Registering Historical Context for Question Answering in a Blocks World Dialogue System *

Benjamin Kane, Georgiy Platonov, and Lenhart Schubert

University of Rochester, Rochester, NY 14627, USA
{bkane2,gplatono,schubert}@cs.rochester.edu

Abstract. Task-oriented dialogue-based spatial reasoning systems need to maintain history of the world/discourse states in order to convey that the dialogue agent is mentally present and engaged with the task, as well as to be able to refer to earlier states, which may be crucial in collaborative planning (e.g., for diagnosing a past misstep). We approach the problem of spatial memory in a multi-modal spoken dialogue system capable of answering questions about interaction history in a physical blocks world setting. We employ a pipeline consisting of a vision system, speech I/O mediated by an animated avatar, a dialogue system that robustly interprets queries, and a constraint solver that derives answers based on 3D spatial modelling. The contributions of this work include a semantic parser competent in this domain and a symbolic dialogue context allowing for interpreting and answering free-form historical questions using world and discourse history.

Keywords: Question Answering · Blocks World · Semantic Parsing · Discourse Context.

1 Introduction

Intelligent, task-oriented dialogue agents that interact with humans in a physical setting are a long-standing AI goal and have received renewed attention in the last 10 or 20 years. However, they have generally lacked the sort of recall of earlier discourse and perceived "world" situations and events—an episodic memory—needed to provide a sense of shared contextual awareness and, ultimately, a basis for diagnosing past errors, planning to re-achieve an earlier situation, repeating a past action sequence, etc.

The blocks world domain is an ideal setting for developing prototypes with such capabilities. In this work, we present a speech-based question-answering system for a physical blocks world featuring a virtual agent, that not only models and understands spatial relations but is able to register historical context and answer questions about the session history, such as *"Which block did I just move?"*, *"Where was the Toyota block before I moved it?"*, *"Did the Target block*

* This work was supported by DARPA grant W911NF-15-1-0542, NSF NRT Graduate Training Grant 2019-2020, and NSF EAGER Award IIS-1940981.

ever touch the Texaco block?", *"Was the Twitter block always between two red blocks?"*, etc. Since explicit storage of detailed successive scene models would be difficult to extend to general complex settings as well as being cognitively implausible (people seem to reconstruct past situations from high-level properties [15]), we maintain a compact symbolic record of changes to the world, allowing reconstruction of past states when combined with current spatial observations.

2 Related Work

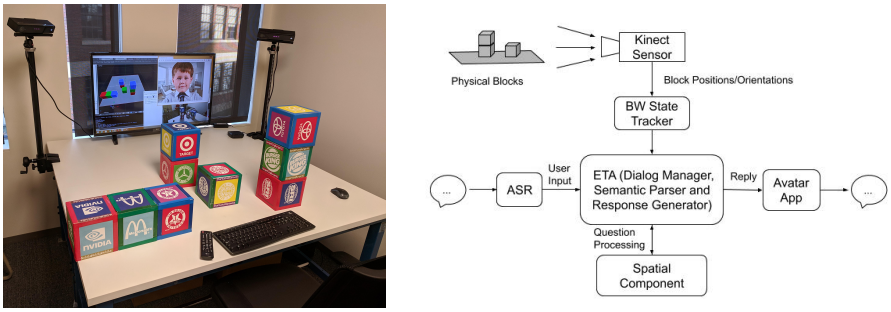
Early studies featuring the blocks world include [18] and [3], both of which maintained symbolic memory of blocks-world states. They demonstrated impressive planning capabilities, but their worlds were simulated, interaction was text-based, and they lacked realistic understanding of spatial relations. Modern efforts in blocks worlds include work by Perera et al. [13], which is focused on learning spatial concepts (staircases, towers, etc.) based on verbally-conveyed structural constraints, e.g., *"The height is at most 3"*, as well as explicit user-given examples and counterexamples. Bisk et al. [2] 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 [10, 12]. A common limitation of these approaches is reliance on unrealistically simple spatial models and domain-specific language formalisms, and in relation to our work, there is no question answering functionality or episodic memory. We are not aware of any recent study in a physical blocks world domain that makes use of spatial memory in answering questions about past states and events.

Outside of the blocks world domain, the TRAINS and TRIPS systems [5, 4] were noteworthy for their dialogue-based problem solving ability in a virtual map environment and their support of planning through temporal reasoning based on Allen Interval Logic [1]. A system aimed at human-like performance on a virtual reality map recall task [11] was based on the LIDA symbolic cognitive architecture and represented spatial context using a grid representation of the world and hierarchical "place nodes" with progressively updated activations. Recent deep-learning based approaches to modelling spatial episodic memory include [16] and [6]. The former uses an unsupervised encoder-decoder model to represent episodic memory as latent embeddings, and show that this model can allow a robot to recall previous visual episodes in a physical scene. The latter introduces a neuro-symbolic Structured Event Memory (SEM) model which is capable of segmenting events in video data and reconstructing past memory items. These methods, however, do not readily lend themselves to use for reasoning about historical relations or interactions in a blocks world question answering system.

3 Blocks World System and Eta Dialogue Manager

Fig. 1a, 1b 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. The blocks are color-coded as green, red, or blue, and marked with corporate logos which serve 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 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.



(a) Blocks world setup (b) Dialogue pipeline

Fig. 1: System overview.

The Eta dialogue manager (DM) is responsible for semantic parsing and dialogue control. Eta is designed to follow a modifiable dialogue schema, the contents of which are formulas in episodic logic [17] with open variables describing successive steps (events) expected in the course of the interaction. These are either realized directly as instantiated actions, or expanded into sub-schemas. [1]

In order to instantiate schema steps and interpret user inputs, the DM uses *hierarchical pattern transduction*, similarly to the mechanism used by the LISSA system [14] to extract context-independent *gist clauses* given the prior utterance. Transduction hierarchies specify patterns at their nodes to be matched to input, with terminal nodes providing result templates, or specifying a subschema. The pattern templates look for particular words or word features (including "wildcards" matching any word sequence of some length). Eta uses gist clause extraction for tidying-up the user's utterance, and then derives an *unscoped logical form* (ULF) [9] (a preliminary form of the episodic logic syntax of the dialogue schema) from the tidied-up input. ULF differs from similar semantic representations, e.g., AMR, in that it is close to the surface form of English, type-consistent, and covers a rich set of phenomena. To derive ULFs, we introduced semantic composition into the transduction trees. The resulting parser is quite efficient and accurate for the domain at hand. The input is recursively broken into constituents, such as a VP segment, until a lexical subroutine supplies ULFs for individual words, which are propagated back up and composed into larger expressions by the "calling" node. The efficiency and accuracy of the approach

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

lies in the fact that hierarchical pattern matching can segment utterances into meaningful parts, so that backtracking is rarely necessary.

An example of a transduction tree being used for parsing a historical question into ULF is shown and described in Figure 2. As can be seen from this example, the resulting ULF retains much of the surface structure, but uses semantic typing and adds operators to indicate plurality, tense, aspect, and other linguistic phenomena. Eta also has a limited coreference module utilizing syntactic constraints, recency, and other heuristics.

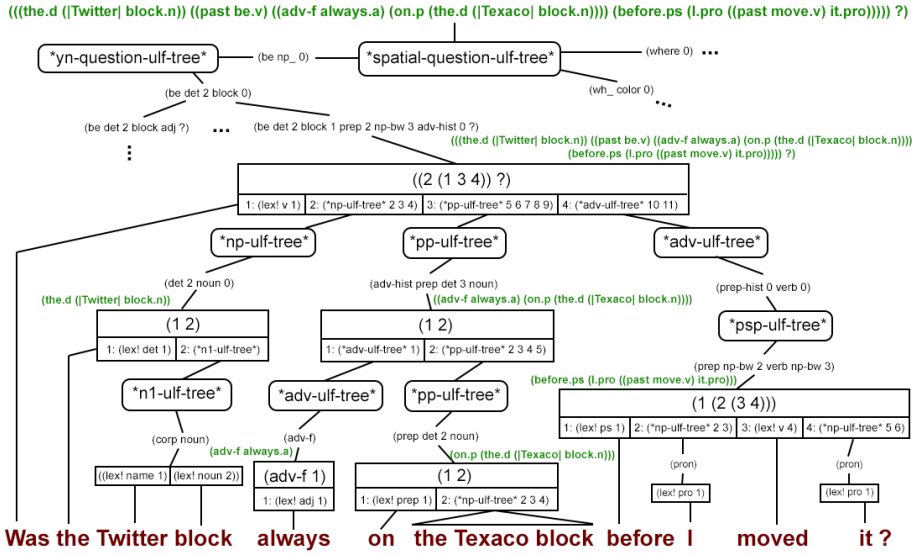


Fig. 2: An example ULF parse, with the input shown in red, and the resulting ULF (at each composition step) shown in green. The nodes with rectangles represent ULF composition nodes, where the numbers in the upper box correspond to the indices of the lower boxes (if no upper boxes, simple concatenation). All unframed nodes are patterns to be matched to the corresponding span of input text.

4 Historical Question-Answering

To answer historical questions, the DM needs to maintain a dialogue context including some sort of spatial episodic memory, so that the ULF obtained from parsing can be resolved into operations over this episodic memory. Based on the cognitive considerations discussed in [15], we maintain a high-level symbolic memory with which the agent can reconstruct past scenes, rather than a detailed visual or vector-based memory.

The vision system records the centroid coordinates and moves of blocks in real time. On the DM side, a “perceive-world” action in the schema causes the

DM to request ULF perceptions from the vision system. We rely on a simple linear, discrete time representation. The temporal entities (|Now0| etc.) are related to each other and to perceived actions propositionally, making use of the episodic operators described in [17]. Based on this context, the DM can efficiently reconstruct a scene at any past time by backtracking from perceived block locations, and evaluate approximate spatial relationships based on centroid coordinates. A simplified example is shown in the top half of Figure 3.

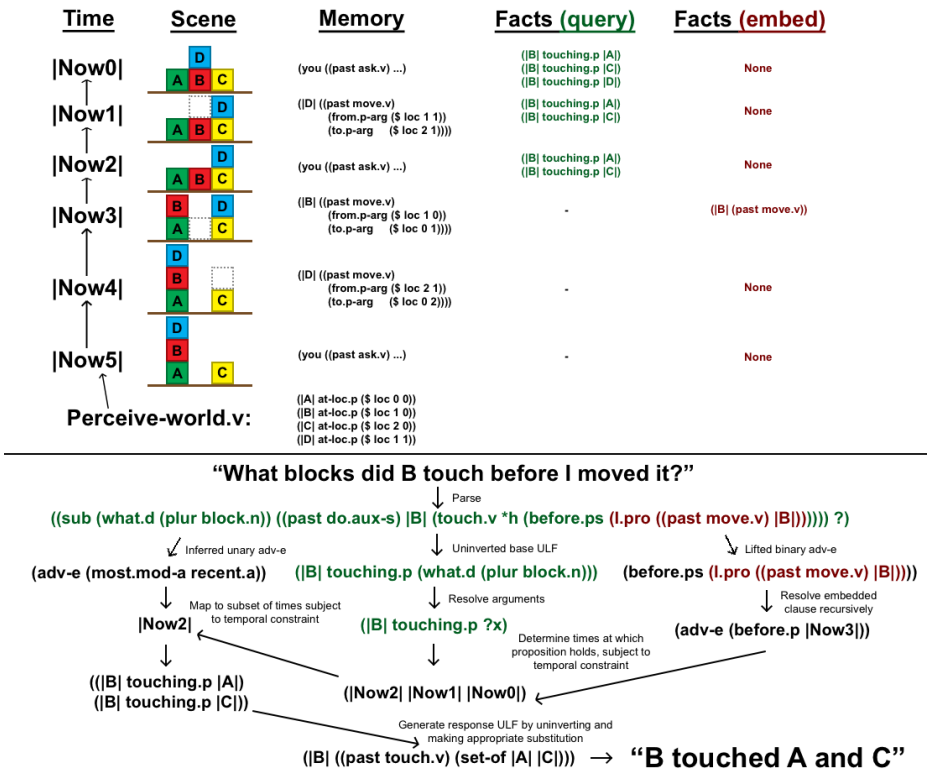


Fig. 3: A simplified example of how the context is represented and how the DM uses the context to compute relations given temporal constraints (top half), and an example of the DM determining an answer from a specific historical query (bottom half).

An example of answering a historical question given a ULF parse is shown in the bottom half of Figure 3. Phrases with head types "adv-e", "adv-f", and "ps", indicate temporal constraints that are applied during the scene reconstruction algorithm, and their semantic types allow them to be lifted to the sentence level. Furthermore, unary modifiers (including frequency modifiers) map a set of times to a subset of those times, whereas binary modifiers take two times and map to

a truth value. Any constraint may be further modified by a "mod-a" term (e.g. "just.mod-a"), which modifies how that mapping is applied.

Note that the example in Figure 3 is ambiguous; the answer could be "A, D, C" or "A, C". In fact, we found that many natural historical questions are similarly underspecified, presenting a major source of difficulty. To deal with this issue, the DM's pragmatic module attempts to infer temporal constraints in these ambiguous cases – in this particular example, Eta would infer the constraint "most recently". The algorithm shown extracts the uninverted base ULF relation, where arguments are represented as entities or variables (possibly with restrictions for noun modifiers). This base ULF, and any temporal constraints, are used to compute a list of times with attached facts through backtracking over past times. In the case of a binary constraint with a complex noun phrase or relative clause, this algorithm is applied recursively (as shown by the red constituents in Figure 3). The algorithm would likewise be applied recursively in the case of a query where the historical content is embedded in a noun phrase, e.g. "the first block that I moved".

Once a list of final times is obtained, an answer is generated by making the appropriate substitutions in the query ULF, applying syntactic transformations, and converting to surface form. If no relations are obtained, the DM's pragmatics module will attempt to respond to any presuppositions of the query, based on the work in [8]. For instance, if the query is "What block was the Twitter block on?", Eta will respond "The Twitter block wasn't on any block."

5 Evaluation and Discussion

Since the COVID-19 pandemic made testing the physical blocks world system on-site impossible, the authors had to resort to evaluating using a virtual environment that mirrors our setup, sans the physical block tracker and the audio I/O. However, as the crucial components being evaluated (parser, DM, and spatial context) were unchanged, the results should not be affected.

We enlisted 4 student volunteers for the user study, both native and non-native English speakers. The participants were instructed to move the blocks around and ask general questions about changes in the world, with no restrictions on wording. After the system displayed its answer, the participants were asked to provide feedback on its quality by marking it as correct, partially correct or incorrect. Each participant contributed about 100 questions or above (primarily historical questions, but also including some non-historical spatial questions). 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. The data is presented in Table 1. Non-historical questions, as well as a few malformed questions, were excluded when computing accuracy.

We find these results encouraging, given the pragmatic richness of the task and the unrestricted form of the questions. About 77% of Eta's answers were judged to be fully correct, with accuracy rising to 80% when including partially

Table 1: Evaluation data.

Total number of questions asked	496
Well-formed historical questions	387
Correct answers	297 (77% of 387)
Partially correct answers	13 (3% of 387)
Incorrect answers	77 (20% of 387)
Number of correctly parsed questions	363 (94% of 387)
Accuracy (correct + partially correct)	80%

correct answers. We find that the semantic parser itself is very reliable, with 94% of grammatical sentences being parsed correctly. Parsing failures accounted for a third of the incorrect answers.

Analyzing the remaining incorrect answers, we find that a major source of error is in the handling of under-specified historical questions, as described in section 4. There are many nuances to how humans naturally interpret these, that are difficult to capture with simple set of pragmatic rules. For example, Eta will plausibly interpret "What blocks did I move before the Twitter block?" as meaning "What blocks did I *recently* move before I moved the Twitter block?"; however if the user instead asked "How many blocks did I move before the Twitter block?", it seems that the questioner really means "How many blocks did I *ever* move before I moved the Twitter block?". Currently, Eta would add "recently" for the latter case, which would be incorrect. In future work, we aim to investigate this phenomenon further and improve the pragmatic inference module to handle these cases correctly. In addition, the tense structure in some more complex questions violated our simplifying assumption of discrete linear time. In future work, we plan to look into the use of more general temporal reasoning systems such as the tense trees described in [7] to enable more robust handling of different aspectual forms and more complex embedded clauses.

6 Conclusion

We have extended a spatial question answering system in a physical blocks world system with the ability to answer free-form historical questions using a symbolic dialogue context, keeping track of a record of block moves and other actions. A custom semantic parser allows historical questions to be parsed into a logical form, which is interpreted in conjunction with the context to generate an answer. We obtained an accuracy of 80%, which we believe is a strong result in view of the free-form and often underspecified nature of the historical questions that users asked, though it also leaves much room for improvement. Overall, we believe that the pragmatic richness and complexity that we've observed in historical question-answering indicates that further work towards representing episodic memory and enabling dialogue systems to reason about historical context will be fruitful in this sparsely researched area.

References

1. Allen, J.F., Ferguson, G.: Actions and Events in Interval Temporal Logic. Journal of Logic and Computation **4**(5), 531–579 (10 1994)
2. Bisk, Y., Shih, K.J., Choi, Y., Marcu, D.: Learning interpretable spatial operations in a rich 3d blocks world. In: 32nd AAAI Conference on Artificial Intelligence (2018)
3. Fahlman, S.E.: A planning system for robot construction tasks. Artificial intelligence **5**(1), 1–49 (1974)
4. Ferguson, G., Allen, J.: Trips: An integrated intelligent problem-solving assistant (08 1998)
5. Ferguson, G., Allen, J., Miller, B.: Trains-95: Towards a mixed-initiative planning assistant. pp. 70–77 (01 1996)
6. Franklin, N.T., Norman, K.A., Ranganath, C., Zacks, J.M., Gershman, S.J.: Structured event memory: a neuro-symbolic model of event cognition. bioRxiv (2019)
7. Hwang, C.H., Schubert, L.K.: Interpreting tense, aspect and time adverbials: A compositional, unified approach. In: Gabbay, D.M., Ohlbach, H.J. (eds.) Temporal Logic. pp. 238–264. Springer Berlin Heidelberg (1994)
8. Kim, G., Kane, B., Duong, V., Mendiratta, M., McGuire, G., Sackstein, S., Platonov, G., Schubert, L.: Generating discourse inferences from unscoped episodic logical formulas. In: Proc. 1st International Workshop on Designing Meaning Representations. pp. 56–65. ACL (Aug 2019)
9. Kim, G.L., Schubert, L.: A type-coherent, expressive representation as an initial step to language understanding. In: Proc. 13th International Conference on Computational Semantics-Long Papers. pp. 13–30 (2019)
10. Kottur, S., Moura, J.M., Parikh, D., Batra, D., Rohrbach, M.: Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. arXiv:1903.03166 (2019)
11. Madl, T., Franklin, S., Chen, K., Trappl, R.: Spatial working memory in the lida cognitive architecture. Proc. International Conference on Cognitive Modelling pp. 384–390 (01 2013)
12. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. arXiv:1904.12584 (2019)
13. Perera, I., Allen, J., Teng, C.M., Galescu, L.: Building and learning structures in a situated blocks world through deep language understanding. In: Proc. 1st International Workshop on Spatial Language Understanding. pp. 12–20 (2018)
14. Razavi, S., Schubert, L., Ali, M., Hoque, H.: Managing casual spoken dialogue using flexible schemas, pattern transduction trees, and gist clauses. In: 5th Ann. Conf. on Advances in Cognitive Systems (May 2017)
15. Rensink, R.: Scene Perception, pp. 151–155 (01 2001)
16. Rothfuss, J., Ferreira, F., Aksoy, E., You, Z., Asfour, T.: Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution. IEEE Robotics and Automation Letters **3** (01 2018)
17. Schubert, L., Hwang, C.: Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language (09 2000)
18. Winograd, T.: Understanding natural language. Cognitive psychology **3**(1), 1–191 (1972)