name	thing, object, shape, fruit
objects_general_name	things, objects, shapes, fruits
block_name	AAA, BBB, CCC, DDD, EEE, JJJ, HHH, JJJ, LLL, KKK, one, two, three.
color	yellow, black, blue, green, red, orange, grey, white, purple
size	small, big, medium, midsize, large, tiny, little
type	circle, oval, square, rectangle, dimond, star, triangle, hexagon, pentagon, watermelon, apple, melon,
types	

(b) List of entities properties supported in SPARTUN

Figure 10: The supported relation expression and entities properties in SPARTUN, which can easily extended based on the target task.