Deep Reinforcement Learning for Dialogue Generation
Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao
and Dan Jurafsky
Seq2Seq for Dialogue

Encode previous message(s) into vector

Decode vector into response

How are you

I am fine
Seq2Seq for Dialogue

Encode previous message(s) into vector

Decode vector into response

Train by maximizing $p(\text{response}|\text{input})$

where the response is produced by a human
Problems with Seq2Seq

How old are you?
Problems with Seq2Seq

How old are you?

16?

I’m 16
How old are you?
16?
You don’t know what you’re saying
I don’t know what you’re talking about

I’m 16
I don’t know what you’re talking about
You don’t know what you’re saying
Problems with Seq2Seq

- How old are you?
- 16?
- You don’t know what you’re saying
- You don’t know what you’re saying
- I’m 16
- I don’t know what you’re talking about
- I don’t know what you’re talking about

reasonable, but unhelpful

generic
probable response != good response
What is a good response?

- Reasonable
- Nonrepetitive
- Easy to answer
What is a good response?

- **Reasonable**: $p(\text{response}|\text{input})$ is high according to seq2seq model

- **Nonrepetitive**: similarity between response and previous messages is low

- **Easy to answer**: $p(\text{“i don’t know”}|\text{response})$ is low

Scoring function: $R(\text{response}) = \text{reasonable\_score} + \text{nonrepetitive\_score} + \text{easy\_to\_answer\_score}$
Reinforcement Learning

Learn from rewards instead of from examples

1. Encode input into a vector

How are you
Reinforcement Learning

Learn from rewards instead of from examples

2. Have the system generate a response
Reinforcement Learning

Learn from rewards instead of from examples

How are you - I don’t know

3. Receive reward $R(\text{response})$

- Train system to maximize reward

$R = -5$
### Quantitative Results

| Setting                                      | Gain  |
|----------------------------------------------|-------|
| single-turn general quality                  | 0.02  |
| single-turn ease to answer                   | 0.12  |
| multi-turn general quality                   | 0.17  |
Qualitative Results

How old are you?

I thought you were 12

I’m 16. Why are you asking?

What made you think so?
Qualitative Results

How old are you?

I thought you were 12

I’m 16. Why are you asking?

What made you think so?

You don’t know what you’re saying

I don’t know what you’re talking about
Conclusion

- Reinforcement learning useful when we want our model to do more than produce a probable human label

- Many more application of RL to NLP!
  - Information extraction, question answering, task-oriented dialogue, coreference resolution, and more