Generating Justifications in a Spatial Question-Answering Dialogue System for a Blocks World

Georgiy Platonov Benjamin Kane Lenhart K. Schubert

Department of Computer Science
University of Rochester
{gplatono, bkane2, schubert}@cs.rochester.edu

Abstract

As AI reaches wider adoption, designing systems that are explainable and interpretable becomes a critical necessity. In particular, when it comes to dialogue systems, their reasoning must be transparent and must comply with human intuitions in order for them to be integrated seamlessly into day-to-day collaborative human-machine activities. Here, we describe our ongoing work on a (general purpose) dialogue system equipped with a spatial specialist with explanatory capabilities. We applied this system to a particular task of characterizing spatial configurations of blocks in a simple physical Blocks World (BW) domain using natural locative expressions, as well as generating justifications for the proposed spatial descriptions by indicating the factors that the system used to arrive at a particular conclusion.

1 Introduction

While black box models like GPT-3 (Brown et al., 2020) demonstrate impressive performance on a variety of isolated benchmarks, they are still subject to significant drawbacks (Marcus, 2018). In particular, as AI systems reach wider adoption, explainability and interpretability become critical features. It is our belief that, instead of focusing exclusively on bigger datasets and models, or cross-modal learning, a somewhat different approach is required, viz., replacement of "tabula rasa" black boxes with architectures that utilize structured representations based around general reasoning, while in a form still amenable to deep learning techniques.

Below, we describe our ongoing work on a system composed of a general-purpose dialogue manager and a spatial specialist module, capable of generating spatial descriptions of configurations in the physical Blocks World domain and supplying justifications of its spatial descriptions. The domain contains several uniquely named blocks placed on a table, where a user can ask questions about relative block locations (e.g., "Is the A block to the right of the B block?") and request clarifications on why particular relations hold. Models for spatial prepositions used by the spatial specialist are probabilistic predicates computed hierarchically, in a tree-like fashion, as a combination of more primitive relations. These primitive relations in the tree hierarchy can be retrieved to provide an explanation for system's outputs. For example, assume that when asked about the location of the block A, the system generates a response of the form "the block A is next to the block B". If queried as to why the system arrived at that particular judgment, the spatial specialist retrieves the underlying component relations from which "next to" is composed (proximity and similar elevation) and returns the relevant relations to the dialogue manager that generates a human-readable response.

2 Related Work

Recent years have seen a push towards explainable AI (Otte, 2013; Samek and Müller, 2019). While classical symbolic AI systems are typically both explainable and interpretable by design, with regard to explainability in a pure neural network setting, many recent efforts have been concentrated around modular neural network architectures (Andreas et al., 2016; Hu et al., 2018; Gupta et al., 2019) and architectures that directly generate the explanations for their own operation (Andreas et al., 2017). The former is concerned, in general, with building a network out of specified blocks (modules) trained to perform particular operation (e.g., finding, filtering, counting, etc.) on the input or process a certain aspect of the task (e.g., recognizing a category vs. recognizing a property such as color, etc.) Explanations of the model's outputs

Proceedings of the Conference on Reasoning and Interaction (ReInAct), pages 53–57

then are derived from clear-cut understanding of the purpose of each module and their interconnections. The latter uses various additional blocks to generate explanations, e.g., in the form of plain English text, based directly on the model's inner state.

Our approach to spatial preposition modeling is inspired by the criteria that have been discussed in linguistically oriented studies (Garrod et al., 1999; Herskovits, 1985; Tyler and Evans, 2003). Studies of human judgements of spatial relations show that overly formal qualitative models with sharp boundaries generally cannot do justice to the usage of locative expressions in natural settings. Our models are implemented along the same general lines as those in (Platonov and Schubert, 2018) and (Richard-Bollans et al., 2020b,a). These studies model prepositions as constructed from more basic physico-geometrical primitives. Modern neural work on spatial relation-learning in the BW domain is exemplified by (Bisk et al., 2018).

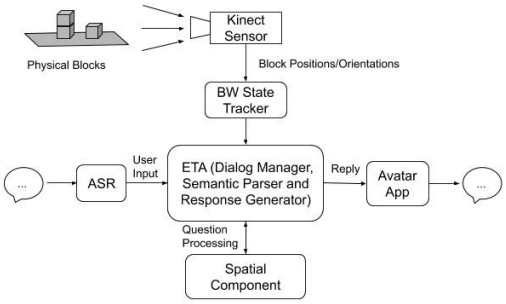
3 Blocks World System and Eta Dialogue Manager

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, and all the spatial predicates are computed based on this 3D scene.

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 (Schubert and Hwang, 2000) 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*

(a) Blocks world setup



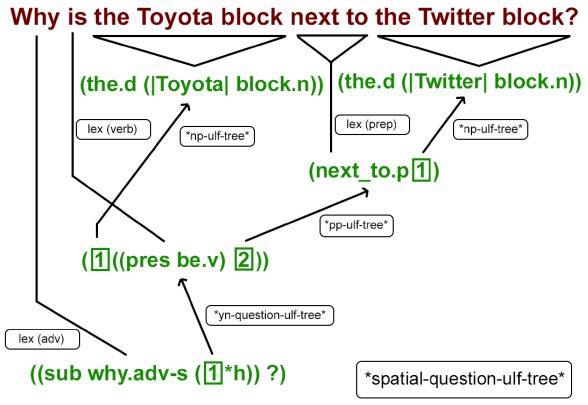
(b) Dialogue pipeline

Figure 1: System overview.

transduction, similarly to the mechanism used by the LISSA system (Razavi et al., 2017) 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) (Kim and Schubert, 2019) (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 lies in the fact

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

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 tense, inversion, and other linguistic phenomena. Eta also has a limited coreference module utilizing syntactic constraints, recency, and other heuristics.




```
((sub why.adv-s (((the.d (|Toyota| block.n))
    ((pres be.v)
      (next_to.p (the.d (|Twitter| block.n)))))
 *h)) ?)
```

Figure 2: An example ULF parse, with the input shown in red, a ULF parse tree shown in green, and the final query ULF shown in blue. The subclauses of a ULF formula are composed by hierarchical pattern-transduction trees (each consisting of a number of patterns to match to a section of the input, together with a composition node), with the help of a lexical subroutine to handle leaf nodes. Edges in the parse tree are labelled with the pattern-transduction tree that handles the corresponding subclause (typically responsible for a particular syntactic category).

Once the final ULF formula is obtained, the dialogue manager queries the spatial specialist module with the formula, and receives a list of relevant preposition factors (in ULF form). The dialogue manager uses a natural language generation (NLG) module to substitute these prepositions into the query ULF before using general linguistic rules to convert this to an answer ULF, which is then mapped to English to produce a verbal explanation.

4 The Spatial Specialist

The spatial specialist contains a family of models for spatial prepositions. Each such model is implemented as a probabilistic predicate, computed hierarchically as a combination of more primitive relations that we call *factors*. These factors typically encode more basic relations that affect whether a particular spatial preposition holds. They are usually either different senses of the same preposition or they co-occur with the preposition in most/all configurations that license the usage of that preposition. The set of factors covers various geometric and structural properties, including distances between objects, direction from one object to another, support relations, physical contact, part structure, etc. The factors are combined according to one of several rules, such as multiplying two or more factors, finding the maximum, or taking a linear combination, in order to produce the final value for the preposition.

Some range of sense ambiguity is taken into account by considering different coordinate frames. In particular, for projective relations, such as *to the right/left of*, one can consider deictic, extrinsic and intrinsic senses. The deictic sense is computed based on the viewer's coordinate frame. Here, one object is considered to be in the given relation to another, if its projection onto the viewer's visual plane is in that relation (e.g., to the right of) to the projection of the latter object. The extrinsic sense is based on the global coordinate system imposed by the world, i.e., front-right axes of the table. The intrinsic sense is determined based on the intrinsic coordinate system of the ground object, i.e., A is intrinsically to the right of B if it is on the right side of B. Note that this sense is absent in the blocks world setting since blocks do not have intrinsic orientations, but it is added to support generality of the spatial models. For objects that do not have intrinsic orientations, the factor for the intrinsic sense of the relation is set to 0. When dealing with multiple senses, the model selects the one with the maximal value as an output.

5 Factor Extraction for Justification of Model's Judgments

The tree-of-factors implementation of spatial models allows backwards-generated justifications spatial judgments. Since factors represent higher-level semantic concepts, they can readily be translated into natural language. The tree of factors computed during the forward computation phase is preserved and is traversed in the backward direction starting from the root that represents the value of the

preposition. The mechanism for factor retrieval is as follows. If the combination rule for the current node is a product, then if the node value is greater than 0.5, return all the child nodes; otherwise, return the child node with the smallest value. If the combination rule for the children is a weighted linear combination of factor values, then if the current node value is greater than 0.5, return the highest contributing factor node or nodes (total contribution includes their value and weight); otherwise, return the value of the node with the largest weight. Finally, if the combination rule is the *max* operation, then if the current node value is greater than 0.5, return the child node with maximum value; otherwise return all the child nodes.

As an example of the operation of the explanation procedure, consider the simplified factor network for *to the right of* in Fig. 3.

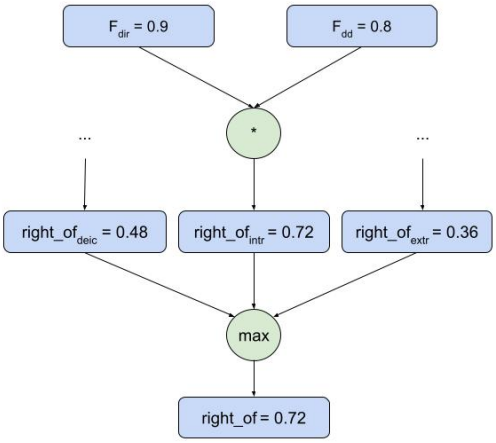


Figure 3: An example of an explanation procedure.

The numbers in the nodes are the respective values of the factors that the node computes. Assume that the system is being asked whether A is to the right of B. Assume further that the final output value is *right of = 0.72*, which corresponds to "yes". Now, if the user inquires why the system arrived at that conclusion, the following process unfolds. The node for the final score for *to the right of* takes the maximum over three values: deictic *right_of*$_{deic}$, intrinsic *right_of*$_{intr}$ and extrinsic *right_of*$_{extr}$. Since the maximum is taken, one of those nodes must be equal to the final value. Hence, the explanatory routine returns the corresponding node and its value (*right_of*$_{intr}$, 0.72). The corresponding interpretation will be (after the dialogue manager generates a response) "A is to the right of B because A is located on the right side of B, according to B's orientation". If asked further as to why the intrinsic relation holds, the system will analyze the intrinsic score's contributing factors, namely F_{dir} (directional factor that defines the "right-side" region for an object) and F_{dd} (distance decay, measuring how far apart the objects are). Since the combination rule used is multiplication and the value of the current node (intrinsic right) is 0.72 (i.e., relation holds), it follows that both factors must hold as well. The system will return the list of the nodes and their values, i.e., [(F_{dir}, 0.9), (F_{dd}, 0.8)] as a result. The straightforward interpretation of the latter would be "A is on the right side of B, because it is located in the general rightward direction w.r.t. to B and it is close enough". This process can continue until leaf nodes are reached, which do not admit further decomposition and are treated as primitives. Alternatively, let $F_{dd} = 0.4$ (A is too far from B). This low value will propagate downstream and affect the *right_of*$_{intr}$ and the final *right of* scores. The system then will supply a negative answer to the original question. If queried, it will return the list of all senses [(right_of$_{deic}$, 0.48), ...] which has a straightforward interpretation of "A is not to the right of B because none of the senses apply". If queried why, say, the intrinsic sense does not apply, the system returns the lowest-value node contributing to the intrinsic sense node, i.e., [(F_{dd}, 0.4)], which translates into "A is too far from B to be on its right side".

6 Conclusion

We described our work in progress concerning a dialogue system incorporating a spatial specialist with spatial semantic models that are based on clear and intuitively-grounded criteria, capable of generating justifications of spatial judgements produced by the system. The spatial subsystem incorporates hierarchical representations of spatial prepositions, constructed using so-called factors - intermediate simpler relations correlating with the occurrences of the prepositions. The explanation system scans the tree of these factors and retrieves the most relevant ones for the given situation. The configuration is inherently interpretable due to factors corresponding to intuitive criteria that seem to underlie the natural usage of prepositions.

Acknowledgments

This work was supported by DARPA grant W911NF-15-1-0542.

References

Jacob Andreas, Anca Dragan, and Dan Klein. 2017. Translating neuralese. *arXiv preprint arXiv:1704.06960*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

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*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2019. Neural module networks for reasoning over text. *arXiv preprint arXiv:1912.04971*.

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

Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 53–69.

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

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Clemens Otte. 2013. Safe and interpretable machine learning: a methodological review. *Computational intelligence in intelligent data analysis*, pages 111–122.

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, 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.

Adam Richard-Bollans, Lucía Gómez Álvarez, and Anthony G Cohn. 2020a. Modelling the polysemy of spatial prepositions in referring expressions. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 703–712.

Adam Richard-Bollans, Brandon Bennett, and A Cohn. 2020b. Automatic generation of typicality measures for spatial language in grounded settings. In *European Conference on Artificial Intelligence*. Leeds.

Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22. Springer.

Lenhart Schubert and Chung Hwang. 2000. 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*.

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