Targeted Data Acquisition for Evolving Negotiation Agents

Minae Kwon, Siddharth Karamcheti, Mariano-Florentino Cuéllar, Dorsa Sadigh
Negotiation is a bargaining process by which a joint decision is made by two parties.
Negotiation is a bargaining process by which a joint decision is made by two parties.
Negotiation is a bargaining process by which a joint decision is made by two parties.
Negotiation is a bargaining process by which a joint decision is made by two parties.

Lawyers in court

Employee negotiating salary

2021 UN climate change conference
Desiderata
Desiderata

(1) Agents that maximize their self-interest
Desiderata

(1) Agents that maximize their self-interest
(2) Agents that can compromise (find Pareto-optimal solutions)
Supervised Learning (SL)

\[ L(\theta) = - \sum_{x,c} \sum_{t} \log p_\theta(x_t | x_{0:t-1}, c) \]

\[ -\alpha \sum_{x,c} \sum_{j} \log p_\theta(o_j | x_{0:t-1}, c) \]
Supervised Learning (SL)

\[ L(\theta) = - \sum_{x,c} \sum_{t} \log p_{\theta}(x_t | x_{0:t-1}, c) \]

\[ -\alpha \sum_{x,c} \sum_{j} \log p_{\theta}(o_j | x_{0:t-1}, c) \]
Supervised Learning (SL)

\[ L(\theta) = - \sum_{x,c} \sum_t \log p_{\theta}(x_t | x_{0:t-1}, c) \]

\[ -\alpha \sum_{x,c} \sum_j \log p_{\theta}(o_j | x_{0:t-1}, c) \]

 utterances context

utterance prediction loss
Supervised Learning (SL)

\[ L(\theta) = - \sum_{x,c} \sum_{t} \log p_{\theta}(x_t | x_{0:t-1}, c) \]

\[ \text{utterance prediction loss} \]

\[ -\alpha \sum_{x,c} \sum_{j} \log p_{\theta}(d_j | x_{0:t-1}, c) \]

\[ \text{final split prediction loss} \]
Supervised Learning (SL)

\[
L(\theta) = - \sum_{x,c} \sum_{t} \log p_\theta(x_t | x_{0:t-1}, c) \]
\[
- \alpha \sum_{x,c} \sum_{j} \log p_\theta(d_j | x_{0:t-1}, c) \]

**utterance prediction loss**

**final split prediction loss**

*Relationship to dataset:* bias inherited from dataset

Lewis et al., Deal or No Deal? End-to-End Learning of Negotiation Dialogues, 2017
Reinforcement Learning (RL)
Reinforcement Learning (RL)
Reinforcement Learning (RL)

Negotiation

Bob (fixed)

Alice (learning)
Reinforcement Learning (RL)

Negotiation

propose(0 buns, 2 puffs, 1 roll)

Bob (fixed)

Alice (learning)

end

Agreement

\[ r_A = 6 \quad r_B = 8 \]
Reinforcement Learning (RL)

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

Alice’s utterances

running mean

propose(0 buns, 2 puffs, 1 roll)

end

Agreement

$r_A = 6 \quad r_B = 8$

Negotiation

Bob (fixed)

Alice (learning)
Reinforcement Learning (RL)

\[ R_A(x_t) = \gamma^{T-t}(r_A - \mu_n) \]

For \( x_t \in X^A \)

**Negotiation**

Bob (fixed)

propose(0 buns, 2 puffs, 1 roll)

Alice’s utterances

running mean
Reinforcement Learning (RL)

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

Alice’s utterances

Running mean

Negotiation

propose(0 buns, 2 puffs, 1 roll)

insist(1 bun, 2 puffs, 2 rolls)

Bob (fixed)

Alice (learning)
**Reinforcement Learning**

### Negotiation

**Bob (fixed)**

**Alice**

- propose(0 books)
- insist(1 bun, 2 pucks)

---

**RL**

<table>
<thead>
<tr>
<th>Alice</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>insist: item0=0 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
<tr>
<td>propose: item0=1 item1=3 item2=1</td>
<td>propose: item0=1 item1=2 item2=0</td>
</tr>
</tbody>
</table>

---

**Alice’s utterances**

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

**running mean**

---

Disagreement?!

Alice: 0 (potential 10)
Bob: 0 (potential 7)
For \( x_t \in X_A \)

\[
R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)
\]
Reinforcement Learning (RL)

\[ R_A(x_t) = \gamma^{T-t}(r_A - \mu_n) \]

For \( x_t \in X^A \)

Alice’s utterances

running mean

Negotiation

propose(0 buns, 2 puffs, 1 roll)

insist(1 bun, 2 puffs, 2 rolls)

..???

Alice (learning)

Bob (fixed)

Bob (fixed)
Reinforcement Learning (RL)

\[ R_A(x_t) = \gamma^{T-t}(r_A - \mu_n) \]

For \( x_t \in X^A \)

**Relationship to dataset:** Alice inherits dataset biases through Bob
Reinforcement Learning (RL)

Negotiation

Bob (fixed)

propose(0 buns, 2 puffs, 1 roll)

insist(1 bun, 2 puffs, 2 rolls)

Bob (fixed)

..???

Alice's utterances

For $x_t \in X^A$

$$R_A(x_t) = \gamma^{T-t}(r_A - \mu_n)$$

running mean

Relationship to dataset: Alice inherits dataset biases through Bob
Mixed RL, SL (\textbf{RL+SL})

Interleave SL training every nth timestep

\begin{itemize}
  \item n=1: RL, SL, RL, SL …
  \item n=2: RL, RL, SL, RL, RL, SL …
\end{itemize}
Mixed RL, SL (\textbf{RL+SL})

Interleave SL training every nth timestep

- $n=1$: RL, SL, RL, SL ...
- $n=2$: RL, RL, SL, RL, RL, SL ...

\textit{Relationship to dataset:} same as SL, bias inherited from dataset
Problem: Low-quality, static datasets!
Problem: Low-quality, static datasets!

Key Insight: Continually improve Bob with expert data!
Targeted Data Acquisition Framework

**Negotiation n**

- **propose(0 buns, 2 puffs, 1 roll)**

- **insist(1 bun, 2 puffs, 2 roll)**
Targeted Data Acquisition Framework

This looks novel!

Negotiation $n$

propose(0 buns, 2 puffs, 1 roll)

insist(1 bun, 2 puffs, 2 roll)

Novelty score:

$$s_n = \min_{x_t \in X^A} \log p_\theta(x_t | x_{0:t-1}, c^A)$$
Targeted Data Acquisition Framework

This looks novel!

Negotiation \( n \)

- **propose**: (0 buns, 2 puffs, 1 roll)
- **insist**: (1 bun, 2 puffs, 2 roll)

**Score \( s_n \)**

\[
s_n = \min_{x_t \in X^A} \log p_\theta(x_t | x_{0:t-1}, c^A)
\]

**Novelty score:**
Targeted Data Acquisition Framework

Alice RL Training

Negotiation n

propose(0 buns, 2 puffs, 1 roll)

insist(1 bun, 2 puffs, 2 rolls)

Score $s_n$
Targeted Data Acquisition Framework

Alice RL Training

Negotiation $n$

Bob

propose(0 buns, 2 puffs, 1 roll)

Alice

insist(1 bun, 2 puffs, 2 rolls)

Score $s_n$

Pick $k=500$
most novel negotiations

35
Targeted Data Acquisition Framework

Alice RL Training

Negotiation $n$

- propose(0 buns, 2 puffs, 1 roll)
- insist(1 bun, 2 puffs, 2 rolls)

Pick $k=500$ most novel negotiations

Score $s_n$
Targeted Data Acquisition Framework

Alice RL Training

Respond to expert annotations:
- Alice: propose(0 buns, 2 puffs, 1 roll)
- Bob: insist(1 bun, 2 puffs, 2 rolls)

Score $s_n$

Expert Annotations

Pick $k=500$ most novel negotiations

Expert: propose(0 buns, 2 puffs, 1 roll)
Bob:(insist(1 bun, 2 puffs, 2 rolls))
Alice: end
Targeted Data Acquisition Framework

**Alice RL Training**

**Negotiation n**

- propose\((0 \text{ buns}, 2 \text{ puffs}, 1 \text{ roll})\)
- insist\((1 \text{ bun}, 2 \text{ puffs}, 2 \text{ rolls})\)

**Score** $s_n$

**Bob**

**Alice**

**Pick k=500 most novel negotiations**

**Expert Annotations**

- propose\((0 \text{ buns}, 2 \text{ puffs}, 1 \text{ roll})\)
- insist\((1 \text{ bun}, 2 \text{ puffs}, 2 \text{ rolls})\)

**Expert**

**end**

**Update dataset**

$\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}'$
Targeted Data Acquisition Framework

**Alice RL Training**

Negotiation $n$

- **Bob**
  - propose($0$ buns, $2$ puffs, $1$ roll)

- **Alice**
  - insist($1$ bun, $2$ puffs, $2$ rolls)

**Bob SL Training**

**Expert Annotations**

- **Bob**
  - propose($0$ buns, $2$ puffs, $1$ roll)

- **Alice**
  - insist($1$ bun, $2$ puffs, $2$ rolls)
  - end

**Update dataset**

$\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}'$

Pick $k=500$ most novel negotiations
Targeted Data Acquisition Framework

Alice RL Training

Negotiation n

propose(0 buns, 2 puffs, 1 roll)

insist(1 bun, 2 puffs, 2 rolls)

Score $s_n$

Pick $k=500$ most novel negotiations

Continue training Alice

Bob SL Training

Expert Annotations

propose(0 buns, 2 puffs, 1 roll)

insist(1 bun, 2 puffs, 2 rolls)

end

Update dataset $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}'$

proposing $0$ buns, $2$ puffs, $1$ roll

insisting $1$ bun, $2$ puffs, $2$ rolls
Evaluation

Can we balance self-interest and Pareto-optimality?
Results with a Simulated Partner

(D1) Self-interest

(higher is better)
Results with a Simulated Partner

(D1) Self-interest

(higher is better)
Results with a Simulated Partner

(higher is better)

(D1) Self-interest
Results with a Simulated Partner

(higher is better)

(D1) Self-interest ✔
Results with a Simulated Partner

(D1) Self-interest ✔

(D2) Pareto-Optimal
Results with a Simulated Partner

(D1) Self-interest ✔

(D2) Pareto-Optimal
Results with a Simulated Partner

(higher is better)

(D1) Self-interest ✔

(D2) Pareto-Optimal ✔
Results with a Simulated Partner

(higher is better)

Advantage

Pareto

Agreement

Novelty

(D1) Self-interest ✔

(D2) Pareto-Optimal ✔
Results with a Human Partner

(D1) Self-interest ✔

(D2) Pareto-Optimal ✔

N=101
Results with a Human Partner

(\textit{higher is better})

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Pareto</th>
<th>Agreement</th>
<th>Novelty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair to You</td>
<td>Fair to Both</td>
<td>Effective</td>
<td>Represent</td>
</tr>
</tbody>
</table>

\(N=101\)
Main Ideas

• Our approach balances self-interest and Pareto-optimality the best.

• This holds true against both simulated and human partners.