A Cooperative Visually Grounded Dialogue Game with a Humanoid Robot

Jordan Prince Tremblay, Ismael Balafrej, Félix Labelle, Félix Martel-Denis,
Éric Matte, Julien Chouinard-Beaupré, Adam Létourneau, Antoine Mercier-Nicol,
Simon Brodeur, François Ferland and Jean Rouat
Dépt. GEGI, U. de Sherbrooke, Québec, Canada
devine.gegi@listes.usherbrooke.ca

This demonstration illustrates how research results in grounded language learning and understanding can be used in a cooperative task between an intelligent agent and a human. The task, undertaken by a robot, is the question answering game GuessWhat?! [1]. The robot interacts with a human player who performs the role of the oracle. The oracle selects an object in a visual scene without disclosing it to the robot. Playing the role of the guesser, the robot asks questions to the person to deduce which object was selected. The guesser's questions are generated using the player's previous answers and visual information. Once the robot has a strong belief, it guesses the object chosen by the player and physically points to it in the scene (Figure 1).

The robotic platform for the demonstration is IRL-1 [2]. By design, it can operate in a typical human environment. It is equipped with a Kinect, differential elastic actuators for its arms, a microphone array, and powered wheels with omnidirectional capability. As IRL-1 has compliant joints, players are protected from physical harm. To begin playing GuessWhat?! with IRL-1, a player approaches the robot. Once the player is close enough, IRL-1 asks if they want to play. After explaining the rules, IRL-1 takes an image of the scene comprising different objects. Then, it turns back to the player and asks if they have selected an object. Upon confirmation, IRL-1 begins to ask them questions. After five questions, the robot reveals its guess. The player validates IRL-1's guess, IRL-1 reacts accordingly and then concludes the game.

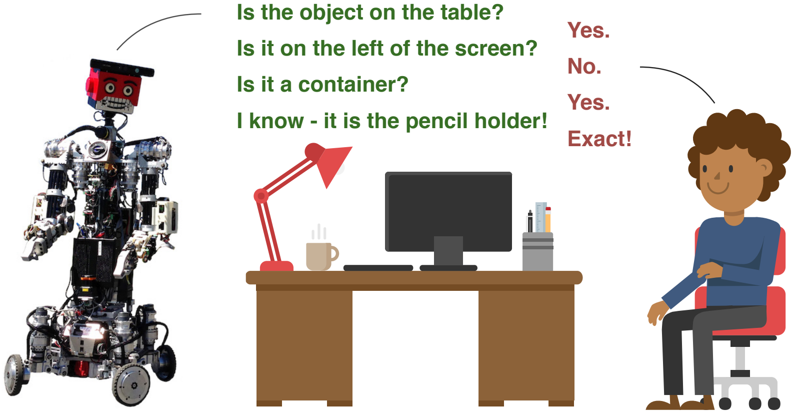
Is the object on the table?
Is it on the left of the screen?
Is it a container?
I know - it is the pencil holder!
Yes.
No.
Yes.
Exact!

Figure 1: Scenario example for the live demonstration with IRL-1 robot

Providing human-robot interactions in the real world requires interfacing GuessWhat?! with speech recognition and synthesis modules, video processing and image recognition algorithms, and the robot's control module (Figure 2). A web-based diagnostic system and the Gazebo simulator[1]

[1] http://gazebosim.org/

accelerated the development of subsystems. Natural human-robot interactions are achieved in soft real-time with the help of a GPU, a necessity because multiple state-of-the-art neural networks are used during sensory and dialogue processing.

One main challenge is adapting GuessWhat?! to work with images outside of MSCOCO's domain [2]. This required implementing a pipeline in ROS [3] which takes images from a Kinect, ensures image quality with blur detection, extracts VGG-16 feature vectors [3], segments objects using Mask R-CNN [4], and extracts position information from the segmented objects. Images from the pipeline are used by GuessWhat?! in tandem with utterances from the player. Snips [4], a private-by-design voice assistant, recognizes whether the player says “Yes”, “No” or “Not Applicable". Snips also provides speech synthesis, converting the questions generated by GuessWhat?! into speech for the player. To identify potential players, OpenPose [5] allows IRL-1 to interact with them throughout the game. The guesser's confidence in its hypothesis determines IRL-1's facial expression during the game. At the end of the game, IRL-1 emotes as a function of the game's result and its confidence in that result.

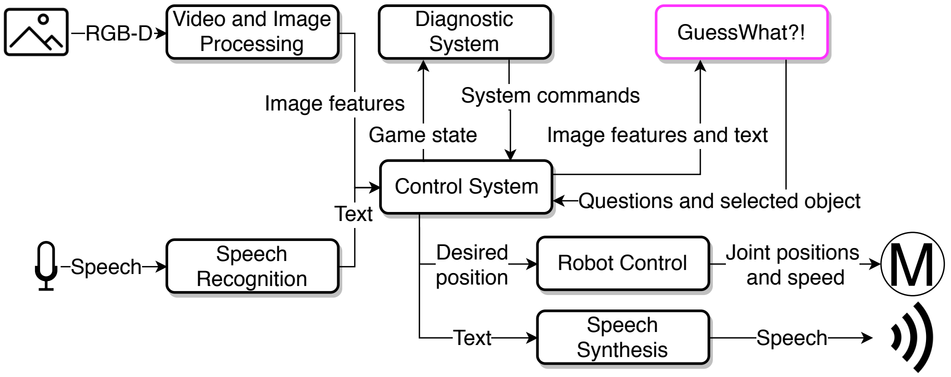

Figure 2: Subsystems diagram: All blocks are developed or integrated around GuessWhat?!

Before the NIPS demonstration, all software and documentation will be released in the public domain. This could be useful as intelligent agents are becoming commonplace and the ability to communicate with people in a given context, such as the home or workplace, becomes imperative. The various functionalities implemented around and, including GuessWhat?!, on IRL-1 would be beneficial to any agent assisting a person in a cooperative task.

Acknowledgments

We would like to thank NVIDIA for a Titan X, IntRoLab for IRL-1, NECOTIS for technical support, CHIST-ERA IGLU & UdeS. Special thanks to F. Strub for his guidance, Mask_RCNN, tf-pose-estimation, tensorflow-vgg, Snips and ROS.

References

[1] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.

[2] F. Ferland, D. Létourneau, A. Aumont, J. Frémy, M.-A. Legault M. Lauria, and F. Michaud. Natural interaction design of a humanoid robot. *Journal of Human-Robot Interaction*, 1(2):118–134, 2013.

[3] K. Simonyan and A. Zisserman. Very deep CNN for large-scale image reco. *CoRR*, abs/1409.1556, 2014.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[5] Z. Cao, T. Simon, S-E Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017.

[2] http://cocodataset.org/

[3] http://www.ros.org/

[4] https://snips.ai/