Towards Open-Domain Conversational AI

YUN-NUNG (VIVIAN) CHEN 陳縝儂

HTTP://VIVIANCHEN.IDV.TW
Iron Man (2008)

What can machines achieve now or in the future?
Language Empowering Intelligent Assistants


Google Assistant (2016)

Why and When We Need?

“I want to chat”
“I have a question”
“I need to get this done”
“What should I do?”

Turing Test (talk like a human)
Information consumption
Task completion
Decision support
Why and When We Need?

“I want to chat”
“I have a question”
“I need to get this done”
“What should I do?”

- Turing Test (talk like a human)
- Information consumption
- Task completion
- Decision support
Why and When We Need?

“I want to chat”
“I have a question”
“I need to get this done”
“What should I do?”

Turing Test (talk like a human)
Information consumption
Task completion
Decision support

• What is today’s agenda?
• Which room is IR workshop in?
• What does IR stand for?
Why and When We Need?

“I want to chat”  Turing Test (talk like a human)
“I have a question”  Information consumption
“I need to get this done”  Task completion
“What should I do?”  Decision support

- Book me the ticket from Taichung to Taipei
- Reserve a table at Din Tai Fung for 5 people, 7PM tonight
- Schedule a meeting with Vivian at 10:00 tomorrow
Why and When We Need?

“I want to chat”
“I have a question”
“I need to get this done”
“What should I do?”

- Turing Test (talk like a human)
- Information consumption
- Task completion
- Decision support

- Is this restaurant worth to try?
- Is the IR workshop good to attend?
Why and When We Need?

“I want to chat”
“I have a question”
“I need to get this done”
“What should I do?”

Turing Test (talk like a human)
Information consumption
Task completion
Decision support

Social Chit-Chat
Task-Oriented Dialogues
Intelligent Assistants

Task-Oriented

- show blockers assigned to Peter
- get me Javier’s contact info
- who does Peter work for?
- add a half hour meeting Friday at 11
- current weather in London
- how’s my Jira sprint status?
- what’s my first meeting on Wed?
- show me next week’s schedule
- when did Jennifer join our company?
- translate “Good morning” to French
- show Rodney’s agenda
- get today’s agenda
- where does Charlotte sit?
- how’s my calendar tomorrow
- what is the local time in Dublin, Ireland?
- when is Jim Harris anniversary?
- cancel my 9 o’clock meeting tomorrow
- how’s my high cycle stories
- Show github leaderboard
- show all my pull requests
- which team does Olivia part of?
- when is Jonathan’s birthday?
- get all Github commits in last 7d

Task-Oriented
Conversational Agents

Chit-Chat
- Seq2seq models
- Seq2seq with conversation contexts
- Knowledge-grounded seq2seq models

Task-Oriented
- Single-domain, system-initiative
- Multi-domain, contextual, mixed-initiative
- End-to-end learning, massively multi-domain
Task-Oriented Dialogue Systems

JARVIS – Iron Man’s Personal Assistant

Baymax – Personal Healthcare Companion
Task-Oriented Dialogue Systems (Young, 2000)

Speech Signal

Hypothesis
Are there any action movies to see this weekend?

Text Input
Are there any action movies to see this weekend?

Language Understanding (LU)
- Domain Identification
- User Intent Detection
- Slot Filling

Semantic Frame
request_movie
genre=action, date=this weekend

Natural Language Generation (NLG)

Text response
Where are you located?

System Action/Policy
request_location

Dialogue Management (DM)
- Dialogue State Tracking (DST)
- Dialogue Policy

Backend Database/Knowledge Providers
Task-Oriented Dialogue Systems (Young, 2000)

Speech Recognition

Hypothesis
are there any action movies to see this weekend

Language Understanding (LU)
- Domain Identification
- User Intent Detection
- Slot Filling

Semantic Frame
request_movie
genre=action, date=this weekend

Natural Language Generation (NLG)

System Action/Policy
request_location

Dialogue Management (DM)
- Dialogue State Tracking (DST)
- Dialogue Policy

Backend Database/Knowledge Providers

Text Input
Are there any action movies to see this weekend?

Text response
Where are you located?
Semantic Frame Representation

- Requires a domain ontology: early connection to backend
- Contains core concept (intent, a set of slots with fillers)

**Restaurant Domain**

```
find me a cheap taiwanese restaurant in oakland
```

```
find_restaurant (price="cheap", type="taiwanese", location="oakland")
```

**Movie Domain**

```
show me action movies directed by james cameron
```

```
find_movie (genre="action", director="james cameron")
```
Backend Database / Ontology

- Domain-specific table
  - Target and attributes

- Functionality
  - Information access: find specific entries
  - Task completion: find the row that satisfies the constraints

<table>
<thead>
<tr>
<th>Movie Name</th>
<th>Theater</th>
<th>Rating</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iron Man Last</td>
<td>Taipei A1</td>
<td>8.5</td>
<td>2018/10/31</td>
<td>09:00</td>
</tr>
<tr>
<td>Iron Man Last</td>
<td>Taipei A1</td>
<td>8.5</td>
<td>2018/10/31</td>
<td>09:25</td>
</tr>
<tr>
<td>Iron Man Last</td>
<td>Taipei A1</td>
<td>8.5</td>
<td>2018/10/31</td>
<td>10:15</td>
</tr>
<tr>
<td>Iron Man Last</td>
<td>Taipei A1</td>
<td>8.5</td>
<td>2018/10/31</td>
<td>10:40</td>
</tr>
</tbody>
</table>
Task-Oriented Dialogue Systems (Young, 2000)

- **Language Understanding (LU)**
  - Domain Identification
  - User Intent Detection
  - Slot Filling

- **Semantic Frame**
  - request_movie
  - genre=action, date=this weekend

- **Dialogue Management (DM)**
  - Dialogue State Tracking (DST)
  - Dialogue Policy

- **Text Input**
  - Are there any action movies to see this weekend?

- **Speech Signal**
  - are there any action movies to see this weekend

- **Text Recognition**

- **Natural Language Generation (NLG)**
  - Where are you located?

- **System Action/Policy**
  - request_location

- **Backend Action / Knowledge Providers**
Language Understanding (LU)

- Pipelined

1. Domain Classification
2. Intent Classification
3. Slot Filling
RNN for Slot Tagging – I (Yao et al, 2013; Mesnil et al, 2015)

- Variations:
  a. RNNs with LSTM cells
  b. Input, sliding window of n-grams
  c. Bi-directional LSTMs

(a) LSTM
(b) LSTM-LA
(c) bLSTM
RNN for Slot Tagging – II \cite{Kurata2016,Simonnet2015}

- Encoder-decoder networks
  - Leverages sentence level information

- Attention-based encoder-decoder
  - Use attention (as in MT) in the encoder-decoder network
Joint Semantic Frame Parsing

- **Sequence-based** (Hakkani-Tur et al., 2016)
  - Slot filling and intent prediction in the same output sequence

- **Parallel** (Liu and Lane, 2016)
  - Intent prediction and slot filling are performed in two branches
## Joint Model Comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Attention Mechanism</th>
<th>Intent-Slot Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint bi-LSTM</td>
<td>X</td>
<td>Δ (Implicit)</td>
</tr>
<tr>
<td>Attentional Encoder-Decoder</td>
<td>✓</td>
<td>Δ (Implicit)</td>
</tr>
<tr>
<td>Slot Gate Joint Model</td>
<td>✓</td>
<td>✓ (Explicit)</td>
</tr>
</tbody>
</table>
Slot-Gated Joint SLU (Goo et al., 2018)

- **Slot Gate**
  \[ g = \sum v \cdot \tanh(c^s_i + W \cdot c^I) \]

- **Slot Prediction**
  \[ y^S_i = \text{softmax}(W^S(h_i + c^S_i) + b^S) \]
  \[ y^S_i = \text{softmax}(W^S(h_i + g \cdot c^S_i) + b^S) \]

  *\( g \) will be larger if slot and intent are better related*
Contextual Language Understanding

For single turn:
- **U**: just sent email to bob about fishing this weekend
- **S**: B-contact_name B-subject l-subject l-subject
- `send_email(contact_name="bob", subject="fishing this weekend")`

For multi-turn:
- **U**₁: send email to bob
- **S**₁: B-contact_name
  - `send_email(contact_name="bob")`

- **U**₂: are we going to fish this weekend
- **S**₂: B-message l-message l-message l-message l-message
  - `send_email(message="are we going to fish this weekend")`
E2E MemNN for Contextual LU (Chen et al., 2016)

U: “i’d like to purchase tickets to see deepwater horizon”
S: “for which theatre”
U: “angelika”
S: “you want them for angelika theatre?”
U: “yes angelika”
S: “how many tickets would you like?”
U: “3 tickets for saturday”
S: “What time would you like?”
U: “Any time on saturday is fine”
S: “okay, there is 4:10 pm, 5:40 pm and 9:20 pm”
U: “Let’s do 5:40”
Role-Based & Time-Aware Attention (Su et al., 2018)

Sentence-Level Time-Decay Attention $\alpha_{u_i}$

Current

History Summary

Role-Level Time-Decay Attention $\alpha_{r_1}$ $\alpha_{r_2}$

Dense Layer

Spoken Language Understanding

Time-Decay Attention Function ($\alpha_u$ & $\alpha_r$)

- Convex
- Linear
- Concave
Context-Sensitive Time- Decay Attention (Su et al., 2018)

Time-decay attention significantly improves the understanding results.
Task-Oriented Dialogue Systems (Young, 2000)

Speech Recognition

Hypothesis
are there any action movies to see this weekend

Natural Language Generation (NLG)

System Action/Policy
request_location

Backend Action / Knowledge Providers

Language Understanding (LU)
- Domain Identification
- User Intent Detection
- Slot Filling

Semantic Frame
request_movie
genre=action, date=this weekend

Dialogue Management (DM)
- Dialogue State Tracking (DST)
- Dialogue Policy

Text response
Where are you located?

Text Input
Are there any action movies to see this weekend?
Hello, how may I help you?

I’m looking for a Thai restaurant.

What part of town do you have in mind?

Something in the centre.

Bangkok city is a nice place, it is in the centre of town and it serves Thai food.

What’s the address?

Bangkok city is a nice place, their address is 24 Green street.

Thank you, bye.

request (restaurant; foodtype=Thai)

inform (area=centre)

request (address)

by ()
DNN for DST

multi-turn conversation

feature extraction

DNN

A slot value distribution for each slot
RNN-CNN DST (Wen et al., 2016)
Hello, how may I help you?

I’m looking for a Thai restaurant.

What part of town do you have in mind?

Something in the centre.

Bangkok city is a nice place, it is in the centre of town and it serves Thai food.

What’s the address?

Bangkok city is a nice place, their address is 24 Green street.

Thank you, bye.

greeting ()

request (restaurant; foodtype=Thai)

request (area)

inform (area=centre)

inform (restaurant=Bangkok city, area=centre of town, foodtype=Thai)

request (address)

inform (address=24 Green street)

bye ()
Dialogue Policy Optimization

- Dialogue management in a RL framework

Environment

User

Natural Language Generation

Action A

Reward R

Language Understanding

Observation O

Dialogue Manager

Agent

Select the best action that maximizes the future reward
E2E Task-Completion Bot (TC-Bot) (Li et al., 2017)

Idea: SL for each component and RL for end-to-end training the neural dialogue system.
User goal – *Two tickets for “Deadpool” tomorrow 9PM at AMC Pacific Place 11 theater, Seattle.*

**RULE BASED SYSTEM**

User: Do you have any ticket for movie ‘Deadpool’ at 9:00 pm?
Agent: Which theater do you want?
User: AMC pacific place 11 theater, please!
Agent: How many tickets do you want?
User: Two!
Agent: What date?
User: Tomorrow!
Agent: Which city do you want?
User: Seattle!
Agent: Do you have any specification for zip code?
User: No, I do not care.
Agent: Do you have any distance constraint?
User: No, I do not care.
Agent: How about video format? 2d? Imax 3d?
User: Any kind is okay.
Agent: Do you have any constraint to price?
User: No!
Agent: Okay, I will issue 2 tickets for you, tomorrow 9:00 pm at AMC pacific place 11 theater, Seattle, movie ‘Deadpool’.
User: Okay, thanks.
Agent: Thanks!

**REINFORCEMENT LEARNING SYSTEM**

User: I want two tickets for movie ‘Deadpool’.
Agent: What is the starttime you prefer?
User: 9:00 PM
Agent: Which day?
User: Tomorrow
Agent: Which city?
User: Seattle
Agent: Which theater do you prefer?
User: AMC Pacific Place 11 theater.
Agent: Okay, I will issue 2 tickets for you, tomorrow 9:00 pm at AMC pacific place 11 theater, Seattle, movie ‘Deadpool’.
User: Okay, thanks.
Agent: Thanks!

**Issue 1: no notion about what requests can be skipped**

Skip the requests the user may not care about to improve efficiency
User goal – Two tickets for “Hail Caesar” tomorrow 7:15PM at AMC Lowes Oak tree 6, Seattle.

**RULE BASED SYSTEM**
User: Do you any ticket for movie ‘Hail Caesar’ at theater AMC Lowes Oak Tree 6?
Agent: What start time do you prefer?
User: 7:15 pm Tomorrow
Agent: Which city you are looking for?
User: Seattle
Agent: How many tickets do you want?
User: Two
Agent: Okay, I have issued two tickets for movie ‘Hail Caeser’ tomorrow 7:50 pm at theater AMC Lowes Oak Tree 6, Seattle.

**REINFORCEMENT LEARNING SYSTEM**
User: Do you any ticket for movie ‘Hail Caesar’ at theater AMC Lowes Oak Tree 6?
Agent: What start time do you prefer?
User: 7:15 pm Tomorrow
Agent: Which city you are looking for?
User: Seattle
Agent: How many tickets do you want?
User: Two
Agent: Okay, I will issue you tickets for movie ‘Hail Caeser’ tomorrow 7:50 pm at theater AMC Lowes Oak Tree 6, Seattle.
User: Oh, no.
Agent: It is 7:50 pm, right?
User: Two tickets for 7:15 pm, please.
Agent: Okay, I will issue you two tickets for movie ‘Hail Caeser’ tomorrow 7:15 pm at theater AMC Lowes Oak Tree 6, Seattle.
User: Thanks!
Agent: Thanks!

**Issue 2: no principle way to deal with LU errors/uncertainty**

Automatically learn when and what to confirm for better robustness
RL in Dialogue Systems

1. Sample inefficient, hard to design reward function, local optima...
2. Real users are expensive
3. Discrepancy between real users and simulators
D3Q: Discriminative Deep Dyna-Q (Su et al., 2018)

- Idea
  - learning with real users with planning
  - add a *discriminator* to filter out the bad experiences
D3Q: Discriminative Deep Dyna-Q (Su et al., 2018)

The policy learning is more robust and shows the improvement in human evaluation.
Task-Oriented Dialogue Systems (Young, 2000)

- **Speech Recognition**
  - Text Input: "Are there any action movies to see this weekend?"
  - Hypothesis: "are there any action movies to see this weekend"

- **Language Understanding (LU)**
  - Domain Identification
  - User Intent Detection
  - Slot Filling
  - Semantic Frame: request_movie, genre=action, date=this weekend

- **Dialogue Management (DM)**
  - Dialogue State Tracking (DST)
  - Dialogue Policy
  - System Action/Policy: request_location
  - Text response: "Where are you located?"

- **Natural Language Generation (NLG)**
  - Backend Action / Knowledge Providers

- **Dialogue State Tracking (DST)**
  - Dialogue Policy
natural Language Generation (NLG)

- Mapping dialogue acts into natural language

\[ \text{inform(name=Seven\_Days, foodtype=Chinese)} \]

Seven Days is a nice Chinese restaurant
Template-Based NLG

- Define a set of rules to map frames to natural language

<table>
<thead>
<tr>
<th>Semantic Frame</th>
<th>Natural Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>confirm()</td>
<td>“Please tell me more about the product you are looking for.”</td>
</tr>
<tr>
<td>confirm(area=$V)</td>
<td>“Do you want somewhere in the $V?”</td>
</tr>
<tr>
<td>confirm(food=$V)</td>
<td>“Do you want a $V restaurant?”</td>
</tr>
<tr>
<td>confirm(food=$V,area=$W)</td>
<td>“Do you want a $V restaurant in the $W.”</td>
</tr>
</tbody>
</table>

**Pros:** simple, error-free, easy to control

**Cons:** time-consuming, rigid, poor scalability
RNN-Based LM NLG (Wen et al., 2015)

Input:
Inform(name=Din Tai Fung, food=Taiwanese)

dialogue act 1-hot representation

conditioned on the dialogue act

Output:
\[
<\text{BOS}> \quad \text{SLOT\_NAME} \quad \text{serves} \quad \text{SLOT\_FOOD} \quad . \quad <\text{EOS}>
\]

delexicalisation

\[
<\text{BOS}> \quad \text{Din Tai Fung} \quad \text{serves} \quad \text{Taiwanese} \quad .
\]
Semantic Conditioned LSTM (Wen et al., 2015)

- Issue: semantic repetition
  - Din Tai Fung is a great Taiwanese restaurant that serves Taiwanese.
  - Din Tai Fung is a child friendly restaurant, and also allows kids.

Idea: using gate mechanism to control the generated semantics (dialogue act/slots)
Issues in NLG

- **Issue**
  - NLG tends to generate *shorter* sentences
  - NLG may generate *grammatically-incorrect* sentences

- **Solution**
  - Generate word patterns in a order
  - Consider *linguistic patterns*
Hierarchical NLG w/ Linguistic Patterns (Su et al., 2018)

**Bidirectional GRU Encoder**
- Input: 
  - name: Midsummer House
  - food: Italian
  - priceRange: moderate
  - near: All Bar One

**Semantic 1-hot Representation**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**GRU Decoder**
- 1. Repeat-input
- 2. Inner-Layer Teacher Forcing
- 3. Inter-Layer Teacher Forcing
- 4. Curriculum Learning

**DECODING LAYER1**
- **NOUN + PROPN + PRON**
  - **1.** Near All Bar One is a moderately priced Italian place it is called **Midsummer House**

**DECODING LAYER2**
- **VERB**
  - **2.** All Bar One is **priced** place it is called **Midsummer House**

**DECODING LAYER3**
- **ADJ + ADV**
  - **3.** All Bar One is **moderately** priced **Italian** place it is called **Midsummer House**

**DECODING LAYER4**
- **4.** Others
  - **4.** All Bar One is a **moderately** priced **Italian** place it is called **Midsummer House**
Evolution Roadmap

Dialogue breadth (coverage)

Dialogue depth (complexity)

I’ve got a cold what do I do?

Tell me a joke.

What is influenza?

I feel sad...

Single domain systems

Extended systems

Multi-domain systems

Open domain systems

What is influenza?
Chit-Chat Social Bots
Neural Response Generation (Sordoni et al., 2015; Vinyals & Le, 2015)

Target: response decoder

Yeah I’m on my way

Source: conversation history

... because of your game?

encoder

\[ \hat{T} = \arg \max_T \{ \log p(T|S) \} \]

Learns to generate dialogues from offline data (no state, action, intent, slot, etc.)
‘tis a fine brew on a day like this! Strong though, how many is sensible?

I'm not sure yet, I'll let you know!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?

I don’t know.

Wow sour starbursts really do make your mouth water... mm drool. Can I have one?

Of course!

Well he was on in Bromley a while ago... still touring.

I don't even know what he's talking about.

32% responses are general and meaningless

“I don’t know”

“I don’t know what you are talking about”

“I don’t think that is a good idea”

“Oh my god”
Wow sour starbursts really do make your mouth water... mm drool. Can I have one?

Of course you can! They’re delicious!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?

I think he’d be a good signing.

‘tis a fine brew on a day like this! Strong though, how many is sensible?

Depends on how much you drink!

Well he was on in Bromley a while ago... still touring.

I’ve never seen him live.
Wow sour starbursts really do make your mouth water... mm drool. Can I have one?

Of course you can! They’re delicious!

Milan apparently selling Zlatan to balance the books... Where next Madrid?

I think he’d be a good signing.

‘tis a fine brew on a day like this! Strong though, how many is sensible?

Depends on how much you drink!

Well he was on in Bromley a while ago... still touring.

I’ve never seen him live.
Real-World Conversations

- Multimodality
  - Conversation history
  - Persona
  - User profile data (bio, social graph, etc.)
  - Visual signal (camera, picture etc.)
  - Knowledge base
  - Mood
  - Geolocation
  - Time

```
Because of your game?
```

EOS

Yeah

I’m

...
### Issue 2: Response Inconsistency

<table>
<thead>
<tr>
<th>message</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>In which city do you live now?</td>
<td>I live in Madrid.</td>
</tr>
<tr>
<td>In which country do you live now?</td>
<td>England, you?</td>
</tr>
<tr>
<td>Where were you born?</td>
<td>I was born in Canada.</td>
</tr>
<tr>
<td>Where are you from?</td>
<td>England, you?</td>
</tr>
<tr>
<td>Where did you grow up?</td>
<td>I grew up in Texas.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>message</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>How old are you?</td>
<td>16 and you?</td>
</tr>
<tr>
<td>What’s your age?</td>
<td>18.</td>
</tr>
<tr>
<td>What is your major?</td>
<td>I’m majoring in psychology</td>
</tr>
<tr>
<td>What did you study in college?</td>
<td>English lit.</td>
</tr>
</tbody>
</table>
Personalized Response Generation (Li et al., 2016)
### Persona Model for Speaker Consistency (Li et al., 2016)

<table>
<thead>
<tr>
<th>Baseline model → inconsistency</th>
<th>Persona model using speaker embedding → consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>message</strong> Where is your hometown?</td>
<td><strong>message</strong> Where is your hometown?</td>
</tr>
<tr>
<td><strong>response</strong> I live in <strong>Indonesia</strong>.</td>
<td><strong>response</strong> I’m from <strong>England</strong>.</td>
</tr>
<tr>
<td><strong>message</strong> Where are you from?</td>
<td><strong>message</strong> Where are you from?</td>
</tr>
<tr>
<td><strong>response</strong> <strong>England</strong>, you?</td>
<td><strong>response</strong> I’m from <strong>England</strong>.</td>
</tr>
<tr>
<td><strong>message</strong> In which city do you live now?</td>
<td><strong>message</strong> In which city do you live now?</td>
</tr>
<tr>
<td><strong>response</strong> I live in <strong>Jakarta</strong>.</td>
<td><strong>response</strong> I live in <strong>London</strong>.</td>
</tr>
<tr>
<td><strong>message</strong> In which country do you live now?</td>
<td><strong>message</strong> In which country do you live now?</td>
</tr>
<tr>
<td><strong>response</strong> I live in <strong>Indonesia</strong>.</td>
<td><strong>response</strong> I live in <strong>England</strong>.</td>
</tr>
<tr>
<td><strong>message</strong> What is your major?</td>
<td><strong>message</strong> What is your major?</td>
</tr>
<tr>
<td><strong>response</strong> <strong>Business</strong>, you?</td>
<td><strong>response</strong> <strong>Business</strong>, You?</td>
</tr>
<tr>
<td><strong>message</strong> What did you study in college?</td>
<td><strong>message</strong> What did you study in college?</td>
</tr>
<tr>
<td><strong>response</strong> <strong>Psychology</strong>, you?</td>
<td><strong>response</strong> I did <strong>business</strong> studies.</td>
</tr>
<tr>
<td><strong>message</strong> How old are you?</td>
<td><strong>message</strong> How old are you?</td>
</tr>
<tr>
<td><strong>response</strong> I’m 18.</td>
<td><strong>response</strong> I’m 18.</td>
</tr>
<tr>
<td><strong>message</strong> What is your age?</td>
<td><strong>message</strong> What is your age?</td>
</tr>
<tr>
<td><strong>response</strong> I’m 16.</td>
<td><strong>response</strong> I’m 18.</td>
</tr>
</tbody>
</table>
Issue 3: Dialogue-Level Optimization via RL

User input (o) \[ \rightarrow \] Language understanding \[ s \] \[ \rightarrow \] Dialogue Manager \[ a = \pi(s) \] \[ \rightarrow \] Collect rewards \[ (s, a, r, s') \]

<table>
<thead>
<tr>
<th>Application</th>
<th>State</th>
<th>Action</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Completion Bots (Movies, Restaurants, …)</td>
<td>User input + Context</td>
<td>Dialog act + slot-value</td>
<td>Task success rate # of turns</td>
</tr>
<tr>
<td>Info Bots (Q&amp;A bot over KB, Web etc.)</td>
<td>Question + Context</td>
<td>Clarification questions, Answers</td>
<td>Relevance of answer # of turns</td>
</tr>
<tr>
<td>Social Bot (Xiaolce)</td>
<td>Conversation history</td>
<td>Response</td>
<td>Engagement(?)</td>
</tr>
</tbody>
</table>
Deep RL for Response Generation *(Li et al., 2016)*

<table>
<thead>
<tr>
<th>Input message</th>
<th>Supervised Learning Agent</th>
<th>Reinforcement Learning Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>How old are you?</td>
<td>I’m 16.</td>
<td>I’m 16, why are you asking?</td>
</tr>
<tr>
<td>What’s your full name?</td>
<td>I have no idea.</td>
<td>What’s yours?</td>
</tr>
<tr>
<td>I don’t want to go home tonight.</td>
<td>Really?</td>
<td>Why?</td>
</tr>
<tr>
<td>Do you have any feelings for me?</td>
<td>I don’t know what you are talking about.</td>
<td>I don’t want to hurt your feelings.</td>
</tr>
<tr>
<td>How much time do you have here?</td>
<td>Not long enough. Sorry, sir.</td>
<td>Ten seconds.</td>
</tr>
<tr>
<td>Shall we get started?</td>
<td>Of course!</td>
<td>Yes, We’ve got a lot of work to do here.</td>
</tr>
<tr>
<td>Do you play football?</td>
<td>No, i don’t</td>
<td>Yes, I love football!</td>
</tr>
<tr>
<td>We’d have to talk to him.</td>
<td>I mean, he’s a good guy</td>
<td>About what?</td>
</tr>
<tr>
<td>How come you never say it?</td>
<td>Because I don’t want to hurt you.</td>
<td>I don’t think it’s a good idea to say it.</td>
</tr>
</tbody>
</table>

- RL agent generates more *interactive* responses
- RL agent tends to end a sentence *with a question* and hand the conversation over to the user
Issue 4: No Grounding \cite{Sordoni et al., 2015; Li et al., 2016}

Neural model learns the general shape of conversations, and the system output is situationally appropriate and coherent.

No grounding into a real calendar, but the “shape” of the conversation is fluent and plausible.

<table>
<thead>
<tr>
<th>H:</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>M:</td>
<td>how are you?</td>
</tr>
<tr>
<td>H:</td>
<td>not bad</td>
</tr>
<tr>
<td>M:</td>
<td>what's wrong?</td>
</tr>
<tr>
<td>H:</td>
<td>nothing really</td>
</tr>
<tr>
<td>M:</td>
<td>wanna talk about it? i 'm here if you wanna talk</td>
</tr>
<tr>
<td>H:</td>
<td>well, i'm just tired</td>
</tr>
<tr>
<td>M:</td>
<td>me too, i'm here if you wanna talk</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H:</th>
<th>would thursday afternoon work sometime?</th>
</tr>
</thead>
<tbody>
<tr>
<td>M:</td>
<td>yeah , sure . just let me know when you’re free.</td>
</tr>
<tr>
<td>H:</td>
<td>after lunch is probably the best time</td>
</tr>
<tr>
<td>M:</td>
<td>okay, sounds good . just let me know when you’re free.</td>
</tr>
<tr>
<td>H:</td>
<td>would 2 pm work for you?</td>
</tr>
<tr>
<td>M:</td>
<td>works for me.</td>
</tr>
<tr>
<td>H:</td>
<td>well let’s say 2 pm then i ‘ll see you there</td>
</tr>
<tr>
<td>M:</td>
<td>sounds good.</td>
</tr>
</tbody>
</table>
The weather is so depressing these days.
I know, I dislike rain too.
What about a day trip to eastern Washington?
Any recommendation?
Try Dry Falls, it’s spectacular!
Image-Grounded Dialogues (Mostafazadeh et al., 2017)

Neural model, trained on 250K Twitter convos:
- Turn 1: tweet + accompanying photo
- Turn 2: tweet is a question

Given data, other signals can be exploited in the same way: program state, database info, geolocation, etc.

Tweet: my poor baby moms saying it 's probably totaled so rip Bessie
NeuroCon: How did you manage to do that?

Tweet: as penitence I offer this
NeuroCon: can I have one of those?

Tweet: some of the damage done
NeuroCon: Is that at your house?
Knowledge-Grounded Responses (Ghazvininejad et al., 2017)

Going to Kusakabe tonight

Conversation History

Dialogue Encoder

Σ

Decoder

Fact Encoder

Try omakase, the best in town

Response

Consistently the best omakase

Amazing sushi tasting [...]

They were out of kaisui [...]

Contextually- Relevant “Facts”
Conversation and Non-Conversation Data

You know any good **A** restaurant in **B**?

**Try C**, one of the best **D** in the city.

You know any good **Japanese** restaurant in **Seattle**?

**Try Kisaku**, one of the best **sushi restaurants** in the city.
Evolution Roadmap

Dialogue breadth (coverage)

I’ve got a cold what do I do?
Common sense system
Tell me a joke.

What is influenza?
Knowledge based system

Empathetic systems
I feel sad...

Dialogue depth (complexity)
Multimodality & Personalization (Chen et al., 2018)

- Task: user intent prediction
- Challenge: language ambiguity

- User preference
  ✓ Some people prefer “Message” to “Email”
  ✓ Some people prefer “Ping” to “Text”

- App-level contexts
  ✓ “Message” is more likely to follow “Camera”
  ✓ “Email” is more likely to follow “Excel”

Behavioral patterns in history helps intent prediction.
High-Level Intention Learning (Sun et al., 2016; Sun et al., 2016)

- High-level intention may span several domains

Users interact via high-level descriptions and the system learns how to plan the dialogues.
Empathy in Dialogue System (Fung et al., 2016)

- Embed an empathy module
  - Recognize emotion using multimodality
  - Generate emotion-aware responses

![Emotion Recognizer Diagram]

- Text
- Speech
- Vision

**Emotion Recognizer**

```
{ "recognition": "Race: Asian Confidence: 65.42750000000001 Smiling: 3.95986 Gender: Female Confidence: 88.9369",
  "race": "Asian",
  "race_confidence": "65.42750000000001",
  "smiling": "3.95986",
  "gender": "Female",
  "gender_confidence": "88.9369"
}
```
Cognitive Behavioral Therapy (CBT)

- Mood Tracking
- Pattern Mining
- Depression Reduction
- Content Providing
- Always Be There
- Know You Well

Daily lessons and check-ins
Quick conversations to feel better
Challenges and Conclusions
The human-machine interface is a hot topic but several components must be integrated!

Most state-of-the-art technologies are based on DNN
  • Requires huge amounts of labeled data
  • Several frameworks/models are available

Fast domain adaptation with scarce data + re-use of rules/knowledge

Handling reasoning and personalization

Data collection and analysis from un-structured data

Complex-cascade systems require high accuracy for working good as a whole
Her (2013)

What can machines achieve now or in the future?
Thanks for Your Attention!

Q & A

Yun-Nung (Vivian) Chen
Assistant Professor
National Taiwan University
y.v.chen@ieee.org / http://vivianchen.idv.tw