IJCSI International Journal of Computer Science Issues, Volume 14, Issue 5, September 2017
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.20943/01201705.3741

Text-to-3D Scene Generation using Semantic Parsing and Spatial Knowledge with Rule Based System

Md. Shahadat Hossain, Abdus Salam

Department of Computer Science, American International University-Bangladesh
Dhaka, Bangladesh

Abstract

Scene Generation plays an important role in digital media to represent a news or a specific domain to the viewers. It's not easy to produce a scene from a text. Text may not completely express the whole situation in digital media. Most of the people are not conscious about the news until it's not visualized to them. Text to 3D scene generation is a process where people do not need to read a news. The 3D Scene will represent the situation. It will help people to conscious about their life. In this paper, we introduce a rule-based framework where scene generated from text using semantic parsing and spatial knowledge. Semantic parsing has identified the templates, objects, and constraints and spatial knowledge have built the relation between object and template. Our rule based framework has identified the uncountable noun and some spatial relations to generate 3D scenes.

Keywords: *Image Processing, Natural Language Processing, Spatial Knowledge, 3D Scene, Semantic Parsing, Rule Based Parsing.*

1. Introduction

Text to 3D scene generation is one of the popular research fields in natural language processing. We do not need to understand language of news, if the text to 3d scene generation is completed successfully. Many creative industries use 3D scene for that. Newspaper industry will use this for reducing reading time consumption. It is impossible to a person to know the every language and gets updated with news or research in current world. On the other hand, every person can interact with the 3D scene. Consequently, to build a realistic system that can illustrate with the world and interact with people, we require such knowledge to communicate with language in context.

The picture suggests a convenient way for the photographer to express their artistic point as well as knowledge. Spatial knowledge is an essential prospect of the world and is implicitly expressed in natural language [1]. It was one of the most expanding challenges in enabling natural communication and grounding language between intelligent systems and people. For example, we consider a system that will execute order as follow "bring a bottle of cold water on the center of the table", it's necessity to do the work with an understanding of likely places for the cold water in the fridge and the water should be placed in the center of the table.

Fig. 1: Generated scene for "bring a bottle of cold water on the center of the table". Note that the system infers the presence of a table and the table should be supported by the bottle.

The Words Eye system [2], Learning Spatial Knowledge [1] and Rich Lexical Grounding [3] had addressed the text to 3D scene generation. Semantic parsing work has a deal with grounding text to physical characteristic and relations [9, 10]. Generating text is referred to objects [11] with conjunctive language to spatial relationships. Semantic parsing process can also impose too many perspectives of text to scene generation.

However, there are many inexistent issues in this area. There is an opportunity to implement learning spatial knowledge and rich lexical grounding with Parts-of-Speech tagging. Unstated facts is an incriminate problem represented by prior work. This problem has been discussed in the limitation part of Learning Spatial Knowledge [1] and Rich Lexical Grounding [3], which has not solved by the community.

We focus on the text-to-3D task to integrate Semantic Parsing with rules based scene generation approach which is covered with the challenging scenario. It's mentioned implicit pragmatics based on a location of an object in the interference. The scene demonstrates the object in the actual location which has not shown in prior work. The user would not need to be wordy with text such as "There is a table which folds with many pieces of paper.'' instead of this we can use "The table covers with the paper". So, we introduce a framework which can represent 3d scene efficiently.

The rest of the paper is organized as follows. The following section discusses about related work. The section 3 discusses about essential information as background study. The section 4 contains methodology with two sub-sections. The section 5 shows the framework description which is our proposed framework. The section 6 discusses about dataset, model and result which are in three subsections. The Section 7 discusses about the limitation of this work. Finally, we conclude our work at section 8.

2. Related Work

The prior work is much resourceful and inspirational for text to 3D scene generation. The recent work has improved a lot using natural language processing technique. There are some issues which is different between 2D and 3D images.

It has been implemented as an application of semantic parsing. Angel X. Chang, Manolis Savva and Christopher D. Mannin [5] have observed that people describe in the text typically relevant and important information. So, the text has divided into 3 parts (Template, Object, and Constraint) before generating 3D scene. They have also introduced some functions to manipulate the scene using Template, Object, and Constraint such as: Select(X), Remove(X), Insert(X), Replace(X, Y), Move(X, ΔX), Scale(X, ΔX), and Orient(X, ΔX).

Grounding text is the most important part in the text to 3D scene generation. It showed that spatial knowledge gives the best output for grounding. The process started with extracting knowledge from text. It's continued with template parsing and extracting constraint. They have arranged the object and predicted most likely type of scene using rules [1]. They extracted dependency pattern from text using Semgrex patterns [6].

It was rules based system which is grounding text with rich lexical grounding. It has split text into multiple parts, extracted object as well as a noun phrase and generated 3D scene [3]. Rule-based parsing component has been described in Chang et al. (2014).

3. Background Study

It's an essential part to clear about different technical terms. It may help to understand the flow of work such us spatial knowledge, grounding, part-of-speech, rules-based-tagger, semantic parsing, etc.

Spatial knowledge is a process where text can be parsed with the binary process (left, right). We use it to build the relation between objects, environment, and constraints. Paradigm, kitchens typically contain kitchen counters on which plates and cups are likely to be found. The type of scene and category of objects condition the spatial relationships that can exist in a scene [1].

Grounding is given the constraints and priors on the spatial relations of the object, transform the scene template into a geometric 3D scene with a set of objects to be instantiated [1].

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc. although generally computational applications use more fine-grained POS tags like 'noun-plural' [4].

Rules-based-tagger, the Brill tagger uses a rule-based approach [Brill 1994] where a set of rules for determining word tags is created as follows (during training): an initial set of naive tags are assigned to the corpus of words, after which transition rules are learned by correcting the falsely identified word-tags. During the tagging process, these rules are applied in order to identify the correct word tag [7].

Semantic parsing is the method of mapping a natural language sentence into a formal illustration of its meaning. A shallow form of semantic representation is a case-role analysis (a.k.a. a semantic role labeling), which identifies roles such as agent, patient, source, and destination (CS, Machine Learning, university of Texas at Austin).

4. Task Description

In text to 3D scene generation, the task is to take a text as an input and generates a 3d scene as an output which is described with inputted text. Moreover, based on the

input text, we extract objects from a dataset of 3d models and put them to generate output scenes.

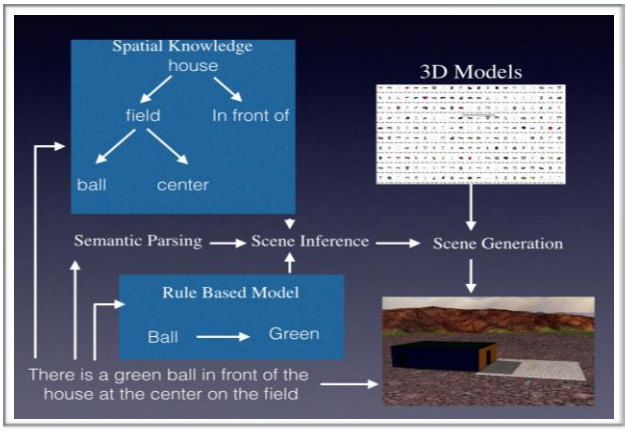

Fig. 2: Overview of our rules based semantic parsing for text-to-3D scene generation. We parse the input text using rules based parsing and semantic parsing which is corresponding to the scene inference. The scene inference uses 3d models of Stanford CoreNLP to generate the scene.

In this paper, we concentrate on the sub problem of the textual term to 3D model references (i.e., selecting the proper object to place on proper location). We introduce a new rule with semantic parsing which is extracting object from inputted text, select it to form data sets and finally, it represents a 3D scene (Fig. 2).

A straight approach to scene generation might conduct object with uncountable nouns from text to retrieve 3D models. However, such a way is likely to demonstrate well. But it ignores spatial relations and attributes. A strong approach is needed to describe the environment to generate the scene accurately with objects. Rules based semantic parsing solves many challenging parts of this task when 3D representation will do accurately.

4.1 Relative Position Priors

We observed some view location with respect to object type and scene location category to propose some relative positions of objects: i.e., the relative position of an object type Tobj is with respect to another object type Tref and the scene location Category Tslc.

$$RPrelpos(x, y \mid Tslc, Tobj, Tref)$$

The position x, y is the centroid of the Scene location. Tobj place on Tslc with respect to the position of Tref. RPrelpos is the relative position of sibling and child parent objects.

4.2 Predefined Spatial Relations

Authors in [1] have seen a set of predefined relations for spatial relations such as left_of, right_of, above, below, front, back, supported_by, supports, next_to, near, inside, outside, faces, left_side, right_side, etc. They have measured those from the viewer's perspective using axis-aligned bounding boxes; bounding boxes making the difference to predict volume overlap or closest distance. We will add some new relations which will give a better output such as center, corner, cover, right_corner, left_corner, right_upper_corner, left_bottom_corner, etc. These spatial relations may place the object on its accurate position with respect to the view of the scene.

5. Framework Description

Rules based parsing is a model to describe the text to the scene. This framework has improved accuracy which has shown in the result. Spatial knowledge and lexical grounding have strong impact on scene generation. Semantic parsing plays an important role to select an object, template and constraint. We have introduced a framework where knowledge has been extracted from the text to cross-match with rules.

The rule-based semantic parsing approach is a 5-step process:

1. Take the input text from user.
2. Rule-based parsing:
   a. Split the text into small sentences.
   b. Extract the objects from small sentences and sort the objects as countable noun phrase and uncountable noun phrase.
   c. Extract adjective of noun and check dependency pattern.
   d. Again extract template from the main text.
   e. Predict additional objects using Association Rules.
   f. Check dependency pattern of a template and additional objects.
   g. Correlate countable objects, uncountable objects, template and additional objects.
3. Parse the text using Semantic Parsing & Spatial Relation. Discover objects, template and additional objects.
4. Compare rules based parsing and semantic parsing and list common objects.
5. Correlate and arrange common objects, template and additional objects.

These properties are later used to query the 3D model database. We use the same model database as Chang et

al. (2014) and also extract spatial relations between objects using the same set of dependency patterns.

6. Discussion

6.1 Dataset

We have used a dataset which has 1128 scenes and 4284 free-form natural language description of these scenes which have been introduced by Chang, Monroe, Savva, Potts, Manning [1]. They used simple online scene design interface that allows users to assemble scenes using available 3D models of common household objects to create this training set. They used a set of 60 seed sentences describing simple configurations of interior scenes as prompts and asked workers on the Amazon Mechanical Turk crowdsourcing platform to create scenes corresponding to these seed descriptions. They asked other workers to describe each scene to obtain more varied description for each scene [1].

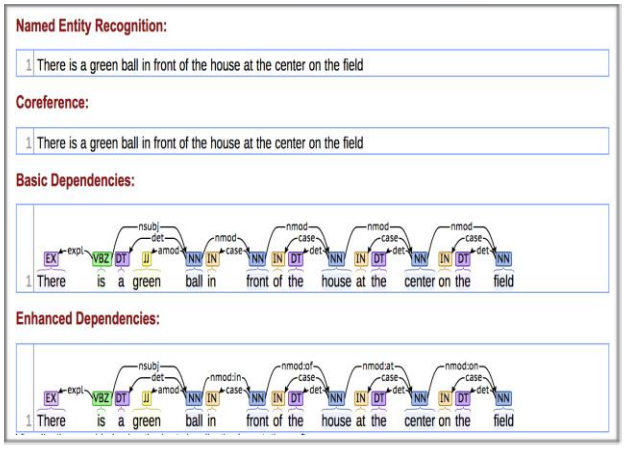

Fig. 3: Basic and Enhanced dependencies for "There is a green ball in front of the house at the center on the field."[12]

6.2 Model

We train a classifier to learn semantic, spatial to create a model for generating scene templates from the text which is used for lexical grounding in paper [1]. The Available dataset is at http://nlp.stanford.edu/data/text2scene.shtml. We combine our learned semantic & spatial with a rule-based scene generation model. To select a better model, the learned semantic & spatial allow us to offer the rule-based model handling scene location and relationships between objects.

6.3 Result

There was a significant effect of scene generation using our framework. We have used Stanford CoreNLP dataset to check our framework. We have shown an expected paradigm of generated scene using our framework. We have also practiced interesting aspects using semantic parsing and spatial knowledge with the rule-based framework.

We will test our framework using 3D model dataset culled from Google 3D warehousing by prior work. It has been done by scene synthesis and containing about 12490 mostly indoor objects [8].

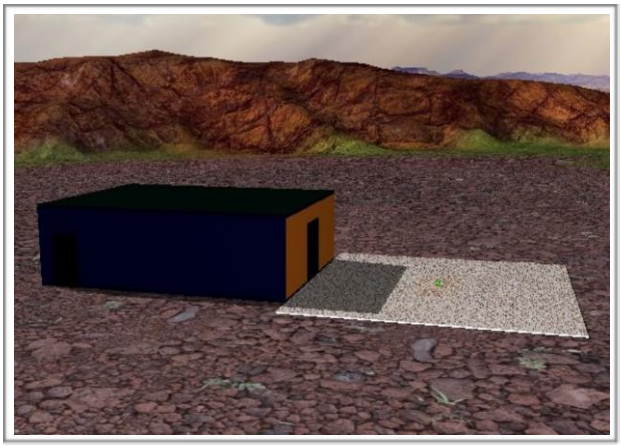

Fig. 4: Scene generated for "There is a green ball in front of the house at the center on the field."

7. Limitation

There are still many objections to rise in the text to scene generation while the framework shows commitment. We did address the difficulties of resolving objects, placing location, uncountable noun. But we didn't test its wide range. We introduce different techniques in the single framework and comparing the result of each technique. But we did not introduce in which method the result of each technique will be compared. On the other hand, we used spatial knowledge to extract location; semantic parsing to extract template, object and constraints; Rule-based model to deal with an uncountable noun of objects. So, a failure case may arise using this framework.

8. Conclusion

We have observed that many error cases appear in the framework which are not generated by our system. Although we have given a solution for lacks of prior work. It solves some lacking prior work such as Position

of an object, uncountable noun of an object. An obvious enhancement would be to expand based on our framework. It would help with semantic parsing and spatial language that are not handled by previous work.

Prior work has relied on 3D scene generation using the rule based method to map objects to build 3D objects where we also introduce a rule-based system. It will surely work for mapping a 3D Scene. We have introduced some spatial function for placing the object according to its description. We have used the Stanford CoreNLP dataset to annotate with natural language description which we believe gives the great result to the real world community. We have presented an approach that will learn data from ground textual descriptions to objects using the corpus. So, there is a scope to build and test the system with a long text what we have not done here.

The scene has been judged by a human to evaluate the grounding approach impacts of generated scene. In addition, we present a comparison of generated objects, template and constraints among spatial knowledge, semantic parsing and rule-based framework which has shown a strong correlation with objects, template and constraint.

We have invented that rule-based grounding can be learned directly from the corpus of 3D scenes and natural language descriptions and that our model can successfully be grounded correlation with objects, template and constraint and enhanced scene generation over baselines adapted from prior work.

References

[1] Angel X. Chang, Manolis Savva, and Christopher D. Manning- Learning spatial knowledge for text to 3D scene generation, In Proceedings of Empirical Methods in Natural Language Processing (EMNLP) (2014).
[2] 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 (2001).
[3] Angel Chang∗ , Will Monroe∗ , Manolis Savva, Christopher Potts and Christopher D. Manning- Text to 3D Scene Generation with Rich Lexical Grounding, In Proceedings of Empirical Methods in Natural Language Processing (EMNLP) (2015).
[4] Kristina Toutanova and Christopher D. Manning- Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger, In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70 (2000).
[5] Angel X. Chang, Manolis Savva and Christopher D. Manning- Semantic Parsing for Text to 3D Scene Generation, In Proceedings of Empirical Methods in Natural Language Processing (EMNLP) (2014).
[6] Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie- Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher D. Manning- Learning alignments and leveraging natural logic. In Proceedings of the ACL-PASCAL Workshop on Textual Entail- ment and Paraphrasing (2007).
[7] Kevin Glass and Shaun Bangay ,"Evaluating parts-ofspeech taggers for use in a text-to-scene conversion system," In Proceedings of SAICSIT '05, Judith Bishop and Derrick Kourie, Eds., pp. 20–28 (2005).
[8] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan- Example-based synthesis of 3D object arrangements, ACM Transactions on Graphics (TOG) (2012).
[9] Jayant Krishnamurthy and Thomas Kollar- Jointly learning to parse and perceive: Connecting natural language to the physical world, Transactions of the Association for Computational Linguistics (2013).
[10]Cynthia Matuszek, Nicholas Fitzgerald, Luke Zettle moyer, Liefeng Bo, and Dieter Fox- A joint model of language and perception for grounded attribute learning, In International Conference on Machine Learning (2012).
[11]Nicholas FitzGerald, Yoav Artzi, and Luke Zettle moyer- Learning distributions over logical forms for referring expression generation, In Proceedings of the Conference on EMNLP (2013).
[12]Stanford Dependencies- https://nlp.stanford.edu/software/stanford-dependencies.html

Md. Shahadat Hossain has completed graduation from AIUB in 2015. He has three years' experience in the field of web development. Currently he is doing masters on intelligent system at AIUB. His objective is to work with a reputed research organization that will increase his knowledge and explore his intelligence and assist him to achieve personal as well as organization goals. He is devoted to Research in the field of Image Processing, Computer Vision & Pattern Recognition, Natural Language processing which are the sub branch of Artificial Intelligence and Algorithms.

Abdus Salam is working as an Assistant Professor at American International University- Bangladesh (AIUB). His research interest includes Data mining, Machine Learning, Semantic Web, Intelligent Systems, and Human Computer Interaction etc. He has received his B.Sc. in Computer Science and Engineering from AIUB and M.Sc. in Computer Science major in Software Technologies from University of Trento, Italy.