A Spoken Dialogue System for Spatial Question Answering in a Physical Blocks World

Georgiy Platonov **Benjamin Kane** **Aaron Gindi** **Lenhart K. Schubert**
Department of Computer Science
University of Rochester
{gplatono, bkane2, agindi, schubert}@cs.rochester.edu

Abstract

A physical blocks world, despite its relative simplicity, requires (in fully interactive form) a rich set of functional capabilities, ranging from vision to natural language understanding. In this work we tackle spatial question answering in a holistic way, using a vision system, speech input and output mediated by an animated avatar, a dialogue system that robustly interprets spatial queries, and a constraint solver that derives answers based on 3-D spatial modeling. The contributions of this work include a semantic parser that maps spatial questions into logical forms consistent with a general approach to meaning representation, a dialogue manager based on a schema representation, and a constraint solver for spatial questions that provides answers in agreement with human perception. These and other components are integrated into a multi-modal human-computer interaction pipeline.

1 Introduction

Despite impressive recent advances of AI in specific, narrow tasks, such as object recognition, natural language parsing and machine translation, game playing, etc., there is still a shortage of multimodal interactive systems capable of performing high-level tasks requiring understanding and reasoning. The blocks world domain, despite its relative simplicity, motivates implementation of a diverse range of capabilities in a virtual interactive agent aware of physical blocks on a table, including visual scene analysis, spatial reasoning, planning, learning of new concepts, dialogue management and voice interaction, and more. In this work, we describe an end-to-end system that integrates several such components in order to perform a simple task of spatial question answering about block configurations. Our goal is dialogue-based question answering about spatial configurations of blocks on a table, in a way that reflects people's intuitive understanding of prepositional spatial relations. The system is able to answer questions such as *"Which blocks are touching some red block?", "Is the X block clear?", "Where is the Y block?",* etc. (where X and Y are unique block labels). Distinctive features of our work: (1) it is an end-to-end system using computer vision and spoken dialogue with an on-screen virtual human; (2) it did not require a large training corpus, only a modest development corpus using naturally posed spatial questions by a few participants; (3) it derives and relies on a 3D representation of the scene; (4) it models spatial relations realistically in terms of meaningful geometric and contextual constraints.

2 Related Work

Early studies featuring the blocks world include (Winograd, 1972) and (Fahlman, 1974), both of which maintained symbolic memory of blocks-world states. They demonstrated impressive planning capabilities, but their worlds were simulated, interaction was text-based, and they lacked a realistic understanding of spatial relations. Modern efforts in blocks worlds include work by Perera et al. (Perera et al., 2018), which is focused on learning spatial concepts (such as staircases, towers, etc.) based on verbally-conveyed structural constraints, e.g., *"The height is at most 3"*, as well as explicit examples and counterexamples, given by the user. Bisk et al. (Bisk et al., 2018) use deep learning to transduce verbal instructions into block displacements in a simulated environment. Some deep learning based studies achieve near-perfect scores on the CLEVR question answering dataset (Kottur et al., 2019; Mao et al., 2019). Common limitation of these approaches is reliance on unrealistically simple spatial models and domain-specific language formalisms, and in relation to our

Proceedings of the SIGdial 2020 Conference, pages 128–131
1st virtual meeting, 01-03 July 2020. ©2020 Association for Computational Linguistics

work, there is no question answering functionality or episodic memory. Our work is inspired by the psychologically and linguistically oriented studies (Garrod et al., 1999; Herskovits, 1985; Tyler and Evans, 2003). Studies of human judgements of spatial relations show that no crisp, qualitative models can do justice to those judgments. The study (Platonov and Schubert, 2018) explored computational models for prepositions using imagistic modeling, akin to the current work. Another study (Bigelow et al., 2015) applied imagistic approach to a story understanding task and employed Blender to create 3D scenes and reason about the relative configuration and visibility of objects in the scene.

3 Blocks World System Overview

Fig. 1, 2 depict our physical blocks world (consisting of a square table with several cubical blocks, two Kinect sensors and a display) and the system's software architecture[1]. The blocks are color-coded as green, red, or blue, and marked with corporate logos, serving as unique identifiers. The system uses audio-visual I/O: the block tracking module periodically updates the block positioning information by reading from the Kinect cameras and an interactive avatar, David, is used for human-machine communication. The block arrangement is modeled as a 3D scene in Blender, which acts as system's "mental image" of the state of the world. Google's Cloud Speech-To-Text API is used for the automatic speech recognition. Its output is processed to fix some common mistakes in the transcripts. The avatar is capable of vocalizing the text and displaying facial expressions, making the flow of conversation more natural than with textual I/O. The spatial component module together with the constraint solver is responsible for analyzing the block configuration with respect to the conditions implicit in the user's utterance. The Eta dialogue manager is responsible for *unscoped logical form* (ULF) generation (see subsection below) and controlling the dialogue flow and transition between phases, such as greeting, ending the session, etc.

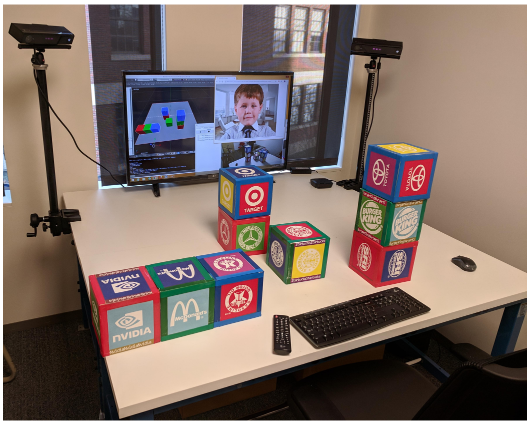

Figure 1: The blocks world apparatus setup.

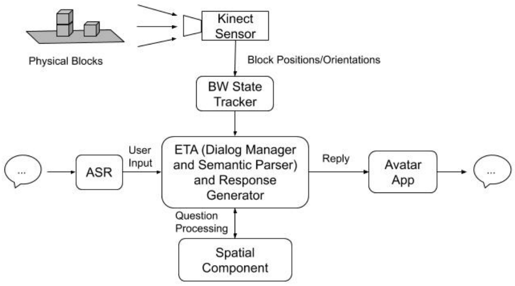

Figure 2: The blocks world dialogue pipeline. The arrows indicate the direction of interaction between the modules.

3.1 Eta Dialogue Manager and Semantic Parser

Eta is a dialogue manager (DM) designed to follow a modifiable dialogue schema, specified using a flexible and expressive schema language. The main contents of a dialogue schema are logical formulas with open variables describing successive steps (events) expected in the course of the interaction, typically speech acts by the system or the user. These are either realized directly as actions (with variables instantiated to particular entities), or, in the case of abstract actions, expanded into subschemas for further processing as the interaction proceeds.[2] A key mechanism used in the course of instantiating schema steps, including interpretation of user inputs, is *hierarchical pattern transduction*. Transduction hierarchies specify patterns at their nodes, with branches from a node providing alternative continuations as a hierarchical match proceeds. Terminal nodes provide result templates, or specify a subschema, a subordinate transduction tree, or some other result. The patterns are simple template-like ones that look for particular words or word features, and allow for "match-anything", length-bounded word spans.

Eta extends the approach implemented in the

[1] The code for Eta and the rest of the system can be found at https://github.com/bkane2/eta and https://github.com/gplatono/BlocksWorld

[2] Intended actions obviated by earlier events may be deleted.

LISSA system (Razavi et al., 2016, 2017). Like the latter, Eta derives English *gist clauses* by preprocessing the input. However, it is only used for handling casual aspects of dialogue such as greetings, and for "tidying up" some inputs in preparation for further processing. Additional regularization is done with a limited coreference module, which can resolve anaphora and referring expressions such as "it", "that block", etc., by detecting and storing discourse entities in context and employing recency and syntactic salience heuristics. This allows Eta to answer some simple follow-up questions like "Where is it now?" From the tidied-up inputs, Eta derives an *unscoped logical form* (ULF) (Kim and Schubert, 2019). ULF is closely related to the logical syntax used in schemas – it is a preliminary form of that syntax, when mapping English to logic. ULF differs from analogs, e.g., AMR, in that it is close to the surface form of English, covers a richer set of semantic phenomena, and does so in a type-consistent way. For example, ULF for the sentence "Which blocks are on two other blocks?" will be (((Which.d (plur block.n)) ((pres be.v) (on.p (two.d (other.a (plur block.n)))))) ?). Resulting ULF retains much of the surface structure, but uses semantic typing and adds operators to indicate plurality, tense, aspect, and other linguistic phenomena. We introduced recursion into hierarchical transduction trees to enable ULF derivation.

3.2 Spatial Relations

We model spatial relations as probabilistic predicates, using 3-D imagistic scene representations. Each predicate is composed of several factors, which represent basic relations that correlate with higher level spatial relation, e.g., if A is on top of B, then (usually) A is above B, and A is in contact with B. Thus, "above-ness" and contact serve as (some of the) factors used in determining "on-ness". After all the contributing factors are computed, their values are combined, e.g., by taking a linear combination, maximal value, etc., depending on the relation. Examples of factors are the scaled distance between centroids, frame size (size of the scene in context, important for judging relative distances), contact, support, certain shapes or types, proportion of the overlap of objects' projections onto the visual plane (for deictic sense of certain relations), etc. Not all factors potentially influencing a relation are relevant in a given situation, so we check various combinations of them that correspond to different usage patterns.

Some factors involve scene statistics, e.g., when determining nearness of A and B, the distribution of other objects is important. First, raw context-independent value is computed, which is then scaled up or down, depending on the raw scores for other objects, e.g., let $near\_raw(A, B) = 0.55$. If B is the closest object to A, i.e., $near\_raw(C, A) < 0.55, \forall C (C \neq B)$, we perceive B as the best near-object of A. Thus, the final score $near(A, B)$ will be boosted by a small (variable) amount.

4 Evaluation

We enlisted 5 volunteers, including native and non-native English speakers. The participants were instructed to ask spatial questions of the general type supported by the system, but without restriction on wording; before their first session they were shown a short demonstration of the expected kind of interaction with the system, including question-answer exchanges. Each session started with the blocks positioned in a row at the front of the table. The participants were instructed to move the blocks arbitrarily to test the robustness and consistency of the spatial models. During each session they were requested to ask 40-50 questions and mark system's answers as correct, partially correct or incorrect. They were asked to indicate separately if no answer could be given due to ASR errors or when the answer (regardless of correctness) seemed to be improperly or oddly phrased. The data are presented in Table 1.

Table 1: Evaluation data.

Total number of questions	388
Bad transcripts due to ASR errors	59
Well-formed transcripts (no ASR errors, or fixed)	329
Correct answers	219 (66.6% of 329)
Partially correct answers	45 (13.7%)
Incorrect answers	65 (18.8%)
The answer was given but sounded unnatural/ungrammatical	25

We found that the system returns correct answer in 67% of the cases. Including partially correct ones, the accuracy rises to 80%. Given that inter-annotator agreement of around 0.72 was observed in (Platonov and Schubert, 2018) for human judgements of prepositional relations on a 5-point Likert scale, our results are reasonable. Such variability is due to the fact that spatial relations are quite vague and people's intuitions differ significantly. Correctness was tracked for both the ULFs produced and the generated spoken answers. The spatial component displays satisfactory sensitivity in terms of the certainty cut-off threshold, i.e., the threshold determining which objects are included seems in accord with human intuitions. Below we present separate evaluation data for the ULF parser.

Table 2: Evaluation data on ULF parsing.

Total number of spatial questions	635
Number of correctly interpreted questions	470
Number of incorrectly interpreted questions	165
Number of incorrect parses due to ASR errors	87
Accuracy	74.02%
Percentage of incorrect parses due to ASR errors	52.73%

Most errors in the ULF parsing are due to either ASR errors, unsupported sentence constructions (e.g., passive voice expressions, some prepositions, etc.), or indexical questions (e.g., “What block did I just move?”).

5 Conclusion and Future Work

We have built a spatial QA system for a physical blocks world, already able to handle a majority of questions in dialogue mode. We are not aware of any other end-to-end system with comparable abilities in QA about spatial relations. Our spatial language model relies on intuitive computational models of spatial prepositions that come close to mirroring human judgments by combining geometrical information with context-specific information about the objects and the scene. This enables natural user-machine interaction. The ongoing work is targeting world history-tracking to enable answering question like “Where was the Toyota block initially?”

Acknowledgments

This work was supported by DARPA grant W911NF-15-1-0542. We thank our team of volunteers for their suggestions and contributions to system evaluation.

References

Eric Bigelow, Daniel Scarafoni, Lenhart Schubert, and Alex Wilson. 2015. On the need for imagistic modeling in story understanding. *Biologically Inspired Cognitive Architectures*, 11:22–28.

Yonatan Bisk, Kevin J Shih, Yejin Choi, and Daniel Marcu. 2018. Learning interpretable spatial operations in a rich 3d blocks world. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Scott Elliott Fahlman. 1974. A planning system for robot construction tasks. *Artificial intelligence*, 5(1):1–49.

Simon Garrod, Gillian Ferrier, and Siobhan Campbell. 1999. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189.

Annette Herskovits. 1985. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9(3):341–378.

Gene Louis Kim and Lenhart Schubert. 2019. A type-coherent, expressive representation as an initial step to language understanding. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 13–30.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

Ian Perera, James Allen, Choh Man Teng, and Lucian Galescu. 2018. Building and learning structures in a situated blocks world through deep language understanding. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 12–20.

Georgiy Platonov and Lenhart Schubert. 2018. Computational models for spatial prepositions. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 21–30.

S.Z. Razavi, M.R. Ali, T.H. Smith, L.K. Schubert, and M.E. Hoque. 2016. The LISSA virtual human and ASD teens: An overview of initial experiments. In *Proc. of the 16th Int. Conf. on Intelligent Virtual Agents (IVA 2016)*, pages 460–463, Los Angeles, CA.

S.Z. Razavi, L.K. Schubert, M.R. Ali, and H.E. Hoque. 2017. Managing casual spoken dialogue using flexible schemas, pattern transduction trees, and gist clauses. In *5th Ann. Conf. on Advances in Cognitive Systems (ACS 2017)*, Rensselaer Polytechnic Institute, Troy, NY.

Andrea Tyler and Vyvyan Evans. 2003. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. Cambridge University Press.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.