Learning Dialogue Generation using Human Feedback

Nabiha Asghar

MAT-Lab Group Meeting, January 11th, 2017

Motivation

- Conversational Agents are all the rage these days
- 2016: Year of the Bots, Year of Conversational Commerce
- Generative Dialogue Models based on Deep Neural Networks
    - Recurrent Networks / LSTM Networks: language modeling (2010, 2012)
    - Sequence to Sequence Framework: machine translation, text summarization, dialogue (Google, 2014)
    - Memory Networks: question answering, language modeling, dialogue (Facebook AI, 2015)
- Limitations of Offline Supervised Learning
    - Short and dull responses, not interesting/engaging
    - Irrelevant, contextually inappropriate, incorrect (if domain-specific)

Goals

- Idea: learn conversational skills like humans, through continuous interaction/feedback
  - Reinforcement Learning, Active Learning with humans in the loop
  - no need to label/annotate huge datasets
  - avoid explicit incorporation of interestingness, relevance, diversity in responses
- Need to explore different types of human involvement/feedback as well as learning strategies
  - "Dialog-based Language Learning", Jason Weston (Facebook AI), NIPS, December 2016
  - "Dialogue Learning with Human-In-The-Loop", Li et al. (Facebook AI), submitted to ICLR 2016
  - Simple QA on short passages or a set of facts

10 Modes of Supervision (Weston, 2016)

human/expert
bot/learner

Task 1: Imitating an Expert Student
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:kitchen
Where is John? A:bathroom

Task 2: Positive and Negative Feedback
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:playground
No, that's incorrect.
Where is John? A:bathroom
Yes, that's right! (+)

Task 3: Answers Supplied by Teacher
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:bedroom
No, the answer is kitchen.
Where is John? A:bathroom
Correct! (+)

Task 4: Hints Supplied by Teacher
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:bathroom
No, they are downstairs.
Where is John? A:kitchen
No, they are upstairs.

10 Modes of Supervision (cont'd)

human/expert
bot/learner

Task 5: Supporting Facts Supplied by Teacher
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:kitchen
Yes, that's right! (+)
Where is John? A:hallway
No, because John moved to the bathroom.

Task 6: Partial Feedback
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:kitchen
Yes, that's right!
Where is John? A:bathroom
Yes, that's correct! (+)

Task 7: No Feedback
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:kitchen
Yes, that's right!
Where is John? A:bathroom
Yes, that's correct!

Task 8: Imitation and Feedback Mixture
Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:kitchen
Where is John? A:bathroom
That's right! (+)

10 Modes of Supervision (cont'd)

human/expert
bot/learner

Task 9: Asking For Corrections

Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:kitchen
Yes, that's right! (+)
Where is John? A:hallway
No, that's not right. A:Can you help me?
Bathroom.

Task 10: Asking For Supporting Facts

Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? A:kitchen
Yes, that's right! (+)
Where is John? A:hallway
No, that's not right. A:Can you help me?
A relevant fact is John moved to the bathroom.

Memory Networks

Hop #1:

$$o_1 = \sum_i p_i^1 m_i, \quad p_i^1 = \text{Softmax}(q^\top m_i).$$

$$u_1 = R_1(o_1 + q)$$

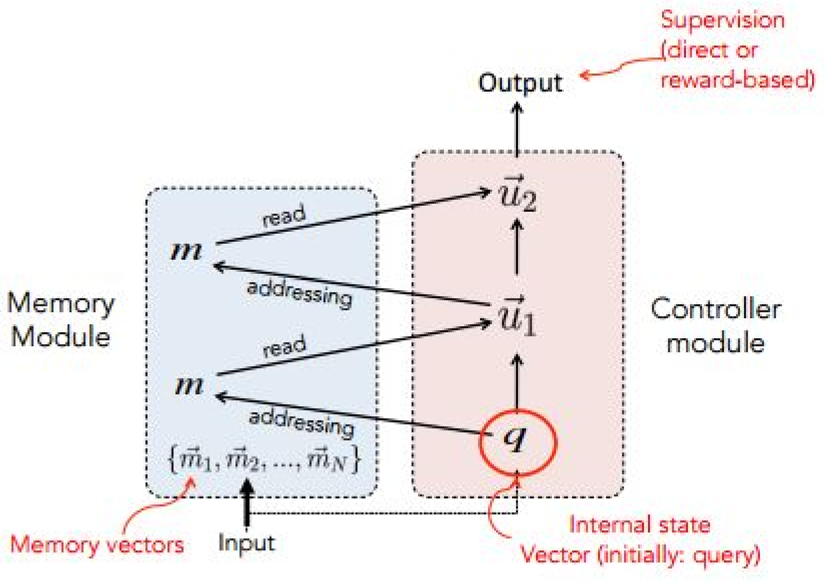
Hop #2:

$$o_2 = \sum_i p_i^2 m_i, \quad p_i^2 = \text{Softmax}(u_1^\top m_i)$$

$$u_2 = R_2(o_2 + u_1)$$

Final output:

$$\hat{a} = \text{Softmax}(u_2^\top A y_1, \ldots, u_2^\top A y_C)$$



Learning Models

- Imitation Learning
  - Essentially supervised learning (message-context-response triples, cross entropy loss function)

Learning Models

- **Imitation Learning**
  - Essentially supervised learning (message-context-response triples, cross entropy loss function)
- **Reward-based Imitation (RBI)**
  - Supervised learning (with cross entropy loss) only on rewarded actions. Discard the rest

Learning Models

- **Imitation Learning**
  - Essentially supervised learning (message-context-response triples, cross entropy loss function)
- **Reward-based Imitation (RBI)**
  - Supervised learning (with cross entropy loss) only on rewarded actions. Discard the rest
- **Forward Prediction (FP)**
  - Given an utterance x from Speaker #1 and answer a by the Learner, predict the response $\bar{x}$ of Speaker #1

Forward Prediction

Hop #1:

$$o_1 = \sum_i p_i^1 m_i, \quad p_i^1 = \text{Softmax}(q^\top m_i).$$

$$u_1 = R_1(o_1 + q)$$

Hop #2:

$$o_2 = \sum_i p_i^2 m_i, \quad p_i^2 = \text{Softmax}(u_1^\top m_i)$$

$$u_2 = R_2(o_2 + u_1)$$

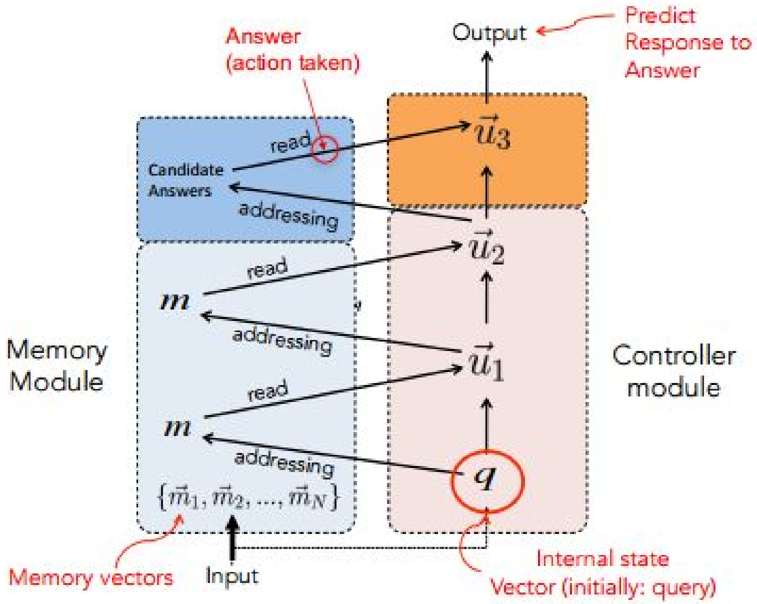
Hop #3:

$$o_3 = \sum_i p_i^3 (A y_i + \beta^* [a = y_i]), \quad p_i^3 = \text{Softmax}(u_2^\top A y_i)$$

$$u_3 = R_3(o_3 + u_2)$$

Final output:

$$\hat{x} = \text{Softmax}(u_3^\top A \bar{x}_1, \ldots, u_3^\top A \bar{x}_{\bar{C}})$$



Forward Prediction

Hop #1:

$$o_1 = \sum_i p_i^1 m_i, \quad p_i^1 = \text{Softmax}(q^\top m_i).$$

$$u_1 = R_1(o_1 + q)$$

Hop #2:

$$o_2 = \sum_i p_i^2 m_i, \quad p_i^2 = \text{Softmax}(u_1^\top m_i)$$

$$u_2 = R_2(o_2 + u_1)$$

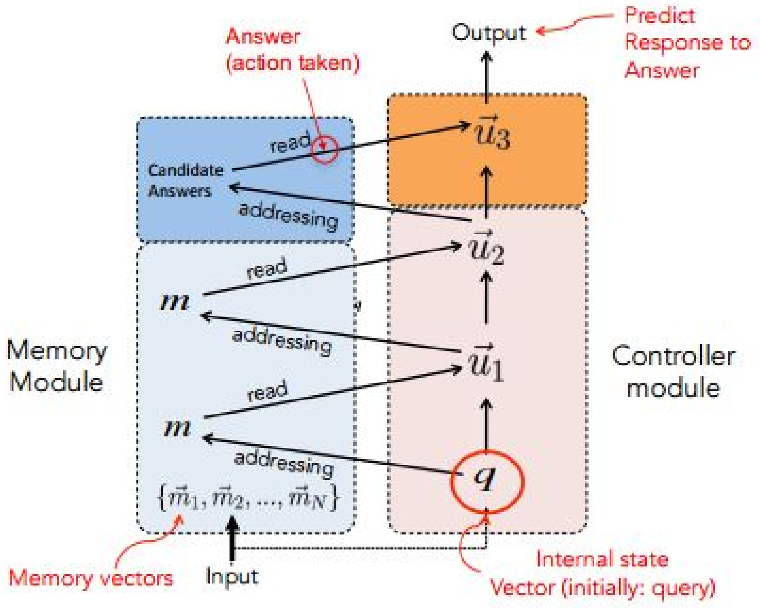
Hop #3:

$$o_3 = \sum_i p_i^3 (Ay_i + \beta^*[a = y_i]), \quad p_i^3 = \text{Softmax}(u_2^\top Ay_i)$$

$$u_3 = R_3(o_3 + u_2)$$

Final output:

$$\hat{x} = \text{Softmax}(u_3^\top A\bar{x}_1, \ldots, u_3^\top A\bar{x}_{\bar{C}})$$



d-dim vector, represents in o_3 the action that was actually selected

Forward Prediction

Hop #1:

$$o_1 = \sum_i p_i^1 m_i, \quad p_i^1 = \text{Softmax}(q^\top m_i).$$

$$u_1 = R_1(o_1 + q)$$

Hop #2:

$$o_2 = \sum_i p_i^2 m_i, \quad p_i^2 = \text{Softmax}(u_1^\top m_i)$$

$$u_2 = R_2(o_2 + u_1)$$

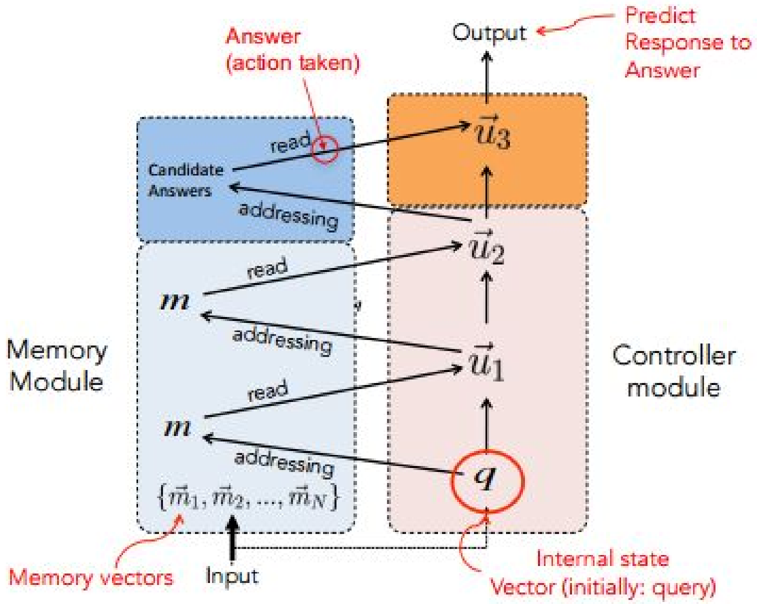
Hop #3:

$$o_3 = \sum_i p_i^3 (Ay_i + \beta^* [a = y_i]), \quad p_i^3 = \text{Softmax}(u_2^\top Ay_i)$$

$$u_3 = R_3(o_3 + u_2)$$

Final output:

$$\hat{x} = \text{Softmax}(u_3^\top A\bar{x}_1, \ldots, u_3^\top A\bar{x}_{\bar{C}})$$



d-dim vector, represents in o_3 the action that was actually selected

a way to compare the most likely answers to x with the given ans ‘a’

Learning Models

- **Imitation Learning**
  - Essentially supervised learning (message-context-response triples, cross entropy loss function)
- **Reward-based Imitation (RBI)**
  - Supervised learning (with cross entropy loss) only on rewarded actions. Discard the rest
- **Forward Prediction (FP)**
  - Given an utterance x from Speaker #1 and answer a by the Learner, predict the response $\bar{x}$ of Speaker#1
  - Cross-entropy loss between $\bar{x}$ and $\hat{x}$.

Learning Models

- **Imitation Learning**
  - Essentially supervised learning (message-context-response triples, cross entropy loss function)
- **Reward-based Imitation (RBI)**
  - Supervised learning (with cross entropy loss) only on rewarded actions. Discard the rest
- **Forward Prediction (FP)**
  - Given an utterance x from Speaker #1 and answer a by the Learner, predict the response $\bar{x}$ of Speaker#1
  - Cross-entropy loss between $\bar{x}$ and $\hat{x}$.
- **Reward-based Imitation + Forward Prediction (RBI+FP)**
  - Mixture of 2 and 3. Shared weights. Use both criteria for gradient descent.

Data

- bAbI dataset: short stories from a simulated world followed by questions
- For each of the 10 supervision tasks, consider a fixed policy for answering questions which gets questions correct with probability π_{acc} .

Evaluation on bAbI dataset

		MemN2N *imitation learning*			MemN2N *reward-based imitation (RBI)*			MemN2N *forward prediction (FP)*			MemN2N *RBI + FP*		
Supervision Type	π_{acc} =	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01
1 - Imitating an Expert Student		100	100	100	100	100	100	23	30	29	99	99	100
2 - Positive and Negative Feedback		79	28	21	99	92	91	93	54	30	99	92	96
3 - Answers Supplied by Teacher		83	37	25	99	96	92	99	96	99	99	100	98
4 - Hints Supplied by Teacher		85	23	22	99	91	90	97	99	66	99	100	100
5 - Supporting Facts Supplied by Teacher		84	24	27	100	96	83	98	99	100	100	99	100
6 - Partial Feedback		90	22	22	98	81	59	100	100	99	99	100	99
7 - No Feedback		90	34	19	20	22	29	100	98	99	98	99	99
8 - Imitation + Feedback Mixture		90	89	82	99	98	98	28	64	67	99	98	97
9 - Asking For Corrections		85	30	22	99	89	83	23	15	21	95	90	84
10 - Asking For Supporting Facts		86	25	26	99	96	84	23	30	48	97	95	91
Number of completed tasks ($\geq 95\%$)		1	1	1	9	5	2	5	5	4	10	8	8

Table 1: Test accuracy (%) on the Single Supporting Fact bAbI dataset for various supervision approachess (training with 1000 examples on each) and different policies π_{acc}. A task is successfully passed if $\geq 95\%$ accuracy is obtained (shown in blue).

Interesting Result: Forward Prediction (predicting the teacher's feedback) works nicely, even though it doesn't use human-labeled rewards

Evaluation on bAbI dataset

	MemN2N *imitation learning*			MemN2N *reward-based imitation (RBI)*			MemN2N *forward prediction (FP)*			MemN2N *RBI + FP*		
Supervision Type π_{acc} =	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01	0.5	0.1	0.01
1 - Imitating an Expert Student	100	100	100	100	100	100	23	30	29	99	99	100
2 - Positive and Negative Feedback	79	28	21	99	92	91	93	54	30	99	92	96
3 - Answers Supplied by Teacher	83	37	25	99	96	92	99	96	99	99	100	98
4 - Hints Supplied by Teacher	85	23	22	99	91	90	97	99	66	99	100	100
5 - Supporting Facts Supplied by Teacher	84	24	27	100	96	83	98	99	100	100	99	100
6 - Partial Feedback	90	22	22	98	81	59	100	100	99	99	100	99
7 - No Feedback	90	34	19	20	22	29	100	98	99	98	99	99
8 - Imitation + Feedback Mixture	90	89	82	99	98	98	28	64	67	99	98	97
9 - Asking For Corrections	85	30	22	99	89	83	23	15	21	95	90	84
10 - Asking For Supporting Facts	86	25	26	99	96	84	23	30	48	97	95	91
Number of completed tasks ($\geq 95\%$)	1	1	1	9	5	2	5	5	4	10	8	8

Table 1: Test accuracy (%) on the Single Supporting Fact bAbI dataset for various supervision approachess (training with 1000 examples on each) and different policies π_{acc}. A task is successfully passed if $\geq 95\%$ accuracy is obtained (shown in blue).

“Dialogue Learning with Human-In-The-Loop”, Li *et al.* 2016

- Use Reinforcement Learning policy instead of fixed policies π_{acc}
- Online and incremental learning (i.e. weights updated after each reward is received)

“Dialogue Learning with Human-In-The-Loop”, Li *et al.* 2016

- Use Reinforcement Learning policy instead of fixed policies π_{acc}
- Online and incremental learning (i.e. weights updated after each reward is received)
- Consider Task 6 (“partial feedback”) only: the teacher replies with positive textual feedback (6 possible templates) when the bot answers correctly, and positive reward is given only 50% of the time. When the bot is wrong, the teacher gives textual feedback containing the answer.

“Dialogue Learning with Human-In-The-Loop”, Li *et al.* 2016

- Use Reinforcement Learning policy instead of fixed policies π_{acc}
- Online and incremental learning (i.e. weights updated after each reward is received)
- Consider Task 6 (“partial feedback”) only: the teacher replies with positive textual feedback (6 possible templates) when the bot answers correctly, and positive reward is given only 50% of the time. When the bot is wrong, the teacher gives textual feedback containing the answer.
- Learning Models: RBI, FP, and REINFORCE (0 and 1)
- Difference between RBI and REINFORCE: former imitates correct behaviour only, latter leverages incorrect behaviour too

“Dialogue Learning with Human-In-The-Loop”, Li *et al.* 2016