Survey On Task Based and Chat Style Systems using Reinforcement Learning

Ketan Desale, Anjali Parihar, Priyanka Nagarkar, Vishakha Bhosale

Abstract— In today's World the Systems that are being used are either Task-based Dialogue systems else it is the Chatbot that fulfills the user requirements. Chatbots Deals with unforeseen input provided by the user to a great extent, Task-based systems are designed for Specified Task mostly. Amazon echo (ALEXA) and Apple's siri have come up with great aspects of combing the chat-based and Task-based System but have some Limitations and shortcomings. To overcome the aforementioned issues this paper addresses different autonomous robotic systems that are either task-based dialogue systems or only chat based systems or the system that uses the combination of task-based and chat style features are studied where Reinforcement Learning (RL) is being employed.

Index Terms— Reinforcement learning (RL); Human Robot Interaction (HRI); Markov decision Process (MDP); Burlap.

1 INTRODUCTION

Spoken dialogue systems[1],[2],[3] that are mostly used nowadays are either the task-based or agenda based, dialogue systems used in Human-Robot Interaction(HRI)[4],[5],[6] also have same functionalities .On the other hand in chat-based systems (chatbots) have limited memory and are mostly focused on entertainment and don't usually support execution of user tasks and transactions . Apple's siri and Amazon echo do combine chat and task-based interaction but cannot be extended to multi-turn dialogue to predict or anticipate the user's unexpected goals or choices. HRI (Human Robot Interaction) is the study where robot interacts and communicates with human in a natural and more engaging way. To create a more integrated approach to dialogue in HRI dialogue should provide entertaining chat as well as multi-step interaction to predict the user goals and intentions for creating system that is more engaging. In [8] Reinforcement learning is used instead of hand-crafted rules to decide when to chat and when to execute user task required by the user in system that uses the novel combination of chat-based and task-based systems .Reinforcement Learning (RL) is also being used for promotion of long-turn conversations so that system can easily and smoothly do transitions between task and non-task content in the system where user engagement is the priority.

2 MATERIALS AND METHODS

This section provides information on Reinforcement Learning used in the systems[7] where user engagement is the priority for long-term conversation so that system can easily and smoothly switch between task based and non-task based content and for the [8]system uses the combination of chat and task-based systems for multi modal systems.

2.1 REINFORCEMENT LEARNING

Reinforcement Learning is one of the machine learning algorithm, it is the branch of AI (Artificial Intelligence). RL(Reinforcement Learning) is used to make the software,software agents and machines to instinctively determine the precise and clear behavior within a specific context, in order to maximize its performance. A Simple reward is given to the agent to learn its behavior from that reward feedback, this is the reinforcement signal. There are variety of different algorithms that deal with this issue. Reinforcement Learning is defined by a particular type of problem, and all of its solutions is being classed as Reinforcement Learning algorithms. In the problem, an agent decides the best action to be selected based on its current state. When this kind of step will be repeated, the problem will be known as a Markov Decision Process. In[8] Reinforcement Learning is used for creating the scalable and extensible approach for combining chat-based and task-based system and according to the customers feedback chat+Task based system was significantly more pleasant and engaging compared to task only system.

2.2 Q-Learning

Q-Learning is one of the Reinforcement Learning technique used in machine learning. This technique does not require any model of the environment. Q-learning can deal with the problems with stochastic transitions and rewards, without requiring adaptations. In[7] conversational processes is modeled as Markov decision processes(MDP) in [7] reinforcement learning algorithms is being used to train a policies, Q-learning is being used because it can handle the discrete states well and learns a Q table that supports a model that makes debugging and interpretation easier. From all the algorithms, Q-learning is a better choice with respect to the implementation of other components of the system .In a reinforcement learning setting, we formulate the problem as $(S, A, R, \gamma, \alpha)$, where S is a set of states that represents the system's environment, in this case the conversation history so far. A is a set of actions available for a particular state. In [7] setting, the actions were the strategies available. By performing an action, the agent moved from one state to the other. Execution of an action in a specific state provided agent

with a reward (which is the numerical score), R(s, a). The goal of the agent was to maximize its total reward. Agent did this by learning which action is optimal to take for each state. The action that was superlative for each state and the action that had the highest long-term reward. In[9]reward that is being given to the agent was the weighted sum of the expected values of the rewards of all future steps starting from the current state, where the discount factor was $\gamma \in (0, 1)$ which traded off the significance of sooner versus later rewards. γ can also be interpreted as the likelihood to succeed (or survive) at every step. The algorithm therefore had a function that calculated the quantity of a state and action combination, $Q: S \times A \rightarrow R$. The central part of the algorithm is a simple value iteration update. It assumes the old value by itself and makes a correction based on the new information at each time step, t. The critical part of the modeling is to design appropriate set of states, actions and a corresponding reward function. In [8] on each consequent turn. Dialogue manager decided an action based on trained Markov Decision Process (MDP) policy policy $\prod^*$. This policy was designed and trained using BURLAP [9], which is a Java-based framework for Reinforcement Learning. The standard Q-Learning algorithm in [10] was used to train the agent, using hand-crafted simulated users emulating that how they could react to each action taken by the agent. Example mentioned in [8]: if the agent responds to a user task utterance (such as "Where can I get a coffee?") with chat, the simulated user will leave the conversation with probability 0.9.For training, the discount factor that is denoted by γ was fixed to 0.99, there the agent cared about long-term rewards, while the learning rate α was kept fixed at 0.1. In order for the agent to explore as much as possible during the initial stages of the training, a policy named greedy policy was followed with an initial of 1 0.9, decaying after each turn i = i+1 .The system's states, denoting the agent's knowledge about its environment at any given time, were represented with 12 features e.g.: Distance, TaskCompleted, UserEngaged, etc. resulting in a state-space of approx. 82944 states for policy learning. The action space consists of 8 actions $a_t \in A$ where A = [PerformTask, Greet, Goodbye, Chat, GiveDirections, Wait, RequestTask, RequestShop]. Most of these actions were converted to text using a mixture of template-based generation and database lookup, and were then synthesized as combinations of speech and robot gestures. PerformTask could be unpacked in several other tasks, depending on the context of the information given. For example if the user requested for a discount or a voucher for a specific shop, the robot presented the image of a voucher in QR code on its screen. The reward function was optimizing for successful task completion as well as higher engagement, and thus awards each completed task with +10, and +5 for each consequent turn. It also penalized when the user abruptly leaves (i.e. without a 'goodbye' phase) with -100. Starting the training process, the initial Q-values were set to 0 (Q(s 0, a 0) = 0). [8]This system was able to discover optimal actions which human designers would have difficulty anticipating, for example: to trigger the chat behavior in particular multi modal state configurations where the user is moving away from the robot and a task was incomplete. In [7] it is mentioned that Q-Learning is sufficient when the number of response candidates are limited. Movie Promotion system in [7] used Q-Learning in Response selection policy to select candidates provided by the two response generator.

2.3 Q-Learning Algorithm

1. Set gamma parameter, and environment rewards in matrix M.
2. Initialize matrix R to zero.
3. For each episode:
4. Select any random state as the initial state.
5. Do While the goal state hasn't been reached.
6. Select one from the all possible actions for the current state.
7. Using this possible action, go to the next state.
8. Get maximum value of R for this next state based on all possible actions.
9. Compute : R(state, action) = M(state, action) + gamma * Max[R(next state, all actions)]
10. Set the next state as the current state.
11. End Do
12. End For

2.4 Human-Robot Interaction

HRI (Human robot interaction) is the study of communication between humans and robots in natural language. HRI(Human robot interaction) is a combined field which combines and contributes from the interaction of human and the computer , AI(artificial intelligence), robotics, natural language understanding, design, and social sciences. HRI(Human Robot Interaction) deals with intelligence of human interaction and communication, many aspects of Human robot interaction are in continuation of human communications topics that are much older than robotics per se. The closer the human and the robot get and the more complicated or detailed the relationship becomes, the more the risk of a human being injured rises. In today's advance world robots employed by manufacturers solve this issue by not letting humans and robots share the workspace at any point of time. Thus the presence of humans is completely not allowed in the robot workspace while it is working.

2.5 Speech Processing

To make the chat style dialogue system [8] uses the collection of AIML files forming the Chabot Rosie3 using the Program-Y4 AIML 2.0 interpreter. The utterance string (pattern) is encoded and send to the Chabot via REST calls, where an appropriate response (template) was formulated and fed back to the call. For switching from chat-based to task-based dialogue system, the task-related state variables are switched based on specific words and phrases used during the interaction.

2.6 Speech Processing Algorithm

In [11] Hidden-Markov models (HMMs) are defined as well admired statistical models use to implement speech recognition technologies. The time variances in the spoken language were modeled as Markov processes with discrete state spaces. Each state produced speech observations according to the probability distribution characteristics of that state. The speech observations took on the discrete or the

continuous values. In either case, the speech observations represented a fixed time duration (frame). The states were not directly observable, which is why the model was called the hidden -Markov mode. The original minimal HMM algorithm was implemented on a floating-point C language program platform running under the Unix operating system. As the result of changes in the technology, a project has been called for to transfer this algorithm from floating-point C language to TMS320C2xx assembly language (ASM). The primary objectives of this project was to have the algorithm running in real time and to have assembly code that is user-supportable.

The speech recognition algorithm illustrated in Figure 1, contains two fundamental parts, which are the acoustic front end and the search algorithm itself. The acoustic front end in [11] was the process for converting sequences of raw-speech data to observation vectors, which represent events existing in a probability space. Search algorithm then discovered the most likely sequence of these events while operating under a set of syntactic constraints.

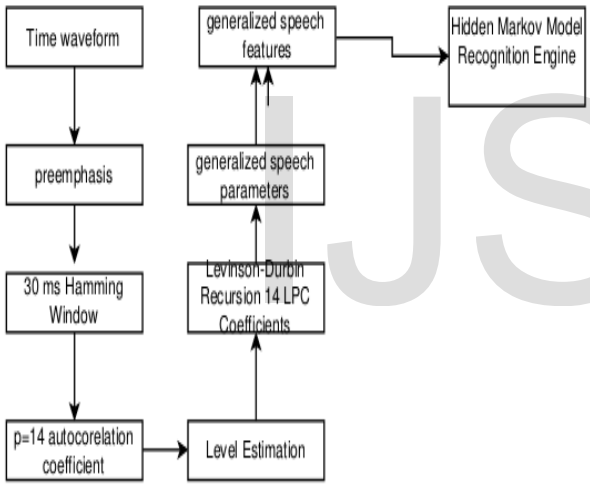

Fig. 1. Block Diagram for Speech Recognition algorithm.

3 Advantages and Disadvantages

In [8]shopping mall customers from which system was tested rated for the hybrid task + chat based condition significantly higher on a 5 point Likert scale regarding pleasant communication of the interaction and meeting their expectations, while all other questions did not show any significant differences between the conditions. Especially the rating for meeting the users' expectations suggested that the system is being able to chat .In addition to just fulfilling the given tasks was more natural to communicate with. The higher ratings of customer satisfaction of the interaction with the system indicated that this kind of dialogue management system is more engaging to interact with. This was supported by the fact that the duration of the interaction was longer in the hybrid task + chat condition even if not significantly. In general, hybrid task + chat based system received higher ratings in the questionnaires in comparison to the task only system. [8] Was the system that presented and evaluated the first approach of a fully autonomous robotic system which uses the novel combination of task-based and chat-style dialogue system? [8] Employed Reinforcement Learning (RL) to create a scalable and extensible approach of combining these tasked-based and chat-style modalities, while being able to easily enrich the used feature vector by including information from the robot's sensors in addition to verbal information. Experiments of the system used in [8] showed that participants found the proposed system more pleasant to interact with and had the feeling that it met their expectations better than a purely task-based version of the same approach. Usually or more often participants interacted longer with the robot without impeding the overall task efficiency, which indicated that this kind of robotic agent was more engaging than a purely task-based one. This presents a first step towards a holistic approach to HRI, being able to not only respond to utterances related to an a-priori defined set of tasks but also being able to chat with the human interaction partner.

4 Framework Description for Dialogue Manager

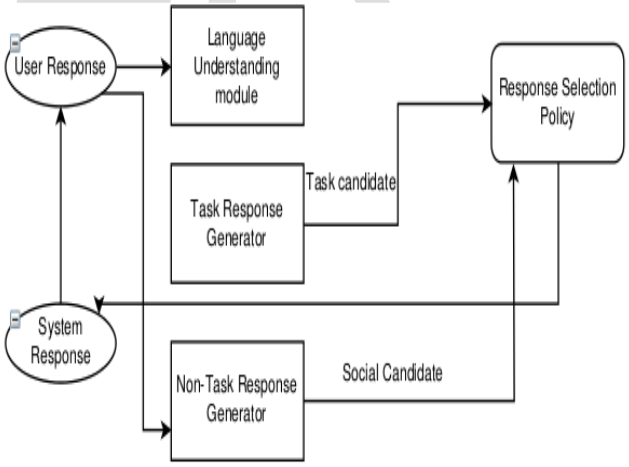

Fig. 2. Framework description

In the system described in[7]the framework contained four major components : language understanding module, task response generator, non-task response generator and response selection policy. Figure 2 describes the flow of information among the components . A user utterance sent by user is sent to both the language understanding module and the non-task response generator. The understanding model then extracted

the useful information that helped the task response generator to produce task-oriented candidates. Simultaneously, the non-task response generator also produced the non-task candidates. At , the response selection policy selected among all the candidates to produce a system response.

5 Results

In [8] Results of the data analysis revealed that participants were successful in performing the given tasks in both experimental conditions, with the average number of completed tasks being 3.98 (SD = 0.95) in the task -based condition and 3.93 (SD= 1.10) in the hybrid task+chat style condition (see Table II). The number of completed tasks was not significantly different between the two conditions (one-sided T-test, p = 0.83), the same holds true for the number of tasks per turn (p = 0.68), system turns (p = 0.28), and the number of actions performed by the robot (p = 0.25). To summaries, the findings showed no significant difference between the two conditions in terms of conversational efficiency, and on average the participants were not adversely affected by the style of conversation when performing their task.

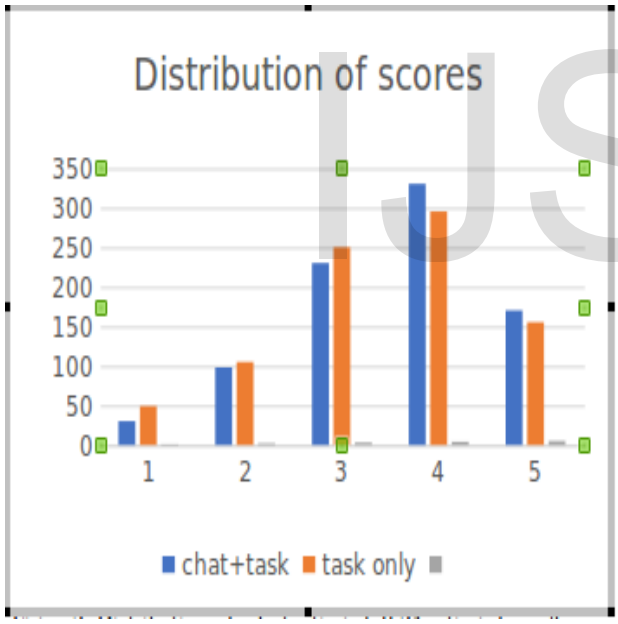

Fig. 3. Distribution of robot ratings (all 21 ratings from the questionnaire are taken into account) in the task-only and hybrid conditions.

TABLE1 :RESULTS OF CONVERSATIONAL EFFICIENCY AND DIALOGUE QUALITY IN TWO CONDITIONS. M DLUE, SD -STANDARD DEVIATION.

Measure	Task only	Chat + Task
Number of Tasks	M= 3.98 SD=0.95	M=3.93 SD=0.74
Number of Actions	M=5.85 SD=0.80	M=5.65 SD=0.74
System Turns	M=18.75 SD=7.74	M=21.05 SD=10.73
Tasks per turn	M=0.24 SD=0.11	M=0.23 SD=0.12
Duration sec.	M=203.99 SD=80.43	M=228.99 SD=117.02
Human Turns	M=20.03 SD=10.69	M=17.10 SD=7.10
Humnas turn per system turn	M=0.91 SD=0.07	M= 0.94 SD=0.05
Confidence of Speech Recognition	M=0.51 SD= 0.01	M=0.51 SD=0.02

REFERENCES

[1]. M. F. McTear, "Spoken dialogue technology: enabling the conversational user interface," CSUR, vol. 34, no. 1, pp. 90–169, 2002.
[2]. V. Rieser and O. Lemon, Reinforcement learning for adaptive dialogue systems. Springer, 2011.
[3]. S. Young, M. Gasic, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," Proceedings of the IEEE, vol. 101, no. 5, pp. 1160–1179, 2013.
[4]. T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on. IEEE, 2010, pp. 259– 266.
[5]. D. L. Chen and R. J. Mooney, "Learning to interpret natural languagessss navigation instructions from observations." in AAAI, vol. 2, 2011, pp. 1–2.
[6]. E. Ferreira and F. Lef`evre, "Reinforcement-learning based dialogue system for human-robot interactions with socially-inspired rewards," Computer Speech & Language, vol. 34, no. 1, pp. 256–274, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2015.03.007
[7]. Z. Yu, A. Black, and A. Rudnicky, "Learning conversational systems that interleave task and non-task content," in arXiv, no arXiv:1703.00099, 2017.
[8]. Ioannis Papaioannou ,Christian Dondrup ,Jekaterina Novikova ,Oliver Lemon,"Hybrid Chat and Task Dialogue for More Engaging HRI Using Reinforcement Learning",in2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) Lisbon, Portugal, Aug 28 - Sept 1, 2017
[9]. J.MacGlashan, "Burlap," 2016. [Online]. Available: http://burlap.cs.brown.edu/
[10]. A. G. Sutton, Richard SBarto, Reinforcement learning. MIT Press, 1998.
[11]. Texas instrumnets Application Report ,Implementing Speech Recognition algorithms on the TMS320C2xx Platform
[12]. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)
[13]. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[14]. I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[15]. K. Elissa, "Title of paper if known," unpublished.
[16]. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[17]. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[18]. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.