7-2023

Learning to ask clarification questions with spatial reasoning

Yang DENG
Singapore Management University, ydeng@smu.edu.sg

Shuaiyi LI

Wai LAM

Citation
DENG, Yang; LI, Shuaiyi; and LAM, Wai. Learning to ask clarification questions with spatial reasoning. (2023). *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 2023 July 23-27*. 2113-2117.
Available at: https://ink.library.smu.edu.sg/sis_research/9106

Learning to Ask Clarification Questions with Spatial Reasoning

Yang Deng
The Chinese University of Hong Kong
Hong Kong SAR, China
ydeng@se.cuhk.edu.hk

Shuaiyi Li
The Chinese University of Hong Kong
Hong Kong SAR, China
syli@se.cuhk.edu.hk

Wai Lam
The Chinese University of Hong Kong
Hong Kong SAR, China
wlam@se.cuhk.edu.hk

ABSTRACT

Asking clarifying questions has become a key element of various conversational systems, allowing for an effective resolution of ambiguity and uncertainty through natural language questions. Despite the extensive applications of spatial information grounded dialogues, it remains an understudied area on learning to ask clarification questions with the capability of spatial reasoning. In this work, we propose a novel method, named SpatialCQ, for this problem. Specifically, we first align the representation space between textual and spatial information by encoding spatial states with textual descriptions. Then a multi-relational graph is constructed to capture the spatial relations and enable spatial reasoning with relational graph attention networks. Finally, a unified encoder is adopted to fuse the multimodal information for asking clarification questions. Experimental results on the latest IGLU dataset show the superiority of the proposed method over existing approaches.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**; **Spatial and physical reasoning**.

KEYWORDS

Asking Clarification Questions, Spatial Reasoning

ACM Reference Format:
Yang Deng, Shuaiyi Li, and Wai Lam. 2023. Learning to Ask Clarification Questions with Spatial Reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592009

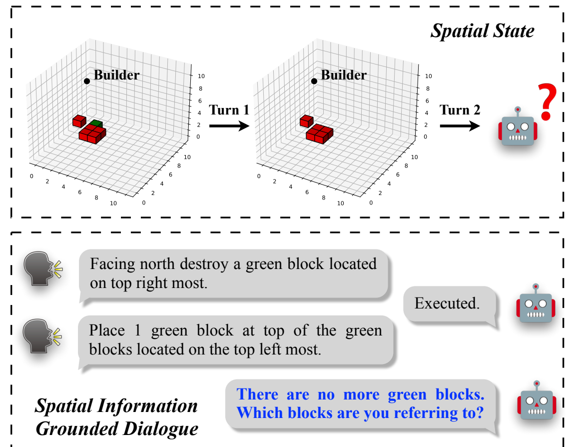

Figure 1: Asking clarification questions (blue) in spatial information grounded dialogues. Best viewed in color.

1 INTRODUCTION

Asking clarification questions, which aims to proactively resolve the ambiguity or uncertainty via natural language questions, has become an essential capability of different types of conversational systems, such as conversational search [2, 28], conversational question answering [4, 10], and conversational recommendation [21, 29]. In general, most studies solely target at identifying uncertainty within the textual context. Due to the extensive applications of conversational systems, several attempts have been made on learning to ask clarification questions grounded on multimodal contexts, such as tables [4], images [8], codes [6], etc. Recently, various conversational systems that involve spatial information emerge, such as conversational POI recommendation [14], multimodal conversational search [15], and embodied instruction-following dialogues [20]. In such dialogues, the uncertainty of context can be largely influenced by spatial information, which is typically represented as the location in a 2D or 3D coordinate. As shown in Figure 1, it is insufficient to determine the uncertainty of the user instruction in the embodied instruction-following dialogue without the current spatial state. Therefore, it attaches great importance to ask clarification questions with the capability of spatial reasoning between the textual instructions and the spatial world state.

There are two main challenges to be tackled for this problem: (1) ***How to bridge the gap between the spatial information and textual information?*** Existing works on spatial reasoning in text typically model the spatial information within a completely different vector space from the textual information and then combine two different types of representations by either concatenation [11, 12] or attention mechanism [22]. However, due to the sparsity in the spatial information, such approaches may turn out to introduce noise when handling clarification question selection that relies on the semantic measurement between the multimodal context and candidate questions. (2) ***How to enable spatial reasoning?*** There are two attributes that are essential to the spatial reasoning process [13], *i.e.*, distance and orientation between concerned objects. As the example in Figure 1, there is rich relational information about orientations during the conversation, such as "north", "top", "left", which plays a crucial role in the context understanding.

The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14200620).

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592009

In the light of these challenges, we propose a novel method for learning to ask clarification questions with spatial reasoning, named SpatialCQ. Specifically, we first represent each object in the spatial state with encoded textual descriptions, so that the spatial information can be learned in the same representation space as other textual information. Then we construct a multi-relational graph to model the interrelationships among different objects in the spatial state, where the object, the distance between objects, and the orientation are regarded as the node, the edge weight, and the relation, respectively. We further employ the relational graph attention network (RGAT) to refine the representations of objects with spatial relations. Finally, we adopt a unified encoder to fuse the multimodal information for clarification need prediction and clarification question selection.

In summary, our contributions are as follows:

- We propose a novel method, SpatialCQ, for asking clarification question with spatial reasoning, which constructs a multi-relation graph for representing 3D locations of objects and adopts RGAT to encode the graph-based spatial information.
- Experimental results on the IGLU dataset show that SpatialCQ effectively incorporates spatial information for improving the performance and outperforms existing approaches on both clarification need prediction and clarification question selection.

2 RELATED WORKS

Asking Clarification Questions. Asking clarification questions is firstly adopted to clarify the potential ambiguity in the user query in conversational information seeking [2, 28]. The problem is typically formulated by two subtasks [1]: clarification need prediction and clarification question generation. Clarification need predication is typically viewed as a binary classification problem for predicting whether the user query is ambiguous. If needed, clarification questions can be either selected from a question bank [1, 2, 30] or generated on the fly [7, 10, 28]. All aforementioned studies mainly target at asking clarification questions grounded on textual data. Some latest studies develop approaches for asking clarification questions based on multi-modal information. Shi et al. [22] propose the LearnToAsk method, which encodes the spatial information with a 3D convolutional neural network (CNN), to only identify the timing of clarifications during the instruction-following dialogues without producing actual clarification questions. To our knowledge, this work is the first attempt to study asking clarification questions grounded on spatial location data.

Spatial Information Grounded Dialogues. Recent years have witnessed several successful applications [14, 15] on conversational systems that are grounded on spatial information. Embodied instruction-following dialogues, which needs to consider both natural language interactions as well as the state of the environment, have become the most popular and widely-studied spatial information grounded dialogues. It covers a wide range of applications, such as collaborative building dialogues [19, 20, 22], navigation dialogues [3, 9], and object manipulation dialogues [8]. Most existing studies focus on the execution of natural language instructions [3, 9, 12], but in real-world applications, the user instructions are often ambiguous or missing necessary information. Mohanty et al. [19] construct the IGLU dataset for this problem, where the world state is provided as 3D locations. In this work, we investigate spatial reasoning methods for asking clarification questions in spatial information grounded dialogues.

Spatial Reasoning in Text. According to different problem settings, various techniques for spatial reasoning in text have been explored and studied [13]. For example, Yang et al. [27] and Jänner et al. [11] treat 2D map-like fully observable world states as the grounded context and process them using CNN. Some researchers further tried to expand the 2D context to 3D simulated environment [12, 22] that necessitates the ability to better learn the representations between cross-modal information. Unlike these settings, where text only constitutes the instructions, another line of spatial reasoning problems [17, 18, 23] focus on fully textual context and aim to understand of spatial concepts in natural language. This potentially brings exploitation of powerful pre-trained language models (PLMs), *e.g.*, BERT [5]. In this work, we focus on scenarios where contexts are described by a list of 3D locations and investigate the cross-modal representation learning with PLMs.

3 METHOD

3.1 Problem Definition

We follow the standard problem definition of asking clarification questions [1]. Given the instruction u from the user, the system first predicts the clarification need labels $l \in \{0, 1\}$, *i.e.*, whether it has sufficient information to execute the described instruction or further clarification is needed, based on the current world state s. If the clarification is needed, the system will select the most appropriate clarification question q for asking by ranking the candidate questions from the question bank. Under the setting of embodied instruction-following dialogues [19], the world state $s = \{(x_i, y_i, z_i, d_i)\}_i^N$ is represented as a list of 3D locations (x, y, z) of N objects with corresponding object descriptions d[1].

3.2 Encoding Spatial World State

3.2.1 Graph Construction with Spatial Relations. To facilitate the spatial reasoning process, it requires to model and aggregate the complex 3D spatial information. To this end, we construct a multi-relational graph to represent the location information of the world state obtained from different spatial relations. The multi-relational graph is denoted as $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{R})$, with nodes $n_i \in \mathcal{N}$, labeled edges (*i.e.*, relations) between node n_i and n_j as $(n_i, r, n_j) \in \mathcal{E}$, where $r \in \mathcal{R}$ is the relation type between two nodes. In our case, we treat each object in the world state as a node in $\mathcal{G}$, with the total number of nodes as N.

To represent the relative location information obtained from 3D directional relations, we employ three adjacency matrices associated with the graph $\mathcal{G}$, with respect to the distance between two objects in each dimension, *i.e.*, orientation. Accordingly, the relation types between two nodes is denoted as $r \in \mathcal{R} = \{\mathsf{x}, \mathsf{y}, \mathsf{z}\}$ that represent the north-south, left-right, and top-bottom relations. Three adjacency matrices can thus be constructed for $\mathcal{G}$:

$$A^{\mathsf{x}}_{i,j} = x_i - x_j, \quad A^{\mathsf{y}}_{i,j} = y_i - y_j, \quad A^{\mathsf{z}}_{i,j} = z_i - z_j, \tag{1}$$

where the edge weight represents the distance between two objects in the corresponding orientation.

[1] In IGLU, the object descriptions can be "builder", "red block", "green block", etc.

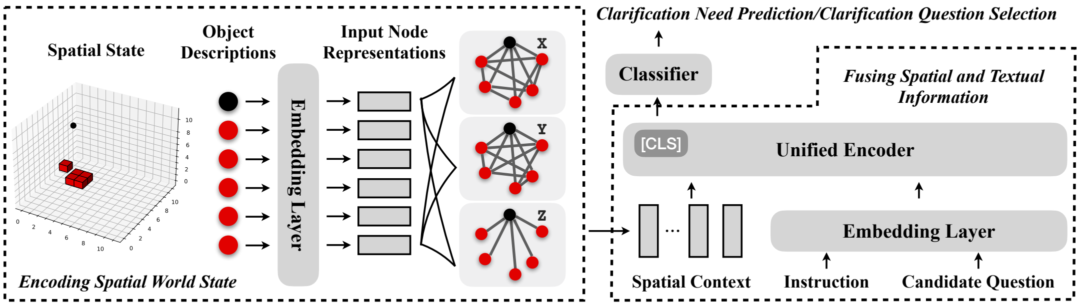

Figure 2: Overview of SpatialCQ.

3.2.2 Relational Graph Attention Network. In order to capture the information from multiple spatial relations with a multi-hop reasoning process, we utilize the Relational Graph Attention Network (R-GAT) to refine the node representations.

Following the graph attention mechanism proposed in [25], the attention weight $\alpha_{i,j}$ indicates the importance of node j's features to node i. For each relation $r \in \mathcal{R}$, we compute the relation-specific attention weights $\alpha^r_{i,j}$ as:

$$\alpha^r_{i,j} = \frac{\exp\left(\text{LeakyReLU}(\boldsymbol{A}^r_{i,j}\boldsymbol{\omega}_r^\top[\boldsymbol{W}_r e_i||\boldsymbol{W}_r e_j])\right)}{\sum_{k\in\mathcal{N}_i^r}\exp\left(\text{LeakyReLU}(\boldsymbol{A}^r_{i,k}\boldsymbol{\omega}_r^\top[\boldsymbol{W}_r e_i||\boldsymbol{W}_r e_k])\right)}, \quad (2)$$

where $\omega_r \in \mathbb{R}^{2d'_h}$ and $\boldsymbol{W}_r \in \mathbb{R}^{d'_h \times d_h}$ are parameters to be learnt for the relation r. e denotes the embeddings for the node. $\mathcal{N}_i$ denotes the set of the neighborhood nodes of node i. $||$ denotes the concatenation operation.

Similar to [24], we employ multi-head attention for the graph attention mechanism. Specifically, K independent attention weights can be calculated based on Equation (2), resulting in the following output node representation for the next layer:

$$e_i^{(l+1)} = \sigma\left(\sum\nolimits_{r\in\mathcal{R}}\frac{1}{K}\sum\nolimits_{k=1}^{K}\sum\nolimits_{j\in\mathcal{N}_i^r}\alpha^{r,k,(l)}_{i,j}\boldsymbol{A}^r_{i,j}\boldsymbol{W}^{(l)}_{r,k}e_j^{(l)}\right), \quad (3)$$

where $\alpha^{r,k,(l)}_{i,j}$ are normalized attention coefficients computed by the k-th head of attention for the relation r, and $W_{r,k} \in \mathbb{R}^{d'_h \times d_h}$ is the corresponding linear transformation matrix to be learnt. In particular, we denote the output node representations in the last layer of the graph attention network as $\hat{e}$:

$$\{\hat{e}_i\}_1^N = \{e_i^{(L)}\}_1^N = \text{RGAT}(\mathcal{G}), \quad (4)$$

where L is the number of graph layers, which can be regarded as the number of reasoning hops. Since each graph layer considers the relation between two adjacent objects in the world state, multiple graph layers can collectively measure the spatial interrelations among multi-hop connected objects in the world state.

3.3 Fusing Spatial and Textual Information

In order to project the spatial information and the textual information into the same representational space, we initialize the node representation $e_i^{(0)}$ in the graph $\mathcal{G}$ with the textual embeddings of the object description d_i by using the same embedding method as the textual input w (*e.g.*, the instruction or the candidate question). Then we adopt a unified encoder to fuse the multi-modal

Dataset	#Sample	Len(Inst.)	#Obj.	%Ambig.	%NS	%LR	%TB
Train	4779	20.19	9.04	14.1	34.2	42.7	51.2
Dev	683	18.54	8.84	10.2	31.8	37.3	41.0
Test	1366	19.63	8.95	10.8	34.6	40.8	48.2

Table 1: The statistics of the IGLU dataset. %NS/LR/TB denote the percentage of instructions that include north-south/left-right/top-bottom information, respectively.

information. Here we take BERT [5] as the encoder for example:

$$\boldsymbol{H} = \textbf{BERT}([e_{\texttt{[CLS]}}; \hat{e}_i; ...; \hat{e}_N; e_{\texttt{[SEP]}}; E_w; e_{\texttt{[SEP]}}]), \quad (5)$$

where $\boldsymbol{H}$ denotes the fused representation for spatial and textual information, and E_w denotes the concatenation of token embeddings of the input sequence w.

3.4 Asking Clarification Question

3.4.1 Clarification Need Prediction. The textual input w for clarification need prediction only includes the user instruction u. After obtaining the fused representation, we build a classifier, which contains a linear transformation and the softmax function, to predict the clarification need label l. The cross entropy is adopted as the objective function:

$$p = \textbf{Softmax}(W_n^\top \boldsymbol{H} + b_n), \quad (6)$$

$$\mathcal{L}_n = -\frac{1}{N}\sum\nolimits_{n=1}^{N}\left(l\log p + (1-l)\log(1-p)\right), \quad (7)$$

where $W_n \in \mathbb{R}^{d_h \times 2}$ and $b_n \in \mathbb{R}^2$ are parameters to be learnt, and d_h is the hidden size of the encoder.

3.4.2 Clarification Question Selection. Differently, the textual input additionally includes the candidate question q for measuring its appropriateness as the clarification question. Therefore, the fused representation can be learned by:

$$\boldsymbol{H} = \textbf{BERT}([e_{\texttt{[CLS]}}; \hat{e}_i; ...; \hat{e}_N; e_{\texttt{[SEP]}}; E_u; e_{\texttt{[SEP]}}; E_q; e_{\texttt{[SEP]}}]). \quad (8)$$

The classifier and the objective function for clarification question selection is the same as clarification need prediction in Eq.(6) and Eq.(7). The selected question is based on the ranked probability p.

4 EXPERIMENT

4.1 Experimental Setups

Datasets & Evaluation Metrics. We evaluate the proposed method on the IGLU dataset [19], which is collected by crowdworkers interacting with Minecraft. Every sample is initialized with a built world

Method	Clarification Need Prediction						Clarification Question Selection					
	Dev			Test			Dev			Test		
	P	R	F1	P	R	F1	MRR@5	MRR@10	MRR	MRR@5	MRR@10	MRR
BM25	-	-	-	-	-	-	0.4348	0.4434	0.4575	0.2373	0.2538	0.2710
$\text{BERT}_{\text{large}}$	0.7283	0.6161	0.6482	0.7649	0.7044	0.7243	0.4669	0.4797	0.4890	0.3204	0.3374	0.3535
$\text{RoBERTa}_{\text{large}}$	0.7176	0.6174	0.6460	0.8034	0.6944	0.7345	0.5701	0.5794	0.5881	0.3882	0.4067	0.4202
BAP	0.4966	0.4988	0.4879	0.5742	0.5047	0.4856	0.0214	0.0232	0.0422	0.0574	0.0691	0.0942
LearnToAsk (BERT)	0.7249	0.6092	0.6398	0.7661	0.7034	0.7328	0.4438	0.4587	0.4707	0.3111	0.3268	0.3473
LearnToAsk (RoBERTa)	0.7326	0.6166	0.6461	0.7963	0.6924	0.7285	0.5554	0.5641	0.5722	0.3765	0.3947	0.4075
SpatialCQ (BERT)	0.7391	**0.6298**	**0.6625**†	0.7784	**0.7111**	0.7383	0.4831	0.4954	0.5083	0.3391	0.3561	0.3724
SpatialCQ (RoBERTa)	**0.7486**	0.6243	0.6587†	**0.8098**	0.7088	**0.7461**†	**0.5879**†	**0.5935**†	**0.6060**†	**0.4034**†	**0.4204**†	**0.4334**†

Table 2: Method comparisons. † indicates statistically significant improvement (p<0.05) over the best baseline.

state from collected multi-turn interactions data, containing the user instruction for the next turn and the current world state. Since only the training set of IGLU has been released[2], we provide a new data split by randomly splitting the training set into train-dev-test split as 7:1:2. The dataset statistics is presented in Table 1. Following previous studies [1, 19], we adopt Macro Precision, Recall, and F1 scores for the evaluation of clarification need prediction, and MRR for clarification question selection.

Compared Methods. Since there is no existing approach directly applied to ask clarification questions with spatial reasoning, we compare the proposed method with two groups of baselines: (i) General baselines with text-only inputs for asking clarification questions, including BM25 [19], BERT [1], and RoBERTa-based Ranker [16]. (ii) Several alternative baselines that can be adapted to the target problem with the capability of incorporating spatial information as follows:

- BAP [12] encodes the spatial information of the world state with a 3D CNN and the textual instructions with GRUs, where the object embeddings are randomly initialized.
- LearnToAsk [22] further improves BAP with a fusion module comprising four major components, two single modality modules and two cross modality modules, to learn contextualized representations for the world state and textual tokens.

Implementation Details. We use $\text{BERT}_{\text{large}}$ and $\text{RoBERTa}_{\text{large}}$ pretrained weights [26]. The learning rate and the dropout rate are set to be 1e-6 and 0.5, respectively. We train up to 15 epochs with mini-batch size 16, and select the best checkpoints based on the F1 score on the validation set.

4.2 Overall Performance

Table 2 summarizes the experimental results of two subtasks. Among the baselines, we observe that BAP barely works, indicating that it is difficult to capture semantic knowledge through simply concatenating textual and spatial representations from two different space. Although LearnToAsk can adopt PLMs, such as BERT and RoBERTa, which largely improve the performance, the spatial information is still underutilized. Compared with text-only models, LearnToAsk achieves only similar performance on clarification need prediction, but worse performance on clarification question selection. Since the clarification question selection task relies more on the semantic measurement between textual and spatial information, it is more sensitive to the spatial knowledge. Overall, SpatialCQ substantially and consistently outperforms all the baselines by effectively incorporating spatial reasoning into learning to ask clarification questions.

[2] https://github.com/iglu-contest/iglu-dataset

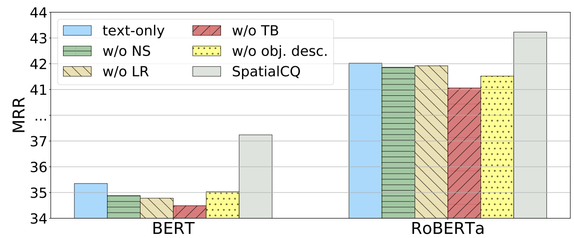

Figure 3: Ablation study on clarification question selection.

4.3 Ablation Study

To verify the effectiveness of the spatial reasoning and the multimodal fusion in SpatialCQ, we present the results of ablation studies in Figure 3. There are several notable observations as follows: (i) Ablating any orientation relation causes a noticeable performance decrease. Among them, the top-bottom relation (w/o TB) has the most significant impact. According to Table 1, the top-bottom relation is the most prevalent information in the instruction, while the other two relations are relatively close in the number of samples as well as the contribution to the final performance. (ii) When using randomly initialized embeddings for objects (w/o obj. desc.), the performance is even worse than their text-only counterparts, indicating that the spatial information is essentially noisy and SpatialCQ effectively aligns the multimodal information.

5 CONCLUSION

In this work, we propose a novel method, named SpatialCQ, for asking clarification questions with spatial reasoning. Specifically, we construct a multi-relational graph that encodes spatial states into textual descriptions for enhancing alignment of representation spaces between the two modalities. RGAT is then utilized for reasoning about spatial relations. Finally, a unified encoder is employed to combine the multimodal information for asking clarification questions. Evaluation results on IGLU dataset demonstrate remarkable advantages of our model compared with existing approaches.

REFERENCES

[1] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail S. Burtsev. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. 4473–4484.

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019*. 475–484.

[3] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*. 12538–12547.

[4] Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: Towards Proactive Conversational Question Answering over Tabular and Textual Data in Finance. *CoRR* abs/2210.08817 (2022).

[5] 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, NAACL-HLT 2019*. 4171–4186.

[6] Zachary Eberhart and Collin McMillan. 2022. Generating Clarifying Questions for Query Refinement in Source Code Search. In *IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2022*. 140–151.

[7] Chang Gao and Wai Lam. 2022. Search Clarification Selection via Query-Intent-Clarification Graph Attention. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022*. 230–243.

[8] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S. Sukhatme. 2022. DialFRED: Dialogue-Enabled Agents for Embodied Instruction Following. *IEEE Robotics Autom. Lett.* 7, 4 (2022), 10049–10056.

[9] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. 2022. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*. 7606–7623.

[10] Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-CoQA: Clarifying Ambiguity in Conversational Question Answering. In *3rd Conference on Automated Knowledge Base Construction, AKBC 2021*.

[11] Michaela Jänner, Karthik Narasimhan, and Regina Barzilay. 2018. Representation Learning for Grounded Spatial Reasoning. *Trans. Assoc. Comput. Linguistics* 6 (2018), 49–61.

[12] Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*. 2589–2602.

[13] Parisa Kordjamshidi, James Pustejovsky, and Marie-Francine Moens. 2020. Representation, Learning and Reasoning on Spatial Language for Downstream NLP Tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, EMNLP 2020*. 28–33.

[14] Changheng Li, Yongjing Hao, Pengpeng Zhao, Fuzhen Zhuang, Yanchi Liu, and Victor S. Sheng. 2021. Tell Me Where to Go Next: Improving POI Recommendation via Conversation. In *Database Systems for Advanced Applications - 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11-14, 2021, Proceedings, Part III*. 211–227.

[15] Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. 675–684.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019).

[17] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*. 4582–4598.

[18] Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. 6148–6165.

[19] Shrestha Mohanty, Negar Arabzadeh, Milagro Teruel, Yuxuan Sun, Artem Zholus, Alexey Skrynnik, Mikhail S. Burtsev, Kavya Srinet, Aleksandr Panov, Arthur Szlam, Marc-Alexandre Côté, and Julia Kiseleva. 2022. Collecting Interactive Multi-modal Datasets for Grounded Language Understanding. *CoRR* abs/2211.06552 (2022).

[20] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative Dialogue in Minecraft. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*. 5405–5415.

[21] Xuhui Ren, Hongzhi Yin, Tong Chen, Hao Wang, Zi Huang, and Kai Zheng. 2021. Learning to Ask Appropriate Questions in Conversational Recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 808–817.

[22] Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to Execute Actions or Ask Clarification Questions. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 2060–2070.

[23] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. StepGame: A New Benchmark for Robust Multi-Hop Spatial Reasoning in Texts. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. 11321–11329.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. 5998–6008.

[25] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018*.

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR* abs/1910.03771 (2019).

[27] Tsung-Yen Yang, Andrew S. Lan, and Karthik Narasimhan. 2020. Robust and Interpretable Grounding of Spatial References with Relation Networks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1908–1923.

[28] Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *WWW '20: The Web Conference 2020*. 418–428.

[29] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*. 177–186.

[30] Ziliang Zhao, Zhicheng Dou, Jiaxin Mao, and Ji-Rong Wen. 2022. Generating Clarifying Questions with Web Search Results. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 234–244.