Robustness, Scalability & Practicality of Conversational AI

September 7th, 2022
Can you help me book a 5-star hotel on Sunday?

For how many people?
Robustness
Task-Oriented Dialogue Systems (Young, 2000)

LU: Language Understanding
DST: Dialogue State Tracking
NLG: Natural Language Generation
DP: Dialogue Policy Learning

Can you help me book a 5-star hotel on Sunday?

For how many people?
Task-Oriented Dialogue Systems (Young, 2000)

Can you help me book a 5-star hotel on Sunday?

For how many people?
Recent Advances in NLP

- Pre-trained models
  - ELMo, BERT, RoBERTa, XLM, GPT, etc.
Lift all lights to Morocco
List all flights tomorrow
Mismatch between Written and Spoken Languages

Goal: ASR-Robust Embeddings

- learning *spoken* embeddings
- better performance on *spoken* language understanding tasks

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Written language</td>
<td>• Spoken language</td>
</tr>
<tr>
<td></td>
<td>• Include recognition errors</td>
</tr>
</tbody>
</table>
Solution: LatticeLM
(Huang & Chen, ACL 2020)

https://github.com/MiuLab/LatticeLM
ASR Lattices for Preserving Uncertainty

- Idea: lattices may include correct words

(Ladhak, et al., 2016)
LatticeLM: Efficient Two-Stage Pre-Training

Stage 1: Pre-Training on Sequential Texts

- What
- a
- day
- <EOS>

Stage 2: Pre-Training on Lattices

- LatticeLSTM
- Linear
- LatticeLSTM

Fine-Tuning

- Max pooling
- Classification
- LatticeLSTM

Spoken Language Understanding Results

- **Intent Prediction**
  - Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)

![Graph showing performance metrics for SNIPS and ATIS datasets. The graph compares different models and their performance.](image-url)
Spoken Language Understanding Results

- **Intent Prediction**
  - Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)

[Diagram showing comparison between SNIPS and ATIS with various models like 1-Best, LatticeLSTM, and LatticeLM]
Spoken Language Understanding Results

- **Intent Prediction**
  - Word Error Rate: 45.6% (SNIPS); 15.6% (ATIS)

### Comparison of Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>SNIPS</th>
<th>ATIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LatticeLSTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LatticeLM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Best</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Best +</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **SNIPS**
- **ATIS**

The chart compares the performance of different methods on the SNIPS and ATIS datasets, showing the word error rates for each.
Spoken Language Understanding Results

- **Dialogue Act Prediction**
  - Word Error Rate: 32.0% (MRDA); 28.4% (SWDA)

![MRDA and SWDA Bar Charts]

- **MRDA**
  - 1-Best + LatticeLSTM
  - LatticeLM

- **SWDA**
  - 1-Best + LatticeLSTM
  - LatticeLM
What if we only have texts from ASR?
Solution:
Contrastive Learning for ASR-Robust Embeddings
(Chang & Chen, INTERSPEECH 2022)

https://github.com/MiuLab/SpokenCSE

Improving ASR Robustness of Embeddings

- Idea: adapt embeddings robust to errors with only textual information

Speech Corpus

<table>
<thead>
<tr>
<th>ASR output</th>
</tr>
</thead>
<tbody>
<tr>
<td>wrack a nice beach</td>
</tr>
<tr>
<td>clean text</td>
</tr>
<tr>
<td>recognize speech</td>
</tr>
</tbody>
</table>

Pre-Training

SLU Corpus

<table>
<thead>
<tr>
<th>ASR output</th>
</tr>
</thead>
<tbody>
<tr>
<td>new email for contact</td>
</tr>
</tbody>
</table>

SLU label

Intent=SendEmail

Fine-Tuning

Improving ASR Robustness of Embeddings

- Idea: adapt embeddings robust to errors with only textual information

---

Contrastive Pre-Training

- Idea: ASR outputs have similar embeddings as their clean texts

- Pre-training objective: \( \mathcal{L}_{pt} = \mathcal{L}_c + \lambda_{mlm} \cdot \mathcal{L}_{mlm} \)

RoBERTa

- Positive
- Negative

\[
\mathcal{L}_c = - \sum_h \log \frac{e^{sim(h, h^+)/\tau_c}}{\sum_{h' \neq h} e^{sim(h, h')/\tau_c}}
\]

Spoken Language Understanding Results

- **SLU data**
  - Synthesized TREC6 (WER=29%) & ATIS (WER=32%)
  - SLURP: Spoken Language Understanding Resources Package (WER=25%)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Class</th>
<th>Avg. Length</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC6</td>
<td>6</td>
<td>8.89</td>
<td>5,452</td>
<td>500</td>
</tr>
<tr>
<td>ATIS</td>
<td>22</td>
<td>11.14</td>
<td>4,978</td>
<td>893</td>
</tr>
<tr>
<td>SLURP</td>
<td>18 * 46</td>
<td>8.89</td>
<td>50,628</td>
<td>10,992</td>
</tr>
</tbody>
</table>
Spoken Language Understanding Results

RoBERTa
Phoneme-BERT*
SimCSE

TREC6
ATIS
SLURP
Spoken Language Understanding Results

- **RoBERTa**
- **Phoneme-BERT**
- **SimCSE**
- **Proposed Pre-training**

Contrastive pre-training improves robustness of spoken embeddings with only texts.
Improving ASR Robustness of Embeddings

Idea: adapting embeddings robust to misrecognitions

- **Pre-Training**
  - Speech Corpus
  - WRACK A NICE BEACH
  - CLEAN TEXT
  - RECOGNIZE SPEECH

- **Fine-Tuning**
  - SLU Corpus
  - NEW EMAIL FOR CONTACT
  - SLU LABEL
    - Intent=SendEmail

Idea: supervised contrastive learning with self-distillation

Supervised Contrastive Learning

- Idea: data with the same label should be close to each other
- Objective:

\[
L_{\text{hard}} = - \sum_i \sum_{j \neq i} 1_{y_i = y_j} \log \frac{e^{\text{sim}(h_i, h_j) / \tau_{sc}}}{\sum_{k \neq i} e^{\text{sim}(h_i, h_k) / \tau_{sc}}}
\]

Self-Distillation

- **Issue:** misrecognitions may lead to wrong or vague intents
- **Objective:** $\mathcal{L}_d = \sum_i KL(p_i^{t-1} \| p_i^t)$

Supervised Contrastive with Self-Distillation

- **Issue**: noisy labels also affect $\mathcal{L}_{hard}$
  \[
  \mathcal{L}_{soft} = - \sum_i \sum_{j \neq i} (p_i^{t-1} \cdot p_j^{t-1}) \log \frac{e^{sim(h_i,h_j)/\tau_{sc}}}{\sum_{k \neq i} e^{sim(h_i,h_k)/\tau_{sc}}}
  \]

- **Fine-tuning objective**: $\mathcal{L}_{ft} = \mathcal{L}_{ce} + \lambda_d \mathcal{L}_d + \lambda_{hard} \mathcal{L}_{hard} + \lambda_{soft} \mathcal{L}_{soft}$

![Diagram showing the contrastive learning and self-distillation process with prediction distributions and similarity values.](diagram.png)
Spoken Language Understanding Results

- **TREC6**
  - RoBERTa
  - Phoneme-BERT*
  - SimCSE
  - Proposed Pre-training

- **ATIS**

- **SLURP**
Spoken Language Understanding Results

- **TREC6**
  - RoBERTa
  - Phoneme-BERT*
  - SimCSE
  - Proposed Pre-training
  - Proposed Pre-training + Fine-tuning

- **ATIS**
  - RoBERTa
  - Phoneme-BERT*
  - SimCSE
  - Proposed Pre-training
  - Proposed Pre-training + Fine-tuning

- **SLURP**
  - RoBERTa
  - Phoneme-BERT*
  - SimCSE
  - Proposed Pre-training
  - Proposed Pre-training + Fine-tuning

Proposed methods improve robustness to ASR errors and label noises
**Robustness**

- ✔ **LatticeLM** for preserving uncertainty
- ✔ **Contrastive learning** with only textual information

- **Contrastive Pre-training** learns error-invariant sentence embeddings

- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

---

**Scalability**

---

**Practicality**

---
Scalability
Can you help me book a 5-star hotel on Sunday?

For how many people?

Issue: limited labelled data hinders scalability
Solution: Data Augmentation

(Lai et al., SIGDIAL 2022)

https://github.com/MiuLab/CUDA

Poster: Today 1pm (National Robotarium)
**Diverse User Dialogue Acts in States**

**Value Substitution**
- No, it needs to serve **British** food and I’d like a reservation for **18:00**.
  - Confirm=False, Inform{ rest-area=center, rest-food=British, rest-time=18:00 }

**CoCo**
- No, it needs to serve **Chinese** food and I’d like a reservation for **18:00**.
  - Confirm=False, Inform{ rest-area=center, rest-food=Chinese, rest-time=18:00 }

**Proposed**
- No, it should serve **Chinese** food and I need to book a table for **2** people.
  - Confirm=False, Inform{ rest-area=center, rest-food=Chinese, rest-people=2 }

**Slot Combination**
- Thank you, can you also find me a hotel with parking near the restaurant?
  - Confirm=True, Inform{ rest-area=center, rest-name=pho bistro, hotel-area=center, hotel-parking=yes }

I recommend Pho Bistro in the center. What time do you plan to book the table?

Recommend \{ rest-name=pho bistro, rest-area=center \}; Request \{ rest-time \}

- **System Turn**
  - **System Recommend**
    - **Informed Slot**
      - **User Reply?**
        - **P_reply**
          - Add to **Inform**
            - **Change Domain?**
              - **P_domain**
                - Select a new domain
                  - Randomly add 1-3 slots from a target domain to **Inform**
      - **Inform** \{ rest-time \}
        - **Inform** \{ hotel-area, hotel-parking \}
    - **User Informed Slots**
      - **Apply Coreference?**
        - **P_coref**
          - corresponding coreference list
            - **Confirm**=True
              - **Inform** \{ rest-name=pho bistro, rest-area=center \}
            - **Confirm**=False
          - pre-defined values
    - **User Confirm?**
      - **P_confirm**
        - **1 − P_confirm**
          - **Confirm**=True
          - **Confirm**=False
    - **Confirm**=True
      - **Inform** \{ rest-name=pho bistro, rest-area=center \}
    - **Inform** \{ rest-time \}
      - **Inform** \{ hotel-area=center, hotel-parking=yes \}
      - **Confirm**=True, 
        - **Inform** \{ rest-name=pho bistro, rest-area=center, rest-time=7pm, hotel-area=center, hotel-parking=yes \}
I recommend Pho Bistro in the center. What time do you plan to book the table?

Confirm=True,
Inform{ rest-name=pho bistro, rest-area=center, rest-time=7pm, hotel-area=center, hotel-parking=yes }

1. Thank you, 7pm works for me. Can you also find me a hotel with parking near the restaurant?
   Slot Appearance: ✔️ Value Consistency: (span) ✔️ (boolean) ✔️
2. Thank you, 7pm works for me. Can you also find me a hotel without parking near the restaurant?
   Slot Appearance: ✔️ Value Consistency: (span) ✔️ (boolean) ✗
3. Thank you, 7pm works for me. Can you also find me a hotel with parking in the center of the town?
   Slot Appearance: ✔️ Value Consistency: (span) ✗ (boolean) ✔️
4. Thank you. Can you also find me a hotel with free wifi near the restaurant?
   Slot Appearance: ✗ Value Consistency: (span) -- (boolean) --

→ 78% success rate for generation (57% for CoCo)
I recommend Pho Bistro, a popular restaurant in the center.

Recommendation{
  restaurant-name=pho bistro,
  restaurant-area=center }

System Turn $U^\text{sys}_t$

System Act

User Dialogue Act Generation

$\hat{A}_t$

Turn-Level Dialogue Act & State

Confirms=True
Inform{
  rest-name=pho bistro,  
  rest-area=center, 
  hotel-area=near the restaurant, 
  hotel-parking=yes }
## DST Performance

### MultiWOZ 2.1

<table>
<thead>
<tr>
<th></th>
<th>TripPy</th>
<th>TRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>57.72</td>
<td>44.08</td>
</tr>
<tr>
<td>Value Substitution</td>
<td>59.48</td>
<td>43.76</td>
</tr>
<tr>
<td>CoCo (Li et al., 2021)</td>
<td>60.46</td>
<td>43.53</td>
</tr>
<tr>
<td>CUDA</td>
<td>62.93†</td>
<td>44.86†</td>
</tr>
</tbody>
</table>

CUDA improves many trackers and achieves SOTA in MultiWOZ 2.1

### CoCo+ (rare)

<table>
<thead>
<tr>
<th></th>
<th>TripPy</th>
<th>TRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>28.38</td>
<td>16.65</td>
</tr>
<tr>
<td>Value Substitution</td>
<td>39.42</td>
<td>16.42</td>
</tr>
<tr>
<td>CUDA</td>
<td>48.83†</td>
<td>17.79†</td>
</tr>
</tbody>
</table>

CUDA shows better robustness for rare state combinations
Robustness

- LatticeLM for preserving uncertainty
- Contrastive learning with only textual information

- Contrastive Pre-training learns error-invariant sentence embeddings
- Supervised CL with Self Distillation improves robustness to noises from ASR and labels

Thank you, can you also find me a hotel with parking near the restaurant?

Scalability

- CUDA augments DST data with diverse user behaviors
  - Efficiency of augmentation
  - Generalizability to different datasets/domains
  - SOTA in MultiWOZ 2.1
  - Potential for user simulators

Practicality
Practicality
Salesperson: Discover Business Potential

- Human salesperson is capable of
  - mining business-related intentions from customers via chatting
  - smoothly promoting the related product through conversations
Two Types of Dialogue Systems

- Open-Domain Chatting
  - Today is a sunny day. It’s great to go on a picnic.

- Task-Oriented
  - How may I help you? Are you interested in listening some songs?
Solution:

SalesBot: Transitioning from Open-Domain to Task-Oriented Dialogues

(Chiu et al., ACL 2022)

https://github.com/MiuLab/SalesBot
Simulated User

My son is 2 and a half. He is handful, but I love him so much.

That's adorable. I have a 2-year-old son and he's the same way. Do you have any pets?

No, I don't. I work too much to have time for one. How about you?

I work a lot too. I'm a stay at home mom. What do you do for work?

I work at a library. It's a lot of fun, and I get to meet lots of interesting people.

That sounds like a fun job. I bet you meet some interesting people at the library. What else do you like to do?

I like to read a lot. I also like to go to the movies. What about yourself?

I'm looking for a movie to watch. A regular showing would be fine.

Are you interested in watching any movie?

I'm in San Francisco, please look for movies there.

There are [COUNT] movies you can watch. What do you think of [MOVIE_NAME]?

Simulated Salesperson

Such data can allow us to train a conversational agent with a salesperson's capability.
SalesBot: Social to Task-Oriented

- Motivation: no existing data with the property
- Approach: simulate the scenarios to generate unlimited data

Open-Domain Dialogue Generation
- Open-Domain BlenderBot
- Open-Domain BlenderBot

Task-Oriented Dialogue (TOD) Generation
- TOD User BlenderBot
- TOD Sales BlenderBot

How to bridge two types of dialogues?

SalesBot: Social to Task-Oriented

Challenges

1. **When** to switch to the task-oriented dialogue system?
   - Task-Oriented (Implicit) Intent Detection

2. **How** to smoothly switch from chit-chat to task-oriented dialogues?
   - Transition Turn Generation

Goal: identify if the user is likely to have task-related intents

Issue: no data with annotated implicit intents

Zero-Shot Intent Detector

Idea: leverage QA system’s capability

Context

I never visit France, but I heard that it is a good place.

Question (FINDATTRCTIONS)

Does the user want to travel there?

intent description: find attractions to visit

Intent-associated questions are naively generated from their descriptions

SalesBot: Social to Task-Oriented

Challenges

1. **When** to switch to the task-oriented dialogue system?
   → **Task-Oriented (Implicit) Intent Detection**

2. **How** to smoothly switch from chit-chat to task-oriented dialogues?
   → **Transition Turn Generation**

## Transition Turn Generation

### Generative-based Generation:
- Training data: OTTers (Source Topic $\rightarrow$ Transition $\rightarrow$ Target Topic)

#### Entity Path: outside - garden – flower

<table>
<thead>
<tr>
<th>User A</th>
<th><strong>Source Topic</strong>: I spend a lot of time <strong>outside</strong>. (Source Topic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>User B</td>
<td><strong>Transition</strong>: I like the outdoors as well, especially <strong>gardening</strong>. It destresses me. <strong>Target Topic</strong>: I enjoy relaxing and getting <strong>flowers</strong>.</td>
</tr>
</tbody>
</table>

#### Entity Path: seafood - Swedish fish – candy

<table>
<thead>
<tr>
<th>User A</th>
<th><strong>Source Topic</strong>: I like <strong>seafood</strong> a lot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>User B</td>
<td><strong>Transition</strong>: Since you like seafood, is <strong>Swedish fish</strong> a candy that you might enjoy? <strong>Target Topic</strong>: I have no self control when it comes to <strong>candy</strong>.</td>
</tr>
</tbody>
</table>

#### Entity Path: engagement - marriage - child

<table>
<thead>
<tr>
<th>User A</th>
<th><strong>Source Topic</strong>: I think I am getting engaged soon.</th>
</tr>
</thead>
<tbody>
<tr>
<td>User B</td>
<td><strong>Transition</strong>: I have two children from a previous <strong>marriage</strong> <strong>Target Topic</strong>: My <strong>children</strong> are my life.</td>
</tr>
</tbody>
</table>

OTTers: One-turn Topic Transitions for Open-Domain Dialogue, ACL2021
SalesBot Simulation Framework

Open-Domain Dialogue Generation

Chit-Chat to Task-Oriented Transition

TOD Generation

Contribution: simulate unlimited dialogues transitioning from chit-chat to task-oriented

Quality?

Human Evaluation

Overall dialogue quality

- **Q1 Relevance**: How relevant is the recommended service to the conversation context?

- **Q2 Aggressiveness**: How aggressive is the salesperson's communication strategy?

- **Q3 Overall**: Do you think the conversation is overall a good example of making a sales recommendation?

Average scores over 3 workers (4000 dialogues)

Robustness

- LatticeLM for preserving uncertainty
- Contrastive learning with only textual information

- **Speech Corpus** → ASR (clean)
  - wrack a nice beach
  - recognize speech
  
- **Pre-Training**

- **SLU Corpus** → ASR (label)
  - new email for contact
  - Intent=SendEmail

- **Contrastive Pre-training** learns error-invariant sentence embeddings
- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

Scalability

- **CUDA** augments DST data with diverse user behaviors
  - Efficiency of augmentation
  - Generalizability to different datasets/domains
  - SOTA in MultiWOZ 2.1
  - Potential for user simulators

Practicality

- **SalesBot** simulates unlimited data bridging two types of agents
  - Reasonable quality
  - Data/simulators for learning (SL/RL) end-to-end agents with sales’ behavior
Robustness

✓ **LatticeLM** for preserving uncertainty
✓ **Contrastive learning** with only textual information

- **Contrastive Pre-training** learns error-invariant sentence embeddings
- **Supervised CL with Self Distillation** improves robustness to noises from ASR and labels

Thank you, can you also find me a hotel with parking near the restaurant?

Scalability

✓ **CUDA** augments DST data with diverse user behaviors
  - Efficiency of augmentation
  - Generalizability to different datasets/domains
  - SOTA in MultiWOZ 2.1
  - Potential for user simulators

Practicality

✓ **SalesBot** simulates unlimited data bridging two types of agents
  - Reasonable quality
  - Data/simulators for learning (SL/RL) end-to-end agents with sales’ behavior
Yun-Nung (Vivian) Chen
Associate Professor, National Taiwan University
y.v.chen@ieee.org / http://vivianchen.idv.tw
Appendix
Lattice Language Modeling

1) LatticeLSTM encodes nodes of a lattice

2) The goal is to predict the outgoing transitions (words) given a node’s representation

- The one-hypothesis lattice reduces to normal language modeling

**Issue:** LatticeLSTM runs prohibitively slow

Ablation Study

\[ L_{pt} = L_c + \lambda_{mlm} \cdot L_{mlm} \]
\[ L_{ft} = L_{ce} + \lambda_d L_d + \lambda_{hard} L_{hard} + \lambda_{soft} L_{soft} \]

<table>
<thead>
<tr>
<th>Pre-Training</th>
<th>Fine-Tuning</th>
<th>SLURP</th>
<th>ATIS</th>
<th>TREC6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Full</td>
<td>85.26</td>
<td>95.10</td>
<td>86.36</td>
</tr>
<tr>
<td>No ( L_{mlm} )</td>
<td>Full</td>
<td>84.83</td>
<td>93.75</td>
<td>85.32</td>
</tr>
<tr>
<td>No ( L_c )</td>
<td>Full</td>
<td>85.15</td>
<td>95.00</td>
<td>85.53</td>
</tr>
<tr>
<td>Full</td>
<td>No ( L_{hard} + L_{soft} )</td>
<td>85.14</td>
<td>94.83</td>
<td>86.08</td>
</tr>
<tr>
<td>Full</td>
<td>No ( L_d + L_{soft} )</td>
<td>84.77</td>
<td>94.75</td>
<td>85.60</td>
</tr>
<tr>
<td>Full</td>
<td>No ( L_{soft} )</td>
<td>84.81</td>
<td>94.65</td>
<td>86.20</td>
</tr>
</tbody>
</table>

All parts in the proposed approach are necessary to achieve better SLU performance.
## Improvement of Different WER

<table>
<thead>
<tr>
<th>Pre-Training</th>
<th>Fine-Tuning</th>
<th>Clean =0</th>
<th>Low (0, 0.16)</th>
<th>Medium (0.16, 0.40)</th>
<th>High &gt;0.4</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>Direct</td>
<td>95.69</td>
<td>92.41</td>
<td>85.89</td>
<td>56.71</td>
<td>83.97</td>
</tr>
<tr>
<td>Phoneme-BERT</td>
<td>Direct</td>
<td>94.97</td>
<td>92.34</td>
<td>85.87</td>
<td>57.20</td>
<td>83.78</td>
</tr>
<tr>
<td>SimCSE</td>
<td>Direct</td>
<td>95.55</td>
<td>93.47</td>
<td><strong>86.82</strong></td>
<td>57.59</td>
<td>84.47</td>
</tr>
<tr>
<td>Proposed</td>
<td>Direct</td>
<td>95.54</td>
<td>93.86</td>
<td>86.68</td>
<td><strong>57.72</strong></td>
<td><strong>84.51</strong></td>
</tr>
<tr>
<td>RoBERTa</td>
<td>Proposed</td>
<td>96.59</td>
<td>94.27</td>
<td>86.70</td>
<td>57.24</td>
<td>84.87</td>
</tr>
<tr>
<td>Phoneme-BERT</td>
<td>Proposed</td>
<td>95.61</td>
<td>93.42</td>
<td>86.87</td>
<td>57.50</td>
<td>84.48</td>
</tr>
<tr>
<td>SimCSE</td>
<td>Proposed</td>
<td>96.57</td>
<td><strong>94.54</strong></td>
<td>87.39</td>
<td>58.01</td>
<td>85.25</td>
</tr>
<tr>
<td>Proposed</td>
<td>Proposed</td>
<td>96.08</td>
<td>94.41</td>
<td><strong>87.63</strong></td>
<td><strong>58.72</strong></td>
<td><strong>85.26</strong></td>
</tr>
</tbody>
</table>

Proposed approach is more effective when WER is higher

Proposed fine-tuning can generalize to diverse pre-training strategies for better SLU results
Task-Oriented Dialogue Generation

- Task-Oriented Simulation
  - Two BlenderBot simulators are additionally trained on
    - user turns to simulate users
    - agent turns to simulate salespersons
  - These turns are taken from task-oriented dialogues.
Transition Turn Generation

- **Template-based Generation:**
  - Use a template sentence to trigger the corresponding task-oriented user reaction

<table>
<thead>
<tr>
<th>Template-based generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User:</strong> I like to read a lot. I also like to go to the movies. What about yourself? - <strong>FindMovies</strong></td>
</tr>
<tr>
<td><strong>Sales:</strong> <em>Do you want to find movies by genre and optionally director?</em></td>
</tr>
<tr>
<td><strong>User:</strong> I’m looking for a movie to watch. A regular showing would be fine.</td>
</tr>
</tbody>
</table>

- **Generative-based Generation:**
  - Re-generate the transition turn for better fluency and diversity

<table>
<thead>
<tr>
<th>Generative-based Re-generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User:</strong> I like to read a lot. I also like to go to the movies. What about yourself?</td>
</tr>
<tr>
<td><strong>Sales:</strong> <em>Are you interested in watching any movie?</em></td>
</tr>
<tr>
<td><strong>User:</strong> I’m looking for a movie to watch. A regular showing would be fine.</td>
</tr>
</tbody>
</table>
Human Evaluation

Transition turn quality

- **Q1 Timing**: Is it a good timing to make the transition?
- **Q2 Relevance**: Is the transition relevant to the conversation context?
- **Q3 Aggressiveness**: Is the transition aggressive?
- **Q4 Overall**: Do you think it is overall a good transition?

All scores above 3 (neutral) demonstrates reasonable quality of the generated data.