Toward an Interleaved Model of Actions and Words in Social Simulation (Extended Abstract)

Jeff Orkin
MIT Media Laboratory
75 Amherst Street, E14-574M
Cambridge, MA 02139
+1 (617)253-1908
jorkin@media.mit.edu

Deb K. Roy
MIT Media Laboratory
75 Amherst Street, E14-574M
Cambridge, MA 02139
+1 (617)253-0596
dkroy@media.mit.edu

ABSTRACT
We present results of semi-automated annotation of dialogue data collected from 5,200 gameplay logs from *The Restaurant Game* (http://www.theRestaurantGame.net), and show that classified dialogue acts can function effectively as a common currency for modeling behavior with interleaved physical actions and words.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language parsing and understanding.*

General Terms
Measurement, Performance, Design, Reliability, Experimentation, Human Factors, Languages, Verification.

Keywords
Social simulation, Modeling natural language, Virtual Agents, Agents in games and virtual environments.

1. INTRODUCTION
Current approaches to implementing natural language dialogue systems in use in the video game industry are labor intensive, requiring designers to anticipate human input and hand-author responses. The dramatic increase in popularity of online games provides an opportunity to teach machines by observing human gameplay, by mining the enormous amount of data generated by thousands (or even millions) of human-human interactions. How can we maximize the utility of this data for agents to exploit at runtime, while minimizing the human labor required to structure and annotate corpora?

We are working toward a long term goal of generating dialogue and behavior for agents based on data collected from human-human interactions. Our approach, influenced by Schank, is to represent context in the form of socio-cultural scripts [3]. Due to the technological limits of the 1970s, Schank's scripts were hand-crafted, and thus subject to limits in the range of behavior that human scripters can possibly anticipate. Hand-crafted scripts are brittle in the face of unanticipated behavior, and are unlikely to cover appropriate responses for the wide range of behaviors exhibited in an open ended restaurant environment. Today, we have the opportunity to do better by *discovering* scripts from human-human interaction traces of online gameplay.

Cite as: Toward an Interleaved Model of Actions and Words in Social Simulation (Extended Abstract), Jeff Orkin, Deb K. Roy, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

Figure 1. Screenshot from *The Restaurant Game.*

We present results of semi-automated annotation of dialogue data collected from *The Restaurant Game* [1,2]. This work illustrates that classified dialogue acts can function effectively as a common currency for modeling interleaved actions and words, given a domain-specific annotation scheme which classifies both illocutionary force and associated propositional content. Our results demonstrate how annotating 2% of the log files from a 5,200 game corpus can produce statistical models of dialogue act sequences with predictive power that outperform models of raw utterance sequences, or utterances clustered by an unsupervised system. High-precision dialogue acts can be integrated cleanly into a model of physical action sequences, preserving the predictive power of the original model of physical actions alone.

2. THE RESTAURANT GAME
We designed *The Restaurant Game* to serve as both a data collection device, and a target platform for simulation of social behavior generated from the human data. Players are anonymously paired online to play the roles of a customer and waitress in a 3D virtual restaurant. Players can move around the

environment, type open ended chat text, and manipulate 47 types of interactive objects through a point-and-click interface. Every object provides the same interaction options: pick up, put down, give, inspect, sit on, eat, and touch. To date, 13,564 people have played *The Restaurant Game*, from which we have collected 9,433 log files of completed two-player games. This paper describes work with a subset of 5,200 game logs. An average game takes about 10-15 minutes, and consists of 84 physical actions, and 40 utterances with an average length of four words each. Player interactions vary greatly, ranging from games where players dramatize what one would expect to witness in a restaurant, to games where players fill the restaurant with cherry pies. We have demonstrated that when immersed in a familiar environment, enough people do engage in common behavior that it is possible for automatic system to learn valid statistical models of typical behavior and language [1].

3. DIALOGUE ACT CLASSIFICATION

We randomly selected 100 game logs from our corpus of 5,200 logs to serve as training data for an SVM-HMM classifier, and we annotated these logs by hand. Each line of dialogue is classified along three axes: *Speech Act*, *Propositional Content*, and *Referent*. Speech Acts categorize utterances by illocutionary force (e.g. question, directive, assertion, greeting, etc.), Propositional Content describes the functional purpose of the utterance, and Referent represents the object or concept that the utterance refers to. Our three labels are combined into a {speech act, content, referent} triple that serves as an abstraction allowing utterances to be clustered semantically, rather than by surface forms, and greatly compresses the space of possible dialogue acts. Evaluation results with 10 fold cross validation show that we classify {speech act, content, referent} with respective accuracies {77.3%, 75.3%, 81.1%}, and we computed kappa coefficients of {0.73, 0.70, 0.89} with another human annotator.

4. PREDICTIVE MODEL EVALUATION

Our long term goal is to generate social behavior for agents based on models learned by observing human-human interactions. These models will allow agents to predict future actions (physical and linguistic) based on recent observations. As a first exploration in this direction, we experimented with simple n-gram statistical models applied to both the surface word level and the speech act "intentional level." We evaluate our dialogue act classification quantitatively by learning three separate dialogue models – based on (1) classified utterances, (2) raw utterances, and (3) automatically clustered utterances – and comparing the predictive power provided by these models. We first compare n-gram models of classified dialogue acts to models of fully automated utterance abstractions that do not incorporate human interpretation (raw and automatically clustered). Next, we integrate these dialogue models with a model of physical interaction, in order to evaluate how these utterance abstractions function as a common currency with physical actions.

Raw utterances yield poor prediction accuracy for n-gram models for all values of *n*, only achieving above 0.1 for bigrams. The k-means algorithm clustered utterances based on the Euclidean distance between feature vectors of unigrams, bigrams, and trigrams observed within the utterances. While clusters do achieve about a 30% increase in prediction accuracy over raw utterances, they fall below that of dialogue acts by over 50%.

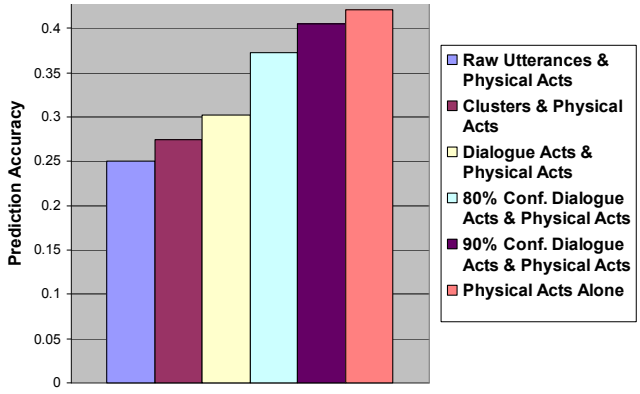

Figure 2. Effect of interleaving physical acts with utterances.

In Figure 2, we illustrate the effect on prediction accuracy of integrating utterance abstractions into a trigram model of physical interaction. The integrated model predicts the next action *or* utterance based on recent observations of interleaved actions and utterances. We find that interleaving physical actions with dialogue acts gives better prediction accuracy than with raw utterances or clusters, and if we filter to only include dialogue acts with at least 90% confidence we can achieve a prediction accuracy negligibly lower than that of physical acts alone (0.41 vs. 0.42), demonstrating that dialogue acts function well as a common currency with physical acts. This is a first step providing a foundation to begin discovering sub-goals, composed of interleaved sequences of actions and utterances.

5. CONCLUSION

Behavioral models generated by observing players of online games and virtual worlds have the potential to produce interactive socially intelligent agents more robust than can be hand-crafted by human designers. While it is possible to automatically learn statistically recurring patterns in surface level behavior, our results demonstrate that we can generate models with stronger predictive power by leveraging a minimal amount of human interpretation to provide annotation of the underlying intentions, in the form of dialogue act triples. The significant increase in predictive power with dialogue acts is evidence of progress towards discovering the socio-cultural scripts that guide social interaction in a restaurant.

6. REFERENCES

[1] Jeff Orkin and Deb Roy. The Restaurant Game: Learning social behavior and language from thousands of players online. *Journal of Game Development*, 3(1), 39-60, 2007.

[2] Jeff Orkin and Deb Roy. Automatic Learning and Generation of Social Behavior from Collective Human Gameplay. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems,* Budapest, Hungary, 2009.

[3] Roger C. Schank and Robert P. Abelson. Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. Lawrence Erlbaum Associates, 1977.