More data and training: an improved baseline for Language Models as Unreliable Spatial Reasoners

Yimin Chen, Jiaxuan Lan, Qiuyi Wei
{yc3294, jl13072, qw2316}@nyu.edu

Abstract

Our main task is to construct a natural language inference benchmark to test the spatial reasoning capabilities of various language models. We focus on evaluating language models on a comprehensive suite of spatial reasoning tasks. These tasks are divided into four categories: motion, orientation, distance, and containment. To complete the tasks, we apply templates to generate thousands of NLI questions and put them into target models. For the extension, we add more non-exhaustive example words and phrases by constructing more complex templates programmatically (with more subjects and localities). The results demonstrate that all families of language models perform poorly on most of the spatial reasoning tasks and these models lack the understanding of the spatial relationship among objects. With more objects being put in, the performance of language models gets even worse.

1 Introduction

Natural language inference is more like recognizing the textual entailment, which enables the computer to identify whether or not the given hypothesis and premise are logically similar, contradicting, or hard to determine. For example, 'There are thousands of men and women in this castle' and 'nobody is in the castle' is a contradiction. Natural language inference is an inevitable and important task because we can develop more precise models by enabling the model to understand the dependencies between sentences. As natural language inference is a classification task, the accuracy rate can be utilized to objectively and effectively evaluate the model's quality. It is easy to focus on semantic understanding and representation, such as generating a good sentence or a suitable sentence vector. Then the result can be more conveniently migrated to other downstream tasks or serve other natural language research.

Language models have shown a large potential power on lots of tasks (Zellers et al., 2018; Durme et al., 2018; Singh et al., 2021). For example, GPT-3, known as generative pre-trained transformer 3, leverages deep learning to create human-like text. GPT-3 is trained to predict the next word in a sentence and obtains a magnificent achievement. In natural language processing, GPT-3 has demonstrated impressive capabilities in writing articles, translating context, and generating codes. It can even learn a person's language pattern and follow this pattern to talk to people. There exists plenty of efficient language models like GPT-3, and these models are tested by massive commonsense reasoning datasets such as taxonomic, spatial-temporal, and qualitative reasoning. In other words, this is hard to study the different types of reasoning individually and why the models perform well on some specific tasks. However, the capabilities of these models remain in question, which requires the creation of increasingly adversarial datasets (Zellers and Cho, 2019; Lin et al., 2021).

Our main goal is to explore one specific reasoning type: spatial reasoning. The existing methods of dataset generation depend on manual generation with some crowd-sourcing which triggers the difficulty of deep study of spatial reasoning. We construct both simple and complex structures of sentence generation on spatial reasoning by synthetic test suits and try to explore different performances of various language models. Our basic hypothesis is that the models will obtain worse performance on all spatial reasoning tasks with the increasing complexity of sentence length and structure.

We split different spatial reasoning and test model individually with each reasoning type to ensure it is easy to diagnose the natural language inference ability of different spatial reasoning types in each model. With the domain of physical commonsense reasoning, mostly spatial reasoning, We-

ston introduces bAbI (Weston et al., 2015), a set of 20 toy tasks that aim to test a model's ability to reason about the properties and relationships of objects in a variety of contexts. Out of the 20 different tasks, Task 17 (Positional reasoning), Task 18(Reasoning about the size), and Task 19(Path-finding) are closely related to spatial reasoning. Our main hypothesis is that all models are not good learners of spatial reasoning.

2 Methods

2.1 Test Suits

As mentioned above, we divided our experiment into Basic (simple) and bAbI (complex) parts. For the Basic part, we fed auto-generated sentences to multiple models to check their performances toward spatial reasoning. For the bAbI part, we generated sentences similar to the ones in the basic part and then fed them to those previous models to check whether their accuracy changed for more complex reasoning tasks. We didn't use the bAbI dataset because it contains too many unrelated tasks, and these tasks contain duplicated premise-hypothesis pairs.

Our Basic test suite contains 33718 pairs of premise and hypothesis (Table 1). These premise-hypothesis pairs can be classified into four types: Motion, Orientation, Distance, and Containment. To generate these premise-hypothesis pairs, we set a basic sentence structure: "{A} is {lexical input} to {B} ", where {A} and {B} are sets of agents. These agents can be people's names (like "James"), buildings (like "the school"), and items (like "a pencil"). The {lexical input} can be different phrases (like "contained in", "north of the", and "close to"). Using this sentence structure, we are able to generate a tremendous amount of spatial reasoning sentences by varying agents and lexical inputs. These samples allow us to test models' performance toward different types of spatial reasoning.

bAbI is introduced as a set of prerequisite toy tasks that aim to produce and build a more reliable dialogue agent. With reference from previous studies and the basic part, we generated a bAbI test suite which contains 29640 premise and hypothesis pairs of similar structure. The only difference is that we added an additional agent {C} to the premise and hypothesized on the spacial relationships between {A} and {C} (Table 1). Also, we didn't create the distance reasoning type for bAbI because it would be ambiguous. For example, when saying {A} is near {B}{B} is near {C} it would be hard to determine the relationship between {A} and {C}Ṫhus all the expected results in this category would be neutral, and it would be meaningless. Consequently, we only focused on the reasoning types: motion, orientation, and containment.

3 Results

We feed our Basic and bAbI test suite to T5, DeBERTa, RoBERTa, ALBERT, UnifiedQAv2, XLNet, and GPT-3 models. Results of accuracy predicted by the best performing model of basic and bAbI test suite are shown in table 2. The ACCwp represents the average accuracy score for each model across all templates with partial credit.

From table 2 , we can see that the best performing model is GPT-3, with ACCpc approximately 59.4% trained with Davinci using 175B parameter. Interestingly, most models using the bAbI test suite performed poorly compared to those using the basic test suite. The highest mean accuracy of the best model of the basic test suite is DeBEERTa with 73.1%, while in bAbI, the best accuracy is just GPT-3 with 59.4%. Moreover, larger parameters tend to perform better for the basic test suite. However, for the bAbI test suite, models like T5, DeBERTa, and RoBERTa perform better with models trained with smaller parameters.

Also, their performance varies significantly across different types of spatial reasoning. Therefore, we used ACCwp to estimate their performance toward different types of spatial reasoning (Figure 1, Figure 2).

3.1 Basic

Most models using the basic dataset resulted in an accuracy of around 40% to 60% (Figure 1). DeBERTa models have the best performance toward each type of spatial reasoning. Considering each type of spatial reasoning separately, DeBERTa models have the highest accuracy rate for motion (around 80%), distance (around 70%), orientation (around 45%), and containment (around 55%) reasoning. Among all models, GPT-3 has the most stable performance across different types of spatial reasoning (around 40% accuracy rate). However, the GPT-3 models' performances are much worse than other models for most types of spatial reasoning.

Additionally, the RoBERTa, XLNet, and T5

	Type	Input	Expected
Basic	Orientation	**P**: The {bar} is {north} of the {library}.	
		H: The {library} is {south} of The {bar}.	entailment
		H: The {library} is {north, east, west} of the {bar}.	contradiction
	Motion	**P**: {James} is {running}.	
		H: {James} is {in motion, stationary}	entailment
		H: {James} is {not moving, not stationary}	contradiction
	Containment	**P**: The {pen} is {in} the {pencil case}.	
		H: The {pencil case} {fit in, contains} the {pen}.	entailment
		H: The {pencil case} {cannot fit in} of the {pen}.	contradiction
	Distance	**P**: The {baseball} is {touching} the {pond}.	
		H: The {pond} is {touching} of the {baseball}.	entailment
		H: The {pond} is {close to, nearby} of the {baseball}.	neutral
bAbi	Orientation	**P**: The {bar} is {north} of the {library}.	
		P: The {library} is {east} of the {gym}.	
		H: The {bar} is {northeast, north, east} of the {gym}.	entailment
		H: The {bar} is {northwest, south, ...} of the {gym}.	contradiction
	Motion	**P**: {James} is {running}.	
		P: {Sophia} is {swimming}.	
		H: {James} is {in motion}.	entailment
		H: {Sophia} is {not in motion}.	contradiction
	Containment	**P**: The {refrigerator} is {larger} than the {computer}.	
		P: The {computer} {cannot fit} in the {small box}.	
		H: The {small box} {cannot contain} the {refrigerator}.	entailment
		H: The {refrigerator} {can fit} in the {small box}.	contradiction

Table 1: Comparison of premises (P) and hypotheses (H) between the Basic and the newly generated bAbI test suite. Curly braces indicate lexical items that can be substituted programmatically using templates, with some non-exhaustive example words and phrases.

	Model	Size(param)	Acc_{wpc}
Basic	GPT-3	XL(175B)	59.5%
	UnifiedQAv2	XL(3B)	39.8%
	T5	XL(3B)	60.3%
	DeBERTa	XL(700M)	**73.1**%
	RoBERTa	L(355M)	66.0%
	ALBERT	L(17M)	38.5%
	XLNet	B(110M)	56.5%
bAbI	GPT-3	XL(175B)	**59.4**%
	UnifiedQAv2	XL(3B)	38.5%
	T5	S(60M)	53.4%
	DeBERTa	B(86M)	52.4%
	RoBERTa	B(125M)	58.8%
	ALBERT	L(17M)	39.7%
	XLNet	B(110M)	58.2%

Table 2: Mean model accuracy with partial credit of the Basic and bAbI test suite, averaged over all reasoning categories, for the best-performed model in each family.

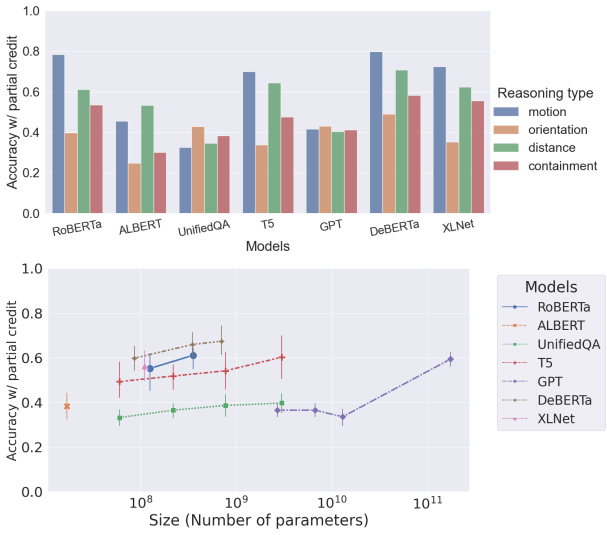

Figure 1: Detailed results of models using basic dataset. Line plot represents ACCwp of models with different size and bar plot represents the mean ACCwp of models under different reasoning type.

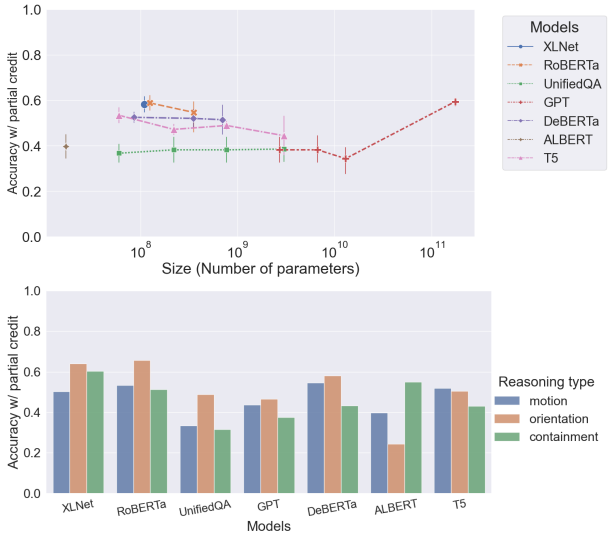

Figure 2: Detailed results of models using bAbI dataset. Line plot represents ACCwp of models with different size and bar plot represents the mean ACCwp of models under different reasoning type.

models demonstrate high performances toward motion (70%-80% accuracy rate) and distance (60%-70% accuracy rate) reasoning. All models' performance increases as the models' size increases (Figure 1). We believe that when using the basic dataset, larger models are associated with better performance in all types of spatial reasoning.

3.2 bAbI

Figure 2 shows the result when using the bAbI dataset to test the models. In the line plot, we can see that the accuracy of all the model is around 40% to 55%, which is lower than our basic dataset. There is not a specific increasing trend when increasing the parameter used to train the model. Only GPT-3's performance increased significantly when using its largest model.

Also, when looking down to specific reasoning types in the bar plot, models such as UnifiedQA and RoBERTa seems to handle orientation reasoning relatively well compared to the poor performance on motion and containment. ALBERT model is another special case where orientation is the worst reasoning type, and containment is the best. In other models, it is usually the opposite.

4 Discussion

The results support our hypothesis that all models are not good learners of spatial reasoning. When testing on different models using both basic and bAbI datasets, none of the models perform satisfactorily (Table 2). Among all types of spatial reasoning using the basic dataset, motion is the reasoning type that most models did well at, and most models perform relatively poorly for orientation spatial reasoning. Compared with the previous experiment, the trend of this current experimental data results is similar, meaning that the language models are bad spatial reasoning learners, which is same as our hypothesis. Additionally, when using the bAbI dataset, there isn't a specific trend of increase in accuracy as the model size increases, which is different from using the basic dataset and the previous study (Figure 2). Models such as DeBERTa and T5 perform better using smaller parameters than the basic test suite. Also, The averaged accuracy of different reasoning types among all model sizes using the bAbI dataset is lower than those using the basic dataset, which can also prove our hypothesis.

5 Conclusion

We created two spatial reasoning test suites to evaluate neural networks' performance toward motion, distance, orientation, and containment reasoning. We believe these four types of spatial can be a good representation of all types of spatial relationships used in daily life. After feeding our test suites to multiple T5, DeBERTa, RoBERTa, ALBERT, UnifiedQAv2, XLNet, and GPT-3 models with different parameters, we realized that for the basic test suite, models with more parameters tend to have better performance. However, smaller models tend to perform better for the bAbI test suite. Besides, our results support our hypothesis that models' performance decreases when sentences' complexity increases. The accuracy rates for the bAbI test suite are all lower than the basic test suite's (Table 2). However, even those state-of-the-art models have relatively high accuracy rates toward specific types of spatial reasoning. All models' overall accuracy toward all types of spatial reasoning is still very low (Figure 1, Figure 2). Consequently, we reached the same conclusion from previous studies that current state-of-the-art language models still need to be improved in conducting spatial reasoning analysis, even with tremendous parameters and training.

6 Contribution

Jiaxuan Lan and Qiuyi Wei were primarily responsible for constructing the suite of tests. All group members helped with testing the models, analyzing the results and writing.

7 Future work

- Construct and test the performance of spatial reasoning QA task or multiple choice questions on GPT or UnifiedQA models
- Test on real-world dataset and compare to see any difference.

References

Benjamin Van Durme, Jianfeng Gao, Jingjing Liu, Kevin Duh, Sheng Zhang, and Xiaodong Liu. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. Com2sense: A commonsense reasoning benchmark with complementary sentences.

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.

Rowan Zellers and Ari Holtzman. Yonatan Bisk. Ali Farhadi. Yejin Cho. 2019. Hellaswag: Can a machine really finish your sentence?

Rowan Zellers, Roy Schwartz, Yejin Choi, and Yonatan Bisk. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference.

8 Appendix

Model		# parameters
GPT-3	B	2.7B
	M	6.7B
	L	13B
	XL	175B
UnifiedQAv2	S	60M
	B	220M
	L	770M
	XL	3B
T5	S	60M
	B	220M
	L	770M
	XL	3B
DeBERTa	B	86M
	L	350M
	XL	700M
RoBERTa	B	125M
	L	355M
ALBERT	L	17M
XLNet	B	110M

Table 3: The number of parameters for each model