A review on robotic cognitive capabilities

Past, present, and future of mapping cognition to humanoid robots

Author Information

Name: Pargorn *Family Name:* Puttapirat *Nickname:* March
Student ID: 3117999011
School of Electronic and Information Engineering
Department of Computer Science and Technology
Email: pargorn@stu.xjtu.edu.cn / pargornp@gmail.com

Abstract

This report reviews key concepts and limitation to the development of cognitive ability to a humanoid robotics. We discuss about incorporation of new knowledge to existing one, creation and storage of commonsense knowledge, and the creation of information schema that would be able to accommodate complex information such as perceived concepts, relationship between entities, and interaction with objects. All of this is to accomplish more meaningful interaction with human, the environment and ultimately construct self-awareness for the robot therefore it will have the ability to act or perform properly in the real-world situation.

Keywords: imitate, perception, sensory, awareness, cognition, mapping,

Introduction

Preface

The development of humanoid robots has been started for a long time as well as cognitive science and engineering field. However, the idea of crashing the two concepts together has been recently developed by the popularization and the need of interdisciplinary studies/expertise. Therefore, in this report, I will discuss how computational cognitive science and engineering (CCSE) could benefit the perception of the robots especially the humanoid robots. Human sensory system is a complex system where the sensors (e.g. eyes for visual, ears auditory sensory, and so on) and the processor (or the brain) has entwined into different system and then develop and work collectively to perceive the world around each individual. While robotic sensory system has many potentials to exceed human ability to perceive the world around them with higher sensor sensitivity, resolution, accuracy, etc. and arguably/increasingly more processing power. Preliminary conclusion could be drawn here that something may have gone wrong in the usage/utilization of acquired data by different sensors equipped with the robots. This report will also take a look into how CCSE studies may help researchers overcome these shortcomings.

Related concepts

In robotics, there are wide range of application and purpose of developing a robot. In this report we will focus on humanoid robot which is generally a robot with its body shape built to resemble the human body. The purpose of such robot is to ultimately interact with real human or the environment where human being lives. With the mentioned obligation, this specific type of robot needs to excel in tasks that human can perform effortlessly which is currently very hard and complex for robots. These tasks could be grouped into two including interacting with human or human-robot interaction (HRI) and interacting with the human friendly environment like in a normal household.

In HRI, the recent challenge is not at the part where robot can achieve thanks to several years of research that have laid the foundation on how to perform each specific task. The hard part is when the robot can interact with human and cause miscommunication. Mistakes can happen all the time and robot could also learn from them. However, it is hard for the robot to tell right from wrong in this situation. This bring the research community of move on to focus on the ability to have "awareness of the mistakes" and "self-correction". Such ability requires higher level of framework, and this is where robotic cognition could solve the problem. For the robot to possess cognition, it needs perception (meaningful interpretation of input from its sensory system), data model (to store and preserve the meaning of the situation, objects, or relationships), and new actuation system.

Conventionally, the robot would rely on recognition, identification, and tracking (RIT) to perform its functions. The new framework to host the cognitive capability would need the rearrangement of existing technique and algorithm to achieve the more complex function. In this report we will discuss two main topic which is the challenges in cognitive capability development and the data modelling method to accommodate the more complex and meaningful information.

In robotic development there could be domain specific robot vs robot with general knowledge. The domain specific robot could include very wide variety of robot, for example, self-driving car (no human interaction), flying drones, and industrial robotic arm in production factory. From the example we can observe that they are programmed to perform specific task which does not require cognition. Humanoid robot tends to be the type with general knowledge. To perform task such as HRI, the robot would need to have general knowledge about the world or at least have access to the information source. This is the correct kind of robot to possess cognitive ability.

In data science, there are efforts to build a structured dataset which is publicly available. These sources of information are often a community created, for example, Cyc by Cycrop, WIKIDATA (Vrandečić & Krötzsch, 2014), and WordNet (Miller, 1995). The availability of these kind of knowledge is a good foundation for the robot to develop its own knowledge upon. Not only the data itself which has been published, the structuring standard is also made such as W3C web ontology guideline (OWL) and resource description framework (RDF). Later, researchers could take the database and build on it to achieve robot cognition.

Challenges of developing robotic cognitive capability

There are several challenges in developing the robot that would have cognitive ability comparable to human. The cognitive capability goes beyond the recognition, identification, and tracking which the more recent computer algorithms have been developed to tackle such problems. While the mentioned capabilities have provided the robot with very wide range of abilities therefore it is helpful and useful. It is also a good foundation for the next step to come which is establishing robot's cognition and understandings of the situation. Current robots especially humanoid robots all have very limited functions compare to what a real human could do. Therefore, it is appropriate to start exploring the limitation and find a solution to create a better robot which can perform well in real-world situations. Some of the limitations are inherit from the data scheme which is defined by conventional means of operation (for example, to RIT tasks), and the other limitations occur from lack of information schema to accommodate the concept. Conventionally, the complex concept such as hot, safe, and beautiful are packed into one keyword or term. This limits it interaction with other proper objects and functions because the concept has been seen as the same type of object in the information schema.

Several limitations will be introduced with some example as follows. The current information schema lack common-sense knowledge (Lemaignan, Ros, Mösenlechner, Alami, & Beetz, 2010) in both declarative -- fire is hot, and procedural -- how to walk upstairs. By

consideration, it can be seen that it is not possible to fit these knowledges into the RIT scheme. To address this problem, it is suggested that new information schema is needed to accommodate such information. The discussion of proposed method throughout the decade will be discussed in the following sections.

Next is about incorporation of newly acquired knowledge to the existing one. The incorporation of the knowledge can be seen from two aspect: First is letting the new perceived information from specific robot (with specific hardware and sensory system) merge into common already structured data which is publicly available such as Wikipedia/WIKIDATA (Vrandečić & Krötzsch, 2014), WordNet (Miller, 1995), or VisualGenome (Krishna et al., 2017) which post challenges about semantic consistencies and technical translation between different information encoding guidelines. In short, when we can make all the robots speak the same language, the developer does not have to create a new data model each time new series of robot is made. Second is to let the robot learn to perceive new knowledge by itself. Since the robots need to understand the situation they are in, they need to know if the specific situation matches the knowledge, which mean they have encountered it before, or it is the new unknown situation. Once this information is acquired, the robots can proceed with appropriate actions to either guess, act on previous similar experience, or ask for help from the humans. After the robot have finished the newly experience situation, the perceived information should be added to the knowledge base for further uses.

At the present, there are no such things as the perfect information schema that can host the semantically-rich and chaos real-world perception and concepts for the robot. Nevertheless, there are efforts to link the concepts together. In (Lemaignan et al., 2010), OpenCyc/ResearchCyc by Cycrop was mentioned to link the semantic concept using so called MicroTheories technique (Nehaniv & Dautenhahn, 2002). Although this was first introduced about a decade ago, the project is still viable and in development until now. Other than using MicroTheories to link the objects, while this is the open research question, the link has been made in the form of solid categorizable link which is specific to some objects, for example person "sit" "on" chair, chicken "hatch" egg, and pilot "fly" air-plane. Categorizing stuff, could be done as shown in Figure 1 . Although this approach could be made to be extendable, it will raise other questions which are how much it can be extended? (the number of words in the dictionary is limited, however new words are being added by the carefully made decision of human committee) who has the authority to add the new kind of link? when the new link is needed? and can the robot establish the new type of link by themselves? The example of good body of links in natural language processing could be seen from WorldNet which requires many iterations and years to develop. Another good example in visual perception is VisualGenome (Krishna et al., 2017) database which provide the link between visually visible objects and has provide some insights to the spatial interaction between them. The sample of structured data in such database is shown in Figure 2. And since there are many sources of information, another related problem is combining them because sources of information that are usually difficult to combine (Lemaignan et al., 2010) such as combining visual perception, geometrical reasoning, common-sense knowledge or human input.

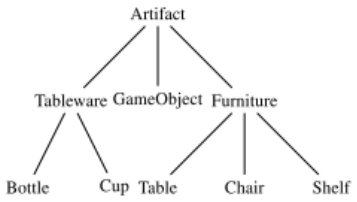

Figure 1 Example of grouping of objects which poses a new question if the number of categories would be ever enough for the real-world application (Lemaignan et al., 2010).

The last open and important challenge I would like to mention in this report is about even registration and attention mechanism. In any events that the robot has recognized, there are likely to be similar events waiting in the knowledge base. With example in human perception, answering homework questions and

closed-book exam questions need a completely different response and behavior while it is the

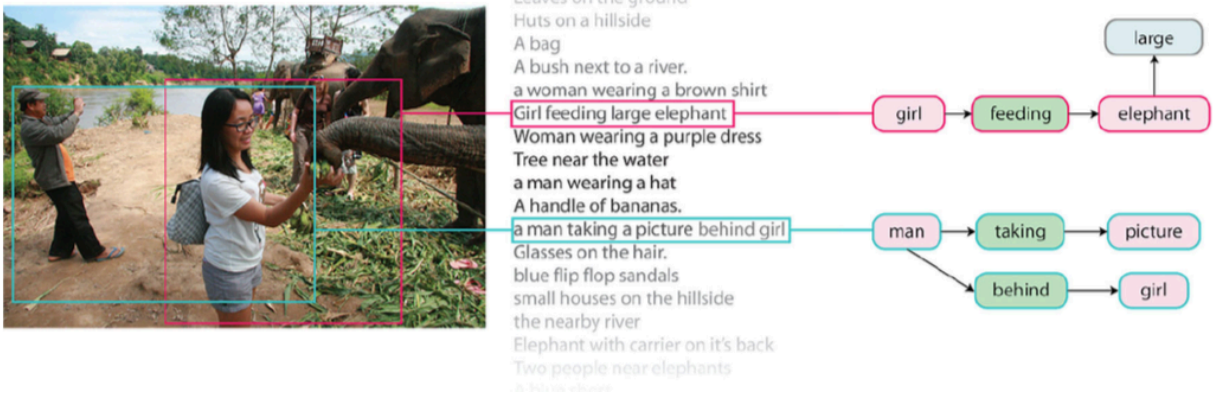

Figure 2 An overview of the data needed to move from perceptual aware- ness to cognitive understanding of images. We present a dataset of images densely annotated with numerous region descriptions, objects, attributes, and relationships. Some examples of region descriptions (e.g. "girl feeding large elephant" and "a man taking a picture behind girl") are shown (Krishna et al., 2017)

same task, question answering. This is the example of similar event which need different behavior. Another example of different events that need same response is the scenario where ones need to stay afloat in the water. If the robot has encountered the situation where it is in the pool and in the open sea, many parameters could show up differently including but not limited to size of the area, amount of water, chemical combination of the water, and existence of waves, all of these parameters could depend upon the robot's sensory system.

It can be seen from the two examples that not all perceived information by the system should be equally processed. It is an open question on What kind of information has to be transferred and what can be omitted? (Taylor, 2015) This brings us to the recent development in machine learning algorithm which is the attention mechanism. It is safe the say that this technique is inspired by the human brain where it focuses on some part of the situation at a given time. The make the central system "pay attention", the information is selectively filtered out depend on several surrounding factors which may came from different sensors or time frame. After the system has receive only the important signal, then it can behave properly. There has been some successful implementation of such mechanism. To build robot with better cognitive ability, this solution could be the key.

The three different data modelling and processing approaches

The beginning of works on questions related to robot's cognitive ability include publication by McCarthy (McCarthy, 2007), Sloman et al. (Sloman, Wyatt, Hawes, Chappell, & Kruijff, 2006) or Levesque and Lakemeyer (Levesque & Lakemeyer, 2008). Most of the challenges of cognitive robotics can be summarized from these articles. In this report, three major milestone of data models will be discussed including the re-identification method which contains layers of conventional RIT tasks, the storage and manipulation of low-level features, and the knowledge base approach.

Re-identification method

The re-identification method would have several intermediate steps which may include RIT tasks along the way between raw data (input from the robot's sensory system) and symbol grounding (the recognition of objects or situation of the robot). In order for this method to work, the features and intermediate steps are carefully

engineered from the researcher's point of view. The human recognizable features are quantized into digital information therefore it become a machine-readable data. The whole process may or may not mimic the thought process of normal human. The example of system that follow natural human behavior would be emotion recognition or more specifically facial expression recognition. As shown in Figure 3 the blue dots represent in the green bounding box are the important landmarks where the system uses to track the location of each component on people's face. These landmarks are effective because their system use the same source of information as normal human eyes would use. Finally, the combination of these data will be turned into recognized facial expression.

Figure 3 Example of facial expression recognition system that rely on RIT tasks. Figure taken from Intelligent Behaviour Understanding Group (iBUG), Department of Computing, Imperial College London (link: https://ibug.doc.ic.ac.uk/research/detection-static-geometric-facial-features/)

Another example that did not follow natural behavior is hand gesture recognition, a good example would be sign language recognition. People with auditory impairment would rely on set of instructions to replace verbal communication. Robot that can understand sign language would need 2D or 3D visual sensor system. Essential components of sign language translation would consist of hand landmark detection and gesture recognition (Poularakis & Katsavounidis, 2014)(Hernandez-Belmonte & Ayala-Ramirez, 2016), elbow and head detection, some tracking algorithm, feature extraction and matching algorithm, and inference engine that has reference to specific language. An example of such workflow by P. Dreuw et al. (Dreuw, Rybach, Deselaers, Zahedi, & Ney, 2007) is shown in Figure 4. It can be seen that these layers of processing mostly rely on conventional RIT tasks.

The advantage of the reidentification method and the reason that makes it popular is that it works well in the past and the framework is clearly explainable even though it is becoming less explainable once deep learning method has come along. After all, the true limitation of this data model is that it is not expandable by the autonomous robot itself. For the robot to achieve awareness or cognition, it needs to be able to perform self-learning and it is the opposite of hand-craft features with hand-craft classification technique such as tuning the hyperparameters in the classification or regression artificial neural networks or SVM model.

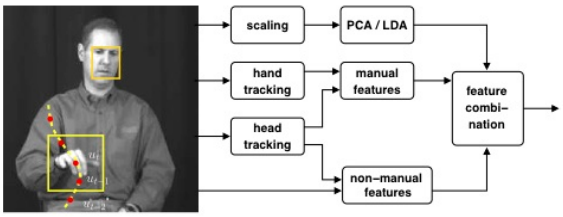

Figure 4 Example of sign language recognition framework (Dreuw et al., 2007)

Storage of low-level facts

The second era of hosting a concept involves the storage of low-level facts. It could also be considered as human-engineered features, however the fluidity of the features themselves are only machine-readable because the features usually contain very high dimension relative to what human could perceive. The example of such features is speeded up robust features (SURF) (Bay, Ess, Tuytelaars, & Van Gool, 2008) and scale-invariant feature transform (SIFT) (Lowe, 2004) descriptor in computer vision (Karami, Prasad, & Shehata, 2015), or Word2Vec (Mikolov, Chen, Corrado, & Dean, n.d.) as in natural language processing. Again, from the

mentioned example, we could see that these features are high-dimension and only machine-readable. Some visualization method could be achieved to see and prove these techniques could accommodate the concept of the object in a proper way as shown in Figure 5 and Figure 6 which shows the visualization of word vector by Word2Vec technique. We can see that similar word or concept can be grouped closer together as a normal human brain would consider it to be.

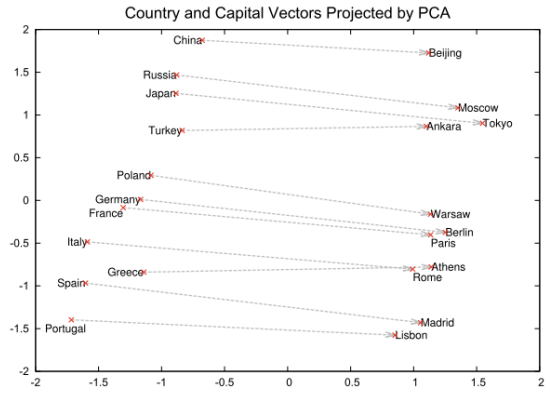

Figure 5 Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means (Mikolov et al., n.d.)

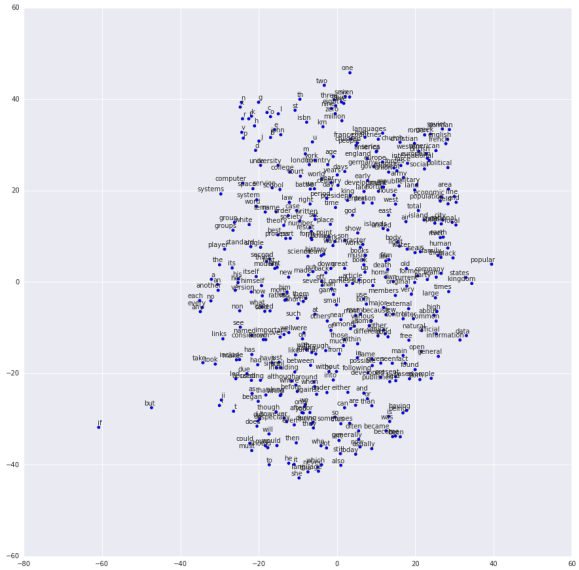

Figure 6 Visualization of word learned embedding using t-SNE. (taken from https://www.tensorflow.org/tutorials/representation/word2vec)

This method has the disadvantages of long-term expandability -- it works now, not sure about the future, because since there is the degree of dimension involve in crafting the feature, it could limit the amount of growing knowledge. To process larger knowledge stored with this kind of method also requires extensively more processing power as well. Finally, the knowledge could not be transferred to different platform which is a common thing to do because of different robot physical limitation, memory and processing power which result in time-constrain problem.

Knowledge ontology with knowledge base approach

There are two main components in this approach including the knowledge which is nicely structured data and the inference engine to drive those knowledges to a good use. In the knowledge base approach there are several publicly available semantically meaningful datasets waiting to be utilized. The examples are mentioned earlier including WordNet, ImageNet (Deng et al., 2009), VisualGenome, etc. The ultimate goal of this section is to discuss how different data model do link the inputs that the robot could acquire via sensory system to infer to correct awareness or perception.

To survey how this information scheme could work, we take a look into different piloting work. In 2009, (Daoutis, Coradeshi, & Loutfi, 2009) try to solve grounded knowledge and common-sense reasoning in their knowledge representation and reasoning system by building their knowledge model directly on the ResearchCyc ontology (including the MicroTheories concept), used in combination with the CYCL language. And, in the meanwhile, (Tenorth & Beetz, 2009) develop KNOWROB which is a knowledge processing framework based on Prolog which is a general-purpose logic programming language associated with artificial intelligence and computational

linguistics. Its underlying storage is based on the W3C web ontology language, derived from OPENCYC. They introduce "the concept of computable relationship" to compute on requesting resource description framework (RDF) triples describing spatial relations between objects, probabilities for certain actions to occur, etc. While computability of the relationship (or the link between entities) enable better scaling or expendability, this prevents on the other hand an efficient use of the reasoner to classify and infer new statements since this generally requires at any time the complete set of statements to be available. The work in this area has shown to us that the well-structured framework could indeed reduce the ambiguity of knowledge base processing. However, the knowledge transfer at this point cannot be transfer automatically by the robot. It needs to be done by human-engineered interface.

Considering all the advantages and disadvantages, the knowledge base approach is the currently most viable solution to store the concept that the robot can perceive and best preserve the true meaning of the situation, awareness, or objects. While building a new knowledge base from the ground up is not an easy job and takes a lot of efforts, thanks to standardized structuring guidelines, researchers can easily build on the existing databases.

Conclusion

In this report we have introduce several concepts about humanoid robots, data modelling, and cognitive ability, then we talk about challenges on how to map cognition to the humanoid robots including lack of information schema to accommodate complex concepts, incorporation of newly acquired knowledge to the existing one, lack of common-sense knowledge, even registration problem, and conclusive implementation of attention mechanism for common and adaptive knowledge. Most of the problem was inherited by the use of conventional processing techniques such as RIT.

The data modelling and processing approach was also discussed. We present three major era of data modelling which include re-identification method which implement RIT tasks to accomplish desired result or procedure, storage of low-level facts which perform better but lacks proper information structuring, and knowledge ontology with knowledge base approach which has a promising future development. Most new datasets (or knowledge base) is constructed according to this information scheme.

Finally, putting cognition into humanoid robot still need a lot more efforts in research and development and there are no complete solution to the problem since the problem is not yet well defined or it may not be defined at all.

References

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359. https://doi.org/10.1016/J.CVIU.2007.09.014

Daoutis, M., Coradeshi, S., & Loutfi, A. (2009). *Grounding commonsense knowledge in intelligent systems*. *Journal of Ambient Intelligence and Smart Environments* (Vol. 1). IOS Press. Retrieved from https://dl.acm.org/citation.cfm?id=2350666

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., & Ney, H. (2007). Speech Recognition Techniques for a Sign Language Recognition System. *Interspeech 2007: 8th Annual Conference of the International Speech Communication Association, Vols 1-4*, 705–708.

Hernandez-Belmonte, U. H., & Ayala-Ramirez, V. (2016). Real-Time Hand Posture Recognition for Human-Robot Interaction Tasks. *Sensors (Basel, Switzerland)*, *16*(1). https://doi.org/10.3390/s16010036

Karami, E., Prasad, S., & Shehata, M. (2015). Image Matching Using SIFT , SURF , BRIEF and ORB : Performance

Comparison for Distorted Images Image Matching Using SIFT , SURF , BRIEF and ORB : Performance Comparison for Distorted Images, (February 2016). https://doi.org/10.13140/RG.2.1.1558.3762

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., … Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-016-0981-7

Lemaignan, S., Ros, R., Mösenlechner, L., Alami, R., & Beetz, M. (2010). ORO, a knowledge management module for cognitive architectures in robotics. *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3548–3553. https://doi.org/10.1109/IROS.2010.5649547

Levesque, H., & Lakemeyer, G. (2008). Chapter 23 Cognitive Robotics. *Foundations of Artificial Intelligence*, *3*(07), 869–886. https://doi.org/10.1016/S1574-6526(07)03023-4

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence*, *171*(18), 1174–1182. https://doi.org/10.1016/J.ARTINT.2007.10.009

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (n.d.). Distributed Representations of Words and Phrases and their Compositionality, 1–9. https://doi.org/10.1162/jmlr.2003.3.4-5.951

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*. https://doi.org/10.1145/219717.219748

Nehaniv, C., & Dautenhahn, K. (2002). The Correspondence Problem. *Imitation in Animals and Artifacts*, 1–40.

Poularakis, S., & Katsavounidis, I. (2014). Finger detection and hand posture recognition based on depth information. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4329–4333). IEEE. https://doi.org/10.1109/ICASSP.2014.6854419

Sloman, A., Wyatt, J., Hawes, N., Chappell, J., & Kruijff, G.-J. (2006). Long term requirements for cognitive robotics. *Cognitive Robotics Papers from the 2006 AAAI Workshop Technical Report WS0603*, (McCarthy), 143–150. Retrieved from http://www.aaai.org/Papers/Workshops/2006/WS-06-03/WS06-03-022.pdf

Taylor, J. M. (2015). Mapping Human Understanding to Robotic Perception. *Procedia Computer Science*, *56*, 514–519. https://doi.org/10.1016/J.PROCS.2015.07.244

Tenorth, M., & Beetz, M. (2009). *KNOWROB - knowledge processing for autonomous personal robots*. *Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems*. IEEE. Retrieved from https://dl.acm.org/citation.cfm?id=1732745

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489