







Deep Reinforcement Learning for Conversational Agents

Nabiha Asghar

MAT-Lab Group Meeting, University of Waterloo

7th November, 2016

The 'Bot Revolution'

- 2016 is being touted as the **Year of the Bots** and **Year of Conversational Commerce**
- Several startups have emerged in 2015-2016:
 -  Chatfuel
 -  msg.ai
 -  message.io Message.io
 -  Birdly
- Recent bot development frameworks:
 -  Facebook
 -  Kik Interactive

The 'Bot Revolution'

Domains:

Advertising, customer services, online shopping, banking, entertainment, news, etc.

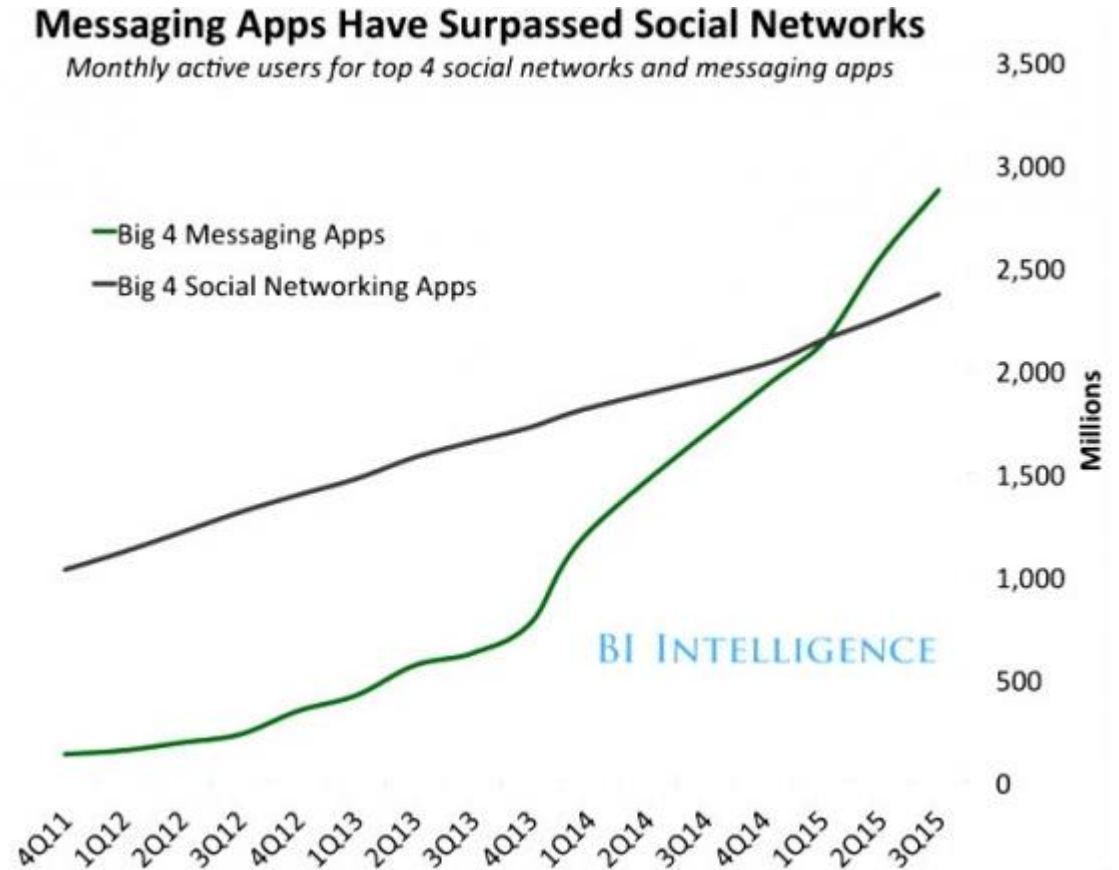
- Wall Street Journal
- **Fox News**
- CNN
- TechCrunch
- Foursquare
- Forbes
- **Uber**
- **Ikea**
- Sephora
- Bank of America
- **Pizza Hut**
- H&M
- WebMD
- **RBS**
- Domino's
- American Express
- Tommy Hilfiger
- **Netflix**
- **London City Airport**
- ToyTalk
- NBC
- Airbnb
- Norton
- **Spotify**
- Tesco
- Evernote
- The Weather Channel
- **The White House**

Bots: Why the Sudden Obsession?

- People don't want to download new apps

Americans download zero apps per month on average

- Superior Human Computer Interface



CAs: Basic Types

- Task-oriented (Domain specific) versus Open Domain
- Retrieval Based versus Generative Models
- Most current systems in production use:
 - retrieval from a database of responses
 - hand-built templates and slot-filling

Generative CAs: Key Challenges

1. Understanding the semantics/meaning of textual prompts
2. Generating meaningful responses to these prompts
3. Maintaining coherence and contextual relevance during a conversation

CAs: Generative Models

- Lately receiving a lot of attention: **Deep Neural Networks**
 - success in Machine Translation
 - can map sequences to sequences (**Seq2Seq** model)
 - a natural way to capture semantics, grammar and context in NLP

Sequence-2-Sequence Model

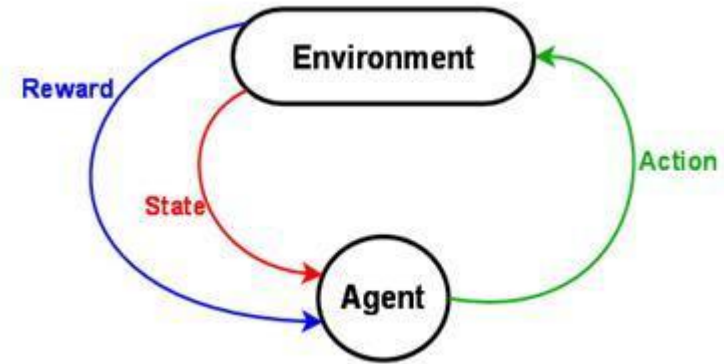
- **me:** Hello?
bot: Hi.
- **me:** How are you?
bot: I'm fine.
- **me:** What's your name?
bot: It's hard to describe.
- **me:** How so?
bot: I'm not sure.
- **me:** What color is the sky?
bot: It's blue.
- **me:** What is the purpose of life?
bot: A gift.
- **me:** Are you intelligent?
bot: Yes, well...
- **me:** Are you a machine?
bot: That's a lie.
- **me:** Are you human?
bot: No, i'm not.
- **me:** What are you?
bot: I'm not sure.
- **me:** Do you plan on taking over the world?
bot: No, i don't.

- **Limitations**

- Generic dull responses (*I don't know, I'm not sure*)
- Infinite loops of repetition
- Short-sighted answers
- Cannot keep the users engaged
- Mutually inconsistent responses

Enter: Reinforcement Learning

- **Idea:** Use Reinforcement Learning (RL) in conjunction with DNNs to solve the NLP challenges



- RL's success:
 - Atari Games, GO
 - MDP/POMDP models to build domain-specific dialogue systems

Today's Discussion

1. Li *et al.* (2016) *“Deep reinforcement learning for dialogue generation”*
2. Su *et al.* (2016) *“Continuously learning neural dialogue management”*
3. Zhao and Eskenazi. (2016) *“Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning”*

Li *et al.* “DRL for dialogue generation”

- Goals:
 1. model the long-term success of a dialogue through RL
 2. produce interesting, diverse, more interactive responses

- Method:
 1. Simulate a dialogue between two virtual agents
 2. Reward sequences that display **informativity**, **coherence** & **ease of answering** (to mimic the true goals of a conversation)

Li *et al.*: Model Overview

- **Backbone:** LSTM encoder-decoder, policy gradient method
- **Stage 1 -- Supervised Learning:** Seq2Seq with Attention, yields $P_{\text{seq2seq}}(s|a)$
 - OpenSubtitles, 80m source-target pairs
 - Cross Entropy Loss function
- **Stage 2 – Supervised Learning:** Change the objective function to **Maximum Mutual Information (MMI)**.
- **Stage 3 -- Reinforcement Learning**
 - Let two agents take turns in a dialogue
 - Maximize the expected future reward $J_{RL}(\theta) = \mathbb{E}_{P_{RL}(a_{1:T})}[\sum_{i=1}^{i=T} R(a_i, [p_i, q_i])]$
where $R(a_i, [p_i, q_i])$ is a weighted combination of informativity, coherence, ease of answering

Aside: Maximum Mutual Information (MMI)

- A diversity promoting objective function
- Traditional Seq2Seq models:

$$\hat{T} = \arg \max_T \{ \log p(T|S) \}$$

- MMI:

$$\hat{T} = \arg \max_T \{ (1 - \lambda) \log p(T|S) + \lambda \log p(S|T) \}$$

- Avoids responses that enjoy unconditionally high probability
- Biases towards the responses that are specific to the given input

Li *et al.*: Reward Function

$$J_{RL}(\theta) = \mathbb{E}_{p_{RL}(a_{1:T})} \left[\sum_{i=1}^{i=T} R(a_i, [p_i, q_i]) \right]$$

$$R(a_i, [p_i, q_i]) = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3$$

$$r_1 = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{\text{seq2seq}}(s|a) \rightarrow \text{Ease of answering}$$

$$r_2 = -\log \cos(h_{p_i}, h_{p_{i+1}}) = -\log \cos \frac{h_{p_i} \cdot h_{p_{i+1}}}{\|h_{p_i}\| \|h_{p_{i+1}}\|} \rightarrow \text{Information Flow}$$

$$r_3 = \frac{1}{N_a} \log p_{\text{seq2seq}}(a|q_i, p_i) + \frac{1}{N_{q_i}} \log p_{\text{seq2seq}}^{\text{backward}}(q_i|a) \rightarrow \text{Semantic Coherence}$$

Reward: Ease of Answering

$$r_1 = -\frac{1}{N_S} \sum_{s \in \mathcal{S}} \frac{1}{N_s} \log p_{\text{seq2seq}}(s|a)$$

- Compute the negative log likelihood of responding to a message with a dull response
- \mathcal{S} = set of dull responses (manually constructed)
- $N_S = |\mathcal{S}|$

Reward: Information Flow

$$r_2 = -\log \cos(h_{p_i}, h_{p_{i+1}}) = -\log \cos \frac{h_{p_i} \cdot h_{p_{i+1}}}{\|h_{p_i}\| \|h_{p_{i+1}}\|}$$

- Penalize semantic similarity between consecutive turns of an agent
- $h_{p_i}, h_{p_{i+1}}$ = encodings of two consecutive messages by an agent

Reward: Semantic Coherence

$$r_3 = \frac{1}{N_a} \log p_{\text{seq2seq}}(a|q_i, p_i) + \frac{1}{N_{q_i}} \log p_{\text{seq2seq}}^{\text{backward}}(q_i|a)$$

- Avoid responses which are highly rewarded but are ungrammatical or incoherent
- Consider the mutual information between the action a and previous turns in the history

Li *et al.*: Model Overview

- **Backbone:** LSTM encoder-decoder, policy gradient method
- **Stage 1 -- Supervised Learning:** Seq2Seq with Attention, yields $P_{\text{seq2seq}}(s|a)$
 - OpenSubtitles, 80m source-target pairs
 - Cross Entropy Loss function
- **Stage 2 – Supervised Learning:** Change the objective function to **Maximum Mutual Information (MMI)**.
 - Treat MMI as reward of the source and each possible response.
- **Stage 3 -- Reinforcement Learning**
 - Let two agents take turns in a dialogue
 - Maximize the expected future reward $J_{RL}(\theta) = \mathbb{E}_{P_{RL}(a_{1:T})}[\sum_{i=1}^{i=T} R(a_i, [p_i, q_i])]$
where $R(a_i, [p_i, q_i])$ is a weighted combination of informativity, coherence, ease of answering

Results

Input Message	Mutual Information Model	RL
How old are you?	I'm 16.	I'm 16, why are you asking?
What's your full name?	I have no idea.	What's yours?
I don't want to go home tonight.	Really?	Why?
Do you have any feelings for me?	I don't know what you are talking about.	I don't want to hurt your feelings.
How much time do you have here?	Not long enough. Sorry, sir.	Ten seconds.
Shall we get started?	Of course!	Yes. We've got a lot of work to do here.
Do you play football?	No, i don't	Yes. I love football!
We'd have to talk to him.	I mean, he's a good guy	About what ?
How come you never say it?	Because I don't want to hurt you.	I don't think it's a good idea to say it.

- More interactive responses
- Generally longer responses

Today's Discussion

1. Li *et al.* (2016) *“Deep reinforcement learning for dialogue generation”*
 - Open domain CAs
 - Supervised Learning: Combine Seq2Seq with an MMI objective function,
 - Reinforcement Learning: two virtual agents, policy-gradient, 3-pronged reward structure
2. Su *et al.* (2016) *“Continuously learning neural dialogue management”*
3. Zhao and Eskenazi. (2016) *“Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning”*



Su et al. “Continuously learning neural dialogue management”

- Live telephone-based spoken dialogue system
- provides restaurant information for Cambridge UK

- Dialogue Acts:

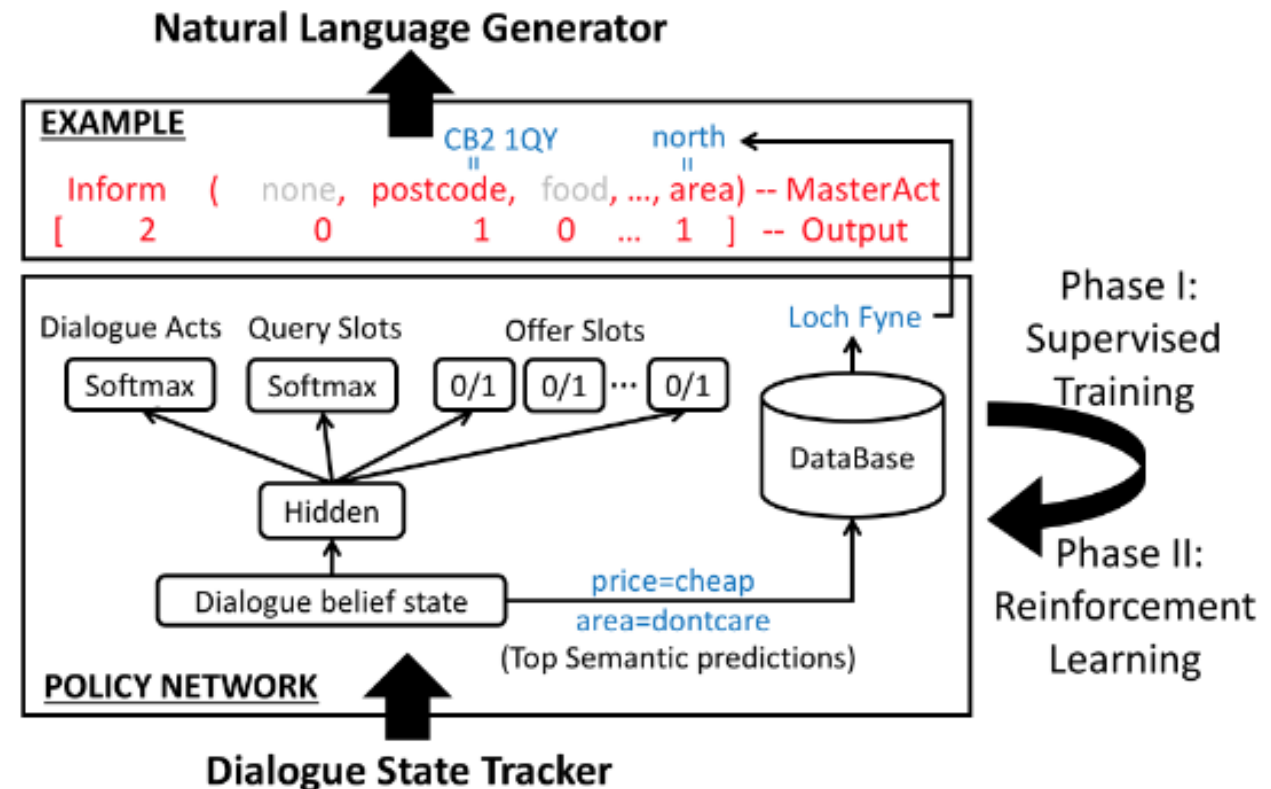
{request, offer, confirm, select, bye}

- Query (search constraint):



{food, price range, area, none}

- Offer slots:



binary predictions within that offer



Today's Discussion

1. Li *et al.* (2016) *“Deep reinforcement learning for dialogue generation”*
 - Open domain conversations
 - Supervised Learning: Combine Seq2Seq with an MMI objective function,
 - Reinforcement Learning: two virtual agents, policy-gradient, 3-pronged reward structure
2. Su *et al.* (2016) *“Continuously learning neural dialogue management”*
 - Domain specific (restaurant; single layer DNN for slot-filling)
 - Supervised Learning: cross-entropy loss
 - Reinforcement Learning: simulated/real users, maximize expected reward
3. Zhao and Eskenazi. (2016) *“Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning”*

Today's Discussion

1. Li et al. (2016) *“Deep reinforcement learning for dialogue generation”*
 - Open domain conversations
 - Supervised Learning: Combine Seq2Seq with an MMI objective function,
 - Reinforcement Learning: two virtual agents, policy-gradient, 3-pronged reward structure
2. Su et al. (2016) *“Continuously learning neural dialogue management”*
 - Domain specific (restaurant; single layer DNN for slot-filling)
 - Supervised Learning: cross-entropy loss
 - Reinforcement Learning: simulated/real users, maximize expected reward
3. Zhao and Eskenazi. (2016) *“Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning”*
 - SL+RL using Deep Recurrent Q Networks (DRQN)