Open-Domain Neural Dialogue Systems

YUN-NUNG (VIVIAN) CHEN 陳縈儂
HTTP://VIVIANCHEN.IDV.TW
Iron Man (2008)

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

- Apple Siri (2011)
- Google Now (2012)
- Microsoft Cortana (2014)
- Google Assistant (2016)
- Amazon Alexa/Echo (2014)
- Facebook M & Bot (2015)
- Google Home (2016)
- Apple HomePod (2017)

Material: http://opendialogue.miulab.tw
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

Material: http://opendialogue.miulab.tw
Why and When We Need?

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

<table>
<thead>
<tr>
<th>Reason</th>
<th>Need</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turing Test (talk like a human)</td>
<td>Information consumption</td>
</tr>
<tr>
<td></td>
<td>Task completion</td>
</tr>
<tr>
<td></td>
<td>Decision support</td>
</tr>
</tbody>
</table>
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 dialogue tutorial in?
• What does ISCSLP stand for?
Why and When We Need?

“I want to chat”
“Turing Test (talk like a human)"
“Information consumption"
“Task completion"

“I have a question”
“Decision support"

“I need to get this done”

“What should I do?”

- Book me the train ticket from Taoyuan 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 bubble tea worth to try?*
- *Is the ISCSLP conference 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
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?

Text response
Where are you located?

Speech Recognition

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

Semantic Frame
request_movie
gender=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

Material: http://opendialogue.miulab.tw
Task-Oriented Dialogue Systems (Young, 2000)

- Speech Recognition
- Language Understanding (LU)
  - Domain Identification
  - User Intent Detection
  - Slot Filling
- Natural Language Generation (NLG)
- Dialogue Management (DM)
  - Dialogue State Tracking (DST)
  - Dialogue Policy
- Backend Database/Knowledge Providers

Speech Signal: Are there any action movies to see this weekend?
Hypothesis: are there any action movies to see this weekend

Semantic Frame:
request_movie
genre=action, date=this weekend

System Action/Policy:
request_location

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)*

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 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:

- RNNs with LSTM cells
- Input, sliding window of n-grams
- Bi-directional LSTMs
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**
  \[ g = \sum v \cdot \tanh(c_i^S + W \cdot c^I) \]
  - \( W^S \): matrix for output layer
  - \( b^S \): bias for output layer
  - \( c_i^S \): slot context vector
  - \( c^I \): intent context vector

- **Slot Prediction**
  \[ y_i^S = \text{softmax}(W^S(h_i + c_i^S) + b^S) \]
  \[ y_i^S = \text{softmax}(W^S(h_i + g \cdot c_i^S) + b^S) \]
  - \( g \) will be larger if slot and intent are better related

- **Diagram**
  - **Slot Sequence**
  - **Word Sequence**
  - **BLSTM**
  - **Slot Gate**
  - **Intent Attention**
  - **Slot Attention**
  - **softmax**
  - **tanh**
  - **\( W \)**: trainable matrix
  - **\( v \)**: trainable vector
  - **\( g \)**: scalar gate value
Contextual Language Understanding

Single Turn

\[ U \quad \text{just} \quad \text{sent} \quad \text{email} \quad \text{to} \quad \text{bob} \quad \text{about} \quad \text{fishing} \quad \text{this} \quad \text{weekend} \]

\[ S \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \rightarrow \quad \text{send\_email}(\text{contact\_name} = \text{"bob"}, \text{subject} = \text{"fishing this weekend"}) \]

Multi-Turn

\[ U_1 \quad \text{send email to bob} \]

\[ S_1 \quad \text{B-contact\_name} \quad \rightarrow \quad \text{send\_email}(\text{contact\_name} = \text{"bob"}) \]

\[ U_2 \quad \text{are we going to fish this weekend} \]

\[ S_2 \quad \text{B-message} \quad \text{l-message} \quad \text{l-message} \quad \text{l-message} \quad \text{l-message} \quad \rightarrow \quad \text{send\_email}(\text{message} = \text{"are we going to fish this weekend"}) \]
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 $\alpha_{ui}$

History Summary
Role-Level Time-Decay Attention $\alpha_{r1}$, $\alpha_{r2}$

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

Spoken Language Understanding

Tourist
Guide

Dense Layer
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?

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?

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

Backend Action / Knowledge Providers

Material: http://opendialogue.miulab.tw
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)
bye ()
DNN for DST

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.

inform(type=restaurant, food=Thai)
inform(area=centre)
request(address)
bye

feature extraction

state of this turn

DNN

A slot value distribution for each slot

multi-turn conversation
RNN-CNN DST (Wen et al., 2016)

(Figure from Wen et al, 2016)
Dialogue Policy Optimization

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)

inform (address=24 Green street)

bye ()
Dialogue Policy Optimization

- Dialogue management in a RL framework

Select the best action that maximizes the future reward
Reward for RL $\cong$ Evaluation for System

- Dialogue is a special RL task
  - Human involves in interaction and rating (evaluation) of a dialogue
  - Fully human-in-the-loop framework
- Rating: correctness, appropriateness, and adequacy

<table>
<thead>
<tr>
<th>Rating Type</th>
<th>Quality Description</th>
<th>Cost Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert rating</td>
<td>high quality, high cost</td>
<td></td>
</tr>
<tr>
<td>User rating</td>
<td>unreliable quality, medium cost</td>
<td></td>
</tr>
<tr>
<td>Objective rating</td>
<td>Check desired aspects, low cost</td>
<td></td>
</tr>
</tbody>
</table>
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

Natural Language
Are there any action movies to see this weekend?

Dialogue Policy
request_location

Knowledge Database
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
Reinforcement Learning Agent (Li et al., 2017)

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 Caesar’ tomorrow **7:50 pm** at theater AMC Lowes Oak Tree 6, Seattle.
User: Oh, no.
Agent: Thanks!

**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 Caesar’ 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 Caesar’ 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
Interactive RL for DM (Shah et al., 2016)

Use a third agent for providing interactive feedback to the DM
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

---

![Diagram of D3Q: Discriminative Deep Dyna-Q](http://opendialogue.miulab.tw)
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 Signal

Hypothesis
are there any action movies to see this weekend

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

Speech Recognition

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

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

Backend Action / Knowledge Providers

Material: http://opendialogue.miulab.tw
Natural Language Generation (NLG)

- Mapping dialogue acts into natural language

```plaintext
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 your 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:
delexicalisation
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)
Contextual NLG (Dušek and Jurčíček, 2016)

- Goal: adapting users’ way of speaking, providing context-aware responses
  - Context encoder
  - Seq2Seq model
Controlled Text Generation (Hu et al., 2017)

- Idea: NLG based on generative adversarial network (GAN) framework
  - c: targeted sentence attributes
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)

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

**Bidirectional GRU Encoder**

Semantic 1-hot Representation
[... 1, 0, 0, 1, 0, ...]

Input Semantics
- name[Midsummer House], food[Italian],
- priceRange[moderate], near[All Bar One]

**DECODING LAYER1**
1. NOUN + PROPN + PRON

**DECODING LAYER2**
2. VERB

**DECODING LAYER3**
3. ADJ + ADV

**DECODING LAYER4**
4. Others

Near All Bar One is a moderately priced Italian place it is called Midsummer House
Evolution Roadmap

Dialogue breadth (coverage)

Dialogue depth (complexity)

Single domain systems

Extended systems

Multi-domain systems

Open domain systems

I've got a cold what do I do?

Tell me a joke.

What is influenza?

I feel sad...
Dialogue Systems

Task-Oriented Dialogue

---

Fully Data-Driven

---

Input $x$ → Understanding (NLU) → State tracker

Input $x$ → Generation (NLG) → Dialog policy

Database

External knowledge

Fully Data-Driven

Input $x$ → Statistical model (e.g., neural) → Output $y$
Chit-Chat Social Bots
Neural Response Generation (Sordoni et al., 2015; Vinyals & Le, 2015)

Source: conversation history

encoder

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

decoder

Target: response

Yeah I’m on my way

Learns to generate dialogues from offline data (no state, action, intent, slot, etc.)
Issue 1: Blandness Problem

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

Of course!

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

I don’t know.

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

I'm not sure yet.

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”
Mutual Information for Neural Generation

- Mutual information objective

\[ \hat{T} = \arg \max_T \left\{ \log \frac{p(S, T)}{p(S)p(T)} \right\} \]

\[ \hat{T} = \arg \max_T \left\{ \log p(T|S) - \lambda \log p(T) \right\} \quad \text{Bayes' rule} \]

\[ \hat{T} = \arg \max_T \left\{ (1 - \lambda) \log p(T|S) + \lambda \log p(S|T) \right\} \]
Mutual Information for Neural Generation

\[
\hat{T} = \arg \max_{T} \left\{ \log p(T|S) - \lambda \log p(T) \right\}
\]

\[
\hat{T} = \arg \max_{T} \left\{ (1 - \lambda) \log p(T|S) + \lambda \log p(S|T) \right\}
\]

antiLM: large \( \lambda \) makes output ungrammatical

used in rescoring (intractable otherwise)
Lexical Diversity

# Distinct Tokens in Generated Targets (Divided by Total #)

- Standard LSTM
- LSTM + MMI (antiLM)
- LSTM + MMI (bidirectional)
MMI for Response Diversity (Li et al., 2016)

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 ...

Material: http://opendialogue.miulab.tw
## 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>
<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)

where do you live in England.

Rob

Word embeddings (50k)

Speaker embeddings (70k)

Tomcoatez

Dreamwalls

Kush_322

D_Gomes25

This_Is_Artful

DigitalDan285

Jinnmeow3

Bob_Kelly2

live okay stay
great good
monday tuesday

england

London u.s.

stay in England.

in

Rob

EOS
### Persona Model for Speaker Consistency (Li et al., 2016)

<table>
<thead>
<tr>
<th>Message</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is your hometown?</td>
<td>I live in <strong>Indonesia</strong>.</td>
</tr>
<tr>
<td>Where are you from?</td>
<td><strong>England</strong>, you?</td>
</tr>
<tr>
<td>In which city do you live now?</td>
<td>I live in <strong>Jakarta</strong>.</td>
</tr>
<tr>
<td>In which country do you live now?</td>
<td>I live in <strong>Indonesia</strong>.</td>
</tr>
<tr>
<td>What is your major?</td>
<td>Business, you?</td>
</tr>
<tr>
<td>What did you study in college?</td>
<td><strong>Psychology</strong>, you?</td>
</tr>
<tr>
<td>How old are you?</td>
<td>I’m <strong>18</strong>.</td>
</tr>
<tr>
<td>What is your age?</td>
<td>I’m <strong>16</strong>.</td>
</tr>
</tbody>
</table>

**Baseline model → inconsistency**

<table>
<thead>
<tr>
<th>Message</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is your hometown?</td>
<td>I’m from <strong>England</strong>.</td>
</tr>
<tr>
<td>Where are you from?</td>
<td>I’m from <strong>England</strong>.</td>
</tr>
<tr>
<td>In which city do you live now?</td>
<td>I live in <strong>London</strong>.</td>
</tr>
<tr>
<td>In which country do you live now?</td>
<td>I live in <strong>England</strong>.</td>
</tr>
<tr>
<td>What is your major?</td>
<td>Business, <strong>You</strong>?</td>
</tr>
<tr>
<td>What did you study in college?</td>
<td>I did <strong>business</strong> studies.</td>
</tr>
<tr>
<td>How old are you?</td>
<td>I’m <strong>18</strong>.</td>
</tr>
<tr>
<td>What is your age?</td>
<td>I’m <strong>18</strong>.</td>
</tr>
</tbody>
</table>

**Persona model using speaker embedding → consistency**
Speak-Role Aware Response \cite{Luan et al., 2017}

Speaker-Independent Conversational Model

Context

Written text

Response

Written text

Speaker Dependent Auto-Encoder Model

Parameter Sharing
Speak-Role Aware Response (Luan et al., 2017)

- Speaker role: support person
- Domain expertise: technical
- Speaking style: polite

**User input:** I am getting a loop back to login page.

**Baseline model:** Ah, ok. Thanks for the info.

**Our model:** I’m sorry to hear that. Have you tried clearing your cache and cookies?

---

Context: have you tried uninstalling nike+ connec? how about docking it to a different computer?

Input: Tried both of those techniques. But nothing works. I can try the other computer again.

**Baseline:** Let me know if there’s anything else I can help you with. Thanks for the help.

**Multitask:** Have you tried using a different browser? Let us know if you have any other questions.

---

Context: What do you see when you have a failed login attempt? Never ending spinner? Loop you back to login page?

Input: I am getting a loop back to login page, windows.

**Baseline:** Ah, ok. Thanks for the info. I’ll check it out. Thanks for the feedback.

**Multitask:** I’m sorry to hear that. Have you tried clearing your cache and cookies?

---

Context: To clarify: your first run after resetting your connection will not post to Facebook. Does this help? Let us know.

Input: That’s correct. I reset it twice! It still doesn’t work.

**Baseline:** Let me know if there’s anything else I can help you with. Thanks for the update.

**Multitask:** I’m sorry to hear that. Are you receiving any error messages? Let us know.
Issue 3: Dialogue-Level Optimization via RL

User input (o) → Language understanding → Dialogue Manager $a = \pi(s)$ → Collect rewards $(s, a, r, s')$ → Optimize $Q(s, a)$ → Response

<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 (XiaoIce)</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 (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.

H: hi
M: how are you?
H: not bad
M: what's wrong?
H: nothing really
M: wanna talk about it? i 'm here if you wanna talk
H: well, i'm just tired
M: me too, i'm here if you wanna talk

H: would thursday afternoon work sometime?
M: yeah , sure . just let me know when you’re free.
H: after lunch is probably the best time
M: okay, sounds good . just let me know when you’re free.
H: would 2 pm work for you?
M: works for me.
H: well let’s say 2 pm then i ‘ll see you there
M: sounds good.
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: some of the damage done
NeuroCon: Is that at your house?

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?
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”

World “Facts”
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**.

Material: [http://opendialogue.miulab.tw](http://opendialogue.miulab.tw)
Knowledge-Grounded Responses  (Ghazvininejad et al., 2017)

A: Visiting the celebs at Los Angeles International Airport (LAX) - [...] w/ 70 others
B: Nice airport terminal. Have a safe flight.

A: Is that [...] in your photos? It’s on my list of places to visit in NYC.
B: Don’t forget to check out the 5th floor, while you are here, it’s a great view.

A: Live right now on [...] Tune in!!!!!
B: Listen to Lisa Paige

A: Been craving Chicken Pot Pie-who has the best? Trying [...] at [...] Must be Change of weather!
B: Love the pasta trattoria.

A: So [...] is down to one copy of Pound Foolish. I’m curious to see if they are re-ordering it.
B: Check out the video feed on 6 and take a picture of the Simpsons on the 3rd floor.

A: I wish [...] would introduce another vegetarian option besides the shroomburger. It’s delicious but kind of ridiculous.
B: This is the best j.crew in the world. Try the lemonade!

A: Just had an awesome dinner at [...] Great recommendation [...] 
B: One of my favorite places I’ve ever been to in NYC. The food is great and the service is lackluster.

Results (23M conversations) outperforms competitive neural baseline (human + automatic eval)
Evolution Roadmap

- **Empathetic systems**
  - I feel sad...

- **Common sense system**
  - What is influenza?
  - Tell me a joke.

- **Knowledge based system**
  - I’ve got a cold what do I do?
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

```
Schedule a lunch with Vivian.
```

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

- Vision
- Speech
- Text

Zara - The Empathetic Supergirl

Material: [http://opendialogue.miulab.tw](http://opendialogue.miulab.tw)
Cognitive Behavioral Therapy (CBT)

- Mood Tracking
- Pattern Mining
- Depression Reduction
- Daily lessons and check-ins
- Quick conversations to feel better
- Content Providing
- Always Be There
- Know You Well
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 requires 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