A Flexible and Scalable Architecture for Human-Robot Interaction

Diego Reforgiato Recupero[0000−0001−8646−6183], Danilo Dessì[0000−0003−3843−3285], and Emanuele Concas[0000−0002−9461−1501]

Department of Mathematics and Computer Science, University of Cagliari
{diego.reforgiato, danilo_dessi}@unica.it, e.concas11@studenti.unica.it

Abstract. Recent developments and advancements in several areas of Computer Science such as Semantic Web, Natural Language Understanding, Knowledge Representation, and more in general Artificial Intelligence have enabled to develop automatic and smart systems able to address various challenges and tasks. In this paper, we present a scalable and flexible humanoid robot architecture which employs artificial intelligent technologies and developed on top of the programmable humanoid robot called Zora. The framework is composed by three different modules which enable the interaction between Zora and a human for tasks such as Sentiment Understanding, Question-Answering, and automatic Object Recognition. The framework is flexible and extensible, and can be augmented by other modules. Moreover, the embedded modules we present are general, in the sense that they can be easily enriched by adding training resources for the presented sub-components. The design of each module consists of two components i) a front-end system which is responsible for the interaction with humans, and ii) a back-end component which resides on server side and performs the heavy computation.

Keywords: Human-Robot Interaction · Natural Language Understanding · Semantic Web · Sentiment Analysis · Artificial Intelligence · Zora.

1 Introduction

Nowadays we are assisting to the spread of robots in many fields to perform tasks in the place of humans. For example they may be found for performing hazardous tasks (e.g. the exploration of volcanoes), helping people with pathological health problems [1] and executing a fairly wide range of tasks (e.g., house cleanings, painting, packaging) with very little human intervention [8]. Next robot generations fit into every day human life raising the need to build systems for a simple engagement between robots and humans. Therefore, robots should be able to accurately assess people's actions and requests, and perform tasks accordingly.

In a human-robot interaction, people typically have high expectations about robot abilities in understanding their requests, contents and feelings [13]. For building this kind of robots, novel platforms that give smart and friendly behaviors to robots must be developed so that robots can have a human appearance.

In the past, robots social behaviour has been simulated through remote control systems so that humans that interact with the robots had the best possible interaction. This approach clearly does not work at large scale because it needs manual intervention of humans [3], limiting the spread of this technology in our daily life. Hence, the study of systems to provide autonomous social robots that can directly interact with people has taken hold for providing services in common tasks day by day.

Recent developments and advancements in several areas of Computer Science such as Semantic Web, Natural Language Understanding, and Knowledge Representation, and more in general in the Artificial Intelligence field, have enabled the development of automatic and smart systems able to address various challenges and tasks, providing robots with much better autonomous and social behaviours. Therefore, in this paper we

- introduce an extensible framework which is current composed by three modules for enabling Zora, the humanoid robotic platform we have employed, to have different skills;
- we provide Zora with a Sentiment Understanding engine for capturing the feelings of a person who is interacting with the robot;
- we provide Zora with a Question-Answering engine to engage the robot in a clever dialog with a person;
- we provide Zora with an Object Recognition engine to detect objects.

Each module exploits results of different research problems [2] we have previously addressed through state-of-the-art Deep Learning technologies on various datasets. In our approach we included the best models and resources we were able to obtain, making a first step for a smart framework for Zora.

2 Background

Machine Learning methods have proved their validity in many fields, and particularly interesting developments have been carried out in Natural Language Understanding and Object Detection. The Natural Language Understanding benefit of the use of Semantic Web technologies such as Stanford CoreNLP [11], Framester [6], FRED [7], and Word Embeddings [12] to move the data representation from simple text to concepts, enabling better models to capture the meaning of natural language expressions. Together with recent developments in Deep Learning they all have obtained great results in those tasks that require human-based cognitive abilities. The powerful technologies may be exploited for powering Conversational Agents, developing systems that can interpret and respond to the input of a user in a smart way. Furthermore, Object Recognition has been also improved with Deep Learning based approaches which are nowadays able to build models that can recognize object with high level of accuracy. The idea behind this paper took inspiration from [9] where a framework able to perform semantic interpretation with the NAO robot was proposed. In the study, authors focused mainly on ontologies both for mapping the user speech and to execute actions accordingly. Differently, we have based our framework on Deep Learning based modules and Semantic Web technologies for providing smart services through the robot Zora.

3 The Proposed Framework

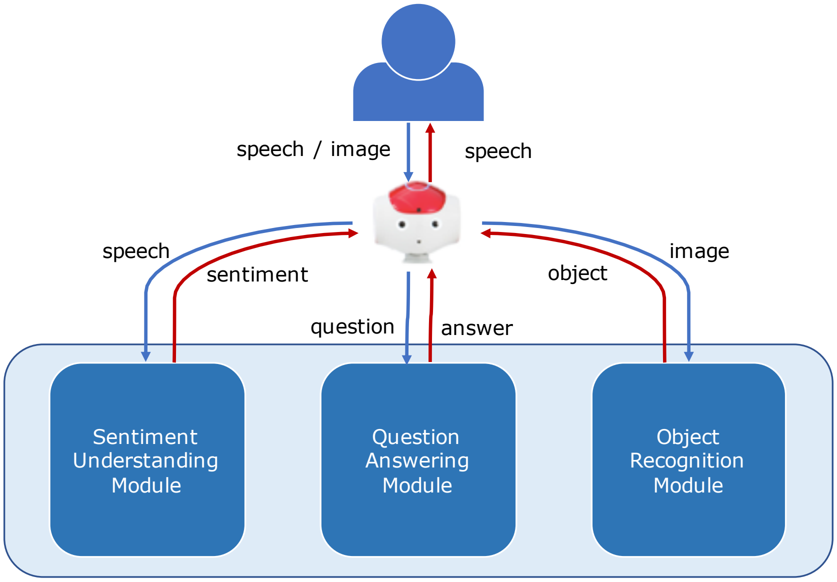
speech / image
speech
speech
sentiment
image
object
question
answer
Sentiment Understanding Module
Question Answering Module
Object Recognition Module

Fig. 1. The schema of the interaction model using Zora.

In this section we describe our framework composed by three modules. Each module is composed by a Choregraphe Component (CC) and a Server Side Support Component (SSSC). CC is very light in terms of required resources (physical space and memory demand), and is directly uploaded and executed by Zora. To note that Zora is a robot which is based on the same infrastructure of NAO, a humanoid robot designed by the Aldebaran Robotics company in France and that presents an intuitive user interface that can be equipped with software packages providing skills to the robot. The SSSC is responsible for the heavy computation on a dedicated server. It runs AI, NLP, Semantic Web and Computer Vision, and any other kind of applications that may be necessary to develop new modules. Each module runs independently from the others, and more than one module at time can be included in the framework. However, a priority order of the modules is set in the framework so that when a user input is received only one module is executed. Furthermore, each module can be activated by the user through the use of *command-words* so that he/she can decide when to start the interaction with the robot which leverages one of the uploaded modules. The current architecture is public accessible[1].

[1] https://github.com/hri-unica

3.1 Sentiment Understanding Module

The Sentiment Understanding [5] Module gives Zora the ability to classify a natural language user's input. It first converts the spoken input of the user in simple text through a speech-to-text internal tool, and then sends the text to the SSSC. The SSSC performs a sentiment polarity detection task adopting a Deep Learning model which classifies the input in one of the following polarity classes: *positive*, *negative*, and *neutral*. A schema of the Deep Learning model is shown in Figure 2(a). The textual data is represented by means of word embeddings which are subsequently fed into the Deep Learning model. The BiLSTM are levels that implement Bidirectional Long-Short Term Memory neural networks that capture patterns of data in both forward and backward direction and, at the same time, can consider long sequences of data by relating the first part of a sentence with its last part. Finally, the model presents an Attention layer which allows to refer to data previously processed instead than forcing the input of the final layer to a single vector. The last layer, *Dense*, combines the result of the whole computation and returns a single output. Afterwards, the result is sent back to the CC which delivers the predicted sentiment to the user. The sentiment model has been trained using the *The Large Movie Review Dataset* [10].

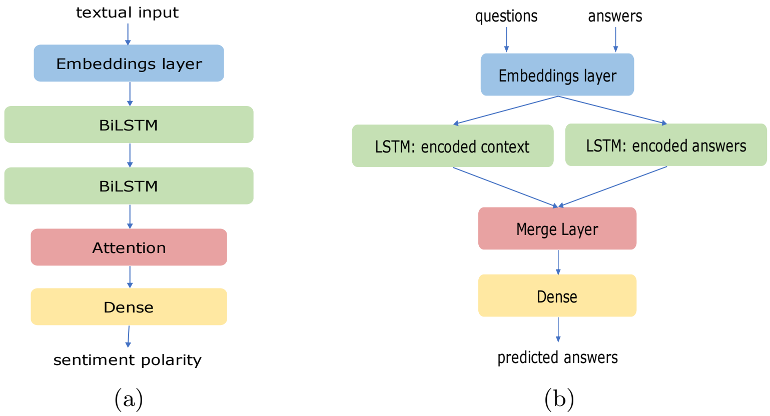

Fig. 2. Deep Learning models of (a) Sentiment Understanding (b) Question-Answering.

3.2 Question-Answering Module

With this module, Zora is able to reply to the user with meaningful responses based on the dialog history with several other users used as training set. The CC interacts with the user through Zora microphones and sends the transcription of his/her voice to the SSSC. The SSSC takes the text, and uses it as an input of a model that has been previous trained. The model is a Deep Learning hierarchy that has been previously trained on the *Cornell Movie Dialogs Corpus* [4] representing the data with Word Embeddings. A schema of the adopted Deep Learning hierarchy is shown in Figure 2(b). It captures syntactical and semantic patterns from a sequence to sequence input defining a context. It uses a LSTM layer to capture the semantic of questions and answers, a Merge layer to bind possible answers to questions, and a Dense layer to predict the best token for the answer. As for the Sentiment Understanding Module, the textual input is represented by word embeddings. The response predicted by the SSSC is sent to the CC which delivers the message to the user.

3.3 Object Recognition Module

For this module, the Zora front camera is used to take a picture of size 640x480. The picture is sent to the SSSC which implements an internal TensorFlow model that is currently able to recognize 91 types of general objects (such as *train*, *person*, *animal*). For the categories *cat* and *dog*, it also recognizes a more fine-grained class (e.g. a dogs breed like *american bulldog*). The SSSC uses two classifiers. One classifier is used to predict the general class. In case the predicted class is one between *dog* or *cat*, it uses a second classifier to further recognize the breed. The general object recognition model has been trained with the Microsoft COCO dataset[2], while the Oxford- IIIT-Pet dataset[3] has been employed for the breed recognition model training. When the elaboration of data has ended by the SSSC and the result has been sent to the CC, the CC tells the user which objects have been recognized. If no objects are recognized, the robot asks the user which objects have been shown and user's annotations are sent to the SSSC which stores the new data. When the interaction is interrupted by the user, the SSSC automatically performs a new training of the model integrating the updated annotated sets with the new data.

4 Conclusions and Future Work

In this paper we have introduced a smart framework which leverages our research in various domains to equip Zora, a completely programmable and autonomous humanoid robot built on top of NAO, with smart human-based skills. We have described three modules which exploit Deep Learning models and Semantic Web technologies to manipulate the input of a user enabling Zora to respond accordingly. The modules allow Zora performing sentiment analysis, generating automatic answers to user's natural language query as in a real dialog, and detecting objects. All modules are based on Deep Learning technologies which are executed on server-side so that the robot is free from the heavy computation. Summing up, each module has (i) a CC which runs on Zora allowing the interaction with the user, and manages inputs and outputs, and (ii) a SSSC which trains Deep

[2] http://cocodataset.org

[3] https://www.robots.ox.ac.uk/˜vgg/data/pets/

Learning model and performs a prediction task on the user's input. As future works, we mainly aim at integrating the framework with modules which allow Zora performing body actions based on the input of the user and interacting with home automation devices such as Google Home. Moreover, we plan to perform different tests to prove the effectiveness and usefulness of the proposed framework.

References

1. Asprino, L., Gangemi, A., Nuzzolese, A.G., Presutti, V., Recupero, D.R., Russo, A.: Ontology-Based Knowledge Management for Comprehensive Geriatric Assessment and Reminiscence Therapy on Social Robots, pp. 173–193. Springer, Cham (2019)
2. Atzeni, M., Reforgiato Recupero, D.: Deep learning and sentiment analysis for human-robot interaction. In: Gangemi, A., Gentile, A.L., Nuzzolese, A.G., Rudolph, S., Maleshkova, M., Paulheim, H., Pan, J.Z., Alam, M. (eds.) The Semantic Web: ESWC 2018 Satellite Events. pp. 14–18. Springer, Cham (2018)
3. Breazeal, C., Takanishi, A., Kobayashi, T.: Social robots that interact with people. Springer handbook of robotics pp. 1349–1369 (2008)
4. Danescu-Niculescu-Mizil, C., Lee, L.: Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In: Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. pp. 76–87 (2011)
5. Dridi, A., Reforgiato Recupero, D.: Leveraging semantics for sentiment polarity detection in social media. International Journal of Machine Learning and Cybernetics **10**(8), 2045–2055 (2019). https://doi.org/10.1007/s13042-017-0727-z, cited By 3
6. Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R.: Framester: a wide coverage linguistic linked data hub. In: European Knowledge Acquisition Workshop. pp. 239–254. Springer (2016)
7. Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with fred. Semantic Web **8**(6), 873–893 (2017)
8. Graetz, G., Michaels, G.: Robots at work. Review of Economics and Statistics **100**(5), 753–768 (2018)
9. Kobayashi, S., Tamagawa, S., Morita, T., Yamaguchi, T.: Intelligent humanoid robot with japanese wikipedia ontology and robot action ontology. In: Proceedings of the 6th international conference on Human-robot interaction. pp. 417–424 (2011)
10. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150 (2011)
11. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Piccolo, L., Mensio, M., Alani, H.: Chasing the chatbots: Directions for interaction and design research (2019)