Dialogue Evaluation via Offline Reinforcement Learning and Emotion Prediction

Dr. Nurul Lubis
Dialog Systems and Machine Learning
Heinrich-Heine University Düsseldorf
Why dialogue?

- Natural language has evolved to facilitate communication
- Dialogue is a prime interactive NLP task
  - Turing poses dialogue as a core AI problem (Turing, 1950)

https://prowritingaid.com/
What makes it challenging?

- Infinite possibilities of how a dialogue can go
  - We can always think of a dialogue that was never produced before
  - Can not be solved with handcrafted rules

- Dialogue can be viewed as an AI-complete problem (Shapiro, 1992)
  - Recognition, reasoning, and generation
What are good dialogue properties?

• Understanding the user
• Handling different (new) topics in a dynamic world
• Understanding emotions and sentiment
• Responding in a human-like manner
• Responding sensibly, intelligently and fluently
• Providing personalised outputs
• …
Dialogue systems are becoming more ubiquitous

What makes one system better than the other?
Dialogue systems

Task oriented dialogues (ToD)
• Centered around fulfilling user goals
• Domain specific

Chit-chat dialogues
• Typical aims are user engagement or entertainment
• Open-ended

I’m looking for a nice restaurant in the center of town
What type of food would you like to have?

How many pets do you have?
I have two dogs and a cat. I love animals.
User-centered criteria (Walker et al., 1997; Lee and Eskenazi, 2012; Ultes et al., 2017)
• Time- and cost-intensive
• Hard to compare
Interactive User Simulator

Interactive user simulator (US) (Schatzmann, 2008; Lin et al., 2021)
  • Not straightforward to build
Automatic evaluation with static corpora

Can we use a test set for dialogue evaluation?

**Practical**
- Easy and fast to compute

**Easily reproducible**
- Suitable for benchmarking
Corpus-based evaluation
Method 1: Response matching

- Match system’s outputs to gold responses from the corpus
  - E.g. N-gram based BLEU (Papineni et al., 2002)
  - Poorly correlates with human judgement (Liu et al., 2016)
    - Dialogue is a one-to-many problem
  - Turn-based, ignores the dialogue as a whole

I’m hungry
What do you want to eat?
Score

What do you want to eat?
I’m hungry
Score

30.06.2023
Nurul Lubis © 2023
Method 2: Predict a score

- Train a model to output a score given a response/dialogue
- Considering dialogue context
- Focus on subjective quality
  - E.g FED (Mehri and Eskanazi, 2020), USR (Mehri and Eskanazi, 2020)
  - Did the user fulfill their goal?

What do you want to eat?
I’m hungry
Method 3: Construct pseudo-dialogue

- Replace golden system turns with generated ones
- Evaluate a dialogue as a whole
- Rules to check whether user goal is fulfilled
  - Objective measure
  - Corpus-specific
- Does pseudo-dialogue still make sense?

Score

I’m hungry

What do you want to eat?

Can you point me to an Italian restaurant?

Vapiano is located at Schifferstr. 169

30.06.2023
Nurul Lubis © 2023
Method 3: Construct pseudo-dialogue

- Context mismatch between user and system turns
  - Pseudo-dialogue is not the result of an interaction
- Overlooks specific types of mistakes
  - May overestimate dialogue policy performance

I’m hungry

Heinemann is a nice cafe in the city center

Can you point me to an Italian restaurant?

Vapiano is located at Schifferstr. 169
Corpus-based evaluation: current challenges

- Not yet strongly correlated with human judgements
- Focus on limited, subjective qualities
- Lack of generalization across datasets and models

• A recent National Science Foundation (NSF) report (Mehri et al., 2022)
Tackling the issues

How can we **solve** current challenges with an **efficient** and **reliable** method to evaluate dialogue systems?

We propose to use offline reinforcement learning (RL) critic as dialogue evaluator.
Hello, dialogue systems
Fundamental tasks in dialogue

initial turn tracking current turn planning final turn

PAST FUTURE

Credit: Prof. Milica Gašić
Modular ToD systems

- Modular approach: Dialogue systems are divided into modules (Williams, 2006; Thomson and Young, 2010)
- Policies can be trained with supervised or reinforcement learning

“I’m looking for an Italian restaurant”

- inform(food=italian)
- food=italian

“Which area do you have in mind?”

- request(area)
Reinforcement Learning (RL)
A gentle primer
RL Primer: Notations

- **Agent**
  - State $s_t$
  - Action $a_t$
- **Environment**
  - Reward $r_t$
- **Trajectory** $\tau = \{(s_1, a_1, s_2, r_1), \ldots, (s_T, a_T, s_{T+1}, r_T)\}$
  Sequence of state, action, reward tuples from a sequence of time steps, e.g. 1 to T. (we assume a finite horizon case)

- **Return** $R_t$
  Discounted cumulative reward: *How much reward have I collected from timestep $t$ until the end?*
  $$R_1 = \sum_{n=1}^{\infty} \gamma^{n-1}r_n$$

- **Policy** $\pi(a|s)$
  Probability distribution over actions in a given state
  *In a given state, which action should I take?*

- **Value functions**
  - $Q^\pi(s, a)$ expected return of being in state $s$, taking action $a$, and following policy $\pi$ afterwards
  - *How good is it to take a particular action in a given state?*

- Through interactions with the environment, the agent tries to find the best policy based on some measure of reward.
- Huge amount of interactions are needed
Learning methods

• **Policy-based**: We learn the policy $\pi_{\theta}(a|s)$...
  • Using parameters $\theta$ to map state to action

• **Value-based**: We learn the value function $Q^\pi(s,a)$...
  • Bellman Equation: the value of any state can be calculated with one-step look ahead, as opposed to having to inspect every future state
    \[ TQ(s_t, a_t) = E_{s_{t+1}}[r_t + \gamma Q(s_{t+1}, a_{t+1})]. \]
  • Act greedily: choose action with the highest value estimate

• **Actor-critic**: We learn both!
  • Learn a policy that maximizes value estimate
Optimizing policies with online RL

Update policy

Policy $\pi_{k+1}$

Rollout data $\{(s_t, a_t, s_{t+1}, r_t)\}$

Rollout

Policy $\pi_k$

$a_t$

$r_t$

$s_t$
Learning from online interaction can be expensive and time consuming
  • Even more than evaluation!

Some environments are high-risk
  • Dialogue systems for emotional distress?

Some behavior we want to learn are highly complex
  • How do we model the environment?

Can we leverage datasets to learn a policy?
Offline RL

Can be better than behavioral policy!

\[ (s_t, a_t, s_{t+1}, r_t) \]

Can use human demonstrations!

Deployment

Rollout

Learn policy

Corpus \( D \)

Rollout data \( \{(s_t, a_t, s_{t+1}, r_t)\} \)
Dialogue Evaluation with Offline Reinforcement Learning

Learning with critic

- Actor and critic are optimized alternatingly throughout training
Actor training

- Start with supervised learning (SL) pre-training to initialize the actor
- Continue training with offline RL
  - For each state, actor predicts the action
  - Critic estimate the value function
  - Actor tries to maximize critic’s estimate

Diagram:

- Corpus $\mathcal{D}$
- $(s_t, a_t, r_t, s_{t+1}) \sim \mathcal{D}$
- $s_t \rightarrow \text{Actor } \pi$
- $\pi(s_t) \rightarrow \text{Critic } Q$
- $Q(s_t, \pi(s_t))$
Critic training

- Critic produces value estimates:
  - $a_t$ comes from data
  - $a_{t+1}$ comes from actor
- Estimate is refined by minimizing the error of Bellman equation

$$\mathcal{L}_{\text{critic}} = (Q(s_t, a_t) - (r_t + \gamma Q'(s_{t+1}, \pi'(s_{t+1})))^2$$
Evaluation with critic

• For any policy, we can train a critic independently after-the-fact
• Use policy to be evaluated to estimate $Q(s_{t+1}, a_{t+1})$
  • Used to compute critic loss
• Use the final critic to estimate Q-values over a test set
  • Average Q-value on the initial states

Query policy to be evaluated $\pi_e$

Critic training

Test set $s_1 \sim D_{test}$
$\pi_e$ $Q(s_1, \pi_e(s_1))$
## Advantages

<table>
<thead>
<tr>
<th>Theoretically grounded</th>
<th>solves context mismatch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontology, data, and model independent</td>
<td>model-based, no hand-coded rules</td>
</tr>
<tr>
<td>State, action, and reward can take any form</td>
<td>adjust the architecture of critic and actor</td>
</tr>
</tbody>
</table>
Experiments
Task-oriented dialogue benchmark

- MultiWoZ corpus (Budzianowski et al., 2018)
  - Information seeking and reservation making
- Multi-domain dialogues
  - Restaurants, hotels, attractions, taxi, train, hospital, police
  - Multiple domain can occur in one dialogue
Metrics to be compared

**Static corpus evaluation**
- BLEU
- Pseudo-dialogue success rate
- Proposed critic

**Interactive evaluation**
- User simulator
- Human Evaluation

Pseudo-dialogue success rate

Proposed critic

Human Evaluation

User simulator

Interactive evaluation
Policy Optimization

Experiments
LAVA + PLAS

• Goal: To use critic’s signal to optimize a dialogue policy via offline RL

• Hypothesis: optimizing critic’s signal will also improve policy performance as measured by existing metrics

• PLAS (Zhou et al., 2020) offline RL algorithm on latent space

• Train a VAE to reconstruct actions found in corpus
Can the critic optimize the policy as measured by established metrics?

- Task-related metrics are **consistently improved** via offline RL on critic’s signal
- Slight decrease on BLEU
  - Trade off between BLEU and success has been observed before (Zhou et al., 2020; Lubis et al., 2021)
Policy Evaluation

Experiments
Dialogue evaluation with offline RL

**Goal**: Investigate critic’s value estimate as an evaluation metric compared to existing ones

**Hypothesis**: Critic can serve as a corpus-based evaluation metric that is better correlated with human judgements
Policies to be evaluated

**SL**
- **AuGPT** (Kulhanek et al., 2021) end-to-end transformer-based dialogue system, large amounts of data and labels
- **HDSA** (Chen et al., 2019) Policy operates on semantic-level action with a dedicated NLG module

**SL + RL**
- **LAVA** (Lubis et al., 2020) Policy with latent action, optimized on corpus-based success rate using RL
- **LAVA + PLAS** (proposed) Policy with latent action, optimized on critic’s signal using offline RL
Corpus-based evaluation

- HDSA has highest BLEU score
  - Trained emphasis on generation
Corpus-based evaluation

- HDSA has highest BLEU score
  - Trained emphasis on generation
- LAVA has highest corpus success rate
  - Optimized with RL on this metric
Corpus-based evaluation

- HDSA has highest BLEU score
  - Trained emphasis on generation
- LAVA has highest corpus success rate
  - Optimized with RL on this metric
- AuGPT has highest Q-value, followed by LAVA + PLAS
  - LAVA + PLAS is optimized on this metric

"Best model" on each corpus-based metric differs!
Interactive Evaluation

- Different trend compared to corpus evaluations
- AuGPT does very well on interactive evaluations
  - Large pre-trained model with data augmentation

![Bar chart showing success rates for AuGPT, HDSA, LAVA, and LAVA + PLAS in User Simulator based and Human trial settings.](chart.png)
Does the critic correlate with human judgement?

- **Corpus-based metrics**
  - Standard corpus-based metrics are **negatively** correlated with human evaluation
  - Our experiment confirm that BLEU has **poor** correlation
  - Our critic has **strong** correlation with human judgement

- **Interactive metrics**
  - User simulator is a good proxy to estimate system performance in human trial

- Critic training has the advantage of being **corpus- and model-independent**
Corpus- and model-independent evaluation

• Can we infer dialogue success from other signals?
• How does a successful dialogue look like?
• How do users behave in a successful dialogue?
• How do users react to a failed dialogue?
• ...
Emotional signal for task-oriented dialogue evaluation
Emotion in task-oriented dialogues


• Emotions are part of a natural human-like dialogue
• However, emotions are mainly studied in **chit-chat** dialogues
• User also expresses emotion as it relates to their goal

Is there something wrong with you? I need a …

Help! I was just robbed! …

I am excited to see some local attractions. …

…. You are doing a wonderful job!
Emotion in task-oriented dialogues


- We identified 7 classes to model user emotion

<table>
<thead>
<tr>
<th>Valence</th>
<th>Elicitor</th>
<th>Conduct</th>
<th>Emotion Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>-</td>
<td>-</td>
<td>Neutral</td>
</tr>
<tr>
<td>Negative</td>
<td>Event/Fact</td>
<td>Neutral/Polite</td>
<td>Fearful, sad, disappointed</td>
</tr>
<tr>
<td>Negative</td>
<td>System</td>
<td>Neutral/Polite</td>
<td>Dissatisfied, disliking</td>
</tr>
<tr>
<td>Negative</td>
<td>User</td>
<td>Neutral/Polite</td>
<td>Apologetic</td>
</tr>
<tr>
<td>Negative</td>
<td>System</td>
<td>Impolite</td>
<td>Abusive</td>
</tr>
<tr>
<td>Positive</td>
<td>Event/Fact</td>
<td>Neutral/Polite</td>
<td>Excited, happy, anticipating</td>
</tr>
<tr>
<td>Positive</td>
<td>System</td>
<td>Neutral/Polite</td>
<td>Satisfied, liking, appreciative</td>
</tr>
</tbody>
</table>

Emotion recognizer for dialogue evaluation
*(Feng et al., under review)*

- **Goal:** Leverage emotion recognition model to infer dialogue system performance

- **Hypothesis:** A model that can recognize emotions can also infer task success via user satisfaction

- **Zero-shot prediction of user satisfaction**

![Diagram showing emotion recognition model with emotion corpus for training and user satisfaction corpus for testing.](image)
What’s the difference?

Could you book me a table for 2 people at the restaurant, tomorrow at 8pm?

Certainly, I have booked it for you. Anything else I could help you with?

Wonderful! That would be all, thanks.
### Zero-shot satisfaction prediction

*(Feng et al., under review)*

<table>
<thead>
<tr>
<th>Model</th>
<th>JDDC</th>
<th>SGD</th>
<th>ReDial</th>
<th>CPPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiGRU (Sun et al., 2021)</td>
<td>17.1</td>
<td>8.6</td>
<td>8.3</td>
<td>27.4</td>
</tr>
<tr>
<td>BERT (Sun et al., 2021)</td>
<td>18.5</td>
<td>4.8</td>
<td>12.5</td>
<td>24.5</td>
</tr>
<tr>
<td>SatAct (Kim and Lipani, 2022)</td>
<td></td>
<td></td>
<td>71.3</td>
<td>16.5</td>
</tr>
<tr>
<td>SatActUtt (Kim and Lipani, 2022)</td>
<td>84.7</td>
<td></td>
<td></td>
<td>73.4</td>
</tr>
<tr>
<td>Zero-shot ERToD</td>
<td><strong>50.1</strong></td>
<td>78.8</td>
<td><strong>78.1</strong></td>
<td><strong>77.6</strong></td>
</tr>
</tbody>
</table>

ERToD achieves state-of-the-art performance on satisfaction prediction.
What happens if we use emotion as reward signal? (Geishauser et al., work in progress)

Emotions provide useful signal to help improve task success
Positive emotion correlates with task success

*(Geishauser et al., work in progress)*

As task success improves, “satisfied” emotion becomes more probable and “dissatisfied” less.
Conclusion

• We propose utilizing **offline RL for dialogue evaluation**
  • Towards **solving current challenges** of dialogue evaluation
  • **Efficient and reliable** metric
  • **Strong correlation** with human judgments

• **Emotional signal** can serve as a proxy to user satisfaction
  • Evaluate user satisfaction with zero-shot satisfaction prediction
  • Emotion as reward signal improves dialogue success
  • A universal, ontology-independent signal

• **Wide-range** of applications
  • Chat-oriented system
  • Human annotations as reward signal
Thank you