Mia: A Web Platform for Mixed-Initiative Annotation

Jarvis Tse[1], Alex C. Williams[2,3], Joslin Goh[1], Timothy Player[3], James H. Brusuelas[4], Edith Law[1]

[1] University of Waterloo, [2] Amazon, [3] University of Tennessee, Knoxville, [4] University of Kentucky, Lexington
jarvis.tse@uwaterloo.ca, acwio@amazon.com, joslin.goh@uwaterloo.ca,
tplayer@vols.utk.edu, james.brusuelas@uky.edu, edith.law@uwaterloo.ca

Abstract

Despite recent advances in AI, the task of integrating machine learning models into interactive experiences remains tedious. This paper introduces *Mia*, a web-based platform that enables researchers to engineer real-time, interactive work experiences with machine learning models. Through *Mia*'s REST API and embedded web interface, *Mia* empowers human-AI interaction researchers to implement task interfaces that collaborative interface agents can both observe and contribute to in real time. To demonstrate *Mia*'s utility, we developed an interface agent that collaborates with crowdworkers on image annotation tasks and conducted a feasibility study. Among other findings, we discovered that the crowdworkers who were supposed to be more altruistic did not provide help to the interface agent more frequently, although they self-reported to feel better than other crowdworkers for doing so. We conclude by discussing other potential usages of *Mia*, such as facilitating hybrid intelligence systems.

Introduction

Human-artificial intelligence (AI) collaboration has become increasingly prominent in modern crowdsourcing contexts. Earlier uses of automation centered around improving annotation quality by means of intelligent task decomposition (Suzuki et al. 2016), task assignment (Roy et al. 2015), and task pricing (Dang and Cao 2013). Other approaches have explored how automation can be integrated into the task itself to reduce the workload for annotators, such as through interactive machine validation (Maninis et al. 2018). Research also suggests that annotators, facilitated by AI, may be significantly more accurate, but could be significantly biased in their annotation behavior as well (Fort and Sagot 2010; Rehbein, Ruppenhofer, and Sporleder 2009).

In this paper, we introduce *Mia*, a versatile mixed-initiative annotation platform. Also, we demonstrate *Mia*'s capability to support mixed-initiative research through a feasibility study, which explores mixed-initiative annotation as well as help-seeking/giving and observational learning.

Mia System Overview

Mia – as abbreviated from Mixed-Initiative Annotation platform – allows researchers to conduct experiments with mixed-initiative interface agents. Precisely, *Mia* is a web-based platform where researchers can create and deploy annotation tasks to be completed by human annotators and/or machine learning models in real time, and is implemented with MeteorJS, MongoDB, and React. *Mia*'s system architecture is displayed in Figure 1. In addition to the standard functionality that often comes with web-based platforms (e.g., user account management), *Mia* provides researchers with three key components: (1) application programming interface (API) hooks for activity observability; (2) an annotation interface for mixed-initiative annotation; and (3) an experimental task design module for designing and deploying mixed-initiative tasks. Each of the three key components will be discussed in detail below.

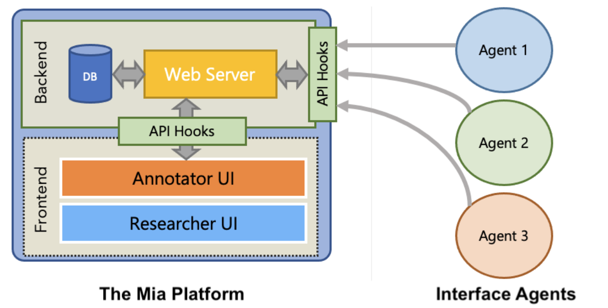

Figure 1: *Mia*'s system architecture. Agent applications are shown to illustrate how they interface with the system.

Mia establishes a rich set of web-based API hooks to synchronize teams of human and machine annotators for multiple ongoing tasks simultaneously. All API hooks operate on the Datagram Delivery Protocol (DDP). Collectively, *Mia*'s API hooks allow client applications and server applications not only to listen for activity, but also contribute to the activity taking place for each ongoing task. Additionally, *Mia*'s API hooks allow interfaces to send and query a variety of information relevant to each ongoing task such as annotation data, activity logs, and meta-information about the task interface (e.g., dimensions of the image being annotated).

Mia deploys its tasks for use with an annotation interface that automatically facilitates synchronization via the system's API hooks. As Figure 2 shows, the interface contains various components, such as an annotation surface where annotations are rendered, an annotation assignment tool (i.e.,

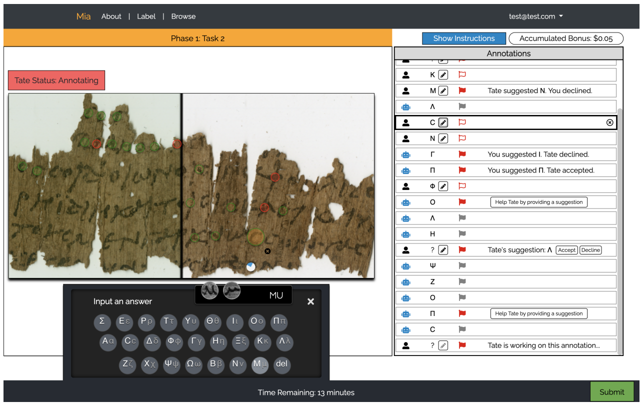

Figure 2: A screenshot of *Mia*'s task interface provided to human annotators.

a virtual keyboard) that human annotators can use to assign labels to annotations, as well as an annotation panel that displays annotation information and can be used by human annotators to select annotations.

Mia provides a set of user interfaces for designing and deploying mixed-initiative annotation tasks in three steps: (1) defining a project that includes an image set; (2) defining mixed-initiative task designs, such as the workload assigned to each annotator for each image (i.e., by dividing each image using a vertical line); and (3) defining experimental instrumentation use, such as by adding questionnaires (e.g., Google forms) to the project's workflow as well as collecting time-stamped telemetry information to keep track of user interface activity (e.g., annotation events).

Feasibility Study

We have conducted a feasibility study by coupling *Mia* with an interface agent named *Tate*. Below, we discuss *Tate*'s design, as well as the feasibility study's design and results.

Design of *Tate*

Tate is an interface agent implemented as a standalone NodeJS application that uses the SimpleDDP JavaScript library to interface with *Mia*'s DDP-based API hooks. Architecturally, *Tate* is sub-divided into two components: (1) an Activity Scheduler that plans *Tate*'s actions and (2) a Message Handler that observes the state of each ongoing task through *Mia*'s API hooks and routes the states to appropriate endpoints in the Activity Scheduler.

Tate's intelligence is simulated. *Tate* is capable of annotating images as well as seeking suggestions from and providing suggestions to human annotators (i.e., by asking and recommending what specific annotations' labels should be). *Tate* may refuse to provide suggestions, such as when it believes an annotation is not located near a letter. *Tate*'s accuracy is configured to be at least 80% in all cases and is significantly more accurate than the pilot study participants. This design decision aims to allow the feasibility study participants to learn from *Tate* by observation.

Feasibility Study Design

Participants were recruited from Upwork. All participants received the same ten ancient Greek papyrus manuscript images to annotate and were randomly assigned to one of the following conditions: (1) annotating five images individually and annotating five images alongside *Tate* while being able to seek suggestions from and provide suggestions to *Tate*; (2) annotating five images individually and annotating five images alongside *Tate* but could not interact with *Tate* except by observing *Tate*'s annotations; and (3) annotating all ten images individually. For each participant, the order that the ten images appeared and whether they worked individually or alongside *Tate* first were randomized, when applicable. We divided each image and all participants annotated the same part of each image. The participants could submit their work any time before or at the time limit. The time limit was enforced to ensure that the whole study did not significantly exceed one hour for each participant. Moreover, questionnaires were used to collect information before, during, and after the participants' annotation tasks.

Feasibility Study Results

Study data were collected from seventeen participants for each of the three conditions. By analyzing the data, we have found that (1) among the participants who had the opportunity to provide help to *Tate*, their altruism scores, as measured using a modified version of a self-report altruism scale (Rushton, Chrisjohn, and Fekken 1981), did not correlate to how frequently they provided help to *Tate* ($p = .88$), although their altruism scores did positively correlate to their agreement levels to the statement "I felt good when I would provide help to *Tate*." ($p = .03$); (2) the participants who worked alongside and were able to seek help from and provide help to *Tate* were significantly more accurate when they self-reported to feel more productive ($p = .03$); and (3) participants who annotated alongside *Tate* but could interact with *Tate* by only observing its annotations annotated significantly faster than the remaining participants ($p = .04$).

Conclusion & Future Work

This paper proposes *Mia*, a mixed-initiative annotation platform. Through a feasibility study, we have demonstrated that *Mia* can be coupled with a standalone AI agent to be used for mixed-initiative annotation research. In the future, *Mia* can be implemented with various machine learning models as well as extended to support more types of annotation tasks (e.g., video annotation tasks with bounding boxes), human-AI interactions (e.g., different help mechanisms), and annotator teams (e.g., human teams and hybrid intelligence teams in which the human complements the AI annotator).

Acknowledgements

We thank Ken Jen Lee, Sangho Suh, and Jessy Ceha from the Augmented Intelligence Lab at the University of Waterloo for their feedback regarding our paper. We also thank the participants of the pilot and feasibility studies, including members of the People, Agents, Interactions, and Systems research group at the University of Tennessee, Knoxville.

References

Dang, K.-H.; and Cao, K.-T. 2013. Towards reward-based spatial crowdsourcing. In *2013 International Conference on Control, Automation and Information Sciences (ICCAIS)*, 363–368. IEEE.

Fort, K.; and Sagot, B. 2010. Influence of pre-annotation on POS-tagged corpus development. In *The fourth ACL linguistic annotation workshop*, 56–63.

Maninis, K.-K.; Caelles, S.; Pont-Tuset, J.; and Van Gool, L. 2018. Deep Extreme Cut: From Extreme Points to Object Segmentation. In *Computer Vision and Pattern Recognition (CVPR)*.

Rehbein, I.; Ruppenhofer, J.; and Sporleder, C. 2009. Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, 19–26.

Roy, S. B.; Lykourentzou, I.; Thirumuruganathan, S.; Amer-Yahia, S.; and Das, G. 2015. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal*, 24(4): 467–491.

Rushton, J. P.; Chrisjohn, R. D.; and Fekken, G. C. 1981. The altruistic personality and the self-report altruism scale. *Personality and individual differences*, 2(4): 293–302.

Suzuki, R.; Salehi, N.; Lam, M. S.; Marroquin, J. C.; and Bernstein, M. S. 2016. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2645–2656.