Towards Next-Generation Intelligent Assistants Leveraging LLM Techniques

Xin Luna Dong
Meta Reality Labs
Redmond, Washington, USA
lunadong@meta.com

Seungwhan Moon
Meta Reality Labs
Redmond, Washington, USA
shanemoon@meta.com

Yifan Ethan Xu
Meta Reality Labs
Austin, Texas, USA
ethanxu@meta.com

Kshitiz Malik
Meta Reality Labs
Burlingame, California, USA
kmalik2@meta.com

Zhou Yu
Columbia University
New York, New York, USA
zy2461@columbia.edu

ABSTRACT

Virtual Intelligent Assistants take user requests in the voice form, perform actions such as setting an alarm, turning on a light, and answering a question, and provide answers or confirmations in the voice form or through other channels such as a screen. Assistants have become prevalent in the past decade, and users have been taking services from assistants like *Amazon Alexa*, *Apple Siri*, *Google Assistant*, and *Microsoft Cortana*.

The emergence of AR/VR devices raised many new challenges for building intelligent assistants. The unique requirements have inspired new research directions such as (a) understanding users' situated multi-modal contexts (e.g. vision, sensor signals) as well as language-oriented conversational contexts, (b) personalizing the assistant services by grounding interactions on growing public and personal knowledge graphs and online search engines, and (c) on-device model inference and training techniques that satisfy strict resource and privacy constraints.

In this tutorial, we will provide an in-depth walk-through of techniques in the afore-mentioned areas in the recent literature. We aim to introduce techniques for researchers and practitioners who are building intelligent assistants, and inspire research that will bring us one step closer to realizing the dream of building an all-day accompanying assistant. Additionally, we will highlight the significant role that Large Language Models (LLMs) play in enhancing these strategies, underscoring their potential to reshape the future landscape of intelligent assistance.

CCS CONCEPTS

• **Computing methodologies** → *Knowledge representation and reasoning*; *Computer vision*; Distributed artificial intelligence; *Discourse, dialogue and pragmatics*; • **Human-centered computing** → **Personal digital assistants**.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599572

KEYWORDS

conversational AI, large language models, multi-modal conversation, knowledge grounding, personalization, federated learning

ACM Reference Format:
Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards Next-Generation Intelligent Assistants Leveraging LLM Techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3580305.3599572

"No one is more cherished in this world than someone who lightens the burden of another." — Joseph Addison

1 OUTLINE OF THE TUTORIAL

1.1 Introduction

We will start with an introduction discussing the basics of virtual assistant, and the new challenges we face in building an AR/VR assistant. We will discuss what is an ideal assistant; that is, *an agent that knows the user and the world, can receive requests from the user in a reactive fashion, or predict the users needs in a proactive fashion, then provide the user the right services at the right time, with the user's permission.*

1.2 Conversational AI Basics

Before diving deep into cutting-edge techniques addressing the novel challenges for AR/VR assistants, we first provide a high-level overview of the conventional language-oriented assistant system. We introduce the overall design of both open-domain end-to-end conversational AI and modularized task-oriented dialog systems, but put more focus on the latter as they are the backbone of virtual assistants in industry. We dive deeper into key components: Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialog State Tracking, Dialog Policy Learning, Natural Language Generation (NLG), and Text-to-Speech (TTS). Due to time constraints, the tutorial will review basic construction with focus on modern modeling designs of each component, and leave implementation details as references. Links to public datasets will also be provided for training and evaluating dialogue-based assistants.

1.3 Multi-modal Context-Aware Conversations

We envision that the next generation of virtual assistants will be able to process multimodal inputs and provide multimodal outputs beyond the traditional NLP stack. Specifically, the AR/VR settings pose a situated multimodal context, where the user and the assistant are continually co-observing the same context, dynamically updated over time. The assistant for AR/VR thus requires (1) understanding of the multimodal context from diverse sources (*e.g.* vision, gestures, sensor signals), and (2) joint grounding of the situated context with the conversational context.

We review the related literature across multiple domains and tasks, including the visual question and answering systems the visual navigation tasks and more general task-oriented multimodal agents. We will then dive deeper into the state-of-the-art modeling techniques that address the key challenges, such as multimodal co-reference resolution, multimodal dialog state tracking, and contextual understanding of user states.

1.4 Knowledge-Enhanced and Personalized Conversations

To better assist a user in a personalized and contextualized manner, it is important for an assistant system to be able to (1) manage personal knowledge through memory grounded dialogue system, and (2) incorporate the world and personal knowledge as part of the grounding context for conversations. Incorporating personal memories as part of conversational interactions will be particularly important for AR devices that reside more closely to users' everyday life, incurring more frequent usage. We will go through in detail the relevant work in the recent literature on inferring new knowledge from unstructured utterances, utilizing memory graphs, knowledge graphs and online-searches for conversational recommendations, question and answering, media retrieval, and knowledge grounded open-ended conversations.

1.5 On-device & Federated Learning for Privacy Preserving Assistant

AR/VR devices have strict privacy constraints because they can access sensitive data from cameras, microphones and other sensors. This necessitates running model inference and training on-device, under challenging compute, memory and power constraints. We first review the state-of-the-art on-device modeling methods used in practice to downsize neural models (*e.g.* quantization, knowledge distillation, automated architecture search, accommodating for the limited memory and compute resources of wearable devices.

We then discuss the unique challenge of training these models while preserving user privacy. We start with an overview of production federated learning systems, and then discuss the problems that arise when training data is spread across edge devices, like the impact of data and device heterogeneity, model personalization, and differential privacy.

1.6 Conclusions & Future Directions

We conclude our tutorial by stating a number of open problems we need to solve to move towards the goal of building next-generation assistants for AR/VR devices. The open problems include 1) on-device machine learning to provide assistant services with sporadic connections, 2) seamless integration of search, question answering, recommendation for information-driven needs, 3) proactive service suggestion at the right time, 4) leveraging public and personal knowledge graphs to improve context-aware services, 5) scalable graph mining from knowledge graphs, social graphs, and behavior graphs for better assistance, and many others.

2 PREVIOUS EDITIONS

Below is a list of the tutorials on similar topics.

- A related tutorial, titled "Deeper Conversational AI", was given at NeurIPS 2020. The authors first reviewed general conversational AI architectures of the key components. Then they deep dived into generational seq2seq deep conversational AI techniques, pointed out limitations and solutions of vanilla models such as lack of diversity, consistency, knowledge etc. Finally, the authors touched upon various challenges and research directions, including reinforcement learning, few/zero shot learning, lifelong learning.
- Another related tutorial titled "Achieving Common Ground in Multi-modal Dialogue" was given at ACL 2020. It reviewed theories and practices of incrementally achieving common ground between a robotic agent and a human participant during an open-domain dialog, with an emphasis on leveraging verbal and non-verbal behavior signals to orchestrate an effective engagements. The authors considered roles of a wide range of modalities including gazing, pointing, nodding, facial expressions and other non-verbal cues.

Our tutorial is more focused on the context of building AR/VR assistants. On the one hand, we go beyond multi-modal conversations to discuss contextual AI and personalized assistant services. On the other hand, we discuss the practical challenges we face and the industrial solutions in building real assistant systems.

Finally, a first edition of this tutorial is accepted by WebConf 2023. Our new version of the tutorial would add how the recent progress on LLM (Large Language Model) can further improve intelligent assistant.

3 AUDIENCE PARTICIPATION

To help the audience understand challenges brought by multi-modality, contextualization, and personalization, we will use one real-world scenario with progressive add-ons throughout the tutorial to illustrate the challenges and opportunities, and demonstrate the technical solutions. In addition, we will set aside time for Q & A in each subsection and encourage the audience to ask questions to ensure they understand the material and stays engaged. All slides will be made available publicly before the start of the tutorial.

4 POTENTIAL SOCIETAL IMPACTS

In recent years, the use of virtual assistants has become increasingly popular, impacting various aspects of people's lives. Our tutorial will encourage discussions on developing virtual assistants with integrity and responsibility in mind. Such considerations include but not limited to ensuring transparency and privacy, improving accessibility, avoiding bias, and promoting inclusivity.