Visual Conversation Starters for Human-Robot Interaction[*]

Ruben Janssens, Thomas Demeester, and Tony Belpaeme

IDLab, Ghent University - imec, Ghent, Belgium
rmajanss.janssens@ugent.be

Abstract. In this demonstration, a Furhat social robot will engage in a conversation with a user and adapt this conversation based on information about the user that the robot has visually identified, such as the apparel of the user. This interaction showcases that data-driven methods in natural language processing and computer vision offer promising possibilities to create personalised, adaptive and natural human-robot interactions, which is essential for robots to support humans in an effective way.

Keywords: conversational agents · human-robot interaction · multimodal interaction · natural language processing · computer vision

1 Introduction

In human-human interactions, we expect the other to see us and to understand the context that we are both in. This is important in order to establish a common ground in these interactions, i.e. the set of propositions in a conversation which we treat as 'true' [10]. For this, situational awareness is necessary: the concept of knowing what is going on around oneself [3] and being able to use that knowledge effectively in an interaction with the environment and with social others.

Yet, when we try to have a conversation with a robotic partner, it is unlikely that the robot is able to meet the criteria we have for human interlocutors, which in turn leads to a less natural interaction and experience. Getting robots to both understand their environment, and allowing them to reference it in a conversation, is a challenging objective in the fields of Human-Robot Interaction (HRI) [2] and Natural Language Processing (NLP). Nevertheless, having robots understand their surroundings, with the ability to weave that understanding into an open-domain conversation, is key to a successful HRI.

This demonstration will show how data-driven methods in NLP and Computer Vision can enable such a visually grounded conversation in an interaction with a social robot. In this interaction, the robot will greet the user with a question that is based on a visual feature of the user, e.g. "How long have you had glasses?". This question will be generated by the Visual Conversation Starter system previously described by the authors [5].

[*] This research received funding from the Flemish Government (AI Research Program).

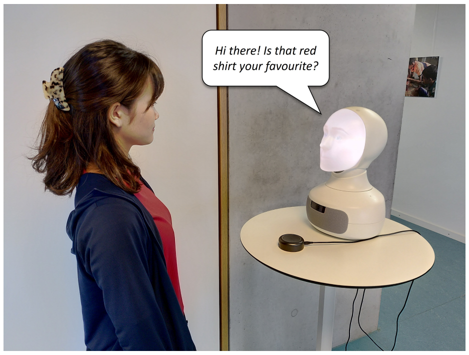


Fig. 1. Illustration of the interaction between a user and a Furhat robot, with the robot inviting the user to have a conversation through a polite question referring to a visual feature of the user.

This research is novel as it constructed a new dataset of images and conversation-starting questions that are appropriate for HRI instead of only human-chatbot interactions, such as previous work by Mostafazadeh and others [8]. Furthermore, it focuses on open-domain conversations instead of specific tasks, e.g. in collaborative robots [6,4]. Finally, we demonstrate the usefulness of data-driven language models for HRI, with image captions as intermediate representation, instead of rule-based systems or scene graphs as intermediate representation [9]. Future work will focus on investigating the effect of such visually grounded conversation starters on the human experience of a human-robot interaction.

2 Demonstration

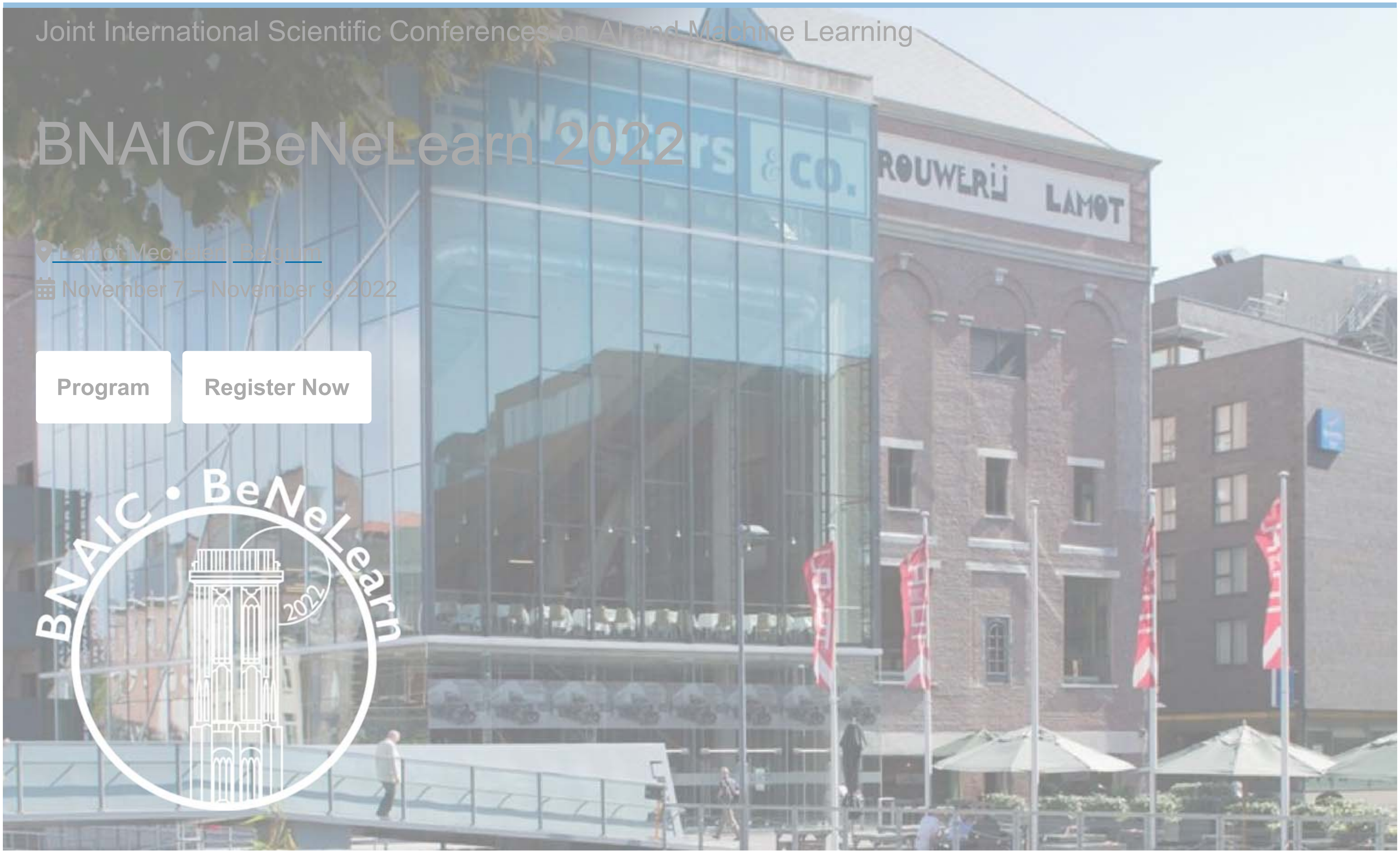
To demonstrate the Visual Conversation Starter system in the wild, we deploy it on a Furhat social robot [1]. Furhat has a camera, text-to-speech software, three-dimensional face with pan-tilt neck, lip syncing, and gaze recognition through which it can 'engage' with people and follow their movements. Fig. 1 shows the setup with the robot and the video accompanying this paper also showcases this demonstrator (accessible at `https://youtu.be/DPTqGCBiMwE`).

The questions are generated based on an image taken by the camera situated in the base of the robot. This image is given to a captioning model [11], which generates a text containing short sentences describing what is visible in the image. This text is then transformed into the question by a version of the BART sequence-to-sequence language model [7] that was fine-tuned by the authors [5].

Both the captioning model and the question generating model are running in a separate VM on a cloud server, each with an NVIDIA GeForce GTX 1080 Ti GPU. A local computer running a Python script connects all the different components by forwarding Furhat's camera feed to the captioning model, sending the caption to the question-generating model, and using that output to drive Furhat's text-to-speech system.

References

1. Al Moubayed, S., Beskow, J., Skantze, G., Granström, B.: Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In: Cognitive behavioural systems, pp. 114–130. Springer (2012)
2. Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M., Šabanović, S.: Human-robot interaction: An introduction. Cambridge University Press (2020)
3. Endsley, M.R., Garland, D.J.: Situation awareness analysis and measurement. CRC Press (2000)
4. Hough, J., Schlangen, D.: Investigating fluidity for human-robot interaction with real-time, real-world grounding strategies. In: Proceedings of the 17th Annual SIGdial Meeting on Discourse and Dialogue (2016)
5. Janssens, R., Wolfert, P., Demeester, T., Belpaeme, T.: "cool glasses, where did you get them?" generating visually grounded conversation starters for human-robot dialogue. In: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction. pp. 821–825 (2022)
6. Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R.: Artificial cognition for social human–robot interaction: An implementation. Artificial Intelligence **247**, 45 – 69 (2017). https://doi.org/https://doi.org/10.1016/j.artint.2016.07.002, http://www.sciencedirect.com/science/article/pii/S0004370216300790, special Issue on AI and Robotics
7. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. CoRR **abs/1910.13461** (2019), http://arxiv.org/abs/1910.13461
8. Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G.P., Vanderwende, L.: Image-grounded conversations: Multimodal context for natural question and response generation. CoRR **abs/1701.08251** (2017), http://arxiv.org/abs/1701.08251
9. Pantazopoulos, G., Bruyere, J., Nikandrou, M., Boissier, T., Hemanthage, S., Sachish, B.K., Shah, V., Dondrup, C., Lemon, O.: Vica: Combining visual, social, and task-oriented conversational ai in a healthcare setting. In: Proceedings of the 2021 International Conference on Multimodal Interaction. p. 71–79. ICMI '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3462244.3479909, https://doi.org/10.1145/3462244.3479909
10. Stalnaker, R.: Common ground. Linguistics and philosophy **25**(5/6), 701–721 (2002)
11. Yang, L., Tang, K., Yang, J., Li, L.J.: Dense captioning with joint inference and visual context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jul 2017)



Joint International Scientific Conferences on AI and Machine Learning

BNAIC/BeNeLearn 2022

Lamot Mechelen, Belgium

November 7 – November 9, 2022

Program

Register Now

BNAIC/BeNeLearn is the reference AI & ML conference for Belgium, Netherlands & Luxembourg. The combined conference will take place from November 7th till November 9th in Mechelen, Belgium and is organized by the University of Antwerp, under the auspices of the Benelux Association for Artificial Intelligence (BNVKI) and the Netherlands Research School for Information and Knowledge Systems (SIKS).

Latest News

Transport on Wednesday November 9

November 4, 2022

On Wednesday **November 9th**, there will be a **national strike in Belgium** which might affect your travel plans back home. Therefore, we advise you to think of some solutions ahead of time, and we will help you to the best of our abilities.

For the time being, **we expect a normal service for most international connections** including Thalys, Eurostar, TGV INOUI and the IC Brussels-Amsterdam trains. The latter also has a stop in Mechelen.

You can check this page for the latest updates:

Latest Info on Strike →

Check out the program

October 11, 2022

We have a full, three day schedule of presentations, posters, demos and invited talks. Find out more on the program page.

Take a look at the accepted submissions

October 5, 2022

A full list of accepted submissions is available on the website

A full list of accepted submissions is available on the website.

Important Dates

- **Welcome coffee**
  → November 7

Sponsors

Universiteit Antwerpen

BNVKI
BAIAI

SIKS

fwo

dpg media

FROOMLE

CREATING AI PERSONALIZED EXPERIENCES

FOODPAIRING®

✉ bnaic2022@uantwerpen.be