Semi-Automated Dialogue Act Classification for Situated Social Agents in Games

Jeff Orkin, Deb Roy,

MIT Media Laboratory, 75 Amherst Street,
Cambridge, MA, USA 02139
{jorkin, dkroy}@media.mit.edu

Abstract. As a step toward simulating dynamic dialogue between agents and humans in virtual environments, we describe learning a model of social behavior composed of interleaved utterances and physical actions. In our model, utterances are abstracted as {speech act, propositional content, referent} triples. After training a classifier on 100 gameplay logs from *The Restaurant Game* annotated with dialogue act triples, we have automatically classified utterances in an additional 5,000 logs. A quantitative evaluation of statistical models learned from the gameplay logs demonstrates that semi-automatically classified dialogue acts yield significantly more predictive power than automatically clustered utterances, and serve as a better common currency for modeling interleaved actions and utterances.

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

While mature graphics hardware, rendering engines, physics simulators, and path planners have leveled the playing field for near-photorealistic visuals in video games and simulations, artificial intelligence methods for social planning, interaction, and communication are poised to take the lead as the differentiating feature in games of the future. Though much progress has been made in navigation and action selection, natural language communication between agents remains a difficult problem, and communication between agents and humans even more so. Dynamic interactive dialogue poses numerous technical challenges, yet also holds the key to enabling

Fig. 1. Screenshot from *The Restaurant Game.*

entirely new genres of games, and broadening the reach of games beyond entertainment into new forms of social simulation.

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. As von Ahn has demonstrated by labeling images with *The ESP Game* [20], the dramatic increase in popularity of online games provides an opportunity to teach machines by observing human gameplay. Multiplayer role-playing games and virtual worlds provide the opportunity for a potentially better approach to developing systems that can understand and generate natural language dialogue, by mining the enormous amount of data generated by thousands (or even millions) of human-human interactions. For example, as of 2008 *World of Warcraft* had over 10 million paying subscribers, *Habbo Hotel* had 9.5 million unique monthly visitors, and *Second Life* had 600,000 unique monthly visitors [4]. Clearly there is an opportunity to collect rich new forms of behavioral data from these players. What is less clear is how to 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 [16]. Due to the technological limits of the 1970s, Schank's scripts were hand-crafted, and thus subject to limitations associated with human authoring. 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 scenario. Today, we have the opportunity to do better by *discovering* scripts from human-human interaction traces of online gameplay.

With these ideas in mind, we launched *The Restaurant Game* (http://theRestaurantGame.net) as a platform for collecting rich physical and linguistic

```
 WAITRESS: "welcome to our fine restaurant"
CUSTOMER: "thanks, it's just me tonight"
WAITRESS: "would you like the seat by the window?"
CUSTOMER: "sounds good"
WAITRESS: "follow me"
CUSTOMER SITSON chair3(Chair)
WAITRESS: "perhaps i should start you off with some water"
CUSTOMER: "that sounds good, can i check out a menu?"
WAITRESS: "sure thing, coming right up"
WAITRESS PICKSUP dyn1733(Menu)
WAITRESS GIVES dyn1733(Menu) TO CUSTOMER
CUSTOMER LOOKSAT dyn1733(Menu)
WAITRESS: "water please"
dyn1741(Water) APPEARS ON bar(Bar)
WAITRESS PICKSUP dyn1741(Water)
WAITRESS PUTSDOWN dyn1741(Water) ON table1(Table)
WAITRESS: "here's your water"
WAITRESS: "i'll give you a minute to look over the menu"
```

Fig. 2. Transcript from a typical interaction in *The Restaurant Game*.

interaction [14]. To date we have collected over 9,400 game logs. The ultimate goal is to replace one of the players of *The Restaurant Game* with an automated agent that has knowledge of possible restaurant scripts at both a surface behavioral and linked deep intentional level. The agent will use this knowledge to guide its interpretation of the other human players' actions, and to plan its own physical actions and utterances while participating in the joint activity of completing a restaurant meal. As a step towards this goal, we address the problem of mapping surface forms of dialogue acts to underlying intentions and evaluate the quality of the resulting model in its ability to predict human dialogue acts at the intentional level.

Previously, we have demonstrated automating agents with data collected from *The Restaurant Game* [15]. In this first iteration of the system, dialogue between agents exhibited frequent non-sequiturs and incorrect responses, due to imitating human-human dialogues relying solely on matching surface forms of utterances. The system had no means of recognizing utterances with unique surface forms but the same semantic function, or utterances with the same surface form but different contextually-dependent semantic functions. There was no representation of the *intent* behind the humans' words. In contrast, players use a point-and-click interface to engage in physical interaction (e.g. highlight the dish and click the PickUp button on a pop-up menu). These interface interactions explicitly represent the player's intent. While the automated system constrained the agents' physical behavior based on learned patterns of these intentional actions, no analogous model existed for sequences of utterances. Ideally, a single model could capture patterns of interleaved physical actions and utterances. A precursor to learning such a model is an intentional representation of utterances that can be interleaved cleanly with physical actions.

In this paper we present results of semi-automated annotation of dialogue data collected from *The Restaurant Game*. We demonstrate 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. We describe a dialogue act classifier that we have trained on 100 log files, and leveraged to automatically annotate the

remaining 5,100. 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 based on surface forms. Finally, we show that 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. While our current system filters out some percentage of utterances from the interleaved model, this is a step toward a solution, and the percentage of included utterances will increase in the future as the classifier improves.

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 two-player games. This paper describes work with a subset of 5,200 game logs. A game takes about 10-15 minutes, and an average game 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. While many players do misbehave, 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 [14].

3 Related Work

We are working toward learning an interleaved model of actions and utterances in an everyday social situation, based on a large corpus of human-human interactions. Here we relate our work to previous research on dialogue modeling and learning from human data, and highlight significant differences.

Gorin et al [6] describe a system that learns to route calls in response to the prompt "How may I help you?", by finding mutual information between routing decisions and n-grams in human speech. Satingh et al [18] developed a dialogue management system that uses reinforcement learning to learn an optimal policy for a phone-based information system about activities in New Jersey based on interactions with human callers. Huang et al [8] trained chatbots by extracting title-reply pairs from online discussion threads. Our work differs from these projects by collecting data from humans *situated* in a (virtual) physical environment, where players dramatize an everyday scenario through a combination of (typed) dialogue and physical interaction,

contributing to learning an interleaved model of actions and utterances, representing a commonsense script of restaurant behavior. Huang's work may point toward an interesting direction for future work; incorporating knowledge extracted from external sources into our model.

McQuiggan and Lester [13] applied a similar methodology to ours (capturing demonstrations between humans in a game environment) to learn models of empathetic behavior, including gestures, posture, and utterances. Their work did not focus on learning open-ended natural language dialogue, and instead incorporated pre-recorded utterances. Gorniak and Roy [7] collected data from pairs of players solving puzzles in *Neverwinter Nights*, and constructed a plan grammar, which could be used to understand utterances between players. Similarly, Fleischman and Hovy [5] leveraged a task model of the game-based Mission Rehearsal Exercise to understand natural language input. In these projects, hand-constructed models of the situation (the plan grammar or task model) helped the system understand language. In contrast, we are training a classifier to understand language, and using classified utterances as building blocks to *learn* the situation model. While our data collection methodology is similar to previous work, we are working toward learning the structure of the situation from data, based on semi-automated annotation and automatic recurrence analysis, rather than hand-crafting a plan grammar or task model. Learning the structure has the potential of producing a more robust model through a less laborious process.

4 Dialogue Act Classification

Players of *The Restaurant Game* communicate by freely typing chat text to one another. While we can capture every utterance transmitted, there is no explicit representation of the intent behind the words of the player. In contrast, players use a point-and-click interface to engage in physical interaction (e.g. highlight the dish and click the PickUp button on a pop-up menu). These interface interactions explicitly represent the player's intent. Human annotation is required to transform utterances into functional units that share a common currency with physical actions – atomic units with explicit representations of intent and semantic function. Unfortunately, human annotation is expensive; it is infeasible to annotate a corpus of thousands of game logs, let alone millions. In this section we describe our approach to semi-automating annotation.

We randomly selected 100 game logs from our corpus of 5,200 logs to serve as training data for a classifier, and we annotated these logs by hand. Each utterance 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.

Labels in the Speech Act axis are similar in function to those found in the widely used DAMSL annotation scheme [2]. Devising our own annotation scheme was necessary in order to also incorporate propositional content and referents, which will be critical to an interactive agent. It is not enough to recognize that an utterance is a

question or directive, the agent needs to understand what it is a question *about* (a problem with the bill, or the desire to see a menu) or a directive to *do* (prepare a steak, or have a seat at a table). Section 5.4 provides quantitative evidence that the inclusion of propositional content and referents maximizes the range of confidently recognized utterances while preserving an agent's ability to predict future actions and utterances based on recent observations.

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. Below we provide the details of our annotation scheme, feature selection, classifier implementation, and classification results.

4.1 Human Annotation

It took one of the authors 56 hours to annotate all 4,295 utterances (of average length four words) observed in 100 games. We developed the list of annotation-labels during the course of annotation. The Speech Act labels were based on Searle's speech acts [17], expanded with Propositional Content and Referent labels to cover the range of utterances frequently observed in our corpus.

All three axes include an OTHER label, applied to utterances that fall outside the scope of typical restaurant conversation, such as nonsense, gibberish, and discussion of the players' personal lives. We applied a NONE label to the Propositional Content and/or Referent axes for utterances that did not require any content or referent specification. For example, the utterance "yes" is annotated as {CONFIRMATION, NONE, NONE}. Table 1 provides the complete lists of labels for each axis, along with their distributions within the 4,295 utterances observed in 100 games. Table 2 provides a sampling of utterances from the 100 training games with their assigned label triples.

4.2 Feature Selection

Each line of dialogue is transformed into a feature vector consisting of features derived from the surface text, and contextual features based on the physical situation of the speakers. Contextual features include the social role of the speaker (waitress or customer), the posture of the speaker (sitting or standing), who the speaker is facing (one or more of: customer, waitress, bartender, chef), and the containing spatial region of the speaker (one or more of the possibly overlapping regions: inside-the-restaurant, outside-the-restaurant, entrance, podium, table, counter, bar, behind bar, kitchen). The physical state of the players is reported explicitly in the game logs. The text-based features primarily consist of indicators for the presence of unigrams, bigrams, and trigrams of words observed to be salient for particular labels, as well as a smaller number of indicators for symbols and punctuation ('?', '!', '$', emoticons, and digits). Salience is computed based on the mutual information between n-grams and labels, where mutual information is a measure of statistical dependence [3]. Mutual information has been applied for text-based feature selection previously [6].

The contextual feature set remains constant for each axis (speech act, content, and referent), while the salient indicators of the text-based feature set are customized for each axis. For each axis, we compute the mutual information between every label and every unigram, bigram, and trigram. The feature set for a classification axis is the compilation of the top 50 unigrams, bigrams, and trigrams for each label. We compute the mutual information between an n-gram and a label as:

$$MI(word, Class) = P(word, Class) * \log\left[\frac{P(word, Class)}{P(word) * P(Class)}\right]$$

Where *Class* refers to a label (e.g. ASSERTION, DIRECTIVE, APPROVE, LAUGH, BILL, MONEY, etc.), and *word* refers to a unigram, bigram, or trigram of words from an utterance.

Table 1. Label distributions and classification accuracy, precision (Pr), and recall (Re).

Speech Act	Dist.	Pr / Re	Content	Dist.	Pr / Re	Referent	Dist.	Pr / Re
ASSERTION	338	0.6 / 0.5	APOLOGIZE	71	0.8 / 0.9	AGE	19	0.6 / 0.5
CONFIRMATION	354	0.9 / 0.8	APPROVE	267	0.7 / 0.6	BILL	106	0.9 / 0.9
DENIAL	90	0.7 / 0.7	BRING	413	0.8 / 0.8	CUSTOMER	5	1.0 / 0.2
DIRECTIVE	1,217	0.8 / 0.9	COMPLAIN	88	0.4 / 0.1	DIET	8	0.0 / 0.0
EXPRESSIVE	724	0.8 / 0.8	CONSOLE	11	0.8 / 0.3	FLOWERS	31	1.0 / 0.8
GREETING	302	0.9 / 0.9	CORRECT	11	0.5 / 0.2	FOOD	1,394	0.9 / 0.9
OTHER	517	0.5 / 0.4	DESIRE	363	0.8 / 0.8	GEOGRAPHY	51	0.9 / 0.3
PROMISE	136	0.9 / 0.8	EXCUSEME	25	0.8 / 0.8	MENU	52	0.9 / 0.9
QUESTION	617	0.8 / 0.9	FAREWELL	110	0.8 / 0.7	MONEY	75	0.8 / 0.6
			FOLLOW	24	0.9 / 0.8	NAME	24	1.0 / 0.3
			GIVE	170	0.8 / 0.7	OTHER	651	0.6 / 0.4
			HELLO	167	0.9 / 0.9	RESTAURANT	20	0.8 / 0.6
			INFORM	176	0.6 / 0.3	SPECIALS	12	0.9 / 0.6
			LAUGH	76	0.8 / 0.9	STAFF	22	0.9 / 0.5
			MOVE	32	0.4 / 0.2	TABLE	37	0.9 / 0.9
			OTHER	643	0.5 / 0.7	TIME	107	0.9 / 0.7
			PICKUP	29	0.5 / 0.3	WAITRESS	21	0.8 / 0.7
			PREPARE	627	0.9 / 0.9			
			REPREMAND	24	0.4 / 0.3			
			SIT	74	0.9 / 0.9			
			STATUS	149	0.7 / 0.4			
			THANK	290	0.9 / 0.9			
			UNDERSTAND	25	0.8 / 0.4			
			YRWELCOME	28	0.8 / 0.8			
CORRECT: 77.3%			CORRECT: 75.3%			CORRECT: 81.1%		
BASELINE: 28.3%			BASELINE: 15.0%			BASELINE: 38.6%		
OVERALL CORRECT:	60.9%				OVERALL BASELINE:	14.3%		

4.3 Classifier Implementation

There have been numerous approaches to automatically classifying speech acts, including neural network classification [12], maximum entropy model classification [1], and Hidden Markov Model (HMM) speech act classification [21]. Our classifier is composed of three independent HMM classifiers, one for each axis (speech act, content, and referent). An HMM classifier exploits transition probabilities in the temporal patterns that emerge in human dialogue to boost classification recognition beyond that of individual utterances. We employed the SVM^{hmm} classifier [9], which combines a Support Vector Machine (SVM) for observation classification with an HMM for learning temporal patterns of hidden states. Words and contextual features function as observations, and the labels themselves are the hidden states. This combination of an SVM and HMM has proven successful for dialogue act classification previously [19].

Table 2. Example labels for utterances in corpus, sorted by classification precision (pr).

Annotation	Utterance	Pr.
{EXPRESSIVE, THANK, MONEY }	"thank you for the tip"	1.0
{ASSERTION, COMPLAIN, FOOD }	"excuse me, i didn't order the cheesecake"	1.0
{PROMISE, BRING, MENU }	"I'll be right back with your menu"	1.0
{DIRECTIVE, PREPARE, FOOD }	"one steak please"	0.9
{GREETING, HELLO, NONE }	"Welcome!"	0.9
{QUESTION, DESIRE, FOOD }	"Would you like a drink to start with?"	0.9
{CONFIRMATION, NONE, NONE }	"okey dokey"	0.9
{PROMISE, BRING, BILL }	"I'll be back with your bill in a moment."	0.8
{DIRECTIVE, FOLLOW, NONE }	"follow me and i will have u seated"	0.8
{ASSERTION, GIVE, NONE }	"there we r sir"	0.8
{DIRECTIVE, SIT, NONE }	"have a seat wherever you want"	0.8
{DIRECTIVE, BRING, FOOD }	"Yes I'll start off with a soup du jour"	0.8
{EXPRESSIVE, YRWELCOME, NONE }	"no problem"	0.8
{QUESTION, DESIRE, TABLE }	"table for one?"	0.8
{EXPRESSIVE, LAUGH, NONE }	"lol"	0.8
{OTHER, OTHER, OTHER }	"i need to complete my quest"	0.5
{OTHER, OTHER, OTHER }	"donfdgdfgdfgdfgadfg"	0.5
{OTHER, OTHER, OTHER }	"some guy wanted to put 400 mb on floppies"	0.5
{OTHER, OTHER, OTHER }	"what are you a vampire?"	0.5
{EXPRESSIVE, APPROVE, FOOD }	"alrighty that was a satisfying dinner"	0.5
{EXPRESSIVE, APPROVE, FOOD }	"yum that lobster is too good"	0.5
{QUESTION, INFORM, SPECIALS }	"any specials today?"	0.0
{ASSERTION, COMPLAIN, NONE }	"its cold"	0.0

4.4 Classification Results

Despite the apparent freedom, players of *The Restaurant Game* tend to constrain their dialogue to social conventions associated with the mutually understood "scripts" of restaurant interaction. This contributes to strong classification results given the challenge of correctly classifying three independent axes capable of producing 4,050 unique triples.

Table 1 presents our classification results, evaluated with 10 fold cross validation. (each fold trained on 90 game logs and tested on 10). For each of the classification

axes, we report the precision and recall of each label, followed by the percentage classified correctly and a comparison baseline. All of the axes perform significantly better than baseline, contributing to 60.9% of the utterances being classified entirely correctly – correct on all three axes. It is notable that a human labeled at least one axis as OTHER in 11.7% of the incorrectly classified utterances. If we focus on the utterances that the human felt were relevant to the restaurant scenario, and ignore these degenerate utterances, the overall percentage correct increases to 70%.

For each label, we tabulate the number of instances in which the label was assigned by the classifier, the number assigned by the human annotator, and the number correctly classified (where the human and classifier agree). Precision is computed by dividing the number correctly classified by the total number assigned by the classifier. Similarly, recall is computed by dividing the number correctly classified by the total number assigned by the human. Baseline represents the percentage classified correctly if we always choose the most common label for each axis (DIRECTIVE, OTHER, and FOOD respectively).

In addition, we evaluated inter-annotator agreement among humans. A volunteer not involved with the development of the classifier annotated 10 game logs (422 utterances). We computed a kappa coefficient of {0.73, 0.70, 0.89} respectively for {speech act, content, referent}, with a mean kappa of 0.77. Kappa between 0.61 and 0.80 is considered substantial agreement [10].

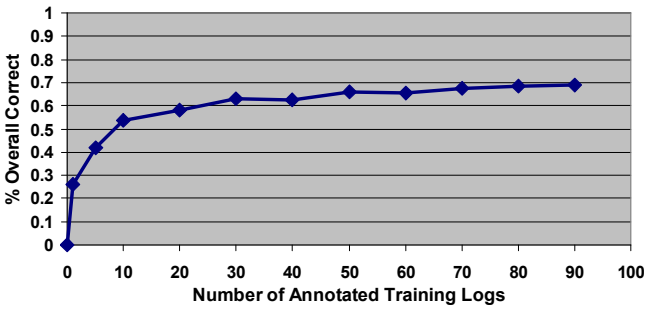

Fig. 3. Effect of training corpus size on classification.

Figure 3 illustrates the effect of the training corpus size. For one particular cut of the data, we plot the overall percent correct when the classifier is trained on between 1 and 90 training log files, and tested on the same set of 10 logs. Given that the human labor involved in annotation is expensive, it appears as though annotating more than 30 game logs yields diminishing returns. Reviewing Table 1, we see that the labels with unsatisfactory results for precision or recall are most often due to sparse data – few examples for these labels in the training data. This suggests that continuing to annotate data is worthwhile, if we can focus human labor on appropriate selections of data. It is likely that a human-machine collaborative effort could lead to significant classification improvements, where the machine requests human assistance on assignments of low precision or recall, and efficiently classifies the rest independently.

5 Predictive Model Evaluation

The fact that we can correctly classify a large proportion of utterances does not guarantee that these dialogue acts are useful for modeling social interaction. In this section, we demonstrate quantitatively that dialogue acts are useful building blocks for learning patterns of interleaved utterances and physical actions. Our long term goal is to generate social behavior for agents based on models learned by observing human-human interactions. These models will guide agents to conform to expected social conventions, and predict future actions (physical and linguistic) of other agents based on recent observations. As a first exploration in this direction, we experimented with simple n-gram statistical models [11] applied to both the surface word level and the speech act "intentional level." In our experiments, we replay the interactions observed between two humans up to some point in a particular game log, and then stop the simulation and predict the next human utterance or action. For each game log in the test set, we slide a window of size *n* over the entire log and count correct predictions of the next utterance or action. These predictions indicate what an agent would do, if guided by these models. In section 5.1 we only predict the next *utterance* based on recent utterances; in section 5.3 we predict the next *action or utterance* based on an interleaved model.

We evaluate our dialogue act classification quantitatively by learning three separate dialogue models – based on (1) classified utterances, (2) automatically clustered utterances, and (3) raw utterances – and comparing the predictive power provided by these models. We first evaluate models of utterance sequences alone. Next, we evaluate interleaved models of physical actions and utterances, in order to evaluate how well these utterance abstractions function as a common currency with physical actions.

5.1 Comparison of Utterance Abstractions

Our original corpus of 5,200 game logs was divided into 100 logs for annotating and training the dialogue act classifier, 4,800 logs for training n-gram models, and 300 logs for evaluating prediction accuracy. After training the dialogue act classifier on 100 logs, we automatically classified all utterances in the remaining 5,100 games in the corpus. There were 312 unique dialogue act triples observed in the 100 annotated logs, with 183 observed in more than one log.

There are 112,650 unique raw utterances observed in the corpus. We clustered these utterances automatically using the k-means algorithm, based on the Euclidean distance between feature vectors of unigrams, bigrams, and trigrams observed within the utterances. We chose k=300, as this number of clusters provides a fair comparison with the 312 unique dialogue act triples.

Figure 4 illustrates that dialogue acts are more predictive than raw utterances or clusters. Prediction accuracy is computed by counting the number of correct predictions of the next observed utterance, cluster, or dialogue act in a bigram, trigram, or 4-gram. The baseline prediction accuracy is computed by counting the number of correct predictions if we always choose the most likely utterance, cluster, or dialogue act found in the training corpus. Raw utterances yield poor prediction

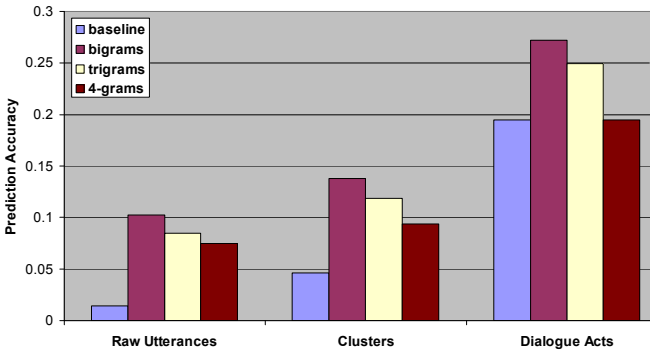

Fig. 4. Prediction accuracy for utterances.

accuracy for n-gram models for all values of *n*, only achieving above 0.1 for bigrams. While clusters do achieve about a 30% increase in prediction accuracy over raw utterances, they fall below that of dialogue acts by over 50%.

5.2 Filtering to Improve Prediction

There is value in knowing what we don't know. Our classifier assigns labels with 60% accuracy. Ideally, we would train the n-gram model with only correctly labeled dialogue acts, rather than introducing noise with those classified incorrectly. Based on the statistics computed in section 4.4, we can interpret precision as our confidence that the classifier has assigned the correct label to an utterance, and exploit this to determine which labels to include in our model, and which to omit. Like a human traveler in a foreign country with limited understanding of the language, the system can grasp onto well understood utterances and exploit them to understand the gist of the interaction. There is no notion of confidence in automatically generated clusters,

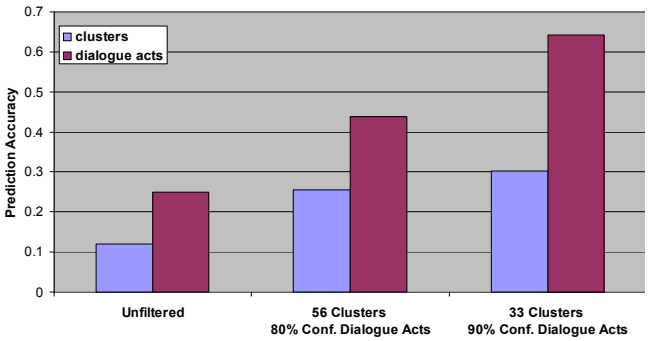

Fig. 5. Prediction accuracy for filtered utterances.

thus clusters cannot be filtered in the same meaningful way. For the sake of comparison, the best we can do is filter clusters by their observation frequency in the training corpus.

If we train an n-gram model based only on dialogue acts with precision of at least 80%, trigram prediction accuracy increases from 0.25 to 0.44. With 90% precision, accuracy increases to 0.64. There are 56 unique dialogue act triples with at least 80% precision covering 51.7% of the training data, and 33 triples with 90% precision covering 28.4% of the training data. See Table 2 for examples of high precision dialogue acts for the restaurant domain – utterances related to ordering food, paying bills, getting seated, and bringing menus. It is not surprising to see an increase in prediction accuracy when we decrease the number of unique labels, and number of utterances labeled. However, Figure 5 illustrates that filtering clusters in a similar way does not yield the same dramatic increase that we observe with dialogue acts. We compare prediction accuracy of the 56 most likely clusters to the 56 dialogue acts with 80% precision, and the 33 most likely clusters to the 33 dialogue acts with 90% precision.

5.3 Integrating Utterances with Physical Acts

We followed the methodology described previously [14] for generating a lexicon of unique physical actions from a corpus of thousands of game logs. By observing the state changes that occur each time a player takes a physical action in each game log, we learn a lexicon of context-sensitive, role-dependent actions (e.g. waitress picks up pie from counter). In our 5,200 logs, we have observed 7,086 unique actions. Based on the learned lexicon, log files are transformed into a sequence of action lexicon indices interleaved with utterances, where utterances may be represented as either clusters or dialogue act triples.

In Figure 6, we illustrate the effect on prediction accuracy of integrating utterances into a trigram model of physical interaction. Based on recent observations of

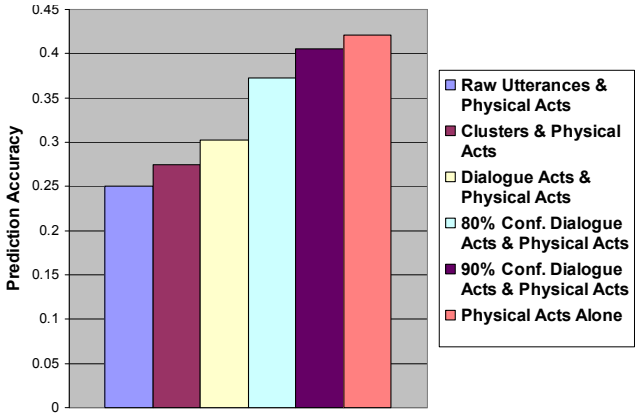

Fig. 6. Effect of interleaving physical acts with utterances.

interleaved actions and utterances, the integrated model predicts the next action *or* utterance. This is a difficult problem, given that we are polluting the lexicon of 7,086 directly observable actions with 112,650 unique utterances, which we can classify with 60% accuracy. 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. While 90% confidence only covers 28.4% of the utterances, these are highly salient utterances for the restaurant domain, and coverage should increase as more data is annotated and the classifier improves. This is a first step toward discovering a higher-level structure of the restaurant scenario, composed of interleaved sequences of actions and utterances.

Ideally integrating dialogue acts with physical actions would yield a higher prediction accuracy than either alone. The current representation of physical interaction clusters similar objects, and abstracts away timing information. In other words, the model does not differentiate between picking up *steak* or *salmon* (both clustered as *food*), and does not need to predict *when* the waitress will depart from the table to pick up food from the kitchen (perhaps after a three utterance exchange, concluding with "I'll be right back with that"). Cleary these details will be important to an automated agent. We are working toward an agent guided by a model of physical interaction that retains these details, and we expect better prediction from the interleaved model of actions and words than from a model of either alone.

5.4 Speech Acts vs. Dialogue Act Triples

Recall that our classification scheme classifies the propositional content and referent in addition to the illocutionary force of each utterance. In this section, we demonstrate that this difference makes a significant impact on our ability to integrate a maximal number and variety of utterances into the predictive model of physical interaction, while preserving predictive power. We compare the predictive power of dialogue act triples to that of speech acts alone, while scrutinizing the percentage of utterances covered as we filter by confidence (aka precision).

Initially, integrating speech acts into the model of physical interaction yields a slightly higher prediction accuracy than integrating dialogue act triples (0.34 vs. 0.30). As we filter out lower confidence speech acts and dialogue acts, the prediction accuracy of both comes closer to that of the physical interaction alone (0.42), and the difference between predictive power of speech acts and dialogue acts becomes negligible. However, as we raise the confidence threshold, the percentage of utterances covered decreases dramatically for the coarser grained speech acts, as seen in Figure 7. We preserve 28.4% of the dialogue acts with 90% confidence, and only 6.5% of speech acts. As seen in Table 1, only GREETING speech acts have above 90% precision, compared to 33 unique dialogue act triples with 90% precision, which employ the full range of speech act labels (see Table 2 for a subset). At 80% confidence, a higher percentage of speech acts are preserved than dialogue acts (63.4% vs. 51.7%), but this comes at the cost of a 5% decrease in predictive power

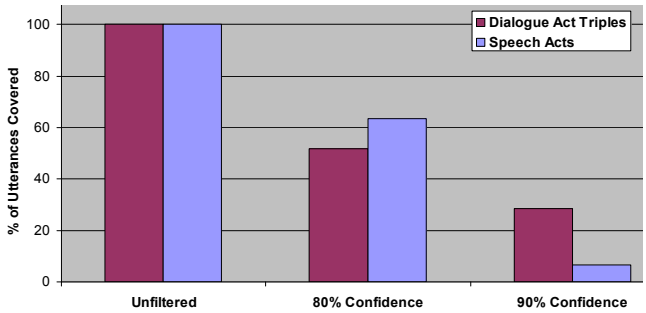

Fig. 7. Percent of utterances covered when filtering.

from physical interactions alone. Annotating with finer grained dialogue act triples provides a means of recognizing a maximal number and range of salient utterances for the restaurant domain, while preserving predictive power.

6 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.

It is likely that our dialogue act classifier could be improved by providing more training data guided by an active learning process, however intention of utterances can never be fully recognized without understanding their role in the higher level structure – the sub-goals of the restaurant scenario composed of interleaved physical and dialogue actions. Our evaluation of integrating physical actions with dialogue models demonstrates the potential for dialogue acts to function as building blocks of sub-goals. Discovering this higher level structure remains a goal for future work. Human annotation will be required to identify intentional sub-goals spanning multiple physical actions and/or dialogue acts, and based on our experience with semi-automated dialogue act annotation, we are optimistic that semi-automation of sub-goal annotation will be possible as well.

References

1. Carvalho, V.R., and Cohen, W.W.: On the Collective Classification of Email Speech Acts. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil (2005)

2. Core, M., and Allen, J.: Coding Dialogs with the DAMSL Annotation Scheme, In Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA (1997)
3. Cover, T.M., and Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, Inc. (1991)
4. Edery, D., and Mollick, E.: Changing the Game: How Video Games are Transforming the Future of Business. FT Press, Upper Saddle River, New Jersey (2008)
5. Fleischman, M. and Hovy, E.: Taking Advantage of the Situation: Non-Linguistic Context for Natural Language Interfaces to Interactive Virtual Environments. Intelligent User Interfaces (2006)
6. Gorin, A., Riccardi, G., and Wright, J.: How may I help you? Speech Communication, Volume 23, 113-127, Elsevier Science (1997)
7. Gorniak, P. and Roy, D.: Speaking with your sidekick: Understanding situated speech in computer role playing games. In Proceedings of Artificial Intelligence and Digital Entertainment (2005)
8. Huang, J., Zhou, M., and Yang, D.: Extracting chatbot knowledge from online discussion forums. In Proceedings of IJCAI (2007)
9. Joachims, T.: SVM^{hmm}: Sequence Tagging with Structural Support Vector Machines. http://www.cs.cornell.edu/People/tj/svm_light/svm_hmm.html (2008)
10. Landis, J.R., and Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics, 33:159-174 (1977)
11. Manning, C.D., and Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Massachusetts (1999)
12. Marineau, J., Wiemer-Hastings, P., Harter, D., Olde, B., Chipman, P., Karnavat, A., Pomeroy, V., Graesser, A., and the Tutoring Research Group: Classification of speech acts in tutorial dialog. In Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference, pp. 65–71 (2000)
13. McQuiggan, S., and Lester, J.: Learning empathy: A data-driven framework for modeling empathetic companion agents. In Proceedings of the 5th International Joint Conference on Autonomous Agents and Multi-Agent Systems. Hakodate, Japan (2006)
14. Orkin, J., and Roy, D.: The Restaurant Game: Learning social behavior and language from thousands of players online. Journal of Game Development, 3(1), 39-60 (2007)
15. Orkin, J. and Roy, D.: 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)
16. Schank, R.C., and Abelson, R.P.: Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. Lawrence Erlbaum Associates (1977)
17. Searle, J.R.: Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge, United Kingdom (1969)
18. Satingh, S., Litman, D., Kearns, M., and Walker, M.: Optimizing Dialogue Management With Reinforcement Learning: Experiments with the NJFun System. In Journal of Artificial Intelligence Research (2002)
19. Surendran, D., and Levow, G.: Dialog Act Tagging with Support Vector Machines and Hidden Markov Models, In Proceedings of Interspeech (2006)
20. von Ahn, L., and Dabbish, L.: Labeling images with a computer game. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 319-326, Vienna, Austria (2004)
21. Woszczyna, M., and Waibel, A.: Inferring linguistic structure in spoken language. In Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan (1994)