Language Models for Human-Robot Interaction

Erik Billing*
erik.billing@his.se
Interaction Lab
University of Skövde
Skövde, Sweden

Julia Rosén
julia.rosen@his.se
Interaction Lab
University of Skövde
Skövde, Sweden

Maurice Lamb
maurice.lamb@his.se
Interaction Lab
University of Skövde
Skövde, Sweden

ABSTRACT

Recent advances in large scale language models have significantly changed the landscape of automatic dialogue systems and chatbots. We believe that these models also have a great potential for changing the way we interact with robots. Here, we present the first integration of the OpenAI GPT-3 language model for the Aldebaran Pepper and Nao robots. The present work transforms the text-based API of GPT-3 into an open verbal dialogue with the robots. The system will be presented live during the HRI2023 conference and the source code of this integration is shared with the hope that it will serve the community in designing and evaluating new dialogue systems for robots.

KEYWORDS

Social Robots, Language Models, Dialogue Systems

ACM Reference Format:
Erik Billing, Julia Rosén, and Maurice Lamb. 2023. Language Models for Human-Robot Interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3568294.3580040

1 INTRODUCTION

Large Language Models (LLM) such as OpenAI GPT-3 [4], Google's LaMDA [9], and HuggingFace's Bloom [6, 7], have received significant public attention over the last years. Behind the success of these models lies a paradigm shift within Natural Language Processing (NLP) where pre-trained language models (PTMs) are used to solve a variety of NLP tasks [8]. After pre-training, these models can be fine-tuned to specific NLP applications using a relatively small amount of training data, or even used 'as is' with extensive prompts.

While the most obvious applications of LLM may be within text-based chatbot systems and article generation [5], these models are currently being applied within a variety of areas, including code generation, copywriting, and product requirement documentation [2]. While they have also received some attention within human-robot interaction (HRI), for example through the humanoid robot Ameca by Engineered Arts [3], there are to our knowledge no freely available implementations of LLP for HRI.

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03.
https://doi.org/10.1145/3568294.3580040

Figure 1: Author 2 in discussion with the GPT-3 powered Pepper robot.

We present one of the first implementations of LLM on two common robot platforms used for HRI, the Nao and Pepper robots from Aldebaran [1]. During the live demo, conference participants can engage in open dialogue with the robot, on a topic of their choice to experience the possibilities and limitations of LLM in a live HRI system.

The presented dialogue system is powered by the *GPT-3 Davinci* model, though flexible APIs and rapid LLM development may mean a more advanced model is presented at the conference. The text-based API of GPT-3 is combined with Google Cloud Speech-to-Text and the NaoQi text-to-speech, transforming the English text based interaction of GPT-3 into an open verbal dialogue with the robot. Figure 1 shows an example from a dialogue between the second author and the Pepper robot. A full recording of the conversation is available at https://youtu.be/zip90jyv1i4.

Section 2 comprises a brief technical description of the system, followed by limitations in section 3. Finally, conclusions and future work are found in section 4.

2 TECHNICAL DESCRIPTION

The source code for the presented dialogue system is available at https://github.com/ilabsweden/pepperchat. The implementation integrates three different services that together produce the complete dialogue system. The NaoQi software, which is the default middleware for the Nao and Pepper robots, is used to control the robot's behaviour and to provide text-to-speech service. Google Cloud Speech-to-Text is used to transform the spoken language into text that can be processed by the LLP. Finally, the OpenAI API to GPT-3 is used to generate responses to spoken input.

From a technical point of view, the dialogue system is implemented as four software components described in more detailed below.

2.1 Chatbot service

The *Chatbot service* keep track of the complete dialogue history and sends it to GPT as a request for a text completion, based on the previous dialogue. Generated completions, i.e. robot utterances, comprise up to 256 tokens, (about 500 characters). When developing this demo, the *text-davinci-002* model was used without fine-tuning and with a stochastic temperature of 70%, producing what we perceived as a suitable trade-off between dialogue variability and consistency with the topic of discussion.

2.2 Chatbot bridge

The NaoQI middleware support extensions (modules) written in Python 2.7. Python 2 is deprecated and as a result, OpenAI only provide bindings for Python 3. To bridge communication between the *Chatbot service* written in Python 3.10 and the NaoQi extension modules written in Python 2.7, ZeroMQ was used. This bridge implements a request-reply pattern that takes input in the form of text from the *Speech recognition module* and return the corresponding reply from the *Chatbot service*.

2.3 Speech recognition module

The *NaoQi ALAudioDevice* is used to capture audio from the robot's microphone. Recordings are segmented through thresholding by volume such that the recording will start and stop when the audio goes above and below specific thresholds, for specified amounts of time. Specific thresholds and time durations used for the present demo are found in the release of the source code, and may need to be adjusted based on environmental noise. Each segmented recording is then sent to Google's cloud service for speech-to-text analysis and the result, if recognized, is forwarded to the *Chatbot bridge*.

2.4 Dialogue module

The *Dialogue module* receives responses from the *Chatbot service* in the form of completions (text responses) produced by the GPT-3 model. The *NaoQi ALAnimatedSpeech* is used to transform received completions into robot speech. The *Dialogue module* also communicates with the *Speech recognition module* in order to pause audio recordings while the robot is speaking, producing a turn based dialogue where the robot is listening for human speech or speaking itself, but never both at the same time.

The dialogue module also integrates a few phrases to corresponding actions by the robot, for example *"I will sit down"* and *"I will stand up"*. Pattern matching is used, allowing some flexibility in formulations. While this aspect of the dialogue system is open to extension, the present implementation should be seen as preliminary.

3 LIMITATIONS

The presented dialogue system should be seen as a verbal proxy for the text-based GPT-3 chatbot, not as a fully integrated dialogue system for the robot. While GPT-3 comes with generic knowledge of robots and is tuned to the specific robot used through custom prompts, the language model has no state information about the robot or its environment. Furthermore, while this system allows for integration of action commands through the dialogue module (section 2.4), the system includes limited action command integration in its present form. Thus, while GPT-3 may produce *"I will clap my hands!"* in response to some user input there is nothing linking this phrase to action. While the current implementation of GPT-3 in the robots functions well within our lab, it is highly dependent on cloud based services which require stable internet connections and carry intrinsic data privacy concerns that will vary depending on context. As the system makes use of several cloud services, privacy constitutes an important consideration. During the demo, visitors are informed that collected audio recordings are processed by cloud services. Video data is used for local processing onboard the robot, but is not stored or transmitted. No sensor data other than audio is shared with the cloud. Future use of the presented system should take questions of access, privacy, and security seriously and make sure that relevant precautions and considerations are made.

4 CONCLUSIONS AND FUTURE WORK

Thanks to the large knowledge base of the Davinci language model and its ability to pick up complex semantic relations over longer sequences of text, the dialogue system resulting from the present integration between GPT-3 and NaoQi allow for open conversation with the Pepper and Nao robots on a large variety of subjects. We believe that the system presented here may be useful for studies specifically investigating dialogue systems of robots, but also for studies of other HRI aspects where a generic ability to speak with the robot is desired. The robot, in its current form, is currently active in data collection and research in the authors' lab, with results from these studies being prepared for publications. The fact that the system is based on several cloud services is a limitation, but it also comes with the advantages that it is relatively easy to use and allows for comparisons between platforms, use cases, and environments.

REFERENCES

[1] Aldebaran. 2022. http://www.aldebaran.com.

[2] Leigh Marie Braswell. 2022. Overview & Applications of Large Language Models (LLMs). https://leighmariebraswell.substack.com/p/overview-and-applications-of-large

[3] Engineered Arts. 2022. https://www.engineeredarts.co.uk/robot/ameca/.

[4] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* 30, 4 (Dec. 2020), 681–694. https://doi.org/10.1007/s11023-020-09548-1

[5] GPT-3. 2020. A robot wrote this entire article. Are you scared yet, human? *The Guardian* (Sept. 2020). https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

[6] Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom Library: Multimodal Datasets in 300+ Languages for a Variety of Downstream Tasks. http://arxiv.org/abs/2210.14712 arXiv:2210.14712 [cs].

[7] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Castagné, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. http://arxiv.org/abs/2211.05100 arXiv:2211.05100 [cs].

[8] Tian-Xiang Sun, Xiang-Yang Liu, Xi-Peng Qiu, and Xuan-Jing Huang. 2022. Paradigm Shift in Natural Language Processing. *Machine Intelligence Research* 19, 3 (June 2022), 169–183. https://doi.org/10.1007/s11633-022-1331-6

[9] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, et al. 2022. LaMDA: Language Models for Dialog Applications. http://arxiv.org/abs/2201.08239 arXiv:2201.08239 [cs].