TaskMAD: A Platform for Multimodal Task-Centric Knowledge-Grounded Conversational Experimentation

Alessandro Speggiorin
University of Glasgow
Glasgow, United Kingdom
alessandro.speggiorin@glasgow.ac.uk

Jeffrey Dalton
University of Glasgow
Glasgow, United Kingdom
jeff.dalton@glasgow.ac.uk

Anton Leuski
Institute for Creative Technologies,
University of Southern California
Los Angeles, USA
leuski@ict.usc.edu

ABSTRACT

The role of conversational assistants continues to evolve, beyond simple voice commands to ones that support rich and complex tasks in the home, car, and even virtual reality. Going beyond simple voice command and control requires agents and datasets blending structured dialogue, information seeking, grounded reasoning, and contextual question-answering in a multimodal environment with rich image and video content. In this demo, we introduce Task-oriented Multimodal Agent Dialogue (TaskMAD), a new platform that supports the creation of interactive multimodal and task-centric datasets in a Wizard-of-Oz experimental setup. TaskMAD includes support for text and voice, federated retrieval from text and knowledge bases, and structured logging of interactions for offline labeling. Its architecture supports a spectrum of tasks that span open-domain exploratory search to traditional frame-based dialogue tasks. It's open-source and offers rich capability as a platform used to collect data for the Amazon Alexa Prize Taskbot challenge, TREC Conversational Assistance track, undergraduate student research, and others. TaskMAD is distributed under the MIT license. [1]

CCS CONCEPTS

• **Computing methodologies** → *Discourse, dialogue and pragmatics*; *Interactive simulation*; • **Information systems** → *Users and interactive retrieval*; • **Human-centered computing** → *Interactive systems and tools.*

KEYWORDS

interactive search; data collection; wizard-of-oz

ACM Reference Format:
Alessandro Speggiorin, Jeffrey Dalton, and Anton Leuski. 2022. TaskMAD: A Platform for Multimodal Task-Centric Knowledge-Grounded Conversational Experimentation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531679

[1] TaskMAD is available at https://github.com/grill-lab/TaskMAD

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531679

1 INTRODUCTION

The rapid growth of virtual personal assistants such as Alexa, Google Assistant, Siri, and Cortana highlights the importance and utility of voice-based task assistants to help accomplish real-world tasks. However, many of these are currently limited to simple intents with basic commands. This is evolving, for example, the recent Amazon Alexa Taskbot Challenge supports users performing complex real-world tasks, such as cooking and DIY with voice, but also with a screen to allow interaction with rich image and video content. These tasks involve multiple steps that encompass unstructured and structured conversational elements including task ranking and selection, task-centric question answering, and require grounded reasoning about structured real-world state (i.e. timers running, what's in the oven). This requires a conversational system and platform that is able to bridge the 'structure chasm' between structured (often frame-based) task systems and open-domain conversational search as well as support diverse types of rich interaction modalities.

Developing these new systems requires new datasets specifically designed for task-specific information-seeking [3]. However, due to the complex nature of these tasks and interactions, it is common to employ a Wizard-of-Oz (WoZ) experiment where system actions are performed by a human-wizard operator. A challenging task is creating a WoZ platform capable of supporting the diverse Conversational Information Seeking (CIS) tasks as well as a wizard-friendly interface that supports efficient interaction and action constraints. In this demo, we focus on developing such an open platform for interactive data collection with a WoZ experimental setup. Its goal is to support diverse CIS tasks [11], from open-domain conversational search over text to rich multimodal real-world tasks like cooking. The aim is to support researchers in creating a new generation of reusable multi-domain datasets that include all elements of a realistic and usable system that supports complex multimodal tasks.

The goal of TaskMAD is not to be an end-to-end framework for building or evaluating conversational systems. For structured dialogues, this exists with ConvLab-2 [12] that is used in the dialog systems community in the DSTC Challenge. For chatbots, the ParlAI [6] framework includes diverse functionalities that support multiple chat models as well as limited asynchronous conversational QA crowdsourcing. However, it focuses on chat and does not support structured tasks, conversational search, and has limited multimodal QA support [8]. Macaw [10] focuses on open-domain CIS with limited support for WoZ and very minimal wizard interfaces as well as lacking support for structured dialogue tasks. In contrast,

TaskMAD is not an overall platform for systems, but a platform for collecting rich task-centric datasets with multimodal elements.

In this work, we present our Task-oriented Multimodal Agent Dialogue (TaskMAD), an interactive platform for multi-modal and contextualized knowledge grounded information-seeking. It builds upon and reuses components and infrastructure from the earlier Agent Dialogue (AD) system [1]. Similar to Agent Dialogue, TaskMAD relies on production-ready tools such as Docker, Kubernetes, gRPC, Google Protocol Buffers, Flask, and search integration to allow scalable WoZ experiments able to interoperate with diverse search and dialogue systems. As detailed below, TaskMAD introduces multiple novel contributions to WoZ data collection that support a wide breadth of CIS tasks beyond previous capability. In the next section, the main contributions of this demo are outlined and described in detail.

2 CONTRIBUTIONS & RESEARCH SCOPE

TaskMAD builds upon the Agent Dialogue core infrastructure but introduces many new changes that span all aspects of the system and tasks supported. A key change is that instead of conversations with a single static document loaded from an excel sheet, TaskMAD integrates with federated search systems and generates the wizard interface dynamically from one or more document results. It adds support for logging wizard search queries and result interactions. Besides text chat, it also supports voice-based interaction for more natural and realistic conversations. There are new wizard and user interfaces redesigned to support rich image and video elements.

A key change is that TaskMAD supports wizard control over structured frame-based task intents using *action-messages* to update state and for operations such as task navigation. The new interface is designed to support these to track evolving task states that can be shown to the user and/or wizard. Also, TaskMAD allows the wizard to compose custom messages independently from the document under consideration introducing the potential for response summarization and generation. TaskMAD is a highly customizable framework, to make task-oriented data collection experiments accessible and scalable.

We summarize these contributions below:

- TaskMAD introduces a new Search API module to allow effective data management and retrieval on custom datasets. This provides fully controlled and parametrized access to domain-specific data.
- We create a new wizard interface by integrating new and easily extendable search capabilities on external data sources. Furthermore, we provide the functionality to allow the wizard to write custom messages and efficiently generate complex context-based responses by selecting text from different sources. This allows the wizard to search and generate knowledge-grounded responses based on the user information need.
- We add a new task and domain-centric user chat interface that provides pre-selected contextual and task-specific information.
- TaskMAD introduces customizable *action messages* that trigger functionalities synchronously on the wizard and chat interfaces. Action messages model changes in intents and resulting system actions. These can be easily modified to provide support for diverse task agents and device behaviors.
- It provides support for multi-modal interactions. More precisely, the current system supports sending and receiving images and videos, with the ability to support audio planned for a future release.
- It adds new structured logging support to monitor the wizard and user interactions by tracking clicks, search queries, search results, and their relative timestamps. This metadata is critical for replicating WoZ experiments offline and for training knowledge-grounded conversational models.

TaskMAD is used for multiple research data collection efforts. This includes ongoing use in the Amazon Alexa Taskbot Challenge, mixed-initiative for the TREC Conversational Assistance Track [2], and ongoing collaboration with researchers at Regensburg and beyond. We envision TaskMAD as a flexible platform that can be used by the research community in order to speed up the data collection process for a wide variety of tasks such as conversational QA, structured dialogues, natural response summarization, generation tasks, and the development of mixed-initiative policies.

3 IMPLEMENTATION & SYSTEM ARCHITECTURE

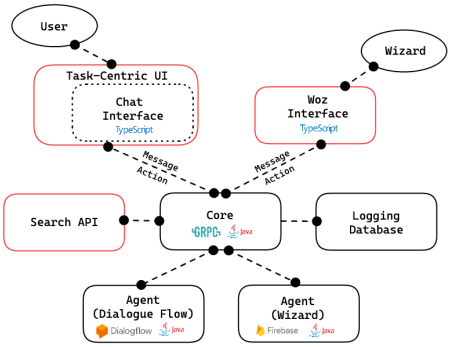

Figure 1: TaskMAD architecture & core modules.

As it is possible to see from Figure 1, TaskMAD is composed of several modules: one or possibly two user interfaces, the user chat and the wizard interface, a custom-built Search API to provide access to external data, one or multiple backend Agents, which provide dialogue management and search functionalities, and a Core Server which handles the communication between components. As previously mentioned, the system uses scalable production-ready technologies such as React, gRPC, Flask, Docker, and Kubernetes, for efficient cloud deployment and to support large-scale experiments.

Search API. One of the core functionalities is the capability of retrieving content from external sources. This is achieved through a Search API module that allows retrieval from custom-built indexes. The module provides easily customizable endpoints for the retrieval of single documents or multiple pages by using any of the standard information retrieval platforms such as Anserini [9] and Terrier

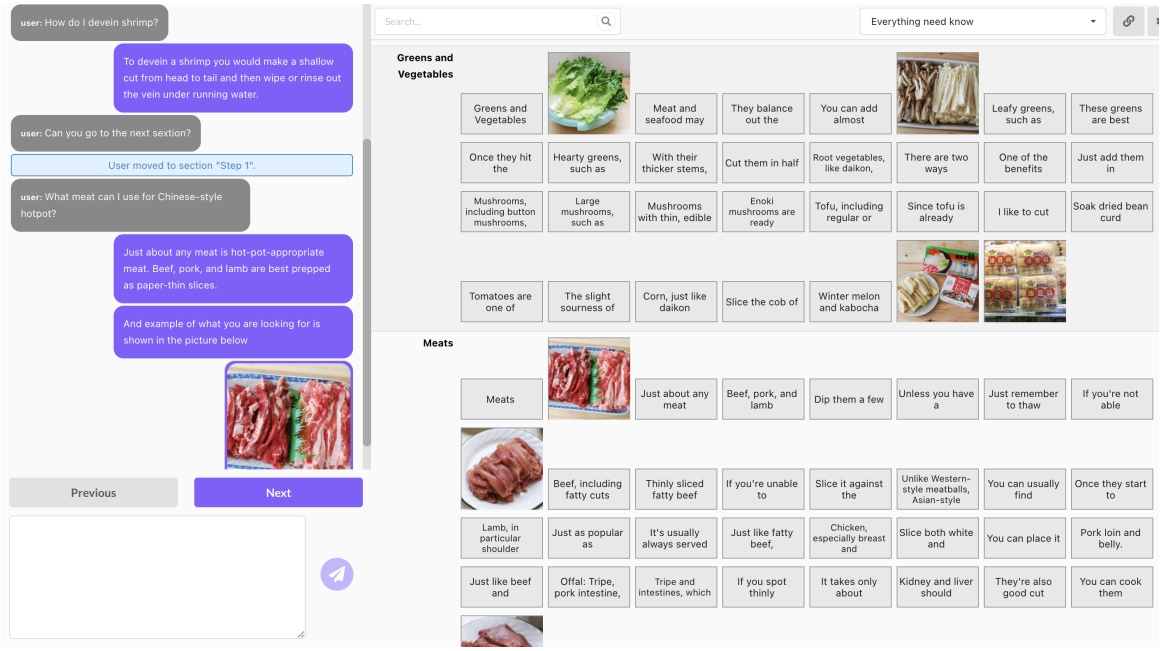

Figure 2: The multimodal task interface for the wizard with task state, search, and action controls.

[7]. In other words, the searcher functionality provides a flexible API structure that can be easily extended and parameterized with custom datasets and search configurations.

Wizard Interface. The wizard interface consists of a modular web app built upon the original infrastructure inspired by earlier WoZ systems developed at the University of Southern California. The structured button system provides control and structure for selecting document-grounded responses [4] that provide passage-level response annotation. The interface, as shown in Figure 2, consists of a set of screens that can be considered self-contained but related documents. This allows the wizard to navigate effectively over a set of predefined correlated articles. Screens are organized into rows containing multiple buttons. Each button has a short label, to help and guide the wizard in navigating the interface, and a tool-tip, illustrating the full button's content when hovering it with the cursor. Furthermore, on the left side of the screen, the chat transcript between the user (grey utterances) and wizard (purple utterances) is displayed.

In the TaskMAD system, we were interested to give the wizard ability to retrieve new information from external corpora and rewrite the retrieved information when presenting it to the user. Due to this fact, the wizard interface includes an input box in the bottom-left corner (see Figure 2). When a wizard selects a button, we append the text of the button to the input box where the wizard can edit it. We keep track of the buttons the wizard uses to compose the response to record the provenance of the response and the edit actions applied to be able to use as data for learning summarization and contextualization.

The new wizard UI adds rich search capabilities. In addition to searching the existing buttons on the screen, performing a query in the top-left input box retrieves relevant results/paragraphs from the Search API module. Selecting a button retrieved from the Search API opens an overlay modal box shown in Figure 3. The modal provides contextual information to the wizard by presenting the retrieved paragraph (highlighted in yellow) and the content of its source page organized into sections. Selecting modal check-boxes appends their content to the input box. The interface also supports buttons that display images (see Figure 2). When a wizard clicks on an image button, the system sends the image to the chat.

Lastly, we define system support for customizable action messages. More precisely, messages of different types can be sent in order to trigger system action behavior. For instance, the buttons Next and Previous in Figure 2 allow the wizard to update the left panel in the chat interface, shown in Figure 4, in order to guide users through steps in a structured task.

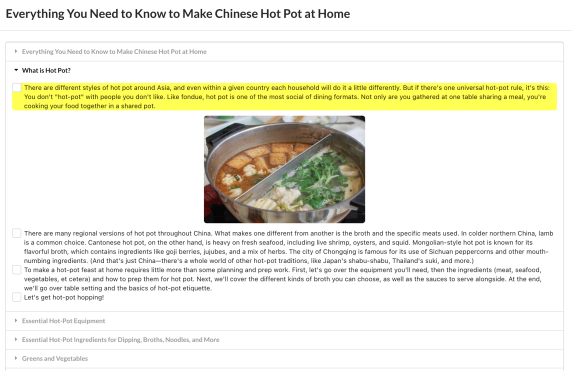

Figure 3: Dynamic search result interface showing content from documents or knowledge base from external APIs.

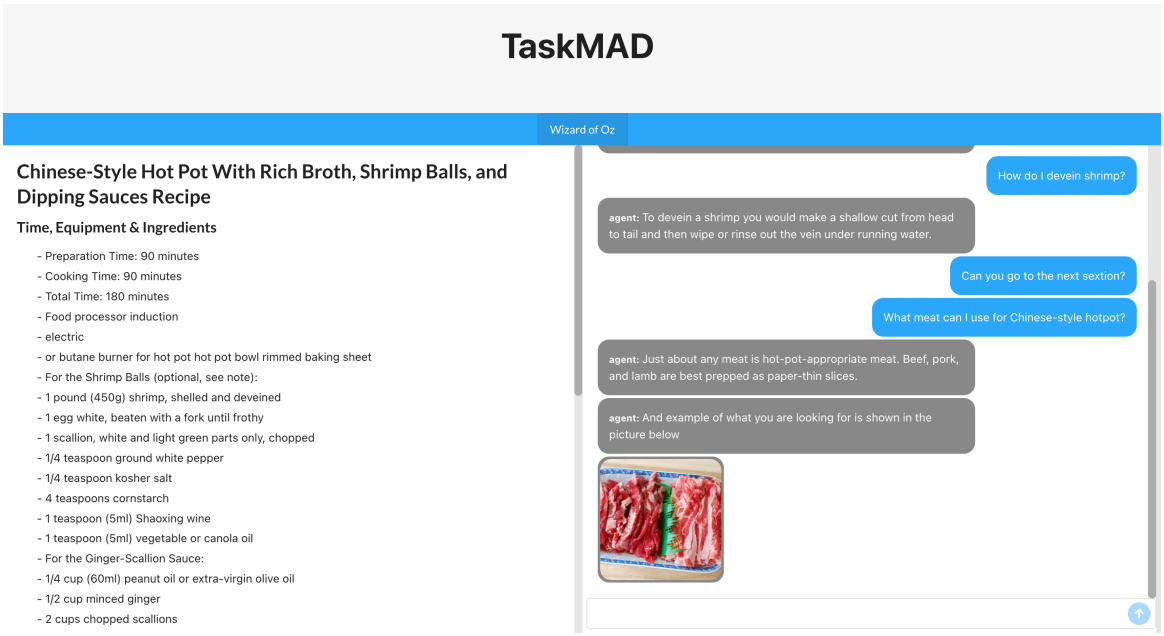

Figure 4: User interface with contextual task context and conversation history.

User Chat Interface. Motivated by the previous limitations of the Agent Dialogue chat interface, we created a new one enriched with significantly new and updated features that include multi-modal rendering, text-to-speech support, and contextual/task-centric information. As shown in Figure 4, we extend the chat by implementing a separate panel with the conversational task context. The interface supports multi-step transitions making it suitable for process-based tasks (such as cooking or DIY). Navigation between steps can be configured to be controlled by the user or by the wizard (through action messages).

Agent Dialogue Core & Agent Interface. TaskMAD reuses the scalable Agent Dialogue Core that acts as a gateway among multiple components [1]. Messages, which are bidirectionally sent from the front-end and the agents, are orchestrated by a central backend gRPC Java-based server. Most of this core functionality is reused from the original AD system with new integrations for search API and Alexa system support.

Logging. Structured logging of behavior plays a crucial role in any interactive system. However, logging is usually neglected leading to noisy and inconsistent data. For this reason, we developed a logging infrastructure specific to conversational WoZ experiments, inspired by previous work from web UI logging [5]. The logging module keeps track of messages sent and their associated timestamp as well as button clicks, wizard searched queries, action intervals, and their correlation to the produced messages. For conversational search tasks, this includes logging wizard interactions with queries, search results, and rewritten responses. This supports generating labeled data for response ranking, intent detection, conversational summarisation, NLG, and mixed-initiative.

4 PILOT STUDIES

TaskMAD was developed based on feedback and iteration from previous experience conducting studies with the earlier AD system. A study with 30 participants found that the quality of the conversations in AD suffered from a lack of task context and the wizard interface had limitations resulting in slow responses. Further, AD lacked key features and capabilities. These include dynamic search on external collections with support for editing that allows the wizard to rapidly generate task-specific and knowledge-grounded responses from multiple sources. Task-centric conversations required multimodal capabilities. Lastly, it needed support for text-to-speech to provide an engaging and more natural experience. The resulting TaskMAD system is currently being piloted in the Alexa Taskbot challenge as well as in collaboration with the University of Regensburg for undergraduate research.

5 CONCLUSION

In this demo, we describe TaskMAD, a new platform for multi-modal and task-specific data collection and open-domain search. TaskMAD aims to provide support for conducting large-scale Wizard-of-Oz experiments for knowledge-grounded information seeking. Moreover, TaskMAD addresses the need for context and task-specific datasets by providing a platform for experimentation and data collection with support for multi-modal task interactions, access to external text using search integration, scalable and modular user interfaces, and extensive structured logging functionality. TaskMAD represents a step towards the development of future interactive agent systems by supporting interactions that are not possible with existing systems on their own. It supports creating data for the next generation of rich multi-modal agents that are grounded in knowledge and real-world environments.

6 ACKNOWLEDGMENTS

We would like to thank David Elsweiler and Alexander Frummet from the University of Regensburg for their support and contributions during the development of this project. This work was supported in part by the Turing AI Acceleration fellowship and by the Amazon Research Award on Knowledge-Grounded Conversational Product Information Seeking. This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

[1] Adam Czyzewski, Jeffrey Dalton, and Anton Leuski. 2020. Agent Dialogue: A Platform for Conversational Information Seeking Experimentation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2121–2124. https://doi.org/10.1145/3397271.3401397

[2] Jeff Dalton, Mohammad Aliannejadi, Leif Azzopardi, Paul Owoicho, Johanne Trippas, and Svitlana Vakulenko. [n.d.]. TREC Conversational Assistance Track (CAsT). https://www.treccast.ai/. Accessed: 2022-02-21.

[3] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. CAsT-19: A Dataset for Conversational Information Seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1985–1988. https://doi.org/10.1145/3397271.3401206

[4] Anton Leuski. 2018 (accessed October 24, 2021). *Wizard of Oz tool.* ICT. https://nld.ict.usc.edu/woz/doc/.

[5] David Maxwell and Claudia Hauff. 2021. LogUI: Contemporary Logging Infrastructure for Web-Based Experiments. In *Proceedings of the 43th ECIR.* (In press).

[6] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476* (2017).

[7] Iadh Ounis, Giambattista Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier Information Retrieval Platform. *Lecture Notes in Computer Science* 3408, 517–519. https://doi.org/10.1007/978-3-540-31865-1_37

[8] Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020. Multi-Modal Open-Domain Dialogue. *CoRR* abs/2010.01082 (2020). arXiv:2010.01082 https://arxiv.org/abs/2010.01082

[9] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *J. Data and Information Quality* 10, 4, Article 16 (Oct. 2018), 20 pages. https://doi.org/10.1145/3239571

[10] Hamed Zamani and Nick Craswell. 2019. Macaw: An Extensible Conversational Information Seeking Platform. *CoRR* abs/1912.08904 (2019). arXiv:1912.08904 http://arxiv.org/abs/1912.08904

[11] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. *CoRR* abs/2201.08808 (2022). arXiv:2201.08808 https://arxiv.org/abs/2201.08808

[12] Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. *CoRR* abs/2002.04793 (2020). arXiv:2002.04793 https://arxiv.org/abs/2002.04793