MIDA: GPT-powered Furhat Robot as a Multilingual Information Desk Assistant

LT2318 — Project Proposal — Bogdan Laszlo, Aruna Elentari

Project Summary

Our research project explores the space of human-robot interactions in the context of tourism. Specifically, it involves integrating the Furhat robot with GPT-4 to develop a Multilingual Information Desk Assistant, or MIDA. We would like to capitalize on Furhat's highly expressive face, human-like head movements, onboard sensors for audio and visual perception, over 40 supported spoken languages and the dialogue flow model, as well as GPT-4's multimodal capabilities, to design a multilingual conversational agent.

The project encompasses two aspects of machine learning: natural language processing (NLP) and computer vision. Our primary goal is for MIDA to promptly respond to users' speech, and answer their questions in a timely, succinct and satisfactory manner. First, MIDA should be able to realize that a user addresses it in a specific language, and to start conversing with the user in that language. If the language is not supported by Furhat, MIDA should politely explain that to the user in English, which will be the default language of interaction. It should also be able to interact with multiple users, and smoothly switch between users, both in terms of language and context. We will focus on active listening via a friendly dialogue (through the use of prompting), zooming in on the relevant information and gracefully handling missing information, turn-skipping or irrelevant information.

The second part of the project involves image recognition and classification, so that the robot is able to react to the visual input provided by the user, including the human form, maps and flags. The ability to process visual information will provide grounding for MIDA and help it extract relevant information from the dialogue. Furhat's ability to follow a user's gaze and have a friendly expression will contribute to rapport between the user and MIDA. While providing useful and relevant information is extremely important, one should not underestimate the pleasantness of interaction in the overall user experience. For this, we will use a number of gestures supported by Furhat, such as smiling, blinking, nodding and appropriate facial expressions.

The overall aim of the project is to evaluate the importance of human-like faces and facial expressions in human-robot interaction, gain skills in configuring robots that converse smoothly with humans and faithfully answer their questions, and develop a general framework for the combined use of LLMs and social robots. We chose the Furhat robot because we received introductory training on it, and it appears easy to program. Both Furhat and GPT-4 have proprietary software, which limits our exploration, but there are other open-source models and robots that might benefit from any insights from this project.

Objective

The objective of this project is to create a prototype artificial multilingual travel information desk agent, which can aid tourists in a human-like fashion. The automatizing of the tourism industry is long overdue, as its

effects would be positively felt, both regarding the economic output and the overall quality of its services by reducing language barriers(Kolar and Kumar, 2023), (Martins et al., 2020), enhancing the accessibility of remote points of interest(Amessafi et al., 2017) or just providing much needed support to overlooked destinations(Garrido et al., 2017).

This artificial agent is designed to communicate effectively and intuitively with tourists in their native languages or widely spoken international languages, offering a significant advantage in the realm of tourist assistance. Moreover, by integrating the capabilities of Large Language Models (LLMs) and computer vision technologies, the agent will not be constrained by traditional limitations. It will be equipped to respond to queries both verbally and visually(Abbo and Belpaeme, 2023). This dual-mode interaction ensures that tourist inquiries are addressed in a comprehensive and varied manner, thus improving the overall user experience.

Available Data and Resources

The project will base its flag recognition model on data originating from the (Avanti, 2023) dataset, which may or may not be further augmented with more images of flags for further improving the model's performance.

The flag detection model itself will be built based on (Vukašin Manojlović, 2020), which will either be adapted to work with a different dataset or will serve as inspiration for creating a new model from scratch.

The LLM would be provided through OpenAI's GPT4-vision API, allowing both advanced speech understanding, providing an extensive knowledge base, multilingualism and image processing capabilities (OpenAI, 2023).

The speech synthesis and recognition alongside the real-time vision capture will be provided by the Furhat robot which has these capabilities out of the box(Furhat Robotics, 2023).

Technical Features and Capabilities

Our project hinges on performing several tasks to fully achieve its goals.

Its features must be contained within a Furhat skill that can easily be toggled on, modified and swapped, leveraging Furhat's innate developer-friendly aspect.

This skill must contain the following capabilities:

1. Object detection in order to identify people, flags, and potential maps from the environment.
2. Once the objects of interest are identified, MIDA must shift its gaze towards the person initiating the conversation.
3. Upon flag detection, it must classify it to switch to its assigned language. This is based on the robot's pre-existing knowledge, but it should also be updateable through user interaction (MIDA has a preset list of supported languages, and it can associate new flags to said languages for future detection).
4. On successful language detection, MIDA must switch over to it. This is done both through altering the speech detection and synthesis internal configurations (so that it would detect and speak the new language

properly), but also by adjusting the prompt given to the GPT API accordingly, so it would generate text based on the selected language.

Additionally, MIDA should also read maps if prompted to. This is initiated verbally by the user, after which the user is prompted by MIDA to present a map. Once the map is presented, MIDA will capture a snapshot of it, sending the image data through the GPT4 API for interpreting and obtaining an answer to the user query, which would later be relayed back to the user.

Expected Results and Evaluation

The expected result is to create a **Furhat** skill that would allow the robot to be able to dynamically adjust and accommodate potential tourists. This would include:

- **multilingual capabilities**: switching languages upon detecting environmental **triggers**
- **multimodal capabilities**: ability to process both speech and image **information**
- **general dialogue capability**: ability to hold a free flow conversation, alongside **backchannels**

The project will be evaluated using the following methods:

- The flag classification and object detection will be evaluated quantitatively using conventional accuracy, precision, recall and f1 scores, with the added mention that they will be tested under different environments and lighting conditions.
- MIDA's dialogue capabilities will be evaluated qualitatively by analyzing aspects such as speed of response, verbosity of response, accuracy of information, facial expression relevance.

Technical Challenges

The challenges that are expected to arise in the process of developing MIDA are the following:

- Developing the Object Detection Model, choosing the right neural-network architecture, making sure that the training (or fine-tuning) data is relevant for its projected use-case and that it offers sufficiently good performance.
- Designing and implementing the new flag learning capability.
- Prompt engineering related difficulties to ensure that the conversation structure is simultaneously kept both on-topic but flexible enough to provide a human-like experience to the user. This will involve ensuring that the conversation boundaries are enforced, and that the conversation will be steered back to the allowed topics as discreetly and organically as possible.
- Integrating the aforementioned modules to work seamlessly within Furhat's framework.

Ethical Considerations

Some ethical issues that may arise from the result of our project may be the following:

- **Negative discrimination based on language/nationality** - this could entail the specific mistreatment of tourists based on their culture or country of origin, ranging from outright refusing to provide information, providing erroneous information, or providing information based on preconceived and stereotypical notions about the user's interests based on their culture or country of origin.
- **Overt promotion of points of interest based not on tourist preference or query relevance, but on monetary incentives** - sending tourists towards restaurants which funded the robot service, in the detriment of other establishments which would be a better fit to the user's query.
- **MIDA's computer vision capabilities could be harnessed by various actors (government, law enforcement or owner company)** - user information gathering for purposes such as user profiling, police surveillance and data harvesting and monetizing.

References

Giulio Antonio Abbo and Tony Belpaeme. I was blind but now i see: Implementing vision-enabled dialogue in social robots. https://arxiv.org/pdf/2311.08957.pdf, nov 2023. arXiv preprint.

Hanane Amessafi, Reda Jourani, Adil Echchelh, and Houssain Yakhlef. Building a smart interactive kiosk for tourist assistance. *Transactions on Machine Learning and Artificial Intelligence*, 5, 08 2017. doi: 10.14738/tmlai.54.3326.

Avanti. Flags dataset. https://universe.roboflow.com/avanti-nlyef/flags-0viie, jun 2023. URL https://universe.roboflow.com/avanti-nlyef/flags-0viie. visited on 2023-11-30.

Furhat Robotics. Furhat robot documentation, 2023. URL https://www.furhatrobotics.com/docs.

Piedad Garrido, Javier Barrachina, Francisco J. Martinez, and Francisco J. Seron. Smart tourist information points by combining agents, semantics and ai techniques. *Computer Science and Information Systems*, 14(1): 1–23, 2017. doi: 10.2298/CSIS150410029G.

Sanjana Kolar and Rohit Kumar. Multilingual tourist assistance using chatgpt: Comparing capabilities in hindi, telugu, and kannada, 2023.

André F. T. Martins, Joao Graca, Paulo Dimas, Helena Moniz, and Graham Neubig. Project MAIA: Multilingual AI agent assistant. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 495–496, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.68.

OpenAI. Openai vision api documentation, 2023. URL https://platform.openai.com/docs/guides/vision.

Sanja Mijović Vukašin Manojlović. flagnet. https://github.com/iamvukasin/flagnet, 2020.