Towards Lifelong Object Learning by Integrating Situated Robot Perception and Semantic Web Mining

Jay Young[1] and **Valerio Basile**[2] and **Lars Kunze**[1] and **Elena Cabrio**[2] and **Nick Hawes**[1]

Abstract.
Autonomous robots that are to assist humans in their daily lives are required, among other things, to recognize and understand the meaning of task-related objects. However, given an open-ended set of tasks, the set of everyday objects that robots will encounter during their lifetime is not foreseeable. That is, robots have to learn and extend their knowledge about previously unknown objects on-the-job. Our approach automatically acquires parts of this knowledge (e.g., the *class* of an object and its *typical location*) in form of ranked *hypotheses* from the Semantic Web using contextual information extracted from observations and experiences made by robots. Thus, by integrating situated robot perception and Semantic Web mining, robots can continuously extend their object knowledge beyond perceptual models which allows them to reason about task-related objects, e.g., when searching for them, robots can infer the most likely object locations. An evaluation of the integrated system on long-term data from real office observations, demonstrates that generated hypotheses can effectively constrain the meaning of objects. Hence, we believe that the proposed system can be an essential component in a lifelong learning framework which acquires knowledge about objects from real world observations.

1 Introduction

It is crucial for autonomous robots working in human environments such as homes, offices or factories to have the ability to represent, reason about, and learn new information about the objects in their environment. Current robot perception systems must be provided with models of the objects in advance, and their extensibility is typically poor. This includes both perceptual models (used to recognise the object in the environment) and semantic models (describing what the object is, what it is used for etc.). Equipping a robot *a priori* with a (necessarily closed) database of object knowledge is problematic because the system designer must predict which subset of all the different domain objects is required, and then build all of these models (a time-consuming task). If a new object appears in the environment, or an unmodelled object becomes important to a task, the robot will be unable to perceive, or reason about, it. The solution to this problem is for the robot to *learn on-line about previously unknown objects*. This allows robots to autonomously extend their knowledge of the environment, training new models from their own experiences and observations.

The online learning of perceptual and semantic object models is a major challenge for the integration of robotics and AI. In this paper we address one problem from this larger challenge: given an observation of a scene containing an unknown object, can an autonomous system predict the semantic description of this object. This is an important problem because online-learnt object models ([5]) must be integrated into the robot's existing knowledge base, and a structured, semantic description of the object is crucial to this. Our solution combines semantic descriptions of perceived scenes containing unknown objects, with a distributional semantic approach which allows us to fill gaps in the scene descriptions by mining knowledge from the Semantic Web. Our approach assumes that the knowledge onboard the robot is a subset of some larger knowledge base, i.e. that the object is not unknown beyond the robot's pre-configured knowledge. To determine which concepts from this larger knowledge base might apply to the unknown object, our approach exploits the spatio-temporal context in which objects appear, e.g. a teacup is often found next to a teapot and sugar bowl. These spatio-temporal co-occurrences provide contextual clues to the properties and identity of otherwise unknown objects.

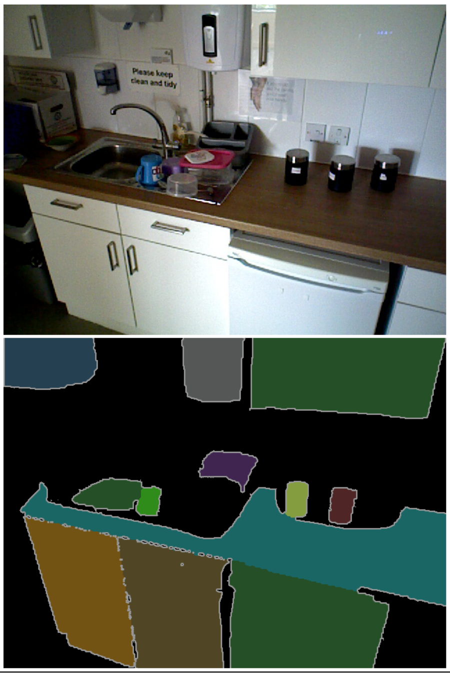
This paper makes the following contributions:

- a novel distributional semantics-based approach for predicting both the semantic identity of an unknown, everyday object based on its spatial context and its most likely location based on semantic relatedness;
- an extension to an existing semantic perception architecture to provide this spatial context; and
- an evaluation of these techniques on real-world scenes gathered from a long-term autonomous robot deployment.

The remainder of the paper is structured as follows. In Section 2, we first state the problem of acquiring semantic descriptions for unknown objects and give an overview of our approach. We then discuss related work in Section 3. In Section 4, we describe the underlying robot perception system and explain how it is integrated with a Semantic Web mining component. Section 5 describes how the component generates answers/hypotheses to web-queries from the perception module. In Section 6, we describe the experimental setup and present the results. Before we conclude in Section 8, we provide a detailed discussion about our approach in Section 7. We also make available our data set and software source code for the benefit of the community at: http://github.com/aloof-project/

[1] Intelligent Robotics Lab, School of Computer Science, University of Birmingham, United Kingdom, {j.young, l.kunze, n.a.hawes}@cs.bham.ac.uk

[2] INRIA Sophia Antipolis Méditerranée, Sophia Antipolis, France, {valerio.basile, elena.cabrio}@inria.fr



Room	kitchen
Surface	coutertop
Furniture	refrigerator, kitchen cabinet, sink
Small Objects	bowl, teabox, instant coffee, water boiler, mug

Figure 1. Perceived and interpreted kitchen scene, with various objects.

2 Problem Statement and Methodology

2.1 Problem Statement

The problem we consider in this work can be summarized as follows: *Given the context of a perceived scene and the experience from previous observations, predict the class of an as 'unknown' identified object.* The context of a scene can include information about the types and locations of recognized small objects, furniture, and the type of the room where the observation has been made.

In this paper we use the following running example (Figure 1) to illustrate the problem and our approach:

> While operating 24/7 in an office environment, a robot routinely visits the kitchen and scans all surfaces for objects. On a kitchen counter it finds several household objects: a bowl, a teabox, a box of instant coffee, and a water boiler. However, one of the segmented objects, a mug, cannot be identified as one of the known object classes. The robot's task is to identify the unknown object solely based on the context of the perceived scene and scenes that have been previously perceived and in which the respective object was identified.

The problem of predicting the class of an object purely based on the context can also be seen as *top-down reasoning* or *top-down processing* of information. This stands in contrast to data-driven bottom-up processing where, for example, a robot tries to recognize an object based on its sensor data. In top-down processing, an agent, or the robot, has some expectations of what it will perceive based on commonsense knowledge and its experiences. For example, if a robot sees a fork and a knife close to each other, and a flat unknown object with a square bounding box next to them, it might deduce that the unknown object is probably a plate. In the following, we refer to this kind of processing which combines top-down reasoning and bottom-up perception as *knowledge-enabled perception*.

Systems such as the one described in this paper are key components of integrated, situated AI systems intended for life-long learning and extensibility. We currently develop the system with two main use-cases in mind, both stemming from the system's capability to suggest information about unknown objects based on the spatial context in which they appear. The first use case is as part of a crowdsourcing platform, allowing humans that inhabit the robot's environment to help it label unknown objects. Here, the prediction system is used to narrow down the list of candidate labels and categories to be shown to users to select from alongside images of unknown objects the robot has encountered. Our second use case will be to help form more informative queries for larger machine learning systems, in our case an image classification system trained on extensive, though categorised, image data from websites like Amazon. Here, having some hints as to an object's identity, such as a distribution over a set of possible labels or categories it might belong to or be related to, could produce a significant speed boost by letting the classification system know what objects it does *not* have to test against. In this case, we aim to use the system to help a robot make smarter, more informed queries when asking external systems questions about the world.

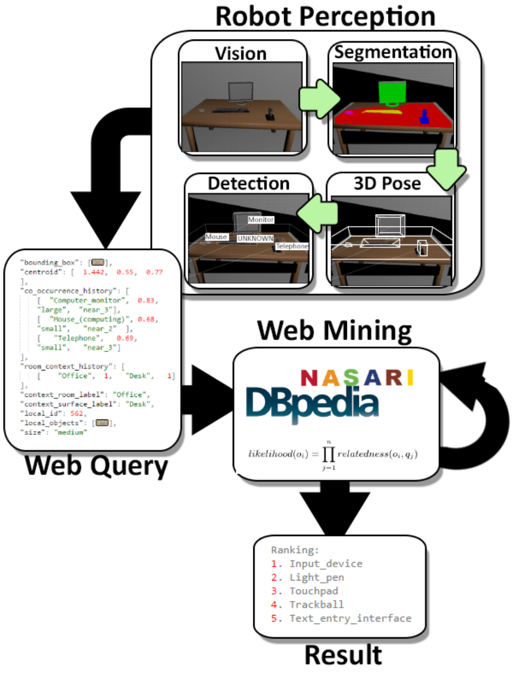
2.2 Our Approach

In this work we address the problem of predicting information about the class of an object based on the perceived scene context by mining the Semantic Web. The extracted scene context includes a list of recognized objects and their spatial relations among each other, plus additional information from a semantic environment map. This information is then used to mine potential object classes based on the semantic relatedness of concepts in the Web. In particular, we use DBpedia as a resource for object knowledge, and will later on use WordNet to investigate object taxonomies. The result of the web mining component is a ranked list of potential objects classes, expressed as DBPedia entries, which allows us access to further information beyond just the class of an object, such as categorical knowledge. An overview of the entire developed system is given in Figure 2.

Overall, we see our context-based class prediction approach as a means to restrict the number of applicable classes for an object. The aim of our knowledge-enabled perception system is not to replace a bottom-up perception system but rather to complement it as an additional *expert*. For example, in the context of a crowdsourcing-based labeling platform our system could generate label suggestions for users. Thereby labeling tasks can be performed in less time and object labels would be more consistent across users. Hence, we belief that our system provides an essential functionality in the context of lifelong object learning.

Before we present related work in Section 3 we briefly discuss various resources of object knowledge.

Resources for object knowledge To provide a common format for object knowledge, and to access the wide variety of structured knowledge available on the Web, we link the observations made by the robot to DBpedia concepts. DBpedia [2] is a crowd-sourced community effort started by the Semantic Web community to extract structured information from Wikipedia and make this information available on the Web. DBpedia has a broad scope of entities covering different domains of human knowledge: it contains more than 4 million things classified in a consistent ontology and denoted by a URI-based



```
"bounding_box": [▭],
"centroid": [  1.442,  0.55,  0.77
],
"co_occurrence_history": [
    [  "Computer_monitor",  0.83,
    "large",  "near_3"],
    [  "Mouse_(computing)", 0.68,
    "small",   "near_2"  ],
    [   "Telephone",   0.69,
    "small",   "near_3"]
],
"room_context_history": [
    [    "Office",  1,    "Desk",   1]
],
"context_room_label": "Office",
"context_surface_label": "Desk",
"local_id": 562,
"local_objects": [▭],
"size": "medium"
```

$$likelihood(o_i) = \prod_{j=1}^{n} relatedness(o_i, q_j)$$

```
Ranking:
1. Input_device
2. Light_pen
3. Touchpad
4. Trackball
5. Text_entry_interface
```

Figure 2. System overview. The robot perception component identifies all object candidates within a scene. All object candidates that can be recognized are labled according their class, all remaining objects are labeled as 'unknown'. Furthermore, the component computes the spatial relations between all objects in the scene. Together with context information from a semantic environment map, the robot generates a query to a web service which is processed by the Semantic Web mining component. Based on the semantic relatedness of objects the component provides a ranked list of the potential classes for all unknown objects.

reference of the form `http://dbpedia.org/page/Teapot` for the Teapot concept. DBpedia supports sophisticated queries (using an SQL-like query language for RDF called SPARQL) to mine relationships and properties associated with Wikipedia resources. We link the objects that the robot can encounter in natural environments to DBpedia concepts, thus exploiting this structured, ontological knowledge.

BabelNet [16] is both a multilingual encyclopedic dictionary and a semantic network which connects concepts and named entities in a very large network of semantic relations (about 14 million entries). BabelNet covers and is obtained from the automatic integration of several resources, such as WordNet [6], Wiktionary and Wikipedia. Each concept contained in BabelNet is represented as a vector in a high-dimensional geometric space in the NASARI resource, that we use to compute the semantic relatedness among objects.

3 Related Work

To obtain information about unknown objects from the Web, a robot can use perceptual or knowledge-based queries. Future systems will inevitably need to use both. In this paper we focus on the knowledge-based approach, but this can be seen as complementary to systems which use image-based queries to search databases of labelled images for similarity, e.g. [17].

Although the online learning of new *visual* object models is currently a niche area in robotics, some approaches do exist [5, 7]. These approaches are capable of segmenting previously unknown objects in a scene and building models to support their future re-recognition. However, this work focuses purely on visual models (what objects look like), and does not address how the learnt objects are described semantically (what objects are).

The RoboSherlock framework [1] (which we build upon, see Section 4.1) is one of the most prominent projects to add semantic descriptions to objects for everyday environments, but the framework must largely be configured *a priori* with knowledge of the objects in its environment. It does support more open ended performance, e.g. through the use of Google Goggles, but does not use spatial or semantic context for its Web queries, only vision. The same research group pioneered Web and cloud robotics, where tools such as KNOWROB [20] (also used in RoboSherlock) both formalised robot knowledge and capabilities, and used this formal structure to exploit the Web for remote data sources and knowledge sharing. In a more supervised setting, many approaches have used humans to train mobile robots about new objects in their environment [9, 19] and robots have also used Web knowledge sources to improve their performance in closed worlds, e.g. the use of object-room co-occurrence data for room categorisation in [10].

The spatial organisation of a robot's environment has also been previously exploited to improve task performance. For example, [21, 12] present a system in which the previous experience of spatial arrangements of desktop objects is used to refine the results of a noisy object categorisation system. This demonstrates the predictive power of spatial arrangements, which is something we also exploit in this paper. However this prior work matched between scenes in the same environment and input modality. In our work we connect spatial arrangements in the robot's situated experience to structured knowledge on the Web.

Our predictions for unknown objects rely on determining the semantic relatedness of terms. This is an important topic in several areas, including data mining, information retrieval and web recommendation. [18] applies ontology-based similarity measures in the robotics domain. Background knowledge about all the objects the robot could encounter, is stored in an extended version of the KNOWROB ontology. Then, WUP similarity [22] is applied to calculate relatedness of the concept types by considering the depth of the concepts and the depth of their lowest common super-concept in the ontology. [14] presents an approach for computing the semantic relatedness of terms using ontological information extracted from DBpedia for a given domain, using the results for music recommendations. Contrary to these approaches, we compute the semantic relatedness between objects by leveraging the vectorial representation of the DBpedia concepts provided by the NASARI resource [3]. This method links back to earlier distributional semantics work (e.g. Latent Semantic Analysis [13]) with the difference that here concepts are represented as vectors, rather than words.

4 Situated Robot Perception

4.1 The RoboSherlock Framework

To be able to detect both known and unknown objects in its environment a robot must have perceptual capabilities. Our perception pipeline is based on the *RoboSherlock framework* [1], an open-source framework for implementing perception systems for robots, geared towards interaction with objects in human environments. The use of RoboSherlock provides us with a suite of vision and perception algorithms. Following the paradigm of Unstructured Information Management (as used by the IBM Watson project), RoboSherlock approaches perception as a problem of content analysis, whereby sensor data is processed by a set of specialised information extraction and processing algorithms called *annotators*. The RoboSherlock perception pipeline is a sequence of annotators which include plane segmentation, RGB-D object segmentation, and object detection algorithms. The output of the pipeline includes 3D point clusters, bounding boxes of segmented objects (as seen in Figure 2), and feature vectors (colour, 3D shape and texture) describing each object. These feature vectors are important as they allow the robot to track unknown objects as it takes multiple views of the same scene. Though in this paper we work with a collected and annotated dataset, we do not require the segmentation or 3D object recognition steps RoboSherlock can provide via LINE-MOD-3D [11], though this component is used in our full Robot and Simulated system where a range of perception algorithms are connected and used instead of dataset input. We make use of all other RoboSherlock capabilities the pipeline to process the data and provide a general architecture for our representation and extraction of historical spatial context, web query generation and the application of Qualitative Spatial Relations, which we will discuss in a following section.

4.2 Scene Perception

In this paper we assume the robot is tasked with observing objects in natural environments. Whilst this is not a service robot task in itself, it is a precursor to many other task-driven capabilities such as object search, manipulation, human-robot interaction etc. Similar to prior work (e.g. [18]) we assume that the robot already has a semantic map of its environment which provides it with at least annotations of supporting surfaces (desks, worktops, shelves etc.), plus the semantic category of the area in which the surface is located (office, kitchen, meeting room etc.). Surfaces and locations are linked to DBpedia entries just as object labels are, typically as entities under the categories `Furniture` and `Room` respectively.

From here, we have access to object, surface and furniture labels described by the data, along with 3D bounding boxes via 3D point data. In the kitchen scene the robot may observe various objects typical of the room, such as a refrigerator, a cabinet, mugs, sugar bowls or coffee tins. Their positions in space relative to a global map frame are recorded and we can then record the distance between objects, estimate their size (volume) and record information about their co-occurrences, and the surfaces upon which they were observed, by updating histograms attached to each object.

In the following we assume that each scene only contains a single unknown object, but the approach generalises to multiple unknown objects treated independently. Joint inference over multiple unknown objects is future work.

4.3 Spatial and Semantic Context Extraction

In order to provide additional information to help subsequent components predict the unknown object, we augment the scene description with additional spatial and semantic *context* information, describing the relationships between the unknown object and the surrounding known objects and furniture. This context starts from the knowledge we already have in the semantic map: labels for the room and surface the object is supported by.

We make use of *Qualitative Spatial Relations* (QSRs) to represent information about objects [8]. QSRs discretise continuous spatial measurements, particularly relational information such as the distance and orientation between points, yielding symbolic representations of ranges of possible continuous values. In this work, we make use of a qualitative distance measure, often called a Ring calculus. When observing an object, we categorise its distance relationship with any other objects in a scene with the following set of symbols: $near_0, near_1, near_2$, where $near_0$ is the closest. This is accomplished by placing sets of thresholds on the distance function between objects, taken from the centroid of the 3D cluster. For example, this allows us to represent that the mug is closer to the spoon than the kettle ($near_0(mug, spoon)$ $near_2(mug, kettle)$) without using floating-point distance values based on noisy and unreliable readings from the robot's sensors. The RoboSherlock framework provides a measure of the qualitative size of objects by thresholding the values associated with the volume of 3D bounding-boxes around objects as they are observed. We categorise objects as $small, medium, large$ in this way, allowing the robot to represent and compare object sizes. Whilst our symbolic abstractions are currently based on manual thresholds, approaches exist for learning parametrisations of QSRs through experience (e.g. [23]) and we will try this in the future. For now, we choose parameters for our qualitative calculi tuned by our own knowledge of objects in the world, and how they might relate. We use $near_0$ for distances in cluster space lower than 0.5, $near_1$ for distances between than 0.5 and 1.0, $near_2$ for distances between 1.0 and 3.5 and $near_3$ for distances greater than 3.5.

As the robot makes subsequent observations, it may re-identify the same unknown object in additional scenes. When this happens we store all the scene descriptions together, providing additional context descriptions for the same object. In Figure 3 we show part of the data structure describing the objects that co-occured with a plate in a kitchen, and their *most common* qualitative spatial relations.

```
"co_occurrences": [
  ["Coffee",  0.5, "near_0" ],
  ["Kitchen_side", 1.0,"near_0"  ],
  ["Kitchen_cabinet", 1.0,"near_1" ],
  ["Fridge", 0.625,"near_1" ],
  ["Teabox", 0.625,"near_0" ],
  ["Waste_container",  0.375,"near_2" ],
  ["Utensil_rack", 0.625, "near_1" ],
  ["Sugar_bowl_(dishware)",  0.625,"near_0" ]
   ,
  ["Electric_water_boiler", 0.875,"near_1" ],
  ["Sink", 0.625, "near_1"   ] ],
   "context_history": [
      ["Kitchen", 1.0, "Kitchen_counter",1 ],
      [   "Office", 0.0,"Desk", 0 ]],
  "context_room_label": "Kitchen",
  "context_surface_label": "Kitchen_counter",
```

Figure 3. An example data fragment taken from a series of observations of a Plate in a series of kitchen scenes, showing object, furniture, room and surface co-occurence

5 Semantic Web Mining

For an unknown object, our aim is to be able to provide a list of likely DBpedia concepts to describe it, and we will later consider and compare the merits and difficulties associated with providing object *labels* and object *categories*. As this knowledge is not available on the robot (the object is *locally* unknown), it must query an external data source to fill this knowledge gap. We therefore use the scene descriptions and spatial contexts for an unknown object to generate a query to a Web service. In return this service provides a list of the possible DBpedia concepts which may describe the unknown object. We expect the robot to use this list in the future to either automatically label a new object model, or to use the list of possible concepts to guide a human through a restricted (rather than open-ended) learning interaction.

The Web service provides access to object- and scene-relevant knowledge extracted from Web sources. It is queried using a JSON structure sent via an HTTP request (shown in Figure 2). This structure aggregates the spatial contexts collected over multiple observations of the unknown object. In our current work we focus on the co-occurrence structure. Each entry in this structure describes an object that was observed with the unknown object, the ratio of observations this object was in, and the spatial relation that most frequently held between the two. The room and surface fields describe where the observations were made.

Upon receiving a query, the service computes the *semantic relatedness* between each object included in the co-occurrence structure and every object in a large set of candidate objects from which possible concepts are drawn from (we discuss the nature of this set in Section 6).

This semantic relatedness is computed by leveraging the vectorial representation of the DBpedia concepts provided by the NASARI resource [3]. In NASARI each concept contained in the multilingual resource BabelNet [16] is represented as a vector in a high-dimensional geometric space. The vector components are computed with the *word2vec* [15] tool, based on the cooccurrence of the mentions of each concept, in this case using Wikipedia as source corpus.

Since the vectors are based on distributional semantic knowledge (based on the *distributional hypothesis*: words that occurr together often are likely semantically related.), vectors that represent related entities end up close in the vector space. We are able to measure such relatedness by computing the inverse of the cosine distance between two vectors. For instance, the NASARI vectors for `Pointing_device` and `Mouse_(computing)` have relatedness 0.98 (on a continuous scale from 0 to 1), while `Mousepad` and `Teabox` are 0.26 related.

The system computes the aggregate of the relatedness of a candidate object to each of the scene objects contained in the query. Using relatedness to score the likely descriptions of an unknown object follows from the intuition that related objects are more likely than unrelated objects to appear in a scene, e.g., to identify a `Teapot` is more useful to know that there is a `Teacup` at the scene rather than a `Desk`.

Formally, given n observed objects in the query $q_1, ..., q_n$, and m candidate objects in the universe under consideration $o_1, ..., o_m \in O$, each o_i is given a score that indicates its likelihood of being the unknown object by aggregating its relatedness across all observed objects. The aggregation function can be as simple as the arithmetic mean of the relatedness scores, or a more complex function. For instance, if the aggregation function is the product, the likelihood of an object o_i is given by:

$$likelihood(o_i) = \prod_{j=1}^{n} relatedness(o_i, q_j)$$

For the sake of this work, we experimented with the product as aggregating function. This way of aggregating similarity scores gives higher weight to highly related pairs, as opposed to the arithmetic mean, where each query object contributes equally to the final score. The idea behind this choice is that if an object is highly related to the target it should be regarded as more informative.

The information carried by each query is richer than just a bare set of object labels. One piece of knowledge that can be exploited to obtain a more accurate prediction is the relative position of the observed objects with respect to the target unknown object. Since this information is represented as a discrete level or proximity (from `near_0` to `near_3`), we can use this as a threshold to determine whether or not an object should be included in relatedness calculation. In this work we discard any object related by `near_3`, based on the intuition that the further away an object is spatially, the less related it is. Section 6.2 includes an empirical investigation into approach.

For clarity, here we present an example of execution of the algorithm described above on the query corresponding to the kitchen example seen throughout the paper. The input to the Web module is a query containing a list of pairs (object, distance): (`Refrigerator`, 3), (`Kitchen_cabinet`, 3), (`Sink`, 3), (`Kitchen_cabinet`, 3), (`Sugar_bowl_(dishware)`, 1), (`Teabox`, 1), (`Instant_coffee`, 2), (`Electric_water_boiler`, 3). For the sake of readability, let us assume a set of candidate objects made only of three elements: `Tea_cosy`, `Pitcher_(container)` and `Mug`. Table 1 show the full matrix of pairwise similarities.

	Tea_cosy	Pitcher_(container)	Mug
Refrigerator	0.473	0.544	0.522
Sink	0.565	0.693	0.621
Sugar_bowl_(dishware)	0.555	0.600	0.627
Teabox	0.781	0.466	0.602
Instant_coffee	0.821	0.575	0.796
Electric_water_boiler	0.503	0.559	0.488
product	**0.048**	**0.034**	**0.047**

Table 1. Object similarity of the three candidates `Tea_cosy`, `Pitcher_(container)` and `Mug` to the objects observed at the example kitchen scene. The last line shows the similarity scores aggregated by product.

Among the three candidates, the one with highest aggregated score is `Tea_cosy`, followed by `Mug` and `Pitcher_(container)`. For reference, the ground truth in the example query is `Mug`, that ended up second in the final ranking returned by the algorithm.

We can also alter the performance of the system using the *frequency* of the objects returned by the query. The notion of frequency, taken from [4], is a measure based on the number of incoming links in the Wikipedia page of an entity. Using this measure we can choose to filter uncommon objects from the results of the query, by thresholding with a given frequency value. In the example above, the frequency counts of `Tea_cosy`, `Pitcher_(container)` and `Mug` are respectively 25, 161 and 108. Setting a threshold anywhere between 25 and 100 would filter `Tea_cosy` out of the result, moving up the ground truth to rank 1. Similarly, we can filter out objects that are too far from the target by imposing a limit on their observed distance. A threshold of 2 (inclusive) for the distance of the objects in the example would exclude `Refrigerator`, `Kitchen_cabinet`, `Sink` and `Electric_water_boiler` from the computation.

Other useful information available from the spatial context includes the label of the room, surface or furniture where the unknown was observed. Unfortunately, in order to leverage such information, one needs a complete knowledge base containing these kind of relations, and such a collection is unavailable at the moment. However, the room and the surface labels are included in the relatedness calculations along with the observed objects.

6 Experiments

In order to evaluate the effectiveness of the method we propose in predicting unknown objects' labels, we perform some experimental tests. In this section we report on the experimental setup and the results we obtained, before discussing them in further detail.

6.1 Experimental Set-up

Our experimental evaluation is an experiment based on a collection of panoramic RGB-D scans taken from an autonomous mobile service robot deployed in a working office for a month. It took these scans at fixed locations according to a flexible schedule. After the deployment we annotated the objects and furniture items in these sweeps, providing each one with a DBpedia concept. This gives us 1329 real world scenes (384 kitchen, 945 office) on which we can test our approach. From this data, our evaluation treats each labeled object in turn as an unknown object in a leave-one-out experiment, querying the Web service with the historical spatial context data for the unknown object similar to that shown in Figure 3.

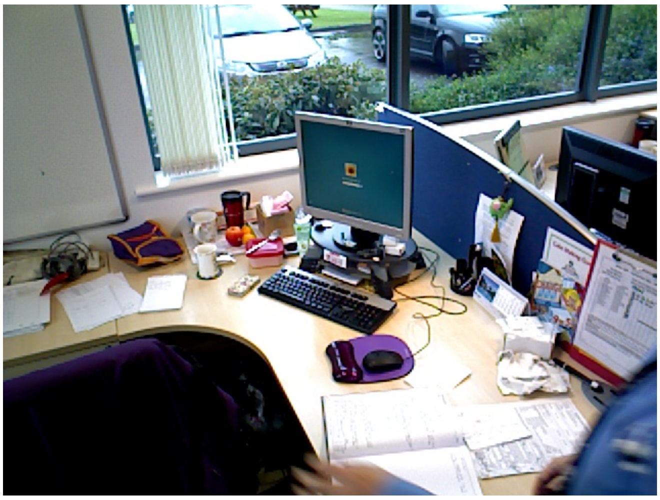


Figure 4. An example office scene as an RGB image from our real-world deployment. Our data contains 945 office scenes, and 384 kitchen scenes.

In all of the experiments we compare the ground truth (known label in the data) to the DBpedia concepts predicted by our system. We measure performance based on two metrics. The first *WUP similarity* measures the semantic similarity between the ground truth and the concept predicted as most likely for the unknown object. The second measure is the *ranking* of the ground truth in the list of suggested concepts.

For the experiments, the set of candidate objects (O in Section 5) was created by adding all concepts from the DBpedia ontology connected to the room types in our data by up to a depth of 3. For example, starting from office leads us to office equipment, computers, stationary etc. This resulted in a set of 1248 possible concepts. We set the frequency threshold to 20, meaning we ignored any suggest concept which had a frequency value lower than this. This means uncommon concepts such as `Chafing_dish` (frequency=13) would always be ignored if suggested, but more common ones such as `Mouse_(computing)` (frequency=1106) would be kept.

6.2 Results

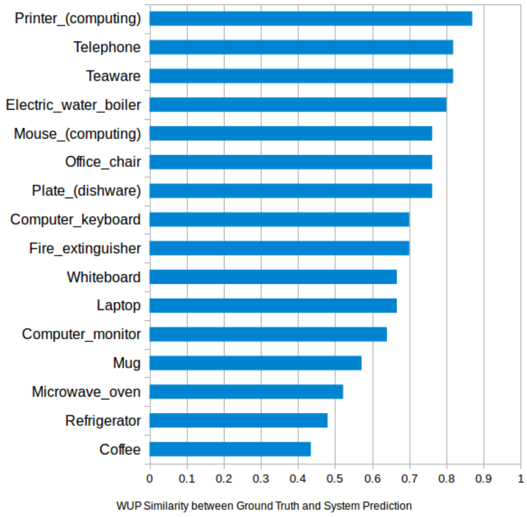


Figure 5. WUP similarity measure between WordNet synsets of ground truth and top-ranked result, with $t = 50$, $p = 2$ using the $prod$ method. Ranks closer to 1 are better. Values closer to 1 indicate similarity, and values closer to 0 indicate dissimilarity.

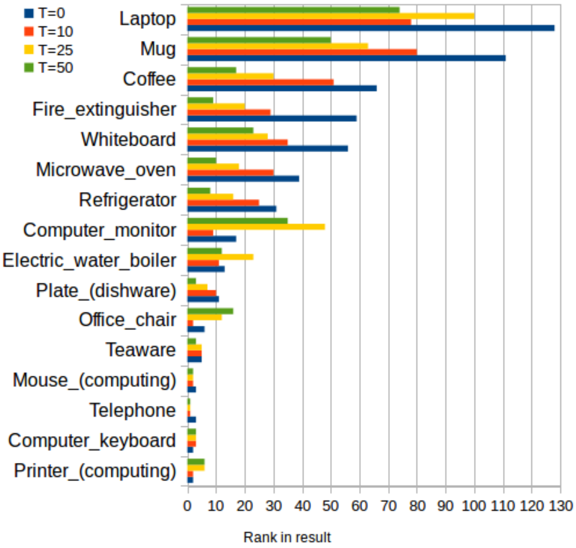


Figure 6. Rank in result by object *category*, matching the highest ranked object with a category shared with the ground truth in the result set, with varying values of the parameter t, with $p = 2$ and the $prod$ method. Ranks closer to 1 are better. Ranking is determined by the position in the result of the first object with an immediate category in common with the ground truth. 56% (9/16) achieve $<= 10$.

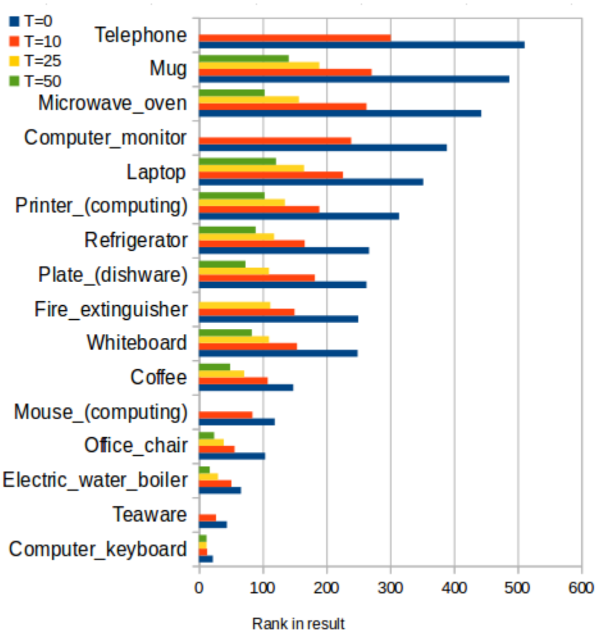


Figure 7. Rank in result by object *label*, matching the label of the ground truth in the result set, with varying values of the parameter t, with $p = 2$ and the *prod* method. Increasing values of T can cause some objects to be excluded from the result set entirely, such as the Teaware or Monitor at T=50

Figure 5 shows the result of calculating the WUP similarity [22] between the WordNet synsets of the ground truth and the top-ranked result from our semantic web-mining system. WUP measures semantic relatedness by considering the depth of two synsets in addition to the depth of their Lowest Common Subsumer (LCS). This means that large leaps between concepts will reduce the eventual similarity score more than small hops might. To do this we used ready available mappings to link DBPedia concepts in our system to WordNet synsets, which are themselves organised as a hierarchy of *is-a* relations. This is in contrast to DBPedia, which is organised as a directed acyclic graph, and while that still means that we could apply the WUP measure to DBPedia nodes directly, WordNet offers a more structured taxonomy of concepts that is more well-suited to this kind of work. This serves to highlight the importance of a multi-modal approach to the use of such ontologies. In the results, the system predicted $Lightpen$ when the ground truth was $Mouse$ producing a WUP score of 0.73, with the LCS being the $Device$ concept, with $Mouse$ and $Lightpen$ having depth 10 and 11 respectively, and $Device$ having depth 8 measured from the root node of $Entity$. In this case, the system suggested an object that fell within 3 concepts of the ground truth, and this is true for the majority of the results in 5. However, in the case of $refrigerator$ as the ground truth, the system suggests $keypad$ as the highest ranked result, producing a WUP score of 0.52. Here, the LCS is at depth 6 with the concept $Artifact$, the ground truth $refrigerator$ is at depth 13 and the prediction $keypad$ is at depth 10. While in this case the node distance between the LCS and the prediction is 4, where in the previous example it was 3, the WUP score is much worse here (0.73 vs 0.52) as there are more large leaps across conceptual space. Our best result in this experiment is for $Printer$ as the ground truth, for which the system suggests $keypad$ again, however the LCS here is the $peripheral$ node at depth 10, where $printer$ is at depth 11 and $keypad$ is at depth 12.

	Mean	Median	Std. Dev	Variance	Range
WUP	0.69	0.70	0.12	0.01	0.43
Category Rank	17.00	9.50	20.17	407.20	73.00
Object Rank	50.93	36.5	50.18	2518.32	141

Figure 8. Statistics on WUP and Rank-in-result data, both for $t = 50, p = 2$ using $prod$

The system suggests a range of objects that are closely related to the unknown object, inferred only from its spatial context and knowledge of the objects and environment around it. From here this allows us to generate a list of candidate concepts which we can use in a second stage of refinement, such as by presentation to a human-in-loop.

Figure 6 shows how frequency thresholding effects the performance of the system. In this experiment we consider the position in the ranked result of the first object with an immediate parent DBPedia category in common with the ground truth. Doing so essentially maps the larger set of object labels to a smaller set of object categories. This is in contrast to considering the position in the result of the specific ground truth label, as shown in Figure 7, and allows us to generate a ranking over *categories of objects*. To ensure categories remain relevant to the situated objects we are interested in, we prune a number DBPedia categories such as those listing objects invented in certain years, or in certain countries. We regard these as being overly broad, and provide a more abstract degree of semantic knowledge about objects than we are interested in. As such, we retrieve the rank-in-result of the first object that shares an immediate DBPedia category with the ground truth, which in the case of $Electric\ water\ boiler$ turns out to be $Samovar$, a kind of Russian water boiler, as both share the immediate ancestor category $Boilers\_(cookware)$. The $Samovar$, and thus the boiler category, appears at rank 12, whereas the specific label $Electric\ water\ boiler$ appears near the end of the result set of 1248 objects, which covers 641 unique DBPedia categories. In our results, categories associated with 9 of the 16 objects (56%) appear within the result's top 10 entries. Here as we filter out uncommon words by increasing the filter threshold T we improve the position of the concept in the list. Whilst this allows us to definitely remove very unlikely answers that appear related due to some quirk of the data, the more we also start to reduce the ability of the robot to learn about certain objects. This is discussed further in Section 7.

Unlike WordNet synsets and concepts, DBPedia categories are more loosely defined and structured, being generated from WikiPedia, but this means they are typically richer in the kind of semantic detail and broad knowledge representation that may be more suitable for presentation to humans, or more easily mapped to human-authored domains. While WordNet affords us access to a well-defined hierarchy of concepts, categories like *device* and *container* are fairly broad, whereas DBPedia categories such as $Video\_game\_control\_methods$ or $Kitchenware$ describe a smaller set of potential objects, but may be more semantically meaningful when presented to humans.

7 Discussion

Overall, whilst the results of our object category prediction system show that it is possible for this novel system to generate some good predictions, the performance is variable across objects. There are a number of factors that influence performance, and lead to this variability. The first issue is that the current system does not rule out suggestions of things it already knows. For example if the unknown object is a keyboard, the spatial context and relatedness may result in a top suggestion of a mouse, but as the system already knows about that, it is probably a less useful suggestion. However, it is possible that the unknown object could be a mouse, but has not been recognised correctly. Perhaps the most fundamental issue in the challenge of predicting objects concepts from limited information is how the limit the scope of suggestions. In our system we restricted ourselves to 1248 possible concepts, automatically selected from DBpedia by ontological connectivity. This is clearly a tiny fraction of all the possible objects in existence. On one hand this means that our autonomous robot will potentially be quite limited in what it can learn about. On the other hand, a large number of this restricted set of objects still make for highly unlikely suggestions. One reason for this is the corpus-based automatically-extracted nature of DBpedia, which means that it includes interesting objects which may never be observed by a robot (e.g. `Mangle_(machine)`). More interestingly though is the effect that the structure of the ontology has on the nature of suggestions. In this work we have been using hierarchical knowledge to unpin our space of hypotheses (i.e. the wider world our robot is placed within), but have not addressed this within our system. This leads to a mismatch between our expectations and the performance of the system with respect to arbitrary precision. For example, if the robot sees a joystick as an unknown object, an appropriate DBpedia concept would seem (to us) to be `Controller_(computing)` or `Joystick`. However, much more specific concepts such as `Thrustmaster` and `Logitech_Thunderpad_Digital` are also available to the system in its current form. When learning about an object for the first time, it seems much more useful for the robot to receive a suggestion of the former kind (allowing it to later refine its knowledge to locally observable instances) than the latter (which unlikely to match the environment of the robot). Instead, returning the *category* of the ranked objects our system suggests allows us to go some way towards this as shown in Figure 6, but still provides us a range of possible candidate categories – though narrowed down from 641 possible categories, to in some cases less than 5. As such, from here we can switch to a secondary level of labelling: that of a human-in-loop. We will next integrate the suggestion system with a crowd-sourcing platform, allowing humans that inhabit the robot's environment to help it label unknown objects. The suggestion system will be used to narrow down the list of candidate categories that will be shown to users as they provide labels for images of objects the robot has seen and learned, but has not yet labelled. While further work is necessary to refine the current 56% of objects that have a category in the top-10 ranked result, we expect that the current results will be sufficient enough to allow a human to pick a good label when provided a brief list of candidates and shown images of the unknown objects. Such systems are crucial for life-long situated learning for mobile robot platforms, and will allow robot systems to extend their world models over time, and learn new objects and patterns.

The issue of how to select which set of possible objects to draw suggestions from is at the heart of the challenge of this work: make the set too large and it is hard to get good, accurate suggestions, but make it too small and you risk ruling out objects that your robot may need to know about. Whilst the use of frequency-based filtering improved our results by removing low-frequency outliers, more semantically-aware approaches may be necessary to improve things further. Further improvements can be made, for instance we largely do not use current instance observations about the object, but prefer its historical context when forming queries. This may be the wrong thing to do in some cases, in fact it may be preferable to weight observations of object context based on their recency. The difference between historical context and the context of an object in a particular instance may provide important contextual clues, and allow us to perform other tasks such as anomaly detection or boost the speed of object search tasks.

One issue we believe our work highlights is the need to integrate a multi-modal approach to the use of differing corpora and ontologies. For instance, the more formal WordNet hierarchy was used to calculate the semantic relatedness of our experiment results, rather than the less formal DBPedia ontology. However we hold that the DBPedia category relationships are more useful in the human-facing component of our system. There exist other ontologies such as YAGO which integrates both WordNet and DBPedia, along with its own category system, that will certainly be of interest to us in the future as we seek to improve the performance of our system. One of our primary goals is to better exploit the hierarchical nature of these ontologies to provide a way of retrieving richer categorical information about objects. While reliably predicting the specific object label from spatial context alone is difficult, we *can* provide higher-level ancestor categories that could be used to spur further learning or improve previous results. As such, we view the prediction process as one of matching the characteristics of a series of increasingly more specific categories to the characteristics of an unknown object, rather than immediately attempting to match the specific lowest-level characteristics and produce the direct object label. This requires an ontology both formally-defined enough to express a meaningful hierarchy of categories for each item, *and* broad enough to provide us mapping to a large set of common-sense categories and objects. It is not clear yet which combination of existing tools will provide the best route to accomplishing this.

8 Conclusions

In this paper we presented an integrated system for solving a novel problem: the suggestion of concept labels for unknown objects observed by a mobile robot. Our system stores the spatial contexts in which objects are observed and uses these to query a Web-based suggestion system to receive a list of possible concepts that could apply to the unknown object. These suggestions are based on the relatedness of the objects observed with the unknown object, and can be improved by filtering the results based on both frequency and spatial proximity. We evaluated our system data from real office observations and demonstrated how various filter parameters changed the match of the results to ground truth data.

We showed that the suggestion systems provides object label suggestions with a reasonably high degree of semantic similarity, as measured by WUP similarity on WordNet synsets. We also achieved success in retrieving the *categories* of objects, rather than their direct labels. In the future we will explore the hierarchical nature of the knowledge used for object concept suggestions, explore different corpora and ontologies to base the suggesting system on, and perform a situated evaluation of our system on a mobile robot with additional perceptual learning capabilities, and crowd-sourcing functionality to label objects on-line with the help of humans using the suggestion system.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 600623, STRANDS, and under the ALOOF project (CHIST-ERA program).

REFERENCES

[1] Michael Beetz, Ferenc Bálint-Benczédi, Nico Blodow, Daniel Nyga, Thiemo Wiedemeyer, and Zoltán-Csaba Marton, 'Robosherlock: Unstructured information processing for robot perception', in *ICRA*, (2015).

[2] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann, 'DBpedia - a crystallization point for the web of data', *Web Semant.*, **7**(3), 154–165, (September 2009).

[3] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli, 'Nasari: a novel approach to a semantically-aware representation of items.', in *HLT-NAACL*, eds., Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, pp. 567–577. The Association for Computational Linguistics, (2015).

[4] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes, 'Improving efficiency and accuracy in multilingual entity extraction', in *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, (2013).

[5] T. Faeulhammer, R. Ambrus, C. Burbridge, M. Zillich, J. Folkesson, N. Hawes, P. Jensfelt, and M. Vincze, 'Autonomous learning of object models on a mobile robot', *IEEE RAL*, **PP**(99), 1–1, (2016).

[6] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, Bradford Books, 1998.

[7] Ross Finman, Thomas Whelan, Michael Kaess, and John J Leonard, 'Toward lifelong object segmentation from change detection in dense rgb-d maps', in *ECMR*. IEEE, (2013).

[8] L. Frommberger and D. Wolter, 'Structural knowledge transfer by spatial abstraction for reinforcement learning agents', *Adaptive Behavior*, **18**(6), 507–525, (December 2010).

[9] Guglielmo Gemignani, Roberto Capobianco, Emanuele Bastianelli, Domenico Bloisi, Luca Iocchi, and Daniele Nardi, 'Living with robots: Interactive environmental knowledge acquisition', *Robotics and Autonomous Systems*, (2016).

[10] Marc Hanheide, Charles Gretton, Richard Dearden, Nick Hawes, Jeremy L. Wyatt, Andrzej Pronobis, Alper Aydemir, Moritz Göbelbecker, and Hendrik Zender, 'Exploiting probabilistic knowledge under uncertain sensing for efficient robot behaviour', in *IJCAI'11*, Barcelona, Spain, (July 2011).

[11] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, 'Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes', in *IEEE ICCV*, (2011).

[12] Lars Kunze, Chris Burbridge, Marina Alberti, Akshaya Tippur, John Folkesson, Patric Jensfelt, and Nick Hawes, 'Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding', in *IEEE IROS*, Chicago, Illinois, US, (September, 14–18 2014).

[13] Thomas K Landauer and Susan T. Dutnais, 'A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge', *PSYCHOLOGICAL REVIEW*, **104**(2), 211–240, (1997).

[14] José Paulo Leal, Vânia Rodrigues, and Ricardo Queirós, 'Computing Semantic Relatedness using DBPedia', in *1st Symposium on Languages, Applications and Technologies*, eds., Alberto Simões, Ricardo Queirós, and Daniela da Cruz, volume 21 of *OpenAccess Series in Informatics (OASIcs)*, pp. 133–147, Dagstuhl, Germany, (2012). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*, (2013).

[16] Roberto Navigli and Simone Paolo Ponzetto, 'Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network', *Artificial Intelligence*, **193**(0), 217 – 250, (2012).

[17] J Philbin, 'Lost in quantization: Improving particular object retrieval in large scale image databases', in *CVPR 2008*, pp. 1–8, (June 2008).

[18] M.J. Schuster, D. Jain, M. Tenorth, and M. Beetz, 'Learning organizational principles in human environments', in *ICRA*, pp. 3867–3874, (May 2012).

[19] Shuran Song, Linguang Zhang, and Jianxiong Xiao, 'Robot in a room: Toward perfect object recognition in closed environments', *CoRR*, **abs/1507.02703**, (2015).

[20] Moritz Tenorth, Lars Kunze, Dominik Jain, and Michael Beetz, 'KNOWROB-MAP – knowledge-linked semantic object maps', in *IEEE-RAS ICHR*, pp. 430–435, Nashville, TN, USA, (December 6-8 2010).

[21] Akshaya Thippur, Chris Burbridge, Lars Kunze, Marina Alberti, John Folkesson, Patric Jensfelt, and Nick Hawes, 'A comparison of qualitative and metric spatial relation models for scene understanding', in *AAAI'15*, (January 2015).

[22] Zhibiao Wu and Martha Palmer, 'Verbs semantics and lexical selection', in *ACL*, ACL '94, pp. 133–138, Stroudsburg, PA, USA, (1994). Association for Computational Linguistics.

[23] Jay Young and Nick Hawes, 'Learning by observation using qualitative spatial relations', in *AAMAS 2015*, (May 2015).