General Dialogue Topics

NAIST
Assistant Professor
Koichiro Yoshino
Conventional dialogue systems

- The user utterance is understood as an actual dialogue state
  - By considering history
- The system action is decided according to the dialogue state and its confidence
  - Generate an utterance according to the decided system action

The user says:

I'd like to take Kintetsu-line from Ikoma

System response:

1 ask $TO\_GO$ = ???
2 inform $NEXT$

Diagram:
- $FROM=Ikoma$
- $LINE=Kintetsu$
- $FROM=Ikoma$
- $TO\_GO=???
- $LINE=Kintetsu$
What NCM does

Give-up understanding and managing

Where do you want to go?

Language generation

Knowledge base, model

Dialogue management

$FROM=Ikoma
$LINE=Kintetsu

1 ask $TO_GO
2 inform $NEXT
...

I’d like to take Kintetsu-line from Ikoma

$FROM=Ikoma
$LINE=Kintetsu

Where do you want to go?

Give-up understanding and managing

Language generation

Knowledge base, model

Dialogue management

1 ask $TO_GO
2 inform $NEXT
...

I’d like to take Kintetsu-line from Ikoma

$FROM=Ikoma
$LINE=Kintetsu

Where do you want to go?
Speech recognition with DNN in early stage

- Conventional ASR architecture

\[
\arg\max_W P(W|X) = \arg\max_W P(X|W)P(W)
\]

\(W\) is word sequence and \(X\) is speech

GMM-HMM

DNN-HMM

Acoustic model

Language model
What is output of the ASR?

• **N-best hypotheses of speech recognition results**
  - with posterior probabilities, which is calculated from likelihoods of acoustic model and language model
    
    | Probability | Hypothesis                              |
    |-------------|-----------------------------------------|
    | 0.7         | I want to take a flight to Austin       |
    | 0.2         | I want to take a flight to Boston       |
    | 0.05        | I want to take applied to Austin        |
    |             | ...                                     |

• **ASR results often contain errors**
  - Insertion, deletion, replacement, ...

• **We have to consider the error in post processes (SLU, DM, ...)**
Spoken language understanding (SLU) and dialogue management (DM)

- **Language understanding**
  - Convert the user utterance into machine-readable expressions
  
  I want to take Kintetsu from Ikoma

- **Dialogue management**
  - Decide the next system action from the SLU result and dialogue history

**SLU result**

Train_info{$FROM=Ikoma,$LINE=Kintetsu}

**Dialogue state**

Train_info{
  $FROM=Ikoma
  $TO_GO=Namba
  $LINE=Kintetsu
}

**Action decision**

1 inform $NEXT_TRAIN
2 ask $TO_GO
...
Dialogue state tracking and action decision

- As mentioned before, **ASR results often contain errors**
  - SLU results are probably affected by the ASR error
  - SLU module also causes error

**Dialogue state tracking**

I'd like to go to Baba with Kintetsu

*history*

| $FROM=???, $TO_GO=Namba, $LINE=???, $FROM=???, $TO.GO=Baba, $LINE=Kintetsu |

**Action decision**

1. inform $NEXT_TRAIN
2. ask $TO_GO

The system need to select "ask $TO.GO" or "confirmation" action if the recognition result may contain critical errors
Language generation systems

• **Generate a sentence given a system action**

  ![Diagram](image)

  Ask $TO\_GO \rightarrow \text{Language generation} \rightarrow \text{Where will you go?}

• **Difficulty of generation**
  
  – **Appropriateness**: Outputs contain the contents that is decided by the dialogue manager
  
  – **Naturalness**: Outputs is natural
  
  – **Understandability**: Outputs should be easy to understand
  
  – **Variation**: Outputs contain some variations of expression
Problems in existing systems

- **Turn-taking is not natural**
  - Based on voice activity detection (VAD)

- **Need to define ontology**
  - Handcrafting for any new domains

- **Dialogue strategy in a new space**
  - RL-based optimization can be used if we define states and actions

- **Controllable neural language generation**
  - Only using cross-entropy loss
Our approaches

• **Turn-taking is not natural**
  – Based on understanding results of the system

• **Need to define ontology**
  – Design of language understanding space

• **Dialogue strategy in a new space**
  – Information seeking for argumentation dialogue

• **Controllable neural language generation**
  – Use seqGAN and label aware objective
Incremental understanding system

- Incremental system that receives a word on each time-step
  - Multi-layer perceptron classifiers given the hidden layer of LSTM
  - Cambridge restaurant navigation system (DSTC2)
Re-labeling

- Make a training data of turn taking by comparing DST results on any time steps with the last result

Any differences?

- Yes $\rightarrow$ the system still need to wait future words
- No $\rightarrow$ the system can take a turn at this moment!
Incremental turn taking decider

- Comparing NLU results on between current input and the point the utterance ends
  - No difference: 0 / Different: 1 $\rightarrow$ supervised learning
  - System can take a turn if “0” is predicted
Results on DSTC2 dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LecTrack [9]</td>
<td>0.63</td>
<td>0.74</td>
<td>0.90</td>
<td>0.19</td>
<td>0.96</td>
<td>0.08</td>
<td>0.62</td>
<td>0.75</td>
<td>0.92</td>
<td>0.15</td>
<td>0.96</td>
<td>0.07</td>
</tr>
<tr>
<td>iDST.ASR (r = 1.0)</td>
<td>0.64</td>
<td>0.53</td>
<td>0.90</td>
<td>0.17</td>
<td>0.96</td>
<td>0.07</td>
<td>0.63</td>
<td>0.56</td>
<td>0.92</td>
<td>0.13</td>
<td>0.97</td>
<td>0.06</td>
</tr>
<tr>
<td>iDST.TRA (r = 1.0)</td>
<td>0.87</td>
<td>0.23</td>
<td>0.94</td>
<td>0.10</td>
<td>0.99</td>
<td>0.02</td>
<td>0.82</td>
<td>0.30</td>
<td>0.94</td>
<td>0.09</td>
<td>0.99</td>
<td>0.02</td>
</tr>
<tr>
<td>iDST.ASR (r = 0.6)</td>
<td>0.57</td>
<td>0.61</td>
<td><strong>0.89</strong></td>
<td>0.18</td>
<td>0.86</td>
<td>0.23</td>
<td>0.56</td>
<td>0.62</td>
<td>0.91</td>
<td>0.14</td>
<td>0.86</td>
<td>0.21</td>
</tr>
<tr>
<td>iTTD.ASR (d = 0.85)</td>
<td><strong>0.59</strong></td>
<td>0.60</td>
<td>0.88</td>
<td>0.19</td>
<td><strong>0.91</strong></td>
<td>0.16</td>
<td><strong>0.58</strong></td>
<td>0.61</td>
<td>0.91</td>
<td>0.15</td>
<td>0.91</td>
<td>0.15</td>
</tr>
<tr>
<td>iDST.TRA (r = 0.6)</td>
<td>0.77</td>
<td>0.34</td>
<td><strong>0.93</strong></td>
<td>0.11</td>
<td>0.88</td>
<td>0.18</td>
<td>0.73</td>
<td>0.39</td>
<td><strong>0.94</strong></td>
<td>0.10</td>
<td>0.88</td>
<td>0.18</td>
</tr>
<tr>
<td>iTTD.TRA (d = 0.85)</td>
<td><strong>0.80</strong></td>
<td>0.31</td>
<td>0.92</td>
<td>0.12</td>
<td><strong>0.91</strong></td>
<td>0.15</td>
<td><strong>0.76</strong></td>
<td>0.37</td>
<td>0.93</td>
<td>0.11</td>
<td><strong>0.91</strong></td>
<td>0.15</td>
</tr>
</tbody>
</table>

- DST accuracy itself was improved by the incremental process
  - 80-97% for each slots, if we use transcription
- r=0.6: the system interrupt at 60% utterance
- d=0.85: the system interrupt if iTTD conf. is bigger than 0.85
- Comparable scores to waiting any words by users
Analysis

- Adaptive turn taking can manage both NLU accuracy and interrupting
On-going project: Language understanding based on events

• Ontology-based NLU space requires handcrafting to define
  – Each domain requires own ontology
  – Generation also requires handcrafting

• Idea: using event (P-A) as understanding space
  – Will work for any domain that parsers can work
  – Coverage is limited
    • Difference between “go to see” and “visit to see”

Frame (slot-value)
- Act: Request
- Type: Chinese restaurant
- Price_range: don’t care
- Count: 2
- Kids_allowed: NULL

Event (predicate-argument)
- Cherry blossom
- Go to see
- Next week
- Speaker
New dialogue domain: argumentation

• Argumentation
  “he drove a car” and “alcohol was detected from his breath”; thus “he did drunk driving”

• Claim:
  – he did drunk driving

• Supportive facts:
  – he drove a car
  – alcohol was detected from his breath
Information seeking for argumentation

- **Collecting supportive facts through a dialogue**
  - Did he drive a car?/
  - Was any alcohol detected from his bless?

  "he drove a car" and "alcohol was detected from his bless"; thus "he did drunk driving"

- **A lot of possible questions**
  - Policy is trained to decide
    - "which action will the system ask on which situation"

Presented at AAAI2019 DeepDIAL-WS
He did drunk driving, because alcohol was detected from his breath.

Collected facts:
- Alcohol was detected from his breath.

Claim:
- He drove a car.

Relevance: 0.1

Answered by weighted abduction [Ovchinnikova+, 14]
Information seeking based on Markov decision process

He did drunk driving, because alcohol was detected from his breath

- Alcohol was detected from his breath

Collected facts

claim

Did he drove a car?

Yes

- He drove a car

What he know

answerer

Reitionality 0.1

• He drove a car
Information seeking based on Markov decision process

He did drunk driving, because alcohol was detected from his breath

Reitionality 0.1

Did he drove a car?

Yes

• He drove a car

What he know

Raitionality 1.0

reward: \( r(s_t, a_t, s_{t+1}) \)

He did drunk driving, because alcohol was detected from his breath, and he drove a car

Collected facts

Claim

system

He did drunk driving, because alcohol was detected from his breath

• Alcohol was detected from his breath

system

• He drove a car

He drove a car

What he know

Answerer

system

• Alcohol was detected from his breath

• He drove a car

system
Strategy optimized by deep reinforcement learning

Comparison of results with different time limit in Legal Discussion Domain

Success of argumentation

Dialogue turns

- DDQN
- DFS
- BFS

0 10 20 30 40 50
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Dialogue example

- Dialogue on twenty angry men dataset
  - System tries to construct a rational argument to claim “the boy is not guilty”
  - System can collect facts from the dialogue partner

<table>
<thead>
<tr>
<th>Step</th>
<th>Speaker</th>
<th>Utterance</th>
<th>Rationality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sys</td>
<td>Did not the woman across the street see the murder committed?</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Ans</td>
<td><strong>The woman across the street did not see the murder committed.</strong></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Sys</td>
<td>Did not the old man hear the kid yell, “I’m gonna kill you”?</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Ans</td>
<td><strong>The old man did not hear the kid yell, “I’m gonna kill you”.</strong></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Sys</td>
<td>Is the old man a liar?</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Ans</td>
<td>I do not know.</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Sys</td>
<td>Did not the boy crouch down and stab a taller man in the chest?</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Ans</td>
<td>I do not know.</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Sys</td>
<td>Did not the woman across the street see the boy stab his father?</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Ans</td>
<td>I do not know.</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Sys</td>
<td>Did not the boy admit buying the knife?</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>Ans</td>
<td><strong>The boy did not admit buying the knife.</strong></td>
<td></td>
</tr>
</tbody>
</table>
Language generation for dialogue systems

• Contents to be contained in the generation results are decided by dialogue manager

![Diagram]

• There are some works to generate sentences given an action
SC-LSTM by Wen et al., 2015

- recurrent hidden layer
- embedding of a word
- 1-hot dialog act and slot values

How to say (=LM)

What to say (contents)
Sample generations by SC-LSTM

Figure 3: Examples showing how the SC-LSTM controls the DA features flowing into the network via
Context-aware NLG

- Sequence-to-sequence modeling of generation
  - Change the response according to the dialogue context

```
<table>
<thead>
<tr>
<th>Prepending context</th>
<th>DA encoder</th>
<th>Context encoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>is there a schedule</td>
<td>for NUMBER AMPM</td>
<td>inform_no_match</td>
</tr>
<tr>
<td>inform(line=M102, direction=Herald Square, vehicle=bus, departure_time=9:01am, from_stop=Wall Street)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Take bus line M102 from Wall Street to Herald Square at 9:01am

There is a bus at 9:01am from Wall Street to Herald Square using line M102.
```

preceding user utterance
is there another option
context-aware additions

তypical NLG

contextually bound response
What the problem?

- Existing generation systems are trained by softmax-cross entropy-loss to words
  - No guarantee to contain given information by system action

**Input**

Candidates: 2
Area: Düsseldorf
Pets: allow

**Generation**

There are 2 hotels that allow pets?

**Training data**

There are 2 hotels in Düsseldorf that allows pets
Controlling generation results with condition

- We built a generation system based on
  - generative adversarial network (Seq-GAN) and
  - label-aware objective
- We only controlled by dialogue acts of the system
  - The system itself is NCM

Diagram:
- User
- 1) Language Understanding
  DA: <Question>
  Whether=?
  Time=Question
- 2) Dialogue Management
  DA: <Answer>
  Whether=Sunny
  Time=Tomorrow
- 3) Response Generation
- Dialogue system

Questions:
- what's the forecast like tomorrow?
- Tomorrow is clear skies!!
Generation based on SeqGAN

- Sequential generative adversarial network is a technique to evaluate whole of sentence (not word-by-word)

- Discriminator predicts two classes (real/fake)

- Generated receives reward to the generation sequence
  - Reinforcement learning is used (n-step delayed reward)
Cross-entropy loss and SeqGAN

• Generation model based on softmax-cross entropy calculates loss for each word

\[ w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5 \quad w_6 \quad w_7 \quad w_8 \]

\[ \bigcirc \quad \bigcirc \quad \times \quad \bigcirc \quad \bigcirc \quad \times \quad \bigcirc \quad \bigcirc \]

• SeqGAN only calculate feedback at the end of sequence

\[ w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5 \quad w_6 \quad w_7 \quad w_8 \]

real/fake
Label aware objective

- SeqGAN only distinguish real or fake
  - We extended the discriminator to multi-class classification to know the generation result is based on input or not
Both naturalness and controllability were improved

- Explicit penalty to the condition improved the controllability
- Adversarial learning improved naturalness
Summary

- We introduced basic architecture of spoken dialogue systems, and tackled several problems of existing systems
  - Turn-taking is not natural
  - Need to define ontology
  - Dialogue strategy in a new space
  - Controllable neural language generation