This the accepted version of a book chapter. Please cite it as follows: Reforgiato Recupero, D., Boi, L. (2025). Integrating Action Robot Ontology for Enhanced Human-Robot Interaction: A NAO Robot Case Study. In: Meroño Peñuela, A., et al. The Semantic Web: ESWC 2024 Satellite Events. ESWC 2024. Lecture Notes in Computer Science, vol 15344. Springer, Cham.
https://doi.org/10.1007/978-3-031-78952-6_47

Integrating Action Robot Ontology for Enhanced Human-Robot Interaction: A NAO Robot Case Study

Diego Reforgiato Recupero[0000−0001−8646−6183] and Lorenzo Boi

Department of Mathematics and Computer Science, University of Cagliari. Via Ospedale 72, 09124 Cagliari, Italy. diego.reforgiato@unica.it

Abstract. This paper presents an approach that allows the NAO humanoid robot to respond to a question from a user and gesticulate depending on the text that it is saying. The question might also be an action command spoken by the user that the robot recognizes and executes. A Large Language Model is integrated within the approach to provide the question-answering capabilities. For the action commands, we have used an action robot ontology that we have defined in past work. We have extracted the pertinent classes and individuals and generated a three-word string for each action that is matched semantically with the users text. Moreover, as far as the action commands are concerned, the system can work in two modes: STATELESS and STATEFUL. When in STATEFUL mode, the robot knows its current posture and performs the command only if it is compatible with its current state.

Keywords: Action Robot Ontology · Human-Robot Interaction · Natural Language Processing · Large Language Models

In the realm of robotics and Artificial Intelligence (AI), recent advancements have given rise to a multitude of robot-centric applications. There is a growing conviction that there is a 50% probability of AI surpassing human capabilities across all tasks within 45 years, eventually leading to the automation of all human jobs in 120 years, as noted by authors in [6]. Social robots are rapidly gaining prominence and are now being deployed in various countries, serving diverse purposes. The overarching objective of social robots is to enhance interaction with humans, aiming for more effective and efficient engagements.

On the one hand, Large language models (LLMs) have emerged as transformative tools within the realm of robotics applications, playing a pivotal role in augmenting the capabilities of robotic systems [12]. The integration of LLMs, such as OpenAI's GPT-4, into robotics research and development has opened new avenues for enhanced human-robot interaction, cognitive processing, and autonomous decision-making. One notable application of LLMs in robotics involves natural language understanding, enabling robots to interpret and respond to human commands with unprecedented accuracy [10,11]. This linguistic proficiency facilitates more intuitive and user-friendly interfaces, allowing users to communicate with robots using everyday language. This not only simplifies the

user experience but also broadens the accessibility of robotic technologies to individuals with varying levels of technical expertise.

On the other hand, the advent of humanoid robots, exemplified by models like NAO, has sparked a rapid proliferation across various domains. Robots are increasingly being leveraged for a multitude of tasks, catalyzing a surge in interdisciplinary research aimed at exploring their integration into diverse applications [1,7,3,2,4,8,5]. From the controlled environments of research laboratories to the dynamic landscapes of real-world scenarios, humanoid robots are making their presence felt. Their adaptability and versatility make them invaluable assets in fields as varied as healthcare, education, entertainment, and beyond. Researchers and practitioners alike are actively exploring the potential of these robots to augment human capabilities and improve efficiency across a wide spectrum of tasks.

In this paper, we propose an innovative approach that leverages the question-answering capabilities of LLMs to facilitate dynamic conversations between users and a NAO humanoid robot. The user's response is intelligently parsed into sub-sentences, which are then articulated by the robot. If a given sub-sentence implies an action corresponding to the robot's ontology, the robot seamlessly executes the action. Otherwise, the robot engages in the standard *Animated Say* animation. Additionally, users have the option to issue action commands directly. In such cases, by leveraging the semantic similarity between the sentence embeddings and three-word strings created from the action robot ontology introduced in [9], the robot determines the appropriate action to perform. This results in a fluid and natural interaction between the user and the robot. The scripts developed for the Action Recognition Engine and the Choregraphe script are freely available in a public repository[1]. Additionally, a video showcasing an example of the interaction can be accessed publicly[2].

1 How it works

The architecture of the approach proposed in this paper is illustrated in Figure 1. The NAO robot is situated within the same local area network as a server hosting an Action Recognition engine that we have developed and that queries the elements from the action robot ontology. Efficient communication between the NAO and the server is facilitated through a router. The NAO executes a Choregraphe[3] program that we have designed to perform the following actions. The robot initiates interaction by asking the user to speak. Subsequently, it waits for the user to articulate a response. Once the user speaks, the robot records everything he/she says. The recorded audio is then transmitted to OpenAI Speech-to-

[1] `https://github.com/loriboi/zoraProject`

[2] `https://www.youtube.com/watch?v=hEC9EHhjVe4&feature=youtu.be`. We have edited the video to remove the instances when the robot was waiting for responses from the network. We can provide a link to the unedited video, which contains all the original footage.

[3] `https://www.robotlab.com/choregraphe-download-resources`

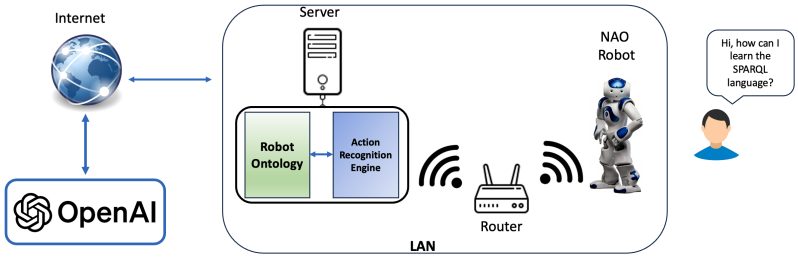

Fig. 1. Architecture of the proposed approach.

text capability[4] for speech-to-text conversion. OpenAI Speech-to-text promptly returns the corresponding text to the robot based on the recorded audio. Following this, the robot sends the obtained text to the Action Recognition engine on the server. The robot then waits for the output from the Action Recognition engine. If the output from the engine is identified as a command action, then the robot proceeds to execute it. However, if the text is not recognized as an action command, the robot forwards it to OpenAI ChatGPT to generate an appropriate response through its question-answering capabilities. The robot retrieves the response from OpenAI ChatGPT and promptly forwards it back to the Action Recognition engine for further processing. Subsequently, the robot receives a dictionary from the Action Recognition engine. Each entry in the dictionary contains a pair: each sub-sentence extracted from the response generated by ChatGPT and the associated action to be performed.

The Action Recognition engine, operating on the server, undertakes the following tasks. Initially, it awaits a text sent from the robot. Upon receipt, it computes the semantic similarity between the received text and all possible actions defined in the action command ontology we have defined[5]. Specifically, the **text representing each action** is transformed into embeddings using the *bert-base-nli-mean-tokens* Sentence Transformer[6]. Subsequently, the semantic similarities between the text and all actions are sorted in decreasing order. If the first element in the list is higher than an empirically determined fixed threshold of 0.8, the engine communicates the identified action (corresponding to the first element in the list) back to the robot for execution. However, if no match is found with any action of the ontology, the engine returns a specific code, signaling the absence of a match, and awaits further text from the robot. Following this, the Action Recognition engine retrieves the text from the robot, which will return a response to the user's input obtained from ChatGPT, and divides the response into sub-sentences. Constructing a dictionary, each entry corresponds to a sub-sentence and is associated with the closest action based on semantic similarity.

[4] `https://platform.openai.com/docs/guides/speech-to-text`
[5] `https://github.com/Fspiga13/Humanoid-Robot-Obeys-Human-Action-Commands-through-a-Robot-Action-Ontology`
[6] `https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens`

If, for a certain entry, the semantic similarity falls below the threshold, the action linked to the underlying sub-sentence will be the *Animated Say* of NAO. This process highlights the engine's intricate semantic analysis, action determination, and response generation capabilities within the proposed framework.

To define the **text representing each action** previously mentioned, we analyzed the action command ontology referenced earlier. Initially, we identified the `BodyPartWord` individuals along with their corresponding `keywords` and `synonyms`. For instance, `arm` is a `BodyPartWord` individual with `keywords` and `synonyms` such as *arm, appendage, bicep, forearm, forelimb*. Next, we retrieved all classes that are subclasses of `BaseAction` and `SimpleAction`, including `ArmAction`, `HandAction`, `HeadAction`, `LegAction`, `Walk` and `Posture`. Subsequently, we extracted individuals from these classes, such as `LegDown, LegUp, HeadDown, HeadForward, HeadLeft, HeadRight, HeadUp, ArmDown, ArmForward, ArmSide, ArmUp, Crouch, LyingBack, LyingBelly, WalkBackward`, etc. Let this set be denoted as S. To collect the objects of the action and formulate a string representation, for each $s \in S$, we extracted elements $w \in W$ and $b \in B$ such that $\{s,$ `involves`, $t\}$ and $\{t,$ `uses`, $w\}$ and $\{t,$ `bodySide`, $b\}$ and $\{w,$ `is_a`, `BodyPartWord`$\}$. Also, to gather other elements of the action, for each $s \in S$, we collected all $a \in A$ such that $\{s,$ `uses`, $a\}$ and $\{a,$ `is_a`, `ActionWord`$\}$. Finally, we obtained the `keyword` and `synonym` values of each element from W, B and A. This process allowed us to formulate all possible combinations for each potential action recognized by the ontology, using the body parts W, the side of the body B, and the remaining action words A.

For instance, considering the `ArmDown` instance, we would derive the sets $W = \{arm, bicep, forearm, hand, claw, paw, etc.\}$, $B = \{left, right\}$, and $A = \{down, drop, lower, etc.\}$ Each combination of values extracted from these three sets would generate a three-word string corresponding to the `ArmDown` instance. The combination with the highest similarity to the user's text is then retrieved: it represents the robot's action corresponding to the user's input.

Additionally, concerning the action commands, the system operates in two modes: STATELESS and STATEFUL. In STATELESS mode, the robot executes each human expression correctly interpreted as an action command, and then reverts to its default posture. In STATEFUL mode, the robot is aware of its current posture and executes a command only if it is compatible with its existing state. In this mode, the robot does not return to its default posture. A sequence of action commands can be given to the robot. For instance, in STATEFUL mode, the user might instruct the robot to stand on its left leg, then ask a question that the robot responds to using ChatGPT and then give one more action to walk. However, the last action will not be performed due to its incompatibility with the current state (left leg raised). The list of incompatibilities has been taken from the defined robot action ontology.

References

1. Alonso, R., Bonini, A., Recupero, D.R., Spano, L.D.: Exploiting virtual reality and the robot operating system to remote-control a humanoid robot. Multim. Tools

Appl. **81**(11), 15565–15592 (2022). https://doi.org/10.1007/S11042-022-12021-Z, https://doi.org/10.1007/s11042-022-12021-z
2. Alonso, R., Concas, E., Recupero, D.R.: A flexible and scalable social robot architecture employing voice assistant technologies. In: Carolis, B.N.D., Gena, C., Lieto, A., Rossi, S., Sciutti, A. (eds.) Proceedings of the Workshop on Adapted intEraction with SociAl Robots, cAESAR 2020, Cagliari, Italy, March 17, 2020. CEUR Workshop Proceedings, vol. 2724, pp. 36–40. CEUR-WS.org (2020), https://ceur-ws.org/Vol-2724/paper10.pdf
3. Atzeni, M., Recupero, D.R.: Multi-domain sentiment analysis with mimicked and polarized word embeddings for human-robot interaction. Future Gener. Comput. Syst. **110**, 984–999 (2020). https://doi.org/10.1016/J.FUTURE.2019.10.012, https://doi.org/10.1016/j.future.2019.10.012
4. Cauli, N., Recupero, D.R.: Video action recognition and prediction architecture for a robotic coach (short paper). In: Consoli, S., Recupero, D.R., Riboni, D. (eds.) Proceedings of the First Workshop on Smart Personal Health Interfaces co-located with 25th International Conference on Intelligent User Interfaces, SmartPhil@IUI 2020, Cagliari, Italy, March 17, 2020. CEUR Workshop Proceedings, vol. 2596, pp. 69–77. CEUR-WS.org (2020), https://ceur-ws.org/Vol-2596/paper6.pdf
5. Gerina, F., Massa, S.M., Moi, F., Recupero, D.R., Riboni, D.: Recognition of cooking activities through air quality sensor data for supporting food journaling. Hum. centric Comput. Inf. Sci. **10**, 27 (2020). https://doi.org/10.1186/S13673-020-00235-9, https://doi.org/10.1186/s13673-020-00235-9
6. Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: Viewpoint: When will ai exceed human performance? evidence from ai experts. Journal of Artificial Intelligence Research **62**, 729–754 (2018)
7. Recupero, D.R.: Technology enhanced learning using humanoid robots. Future Internet **13**(2), 32 (2021). https://doi.org/10.3390/FI13020032, https://doi.org/10.3390/fi13020032
8. Recupero, D.R., Dessì, D., Concas, E.: A flexible and scalable architecture for human-robot interaction. In: Chatzigiannakis, I., de Ruyter, B.E.R., Mavrommati, I. (eds.) Ambient Intelligence - 15th European Conference, AmI 2019, Rome, Italy, November 13-15, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11912, pp. 311–317. Springer (2019). https://doi.org/10.1007/978-3-030-34255-5_21, https://doi.org/10.1007/978-3-030-34255-5_21
9. Recupero, D.R., Spiga, F.: Knowledge acquisition from parsing natural language expressions for humanoid robot action commands. Inf. Process. Manag. **57**(6), 102094 (2020). https://doi.org/10.1016/J.IPM.2019.102094
10. Yoshikawa, N., Skreta, M., Darvish, K., Arellano-Rubach, S., Ji, Z., Bjørn Kristensen, L., Li, A.Z., Zhao, Y., Xu, H., Kuramshin, A., Aspuru-Guzik, A., Shkurti, F., Garg, A.: Large language models for chemistry robotics. Autonomous Robots **47**(8), 1057–1086 (2023). https://doi.org/10.1007/s10514-023-10136-2
11. Zeng, A., Ichter, B., Xia, F., Xiao, T., Sindhwani, V.: Demonstrating large language models on robots. In: Bekris, K.E., Hauser, K., Herbert, S.L., Yu, J. (eds.) Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023 (2023). https://doi.org/10.15607/RSS.2023.XIX.024
12. Zhang, C., Chen, J., Li, J., Peng, Y., Mao, Z.: Large language models for human–robot interaction: A review. Biomimetic Intelligence and Robotics **3**(4), 100131 (2023). https://doi.org/https://doi.org/10.1016/j.birob.2023.100131