SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning

Roshanak Mirzaee[★] **Hossein Rajaby Faghihi**[★] **Qiang Ning**[♠*] **Parisa Kordjamshidi**[★]
[★]Michigan State University [♠]Amazon
{mirzaeem,rajabyfa,kordjams}@msu.edu qning@amazon.com

Abstract

This paper proposes a question-answering (QA) benchmark for spatial reasoning on natural language text which contains more realistic spatial phenomena not covered by prior work and is challenging for state-of-the-art language models (LM). We propose a distant supervision method to improve on this task. Specifically, we design grammar and reasoning rules to automatically generate a spatial description of visual scenes and corresponding QA pairs. Experiments show that further pre-training LMs on these automatically generated data significantly improves LMs' capability on spatial understanding, which in turn helps to better solve two external datasets, bAbI, and boolQ. We hope that this work can foster investigations into more sophisticated models for spatial reasoning over text.

1 Introduction

Spatial reasoning is a cognitive process based on the construction of mental representations for spatial objects, relations, and transformations (Clements and Battista, 1992), which is necessary for many natural language understanding (NLU) tasks such as natural language navigation (Chen et al., 2019; Roman Roman et al., 2020; Kim et al., 2020), human-machine interaction (Landsiedel et al., 2017; Roman Roman et al., 2020), dialogue systems (Udagawa et al., 2020), and clinical analysis (Datta and Roberts, 2020).

Modern language models (LM), e.g., BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and XLNet (Yang et al., 2019) have seen great successes in natural language processing (NLP). However, there has been limited investigation into *spatial reasoning capabilities of LMs*. To the best of our knowledge, bAbI (Weston et al., 2015) (Fig 9) is the only dataset with direct textual spatial question answering (QA) (Task 17), but it is synthetic and overly simplified: (1) The underlying scenes are spatially simple, with only three objects and relations only in four directions. (2) The stories for these scenes are two short, templated sentences, each describing a single relation between two objects. (3) The questions typically require up to two-steps reasoning due to the simplicity of those stories.

To address these issues, this paper proposes a new dataset, SPARTQA[1] (see Fig. 1). Specifically, (1) SPARTQA is built on NLVR's (Suhr et al., 2017) images containing more objects with richer spatial structures (Fig. 1b). (2) SPARTQA's stories are more natural, have more sentences, and richer in spatial relations in each sentence. (3) SPARTQA's questions require deeper reasoning and have four types: *find relation* (FR), *find blocks* (FB), *choose object* (CO), and *yes/no* (YN), which allows for more fine-grained analysis of models' capabilities.

We showed annotators random images from NLVR, and instructed them to describe objects and relationships not exhaustively at the cost of naturalness (Sec. 3). In total, we obtained 1.1k unique QA pair annotations on spatial reasoning, evenly distributed among the aforementioned types. Similar to bAbI, we keep this dataset in relatively small scale and suggest to use as little training data as possible. Experiments show that modern LMs (e.g., BERT) do not perform well in this low-resource setting.

This paper thus proposes a way to obtain distant supervision signals for spatial reasoning (Sec. 4). As spatial relationships are rarely mentioned in existing corpora, we take advantage of the fact that spatial language is grounded to the geometry of visual scenes. We are able to automatically generate stories for NLVR images (Suhr et al., 2017) via our newly designed context free grammars (CFG) and context-sensitive rules. In the process of story generation, we store the information about all ob-

[*]Work was done while at the Allen Institute for AI.

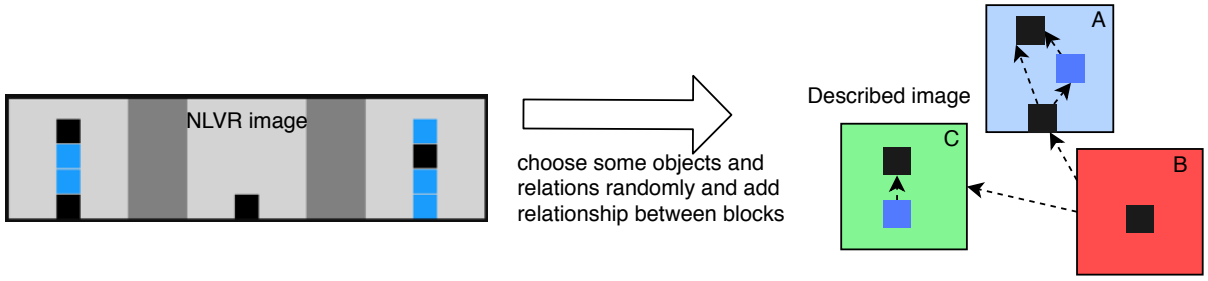
[1]*SPAtial Reasoning on Textual Question Answering.*

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4582–4598
June 6–11, 2021.

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

(a) An example story and corresponding questions and answers.



(b) An example NLVR image and the scene created in Fig. 1a, where the blocks in the NLVR image are rearranged.

Figure 1: Example from SPARTQA (specifically from SPARTQA-AUTO)

jects and relationships, such that QA pairs can also be generated automatically. In contrast to bAbI, we use various spatial rules to infer new relationships in these QA pairs, which requires more complex reasoning capabilities. Hereafter, we call this automatically-generated dataset SPARTQA-AUTO, and the human-annotated one SPARTQA-HUMAN.

Experiments show that, by further pretraining on SPARTQA-AUTO, we improve LMs' performance on SPARTQA-HUMAN by a large margin.[2] The spatially-improved LMs also show stronger performance on two external QA datasets, bAbI and boolQ (Clark et al., 2019): BERT further pretrained on SPARTQA-AUTO only requires half of the training data to achieve 99% accuracy on bAbI as compared to the original BERT; on boolQ's development set, this model shows better performance than BERT, with 2.3% relative error reduction.[3]

Our contributions can be summarized as follows. First, we propose the first human-curated benchmark, SPARTQA-HUMAN, for spatial reasoning with richer spatial phenomena than the prior synthetic dataset bAbI (Task 17).

Second, we exploit the scene structure of images and design novel CFGs and spatial reasoning rules to automatically generate data (i.e., SPARTQA-AUTO) to obtain distant supervision signals for spatial reasoning over text.

Third, SPARTQA-AUTO proves to be a rich source of spatial knowledge that improved the performance of LMs on SPARTQA-HUMAN as well as on different data domains such as bAbI and boolQ.

2 Related work

Question answering is a useful format to evaluate machines' capability of reading comprehension (Gardner et al., 2019) and many recent works have been implementing this strategy to test machines' understanding of linguistic formalisms: He et al. (2015); Michael et al. (2018); Levy et al. (2017); Jia et al. (2018); Ning et al. (2020); Du

[2]Further pretraining LMs has become a common practice and baseline method for transferring knowledge between tasks (Phang et al., 2018; Zhou et al., 2020). We leave more advanced methods for future work.

[3]To the best of our knowledge, the test set or leaderboard of boolQ has not been released yet.

and Cardie (2020). An important advantage of QA is using natural language to annotate natural language, thus having the flexibility to get annotations on complex phenomena such as *spatial reasoning*. However, spatial reasoning phenomena have been covered minimally in the existing works.

To the best of our knowledge, Task 17 of the bAbI project (Weston et al., 2015) is the only QA dataset focused on textual spatial reasoning (examples in Appendix F). However, bAbI is synthetic and does not reflect the complexity of the spatial reasoning in natural language. Solving Task 17 of bAbI typically does not require sophisticated reasoning, which is an important capability emphasized by more recent works (e.g., Dua et al. (2019); Khashabi et al. (2018); Yang et al. (2018); Dasigi et al. (2019); Ning et al. (2020)).

Spatial reasoning is arguably more prominent in multi-modal QA benchmarks, e.g., NLVR (Suhr et al., 2017), VQA (Antol et al., 2015), GQA (Hudson and Manning, 2019), CLEVR (Johnson et al., 2017). However, those spatial reasoning phenomena are mostly expressed naturally through images, while this paper focuses on studying spatial reasoning on natural language. Some other works on visual-spatial reasoning are based on geographical information inside maps and diagrams (Huang et al., 2019) and navigational instructions (Chen et al., 2019; Anderson et al., 2018).

As another approach to evaluate spatial reasoning capabilities of models, a dataset proposed in Ghanimifard and Dobnik (2017) generates a synthetic training set of spatial sentences and evaluates the models' ability to generate spatial facts and sentences containing composition and decomposition of relations on grounded objects.

3 SpartQA-Human

To mitigate the aforementioned problems of Task 17 of bAbI, i.e., simple scenes, stories, and questions, we describe the data annotation process of SpartQA-Human, and explain how those problems were addressed in this section.

First, we randomly selected a subset of NLVR images, each of which has three blocks containing multiple objects (see Fig 1b). The scenes shown by these images are more complicated than those described by bAbI because (1) there are more objects in NLVR images; (2) the spatial relationships in NLVR are not limited to just four relative directions as objects are placed arbitrarily within blocks.

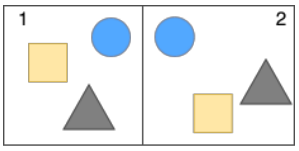


Figure 2: For "A blue circle is above a big triangle. To the left of the big triangle, there is a square," if the question is: "Is the square to the left of the blue circle?", the answer is neither Yes nor No. Thus, the correct answer is "Do not Know" (DK) in our setting.

Second, two student volunteers produced textual description of those objects and their corresponding spatial relationships based on these images. Since the blocks are always horizontally aligned in each NLVR image, to allow for more flexibility, annotators could also rearrange these blocks (see Fig. 1a). Relationships between objects within the same block can take the forms of relative direction (e.g., left or above), qualitative distance (e.g., near or far), and topological relationship (e.g., touching or containing).

However, we instructed the annotators not to describe all objects and relationships, (1) to avoid unnecessarily verbose stories, and (2) to intentionally miss some information to enable more complex reasoning later. Therefore, annotators describe only a random subset of blocks, objects, and relationships.

To query more interesting phenomena, annotators were then encouraged to write questions requiring detecting relations and reasoning over them using multiple spatial rules. A spatial rule can be one of the transitivity ($A \rightarrow B, B \rightarrow C \Rightarrow A \rightarrow C$), symmetry ($A \rightarrow B \Rightarrow B \rightarrow A$), converse ($(A,\ R,\ B) \Rightarrow (B,\ reverse(R),\ A)$), inclusion ($obj1\ in\ A$), and exclusion ($obj1\ not\ in\ B$) rules.

There are four types of questions (Q-Type). (1) *FR*: find relation between two objects. (2) *FB*: find the block that contains certain object(s). (3) *CO*: choose between two objects mentioned in the question that meets certain criteria. (4) *YN*: a yes/no question that tests if a claim on spatial relationship holds.

FB, FR, and CO questions are formulated as multiple-choice questions[4] and receive a list of candidate answers, and YN questions' answer is choosing from Yes, No, or "DK" (Do not Know). The "DK" option is due to the open-world assumption of the stories, where if something is not described

[4]CO can be considered as both single-choice and multiple-choices question.

Sets	FB	FR	YN	CO	Total
SPARTQA-HUMAN:					
Test	104	105	194	107	510
Train	154	149	162	151	616
SPARTQA-AUTO:					
Seen Test	3872	3712	3896	3594	15074
Unseen Test	3872	3721	3896	3598	15087
Dev	3842	3742	3860	3579	15023
Train	23654	23302	23968	22794	93673

Table 1: Number of questions per Q-TYPE

in the text, it is not considered as false (See Fig. 2).

Finally, annotators were able to create 1.1k QA pairs on spatial reasoning on the generated descriptions, distributed among the aforementioned types. We intentionally keep this data in a relatively small scale due to two reasons. First, there has been some consensus in our community that modern systems, given their sufficiently large model capacities, can easily find shortcuts and overfit a dataset if provided with a large training data (Gardner et al., 2020; Sen and Saffari, 2020). Second, collecting spatial reasoning QAs is very costly: The two annotators spent 45-60 mins on average to create a single story with 8-16 QA pairs. We estimate that SPARTQA-HUMAN costed about 100 human hours in total. The expert performance on 100 examples of SPARTQA-HUMAN's test set measured by their accuracy of answering the questions is 92% across four Q-TYPEs on average, indicating its high quality.

4 Distant Supervision: SPARTQA-AUTO

Since human annotations are costly, it is important to investigate ways to generate distant supervision signals for spatial reasoning. However, unlike conventional distant supervision approaches (e.g., Mintz et al. (2009); Zeng et al. (2015); Zhou et al. (2020)) where distant supervision data can be selected from large corpora by implementing specialized filtering rules, spatial reasoning does not appear often in existing corpora. Therefore, similar to SPARTQA-HUMAN, we take advantage of the ground truth of NLVR images, design CFGs to generate stories, and use spatial reasoning rules to ask and answer spatial reasoning questions. This automatically generated data is called SPARTQA-AUTO, and below we describe its generation process in detail.

Story generation Since NLVR comes with structured descriptions of the ground truth locations of those objects, we were able to choose random blocks and objects from each image programmatically. The benefit is two-fold. First, a random selection of blocks and objects allows us to create multiple stories for each image; second, this randomness also creates spatial reasoning opportunities with missing information.

Once we decide on a set of blocks and objects to be included, we determine their relationships: Those relationships between blocks are generated randomly; as for those between objects, we refer to the ground truth of these images to determine them.

Now we have a scene containing a set of blocks and objects and their associated relationships. To produce a story for this scene, we design CFGs to produce natural language sentences that describe those blocks/objects/relationships in various expressions (see Fig. 3 for two portions of our CFG describing relative and nested relations between objects).

The big black shape is above the medium triangle.

S → *<Article> <Object> is <Relation> <Article> <Object>.*

Article → *the | a*
Relation → *above | left | ...*
Object → *<Size>* <Color>* <Shape| Ind_shape>*
Size → *small | medium | big*
Color → *yellow | blue | black*
Shape → *square | triangle | circle*
Ind_shape → *shape | object | thing*

(a) Part of the grammar describing relations between objects

The big black shape is above the object that is to the right of the medium triangle

S → *<Article> <Object> is <Relation> <Article> <Object>.*

Object → *<Size>* <Color>* <Shape| Ind_shape> | <Ind_shape> that is <Relation> <Object>*

(b) Part of the grammar describing nested relationships.

Figure 3: Two parts of our designed CFG

Being grounded to visual scenes guarantees spatial coherency in a story, and using CFGs helps to have correct sentences (grammatically) and various expressions. We also design context-sensitive rules to limited options for each CFG's variable based on the chosen entities (e.g. black circle), or what is described in the previous sentences (e.g. Block A has *a* circle. *The* circle is below *a* triangle.)

Question generation To generate questions based on a passage, there are rule-based sys-

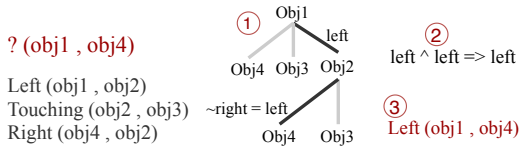


Figure 4: Find the implicit relation between $obj1$ and $obj4$ by *Transitivity* rule. (1) Find a set of objects that have a relation with $obj1$. Continue the same process on the new set until $obj4$ is found. (2) Get the union of the intermediate relations between these two objects and it is the final answer.

tems (Heilman and Smith, 2009; Labutov et al., 2015), neural networks (Du et al., 2017), and their combinations (Dhole and Manning, 2020). However, in our approach, during generating each story, the program stores the information about the entities and their relationships. Thus, without processing the raw text, which is error-prone, we generate questions by only looking at the stored data. The question generation operates based on four primary functionalities, *Choose-objects*, *Describe-objects*, *Find-all-relations*, and *Find-similar-objects*. These modules are responsible to control the logical consistency, correctness, and the number of steps required for reasoning in each question.

Choose-objects randomly chooses up to three objects from the set of possible objects in a story under a set of constraints such as preventing selection of similar objects, or excluding objects with relations that are directly mentioned in the text.

Describe-Objects generates a mention phrase for an object using parts of its full name (presented in the story). The generated phrase is either pointing to a unique object or a group of objects such as "the big circle," or "big circles." To describe a unique object, it chooses an attribute or a group of attributes that apply to a unique object among others in the story. To increase the steps of reasoning, the description may include the relationship of the object to other objects instead of using a direct unique description. For example, "the circle which is above the black triangle."

Find-all-relations completes the relationship graph between objects by applying a set of spatial rules such as transitivity, symmetry, converse, inclusion, and exclusion on top of the direct relations described in the story. As shown in Fig. 4, it does an exhaustive search over all combinations of the relations that link two objects to each other.

Find-similar-objects finds all the mentions matching a description from the question to objects in the story. For instance, for the question "is there any blue circle above the big blue triangle?", this module finds all the mentions in the story matching the description “a blue circle”.

Similar to the SPARTQA-HUMAN, we provide four Q-TYPES FR, FB, CO, and YN. To generate FR questions, we choose two objects using *Choose-objects* module and question their relationships. The YN Q-TYPE is similar to FR, but the question specifies one relationship of interest chosen from all relation extracted by *Find-all-relations* module to be questioned about the objects. Since most of the time, Yes/No questions are simpler problems, we make this question type more complex by adding quantifiers (adding “all” and “any”). These quantifiers help to evaluates the models’ capability to aggregate relations between more than two objects in the story and do the reasoning over all find relations to find the final answer. In FB Q-TYPE, we mention an object by its indirect relation to another object using the nested relation in *Describe-objects* module and ask to find the blocks containing or not containing this object. Finally, the CO question selects an anchor object (*Choose-objects*) and specifies a relationship (using *Find-all-relations*) in the question. Two other objects are chosen as candidates to check whether the specified relationship holds between them and the anchor object. We tend to force the algorithm to choose objects as candidates that at least have one relationship to the anchor object. To see more details about different question’ templates see Table 7 in the Appendix.

Answer generation We compute all direct and indirect relationships between objects using *Find-all-relations* function and based on the Q-TYPES generate the final answer.

For instance, in YN Q-TYPE if the asked relation exists in the found relations, the answer is "Yes", if the inverse relation exists it must be "No", and otherwise, it is "DK"[5].

4.1 Corpus Statistics

We generate the train, dev, and test set splits based on the same splits of the images in the NLVR dataset. On average, each story contains 9 sentences (Min:3, Max: 22) and 118 tokens (Min: 66,

[5]The SPARTQA-AUTO generation code and the file of dataset are available at https://github.com/HLR/SpartQA_generation

Max: 274). Also, the average tokens of each question (on all Q-TYPE) is 23 (Min:6, Max: 57).

Table 1 shows the total number of each question type in SPARTQA-AUTO (Check Appendix to see more statistic information about the labels in Tab 8.)

5 Models for Spatial Reasoning over Language

This section describes the model architectures on different Q-TYPEs: FR, YN, FB, and CO. All Q-TYPEs can be cast into a sequence classification task, and the three transformer-based LMs tested in this paper, BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and XLNet (Yang et al., 2019), can all handle this type of tasks by classifying the representation of [CLS], a special token prepended to each target sequence (see Appendix E). Depending on the Q-TYPE, the input sequence and how we do inference may be different.

FR and YN both have a predefined label set as candidate answers, and their input sequences are both the concatenation of a story and a question. While the answer to a YN question is a single label chosen from *Yes, No*, and *DK*, FR questions can have multiple correct answers. Therefore, we treat each candidate answer to FR as an independent binary classification problem, and take the union as the final answer. As for YN, we choose the label with the highest confidence (Fig 8b).

As the candidate answers to FB and CO are not fixed and depend on each story and its question the input sequences to these Q-TYPEs are concatenated with each candidate answer. Since the defined YN and FR model has moderately less accurate results on FB and CO Q-TYPEs, we add a LSTM (Hochreiter and Schmidhuber, 1997) layer to improve it. Hence, to find the final answer, we run the model with each candidate answer and then apply an LSTM layer on top of all token representations. Then, we use the last vector of the LSTM outputs for classification (Fig 8a). The final answers are selected based on Eq. (1).

$$
\begin{aligned}
x_i &= [s, c_i, q] \\
\vec{T_i} &= [\vec{t_1^i}, ..., \vec{t_{m^i}^i}] = LM(x_i) \\
[\vec{h}_1^i, ..., \vec{h}_{m^i}^i] &= \text{LSTM}(\vec{T_i}) \\
\vec{y_i} = [y_i^0, y_i^1] &= \text{Softmax}(\vec{h}_{m^i}^{iT} W)) \\
\text{Answer} &= \{c_i | \arg\max_j (y_i^j) = 1\}
\end{aligned}
\tag{1}
$$

where s is the story, c_i is the candidate answer, q is the question, [ ] indicates the concatenation of the listed vectors, and m_i is tokens' number in x_i. The parameter vector, W, is shared for all candidates.

5.1 Training and Inference

We train the models based on the summation of the cross-entropy losses of all binary classifiers in the architecture. For FR and YN Q-TYPEs, there are multiple classifiers, while there is only one classifier used for CO and FB Q-TYPEs.

We remove inconsistent answers in post-processing for FR and YN Q-TYPEs during inference phase. For instance on FR, *left* and *right* relations between two objects cannot be valid at the same time. For YN, as there is only one valid answer amongst the three candidates, we select the candidate with the maximal predicted probability of being the true answer.

6 Experiments

As fine-tuning LMs has become a common baseline approach to knowledge transfer from a source dataset to a target task, including but not limited to Phang et al. (2018); Zhou et al. (2020); He et al. (2020b), we study the capability of spatial reasoning of modern LMs, specifically BERT, ALBERT, and XLNet, after fine-tuning them on SPARTQA-AUTO. This fine-tuning process is also known as *further pretraining*, to distinguish with the fine-tuning process on one's target task. It is an open problem to find out better transfer learning techniques than simple further pretraining, as suggested in He et al. (2020a); Khashabi et al. (2020), which is beyond the scope of this work. All experiments use the models proposed in Sec. 5. We use AdamW (Loshchilov and Hutter, 2017) with 2×10^{-6} learning rate and Focal Loss (Lin et al., 2017) with $\gamma = 2$ for training all the models.[6]

6.1 Further pretraining on SPARTQA-AUTO improves spatial reasoning

Table 2 shows performance on SPARTQA-HUMAN in a low-resource setting, where 0.6k QA pairs from SPARTQA-HUMAN are used for fine-tuning these LMs and 0.5k for testing (see Table 1 for information on this split).[7] During our annotation, we found that the description of "near to " and "far

[6] All codes are available at https://github.com/HLR/SpartQA-baselines

[7] Note this low-resource setting can also be viewed as a spatial reasoning probe to these LMs (Tenney et al., 2019).

#	Model	FB	FR	CO	YN	Avg
1	Majority	28.84	24.52	40.18	53.60	36.64
2	BERT	16.34	20	26.16	45.36	30.17
3	BERT (Stories only; MLM)	21.15	16.19	27.1	**51.54**	32.90
4	BERT (SPARTQA-AUTO; MLM)	19.23	29.54	**32.71**	47.42	34.88
5	BERT (SPARTQA-AUTO)	**62.5**	**46.66**	**32.71**	47.42	**47.25**
6	Human	91.66	95.23	91.66	90.69	92.31

Table 2: **Further pretraining BERT on SPARTQA-AUTO improves accuracies on SPARTQA-HUMAN**. All systems are fine-tuned on the training data of SPARTQA-HUMAN, but Systems 3-5 are also further pretrained in different ways. System 3: further pretrained on the stories from SPARTQA-AUTO as a masked language model (MLM) task. System 4: further pretrained on both stories and QA annotations as MLM. System 5: the proposed model that is further pretrained on SPARTQA-AUTO as a QA task. Avg: The micro-average on all four Q-TYPES.

from" varies largely between annotators. Therefore, we ignore these two relations from FR Q-TYPE in our evaluations.

In Table 2, System 5, BERT (SPARTQA-AUTO), is the proposed method of further pretraining BERT on SPARTQA-AUTO. We can see that System 2, the original BERT, performs consistently lower than System 5, indicating that having SPARTQA-AUTO as a further pretraining task improves BERT's spatial understanding.

Model	F_1
Majority	35
BERT	50
BERT (Stories only; MLM)	53
BERT (SPARTQA-AUTO; MLM)	48
BERT (SPARTQA-AUTO)	48

Table 3: Switching from accuracy in Table 2 to F_1 shows that the models are all performing better than the majority baseline on YN Q-TYPE.

In addition, we implement another two baselines. System 3, BERT (Stories only; MLM): further pretraining BERT only on the stories of SPARTQA-AUTO as a masked language model (MLM) task; System 4, BERT (SPARTQA-AUTO; MLM): we convert the QA pairs in SPARTQA-AUTO into textual statements and further pretrain BERT on the text as an MLM (see Fig. 5 for an example conversion).

To convert each question and its answer into a sentence, we utilize static templates for each question type which removes the question words and rearranges other parts into a sentence.

We can see that System 3 slightly improves over System 2, an observation consistent with many prior works that seeing more text generally helps an LM (e.g., Gururangan et al. (2020)). The significant gap between System 3 and the proposed System 5 indicates that supervision signals come more from our annotations in SPARTQA-AUTO rather than from seeing more unannotated text. System 4 is another way to make use of the annotations in SPARTQA-AUTO, but it is shown to be not as effective as further pretraining BERT on SPARTQA-AUTO as a QA task.

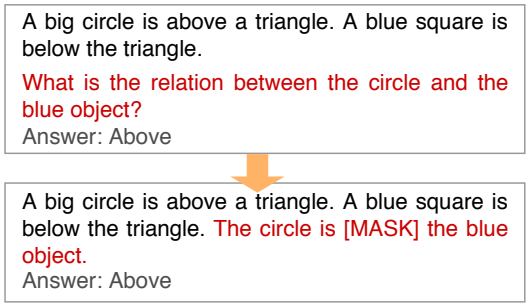


Figure 5: Convert a triplet of (paragraph, question, answer) into a single piece of text for the MLM task.

While the proposed System 5 overall performs better than the other three baseline systems, one exception is its accuracy on YN, which is lower than that of System 3. Since all systems' YN accuracies are also lower than the majority baseline[8], we hypothesize that this is due to imbalanced data. To verify it, we compute the F_1 score for YN Q-TYPE in Table 3, where we see all systems effectively achieve better scores than the majority baseline. However, further pretraining BERT on SPARTQA-AUTO still does not beat other baseline systems, which implies that straightforward pretraining is not necessarily helpful in capturing the complex reasoning phenomena required by YN questions.

The human performance is evaluated on 100 ran-

[8] which predicts the label that is most common in each set of SPARTQA

#	Models	FB			FR			CO			YN		
		Seen	Unseen	Human*	Seen	Unseen	Human*	Seen	Unseen	Human*	Seen	Unseen	Human*
1	Majority	48.70	48.70	28.84	40.81	40.81	24.52	20.59	20.38	40.18	49.94	49.91	**53.60**
2	BERT	87.13	69.38	62.5	85.68	73.71	46.66	71.44	61.09	32.71	78.29	76.81	47.42
3	ALBERT	97.66	83.53	56.73	91.61	83.70	44.76	95.20	84.55	49.53	79.38	75.05	41.75
4	XLNet	**98.00**	**84.85**	**73.07**	**94.60**	**91.63**	**57.14**	**97.11**	**90.88**	**50.46**	**79.91**	**78.54**	39.69
5	Human	85		91.66	90		95.23	94.44		91.66	90		90.69

Table 4: **Spatial reasoning is challenging**. We further pretrain three transformer-based LMs, BERT, ALBERT, and XLNet, on SPARTQA-AUTO, and test their accuracy in three ways: *Seen* and *Unseen* are both from SPARTQA-AUTO, where *Unseen* has applied minor modifications to its vocabulary; to get those *Human* columns, all models are fine-tuned on SPARTQA-HUMAN's training data. Human performance on *Seen* and *Unseen* is the same since the changes applied to *Unseen* does not affect human reasoning.

dom questions from each SPARTQA-AUTO and SPARTQA-HUMAN test set. The respondents are graduate students that were trained by some examples of the dataset before answering the final questions. We can see from Table 2 that all systems' performances fall behind human performance by a large margin. We expand on the difficulty of SPARTQA in the next subsection.

6.2 SPARTQA is challenging

In addition to BERT, we continue to test another two LMs, ALBERT and XLNet (Table 5). We further pretrain these LMs on SPARTQA-AUTO, and test them on SPARTQA-HUMAN (the numbers of BERT are copied from Table 2) and two held-out test sets of SPARTQA-AUTO, *Seen* and *Unseen*. Note that when a system is tested against SPARTQA-HUMAN, it is fine-tuned on SPARTQA-HUMAN's training data following its further pretraining on SPARTQA-AUTO. We use the unseen set to test to what extent the baseline models use shortcuts in the language surface. This set applies minor modifications randomly on a number of stories and questions to change the names of shapes, colors, sizes, and relationships in the vocabulary of the stories, which do not influence the reasoning steps (more details in Appendix C.1).

All models perform worst in YN across all Q-TYPES, which suggests that YN presents a more complex phenomena, probably due to additional quantifiers in the questions. XLNet performs the best on all Q-TYPES except its accuracy on SPARTQA-HUMAN's YN section. However, the drops in *Unseen* and *human* suggest overfitting on the training vocabulary. The low accuracies on human test set from all models show that solving this benchmark is still a challenging problem and requires more sophisticated methods like considering spatial roles and relations extraction (Kordjamshidi et al., 2010; Dan et al., 2020; Rahgooy et al., 2018) to understand stories and questions better.

To evaluate the reliability of the models, we also provide two extra consistency and contrast test sets. **Consistency set** is made by changing a part of the question in a way that seeks for the same information (Hudson and Manning, 2019; Suhr et al., 2019). Given a pivot question and answer of a specific consistency set, answering other questions in the set does not need extra reasoning over the story.

Contrast set is made by minimal modification in a question to change its answer (Gardner et al., 2020). For contrast sets, there is a need to go back to the story to find the new answer for the question's minor variations (see Appendix C.2 for examples.) The consistency and contrast sets are evaluated only on the correctly predicted questions to check if the actual understanding and reasoning occurs. This ensures the reliability of the models.

Table 5 shows the result of this evaluation on four Q-TYPES of SPARTQA-AUTO, where we can see, for another time, that the high scores on the *Seen* test set are likely due to overfitting on training data rather than correct detection of spatial terms and reasoning over them.

6.3 Extrinsic evaluation

In this subsection, we take BERT as an example to show, once pretrained on SPARTQA-AUTO, BERT can achieve better performance on two extrinsic evaluation datasets, namely bAbI and boolQ.

We draw the learning curve on bAbI, using the original BERT as a baseline and BERT further pretrained on SPARTQA-AUTO (Fig. 6). Although both systems achieve perfect accuracy given large enough training data (i.e., 5k and 10k), BERT (SPARTQA-AUTO) is showing better scores given less training data. Specifically, to achieve an accuracy of 99%, BERT (SPARTQA-AUTO) requires

Models	FB	FR		CO		YN	
	Consistency	Consistency	Contrast	Consistency	Contrast	Consistency	Contrast
BERT	69.44	76.13	42.47	16.99	15.58	48.07	71.41
AlBERT	84.77	82.42	41.69	58.42	62.51	48.78	69.19
XLNet	85.2	88.56	50	71.10	72.31	51.08	69.18

Table 5: Evaluation of consistency and semantic sensitivity of models in Table 4. All the results are on the correctly predicted questions of *Seen* test set of SPARTQA-AUTO.

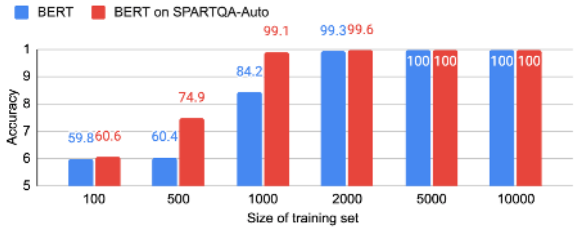


Figure 6: Learning curve of BERT and BERT further pretrained on SPARTQA-AUTO on bAbI.

Model	Accuracy
Majority baseline	62.2
Recurrent model (ReM)	62.2
ReM fine-tuned on SQuAD	69.8
ReM fine-tuned on QNLI	71.4
ReM fine-tuned on NQ	72.8
BERT (our setup)	71.9
BERT (SPARTQA-AUTO)	**74.2**

Table 6: System performances on the dev set of boolQ (since the test set is not available to us). Top: numbers reported in (Clark et al., 2019). Bottom: numbers from our experiments. BERT (SPARTQA-AUTO): further pretraining BERT on SPARTQA-AUTO as a QA task.

1k training examples, while BERT requires twice as much. We also notice that BERT (SPARTQA-AUTO) converges faster in our experiments.

As another evaluation dataset, we chose boolQ for two reasons. First, we needed a QA dataset with Yes/No questions. To our knowledge boolQ is the only available one used in the recent work. Second, indeed, SPARTQA and boolQ are from different domains, however, boolQ needs multi-step reasoning in which we wanted to see if SPARTQA helps.

Table 6 shows that further pretraining BERT on SPARTQA-AUTO yields a better result than the original BERT and those reported numbers in Clark et al. (2019), which also tested on various distant supervision signals such as SQuAD (Rajpurkar et al., 2016), Google's Natural Question dataset NQ (Kwiatkowski et al., 2019), and QNLI from GLUE (Wang et al., 2018).

We observe that many of the boolQ examples answered correctly by the BERT further pretrained on SPARTQA-AUTO require multi-step reasoning. Our hypothesis is that since solving SPARTQA-AUTO questions needs multi-step reasoning, fine-tuning BERT on SPARTQA-AUTO generally improves this capability of the base model.

7 Conclusion

Spatial reasoning is an important problem in natural language understanding. We propose the first human-created QA benchmark on spatial reasoning, and experiments show that state-of-the-art pretrained language models (LM) do not have the capability to solve this task given limited training data, while humans can solve those spatial reasoning questions reliably. To improve LMs' capability on this task, we propose to use hand-crafted grammar and spatial reasoning rules to automatically generate a large corpus of spatial descriptions and corresponding question-answer annotations; further pretraining LMs on this distant supervision dataset significantly enhances their spatial language understanding and reasoning. We also show that a spatially-improved LM can have better results on two extrinsic datasets (bAbI and boolQ).

Acknowledgements

This project is supported by National Science Foundation (NSF) CAREER award #2028626 and (partially) supported by the Office of Naval Research grant #N00014-20-1-2005. We thank the reviewers for their helpful comments to improve this paper and Timothy Moran for his help in the human data generation.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-

and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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 Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936.

Douglas H Clements and Michael T Battista. 1992. Geometry and spatial reasoning. *Handbook of research on mathematics teaching and learning*, pages 420–464.

Soham Dan, Parisa Kordjamshidi, Julia Bonn, Archna Bhatia, Zheng Cai, Martha Palmer, and Dan Roth. 2020. From spatial relations to spatial configurations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5855–5864, Marseille, France. European Language Resources Association.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932.

Surabhi Datta and Kirk Roberts. 2020. A hybrid deep learning approach for spatial trigger extraction from radiology reports. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 50–55, Online. Association for Computational Linguistics.

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.

Kaustubh Dhole and Christopher D. Manning. 2020. Syn-QG: Syntactic and shallow semantic rules for question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 752–765.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. Question Answering is a Format; when is it useful? *ArXiv*, abs/1909.11291.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *IWCS 2017-12th International Conference on Computational Semantics-Long papers*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Hangfeng He, Qiang Ning, and Dan Roth. 2020a. QuASE: Question-answer driven sentence encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8743–8758, Online. Association for Computational Linguistics.

Hangfeng He, Mingyuan Zhang, Qiang Ning, and Dan Roth. 2020b. Foreshadowing the benefits of incidental supervision. *arXiv preprint arXiv:2006.05500*.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-Answer Driven Semantic Role Labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653.

Michael Heilman and Noah A Smith. 2009. Question generation via overgenerating transformations and ranking. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA LANGUAGE TECHNOLOGIES INST.

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

Zixian Huang, Yulin Shen, Xiao Li, Yu'ang Wei, Gong Cheng, Lin Zhou, Xinyu Dai, and Yuzhong Qu. 2019. GeoSQA: A benchmark for scenario-based question answering in the geography domain at high school level. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5866–5871.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. TempQuestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

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

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

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.

Hyounghun Kim, Abhaysinh Zala, Graham Burri, Hao Tan, and Mohit Bansal. 2020. ArraMon: A joint navigation-assembly instruction interpretation task in dynamic environments. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3910–3927, 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).

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural Questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 889–898.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Christian Landsiedel, Verena Rieser, Matthew Walter, and Dirk Wollherr. 2017. A review of spatial reasoning and interaction for real-world robotics. *Advanced Robotics*, 31(5):222–242.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot relation extraction via reading comprehension. In *CONLL*, pages 333–342.

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.

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

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Taher Rahgooy, Umar Manzoor, and Parisa Kordjamshidi. 2018. Visually guided spatial relation extraction from text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 788–794.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. 2020. RMM: A recursive mental model for dialogue navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1732–1745, Online. Association for Computational Linguistics.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438.

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.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Takuma Udagawa, Takato Yamazaki, and Akiko Aizawa. 2020. A linguistic analysis of visually grounded dialogues based on spatial expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 750–765, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

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*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. pages 2369–2380.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589.

A Question Templates and statistics Information

Table 7 shows the templates used to create questions in SPARTQA-AUTO. The "<object>" is a variable replaced by objects from the story (using *Choose-objects* and *Describe-objects* modules), and the "<relation>" variable can be replaced by the chosen relations between objects (using *Find-all-relations* module).

The articles and the indefinite pronouns in each template play an essential role in understanding the question's objective. For example, "Are all blue circles near to a triangle?" is different from "Are there any blue circles near to a triangle?", and "Are there any blue circles near to all triangles?". Therefore, we check the uniqueness of the object definition, using "a" or "the" in proper places and randomly place the terms "any" or "all" in the YN questions to generate different questions.

Table 8 shows the percentage of correct labels in train and test sets. In multi-choice Q-TYPES, more than one label can be true.

B Sentences of the Dataset

Table 10 shows some generated sentences in SPARTQA-AUTO with some specific features that challenge models to understand different forms of relation description in spatial language.

C Additional Evaluation Sets

Here we describe three extra evaluation sets provided with this dataset in more detail, including unseen test, consistency, and contrast sets.

C.1 Unseen Evaluation Set

We propose an unseen test set alongside the seen test of SPARTQA-AUTO to check whether a model is using shortcuts in the language surface by describing objects and relations with new vocabularies in the samples. This set has minor modifications that should not affect the performance of a consistent and reliable model. The modifications are randomly applied on a number of generated stories and questions and include changing names of shapes, colors, sizes, and relationships' names (describing relationships using different language expressions). The modification choices are described in Table 9.

C.2 Contrast and Consistency Evaluation

For probing the consistency and semantic sensitivity of models, we provide two extra evaluation test sets, Consistency and Contrast[9].

Consistency set is made by changing parts of the question in a way that it still asks about the same information (Hudson and Manning, 2019; Suhr et al., 2019). For instance, for the question, "What is the relation between the blue circle and the big shape? Left," we create a similar question in the form of "What is the relation between the big shape and the blue circle? Right". Answering these questions around a pivot question is possible for human without the need for extra reasoning over the story and based on the main questions' answer. Hence, the evaluation on this set shows that models understand the real underlying semantics rather than overfit on the structure of questions.

Contrast set: This set is made by minor changes in a question that changes the answer (Gardner et al., 2020). As an instance, in the question "Is the blue circle below the black triangle? Yes," we create a contrast question "Is the blue circle below all triangles? No" by changing "the black trinagle" to "all triangles". The evaluation on this set shows the robustness of the model and its sensitivity to the semantic changes when there are minor changes in the language surface [10].

D Extra Annotations

Alongside the main SPARTQA-AUTO's stories and questions we provided some extra annotation to help the models to understand the spatial language better.

D.1 Detailed Annotation and Scene-Graphs

Providing in-depth human annotations is quite expensive and time-consuming. In SPARTQA-AUTO, we generated fine-grained scene-graph based on the story. This scene-graph contains blocks' description, their relations, and the objects' attributes alongside their direct relations with each other. The scene-graphs can be used for the models to understand all spatial relations directly mentioned in the textual context. Figure 7 shows an example of this scene-graph. The scene-graph can provide strong supervision for question answering challenges and

[9] for some questions, it is not possible to generate a complementary set

[10] Based on the original contrast set paper, consistency and contrast set should be generated manually to control the semantic change. In our case that we are probing the spatial language understanding of models, we must change parts that affect spatial understanding, which can be implemented by some static rules.

Q-Type	Q-Templates	Candidate answer
FR	what is the relation between <object>and <object>?	Left, Right, Below, Above, Touching, Far from, Near to
CO	What is <relation >the <object>? an <object1>or an <object2>? Which object is <relation >an <object>? the <object1>or the <object2>?	Object1, object2, Both, None
YN	Is (the | a)<object1><relation>(the | a) <object2>? Is there any <object1>s <relation>all <object2>s?	Yes, No, Don't Know
FB	Which block has an <object>? Which block doesn't have an <object>?	Name of blocks, None

Table 7: Questions and answers templates.

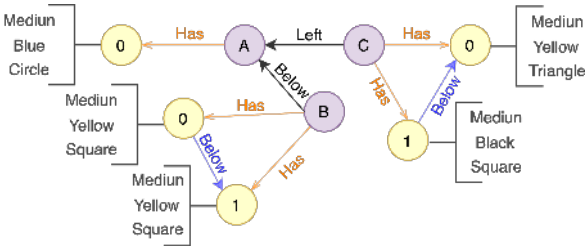


Figure 7: Scene-graph

can be used to evaluate models based on their steps of reasoning and decisions.

D.2 SpRL Annotation

We also provided spatial annotations for each sentence and question, based on Spatial Role Labeling (SpRL) annotation scheme (Kordjamshidi et al., 2010)(Fig. 11). This annotation is generated by hand-crafted rules during the main data generation. SpRL is used for recognizing spatial expressions and arguments in a sentence. This annotation is useful for applications that need to detect and reason about spatial expressions and arguments.

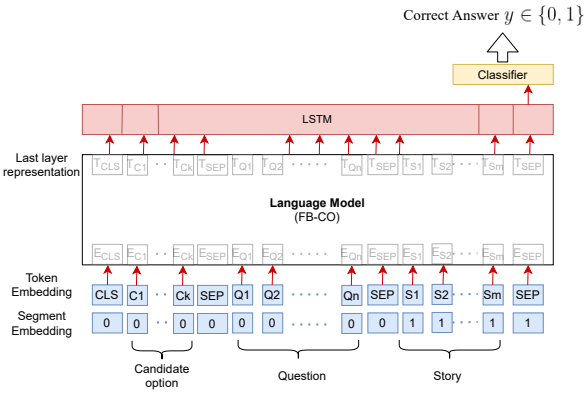
E QA Language Models for Spatial Reasoning over Text

Figures 8a and 8b depict the architecture used for further fine-tuning language models on SPARTQA described in section 5.

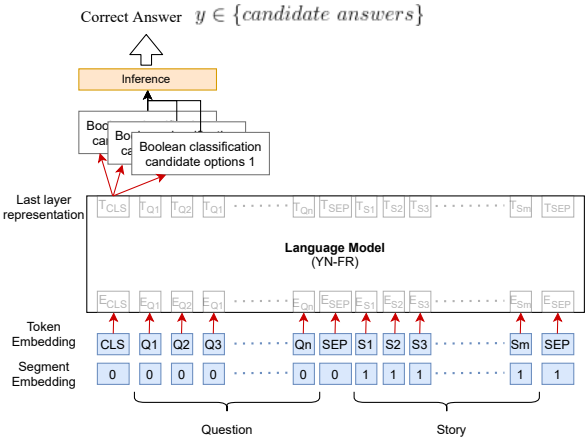
F bAbI and boolQ Datasets

Figure 9 shows an example of the bAbI dataset (Weston et al., 2015) task 17.

To solve task 17 of bAbI , we implement two SpRL+rule-based and neural network models. The



(a) LM_{QA} Architecture for CO and FB Q-TYPEs



(b) LM_{QA} Architecture for FR and YN Q-TYPEs

Figure 8: LM_{QA} for Spatial Reasoning over Text

"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 9: An example of bAbI dataset, task 17.

SpRL+rule-based model first, finds different spa-

Q-TYPE	Candidate Answers	train	test
FR (Multiple Choices)	Left	20.7	17.9
	Right	21.4	16.7
	Above	26.9	25.4
	Below	37.2	42.9
	Near to	5.8	2.9
	Far from	1.3	0.56
	Touching	0.57	0.27
	DK	0.52	0.32
FB (multiple Choices)	A	49.8	49.4
	B	50.1	50
	C	35.1	62
	[]	7.1	90.5
CO (Single choice)	Object1	25.4	26
	Object2	25.3	24.9
	Both	44.3	43.9
	None	4.9	5.0
YN (Single choice)	Yes	53.3	50.5
	No	18.7	23.6
	DK	27.8	25.9

Table 8: The percentage of each correct label in all samples. *The candidate answers for the FB Q-TYPE can be varied, based on its story. **CO can be considered as a multiple choice or single choice question. E.g., in "which object is above the triangle? the blue circle or the black circle?" you can consider two labels with boolean classification on each "blue circle" and "black circle" or consider it as a four labels classification: "blue circle," "black circle," "both of them," and "None of them." *** **DK**, **None**, **[]**, all mean none of the actual labels are correct.

tial relation triplets (Landmark, Spatial-indicator, trajector) for each fact in a story the applies spatial rules over these extracted triplets and report all possible relations between two asked objects. Finally, it checks whether the asked relation existed in the find relation. This model solves task 17 of the bAbI with 100% accuracy.

To implement the neural network approach, we use huggingface implementation of pre-trained BERT (Devlin et al., 2019). We apply a boolean classifier on the output of "[CLS]" token from the last layer of BERT model for each "Yes" and "No" answers (the same as model used on YN question types.) We use Adamw (Loshchilov and Hutter, 2017) optimizer and $2e-6$ learning rate with negative log-likelihood loss objective and train the model on the 10k, 5k, 2k, 1k, 500, and 100 portion of bAbI's training questions. The model yields 100% accuracy on 10k, and 5k and 99% accuracy on 2k and 1k training samples.

Type	**Original Set**	**Unseen Set**
Shapes	Square, Circle, Triangle	Rectangle, Oval, Diamond
Relations	Left, Right, Above, Below	Left side, Right side, Top, Under
Colors	Yellow, Black, Below	Green, Red, White
Size	Small, Medium, Big	Little, Midsize, Large

Table 9: Modifications on the unseen set

Figure 10 shows an example of boolQ dataset. To Answering the questions of this dataset, we use the same setting as neural network model on bAbI to further fine-tune BERT on boolQ.

Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A: Yes. [An example event is given.]

Q: Does France have a Prime Minister and a President?
P: ... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A: Yes. [Both are mentioned, so it can be inferred both exist.]

Figure 10: An example of boolQ dataset.

```
sentence: "Medium blue square number one is touching the bottom edge of this block."
▾ spatial_description: [] 1 item
  ▾ 0:
    ▾ trajector:
        phrase: "medium blue square number one"
        head: "square"
      ▾ properties:
          color: "blue"
          size: "medium"
          name: "number one"
          number: ""
        spatial_property: ""
      ▾ SOT:
          start: 167
          end: 195
    ▾ landmark:
        phrase: "the bottom edge of this block"
        head: "block"
      ▸ properties:
        spatial_property: "the bottom edge"
      ▸ SOT:
    ▾ spatial_indicator:
        phrase: "touching"
        spatial_value: "TPP"
        g_type: "Region"
        s_type: "RCC8"
        polarity: false
        FoR: "Relative"
      ▸ SOT:
```


Figure 11: SpRL annotation for an example sentence from SPARTQA.

Examples	Features
Block A is above Block C **and** B.	Using conjunction to describe relation between more than two blocks.
The small circle is **above** the yellow square **and** the big black shape.	Using conjunction to describe relationships between more than two objects.
The yellow square number one is to the **right** of **and above** the blue circle.	Using conjunction for more than one relation.
Block B has **two medium yellow squares** and **two blue circles**.	Describing a group of objects with the same properties. In the next sentences, they are mentioned by an asigned number. For example, the blue circle number two.
The blue circle is below the object **which is to the right** of the big square.	Using nested relations between objects in their description.
A small blue circle is near to the big circle. **It** is to the left of the medium yellow square.	Using coreferences for an entity described in the previous sentences.
There **is a** block named A. One small yellow square **is** touching the bottom edge of this block.	The verb matches the number of the subject.
What is the relation between black **object** and a big circle?	Using shape, object, and thing, which are a general description of an object. It could be the "black triangle" or the "black circle" mentioned in the story.

Table 10: Particular features of the dataset