Understanding Spatial Knowledge in Natural Language

Lu Cao and Yoshinori Kuno

***Abstract*— Language references knowledge. There is a large body of knowledge conveyed through language such as object categories and colors. Still, much knowledge about the world is obvious for us that would not be expressed explicitly in language, for instance, spatial knowledge. To achieve smooth and pragmatically-charged communication with people, in this paper, we propose an approach for spatial knowledge understanding in natural language.**

I. INTRODUCTION

What makes understanding spatial knowledge challenging in natural language is that the knowledge which is common sense or can be easily inferred to us is neither stated in language nor manually annotated in a laborious effort. For instance, to correctly execute the command "Tell me where my computer is." the robot needs to have a *prior knowledge* about the likely location for the computer on the desk. If we say "Where is the keyboard?" then the robot should be able to *infer the knowledge* – the keyboard is on the desk and in front of the computer.

In this paper, we introduce a simple yet effective approach to understand spatial knowledge in natural language (Figure 1), which is a part of large-scale knowledge processing framework aiming to help robots find objects through non-verbal and verbal knowledge. As a starting point, we learned knowledge priors from state-of-the-art datasets [1][2][3], which is the first contribution of this work (Sect. II). The learned knowledge is converted to machine-readable tuples as axioms interlinked by properties and imported to spatial knowledge base (KB). We would discuss some new improvements of the spatial KB in this paper (Sect. III). Comparing with the one we reported in [4], the new version is able to extract more knowledge from language and resolve more rarely mentioned implicit facts, which would be beneficial in challenging scenarios. This is the second contribution. The main drawback of the natural language parser in [5] is the limitation of semantic forms, which is still far from achieving natural communication with users. The third contribution is that we propose a natural language interface (Sect. IV), enabling to correctly interpret complex syntax structures.

II. LEARNING KNOWLEDGE PRIORS

Inspired by recent work on spatial knowledge learning [1], we collected common sense knowledge about *object categories*, *innate parts* (e.g. a computer has *functional* part – screen), *occurrences* (e.g. a cup is likely to be found in the kitchen, not bedroom) of objects, *support hierarchies* (e.g. a laptop is usually placed on a desk rather than a chair) and *common spatial relations* (e.g. a mouse usually appears to the right of a laptop) between objects.

We learn the knowledge priors from description corpus in three 3D scene datasets: the Standford Text2Scene Spatial Learning [1], Scenes and Descriptions for Text to Scene Generation [2] and SPARE [3]. To obtain clean data, we manually examined the descriptions to eliminate poor ones (e.g. the description which do not use spatial prepositions). As a result, we obtained 3740 descriptions.

A. Object Category Taxonomy

We define a taxonomy for those objects whose names are occurred in descriptions. We start from *basic-level* categories which are the level that we are usually fastest at learning, such as apples and cups. The taxonomy is the benchmark for exploiting domain knowledge for objects. By aligning it with upper-level KB such as WordNet [6], it allows for knowledge understanding and transferring from fine-to-coarse grained categories.

B. Object Innate Part

Innate part of an object determines its natural orientation which establishes an intrinsic reference frame [1]. For each basic category, we examine its *innate* part using the criteria concluded by prior work. For example, objects such as cameras and laptops have *innate* functional parts – lens and screens.

C. Spatial Knowledge

We calculated the similarity between pairs of sentences in different scene types using BLEU-like (n-gram) [8]. The score threshold is set to 0.87, ensuring that the similar sentences are grouped. Within each group, we learn object occurrences, support hierarchies and common spatial relations between objects. For instance, the probability of spatial relation pos between a reference object O_{ref} and a target object O_{tar} given scene type C_s can be defined as:

$$P_{pos} = \frac{count(O_{tar}, O_{ref}, pos|C_s)}{count(O_{tar}, O_{ref}|C_s)} \quad (1)$$

III. THE SPATIAL KNOWLEDGE BASE

We have improved the KB from three aspects: (1) Instead of manually mapping entities, the KB is aligned with SUMO (Suggested Upper Merged Ontology) [9] which has been linked to the WordNet [6] synsets; (2) We develop a `Language` ontology for representing the linguistic meaning of a sentence; and (3) With more rules defined, inference

L.Cao and Y.Kuno are with the Graduate School of science and Engineering, Saitama University, Saitama City, Saitama, 338-8570, Japan. (email: `caolu, kuno @cv.ics.saitama-y.ac.jp`)

[1] In English, three different reference frames are distinguished: *absolute*, *intrinsic* and *relative*. For interesting readers, we direct Levinson [7] for further reading.

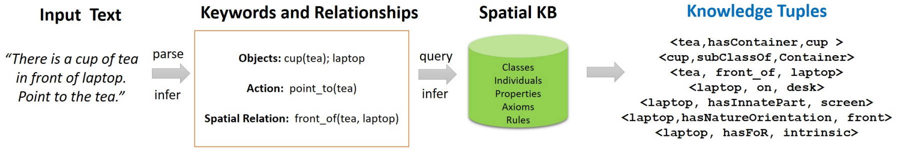
Input Text
"There is a cup of tea in front of laptop. Point to the tea."
parse
infer
Keywords and Relationships
Objects: cup(tea); laptop
Action: point_to(tea)
Spatial Relation: front_of(tea, laptop)
query
infer
Spatial KB
Classes
Individuals
Properties
Axioms
Rules
Knowledge Tuples
<tea,hasContainer,cup >
<cup,subClassOf,Container>
<tea, front_of, laptop>
<laptop, on, desk>
<laptop, hasInnatePart, screen>
<laptop,hasNatureOrientation, front>
<laptop, hasFoR, intrinsic>

Fig. 1. Pipeline of our approach.

TABLE I

EXAMPLES OF TEXTS AND PATTERNS.

Dependency Rule Pattern	Text Example	Attribute
{word:is}>nsubj{} = nsubj>prep({}=prep >plbj{}=pobj)	The cup is to the left of the bottle	`left_of(cup, bottle)`
{tag:VB}=verb>iobj{}=iobj >dobj({}=dobj>prep({}=prep >pobj{}=pobj))	Bring me a cup of coffee.	`bring(agent, cup_of_coffee)`
{}>advmod{}=advmod]>nsubj{} = nsubj	Where is the cup?	`reply(where, cup)`

of implicit facts are improved. Without this ability, users would be much more verbose such as "Bring me a cup of tea. The target is cup not tea. The cup is on the desk." Figure 2 shows the KB architecture. With knowledge extracted and inferred on the fly, a new knowledge hierarchy is generated autonomously. The knowledge is represented in the OWL-DL (Web Ontology Language-Description Logic) language as a collection of RDF (Resource Description of Framework) tuples, for instance `<cup on table>`.

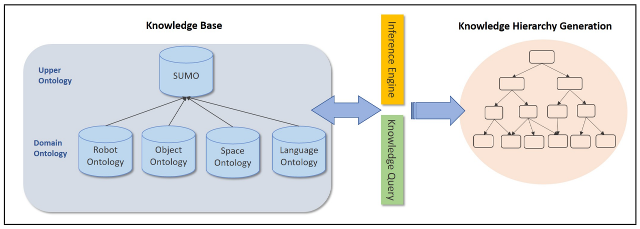

Fig. 2. Knowledge base structure.

The rules are categorized into two types: the R^I rules enables the reasoning within classes, while the R^B rules allow the knowledge to transfer between classes. In the following, we show two examples below.

- **[R^I] Reasoning about phrase.** If there are some nodes match the chain pattern A>prep(B)>pobj(C), then the words should be grouped together as "phrase". By querying the `Object` or `Space` ontology, noun phrases and spatial prepositions are determined.

  Example: a $\text{cup}_{[\text{noun}]}>\text{of}_{[\text{prep}]}>\text{coffee}_{[\text{pobj}]}$ *(noun phrase)*

- **[R^B] Reasoning about spatial relation.** If object O_1 in relation with object O_2, O_2 has the same relation with O_3, then O_1 is in relation with O_3, too. For example, if there is a book to the left of a laptop and a cup is to the left of the book, we can infer that the book is to the left of laptop.

IV. PARSING NATURAL LANGUAGE

Following the pipeline proposed by Chang et al. [1], we would consider more complex syntactic structures. During text parsing, we identify robot's actions, names of objects, attributes and the spatial relations. We first use the Standford Parser pipeline [10] to process input text which provides sentence splitting, part-of-speech tagging and syntactic dependency parsing. The tagged words are filtered with the KB to determine keywords. For example, to identify object categories, we look for nouns (e.g. cup) or noun phrases (e.g. cup of tea). With a set of associated keywords, we use Semgrex pattern [11] over Stanford dependencies to match the semantic form. The attribute types are determined from the KB query and inference. Table 1 shows some examples.

V. CONCLUSIONS

In this paper, we proposed an approach for spatial knowledge understanding in natural language. We first learned knowledge priors corresponding to natural language from datasets. We discussed the improvements of the spatial KB. We also showed the examples of inference of implicit facts based on knowledge priors. Finally, the proposed natural language parser is able to process various sentence structures and semantic forms. We are developing a dialog system for robots. An interesting line of future work is the generation of referring expressions (GRE).

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 26240038.

REFERENCES

[1] A. X. Chang, M. Savva, and C. D. Manning, "Learning spatial knowledge for text to 3d scene generation." in *EMNLP*, 2014, pp. 2028–2038.

[2] A. Chang, W. Monroe, M. Savva, C. Potts, and C. D. Manning, "Text to 3d scene generation with rich lexical grounding," in *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2015.

[3] L. Kunze, T. Williams, N. Hawes, and M. Scheutz, "Spatial referring expression generation for hri: Algorithms and evaluation framework," 2017.

[4] L. Cao, H. Fukuda, A. Lam, and Y. Kuno, "Communicating spatial knowledge in japanese for interaction with autonomous robots," in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*. IEEE, 2017, pp. 696–703.

[5] L. Cao, A. Lam, Y. Kuno, and D. Kachi, "Understanding spatial knowledge: An ontology-based representation for object identification," *IIEEJ transactions on image electronics and visual computing*, vol. 3, no. 2, pp. 150–163, 2015.

[6] G. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.

[7] S. C. Levinson, "Frames of reference and molyneuxs question: Crosslinguistic evidence," *Language and space*, vol. 109, p. 169, 1996.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[9] A. Pease, I. Niles, and J. Li, "The suggested upper merged ontology: A large ontology for the semantic web and its applications."

[10] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423–430.

[11] N. Chambers, D. Cer, T. Grenager, D. Hall, C. Kiddon, B. MacCartney, M.-C. De Marneffe, D. Ramage, E. Yeh, and C. D. Manning, "Learning alignments and leveraging natural logic," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics, 2007, pp. 165–170.