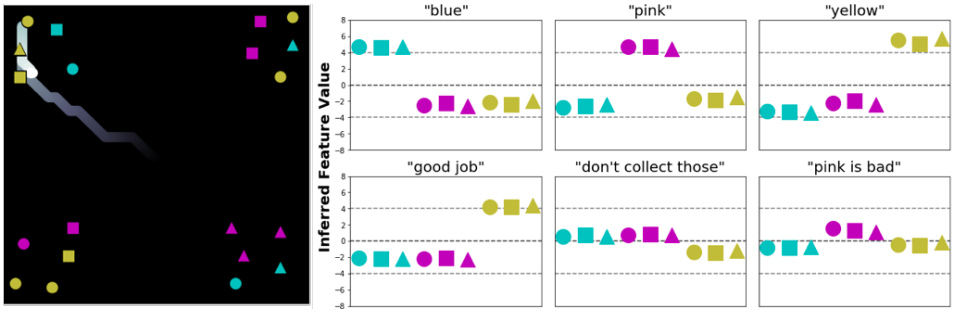

Figure 4: Left: A trajectory from our experiment. Right: Inference network output given this trajectory. Top row: the model learns to map feature-related tokens ("Descriptive" feedback) directly to rewards, independent of the trajectory. Bottom left / center: the model maps praise and criticism ("Evaluative" feedback) through the feature-counts from the trajectory. Bottom right: a failure mode. "Descriptive" feedback with negative sentiment, a rare speech pattern, is not handled correctly.

6.2 Learning from Different Forms of Feedback

To characterize model learning from different "forms" of feedback, we design a second evaluation independent of the experiment structure. Our "episode" sequence is as follows: we draw a (u, τ) tuple at random from the human-human experiment, *update* each model, and have it *act* on our 100 pre-generated test levels. We take its averaged normalized score on these levels. We repeat this procedure 5 times for each cross-validation fold, ensuring the learned model is always tested on its hold-out teachers and rewards. This draws feedback from a variety of teachers and tests learners on a variety of level configurations, giving a picture of overall learning trends. Normalized scores over test levels are shown in Fig. 3 and Table 2 ("Interaction Sampling"). All models improve when learning from the entire corpus ("All") versus individual teachers ("Experiment"). The inference network improves most dramatically, suggesting it may be vulnerable to idiosyncratic communication styles used by individual teachers. We then use our feedback classifier (Section 5.1) to expose models to only a single form of feedback. This reveals that our "pragmatic" augmentations help most on "Descriptive" feedback, which is critical for early learning in the experiment. Finally, we explore our inference network's contextualization process (Fig. 4). It learns to map "Evaluative" feedback through its prior behavior and typical "Descriptive" tokens directly to the appropriate features. We also confirm failure modes on rarer speech patterns, most notably descriptive feedback with negative sentiment. This suggests the learned approach would benefit from more data.

7 Conclusion

We presented two methods to recover latent rewards from naturalistic language: using aspect-based sentiment analysis and learning an end-to-end mapping from utterances and context to rewards. We find that three implementations of these models all learn from live human interactions. The "pragmatic" model in particular achieves near-human performance, highlighting the role of implicature in natural language. We also note that the inference network's performance varies qualitatively across evaluation modes: it outperforms the "literal" model when tested on the whole corpus, but ties it when playing with individual humans ("Interaction Sampling" vs "Experiment - Live"). This underscores the importance of evaluation in realistic interaction settings. We see several future research directions. First, our sentiment models could be improved via theory-of-mind based pragmatics, while our end-to-end approach could benefit from stronger language models (recurrent networks or pre-trained embeddings). Hybridizing sentiment and learned approaches (**??**) could offer the best of both. We also see potential synergies with instruction following: treating commands as "Imperative" feedback could provide a preference-based prior for interpreting future instructions. Finally, we anticipate extending our approach to more complex MDPs in which humans teach both rewards and transition dynamics (**?**). In general, we hope the methods and insights presented here facilitate the adoption of truly natural language as an input for learning.

Acknowledgements

We thank our anonymous reviewers for their thoughtful and constructive feedback. This work was supported by NSF grants #1545126 and #1911835, and grant #61454 from the John Templeton Foundation.

Ethics Statement

Equipping artificial agents with the capacity to learn from linguistic feedback is an important step towards value alignment between humans and machines, with the end goal of supporting beneficial interactions. However, one risk is expanding the set of roles that such agents can play to those requiring significant interaction with humans – roles currently restricted to human agents. As a consequence, certain jobs may be more readily replaced by artificial agents. On the other hand, being able to provide verbal feedback to such agents could expand the group of people able to interact with them, creating new opportunities for people with disabilities or less formal training in computer science.

Modi voluptatibus dolor, vel similique incidunt excepturi quisquam, libero cumque saepe.Eaque iusto inventore quod at error assumenda, quod inventore recusandae quibusdam rem unde harum

ipsum.Ea dicta culpa neque aliquid nulla dolore impedit, amet minus beatae sunt rem?