Advancing Robotic Perception with Perceived-Entity Linking

Mark Adamik[1][0000−0002−7977−3617], Romana Pernisch[1,2][0000−0001−8590−1817], Ilaria Tiddi[1][0000−0001−7116−9338], and Stefan Schlobach[1][0000−0002−3282−1597]

[1] Computer Science Department, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[2] Discovery Lab, Elsevier, Amsterdam, The Netherlands
{m.adamik, r.pernisch, i.tiddi, k.s.schlobach}@vu.nl

Abstract. The capabilities of current robotic applications are significantly constrained by their limited ability to perceive and understand their surroundings. The Semantic Web aims to offer general, machine-readable knowledge about the world and could be a potential solution to address the information needs of robotic agents. We introduce the Perceived-Entity Linking (PEL) problem as the task of recognizing entities and linking the sensory data of an autonomous agent to a unique identifier in a target knowledge graph. We provide a formal definition of PEL, and propose a PEL baseline based on the YOLO object detection algorithm and a conventional entity linking method as an initial attempt to solve the task. The baseline is evaluated by linking the concepts contained in MS COCO and VisualGenome datasets to WikiData, DBpedia and YAGO as target knowledge graphs. This study makes a first step in allowing robotic agents to leverage the extensive knowledge contained in general-purpose knowledge graphs.

Keywords: Robotic Perception · Knowledge Graphs · Semantic Web

1 Introduction

The new developments in Artificial Intelligence (AI) and Robotics aim to develop autonomous agents that can operate in the ever-changing world beyond the confines of laboratory setups. Physical autonomous agents such as robots deployed outside of a controlled environment need to acquire some form of signal from their surroundings that constitutes the system's perception. These percepts need to be interpreted in order to serve as a meaningful basis for the decisions the robot needs to make to achieve its task. The decisions are often informed by some background knowledge [6], that may be either manually embedded in the control architecture by the designers of the system or drawn from some external knowledge stored in a knowledge repository [15]. A considerable effort has been directed towards improving these knowledge repositories with common-sense knowledge, which robotic agents often lack [10].

The ambitious vision of the Semantic Web [7] is to transform the vast amount of unstructured information found on the internet into a machine-readable global

Fig. 1. A depiction of the overall vision of the PEL task.

knowledge repository. Several existing projects aim at constructing such knowledge repositories, e.g. WikiData [31], DBpedia [22] and YAGO [28], where information is structured in the form of knowledge graphs. The information within these graphs is often drawn from Wikipedia, WordNet, web scrapers and other existing data sources, and often contains information that can be considered common sense. Therefore, this information could serve as a potentially useful background knowledge repository for autonomous agents trying to operate in an unknown environment. However, to utilize the information stored within these resources, robots will have to establish a link between these repositories and the surrounding environment.

In this work, we focus on the problem of establishing the link between the percepts arriving from various sensors of the robotic agents and knowledge repositories on the Semantic Web. The main research problem guiding the efforts of the paper can be summarized as:

> ***Research Problem:*** *How can a robotic agent that is equipped with a set of sensors associate the incoming sensory data with a corresponding unique identifier in a target knowledge graph?*

To answer this question, we start by defining Perceived-Entity Linking (PEL) as the problem of linking the percepts of agents to entities in a knowledge base that correspond to those percepts. An example of this can be seen in Figure 1, where the object recognition and visual scene understanding capabilities of robotic agents are enhanced by linking the sensory data to target knowledge graphs (i.e. WikiData). The task is inspired by the Entity Linking (EL) and Entity Typing (ET) problems of the Named Entity Recognition task of Natural Language Processing (NLP). We showcase a solution that serves as a preliminary baseline, focusing on examining how the conventional entity linking methods could be adopted to solve the PEL problem. To evaluate the baseline, we conduct an experiment using the concepts contained in the MS COCO [21] object recognition

and Visual Genome [19] scene graph datasets[3]. Our work therefore offers the following contributions:

- We introduce and define the problem of PEL, providing a mathematical formulation of the task and describing motivating scenarios.
- We implement a first PEL system relying on conventional entity linking methods, and evaluate it to establish a baseline for the PEL task.

The rest of the paper is structured as follows. Section 2 briefly overviews the related work, while Section 3 provides the motivation and a formal definition of the Perceived-Entity Linking problem. In Section 4, we present a baseline PEL system, and evaluate it, as well as discuss the results in Section 5. Section 6 concludes the paper and describes directions for future work.

2 Related Work

Knowledge-driven robotics is a sub-field within robotics that is concerned with ontology-based robotic applications (for a comprehensive survey, consult [2,1]). We focus on robotic systems that utilize some form of ontology and examine how they perform the linking process between their perception system and the knowledge base.

Despite the longstanding nature of the Semantic Web, only limited research focuses on its combination with robotic agents. One of the earliest approaches aiming to use the Semantic Web to ground robot perception is [27], which focuses on describing the RoboCup domain using the Resource Description Framework (RDF). Fischer et al. [15] use a combination of computer vision algorithms (YOLO) and external knowledge sources (WikiData and WordNet) to provide suggestions for tools and actions when encountering unknown objects and actions. Daoutis et al. [11] integrate Cyc with the perceptual system of a robot and then use Cyc to interact with the users about the percepts using natural language. In their work, ambiguities that arise when linking percepts to concepts in Cyc are resolved manually.

Young et al. [35] proposed a lifelong object learning system that utilizes DBpedia to suggest labels for unknown perceived objects by taking the surrounding known objects into account. The authors expanded the previous work in [36], where they utilized the context-based web-mining results to filter the candidate labels provided by the deep-learning-based vision system. Although both of these systems would serve as an example that is partially similar to the PEL task, the problem is only implicitly addressed and no analysis is provided.

As one of the foundational knowledge-driven robotic perception system, RoboSherlock [3,5] of the larger KnowRob infrastructure [29] proposes a set of annotators to provide a richer semantic description of object attributes such as color, shape, size, etc. Although KnowRob follows Semantic Web standards

[3] The results of the experiment, as well as the implementation of the architecture can be openly accessed online `https://github.com/Dorteel/pel`

such as OWL and RDF, there is no implementation, where external knowledge sources are exploited. RoboEarth [30,32] was an ambitious project running between 2010 and 2014 that aimed to create a "World Wide Web for Robots", by allowing autonomous agents to access a shared database designed to store common-sense knowledge. However, after the project ended, this database became inaccessible. This further reinforces our argument that robots should rely on the publicly maintained Semantic Web repositories.

Also relevant to the problem of PEL is the work done on multi-modal [16] and visual entity linking [37], that combine textual and visual representations to increase the efficiency of entity linking methods. The main focus of these attempts is linking Named Entities such as celebrities and brands, and do not consider common objects in the context of embodied agents.

Lastly, some examples of tasks similar to PEL can be found in the domains of Internet of Things (IoT) and ambiguous computing. In [25], the authors use a semi-manual process to link entities from lifelogs to their corresponding entities in WikiData based on string matching, which was manually corrected to enable semantic search over the lifelogs. Wu et al. [33] proposes Semantic Web of Things, where a machine learning-based entity linking method links sensory metadata to DBpedia. Their results suggest that more knowledge engineering is needed to increase the accuracy of the linking.

While certain systems implicitly incorporate a linking process between the robots' sensors and the utilized knowledge repositories, there is currently a lack of comprehensive exploration, definition, and formalization of the problem in existing research. Moreover, no framework exists that would systematically examine and address this linking problem, impeding a deeper understanding and development of solutions.

3 Problem Definition and Motivation

We begin by describing the PEL problem, delineating related concepts, and defining the perception pipeline in robotic agents. We follow with defining PEL and formulating its components. Ultimately, we present three motivating scenarios for the task.

3.1 Problem Description

The main purpose behind this task is to equip agents with common-sense knowledge about the environment. We propose to achieve this by relating the incoming sensory data to a unique identifier in a target knowledge repository, as shown in Figure 2.

Related concepts. One of the most well-known concepts is the *symbol grounding problem*, which is closely related and originally proposed by Harnad in 1990 [17]. It addresses the challenge of connecting symbols, such as words or other abstract representations to objects and concepts in the real world. *Object*

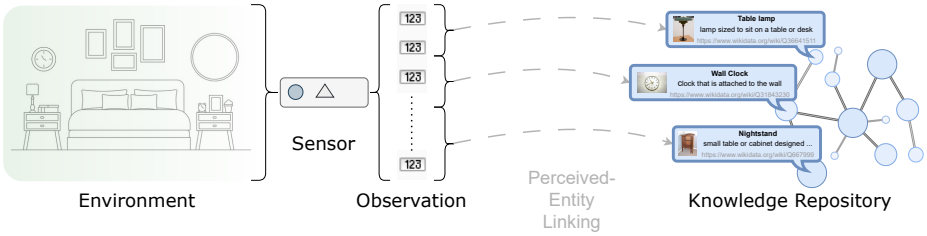

Fig. 2. The general problem addressed by PEL concerns the linking of regions in the sensory data to entities in a target knowledge graph.

recognition is a task originating from the field of cognitive sciences and taken up by computer vision, that concerns a set of tasks involving object identification and categorization [24]. Balogh et al. [4] describes *entity linking* as a task that concerns the recognition and linking of entity mentions in a text to the corresponding entities in a target knowledge repository. From this perspective, the task of PEL can be seen as a specialized subset of object recognition. More specifically, PEL sits at the intersection of object recognition and entity linking, offering a unique approach to the grounding problem.

General perception pipeline. The general robotic perception process can be described as follows. A robot r is equipped with a set of sensors S^r. Given a set of world states $\mathcal{X}$, there exists a set of entities (objects) $\mathcal{E}_\mathcal{X}$, a set of attribute types $\mathcal{A}$ that the entities could possess, and a set of relationship types P that can exist between entities. We define f_o^s as an observation process performed by sensor $s \in S^r$, that yields a noisy representation of part of the world state $f_o^s : \mathcal{X} \mapsto \mathcal{O}$, where $\mathcal{O} = \{o_1, ..., o_n\}$. To maintain the generality of the representations, observation o_i here is represented as an n-dimensional tensor of real numbers $o_i = \{\mathbf{X} \mid \mathbf{X} \in \mathbb{R}^{d_1 \times d_1 \times ... \times d_n}\}$. The dimensionality of the observation depends on the sensor s providing the observations. The task of a general perception pipeline is then to determine which entities, attributes and relationships are contained within the observation. The description of observations is then used as a starting point for the problem definition.

3.2 Problem Definition

We provide an initial definition of the problem by adapting the entity linking task definition from [4] to encompass the sensory aspect of the problem.

Definition 1 (Perceived-Entity Linking). *Given an observation process providing sensory data, Perceived-Entity Linking is the task of* ***recognizing*** *entities in the sensory data and* ***mapping*** *them to the corresponding entries in a knowledge repository.*

Using this definition, we divide PEL into three steps: an observation function f_o, a recognition function f_r and a mapping function f_m. As the observation

function is usually sensor-dependent and contained within the hardware level, we will henceforth focus on the functions f_r and f_m.

Recognition process. To address the recognition step, we first refer to the formulation provided by Besl et al. [8]. They describe the different components that a general object recognition system should possess. Most notably, the authors emphasize the importance of a symbolic description process that takes place during object recognition, where the relevant features are extracted from the sensory data and described symbolically.

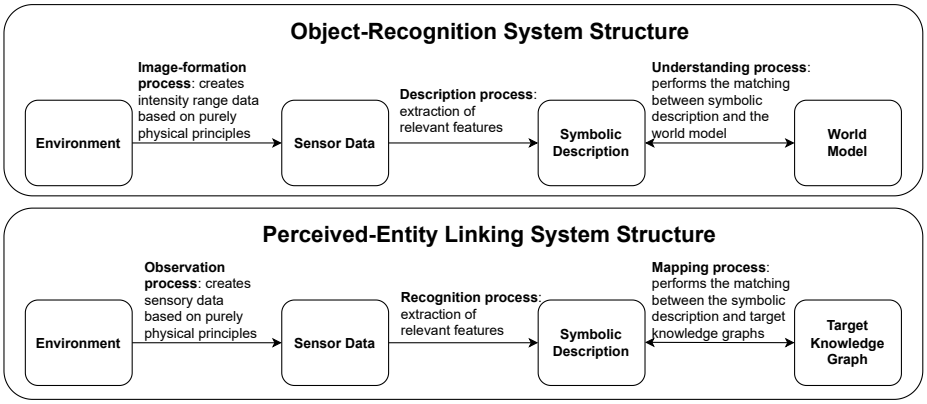

Fig. 3. An overview of the PEL system describing the components of the task from a visual object recognition perspective (the object-recognition system structure is adopted from [8]).

These symbolic descriptions are then matched to a world model through an 'understanding process', which, in turn, influences the symbolic descriptions. A key advantage of adopting this approach is that it highlights how the target knowledge bases in question also represent information in a symbolic format. As stated in [8], the term *recognition* refers to knowledge of something already known, which implies the existence of some form of prior world model. The model of the world considered in the object recognition process is analogous to the target knowledge graph we want to link the sensory data to during the PEL task. A depiction of the object recognition process as envisioned in [8], and its comparison to the PEL task can be seen in Figure 3.

In state-of-the-art approaches for object recognition based on neural networks, the *world model* and *understanding processes* steps seem to have become implicit and essentially a "black box", while the task of object recognition has evolved into a sub-task of scene graph generation [20]. Despite these advances, we contend that contemporary visual scene understanding approaches [38] do not consider the understanding process as conceptualized by [8], but generate

detailed symbolic descriptions in the form of scene graphs. Therefore, to characterize the recognition process of PEL more generically, we adopt the definition of a scene graph from [38]. As scene graphs are predominantly concerned with visual data a more general formulation is proposed in the form of *observation graphs* to encompass a wider variety of sensors.

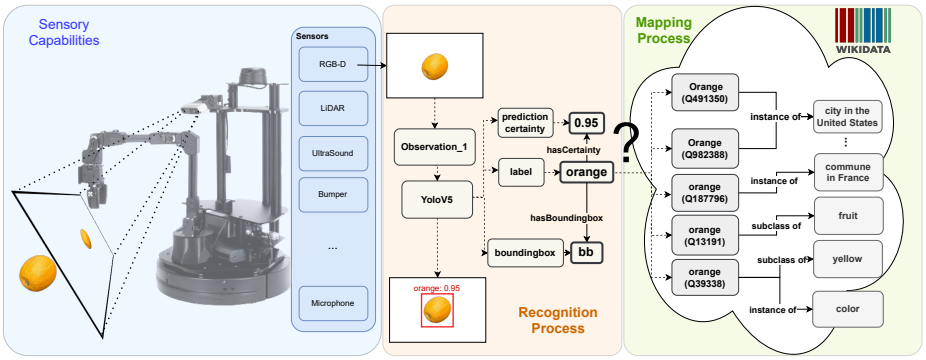

Fig. 4. An example of the PEL problem. Given an object recognition algorithm, such as the YOLO [9] algorithm and the resulting output, how should the mapping process f_m link the entity e_{G_o} to the corresponding entity $e_{G_t} \in G_t$.

Formally, given an observation o provided by sensor S, an observation graph describing the observation is defined as a set of triples $G_o \subseteq E_o \times P \times (E_o \cup A)$ where E_o is the set of entities, A denotes the attribute set and P denotes the relation set. Each entity $e \in E_o$ is denoted by a literal $l_e \in L$ representing the label of the entity, and is grounded to the observation by a location and a time attributes $a_e^{loc}, a_e^{time} \in A$, where $a_e^{loc} = \{\mathbf{x} | \mathbf{x} \in \mathbb{R}^n\}$ and $a_e^{time} = \{\mathbf{x} | \mathbf{x} \in \mathbb{R}\}$. With this formulation, the location attribute is not constrained to the form of a bounding box but could encompass several representation formats, including the implicit assumption of an image classification process, where the location of the entity could be considered the entire image. The recognition process $f_r : \mathcal{O} \mapsto \mathcal{G}$ therefore associates entities with a region of the sensory input (e.g. a segmentation map or a bounding box).

Mapping process. As mentioned before, the mapping process in PEL can be considered analogous with the understanding process outlined in [8], where the world model is represented by the target knowledge graph. Although several definitions of knowledge graphs exist [14], to allow for a general description with the least assumptions of the underlying format, we define a target knowledge graph in line with [18], as a tuple $G_t := (N_t, L_t)$, where N_t is a finite set of

nodes, and L_t is a finite set of links between nodes such that $L_t \subseteq N_t \times P_t \times N_t$, where P_t is a set of properties.

To define the mapping process, a definition of entity linking provided in [26] is adopted with the inclusion of linking entities to classes as well as entities in G_t. Therefore, given a set of entities in a target knowledge graph E_{G_t} and a set of entities in the observation graph E_{G_o}, the mapping process f_m aims to map each entity $e_i \in E_{G_o}$ to its corresponding entity $e_j \in E_{G_t}$, where cases of $e_{G_o} \notin E_{G_t}$ are mapped to $l_\emptyset$. Hence, the mapping process $f_m : G_o \times G_t \mapsto G_l$ results in a linked knowledge graph $G_l = (E_{G_o}, P, E_{G_t})$.

Fig. 5. An example image is taken from the MS COCO, where instances of chairs are annotated (*left*). The task of an ideal Perceived-Entity Linking system is to recognize the different types of chairs and provide a more fine-grained resolution of class labels. An example of this using WikiData is provided on the (*right*).

3.3 Motivating Scenarios

Perceived-Entity Typing. As opposed to the conventional entity linking task of natural language processing, where there is a single correct entity that the mention in the text refers to, the symbolic descriptors in PEL can yield a class label, which then can refer to one of multiple possible sub-classes. An example of this can be seen in Figure 5, where the MS COCO dataset annotates different types of chairs with the umbrella term *chair*. A robotic agent should instead distinguish between an office chair and an armchair, to enable reasoning over their differences (e.g. an office chair with wheels could be moved around easier than an armchair).

Perceived-Entity Recognition. In this task, the robot needs to link instances of objects from the perceptual modalities to instances in the target knowledge graph. This task exhibits a more pronounced connection to situated scene understanding, wherein object instances possess instance-specific attributes (i.e. location, orientation, colour) that can be exploited. This does not necessarily

apply to more general class characteristics that are found on the general-purpose knowledge graphs. An example of this could be a mobile robot operating in a household environment, tasked with fetching a cup for a person. In this scenario, different instances of cups could exist, with each person living in the household owning a specific or favourite cup. The perceived objects should be linked to previously known, distinguishable instances of the cups related to the person, instead of the general *cup* class. While recognizing the significance of this task, especially in the context of autonomous agents, we defer its in-depth exploration to future work.

Attribute-based Perceived-Entity Typing. In some cases, the autonomous agent might be equipped with sensors that perceive the existence of physical entities in the surroundings (i.e. an RGB-D camera with a point cloud clustering) and can infer certain attributes, such as size, color or shape, but is unable to assign types to the entity. In such a case, the perceived-entity linker should be able to infer the type of an object, given a sufficient set of attributes describing it in the target knowledge graph. For example, a kitchen helper robot should be able to differentiate between oranges and grapefruits using the *colour* and *size* attributes.

4 Baseline System Design

Having defined the PEL task, we implement a simple PEL baseline composed of a recognition process f_r and a mapping process f_m. Given that the recognition process f_r can be achieved in multiple ways, we opt for the simplest scenario, where recognition is performed by the off-the-shelf YOLO algorithm [9]. As seen in Section 3.3, a common challenge across all the scenarios motivating PEL is establishing the correct mapping between the labels provided in the observation graph, and the classes or entities in the target knowledge graph. This raises the question of how to map literals representing entities in the observation graph to entities or classes of the target knowledge graph. Inspired by the conventional NLP pipeline [4,26], we implemented a mapping process f_m composed of a candidate selection process f_c and a disambiguation process f_d. These will be explained in the rest of the section, while a depiction of this overall pipeline can be seen in Figure 6.

4.1 Candidate Selection

This sub-task considers the generation of candidate entities from the target knowledge graph G_t for each entity $e \in E_{G_o}$. The selection strategy aims to consider alternate spellings and variations of the words representing the entities. To achieve this, we utilize WordNet [23], where first the synsets corresponding to e_{G_o} are retrieved, and a dictionary is generated using the alternate lemmas of a synset. The steps the algorithm takes can be seen in Algorithm 1. In order to focus on accounting for alternate spellings, the most similar entities to

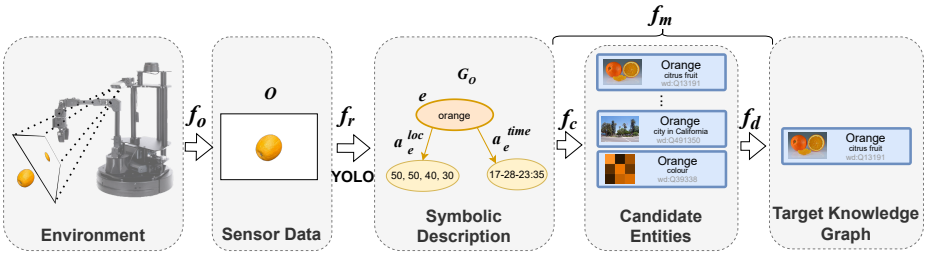

Fig. 6. f_o: Observation process, f_r: Recognition processes extract the relevant features and provide symbolic descriptions. We utilize the YOLO [9] object recognition algorithm to provide the entities e in the form of an observation graph. The mapping process f_m is divided into two processes: The candidate selection process f_c selects a list of candidates that could correspond to the relevant entity in the target knowledge graph. The Disambiguation process f_d aims to resolve the ambiguities and selects the best candidate to establish the link to.

the label are selected, using the Levenshtein edit-distance to calculate the similarity between two strings. Once a dictionary has been generated, all entities returned from G_t are added as potential candidates. As several candidates exist, disambiguation is needed to find the related entity.

Algorithm 1 Candidate Selection

```
function SELECT_CANDIDATES(e_{G_o})
    synsets ← GET_SYNSET(e_{G_o})                ▷ Find synsets of entity
    candidates ← []
    if ¬synsets then
        return e_{G_o}
    end if
    for sn in synsets do
        names ← GET_NAMES(sn)
        for each name do
            sim_score ← GET_SIMILARITY(e_G, name)
        end for
        candidates ← sim_score, name
    end for
    candidates ← sorted(candidates)
    return candidates
end function
```

4.2 Disambiguation

Disambiguation corresponds to narrowing down the candidates resulting from the previous step to either a single entity or no entity. The simplest solution is to

select the most common-sense answer, for which commonness can be a suitable first candidate and provide a baseline. Other suitable approaches are examining the coherence among the already linked entities in the observation and the candidate entities, and observing the attributes of the entities provided by the recognition process. In our approach, a ranking is established based on relevant features, namely the similarity of the label for the class of each target candidate e_{i,G_t} and the perceived entity e_{G_o}. This approach aims to exploit the fact that the class relationships between these two entities could be a distinguishing feature. This design choice is motivated by the fact that in an embodied agent scenario, the prior importance of entities that correspond to physical entities should be considerably higher than that of abstract entities. As an example, the entities in WikiData corresponding to *orange* contain both the city in California and the fruit (see Figure 4). The algorithm is described in Algorithm 2. As a naive approach, the disambiguation process calculates a ranking based on the similarity, where the class synset with the shortest path to e_{G_o} in the WordNet taxonomy is considered as a similarity measure. The candidate to be linked then is the one with the highest similarity score.

Algorithm 2 Disambiguation

```
function DISAMBIGUATE(e_{G_o}, candidates, G_t)
    e_G_ss ← GET_SYNSET(e_{G_o})
    scores ← []
    for each c in candidates do
        candidate_classes ← GET_CLASS(candidate)
        high_score ← 0
        for each class in candidate_classes do
            class_ss ← GET_SYNSET(class)
            for each ss in class_ss do
                sim_score ← max(GET_SIMILARITY(class_ss, e_G_ss))
                high_score ← max(sim_score, high_score)
            end for
        end for
        scores[] ← (high_score, c)
    end for
    candidates ← sorted(candidates)
    return candidates
end function
```

5 Evaluation

In this section, we aim at evaluating the effectiveness of the mapping method of the PEL baseline described in Section 4. In other words, the goal of our evaluation is to assess how well a method based on conventional entity linking can map concepts provided by the recognition process to target knowledge graphs.

5.1 Experimental Design

We provide an evaluation of our method by performing a component analysis, e.g. comparing the effectiveness of the queries to the mapping process first by performing only candidate selection, then adding the disambiguation step. As a test-bed for the PEL task, we use two datasets. The first dataset is the 80 annotating labels of the MS COCO dataset [21] commonly employed by various object detection algorithms utilized in the recognition process f_r. The Visual Genome Scene Graph dataset [19] is used as the second dataset to assess the scalability of our method. The target knowledge graphs we use in our experiments are WikiData, DBpedia and YAGO. The task is an ad-hoc entity retrieval with the keywords being the labels of the dataset. The following queries are used for the retrieval:

- **Simple query (S)**: Ask for any entity with any property, where the object o corresponds to the e_{G_o} label encompassed with the @en language tag.
- **Simple query with links (S_{links})**: The previous query is extended by counting the outgoing links from the entity e_{G_t}, and ordering the results in a descending order to encourage more common elements.
- **Label query (L):** In this query, only relations with `rdfs:label` predicates are considered where the object corresponds to the e_{G_o} label, with an added @en language tag.
- **Label query with links(L_{links})** Similarly to the simple queries, the results of the previous query are ordered based on the number of outgoing edges.

Ground Truth. In order to establish the ground truth for the MS COCO objects, we manually searched for entities in the target knowledge graphs. The entries were left empty when we did not find any concepts corresponding to the labels. In some cases, no exact concepts were found, but viable alternatives were available. For example, none of the knowledge graphs contain the concept of a *sports ball*, but all of them contained the concept of a *ball*, which we defined as a desirable concept to map to.

Establishing the ground truth for the Visual Genome dataset is more straightforward, as it contains mappings for all the concepts within the scene graphs to WordNet 3.1. We utilize these mappings to find the corresponding WikiData entities using the matching WordNet 3.1 Synset ID (P8814) property. The DBpedia and YAGO entries are then collected by the corresponding `owl:sameAs`, resulting in 2,399 entries. The labels are yielded by taking the synset's lemma. Among the 76,340 object classes identified in Visual Genome, 2,399 classes are consistently represented across all three knowledge graphs. Given that a representative sample size of 2,328 classes is required for a 2% margin of error and a 95% confidence level, the 2,399 classes included in the experiment are considered sufficient. The curated datasets are available for re-use in the online repository.

Evaluation metrics. The main metric used for evaluation is the accuracy of the linking, measured in terms of the ratio of correctly linked mentions and all the links generated by the method.

5.2 Results

The results of the experiment indicate that both the knowledge graph under consideration and the query types influence the accuracy of the mapping process. In order to provide more insight into the workings of the system, we also considered cases when the ground truth candidate was in the hit@3 and hit@5 entities returned by the queries. A summary of the results can be found in Table 1.

MS COCO. Both the candidate selection and disambiguation steps outperform the simple queries in most cases when the hit@3 and hit@5 candidates are considered. A significant increase in performance when considering hit@1 can only be observed in the case of WikiData, where using query L with the `rdfs:label` predicate yields an accuracy of 0.69, corresponding to the correct mappings in 55 out of the 80 labels. When the candidate selection step is used without disambiguation, a considerable drop in performance can be observed. This suggests that in each G_t, there are entities corresponding to labels with many outgoing links, which push the correct entity further down the list.

As an additional step, we compared the results to DBpedia Spotlight[4] [22], and AIDA[5] [34], an off-the-shelf entity linking system for disambiguating named entities for DBpedia and YAGO2 respectively (WikiData did not have such easily accessible entity linking systems). When attempting to link the labels of MS COCO, DBpedia Spotlight achieved an accuracy of 0.525, whereas AIDA did not result in any hits, which indicates that most of these methods focusing on Named Entity Recognition do not translate well to common-sense objects.

Visual Genome. With most query types and knowledge graphs, the average correctly linked entities increase significantly, indicating that the conventional entity linking pipeline increases the accuracy of the linking even when with greater number of labels. Using DBpedia as G_t, a 0.51 hit@1 accuracy is achieved using both the candidate selection and disambiguation methods, whereas for WikiData and YAGO, the candidate selection method outperformed the disambiguation method. For WikiData, using the queries S_{links} and L_{links} with `rdfs:label` and for YAGO, the simple query yielded the highest results, with 45% of the labels linked to the correct target entity for both knowledge graphs.

5.3 Discussion

In certain cases, a performance drop can be observed when considering only the candidate selection step, especially when results are sorted based on the outgoing links. This indicates that incoming links might not be a good indicator of the candidates' correctness.

[4] http://spotlight.dbpedia.org/
[5] http://www.mpi-inf.mpg.de/yago-naga/aida/

Table 1. Comparison of accuracy results across MS COCO and Visual Genome datasets using different conditions and processes. The table shows the hit@1, hit@3, and hit@5 metrics for three knowledge bases: WikiData, DBpedia, and YAGO, evaluated for various query types and processes. f_q represents running the mapping function using only the query without the candidate selection or disambiguation steps.

Method		**WikiData**			**DBpedia**			**YAGO**		
Query	**Process**	**hit@1**	**hit@3**	**hit@5**	**hit@1**	**hit@3**	**hit@5**	**hit@1**	**hit@3**	**hit@5**
MS COCO Dataset										
S	f_q	0.08	0.19	0.4	0.19	0.4	0.48	0.49	0.59	0.7
	f_c	0.07	0.23	0.41	0.19	0.4	0.47	0.51	0.65	0.72
	$f_c + f_d$	0.07	0.23	0.41	0.19	0.41	0.49	0.51	0.69	0.71
S_{links}	f_q	0.59	**0.94**	0.94	0.24	0.65	0.75	**0.71**	**0.78**	0.78
	f_c	0.59	0.71	0.95	0.24	0.19	0.56	0.62	0.75	**0.8**
	$f_c + f_d$	0.59	**0.94**	**0.97**	0.24	0.65	0.75	**0.71**	**0.78**	**0.8**
L	f_q	0.5	0.7	0.65	0.68	0.84	0.84	0.4	0.57	0.65
	f_c	**0.69**	0.75	0.8	0.69	0.86	0.86	0.53	0.62	0.68
	$f_c + f_d$	**0.69**	0.78	0.79	0.69	**0.88**	**0.88**	0.55	0.61	0.68
L_{links}	f_q	0.51	0.8	0.8	**0.84**	0.84	0.84	0.64	0.65	0.65
	f_c	0.14	0.66	0.82	0.45	0.85	0.86	0.28	0.64	0.69
	$f_c + f_d$	0.51	0.8	0.84	**0.84**	**0.88**	**0.88**	0.64	0.68	0.69
Visual Genome Dataset										
S	f_q	0.01	0.07	0.17	0.09	0.16	0.2	0.37	0.5	0.55
	f_c	0.04	0.14	0.25	0.13	0.22	0.27	**0.45**	0.6	0.66
	$f_c + f_d$	0.22	0.34	0.37	0.14	0.22	0.28	0.38	0.55	0.64
S_{links}	f_q	0.09	0.36	0.64	0.08	0.2	0.31	0.26	0.53	0.66
	f_c	0.1	0.31	0.5	0.11	0.22	0.33	0.27	0.54	**0.68**
	$f_c + f_d$	0.31	0.5	0.62	0.22	0.31	0.38	0.4	**0.59**	0.7
L	f_q	0.41	0.46	0.48	0.37	0.48	0.49	0.36	0.44	0.47
	f_c	**0.45**	**0.52**	0.54	0.4	0.59	0.62	0.43	0.53	0.58
	$f_c + f_d$	0.29	0.44	0.52	**0.51**	0.62	0.63	0.42	0.54	0.59
L_{links}	f_q	0.08	0.43	**0.68**	0.34	**0.63**	**0.66**	0.25	0.47	**0.68**
	f_c	0.05	0.2	0.35	0.24	0.47	0.54	0.19	0.37	0.56
	$f_c + f_d$	0.24	0.39	0.58	0.44	0.61	0.61	0.39	0.53	0.61

However, the overall approach results in some performance increases, which indicates that a more sophisticated approach along the lines of conventional entity-linking processes could be investigated.

Although in this PEL baseline only WordNet alternatives were considered to generate the candidates, in a more comprehensive system additional attributes resulting from the recognition processes could be utilized. If, for example, the detection contains a segmented area from the depth camera of the robot, a color descriptor could extract the dominant color of the region, and the knowledge graph could be searched for entities that have a corresponding property.

Overall, the results of the linking process when using MS COCO are higher. A potential explanation could be that the dataset is manually curated, and is generally cleaner. Furthermore, it contains more general objects, whereas Visual Genome sometimes contains entities that could not be considered common sense.

In our experiments, labels were examined individually. For a more complete disambiguation, contextual features based on the surroundings of the robot could be influential when determining the correct entity the labels should be linked to. To tackle this issue, alternative approaches like semantic relatedness, which assesses the similarity between entities in G_t and the attributes associated with e_{G_o}, can be employed. Semantic relatedness can be indicated by matching the attributes of entities, as suggested by Dredze et al [12]. Lastly, modality matching, which considers the fact that sensory data may contain information corresponding to multiple modalities, could also be considered for the disambiguation process. This could be useful when using target knowledge graphs representing entities beyond textual format, such as WikiData and DBpedia containing images and sound information.

6 Conclusion

In this paper, we introduce the Perceived-Entity Linking task, which involves mapping the entities detected by an autonomous agent's sensors to a target knowledge graph. We define this problem as the integration of recognition and mapping processes, and we outline how these processes can be combined to enhance sensory data interpretation. Furthermore, we discussed a few motivating scenarios where this approach can be applied. As a technical contribution, we present a PEL baseline, where our research explores the effectiveness of a conventional entity linking strategy, specifically how well object labels from the MS COCO and Visual Genome datasets can be mapped to entities in knowledge graphs such as WikiData, DBpedia, and YAGO. The results establish a benchmark for how autonomous agents can leverage existing knowledge structures to improve their interaction with the physical world.

Future work can extend PEL in several directions. In the recognition process, a combination of symbolic descriptors can be deployed, similar to robotic vision systems such as [5], where the observation graph could be generated by combining several computer vision algorithms. On the mapping process, the current formulation makes it possible to investigate different ontology alignment meth-

ods [13] that could aid in finding corresponding entities between the observation graph and the target knowledge graph. Lastly, including the context of the perceived entities (e.g. what other entities and attributes are recognized by the descriptors) could significantly boost the performance of the mapping process.

In conclusion, this study has explored strategies for enabling robotic agents to harness the Semantic Web's extensive knowledge. By focusing on advancing common-sense reasoning and perceptual understanding, we aim to develop methodologies that significantly enhance the capabilities of autonomous agents.

References

1. Aguado, E., Sanz, R.: Using ontologies in autonomous robots engineering. Robotics Software Design and Engineering **71** (2021)
2. Alarcos, A.O., Beßler, D., Khamis, A.M., Gonçalves, P.J.S., Habib, M.K., Bermejo-Alonso, J., Barreto, M., Diab, M., Rosell, J., Quintas, J., Olszewska, J.I., Nakawala, H., de Freitas, E.P., Gyrard, A., Borgo, S., Alenyà, G., Beetz, M., Li, H.: A review and comparison of ontology-based approaches to robot autonomy. Knowl. Eng. Rev. **34**, e29 (2019). https://doi.org/10.1017/S0269888919000237, `https://doi.org/10.1017/S0269888919000237`
3. Bálint-Benczédi, F., Worch, J., Nyga, D., Blodow, N., Mania, P., Márton, Z., Beetz, M.: Robosherlock: Cognition-enabled robot perception for everyday manipulation tasks. CoRR **abs/1911.10079** (2019), `http://arxiv.org/abs/1911.10079`
4. Balog, K.: Entity-Oriented Search, The Information Retrieval Series, vol. 39. Springer (2018). https://doi.org/10.1007/978-3-319-93935-3, `https://doi.org/10.1007/978-3-319-93935-3`
5. Beetz, M., Balint-Benczedi, F., Blodow, N., Nyga, D., Wiedemeyer, T., Marton, Z.: Robosherlock: Unstructured information processing for robot perception. In: IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015. pp. 1549–1556. IEEE (2015). https://doi.org/10.1109/ICRA.2015.7139395, `https://doi.org/10.1109/ICRA.2015.7139395`
6. Beetz, M., Beßler, D., Haidu, A., Pomarlan, M., Bozcuoglu, A.K., Bartels, G.: 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 2018, Brisbane, Australia, May 21-25, 2018. pp. 512–519. IEEE (2018). https://doi.org/10.1109/ICRA.2018.8460964, `https://doi.org/10.1109/ICRA.2018.8460964`
7. Berners-Lee, T., Hendler, J.A., Lassila, O.: The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. In: Seneviratne, O., Hendler, J.A. (eds.) Linking the World's Information - Essays on Tim Berners-Lee's Invention of the World Wide Web, ACM Books, vol. 52, pp. 91–103. ACM (2023). https://doi.org/10.1145/3591366.3591376, `https://doi.org/10.1145/3591366.3591376`
8. Besl, P.J., Jain, R.C.: Three-dimensional object recognition. ACM Comput. Surv. **17**(1), 75–145 (1985). https://doi.org/10.1145/4078.4081, `https://doi.org/10.1145/4078.4081`
9. Bochkovskiy, A., Wang, C., Liao, H.M.: Yolov4: Optimal speed and accuracy of object detection. CoRR **abs/2004.10934** (2020), `https://arxiv.org/abs/2004.10934`

10. Brachman, R.J., Levesque, H.J.: Toward a new science of common sense. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 12245–12249. AAAI Press (2022). https://doi.org/10.1609/AAAI.V36I11.21485, https://doi.org/10.1609/aaai.v36i11.21485
11. Daoutis, M., Coradeschi, S., Loutfi, A.: Grounding commonsense knowledge in intelligent systems. J. Ambient Intell. Smart Environ. **1**(4), 311–321 (2009). https://doi.org/10.3233/AIS-2009-0040, https://doi.org/10.3233/AIS-2009-0040
12. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Huang, C., Jurafsky, D. (eds.) COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China. pp. 277–285. Tsinghua University Press (2010), https://aclanthology.org/C10-1032/
13. Ehrig, M.: Ontology Alignment: Bridging the Semantic Gap, Semantic Web and Beyond: Computing for Human Experience, vol. 4. Springer (2007). https://doi.org/10.1007/978-0-387-36501-5, https://doi.org/10.1007/978-0-387-36501-5
14. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: Martin, M., Cuquet, M., Folmer, E. (eds.) Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016. CEUR Workshop Proceedings, vol. 1695. CEUR-WS.org (2016), https://ceur-ws.org/Vol-1695/paper4.pdf
15. Fischer, L., Hasler, S., Deigmoeller, J., Schnuerer, T., Redert, M., Pluntke, U., Nagel, K., Senzel, C., Ploennigs, J., Richter, A., Eggert, J.: Which tool to use? grounded reasoning in everyday environments with assistant robots. In: Steinbauer, G., Ferrein, A. (eds.) Proceedings of the 11th Cognitive Robotics Workshop 2018, co-located with 16th International Conference on Principles of Knowledge Representation and Reasoning, CogRob@KR 2018, Tempe, AZ, USA, October 27th, 2018. CEUR Workshop Proceedings, vol. 2325, pp. 3–10. CEUR-WS.org (2018), https://ceur-ws.org/Vol-2325/paper-03.pdf
16. Gan, J., Luo, J., Wang, H., Wang, S., He, W., Huang, Q.: Multimodal entity linking: A new dataset and A baseline. In: Shen, H.T., Zhuang, Y., Smith, J.R., Yang, Y., César, P., Metze, F., Prabhakaran, B. (eds.) MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021. pp. 993–1001. ACM (2021). https://doi.org/10.1145/3474085.3475400, https://doi.org/10.1145/3474085.3475400
17. Harnad, S.: Symbol grounding problem. Scholarpedia **2**(7), 2373 (2007). https://doi.org/10.4249/SCHOLARPEDIA.2373, https://doi.org/10.4249/scholarpedia.2373
18. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J.F., Staab, S., Zimmermann, A.: Knowledge graphs. ACM Comput. Surv. **54**(4), 71:1–71:37 (2022). https://doi.org/10.1145/3447772, https://doi.org/10.1145/3447772

19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. CoRR **abs/1602.07332** (2016), http://arxiv.org/abs/1602.07332
20. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 1270–1279. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.142, https://doi.org/10.1109/ICCV.2017.142
21. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48, https://doi.org/10.1007/978-3-319-10602-1_48
22. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Ghidini, C., Ngomo, A.N., Lindstaedt, S.N., Pellegrini, T. (eds.) Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011. pp. 1–8. ACM International Conference Proceeding Series, ACM (2011). https://doi.org/10.1145/2063518.2063519, https://doi.org/10.1145/2063518.2063519
23. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM **38**(11), 39–41 (1995). https://doi.org/10.1145/219717.219748, https://doi.org/10.1145/219717.219748
24. Mozer, M.: Object recognition: Theories. In: Smelser, N.J., Baltes, P.B. (eds.) International Encyclopedia of the Social and Behavioral Sciences, pp. 10781–10785. Pergamon, Oxford (2001). https://doi.org/https://doi.org/10.1016/B0-08-043076-7/01459-5, https://www.sciencedirect.com/science/article/pii/B0080430767014595
25. Rossetto, L., Baumgartner, M., Ashena, N., Ruosch, F., Pernischová, R., Bernstein, A.: Lifegraph: A knowledge graph for lifelogs. In: Gurrin, C., Schoffmann, K., Jonsson, B.T., Dang-Nguyen, D., Lokoc, J., Tran, M., Hürst, W. (eds.) Proceedings of the Third ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2020, Dublin, Ireland, June 8-11, 2020. pp. 13–17. ACM (2020). https://doi.org/10.1145/3379172.3391717, https://doi.org/10.1145/3379172.3391717
26. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2015). https://doi.org/10.1109/TKDE.2014.2327028, https://doi.org/10.1109/TKDE.2014.2327028
27. Stanton, C.J., Williams, M.: Grounding robot sensory and symbolic information using the semantic web. In: Polani, D., Browning, B., Bonarini, A., Yoshida, K. (eds.) RoboCup 2003: Robot Soccer World Cup VII. Lecture Notes in Computer Science, vol. 3020, pp. 757–764. Springer (2003). https://doi.org/10.1007/978-3-540-25940-4_75, https://doi.org/10.1007/978-3-540-25940-4_75
28. Tanon, T.P., Weikum, G., Suchanek, F.M.: YAGO 4: A reason-able knowledge base. In: Harth, A., Kirrane, S., Ngomo, A.N., Paulheim, H., Rula, A., Gentile, A.L., Haase, P., Cochez, M. (eds.) The Semantic Web - 17th Inter-

national Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12123, pp. 583–596. Springer (2020). https://doi.org/10.1007/978-3-030-49461-2_34, https://doi.org/10.1007/978-3-030-49461-2_34

29. Tenorth, M., Beetz, M.: Knowrob: A knowledge processing infrastructure for cognition-enabled robots. Int. J. Robotics Res. **32**(5), 566–590 (2013). https://doi.org/10.1177/0278364913481635, https://doi.org/10.1177/0278364913481635
30. Tenorth, M., Perzylo, A.C., Lafrenz, R., Beetz, M.: Representation and exchange of knowledge about actions, objects, and environments in the roboearth framework. IEEE Trans Autom. Sci. Eng. **10**(3), 643–651 (2013). https://doi.org/10.1109/TASE.2013.2244883, https://doi.org/10.1109/TASE.2013.2244883
31. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014). https://doi.org/10.1145/2629489, https://doi.org/10.1145/2629489
32. Waibel, M., Beetz, M., Civera, J., D'Andrea, R., Elfring, J., Gálvez-López, D., Häussermann, K., Janssen, R., Montiel, J.M.M., Perzylo, A., Schießle, B., Tenorth, M., Zweigle, O., van de Molengraft, R.: Roboearth. IEEE Robotics Autom. Mag. **18**(2), 69–82 (2011). https://doi.org/10.1109/MRA.2011.941632, https://doi.org/10.1109/MRA.2011.941632
33. Wu, Z., Xu, Y., Yang, Y., Zhang, C., Zhu, X., Ji, Y.: Towards a semantic web of things: A hybrid semantic annotation, extraction, and reasoning framework for cyber-physical system. Sensors **17**(2), 403 (2017). https://doi.org/10.3390/S17020403, https://doi.org/10.3390/s17020403
34. Yosef, M.A., Hoffart, J., Bordino, I., Spaniol, M., Weikum, G.: AIDA: an online tool for accurate disambiguation of named entities in text and tables. Proc. VLDB Endow. **4**(12), 1450–1453 (2011), http://www.vldb.org/pvldb/vol4/p1450-yosef.pdf
35. Young, J., Basile, V., Kunze, L., Cabrio, E., Hawes, N.: Towards lifelong object learning by integrating situated robot perception and semantic web mining. In: Kaminka, G.A., Fox, M., Bouquet, P., Hüllermeier, E., Dignum, V., Dignum, F., van Harmelen, F. (eds.) ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016). Frontiers in Artificial Intelligence and Applications, vol. 285, pp. 1458–1466. IOS Press (2016). https://doi.org/10.3233/978-1-61499-672-9-1458, https://doi.org/10.3233/978-1-61499-672-9-1458
36. Young, J., Kunze, L., Basile, V., Cabrio, E., Hawes, N., Caputo, B.: Semantic web-mining and deep vision for lifelong object discovery. In: 2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017. pp. 2774–2779. IEEE (2017). https://doi.org/10.1109/ICRA.2017.7989323, https://doi.org/10.1109/ICRA.2017.7989323
37. Zheng, Q., Wen, H., Wang, M., Qi, G.: Visual entity linking via multi-modal learning. Data Intell. **4**(1), 1–19 (2022). https://doi.org/10.1162/DINT_A_00114, https://doi.org/10.1162/dint_a_00114
38. Zhu, G., Zhang, L., Jiang, Y., Dang, Y., Hou, H., Shen, P., Feng, M., Zhao, X., Miao, Q., Shah, S.A.A., Bennamoun, M.: Scene graph generation: A comprehensive survey. CoRR **abs/2201.00443** (2022), https://arxiv.org/abs/2201.00443