

Abstract

In accordance with our vision, we will propose a radically different architectural approach to learning generative models of human-level conversation. The motivation assumes that a conversation can be modeled as a game between two players attempting to maximize mutual engagement reward under a minimum information constraint. In the framework of reinforcement learning, we can view the Social Bot as an agent, π , hereafter referred to as the conversationalist, whose environment is the conversation. An additional constraint on π is that of one-shot learning, the ability to engage details pertinent to the conversation, drawing from a model-free source of external information. Under this paradigm, we develop the architecture of the Alexa Social Bot by proposing a novel procedure for information lookup and internalization via document embeddings in the regime of deep reinforcement learning with memory augmented metalearning.

1. Model

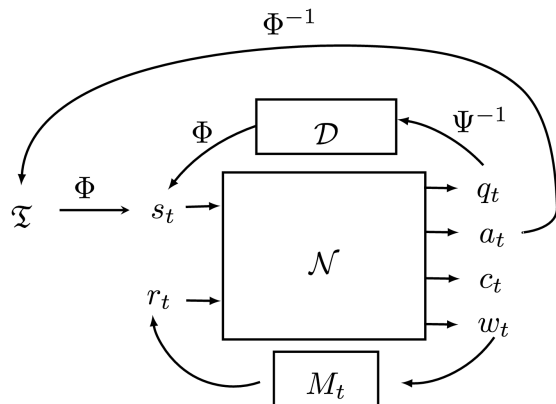


Figure 1. System architecture

Formally, we model a conversation as an *environment*, which is a Markov decision process, generally represented as a tuple $E = (\mathcal{S}, \mathcal{A}, \mathcal{R}, T, r)$ where \mathcal{S} is the state space, \mathcal{A} is the action space, \mathcal{R} is the reward

space, and T, r are the transition and reward functions respectively¹. Given an environment, a (deterministic) agent is a map $\pi : \mathcal{S} \rightarrow \mathcal{A}$ which takes actions conditioned on state space. In the context of a *conversational environment*, the state space is some featurization of the dialogue produced by the other party (the user) with whom Alexa communicates. Correspondingly the action space is the dialogue that Alexa produces in response. As aforementioned, a conversationalist is rewarded under r for maximizing engagement constrained to a minimal information theoretic throughput. Given the difficulty of defining such a function, we will learn r based on actual conversation as depicted in Section 3; thus we suppose that r is well defined.

Depicted in Figure 1 we separate the approximation of an optimal policy π into three components: dialogue representation, internal knowledge representation, and information lookup.

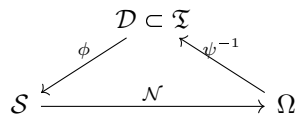
1. Dialogue representation is an embedding $\phi : \mathcal{T} \rightarrow \mathcal{S} = \mathbb{R}^n$ from dialogue text into some suitable vector representation of semantics conditioned on context. Respecting that this task is still an active area of research, we will model ϕ and its inverse by experimenting with a variety of different approaches such as thought vectors [6] and sparse distributed representations of syntax trees [10].
2. Our requirement that π actively learn, manipulate, and act on an internal episodic knowledge representation of its current conversation, is analogous to the problem of one-shot learning. Therefore we approximate an optimal π using a deep differentiable neural computer (DNC)[5][4], otherwise known as memory augmented network as motivated by the results of (Santoro et al., 2016).

In its simplest form the DNC is a map $\mathcal{N} : (s_t, r_t) \mapsto (a_t, c_t, w_t)$ where r_t, w_t are the inputs and outputs from *read* and *write* heads respectively. At every time step \mathcal{N} writes to an internal memory matrix M_t at a specific address as specified by the controller output c_t . The DNN representation of \mathcal{N} will be implemented using a combination of FC and LSTM layers so as to provide π with the ability to maintain an temporally conditioned representation of state. As per the training regime to be specified later, we will condition the DNC to use its memory augmentation to store an internal representation of facts learned during the course of a conversational episode in E .

¹For a high level overview of the MDP learning set up, the reader is referred to [1]

3. Lastly, we describe a novel approach to external information lookup which is central to the social bot's ability to communicate details about popular topics. We define a set of documents $\mathcal{D} \subset \mathfrak{T}$ and assume \mathcal{D} is *locally* homeomorphic to some linear manifold $\Omega \subset \mathbb{R}^m$. Then let $\psi : \mathcal{D} \rightarrow \Omega$ be that embedding. Finding such an embedding is a recent but promising development in NLP, with application to unsupervised search[2].

We wish to augment π with the capability to request external information in a format compatible with its internal knowledge representation. Thus we add to \mathcal{N} a query output $q_t \in \Omega$. Procedurally we find the nearest document in Ω , say d_t , and then use the text of the document as dialogue input into \mathcal{N} via ϕ ; that is,



Requiring that π only learn which class of documents are useful for responding in an engaging fashion is a much milder condition than assuming that \mathcal{N} will encode the details of every document in \mathcal{D} . This novel approach to information lookup primes the memory of \mathcal{N} , M_t so that the details of a particular topic of discussion are at hand during an episode, much in the same way that episodic conversational details are likewise written to M_t .

2. Training

With the core architectural components of the model developed, we draw from a state of the art development in RL, inverse reinforcement learning[7], to impart natural and expert level conversational skill on the Social Bot.

The principle idea behind inverse reinforcement learning is that behavioral cloning is limited to a biased distribution of only expert examples; that is, if the social bots model is trained to exactly replicate a human level conversationalist's response to different dialogues, then its ability to extrapolate beyond that known set of dialogues is unstable due to the severe nonlinearity of the model itself. In answer to this, inverse reinforcement learning attempts to infer the reward function of the agent being cloned. Knowing the reward function which an expert policy maximizes across a variety of states allows that policy to be cloned in a robust and dynamical fashion independent of the particular states in which it has acted; the cloned policy learns not about the particular examples to which it is

tuned, but about the general principles which govern the expert policy.

In the context of the social bot, learning to communicate by understanding the forces which govern human level conversation as opposed to mimicking the conversation itself has immediate and fundamental benefits which have not yet been utilized in natural language processing. Therefore we propose the following training regime for our model.

To train parts (1) and (2) in our model, we take a variety of data sources containing conversation as described in the technical approach document (Public Conversation Corpus, News Interview Corpus, Late Night Show Corpus) and fix a conversationalist and an environment; that is, in all instances of conversation we specify a party $\hat{\pi}$ whose maximal reward function, \hat{r} is to be learned, and a party who for all intents and purposes acts as E . Then with respect to the inversed party, we learn the action-value function $\hat{Q} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$ which estimates the expected future reward of $\hat{\pi}$ recursively

$$\hat{Q}(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim E} \left[\hat{r}(s_t, a_t) + \gamma \hat{Q}(s_{t+1}, \mu(s_{t+1})) \right]$$

In addition to learning the action-value function, as prescribed in our technical approach, we learn a critic network $\epsilon : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ in the style of Generative Adversarial Networks [3] which estimates the engagement of the conversation as per the Mechanical Turk dataset. Accompanying ϵ is a prior on the density of engagement labels in state space, $P(\epsilon|s)$, which we use to derive the following update rule.

As per the standard deterministic policy gradient [9] regime, assume that \mathcal{N} is parameterized by θ and then maximize the engagement and reward of a conversation by following the gradient

$$\theta' \leftarrow \theta + \left(\nabla_a \hat{Q}(s, a) + P(\epsilon|s) \nabla_a \epsilon \right) \nabla_\theta \mathcal{N}.$$

Lastly, to train information lookup (3) we use the Washington Post News Interview dataset which contains conversations and coresponding news articles (documents in \mathcal{D}). Holding constant a proportion of the parameters of θ , we reward the network for producing conversation which maximizes the reward function of the interviewer (inverse RL) after feeding the particular document in question through \mathcal{N} via ϕ . Limiting the dimensions in parameter space on which \mathcal{N} can learn requires the policy π to internalize pertinent information in its memory M_t . This regime is analogous to that depicted in (Santoro et al., 2016).

References

- [1] Richard Bellman. “A Markovian Decision Process”. In: *Indiana Univ. Math. J.* 6 (4 1957), pp. 679–684. ISSN: 0022-2518.
- [2] Andrew M. Dai, Christopher Olah, and Quoc V. Le. “Document Embedding with Paragraph Vectors”. In: *CoRR* abs/1507.07998 (2015). URL: <http://arxiv.org/abs/1507.07998>.
- [3] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680. URL: <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>.
- [4] Alex Graves, Greg Wayne, and Ivo Danihelka. “Neural Turing Machines”. In: *CoRR* abs/1410.5401 (2014). URL: <http://arxiv.org/abs/1410.5401>.
- [5] Alex Graves et al. “Hybrid computing using a neural network with dynamic external memory”. In: *Nature* (2016).
- [6] Ryan Kiros et al. “Skip-Thought Vectors”. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 3294–3302. URL: <http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf>.
- [7] Andrew Y Ng, Stuart J Russell, et al. “Algorithms for inverse reinforcement learning.” In: *ICML*. 2000, pp. 663–670.
- [8] Adam Santoro et al. “One-shot Learning with Memory-Augmented Neural Networks”. In: *CoRR* abs/1605.06065 (2016). URL: <http://arxiv.org/abs/1605.06065>.
- [9] David Silver et al. “Deterministic Policy Gradient Algorithms”. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. Ed. by Tony Jebara and Eric P. Xing. JMLR Workshop and Conference Proceedings, 2014, pp. 387–395. URL: <http://jmlr.org/proceedings/papers/v32/silver14.pdf>.
- [10] Dani Yogatama et al. “Learning Word Representations with Hierarchical Sparse Coding”. In: *CoRR* abs/1406.2035 (2014). URL: <http://arxiv.org/abs/1406.2035>.