Learning Avatar Behaviours
Project for MAS 732

Peter Gorniak
MIT Media Lab
20 Ames Street
Cambridge, MA
02139
pgorniak@media.mit.edu

Abstract

To engage in discourse, a participant must map the discourse functions he or she wants to achieve into surface level realizations. These mappings are often treated as rule-based and general in work on discourse and computational discourse. Here, I demonstrate a proof-of-concept system that learns such mappings stochastically in a reinforcement learning paradigm. Learning these mapping has the advantage that it takes some of the work out of the mapping design and allows for more personalized systems. Approaching the problem stochastically deals with the inherent brittleness of rule-based system and allows reasoning about beliefs as well as graceful degradation. The implementation discussed here is based on BodyChat, a virtual chat system with semi-autonomous avatars.

1 Introduction

Much of the work in discourse analysis and synthesis involves mapping from the underlying discourse structure to surface level phenomena, and vice versa. These mappings may be hypothesized from observing and clustering surface level phenomena, either intuitively or statistically or by reasoning about the mappings directly. The result is typically used heuristically both in analysis and generation of discourse. Such mappings are also usually static in that they do not easily incorporate a new source of information (like additional discourse features), in that they do not account for individual preferences, and in that they do not model either time dependency or uncertainty about the current state of discourse.

As a result, while these heuristic mappings do provide enhanced understanding of discourse functions and their realizations, they require a large amount of work on the part of the designer and still sometimes fail to produce natural discourse behaviour. These failures may happen due to ignoring uncertainty about the current discourse state (due to failure of any component of the system, from speech recognizer to semantic interpreter). They may also happen due to limitations in the mapping itself in terms of the number and types of discourse phenomena it draws from, the number and types of surface level realizations it maps to, and the appropriateness and generality of the mapping itself.

1.1 Problems in Discourse Modeling

Thus, the problem consists of three parts, each of which has seen some prior related work:

Failure to Account for Uncertainty Rules are a necessary framework underlying and constraining any probabilistic system. While the papers in (Klavans and Resnik, 1996) summarize this tradeoff nicely and argue for a hybrid between the two, they fail to emphasize that uncertainty is a phenomenon important for all aspects of discourse, both in terms of modalities (speech, gesture, facial expressions, etc.) and in terms of levels within and across modalities. For example, a discourse participant may not understand words at an acoustic level, or may have problems understanding a sentence due to lack of topical knowledge. While a rule-based system likely fails here, a probabilistic system can be aware of its uncertainty, and try to disambiguate the event in one of multiple ways. An appropriate probabilistic model of these channels should be able to naturally use information in one channel to disambiguate another. For example, it may use recognition of a deictic gesture both to identify referents for lingustic expressions, and to improve understanding of the speech accompanying the gesture. A natural way to think about incorporating various sources of probabilistic knowledge is to view a discourse participant as an agent making decisions under uncertainty, which is by now a well-studied field as summarized in (Boutilier et al., 1999). This way of thinking also has as almost a by-product the advantage that the agent is in some sense aware of its own confusion, and can decide to take disambiguating actions. (Paek and Horvitz, 2000) is an example of exactly this type of agent. In this paper, I present an agent based on the Markov Decision Process framework for decision making under uncertainty (Boutilier et al., 1999).

Hand-Crafted Behaviour Mappings There still remains the problem of supplying a probabilistic decision-making process with its structure and probability parameters. While these may be estimated empirically in an a priori fashion, this only yields a static and impersonal decision making process. Ideally, one would like the agent to be adaptive and personalizable, both in the sense of adapting it to one's personal preferences as well as in the sense of infusing it with personality. Classes of personalities and emotional modeling as in (Ball and Breese, 2000) may yield some advances here, but it would be far more convenient to let the agent learn its behaviours autonomously during discourse interaction. Luckily, learning algorithms to perform that type of learning in Markov Decision Processes exist under the name of Reinforcement Learning (Sutton and Barto, 1998). The agent presented here uses the Q learning algorithm (Watkins, 1989).

Overspecialized Discourse States Most work on discourse relies on some notion of discourse state. However, there exists no formal way of integrating the various discourse state representations, such as (Grosz et al., 1989), to let a discourse participant integrate the various foci such representations impose on discourse modeling. While Markov Decision Processes do not solve this problem automatically, they at least provide a framework to fit the various existing discourse state representations into, as well as a large toolbox of representation and optimization techniques ranging from Bayesian Networks to clustering algorithms structured representations to efficiently model and analyse the discourse state space (Boutilier et al., 1999). In addition to uniting existing work, these state representation techniques also provide obvious ways of incorporating new source of information.

1.2 Discourse State vs. Function

I should note that I conflate the notion of discourse state with that of discourse function here. This is a debatable step. Here, it allows me to equate actions in the Markov Decision Process sense to surface level behaviours in the discourse sense, and to learn a mapping from discourse states (or functions) to actions (or behaviours). One might argue that the mapping from discourse states to discourse functions is a mapping that needs to be learned first. I work under the assumption here that it is possible to make the discourse state description rich enough (including goals of the participants, context, etc.) so that it subsumes discourse functions. For example, a state that includes the fact that the discourse participants are 20 feet apart, that they know each other well, that both wish to speak to each other, and that they made eye contact in the last timestep completely determines the discourse function of being engaged in a specific version of a distance greeting. On the other hand, one could also claim that the discourse function should be the primary level of action decided upon by a discourse participant. I currently have no strong arguments to support one decision more than the other, except for the fact that there is no simple decision making model that will learn two layered action mappings at once.

1.3 Setting

In this paper, I describe a proof-of-concept implementation of a learning computational discourse participant. Its design is phrased in terms of an agent having to make decisions under uncertainty and uses Q learning over a Markov Decision Process to arrive at a policy connecting discourse states to actions. This is proof-of-concept work and is made possible yet at the same time limited by the following design decisions:

1. To solve sensing and real-world modeling problems, the agent is situated in a simulated 3D environment.

2. To deal with the immense variety of complicated discourse phenomena, the agent only learns a tiny subset of possible discourse state to action mapping, namely only some of the gestures and glances involving conversation initiation and closing. Other functions, as far as they are possible in the impoverished virtual environment, are provided through shared control of the agent with a human being.

3. The rewards and penalties necessary for reinforcement learning are provided by the user, either through explicit rewarding, or through demonstration of appropriate actions.

4. The state representation used is not itself probabilistic in nature. This makes sense as the environment is far more deterministic than the real world. However, all the work presented here should generalize immediately to, for example, a Bayesian Network representation.

The platform used in this project is the BodyChat system by (Cassel and Vilhjamsson, 1999), a graphical chat system that implemented the same set of mappings from discourse state to surface level behaviours for semi-autonomous avatars, but in a rule-based way.

The last item needing to be discussed in terms of introduction is the notion of learning discourse behaviours. It should be made clear that there is no direct correspondence between the type of learning presented here, and the learning children undergo when they acquire discourse skills. The BodyChat environment distorts human conversation through the rough and comic-like graphical representation of its avatars and their limited motions, the difference between typed text and speech, and the fact that we are dealing with chat between adults, make the set of features available for receiving feedback about discourse actions radically different from that available to children. I therefore draw no parallels between the two kinds of learning in this paper.

2 Reinforcement Learning for Shared Control Discourse

This section contains a short overview of the formalism and the learning algorithm used in this project. A Markov Decision Process (MDP) is defined by a set of states, a set of actions and the dynamics of the process specified as a set of transition probabilities and a set of expected rewards. These dynamics are specified under the assumption that knowing the state and action at time t is enough to know the probability of receiving a reward r and transitioning to state s' at time $t + 1$. This is the Markov assumption. Solving an MDP involves producing an optimal policy π^* that is a function from states to actions

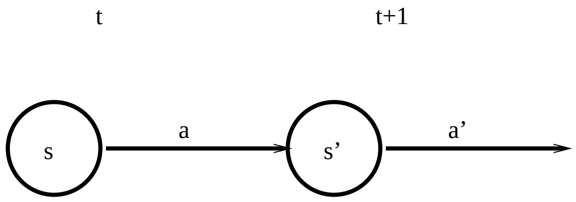

Figure 1: MDP State and Action Structure

(i.e. $\pi^*(s) = a$) and that maximizes the expected future rewards when followed throughout. Mapping this to an agent in an environment, the question answered by an MDP is which action the agent should take in the current state to maximize its expected future rewards. If the MDP is fully specified, the optimal policy can be found using Dynamic Programming. However, in most cases the dynamics of the process are unknown to the agent. This (harder) problem can be solved via Reinforcement Learning algorithms that explore the state and action space to learn about transition probabilities and rewards. There exists a good number of variations on the basic algorithms.

The difference between the problem at hand and the standard MDP formulation is that an MDP assumes full autonomy on the part of the agent. In the discourse problem, this means the agent should decide when and how to gesture, move, speak, etc. However, solving the general discourse problem as a first step seems like an overly ambitious goal. However, it is hard to produce discourse phenomena in isolation (many gestures only occur in the context of speech, for example.) The solution offered by BodyChat is to let a human being drive most discourse activity, like speech and movement of the avatar, but to automate others, like gesture and gaze. Luckily, there exists a class of Reinforcement Learning algorithms known as off-policy learning algorithms, that allow the agent to learn the optimal policy while following another. In our case, the actual policy followed is partially dictated by the user's action decisions. The off-policy algorithm used here is Q-Learning (Watkins, 1989), an off-policy version of temporal difference learning. In temporal difference learning, the agent learns the value of being in a certain state and taking a certain action by exploring the state and action space and updating values based upon the rewards experienced and the current estimates of values for future states. The Q learning update rule is the following:

$$Q(s,a) \leftarrow Q(s_t,a_t) + \\ \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)]$$

The Q-value computed here corresponds to the value of taking action a in state s and following a certain policy to choose actions in the future. In turn, the Q values induce a policy as an agent can simply always choose the action with the maximal Q value in the current state. The goal of finding the optimal policy thus translates into the goal of finding Q values that induce the optimal policy. The Q learning rule does this in a straighforward way: $\max_a Q(s_{t+1},a)$ is an estimate of how good taking the best action at the next timestep is. This, plus the next reward experienced, r_{t+1} is the difference between the current and the new estimate of $Q(s,a)$. The parameter α is a learning rate parameter, and γ discounts future rewards so that in an ongoing task the expected future rewards are never infinite in value. Q learning is independent of the actual policy followed because of the max operator that computes the value of the best possible future instead of the value of the future dictated by the policy being followed. This algorithm is proven to converge under a number of assumptions like complete exploration of the state and action space and an infinite number of visits to all states, all of which do not hold in the case discussed here.

Shared control also problematizes the simple assumptions of MDP's that all actions are possible choices for the avatar and that each timestep contains exactly one state-action pair. In BodyChat, the avatar can be moved freely by the user and the user produces the avatar's speech. I did not endeavour to have the avatar learn these sophisticated behaviours, but they occur in continuous time and produce state changes in continuous time, breaking the discretized view and MDP imposes on the world. To discretize time, timesteps are simply delimited by state change events. For example, when an avatar moves into another avatars 'fringe zone', the last timestep is considered to have lasted from the previous state change to the current one. The rule that only the last action in the timestep counts enforces the one action per timestep limitation. This rule also makes sense in terms of the reward system employed, as will become obvious below.

The last two problems in learning discourse behaviours from reinforcements are those of specifying the rewards given to the agent, and that of speed of convergence of reinforcement learning algorithms. As discussed in the introduction, BodyChat is the wrong setting to try to interpret the actual discourse occurring for perceived rewards and punishments. Instead, the user can personalize the agent by explicitly distributing rewards and punishments in response to avatar behaviour. At every state change, the avatar will choose and action and execute it. If the user takes the avatar to be behaving appropriately, he or she can reward the avatar with a constant positive reward by pressing a 'Good Avatar' button. Or, the user can punish the avatar upon bad behaviour using a 'Bad Avatar' button. Note that reinforcement learning automatically distributes this feedback over time, and thus the user response does not have to occur at every timestep and can evaluate behaviours over time instead of individual actions. While this sort of feedback should eventually let Q learning converge, due to the number of state and actions in even a small system such as this one it is unlikely to do so quickly enough to make this version of BodyChat useable. Intuitively, it takes a good amount of time for the avatar to even try the appropriate actions in the appropriate states, and even longer for it to learn that these are the appropriate actions. Luckily, the fact that control of the avatar is already shared between the learning agent and the user can be exploited to extend the reinforcement learning paradigm to a learning by demonstration solution. In this version, the avatar's action choices are also available to the user. If the avatar picks an inappropriate action in a given state, the user can thus demonstrate the correct action. As only the last action in one timestep is used in the Q update rule, the user's action choice overrides the avatar's action choice. At the same time, an automatic positive reward is issued after every user action. In this way, user action choices are made to look rewarding to the agent and thus Q learning is driven towards them. This speeds up learning immensly and alleviates the need for lengthy exploration of the state and action space as the user can drive learning towards the correct set of behaviours by demonstration and then tweak them by using the explicit feedback options.

3 Discourse States and Actions

The original BodyChat system uses an implicit definition of discourse state relevant to greetings, closings and some behaviours during the main part of a conversation specified by

- distance between pairs of avatars (classified into distant, within a fringe zone, and within a partner zone),
- the avatar that is currently the focus of another avatar (if any)
- whether the avatar is speaking
- whether the avatar in focus is speaking
- whether the avatar's user is typing on the keyboard (taken to mean that he or she is preparing to speak)
- whether the user has declared the avatar to be available for a new conversation or not
- whether certain keywords appear in the speech provided by the user
- whether a special closing symbol appears in the speech, taken to mean the the user wants to break off the conversation

To keep the state space initially small, the state representation employed in the learning avatar makes some of the variables above explicit, but limits them to only two discourse participants. The state defining variables are:

- distance to the other avatar (classified into distant, within a fringe zone, and within a partner zone),
- whether the avatar is speaking
- whether the avatar's user is typing on the keyboard
- whether the user has declared the avatar to be available for a new conversation or not
- the previous state

The last variable is included because the state description is otherwise strongly non-Markovian, as for example the states when moving away from an interlocutor are identical to those moving towards one, yet obviously require different actions and have different dynamics. Note that the state description given also does not incorporate actions by other avatars, making it impossible to learn behaviours that occur as adjacency pairs. However, due to the nature of the representation and algorithms presented here, the state description can easily extended to include any number of variable with a need to change the learning algorithm employed.

From (Goodwin, 1981) it seems reasonable that the chosen state description should allow for learning at least a small subset of the human behaviours used in greetings. Here the extensibility of this approach becomes clear, as additional features, like the social distance between greeters, can easily be incorporated into the state description when deemed appropriate. Indeed, it would even seem possible to evaluate the need for certain factors by examining the policy learned and how it depends on changes in a given variable.

4 BodyChat Interface

The learning version of BodyChat uses the original interface consisting of a 3D view of the virtual world, showing by choice either a bird's eye view, the view of the avatar, or a view over the avatar's shoulder. Speech can be entered through a one line text input window, and avatar actions can be selected from a pull-down menu. For the learning task, I augmented the interface with a button to provide positive rewards ('Good Avatar') and one to provide negative rewards ('Bad Avatar'). As the user has to evaluate his or her own avatar's actions, I also added another always visible view window that shows a frontal view of the avatar, so that the avatar's action choices are always visible. Figure 2 shows the main interface window.

5 Results

While a full evaluation is out of the scope of this project, I did use the system together with some test users and can thus provide some initial evaluation of its various aspects.

5.1 Learning

I hesitate to use the word 'converge' in this context, but teaching by demonstration in the way discussed above appears to be an effective means of quickly letting the learning system acquire appropriate action sequences. For example, from only a few runs through a greeting sequence, first with providing an example of an appropriate action at most steps, then reinforcing appropriate behaviours, the avatar learns to, for example, toss its head for a distance greeting, nod for a close greeting and wave as a goodbye from only a few runs through the situation. Mainly, this is due to the fact that full exploration is unnecessary if the optimal actions are signaled by the user. Of course, the avatar does not learn a full Q function in this way, but as only the maximum Q value in any given state matter for action selection, that seems to be an appropriate shortcut.

Using its simplified state representation this system thus learns a subset of the original BodyChat behaviours successfully. I am confident that with a bit longer training the system could easily learn the full set of BodyChat state to behaviour mappings. More interest-

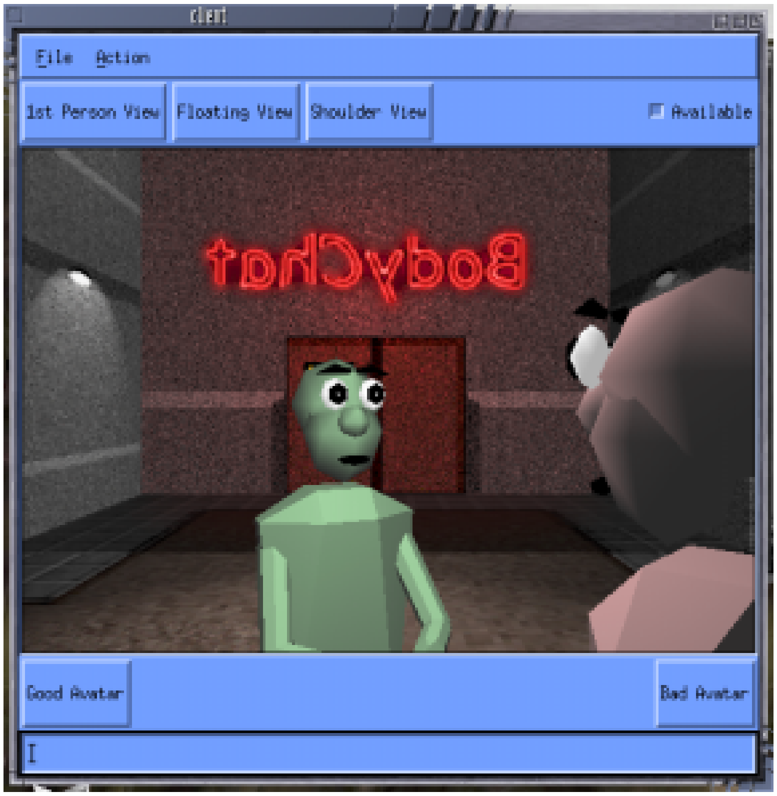

Figure 2: MDP State and Action Structure

ing than simply reproducing the BodyChat environment, would be a setting that allows for more personalization. For example, given sets of gesture versions instead of single gestures, say, 10 versions of waving one's hand, an avatar could be personalized not only in terms of appropriate discourse behaviour, but also somewhat in terms of portrayed attitude and personality. The real question is how well the training and learning methods discussed here scale to general discourse modeling with its large state and action space. Intuitively, training will take a lot longer, perhaps infesiably so at least for the application discussed here. However, when thinking about learning discourse behaviours in general the question arises again whether any setting like BodyChat can be the appropriate one, and whether one should not draw from knowledge about the context and feedback children have access to to design learning algorithms. On the other hand, the large body of work available on concise and efficient representations from the MDP and RE communities promises to yield solutions that should at least delay the point at which the type of training and learning discussed here becomes infeasible.

5.2 Interface

Another main result of this project concerns the importance of interface design for a training system such as this. The user needs at least partial access to the progress of the learning system. Some is provided by the avatar's actions, as displayed in the additional window with a frontal view, but this gives no feeling for the underlying state structure the user helps the avatar navigate. Perhaps signaling explicitly when a reward actually is consumed would satisfy the user, but perhaps a step like the one in (Rich and Sidner, 1998) of making the underlying discourse state representation visible to the user is more appropriate. Interestingly, both tasks are collaborative shared control environments.

6 Conclusion and Future Work

In this paper I presented an application of Reinforcement Learning to the problem of specifying mappings from discourse functions to surface level realizations. The main result is that an avatar in a graphical chat environment can learn simple versions of such mappings using a modified Q Learning approach. The modifications consist of the learning agent sharing control over the avatar with an expert user, and employing actions dictated by the user as a source of positive behavioural examples. These examples fit into the reinforcement learing paradigm by treating the problem as an off-policy learning problem and accompanying user specified actions with automatic rewards, thus quickly driving learning towards the relevant regions of action-state space.

There are several obvious next steps to this project. First, state space and action space should be extended to allow for richer behaviours, both in terms of covering more discourse phenomena as well as providing more varied surface level behaviours. Secondly, the state representation could easily be made more sophisticated than the extensive and explicit representation currently employed. For example, a Bayesian Network could capture the agent's beliefs about the world and incorporate the agent's sensors in a meaningful way, rather than having the state description limited to the current externally dictated variables. This would also make the transition out of the artificial environment into a noisier real environment much smoother. Extensive work on using Bayesian Networks to represent state in Markov Decision Processes already exists. Finally, there remains the question to be answered as to whether the computational learning of discourse behaviours can be made meaningful when compared with the learning human beings perform to acquire the same things. This project can make no such claim, but I believe the idea as such is certainly not out of the realm of the possible.

References

Gene Ball and Jack Breese. 2000. Emotion and personality in a conversational agent. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Chruchill, editors, *Embodied Conversational Agents*. MIT Press.

Craig Boutilier, Thomas Dean, and Steve Hanks. 1999. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of AI Research*, 11:1–94.

Justine Cassel and Hannes Vilhjamsson. 1999. Fully embodied conversational avatars: making communcative behaviors autonomous. *Autonomous Agents and Multi-Agent Sytems Journal*.

Charles Goodwin. 1981. *Conversational Organization*. Academic Press.

Barbara Grosz, Martha Pollack, and Sidner Candace. 1989. Discourse. In *Foundations of Cognitive Science*. MIT Press.

Judith L. Klavans and Philip Resnik, editors. 1996. *The Balancing Act*. MIT Press.

Tim Paek and Eric Horvitz. 2000. Dialog as action under uncertainty. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI2000*.

Charles Rich and Candace L. Sidner. 1998. Collagen: A collaboration amanger for software agents. Technical report, Mitsubishi Electric Information Technology Center America.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning*. MIT Press.

C.J.C.H. Watkins. 1989. *Learning from Delayed Rewards*. Ph.D. thesis, Cambridge University.