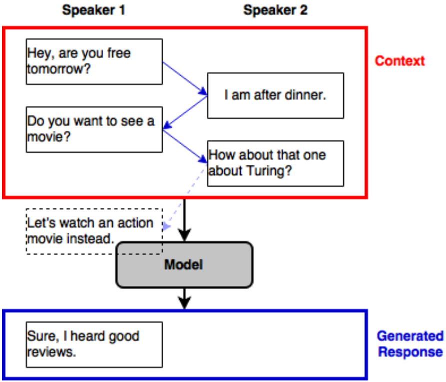
# Training a Discriminator to Compare Generative Dialogue Model

Nicolas A.-Gontier

Reasoning and Learning Lab
McGill University

COMP 652 - Final Project Presentation
April 19th

# Dialogue Evaluation


Speaker 1
Speaker 2
Hey, are you free tomorrow?
I am after dinner.
Do you want to see a movie?
How about that one about Turing?
Context
Let's watch an action movie instead.
Model
Sure, I heard good reviews.
Generated Response

- **Problem:** How can we measure the quality of a response?
  - Previous work: A Dialogue Evaluation Model (R. Lowe, M. Noseworthy, I.V. Serban, N. A.-Gontier, Y. Bengio, and J. Pineau)
  - Adversarial evaluation: this presentation
  - Really a combination of both

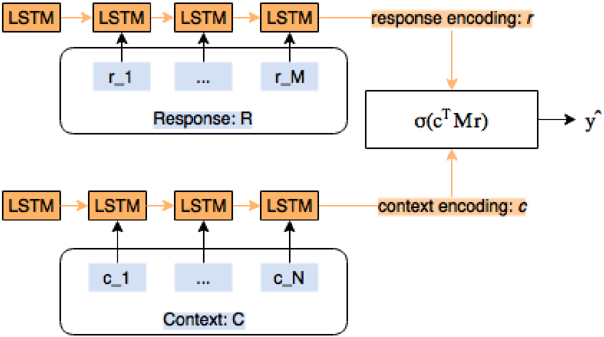
# Dialogue Evaluation

- **Goal**: Train a model that can discriminate between generated responses and true responses
- We will use the Adversarial framework from GANs
- Data-set used: On-line Tweets (˜700,000 conversations for each model)
    - Random sampler (y=0)
    - TF-IDF context-response similarity (y=0)
    - The HRED [1] model with random sampling (y=0)
    - The HRED model with beam-5 sampling (y=0)
    - The VHRED [2] model with random sampling (y=0)
    - The VHRED model with beam-5 sampling (y=0)
- We oversampled true responses (y=1) in training set
- Work inspired by "Adversarial Evaluation of Dialogue Models" (A. Kannan, O. Vinyals, 2017)

[1] I.V. Serban et al. (2015)

[2] I.V. Serban et al. (2016)

# Dual-Encoder Discriminator Network



- **Objective**: minimize the binary cross-entropy:
  $J(\hat{y}, y) = -y * \log P(\hat{y}) - (1 - y) * \log(1 - P(\hat{y}))$

# Results

- Test set scores[3]:

| Model | True negative | False positive |
|---|---|---|
| Random | 67% | 33% |
| TF-IDF | 69% | 31% |
| HRED (random) | 53% | 47% |
| HRED (beam-5) | 99% | 1% |
| VHRED (random) | 62% | 38% |
| VHRED (beam-5) | 96% | 4% |

| True positive | False negative |
|---|---|
| 70% | 30% |

[3] Network is still improving...

# Conclusions & Future work

## Conclusions

- Discriminator networks are strong on generative models.
- Hard to discriminate doesn't mean it's a good dialogue system. Thus the need to combine this with a scoring machine like ADEM [a]

[a] A Dialogue Evaluation Model (R. Lowe, M. Noseworthy, I.V. Serban, N. A.-Gontier, Y. Bengio, and J. Pineau)

## Future work

- Try different embeddings like Tweet2Vec or HRED embeddings.
- Train the discriminator only on true responses and **one** model responses. Maybe an ensemble of these discriminators (each which is only good at a single model type) would do better.