S9276: Towards Open-Domain Conversational AI

YUN-NUNG (VIVIAN) CHEN

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

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

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

- Global Digital Statistics (2018 January)

Total Population: 7.59B
Internet Users: 4.02B
Active Social Media Users: 3.20B
Unique Mobile Users: 5.14B
Active Mobile Social Users: 2.96B

The more **natural** and **convenient** input of devices evolves towards **speech**.
Why and When We Need?

“"I want to chat”
Turing Test (talk like a human)
Social Chit-Chat

“I have a question”
Information consumption

“I need to get this done”
Task completion
Task-Oriented Dialogues

“What should I do?”
Decision support

- What is today’s agenda?
- What does GTC stand for?

- Book me the flight ticket from Taipei to San Francisco
- Reserve a table at Din Tai Fung for 5 people, 7PM tonight

- Is GTC good to attend?
Intelligent Assistants

Task-Oriented
Conversational Agents

Chit-Chat

Task-Oriented

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

- 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 System (Young, 2000)

Speech Recognition

Speech Signal

Hypothesis

are there any action movies to see this weekend

Text Input

Are there any action movies to see this weekend?

Natural Language Generation (NLG)

Text response

Where are you located?

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

Backend Action / Knowledge Providers
Task-Oriented Dialogue System (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)

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

Text response
Where are you located?

System Action/Policy
request_location

Backend Action / Knowledge Providers
Semantic Frame Representation

• Requires a domain ontology: early connection to backend
• Contains core content (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><strong>09:25</strong></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 System (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)

Text response
Where are you located?

System Action/Policy
request_location

Dialog 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
1. Domain Identification
Requires Predefined Domain Ontology

User
find a good eating place for Taiwanese food

Intelligent Agent
Organized Domain Knowledge (Database)

Restaurant DB
Taxi DB
Movie DB

Classification!
2. Intent Detection
Requires Predefined Schema

User: find a good eating place for Taiwanese food

Intelligent Agent

Restaurant DB:
- FIND_RESTAURANT
- FIND_PRICE
- FIND_TYPE

Classification!
3. Slot Filling
Requires Predefined Schema

User:
find a **good** eating place for **taiwanese** food

Restaurant DB:

<table>
<thead>
<tr>
<th>Restaurant</th>
<th>Rating</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rest 1</td>
<td>good</td>
<td>Taiwanese</td>
</tr>
<tr>
<td>Rest 2</td>
<td>bad</td>
<td>Thai</td>
</tr>
</tbody>
</table>

Semantic Frame:

```
FIND_RESTAURANT
rating="good"
type="taiwanese"
```

Intelligent Agent:

```
SELECT restaurant {
rest.rating="good"
rest.type="taiwanese"
}
```

Sequence Labeling:

```
O O B-rating O O O B-type O
```
Slot Tagging (Yao+, 2013; Mesnil+, 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

Slot Tagging (Kurata+, 2016; Simonnet+, 2015)

- Encoder-decoder networks
  - Leverages sentence level information

- Attention-based encoder-decoder
  - Use of attention (as in MT) in the encoder-decoder network
  - Attention is estimated using a feed-forward network with input: $h_t$ and $s_t$ at time $t$
Joint Semantic Frame Parsing

- Slot filling and intent prediction in the same output sequence

Sequence-based (Hakkani-Tur et al., 2016)

- Intent prediction and slot filling are performed in two branches

Parallel (Liu and Lane, 2016)

Slot Filling

- taiwanese
- food
- please

Intent Prediction

- EOS
- B-type
- V
- O

- FIND_REST
## 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 \((\text{Goo}+, 2018)\)

Slot Gate

\[ g = \sum v \cdot \tanh(c_i^S + W \cdot c^I) \]

Slot Prediction

\[ y_i^S = \text{softmax}(W^S(h_i + g \cdot c_i^S) + b^S) \]

\(g\) will be larger if slot and intent are better related
Contextual LU

Domain Identification $\rightarrow$ Intent Prediction $\rightarrow$ Slot Filling

$D$ communication $I$ send_email

$U$ just sent email to bob about fishing this weekend
$S$ B-contact_name B-subject l-subject l-subject
$\rightarrow$ send_email(contact_name="bob", subject="fishing this weekend")

$U_1$ send email to bob
$S_1$ B-contact_name
$\rightarrow$ 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
$\rightarrow$ send_email(message="are we going to fish this weekend")
Contextual LU

• User utterances are highly ambiguous in isolation

Restaurant Booking

Book a table for 10 people tonight.

Which restaurant would you like to book a table for?

Cascal, for 6.

#people  time
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”
Time-Aware & Role-Based Attention (Su+, 2018)

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

Tourist
$\begin{array}{c} u_2 \rightarrow u_1 \rightarrow u_4 \rightarrow u_3 \rightarrow u_5 \rightarrow u_6 \rightarrow u_7 \end{array}$

Guide

Current

History Summary
$\begin{array}{c} \alpha_{u_2} \cdot u_2 \cdot \alpha_{u_4} \cdot u_4 \cdot \alpha_{u_5} \cdot u_5 \end{array}$

Role-Level Time-Decay Attention
$\begin{array}{c} \alpha_{\tau_1} \cdot u_1 \cdot \alpha_{\tau_2} \cdot u_3 \cdot \alpha_{\tau_2} \cdot u_6 \cdot u_6 \end{array}$

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

Convex, Linear, Concave

Dense Layer

Spoken Language Understanding

... $w_t \rightarrow w_{t+1} \rightarrow \ldots \rightarrow w_T$
Task-Oriented Dialogue System (Young, 2000)

**Speech Recognition**
- Speech Signal
- 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

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

**Backend Action / Knowledge Providers**
Dialogue State Tracking

Hello, how may I help you?

I’m looking for a Thai restaurant.
request (restaurant; foodtype=Thai)

What part of town do you have in mind?

Something in the centre.
inform (area=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.
request (address)

Thank you, bye.
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)
inform(area=centre)
request(address)

feature extraction

DNN

state of this turn

A slot value distribution for each slot

multi-turn conversation
RNN-CNN DST (Mrkšić+, 2015)

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

Hello, how may I help you?

request (restaurant; foodtype=Thai)

I’m looking for a Thai restaurant.

request (area)

What part of town do you have in mind?

inform (area=centre)

Something in the centre.

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

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

inform (address=24 Green street)

What’s the address?

request (address)

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

bye ()

Thank you, bye.
Supervised v.s. Reinforcement

- Supervised
  - Learning from teacher
    - "Hello"
    - "Bye bye"
  - Say "Hi"
  - Say "Good bye"

- Reinforcement
  - Learning from critics
    - Hello 😊
    - ......
  - Bad
    - OXX???!
Dialogue Policy Optimization

- Dialogue management in a RL framework

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

Idea: SL for each component and RL for end-to-end training

Are there any action movies to see this weekend?

User Goal
User Model
User Simulation

Natural Language
Dialogue Policy

Database

Neural Dialogue System

User Goal
User Model
User Simulation

Dialogue Policy Learning

Dialogue Policy

Request_location
Reinforcement Learning Agent (Li+, 2017)

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 start time 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+, 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 **7:15 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
Task-Oriented Dialogue System (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

System Action/Policy
request_location

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

Backend Action / Knowledge Providers

Text response
Where are you located?

Natural Language Generation (NLG)
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 NL

<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, poor scalability
RNN-Based LM NLG (Wen+, 2015)

- Dialog act 1-hot representation
- Delexicalisation
- Slot weight tying conditioned on the dialogue act

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

Output: Din Tai Fung serves Taiwanese
Handling Semantic Repetition

- 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**.
- Deficiency in either model or decoding (or both)
- Mitigation
  - Post-processing rules (Oh & Rudnicky, 2000)
  - Gating mechanism (Wen et al., 2015)
  - Attention (Mei et al., 2016; Wen et al., 2015)
Semantic Conditioned LSTM (Wen+, 2015)

- Original LSTM cell

\[ i_t = \sigma(W_{wi}x_t + W_{hi}h_{t-1}) \]
\[ f_t = \sigma(W_{wf}x_t + W_{hf}h_{t-1}) \]
\[ o_t = \sigma(W_{wo}x_t + W_{ho}h_{t-1}) \]
\[ \tilde{c}_t = \tanh(W_{wc}x_t + W_{hc}h_{t-1}) \]
\[ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \]
\[ h_t = o_t \odot \tanh(c_t) \]

- Dialogue act (DA) cell

\[ r_t = \sigma(W_{wr}x_t + W_{hr}h_{t-1}) \]
\[ d_t = r_t \odot d_{t-1} \]

- Modify C_t

\[ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t + \tanh(W_{dc}d_t) \]

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+, 2018)

Idea: gradually generate words based on the linguistic knowledge

**Bidirectional GRU Encoder**
- Name
- Italian
- Price Range

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

```
Input: name[Midsummer House], food[Italian], priceRange[moderate], near[All Bar One]
```

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

```
ENCODER: h_{enc}
```

```
DECODING LAYER1: 1. NOUN + PROP + 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
```

```
All Bar One is moderately priced Italian place it is called Midsummer House
```

```
All Bar One is a moderately priced Italian place it is called Midsummer House
```

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

- Single domain systems
- Extended systems
- Multi-domain systems
- Open domain systems

Dialogue breadth (coverage)

Dialogue depth (complexity)

What is influenza?

I've got a cold what do I do?

Tell me a joke.

I feel sad...
Dialogue Systems

Task-Oriented Dialogue

Fully Data-Driven

Input: $x$

Output: $y$

Understanding (NLU)

State tracker

Generation (NLG)

Dialog policy

Database Memory

External knowledge

Statistical model (e.g., neural)
Chit-Chat Social Bots | Non-Task-Oriented
Neural Response Generation (Sordoni+, 2015; Vinyals & Le, 2015)

... because of your game?

Source: conversation history

encoder

decoder

\[ \hat{T} = \arg \max_T \{ \log p(T|S') \} \]

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, I'll let you know!

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.
Issue 2: Response Inconsistency

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

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

where do you live in england.

Rob

Word embeddings (50k)

Speaker embeddings (70k)

where do you live

Rob

in england

Rob_712

Dreamwalls

Kush_322

D_Gomes25

Tomcoatez

skinnyoflynny2

The_Football_Bar

This_Is_Artful

DigitalDan285

Jinnmeow3

Bob_Kelly2

Rob

in

england

. 

EOS
Persona Model for Speaker Consistency (Li+, 2016)

Baseline model → inconsistency

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

Persona model using speaker embedding → consistency

<table>
<thead>
<tr>
<th>message</th>
<th>Where is your hometown?</th>
</tr>
</thead>
<tbody>
<tr>
<td>response</td>
<td>I’m from <strong>England</strong>.</td>
</tr>
<tr>
<td>message</td>
<td>Where are you from?</td>
</tr>
<tr>
<td>response</td>
<td>I’m from <strong>England</strong>.</td>
</tr>
<tr>
<td>message</td>
<td>In which city do you live now?</td>
</tr>
<tr>
<td>response</td>
<td>I live in <strong>London</strong>.</td>
</tr>
<tr>
<td>message</td>
<td>In which country do you live now?</td>
</tr>
<tr>
<td>response</td>
<td>I live in <strong>England</strong>.</td>
</tr>
<tr>
<td>message</td>
<td>What is your major?</td>
</tr>
<tr>
<td>response</td>
<td><strong>Business</strong>, you?</td>
</tr>
<tr>
<td>message</td>
<td>What did you study in college?</td>
</tr>
<tr>
<td>response</td>
<td>I did <strong>business</strong> studies.</td>
</tr>
<tr>
<td>message</td>
<td>How old are you?</td>
</tr>
<tr>
<td>response</td>
<td>I’m <strong>18</strong>.</td>
</tr>
<tr>
<td>message</td>
<td>What is your age?</td>
</tr>
<tr>
<td>response</td>
<td>I’m <strong>18</strong>.</td>
</tr>
</tbody>
</table>
### Issue 3: Dialogue-Level Optimization via RL

**Application**
- **Task Completion Bots** (Movies, Restaurants, ...)
- **Info Bots** (Q&A bot over KB, Web etc.)
- **Social Bot** (Xiaolce)

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

**Diagram:**
- User input $(o)$
- Language understanding
- Language (response) generation
- **Dialogue Manager**
  - $a = \pi(s)$
  - Collect rewards $(s, a, r, s')$
  - Optimize $Q(s, a)$

**Equation:**
$$a = \pi(s)$$

**Collect rewards** $(s, a, r, s')$
Deep RL for Response Generation (Li+, 2016)

- 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+, 2015; Li+, 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!
Going to Kusakabe tonight

Conversation History

Dialogue Encoder

Σ

Decoder

Fact Encoder

Try omakase, the best in town

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)

Dialogue depth (complexity)

Empathetic systems

I feel sad...

I’ve got a cold what do I do?

Common sense system

Tell me a joke.

What is influenza?

Knowledge based system

What is influenza?
Common Sense for Dialogue Planning (Sun+, 2016)

• High-level intention may span several domains

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

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

Emotion Recognizer

- Vision
- Speech
- Text

Zara - The Empathetic Supergirl

[Image of Zara with emotion recognition interface]

Race recognition output:
```
{  "recognition": "Race: Asian Confidence: 65.42750000000001 Smiling: 3.95896 Gender: Female Confidence: 88.9369" }
```

Made with love by Ice Technologies in collaboration with Hong Kong University of Science and Technology
Cognitive Behavioral Therapy (CBT)

Pattern Mining
- Daily lessons and check-ins

Mood Tracking
- Emoticons

Content Providing
- News
- Always Be There

Depression Reduction
- Character

Know You Well
- Magnifying glass
Summarized Challenges

- 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
- Data collection and analysis from un-structured data
- Complex-cascade systems requires high accuracy for working good as a whole
Framework & Resources

• MiuLab codes are available here: https://github.com/MiuLab/

• Frameworks
  • Tensorflow, PyTorch

• Resources
  • NVIDIA GTX 1070
Her (2013)

What can machines achieve now or in the future?
Q & A

Thanks for Your Attention!

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