Review of:
"Advancing Robotic Perception with Perceived-Entity Linking"
ISWC 2024, Volume 2, Pg 192-209

Adamik et al.

Michael Rice

February 19, 2025

1 Paper Summary

This paper discusses a novel approach to robotic perception, known as Perceived Entity Linking or PEL, which operates by leveraging the Semantic Web to link visually perceived entities to uniquely identifiable entries in target knowledge repositories. The authors propose a fascinating approach, attempting to form the aforementioned link between what the robot perceives and what it knows using the raw sensory data captured by the robot's sensor array. [7]. This publication draws on techniques from both the fields of Computer Vision and Natural Language Processing to remedy several challenges in robotic perception, including entity typing (both with and without entity attributes i.e. size, colour etc.) as well as entity recognition and entity linking.

The primary technique from the field of computer vision deployed in this paper is object detection, in the form of the YOLO (You Only Look Once) algorithm [13]. Additionally, two techniques, inspired by the field of Natural Language Processing (NLP), are employed: Candidate Generation and Disambiguation. The Candidate Generation process is utilized to propose a set of potential perceived entities based on results from the object detection algorithm. For instance, if the label 'Orange' is returned, WordNet [5] is utilized to provide alternative variations or interpretations of the label, such as 'orange' referring to the fruit or 'Orange' referring to the city in California. Following on from this, the Disambiguation process seeks to refine this list of candidates, reducing it to either a single entity or no entity at all. These techniques combine to form a pipeline that operates under the following headings - observation, (entity) recognition and (entity) linking [7] in an attempt to equip robots with a form of 'commonsense' environmental knowledge.

2 Expected Impact of the Paper

I believe that the future impact of the work undertaken in this paper could potentially be significant, particularly in the field of general-purpose robotics. Enhancing 'common-sense' knowledge in robotic perception systems is the key to increasing the generalizability of such systems. The discussed concept of PEL [7], which will likely be central to any potential future impact this paper may have, is enabled by a variety of individual building blocks, which combine to achieve its overall goal.

First, the linking of sensory data to known entities in knowledge bases, could make use of the RDF model [2], which uses subject-predicate-object triples to create links between pieces of information. For example, if an object detection algorithm is fine-tuned to detect car manufacturer, and a Volvo car is being perceived by the robot, one such triple that could represent the situation would be Volvo - rdf:type - Car , as the 'Car' entity is linked to the 'Volvo' entry in the knowledge graph as a subclass. Furthermore, the fact that RDF is a standardized and machine-readable framework could potentially allow for the sharing, understanding and processing of the same information by multiple different systems without issue. RDF also supports ontology languages such as OWL [4], which provide a structured framework for helping robots understand entity types and their relationships. For example, OWL can define that a leaf is a type of plant, enabling more accurate linking of s ensory data to recognized entities and improving overall entity classification.

Secondly, SPARQL [3] and Turtle [1] could potentially be in use here as a framework for the efficient querying and serialization of linked knowledge. These two concepts are critical to the lowering of delays in the linking process, an important factor in the overall success and long-term implications of this work.

3 Primary Weakness

The main weaknesses of this implementation are twofold. First, although the peak achieved accuracy of the PEL process on the sample datasets, (69%), is commendable for an initial attempt, it falls short of the standards

required for real-world deployment. Second, both the size and quality of the knowledge base present significant challenges to any potential real-world deployment. An excessively large knowledge base can negatively impact real-time performance, while a low-quality knowledge base can compromise linking accuracy.

In order to first tackle the issue of linking accuracy, a number of solutions could be proposed. The first could augment the Disambiguation process through the use of a ViT (Vision Transformer) [10]. A ViT could be used to provide general scene context to the robot, enabling a more accurate and dependable process.

Secondly, to address the issues relating to the knowledge base itself, a three-part hybrid solution could be proposed to solve the size issues, in the form of dynamically-constructed[9], context-aware [12] and domain-specific [6] sub-graphs. As large knowledge bases can negatively impact real-time performance, sub-graphs could be constructed in real-time based on the scene's context, which could in turn be extracted using a variety of techniques, not limited to; object detection algorithms, (semantic) SLAM algorithms [11] or Vision Transformer implementations [10]. This would limit the knowledge graph to only what is currently relevant to the robot's surroundings, reducing search and retrieval delays. To actually extract entities from the graph that are related to the scene context, SPARQL [3] queries could again be used. For example, if the scene context is understood to be an object located in a kitchen, queries could be used to extract all entities with a sample 'foundIn' relationship to 'kitchen'. An example being:

```
SELECT * WHERE {?subject foundIn "kitchen".}
```

Finally, the issue of knowledge graph quality is more challenging to address, as it relies on externally available information. However, a solution to this issue could potentially involve the use of in-depth, domain-specific knowledge bases alongside more general, publicly available knowledge bases.

4 Comparison to Other Published Works

4.1 Work 1 : "Know Rob 2.0 — A 2nd Generation Knowledge Processing Framework for Cognition-Enabled Robotic Agents" - Beetz et al.

The first paper I will use for comparison to this work is "Know Rob 2.0" by Beetz et al., a second generation framework created to tackle knowledge representation and reasoning challenges for robotic agents. In contrast to the objective of the paper under review, where the goal was to link perceived entities to a predefined knowledge base, the goal of "Know Rob 2.0" is to 'bridge the gap' between user instructions such as 'open the fridge' and the detailed robotic actions/motions required to perform the task [8]. This process aims to incorporate environmental specifics into the robotic agent's decision-making process, such as the weight of an object that is to be picked up or the orientation it should be held in. In essence, the goal of this implementation is again to provide the robots with a degree of 'common-sense knowledge' about it's surroundings and how to use this knowledge to its advantage.

To further extend Beetz's work, PEL [7], as discussed by Adamik et al., could be integrated into "Know Rob 2.0" in an attempt to more accurately determine/understand objects in the robot's environment, which are necessary to perform the task at hand. This could potentially enhance their ability to perform complex tasks by augmenting the decision-making process with detailed object attributes acquired from the accompanying knowledge base.

4.2 Work 2 : "Grounding Robot Sensory and Symbolic Information Using the Semantic Web" - Stanton and Williams

The second paper I will use for comparison is "Grounding Robot Sensory and Symbolic Information Using the Semantic Web" by Stanton and Williams. This paper concerns a theoretical discussion of symbol grounding (linking internal symbols used in decision-making to the real-world phenomena they refer to) [14] using ontologies (frameworks for knowledge representation and machine understanding of data and relationships) originally designed for the Semantic Web. In comparison to Adamik et al.'s and Beetz et al.'s paper, Stanton and Williams' work is an earlier and more foundational foray into understanding how to map raw sensory data to real-world information. As such, it is interesting to compare it to more recent works. However, Stanton and Williams provide no details on an implementation, offering only a discussion of the broader concepts and challenges, which sets their work apart from the other two papers under review.

In their paper, Stanton and Williams focus on using a soccer-playing robot as a canvas to illustrate their framework and approach. They explore the use of ontologies in the grounding process (e.g. linking an orange blob in the robot's vision to a soccer ball, due to known properties of the ball.) By doing so, robots are able to build complex, internal models of their environments, allowing them to reason and plan as well as predict future states [14]. The role of ontologies in this is to aid in the description of entities is terms of their sensory data, which in turn allows a connection between the 'Perception Layer' and the 'Symbolic Layer' [14].

References

[1] RDF 1.1 Turtle. https://www.w3.org/TR/turtle/.

[2] Resource Description Framework (RDF) Model and Syntax Specification. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[3] SPARQL 1.1 Query Language. https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[4] Web ontology language: OWL. https://www.researchgate.net/publication/2844629_Web_ontology_language_OWL.

[5] WordNet: A lexical database for English: Communications of the ACM: Vol 38, No 11. https://dl.acm.org/doi/10.1145/219717.219748.

[6] Bilal Abu-Salih. Domain-specific Knowledge Graphs: A survey, March 2021.

[7] Mark Adamik, Romana Pernisch, Ilaria Tiddi, and Stefan Schlobach. Advancing Robotic Perception with Perceived-Entity Linking. In *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings, Part II*, pages 192–209, Berlin, Heidelberg, November 2024. Springer-Verlag.

[8] Michael Beetz, Daniel Beßler, Andrei Haidu, Mihai Pomarlan, Asil Kaan Bozcuoğlu, and Georg Bartels. Know Rob 2.0 — A 2nd Generation Knowledge Processing Framework for Cognition-Enabled Robotic Agents. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 512–519, May 2018.

[9] Yaxi Chen and Xuefeng Xing. Constructing Dynamic Knowledge Graph Based on Ontology Modeling and Neo4j Graph Database. In *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 522–525, May 2022.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

[11] Feiya Li, Chunyun Fu, Dongye Sun, Jian Li, and Jianwen Wang. SD-SLAM: A Semantic SLAM Approach for Dynamic Scenes Based on LiDAR Point Clouds, February 2024.

[12] Weiran Pan, Wei Wei, and Xian-Ling Mao. Context-aware Entity Typing in Knowledge Graphs. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2240–2250, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection, May 2016.

[14] Christopher Stanton and Mary-Anne Williams. Grounding Robot Sensory and Symbolic Information Using the Semantic Web. In Daniel Polani, Brett Browning, Andrea Bonarini, and Kazuo Yoshida, editors, *RoboCup 2003: Robot Soccer World Cup VII*, pages 757–764, Berlin, Heidelberg, 2004. Springer.