Dual-Phase Models for Extracting Information and Symbolic Reasoning: A Case-Study in Spatial Reasoning

Roshanak Mirzaee[1], Parisa Kordjamshidi[1]

[1]Department of Computer Science and Engineering, Michigan State University, MI, USA

Abstract

Spatial reasoning over text is challenging as the models need to extract the direct spatial information from the text, reason over those, and infer implicit spatial relations. Recent studies highlight the struggles even large-scale language models encounter when it comes to performing spatial reasoning over text. In this paper, we explore the potential benefits of disentangling the processes of information extraction and reasoning in models to address this challenge. To explore this, we devise various models that disentangle extraction and reasoning (either symbolic or neural) and compare them with SOTA baselines with no explicit design for these parts. Our experimental results consistently demonstrate the efficacy of disentangling, showcasing its ability to enhance models' generalizability within realistic data domains.

Keywords

Spatial Reasoning, Spatial Role Labeling, Disentangling Extraction and Reasoning, Pretrained Language Models

1. Introduction

Despite the high performance of recent pretrained language models (PLMs) on question-answering tasks, solving questions that require multi-hop spatial reasoning is still challenging [1, 2]. Here, we aim to address the limitations of end-to-end PLMs [2] and capitalize on the advantages of fine-grained information extraction [3, 4, 5, 6] in solving Spatial Question Answering (SQA). Thus, we propose models which disentangle the *language understanding* and *spatial reasoning* computations as two separate components. Specifically, we first design a pipeline model, called PistaQ shown in Figure 1-b, that includes trained neural modules for extracting direct fine-grained spatial information from the text and performing symbolic spatial reasoning over them.

We compare this pipeline with two additional models that utilize the same amount of supervision and modules as in the pipeline model. The first model, named BERT-EQ, is simply an End-to-End PLM (Figure 1-a) that uses annotations used in extraction modules in the format of *extra QA* pairs. This model aims to demonstrate the advantages of using separate extraction modules compared to a QA-based approach. The second model called SREQA shown in Figure 1-c, is an End-to-End PLM-based model on relation extraction tasks that has explicit neural layers to disentangle the extraction and reasoning inside the model. This model incorporates a neural spatial reasoner, which is trained to identify all spatial relations between each pair of entities.

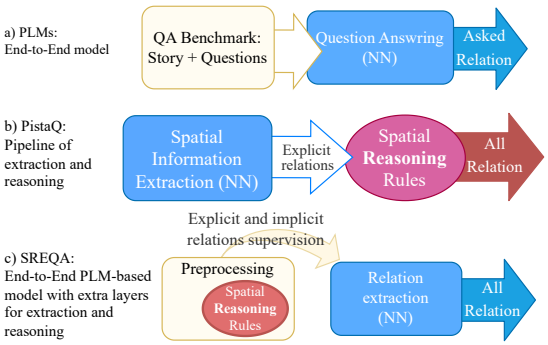

Figure 1: Various models to find the asked relations in spatial question answering task.

We evaluate the proposed models on multiple SQA datasets, SpartQA [2], SpaRTUN, and ReSQ [7] demonstrating the effectiveness of the disentangling approach in controlled and realistic environments. Our results show that disentangling extraction and reasoning benefits deterministic spatial reasoning and improves generalization in realistic domains despite the coverage limitations and sensitivity to noises in the symbolic reasoner. These findings highlight the potential of leveraging language models for the extraction task and emphasize the importance of utilizing explicit reasoning modules rather than solely depending on black-box neural models for reasoning.

2. Results and Conclusion

We devised a series of experiments utilizing PLMs for spatial information extraction coupled with a symbolic reasoner, proposed in [7], for inferring indirect relations.

STRL'23: Second International Workshop on Spatio-Temporal Reasoning and Learning, 21 August 2023, Macao, S.A.R

mirzaeem@msu.edu (R. Mirzaee); kordjams@msu.edu (P. Kordjamshidi)

Table 1
Results on auto-generated datasets. We use the accuracy metric for both YN (Yes/No) and FR (Find Relations) questions.

#	Models	SpaRTUN		SpartQA-Auto	
		YN	FR	YN	FR
1	Majority baseline	53.62	14.23	51.82	44.35
2	GT-PistaQ	99.07	99.43	99.51	98.99
3	BERT	91.80	91.80	84.88	94.17
4	BERT-EQ	90.71	N/A	85.60	N/A
5	SREQA	88.21	83.31	85.11	86.88
6	PistaQ	**96.37**	**94.52**	**97.56**	**98.02**

Table 2
Results on SpartQA-Human. We use accuracy on YN questions and average Precision (P), Recall (R), and Macro-F1 on FR question types. *Using SpaRTUN supervision for further training.

#	Models	YN	FR		
		Acc	P	R	F1
1	Majority baseline	52.44	29.87	14.28	6.57
2	GT-PistaQ	79.72	96.38	66.04	75.16
3	BERT	51.74	30.74	30.13	28.17
4	BERT*	48.95	60.96	49.10	**50.56**
5	GPT3^{Zero_shot}	45.45	40.13	22.42	16.51
6	GPT3^{Few_shot}	60.13	45.20	54.10	44.28
7	GPT3^{Few_shot}+CoT	62.93	57.18	37.92	38.47
8	BERT-EQ	50.34	-	-	-
9	BERT-EQ*	45.45	-	-	-
10	SREQA	53.23	15.68	13.85	13.70
11	SREQA*	46.96	18.70	25.79	24.61
12	PistaQ	**75.52**	72.11	35.93	46.80

To train the extraction modules, we adapt them through training on the corresponding or auxiliary datasets. The outcomes of our experiments provide noteworthy insights:

(1) Tables 1 and 2 show the performance of models on two datasets with controlled experimental conditions, SpaRTUN and SpartQA-Auto. As it is shown, PistaQ outperforms all PLM baselines and SREQA. Our observations in this setting demonstrate that disentangling extraction and symbolic reasoning compared to PLMs enhances the models' reasoning capabilities, even with comparable or reduced supervision. This result suggests that SpRL annotations are more effective in the PistaQ pipeline than when utilized in BERT-EQ in the form of QA supervision. Note that the BERT-EQ uses all the original dataset questions and extra questions created from the full SpRL annotations.

(2) We select ReSQ as an SQA dataset with realistic settings and present the result of models on this dataset in Table 3. The low performance of PistaQ is attributed to , first, the absence of integrating commonsense information in this model and, second, the errors in the extraction modules, which are passed to the reasoning modules. SREQA* surpasses the PLMs trained on QA and QA+SpRL annotation, showcasing the advantage of the design of this model in utilizing QA and SpRL data within explicit extraction layers and the data preprocessing. Also, the better performance of this model compared to PistaQ demonstrates how the end-to-end structure of SREQA can handle the errors from the extraction part while capturing some rules and commonsense knowledge from ReSQ training data that are not explicitly supported in the symbolic reasoner.

Table 3
Result on ReSQ. *Further training on SpaRTUN. The $Zero\_shot$ refers to evaluation without further training on ReSQ or CLEF training data.

#	Models	Accuracy
1	Majority baseline	50.21
2	BERT	57.37
3	BERT*$^{Zero\_shot}$	49.18
4	BERT*	63.60
5	GPT3^{Zero_shot}	60.32
6	GPT3^{Few_shot}	65.90
7	GPT3^{Few_shot}+CoT	67.05
8	BERT-EQ	56.55
9	BERT-EQ*$^{Zero\_shot}$	51.96
10	BERT-EQ*	61.47
11	SREQA	53.15
12	SREQA*$^{Zero\_shot}$	53.32
13	SREQA*	**69.50**
14	PistaQ^{CLEF}	41.96
15	PistaQ $^{SpaRTUN+CLEF}$	47.21
16	Human	90.38

Story:		a photo of a room with white walls , **two single beds** with a night table in between and **a picture** on the wall **above the beds** .
Question:		Are the beds below the picture? Answer: Yes
Story Facts:	BERT	0: ['a picture', 'the beds'], 2:['a'], 1: ['a picture', 'the wall'] Facts: right(2, 1), below(2, 0), near(2, 0)
	GPT3	3: ['two single beds', 'the beds'], 5: ['a picture'], 6: ['the wall', 'the beds'] Facts: above(5, 3), above(5, 6) ...
Queries:	BERT	below(0 , 0)? or below(0 , 1)?
	GPT3	below(3 , 5)? or below(3 , 6)?
Reasoning:	BERT	below(0 , 0) = False, below(0 , 1) = False → Answer = No
	GPT3	below(3 , 5) = True, below(3 , 6) = False → Answer = Yes

Figure 2: An example of using BERT-based SpRL and GPT3 as information extraction in PistaQ on a ReSQ.

(3) Recent research[1] show that powerful LLMs cannot perform well on SQA task. Similarly, our experiments, as shown in Tables 2 and 3 show the lower performance of GPT3.5 compared to humans and our models PistaQ and SREQA. However, we show that harnessing LLMs' potentials in information extraction [8] can yield significant benefits within the disentangled structure of extraction and reasoning. Figure 2 provides a comparison between the BERT-based SpRL extraction modules and GPT3.5 with few_shot prompting in PistaQ. It shows that GPT3.5 extracts more accurate information, leading to correct answers from the reasoning phase.

References

[1] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, P. Fung, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023. arXiv:2302.04023.

[2] R. Mirzaee, H. Rajaby Faghihi, Q. Ning, P. Kordjamshidi, 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, Association for Computational Linguistics, Online, 2021, pp. 4582–4598. URL: https://aclanthology.org/2021.naacl-main.364. doi:10.18653/v1/2021.naacl-main.364.

[3] D. Mollá, M. Van Zaanen, D. Smith, et al., Named entity recognition for question answering (2006).

[4] A. C. Mendes, L. Coheur, P. V. Lobo, Named entity recognition in questions: Towards a golden collection., in: LREC, 2010.

[5] D. Shen, M. Lapata, Using semantic roles to improve question answering, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 12–21. URL: https://aclanthology.org/D07-1002.

[6] H. R. Faghihi, P. Kordjamshidi, C. M. Teng, J. Allen, The role of semantic parsing in understanding procedural text, arXiv preprint arXiv:2302.06829 (2023).

[7] R. Mirzaee, P. Kordjamshidi, Transfer learning with synthetic corpora for spatial role labeling and reasoning, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6148–6165. URL: https://aclanthology.org/2022.emnlp-main.413.

[8] T. Shen, G. Long, X. Geng, C. Tao, T. Zhou, D. Jiang, Large language models are strong zero-shot retriever, arXiv preprint arXiv:2304.14233 (2023).