Lazy Human AI teams

vihang@umich.edu, smendke@umich.edu

Introduction

In today's world when AI systems have begun to dominate industries and organizations, it becomes necessary to think about the role such systems will play in a society largely centered around humans. There is no doubt about the positive impact of data driven methodologies in our society with AI as a tool already revolutionizing technology and leaving its mark in the fields of healthcare, transportation, agriculture, weather forecast and many more. Yet the promise of AI offers much more, being deployed as a team member to support human decisions and achieving performances in Human-AI teams towards a shared goal that neither an AI nor the human could achieve alone. As Human-AI teams become increasingly important to assist human decision making, we try to study such teams and factors that influence their development through shared experiences.

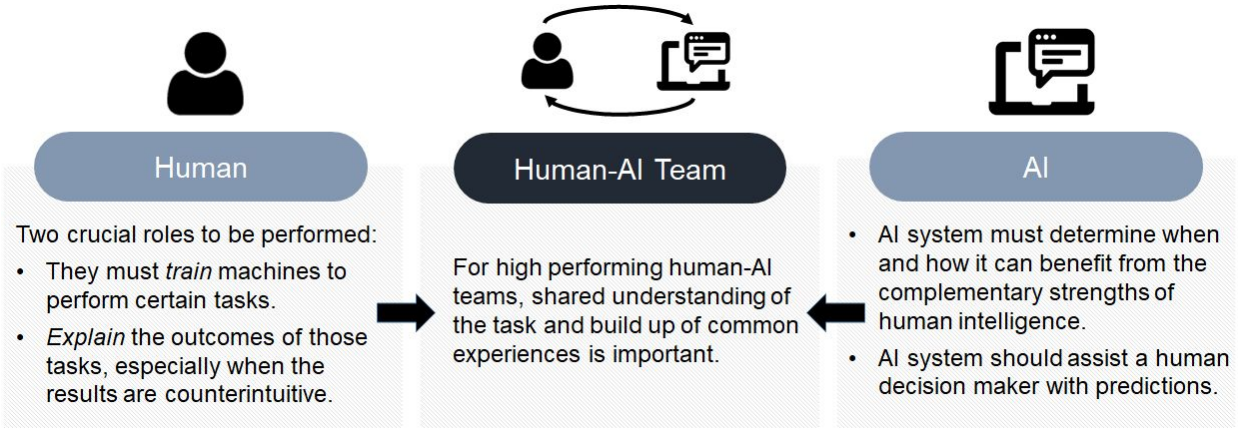

Figure 1: Crucial role of humans and AI in Human-AI teams.

Often when research talks about Human-AI teams, these teams are studied with the view that humans are oracles and focus on accounting for AI mistakes, explaining its error boundaries so that humans can develop better insights into the functioning of AI systems and decide when to trust the agent and when to not. Little focus is given to the human component of such teams. We believe it to be equally important to account for human factors while designing AI systems for optimal teams. We are inspired by this idea in psychology that humans are competent by nature and our mental effort is parsimonious. Humans are also prone to errors as tasks become more and more cognitively expensive. Thus we are motivated to build lazy human-AI teams that can perform better overall. To study this idea, we explore an important human factor namely human effort [Table 1]. As accuracy is mostly used in literature as a criteria for performance, we evaluate our results considering human effort to be an additional task dimension alongside model accuracy. Finally the metric, overall team performance is defined to account for both accuracy and effort in a single measure of success. Effective development of teams requires us to be clear about how humans can most effectively augment machines, how machines can enhance what humans do best, and how to redesign processes to support this partnership. Our study offers a small directional step towards this goal.

Metric	Definition
Model Accuracy (Acc)	Measured on a held out dataset at each iterative step with ground truth labels annotated by experts.
Human Effort (HE)	Defined as being proportional to the time (s) it takes the user to complete all annotations. Average human effort is this average for single annotation.
Overall Team Performance (OTP)	Weighted sum of human effort and achieved model accuracy. (weights were decided such that both task dimensions are at the same scale) $$OTP = Acc/100 + 1000/HE$$ *(higher is better)

Table 1: Definitions of metrics used and evaluated throughout this report.

Platform for Studying Human-AI Teams

Given the similarity of related work in active learning with our proposed goal and the notion of effort, we consider the task of text annotation for understanding Human-AI teams. Currently, there exist three common annotation tools used in the market, with prodigy being the closest to our considerations. Although they consider the idea of reducing mental effort, they don't account for it as a model parameter but we do. Here, the shared goal of the team is to achieve the maximum possible accuracy on an unknown test dataset. The updates to the model occur in batches. These updates in active learning could be summarized as follows:

a) Collect initial training data T1 and train a model h on T1.
b) Collect additional data to create T2 using the sampling strategy, where T1 $\subset$ T2.
c) Train h on T2 and repeat until stopping criteria is met.

A very simple interface was designed to carry out our experiments. We iteratively improved upon our design for the interface to provide maximum fluidity while annotating text samples. Refer to the appendix for the task workflow and details about our interface.

Dataset: The data samples used in this study are open ended dialogue utterances that come from the switchboard corpus (a dataset of telephonic conversations between people). The original dataset consists of about 200,000 utterances to be annotated into 41 possible dialogue sets. We derive a feasible subtask sampling about 500 utterances for training to be annotated into 4 possible dialogue acts: Answer, Paraphrase, Statement or Opinion. For text, it is intuitive to realize that there is a strong correlation between text sample length and the mental effort [Figure 2] it would require to annotate that sample. To account for this factor, all our datasets were uniformly sampled to contain an equal distribution of short, medium and long length text samples. The created dataset is class imbalanced with 40% samples being statements and the rest being uniformly sampled in alignment with the original dataset. Along with training data, we use 30 text samples as our test data and 30 text samples with annotated ground truth effort estimates for development of the study. This is a microtask that comes naturally to humans and could potentially be viable to analyse Human effort in complex tasks as well.

Experiments

Human Baseline: How do we allow a model to account for human efforts and what could a possible reference for effort estimation of data samples be? We collected a small ground truth dataset of 30 text samples from 3 human annotators to collect data on effort estimation and get a human baseline of achievable accuracy on this task. Final effort is the average over all annotations. For analysis, we also bin the effort into three broad categories: low (< 8s), medium (>= 8s, < 16s) and high (>= 16s) and assign training examples to each category using a k-nearest neighbor approach.

Task Dimensions	Accuracy (%)	Average Human Effort (s)
Human Annotations	52.00 ± 2.00	12.05 ± 0.45

Table 2: Baseline human accuracy and effort estimates

Baseline Experiments: We use active learning as the baseline team performance. We conducted experiments with 3 sampling strategies: Least confidence sampling, margin sampling and maximum entropy sampling along with random sampling of data. The final results are presented for the best performing maximum entropy sampling strategy. To get insights into the hybrid nature of human-AI teamwork we attempt to answer two research questions:

1) Can accounting for human factors, specifically the notion of human effort help build better human-AI teams? And at what scale should this notion be considered?

The machine learning community has long studied the paradigm of active learning, that accounts for effort on the scale of number of training examples. Although this works well for models trained on homogenous datasets, we argue that as agents become more general, our tasks and datasets will keep changing to heterogeneous forms and effort needs to be accounted for at the scale of each data sample.

To answer this question, the following study was conducted: Using the k-nearest neighbor approach, each training sample was assigned an effort value from the effort ground truth dataset. Rather than the model centric approach of active learning (where annotation is collected for samples the model is least confident about), a user centric approach (the model tries to reduce the effort human has to invest) is considered for sampling. Initial experiments with naive sampling of data points according to increasing effort values did not result in high performing teams. Instead we consider random sampling from each of the three broad effort categories (from low to high effort) i.e., the user is presented with random data samples from an effort category and we do not sample data from a higher effort category until annotations for all points within the lower effort category are collected. We observed such a strategy to boost the overall team performance reducing required human effort to be invested by 3x while achieving the same model accuracy. We believe this is the case as low effort annotations are cognitively less complex, and users can provide more accurate annotations. This inturn leads to faster model convergence for the task. Although this is not sufficient to draw concrete conclusions, there is preliminary evidence that optimizing for human effort with model accuracy can lead to significant gains. We leave an analysis on multiple datasets with more users as future work.

Learning	Accuracy (%)	Human Effort (s)	OTP	Average Human Effort
Active	51.85	4635.36	0.7345	13.32
Random	51.58	6503.33	0.6715	13.56

Table 3: Evaluation of baseline models. To ensure validity, average over 5 experiments is presented.

Display	Accuracy	Human Effort (s)	OTP	Average Human Effort
w accuracy	52.10	1239.6	1.3264	10.33
w/o accuracy	51.85	1239.6	1.3249	10.33

Table 4: Evaluation for the proposed experiment. To evaluate our design choices for the interface, we conduct two experiments, one “with displaying” current model accuracy to the user and one without (*the experiment “with displaying” accuracy is for a single user)

2) If such a paradigm is indeed useful, does an AI model optimized for human factors influence the users mental model in any way? Research in human teams have demonstrated that humans develop shared mental models to improve coordination and effectively achieve high performance in tasks. Analogously, understanding of mental models is also necessary to improve overall Human-AI team performance. Mental models are a useful construct for humans understanding about the AI system and are related to developing trust and making these systems more interpretable for us. For effective team functioning it is important to have a shared understanding of the task and further build up on the common experiences thereby resulting in high performing Human-AI teams. It’d be interesting to see the impact of our studies on development of such mental models.

We propose this study as follows: Along with optimizing for required effort in human annotation, allow the user to skip certain examples or provide annotations. The second version of our interface was developed accordingly. At each iteration the user is displayed with the current model accuracy and the expected effort from him in annotating that particular example. Evaluating the overall team performance in such a setting will allow us to understand user expectations of the model and help reduce human mistakes when the user is not confident about his annotations thus providing for better human-AI teams. We were not able to complete this study given the time constraints and leave it as future work.

Conclusion

We presented a study analysing human-AI teams from the perspective of optimizing for invested human effort during the learning process. Initial experiments seem to indicate that annotation tasks with varying annotation difficulty for data samples can benefit greatly by enabling models to account for human factors, specifically human effort. This study also provides possible research directions to be able to realize the full potential of Human-AI teams.

References

[1] Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance
[2] Learning to Maximize Accuracy vs. Effort in Interactive Information Retrieval
[3] Measuring Mental models
[4] The Impact of Multimodal Communication on a Shared Mental Model, Trust, and Commitment in Human–Intelligent Virtual Agent Teams

Appendix

Interface Design

Figure 2: Interface design

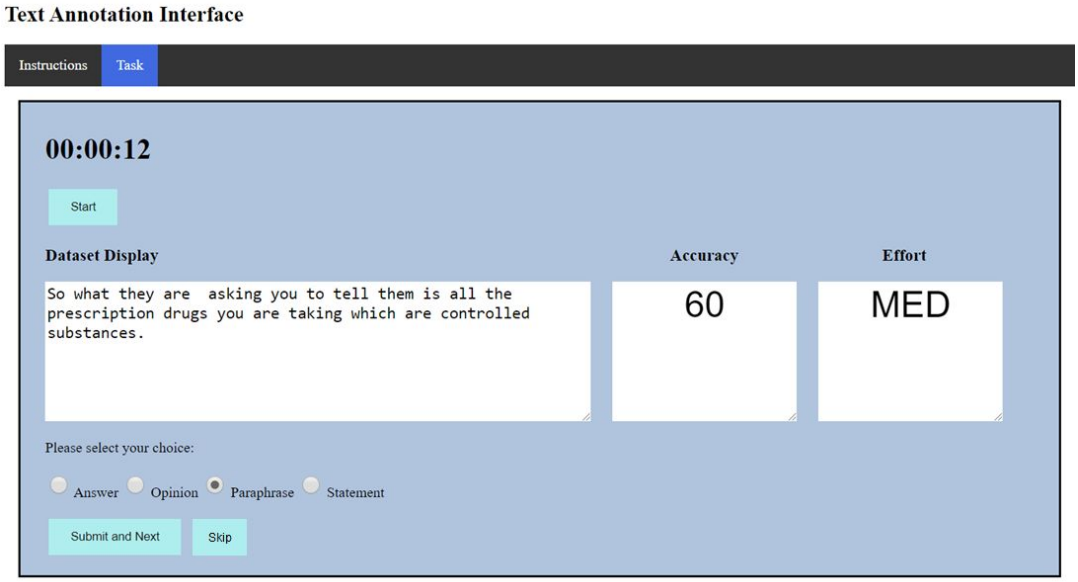

Figure 3: Interface design displaying model accuracy and expected effort

We designed two interfaces to perform our studies. The first interface design enables the human annotator to classify open ended dialogue by selecting one of the four possible dialogue acts. The second interface displays model accuracy and expected effort for human annotator's reference and allows them to either classify the dialogue act by selecting an appropriate intent or skip it, indicating agreement with model prediction. We were only able to complete our study for the first proposed research question that uses first interface (figure 2).

Workflow

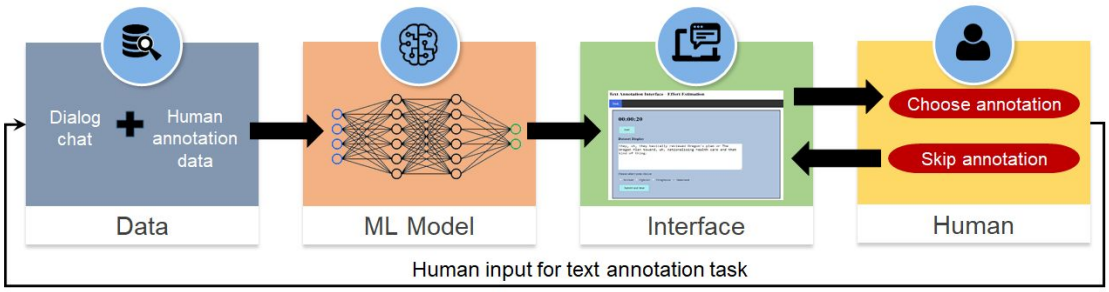

Figure 4: Workflow showing our platform for studying Human-AI Teams

Data Analysis

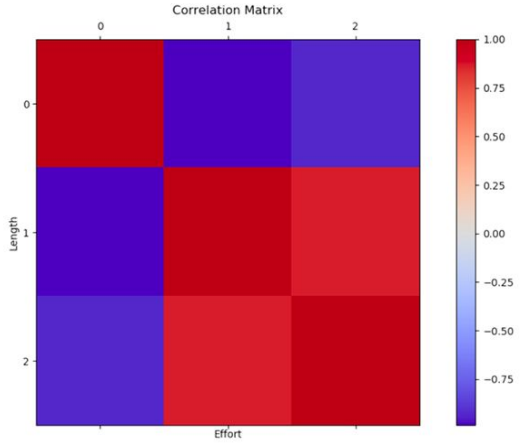

Figure 5: Correlation between text sample length and Human effort invested for the task

	Low	Medium	High
Statement	0.180	0.810	0.000
Answer	0.375	0.500	0.125
Opinion	0.125	0.500	0.375
Paraphrase	0.167	0.500	0.333

Table 5: Effort probabilities for each class in the dataset

Test curves for the experiments

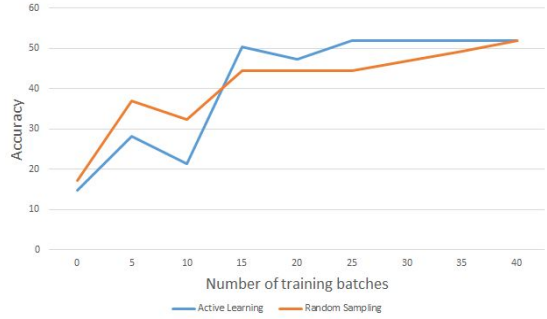

Figure 6: Baseline experiments

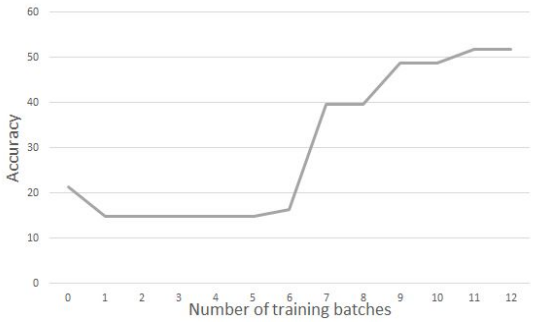

Figure 7: Proposed experiment