Joint System and User Modelling

Hsien-chin Lin and Christian Geishauser
Dialogue Systems and Machine Learning Group

18.06.2021
Outline

- Part I: Multi-agent reinforcement learning for dialogue
  - What is multi-agent RL?
  - How has it been applied to dialogue?
  - Emergent communication
- Part II: Data augmentation and evaluation by machine-to-machine conversation
  - How to collect more dialogue data?
  - Evaluate dialogue systems by machine-to-machine interaction
Part I: Multi-agent reinforcement learning
Hide-and-seek

Source: YouTube, Multi-Agent Hide and Seek
Why train a user agent with RL?

- Building a rule-based user simulator is labour-intensive and hard to maintain
- Supervised trained user agents only know how to act in situations seen in the data
Why train a user agent with RL?

- Building a rule-based user simulator is labour-intensive and hard to maintain
- Supervised trained user agents only know how to act in situations seen in the data
- Advantages of using RL for improving user agent:
  - Can explore situations that haven’t been observed in the data and learn how to act there
  - Can optimise discrete metrics, for instance:
    - Were all constraints and requests communicated?
    - Did it inform the constraints before requesting information?
  - Utilize RL algorithms
    - Optimise for diversity to obtain different user agents, optimise for maximum entropy for diverse behaviour, ...
Multi-agent Reinforcement Learning

- Train a user agent with RL
- Actual goal is to obtain a good dialogue system policy
- Train both user agent and system agent using RL
- We have two RL agents that need to be trained -&gt; multi-agent RL setting
Multi-agent Reinforcement Learning

- **Single agent RL**
  - single agent that interacts with the environment
  - optimal strategy: maximise expected return for a given environment

- **Multi agent RL:**
  - multiple agents that interact with the environment
  - optimal strategy: varies depending on the behaviour of other agents that might change over time
  - -> environment becomes non-stationary for each individual agent
  - constantly need to adapt to environment changes
Multi-agent Reinforcement Learning

- Stochastic game (Markov game)
  - $N$: number of agents
  - $S$: state space
  - $A = A_1 \times \ldots \times A_N$ joint action space
  - $R$: reward function, emitting reward for every agent $i$
  - $P$: transition probability distribution
  - $O = O_1 \times \ldots \times O_N$ joint observation space
  - $\gamma$: discount factor

Centralised learning

- if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent
Multi-agent Reinforcement Learning

- Centralised learning
  - if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent
  - action space grows exponentially with number of agents
  - not applicable when each agent receives different observations
  - Multi-action policies can be considered as these (one agent per action, domain, ...)


Multi-agent Reinforcement Learning

- **Centralised learning**
  - if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent
  - action space grows exponentially with number of agents
  - not applicable when each agent receives different observations
  - Multi-action policies can be considered as these (one agent per action, domain, ...)

- **Independent learners**
  - each agent independently learns its own policy, treating other agents as part of the environment
Multi-agent Reinforcement Learning

- **Centralised learning**
  - if all agents observe the full state, we can model cooperative multi-agent system as single meta-agent
  - action space grows exponentially with number of agents
  - not applicable when each agent receives different observations
  - Multi-action policies can be considered as these (one agent per action, domain, ...)

- **Independent learners**
  - each agent independently learns its own policy, treating other agents as part of the environment
  - Example: Independent Q-learning (IQL)
  - Does not suffer from exponential growth of action space
  - each agent only needs its local observation
  - Problem: non-stationarity of the environment (experience replay methods not straightforward)
Centralized training with decentralized execution (CTDE)

- assume you can observe the full state during training, but only local observations during testing
- Critic is only necessary during training
Centralized training with decentralized execution (CTDE)

- assume you can observe the full state during training, but only local observations during testing
- Critic is only necessary during training
- \(\rightarrow\) learn centralised critic operating on the full space
- \(\rightarrow\) learn actors independently on their respective smaller space (important in dialogue)
Centralized training with decentralized execution (CTDE)

- assume you can observe the full state during training, but only local observations during testing
- Critic is only necessary during training
- \(-\) learn centralised critic operating on the full space
- \(-\) learn actors independently on their respective smaller space (important in dialogue)
- helps to resolve the non-stationarity of the critic

\[ g = \nabla_{\theta} \log \pi(u|\tau_t)(r + \gamma V(s_{t+1}) - V(s_t)) \]
Dialogue system and user simulator are both agents that can be optimised using RL

- common goal: successfully complete the conversation, i.e. fulfilling the user goal
  - user must provide information about the problem to be solved
  - system must solve the problem using goal information that was provided by the user
Multi-agent RL for Dialogue

- Dialogue system and user simulator are both agents that can be optimised using RL
  - common goal: successfully complete the conversation, i.e. fulfilling the user goal
    - user must provide information about the problem to be solved
    - system must solve the problem using goal information that was provided by the user

- Games like Hide-and-Seek:
  - agents can come up with their own behaviour to solve the game
  - in dialogue: agents should not come up with an artificial language to complete the conversation
Multi-agent RL for Dialogue

- Liu, Lane, 2017 (IterDPL)
  - train agents iteratively to deal with non-stationarity

- Papangelis et. al., 2019
  - Model the dialogue as stochastic game and optimise via WoLF-PHC
    - slow learning rate when winning, high learning rate when loosing
  - train agents concurrently

- Both show improved performance over supervised trained agents
  - using DSTC2 data for supervised training
Multi-agent RL for Dialogue

- CTDE for dialogue
  - user simulator has own observation such as the goal
  - dialogue system has own observation such as database results
  - -> no meta-agent that operates on the shared space

- Critic that has information from user and system can better estimate task success
CTDE for dialogue
- user simulator has own observation such as the goal
- dialogue system has own observation such as database results
- -> no meta-agent that operates on the shared space

Critic that has information from user and system can better estimate task success

Takanobu et. al., 2020 (MADPL)
- use centralised critic to estimate the shared goal of task success
- use dedicated critics for the private goals of user and dialogue system
  - penalty when requested info is not provided immediately; task success rate based on user agent description
  - penalty when requesting too early; reward when all constraints and requests are communicated
- work in MultiWOZ setting
Multi-agent RL for Dialogue

- Takanobu et. al., 2020 (MADPL)
  \[ \nabla_{\phi} \log \pi_{\phi}(a^S | s^S) [A^S(s^S) + A^G(s)] \\
\nabla_{\omega} \log \mu_{\omega}(a^U | s^U) [A^U(s^U) + A^G(s)] \n\]

- CRL uses centralised critic

- IterDPL trains the two agents iteratively instead of concurrently

- All algorithms are able to improve success

<table>
<thead>
<tr>
<th>System</th>
<th>User</th>
<th>Turns</th>
<th>Inform</th>
<th>Match</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL</td>
<td>SL</td>
<td>6.34</td>
<td>73.08</td>
<td>82.58</td>
<td>49.7</td>
</tr>
<tr>
<td>SL</td>
<td>RL</td>
<td>8.75</td>
<td>76.86</td>
<td>76.28</td>
<td>60.2</td>
</tr>
<tr>
<td>RL</td>
<td>SL</td>
<td>6.20</td>
<td>72.84</td>
<td>79.15</td>
<td>51.1</td>
</tr>
<tr>
<td>RL</td>
<td>RL</td>
<td>7.92</td>
<td>75.96</td>
<td>70.37</td>
<td>58.7</td>
</tr>
<tr>
<td>CRL</td>
<td></td>
<td>8.13</td>
<td>68.29</td>
<td>89.71</td>
<td>66.6</td>
</tr>
<tr>
<td>IterDPL</td>
<td></td>
<td>8.79</td>
<td>74.01</td>
<td>81.04</td>
<td>64.6</td>
</tr>
<tr>
<td>MADPL</td>
<td></td>
<td><strong>8.96</strong></td>
<td><strong>76.26</strong></td>
<td><strong>90.98</strong></td>
<td><strong>70.1</strong></td>
</tr>
</tbody>
</table>
Multi-agent RL for Dialogue

- Evaluation with **rule-based user simulator** during multi-agent training
  - erratic behaviour
  - large action space of system
  - more possible optimal solutions?
Multi-agent RL for Dialogue

- Evaluation with **rule-based user simulator** during multi-agent training
  - erratic behaviour
  - large action space of system
  - more possible optimal solutions?
- Evaluation with **rule-based system** during multi-agent training
  - rule-based system is able to interact with user agent
  - easier to learn a good user agent
Multi-agent RL for Dialogue

- Evaluation with **rule-based user simulator** during multi-agent training
  - erratic behaviour
  - large action space of system
  - more possible optimal solutions?

- Evaluation with **rule-based system** during multi-agent training
  - rule-based system is able to interact with user agent
  - easier to learn a good user agent

- Need rule-based methods for this evaluation
  - exactly what should be omitted
Human evaluation

- humans compare dialogues generated by different algorithms and give preference
Multi-agent RL for Dialogue

- Human evaluation
  - humans compare dialogues generated by different algorithms and give preference

- Role of supervised pre-training
  - training only a system agent from scratch with a rule-based simulator is already challenging
  - training both concurrently makes it much more difficult
  - supervised pre-training makes RL training possible in the first place
  - supervised pre-training gives already a good bias towards the solution one wants to have
  - can we squeeze more out of the data? (Imitation learning, Inverse RL, ...)
Multi-agent RL for Dialogue

- Human evaluation
  - humans compare dialogues generated by different algorithms and give preference

- Role of supervised pre-training
  - training only a system agent from scratch with a rule-based simulator is already challenging
  - training both concurrently makes it much more difficult
  - supervised pre-training makes RL training possible in the first place
  - supervised pre-training gives already a good bias towards the solution one wants to have
  - can we squeeze more out of the data? (Imitation learning, Inverse RL, ...)

- Can we train entirely without data?
  - and use that bootstrapped policy to interact/learn with real humans
Emergent Communication in multi-agent RL

- Emergent communication: Language arises because agents must communicate in order to solve a task
  - language that is learned has solely a functional purpose
Emergent Communication in multi-agent RL

- Emergent communication: Language arises because agents must communicate in order to solve a task
  - language that is learned has solely a functional purpose

- Common task: Speaker and listener communicate
  - speaker needs to navigate listener to a certain spot in the map
  - speaker needs to describe a target image that the listener has to select
  - user agent and system communicate goal and information?
Emergent communication: Language arises because agents must communicate in order to solve a task
- language that is learned has solely a functional purpose

Common task: Speaker and listener communicate
- speaker needs to navigate listener to a certain spot in the map
- speaker needs to describe a target image that the listener has to select
- user agent and system communicate goal and information?

Typical questions
- Is the emerged language interpretable?
- is it of compositional nature?
Emergent Communication in multi-agent RL

- Kottur et. al. 2017
  - Natural language does not emerge “naturally” in multi-agent dialog
  - Large enough vocabulary results in symbols mapped to instances
  - Small vocabulary still non-compositional and hard to interpret
Emergent Communication in multi-agent RL

- Kottur et. al., 2017
  - Natural language does not emerge “naturally” in multi-agent dialog
  - Large enough vocabulary results in symbols mapped to instances
  - Small vocabulary still non-compositional and hard to interpret

- Lazaridou et. al., 2020
  - Language models infer structural properties of language from text corpora
    - Ignore the functional aspect of communication
  - Multi-agent communication focuses only on using language as a utility
    - No “natural” language
  - Functional learning: “what to say”
  - Structural learning: “how to say”
Part II:
Data augmentation and evaluation
Data augmentation

- Learning from human
  - Time consuming and costly
  - Cold start
Data augmentation

- Learning from human
  - Time consuming and costly
  - Cold start
- Data collection
  - Annotation (MultiWOZ 2.0, 2.1, 2.2, 2.3, 2.4)
Data augmentation

- Learning from human
  - Time consuming and costly
  - Cold start
- Data collection
  - Annotation
- Coverage and bias

MultiWoz 2.0 (Budzianowski et al., 2018)
User simulation

- Rule-based user simulator
- Pros
  - data free
  - interpretable
- Cons
  - Require expert knowledge
  - Domain dependent
  - Not human like
  - Hard to build for complicated domains

Agenda-based user simulator (Schatzmann et al. 2007)
User simulation

Data-driven user simulator

Pros
- Learn from data
- Domain independent (TUS)

Cons
- Bias on the corpus
- Exploration limited by the corpus
- Zero-shot transfer is still challenging

TUS (Lin et al., 2021)
**User simulation**

- Data-driven user simulator
- **Pros**
  - Domain independent (TUS)
  - Learn from data
- **Cons**
  - Bias on the corpus
  - Exploration limited by the corpus
  - Zero-shot transfer is still challenging

<table>
<thead>
<tr>
<th>Training</th>
<th>Removed data(%)</th>
<th>Attr</th>
<th>Hotel</th>
<th>Rest</th>
<th>Taxi</th>
<th>Train</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUS-noAttr</td>
<td>32.20</td>
<td>0.69</td>
<td>0.64</td>
<td>0.81</td>
<td>0.65</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>TUS-noTaxi</td>
<td>19.60</td>
<td>0.63</td>
<td>0.61</td>
<td>0.81</td>
<td>0.61</td>
<td>0.70</td>
<td>0.74</td>
</tr>
<tr>
<td>TUS-noRest</td>
<td>45.21</td>
<td>0.62</td>
<td><strong>0.66</strong></td>
<td>0.80</td>
<td>0.56</td>
<td>0.75</td>
<td>0.76</td>
</tr>
<tr>
<td>TUS-noTrain</td>
<td>36.95</td>
<td>0.64</td>
<td>0.65</td>
<td>0.78</td>
<td><strong>0.67</strong></td>
<td>0.62</td>
<td>0.73</td>
</tr>
<tr>
<td>TUS-noHotel</td>
<td>40.15</td>
<td>0.59</td>
<td>0.59</td>
<td>0.76</td>
<td><strong>0.61</strong></td>
<td>0.54</td>
<td>0.69</td>
</tr>
<tr>
<td>TUS</td>
<td>0</td>
<td>0.69</td>
<td>0.68</td>
<td>0.81</td>
<td>0.66</td>
<td>0.77</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Generating datasets through self-play

- To collect data for a new domain...
- Problems of Wizard-of-Oz setup
  - might not cover all the expected interactions
  - might contain dialogues unfit for use (too simple or too complex)
  - annotation errors
- Including automation and crowdsourcing to collect datasets
Generating datasets through self-play

machine-to-machine (Shah et al. 2018)
Generating datasets through self-play

- Crowd workers are asked to rewrite the machine-generated conversations.
- They are encouraged to use linguistic phenomena like coreference (“Reserve that restaurant”) and lexical entrainment.
- Second round of crowdsourcing for validation the annotation and utterances.
Generating datasets through self-play

- Crowd workers are asked to rewrite the machine-generated conversations.
- They are encouraged to use linguistic phenomena like coreference (“Reserve that restaurant”) and lexical entrainment.
- Second round of crowdsourcing for validation the annotation and utterances.

<table>
<thead>
<tr>
<th></th>
<th>Machine-to-Machine</th>
<th>Wizard-of-Oz</th>
</tr>
</thead>
<tbody>
<tr>
<td>conversation policy</td>
<td>based on models</td>
<td>real users</td>
</tr>
<tr>
<td>annotation</td>
<td>easy</td>
<td>expensive</td>
</tr>
<tr>
<td>coverage</td>
<td>controllable</td>
<td>interesting</td>
</tr>
</tbody>
</table>
The difference between human trial and corpust

MultiWOZ 2.0

u: I am looking to book a train that is leaving from Cambridge to Bishops Stortford on Friday.

s: There are a number of trains leaving throughout the day. What time would you like to travel?

u: I want to get there by 19:45 at the latest.

s: Okay! The latest train you can take leaves at 17:29, and arrives by 18:07. Would you like for me to book that for you?

u: Yes please. I also need the travel time, departure time, and price.

...
Evaluation

- Human evaluation
  - Time-consuming and costly (500 dialogue/day)
  - coverage
Evaluation

- **Human evaluation**
  - Time-consuming and costly
  - coverage

- **Self-play evaluation**
  - The user model is far from the real users
Evaluation

- Human evaluation
  - Time-consuming and costly
  - coverage
- Self-play evaluation
  - The user model is far from the real users
- How to speed up the evaluation?
Self-play for chit chat bots

- The chatbot can talk to itself
- Automatic matrix (Ghandeharioun et. al. 2019)
  - perplexity
  - embedding matrix
  - KL-Divergence between the posterior and the prior distribution.
- The dialogue can also be rated by crowd workers
  - engagingness
  - interestingness
  - knowledge
  - humanness

ACUTE-EVAL (Li et al. 2019)
Generating challenge datasets

- With complete control of synthetic data generated by dialogue self-play, we can generate unseen patterns in the test set (Majumdar et al. 2019)
Generating challenge datasets

- With complete control of synthetic data generated by dialogue self-play, we can generate unseen patterns in the test set (Majumdar et al. 2019)
- Out of template
Generating challenge datasets

- With complete control of synthetic data generated by dialogue self-play, we can generate unseen patterns in the test set (Majumdar et al. 2019)
- Out of template
- Out of Pattern
  - Turn compression
  - New api
  - Reordering the slot filling
  - Another Slot (irrelevant slots)
  - Audit more: When the system requests a new slot, the user provides another slot-value pair in addition
Generating challenge datasets

- **Experiment**
  - Single End-to-End Memory Network (SMN) (Bordes et al. (2017))
  - Multiple End-to-End Memory Network (MMN)

<table>
<thead>
<tr>
<th>Test Case</th>
<th>IT</th>
<th>OOT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SMN</td>
<td>MMN</td>
</tr>
<tr>
<td>Non OOP</td>
<td>88.62</td>
<td>90.17</td>
</tr>
<tr>
<td>Turn Comp.</td>
<td>27.80</td>
<td>55.00</td>
</tr>
<tr>
<td>New API</td>
<td>7.42</td>
<td>7.83</td>
</tr>
<tr>
<td>Reordering</td>
<td>54.50</td>
<td>45.50</td>
</tr>
<tr>
<td>Another Slot</td>
<td>38.00</td>
<td>25.00</td>
</tr>
<tr>
<td>Audit More</td>
<td>15.50</td>
<td>34.00</td>
</tr>
<tr>
<td>OOP Avg.</td>
<td>28.64</td>
<td>33.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SMN</th>
<th>MMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non OOP</td>
<td>87.39</td>
<td>88.27</td>
</tr>
<tr>
<td>Turn Comp.</td>
<td>27.90</td>
<td>54.70</td>
</tr>
<tr>
<td>New API</td>
<td>8.17</td>
<td>6.67</td>
</tr>
<tr>
<td>Reordering</td>
<td>54.00</td>
<td>41.50</td>
</tr>
<tr>
<td>Another Slot</td>
<td>41.50</td>
<td>27.50</td>
</tr>
<tr>
<td>Audit More</td>
<td>16.00</td>
<td>35.00</td>
</tr>
<tr>
<td>OOP Avg.</td>
<td>29.51</td>
<td>33.07</td>
</tr>
</tbody>
</table>
References

- Liu, Lane, 2017: Iterative Policy Learning in End-to-End trainable task-oriented neural dialog models
- Papangelis et. al., 2019: Collaborative Multi-Agent Dialogue Model Training via Reinforcement Learning
- Takanobu et. al., 2020: Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition
- Kottur et. al., 2017: Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog
- Lazaridou et. al., 2020: Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning
- Baker et. al., 2020: Emergent Tool Use from Multi-Agent Autocurricula
- Shah et. al. 2018: Building a conversational agent overnight with dialogue self-play
- Ghandeharioun et al. 2019: Approximating interactive human evaluation with self-play for open-domain dialog systems
- Li et al. 2019: Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons
- Majumdar et al. 2019: Generating Challenge Datasets for Task-Oriented Conversational Agents through Self-Play