Answering questions about the relative positions of objects in the 3D scene

CSC447 Project Report

Georgiy Platonov

December 2016

Introduction

In this project the author tackles the problem of answering a queries about relative objects positions inside the 3D scene, posed in the natural language. Spatial reasoning is a core human skill that is heavily used in our everyday activities. It, thus, has a critical importance to the field of human-computer interaction, where it is a key capability to understanding descriptions and instructions. The idea and motivation behind this work is to rely on computational models of spatial prepositions, used in English, in order to check the implicit constraints, contained in the question, to locate the required objects, present in the scene. We are going to work in the toy blocks world domain, consisting of a plane and a small set of cubical blocks of the same size.

Related work

There are several works that inspired the current research. First of all, the computational models for prepositions draw the ideas from the works by Talmy[11] and, especially, Herskovits[8] and, also, their evolution in Tyler and Evan's book [12]. These works describe certain categories of properties of the figure and the ground that affect the interpretation of a particular preposition. For example, when we want to express the fact that two objects is close to each other, we rely on their size, salience and mobility to put them in the right order ("a dog is near the house", not "A house is near to the dog"). In the same way, different geometrical, physical and functional (in Herskovits' terminology) properties affect the applicability of other locative expressions. In terms of scene reconstruction based on the natural language textual representation, the closest are the works by Coyne and Sproat[6, 5] and Bigelow et. al[1]. Coyne and Sproat's Wordseye is one of the best text-to-scene systems up to date, while Bigelow and his collaborators studied application of modeling the 3D scenes in Blender[2] for story understanding. Another work, very close in spirit to the current one is [9]. Regarding the tools to be used, the current project will rely on the Blender

for the purposes of scene modeling. The TRIPS parser is going to be used for parsing the questions [7]. It should be noted that the approach in this project is different from the approaches taken typically in the field. Most of the work on spatial reasoning is concentrated either on developing some form of abstract formal framework, e.g., topological or mereotopological[4, 3], or on using the connectionist ideas, e.g., [10]. The first approach, while allowing for rigorous derivations, is not natural and does not fully reflect the usage of spatial prepositions in the natural language. The second one, on the other hand, lack clarity and relies mainly on the superficial features, without taking into account the nature of the objects under consideration.

High-level overview

The developed system takes a set of questions about relative locations of objects and for each question tries to evaluate the spatial constraints contained in that question in the form of prepositions and output the answer to that question. A typical query about the relative positions of two or more objects can be one of two kinds that we will call confirmation queries and identification queries. The confirmation queries fully specify the set of objects and spatial relations between them and require determine whether or not given set of relations holds with respect to the given objects (a "Yes/No" question). An example of such a question would be "Is the red cube near the black one?". An identification query consists of a set of objects and constraints and requires to determine the set of objects, satisfying that set of constraints with respect to given objects (a "Wh-" question), e.g., "What cube is above the red cube?". Our system is capable of answering both types of questions.

It relies on two crucial components: the Blender 3D modeling software and the TRIPS parser. The general pipeline is as follows. The system reads the question from input file and then proceeds to parse it using the TRIPS parser. The TRIPS logical form, representing the sentence is extracted from the output of the parser. That form is used as a main tool in analyzing the query since it represents the semantic roles of different tokens and connections between them in a convenient, tractable and detailed manner. From that logical form the formal query structure is built. The formal query consists of the query type and a list of relations, encoded in the question. Each relation represented with three data elements: its type (e.g., *above, on*), its figure and its ground. Then, depending on the type of query, the computational models for the spatial relations are applied to the 3D scene, loaded into Blender. If dealing with the confirmation query, the objects mentioned in the query are located in the scene and the corresponding relation is evaluated. Depending on whether the relation holds or not, the system outputs "YES" or "NO". In the other case, all objects in the scene are checked if they satisfy the set of given constraints or not. Those that satisfy form the output.

It should be noted that the final system was significantly scaled down in its capabilities compared to the original proposal. The first serious limitation is that the model world was switched from the "room world", i.e., the one

containing the everyday objects, like furniture, etc., to the blocks world. The second significant drawback is the giving up on the scene generation from the textual description. The main reason for that are problems stemming from the sampling the positions of the objects, satisfying the particular relation. Say, if we want to generate a scene with the green block being above the black one, we sample the positions of the two blocks, i.e., we place these two blocks in the scene and check the value of the the *above* predicate with these two blocks as arguments. If the value of the predicate is sufficiently high, the two blocks are said to satisfy this constraint, and the system moves on and tries to satisfy the next one. Otherwise, it withdraws the blocks and tries to sample the new positions for them. For the configuration to be considered acceptable, it does not have to correspond to the perfect schema for that relation, just be good enough (e.g., be sufficiently "above"). This leads to an offset, accumulating with each additional constraint, thus rendering the final result quite messy and far from the intended scene.

Evaluation and testing

The system was tested using three spatial relations, *on, above(below)* and *near*. Below a sample blocks world scene is depicted. The scene can be easily modified in Blender by moving or adding/deleting blocks. The scene contains ten colored blocks (black, blue, brown, gray, green, orange, pink, purple, red and yellow one). In what follows we consider the capabilities and limitations of the system by analyzing a set of example queries.

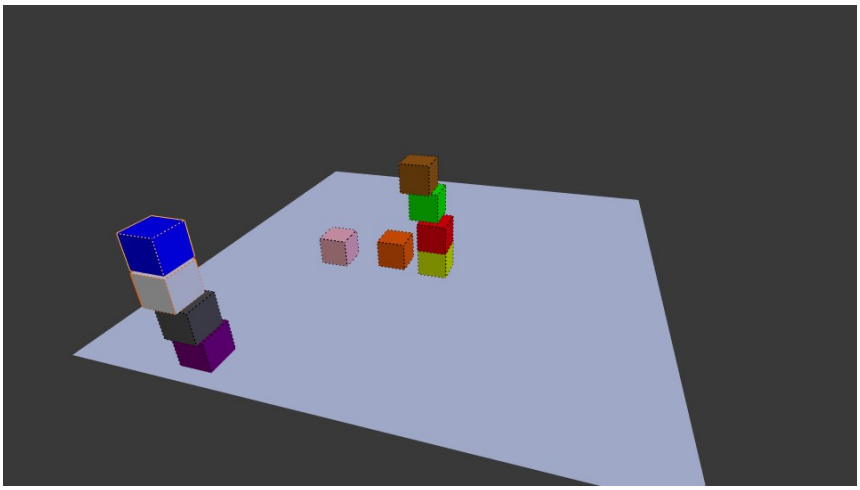

Figure 1: The 3D scene used for testing

Consider the following the of test queries:

1. Is the blue cube on the green one?

2. Is the brown block above the orange block?
3. Is the brown block above the pink?
4. What is above the yellow block?
5. Is it true that the gray cube is on the black cube?
6. What is the cube that is near the gray one and above the purple one?

The sample output for some of the tests can be seen below.

```
======================================================================

QUERY: Is the blue cube on the green one?

PROCESSING...
RESULT:
NO
======================================================================

QUERY: Is the red block on the yellow block?

PROCESSING...
RESULT:
YES
======================================================================

QUERY: Is the brown block above the orange block?

PROCESSING...
RESULT:
YES
======================================================================

QUERY: Is the brown block above the pink?

PROCESSING...
RESULT:
NO
======================================================================

QUERY: Is the brown block above orange block?

PROCESSING...
ERROR: Figure or ground is not fully specified or has not been parsed properly.
======================================================================

QUERY: What is above the yellow block?

PROCESSING...
RESULT:
THE RED CUBE
THE BROWN CUBE
THE GREEN CUBE
======================================================================
```

Figure 2: Sample output

The first test query is an example of the simplest question that can be asked about the blocks world. As can be seen from the scene, it is quite far from the green one to be considered as being *on* it, and the system outputs "NO" for this test case. Note that, since the spatial relations are not well determined, which makes it hard to judge whether the particular relation actually holds between two objects and different responders might disagree. This is illustrated by the second and third examples. From the picture it is clear that the brown cube is indeed above (one might say "directly above") the orange one. In the third example, the brown cube is elevated higher than the pink one. However, the system replies "NO", as the brown block is too far and the *above* predicate returns returns value too small for acceptance. Test case four demonstrates

the result of asking an identification query. Given this query, the system proceeds to determine what blocks satisfy the given constraints. As is visible from the screen-shot, three blocks unambiguously satisfy the query and, thus, are present in the output. Test case five shows a similar query to test cases 1-3, but using a slightly different wording. The result (not shown in the screen shot) is "YES". Finally, the last test case shows a somewhat more complex type of query, containing two constraints. Both, the blue and the black blocks satisfy these constraints and are present in the result list.

Unfortunately, the current version of the system is still very limited and fragile. First of all, it relies heavily on the particular patterns in the TRIPS logical forms, viz. is expects a particular format of each spatial preposition (its figure and ground) to be extracted properly from the sentence. Consider two example sentences, "Is the brown block above orange block?" and "Is the brown block above the orange block?". The only difference between them is the presence of the article before "orange", but this leads to completely disparate logical forms, so that the system fails to analyze the first sentence. Another problem is asking the queries involving several objects, like "What is above the black and the blue cube?". The current version is unable to process multiple grounds or figures properly.

Running the system

This section here is in lieu of the readme file and contains instructions on running the code and formatting the input. The project consists of the following files:

- init.py - executes main.py as a script in Blender
- main.py - main code of the project
- blocks_world.blend - Blender scene containing the blocks world
- tests - contains the test cases
- frames.xml - contains frames, used in the project

Note that all the files must be in the same directory for the code to run. The test cases are contained in the file named "tests", in a simple text format, separated by a new line character. You can easily add or remove the test cases. In order to run the code, Blender must be installed on your OS. If blender is installed, simply execute the init.py in the terminal to run the system: "*python init.py*". The output will be displayed in the terminal window.

Conclusion

This project was planned as a long term one, with the idea to get the initial results and then expand the system into something more powerful, e.g. an inference engine, capable of answering questions about short stories, that require spatial thinking. As it stands, the system is too rigid. However, the TRIPS logical forms that were used in the project were proven to be content-rich enough to be a good tool for such a task. The main drawback, at this point is the lack of flexibility in determining the relation's arguments (i.e., figure and ground of the

spatial preposition). The initial plan was to use the frames to represent different structural patterns of the prepositions. This idea, however, was abandoned in favor of a direct figure and ground extraction from the logical form links. It was not clear from the beginning, but this approach does not always work properly, as the structure of the arguments might differ significantly. In the future, though, the frames are planned to be included back.

References

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

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam,

[3] Anthony G Cohn. Qualitative spatial representation and reasoning techniques. In *KI-97: Advances in Artificial Intelligence*, pages 1–30. Springer, 1997.

[4] Anthony G Cohn and Jochen Renz. Qualitative spatial representation and reasoning. *Handbook of knowledge representation*, 3:551–596, 2008.

[5] Bob Coyne, Alex Klapheke, Masoud Rouhizadeh, Richard Sproat, and Daniel Bauer. Annotation tools and knowledge representation for a text-to-scene system. In *COLING*, pages 679–694, 2012.

[6] Bob Coyne and Richard Sproat. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 487–496. ACM, 2001.

[7] Myroslava Dzikovska, Mary Swift, James Allen, and William de Beaumont. Generic parsing for multi-domain semantic interpretation. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 196–197. Association for Computational Linguistics, 2005.

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

[9] Rabiah Abdul Kadir, Abdul Rahman Mad Hashim, Rahmita Wirza, and Aida Mustapha. 3d visualization of simple natural language statement using semantic description. In *International Visual Informatics Conference*, pages 36–44. Springer, 2011.

[10] Baolin Peng, Zhengdong Lu, Hang Li, and Kam-Fai Wong. Towards neural network-based reasoning. *arXiv preprint arXiv:1508.05508*, 2015.

[11] Leonard Talmy. How language structures space. In *Spatial orientation*, pages 225–282. Springer, 1983.

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