Spoken Dialogue System (SDS) for a Human-like Conversational Robot

ERICA

Tatsuya Kawahara
(Kyoto University, Japan)
Limitation of Current (deployed) SDS

- Machine-oriented constrained dialogue
  - Think over what system can [conceptual constraint]
  - Utter one simple sentence [linguistic constraint]
  - with clear articulation [acoustic constraint]
  - and wait for response [reactive model]

- Big gap from human (or ideal) dialogue
  - Human tourist guide, Concierge at hotels
Human-Machine Interface *(Current SDS)*

- Constrained speech/dialog
  - Half duplex and reactive
  - One sentence per one turn
  - System responds only when user asks

People are aware they are talking to a machine.

Human-Human Communication

- Natural speech/dialog
  - Duplex and interactive
  - Many sentences per one turn
  - Backchannels

Human is the most natural interface! ➔ Human-like Robot
Android ERICA Project started in 2016

http://sap.ist.i.kyoto-u.ac.jp/erato/
JST ERATO Symbiotic Human-Robot Interaction Project (2014-2020)

- **Goal:** Autonomous android who behaves and interacts just like a human
  - Facial look and expression
  - Gaze and gesture
  - Natural spoken dialogue
- **Criterion:** Total Turing Test
  - Convince people it is comparable to human, or indistinguishable from remote-operated android
- **Science:**
  - Clarify what is missing or critical in natural interaction
- **Engineering Applications:**
  - Replace social roles done by human（感情労働）
  - Conversation skill training
Android ERICA
with flowers with microphones & camera
Tasks of ERICA

× Information services $\rightarrow$ smart phones

× Move objects $\rightarrow$ conventional robots
  × ERICA cannot move except for gestures

× Chatting $\rightarrow$ ChatBot
  × Should involve physical presence and non-verbal communication

• Social Interaction
Social Roles of ERICA

- Counseling
- Interview
- Receptionist, Secretary
- Guide, Companion
- Newscaster

Role of Listening

Role of Talking (to)

Shallow and short interaction

- One person
- Several persons
- Many people
Research Topics

Robust Speech Recognition (ASR) → Flexible Dialogue

(1) Front-end (hands-free input) → (2) Back-end (spontaneous speech model) → (3) Understanding and Generation → (4) Turn-taking & Backchannel → (6) Interaction corpus

(5) Speech Synthesis

Machine learning & evaluation
Challenge in Speech Recognition

Conversational Speaking-style

Query/command (one-sentence)

Lecture & Meeting
   - Parliament
   - Video lecture

93%
90%

Humanoid Robot
   - Close-talk 82%
   - Gun-mic 72%
   - Distant 66%

Smartphone
   - Voice search
   - Apple Siri

90%

Home appliance
   - Amazon Echo
   - Google Home

90%

Close-talking Input Distant
Real Problem in Distant Talking

• When people speak without microphone, speaking style becomes so casual that it is not easy to detect utterance units.
  – Not addressed in conventional “challenges”
  – Circumvented in conventional products
    • Smartphones: push-to-talk
    • Smart speakers: magic word “Alexa”, “OK Google”
    • Pepper: talk when flash
Latency is Critical for Human-like Conversation

• Turn-switch interval in human dialogue
  – Average ~500msec
  – 700msec is too late
  → difficult for smooth conversation (cf.) oversea phone

• Cloud-based ASR cannot meet requirement

• Recent End-to-End (acoustic-to-word) ASR
  – 0.03xRT [ICASSP18]

• All downstream NLP modules must be tuned
Features in Speech Synthesis

• Very high quality
• Conversational style rather than text-reading
  – Questions (direct/indirect)
• A variety of non-lexical utterances with a variety of prosody
  – Backchannels
  – Fillers
  – Laughter
• http://voicetext.jp (ERICA)
Human-like Dialogue Features

• Hybrid Dialogue Structure
• Mixed-initiative
• Natural turn-taking
• Backchanneling
• Non-lexical utterances
• Non-verbal information (in spoken dialogue)
Hybrid of Different Dialogue Modules

• State-transition flow (hand-crafted)
  – Used in limited task domain
  – Deep interaction but works only in narrow domains
  – Cannot cope beyond the prepared scenario
• Question-Answering
  – Used in smartphone and smart speakers
  – Wide coverage but short interaction
  – Cannot cope beyond the prepared DB
• Statement-Response
  – Used in ChatBot
  – Wide coverage but shallow interaction
  – Many irrelevant OR only short formulaic responses
• Systems were not convincing and engaging!
• Dialogues were not realistic!
Real Problems in non-task-oriented SDS

• System often generates boring (safe) OR irrelevant (challenging) dialogue.

• Sensible adults (college students) hesitate to talk to robots.

• Attendants and Receptionists involve shallow interaction for easy task.
  – These robots are being deployed.
Our Solutions

- Realistic social role given to ERICA
- So matched users will be seriously engaged
- “Social interaction” task
  - Dialogue itself is task
    - Mutual understanding or appealing
  - (cf.) tasks solved via spoken dialogue
    - query or transaction
  - Not just chatting
  - Must be engaged by users as well as the robot
  - Face-to-face (physical presence) is important
Dialogue with Android ERICA in WOZ setting
Task 1: **Attentive Listening**

- ERICA mostly listens to senior people
  - Topics on memorable travels and recent activities
  - Encourages users to speak
Task 2: **Job Interview (Practice)**

- ERICA plays a role of interviewer
  - asks questions, which are answered by users
  - makes additional questions according to initial answers
  - provides a realistic simulation, or replace human
- Users need to appeal themselves
Task 3: **Speed Dating (Practice)**

- ERICA plays a role of female participant
  - asks questions to users AND answers questions by users on topics such as hobbies, favorite foods and music
  - provides a realistic simulation by not being too friendly
  - gives proper feedbacks according to the dialogue
- Users need to not only appeal but also listen

Relaxed, but somewhat nervous

Physical presence and face-to-face is important!
## Comparison of 3 Tasks

<table>
<thead>
<tr>
<th></th>
<th>Attentive Listening</th>
<th>Job interview</th>
<th>Speed Dating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue Initiative</td>
<td>User</td>
<td>System</td>
<td>Both (mixed)</td>
</tr>
<tr>
<td>Utterance mostly by</td>
<td>User</td>
<td>User</td>
<td>Both</td>
</tr>
<tr>
<td>Backchannel by</td>
<td>System</td>
<td>System</td>
<td>Both</td>
</tr>
<tr>
<td>Turn-switching</td>
<td>Rare</td>
<td>Clear</td>
<td>Complex</td>
</tr>
<tr>
<td># dialogue sessions</td>
<td>19</td>
<td>30</td>
<td>33</td>
</tr>
</tbody>
</table>
## Comparison of 3 Tasks

<table>
<thead>
<tr>
<th></th>
<th>Attentive Listening</th>
<th>Job interview</th>
<th>Speed Dating</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Utterance by User</td>
<td>64%</td>
<td>53%</td>
<td>49%</td>
</tr>
<tr>
<td>%Occurrence of system backchannel</td>
<td>38%</td>
<td>19%</td>
<td>19%</td>
</tr>
<tr>
<td>%Turn-switching</td>
<td>19%</td>
<td>30%</td>
<td>37%</td>
</tr>
<tr>
<td>Turn-switch time</td>
<td>454msec</td>
<td>629msec</td>
<td>548msec</td>
</tr>
</tbody>
</table>
Challenge: Total Turing Test

1. Can we generate same responses for a corpus collected via WOZ? [objective evaluation]
2. Can autonomous ERICA satisfy subjects in a same level as WOZ? [subjective evaluation]
Attentive Listening System
Attentive Listening

• People, esp. senior, want someone to listen.
• Talking by remembering is important for maintaining communication ability.

• System (robot), which listens and encourages the subject to talk more
  – Need to respond to anything
  – Does not require large knowledge base
  – Empathy and entrainment is important
Challenge: Total Turing Test of Attentive Listening System

• Can robot be a counselor?
  – Ishiguro thinks so
• Almost all senior subjects believed to be talking to ERICA during data collection in WOZ setting.

1. Can we generate same responses for a corpus collected via WOZ? [objective evaluation]
2. Can autonomous ERICA satisfy subjects in a same level as WOZ? [subjective evaluation]
Flow of Attentive Listening System

- Elaborating Question
- Partial Repeat
- Statement Assessment
- Formulaic Response
- Backchannel

- Speech recognition
- Focus detection
- Sentiment analysis
- Prosody

Response Selection
Elaborating Question and Partial Repeat based on Focus Word

- Detect a focus word
- Try to combine with WH phrases for a plausible question
  
  “I went to a conference.”

  - Which conference
  - When is conference
  - Where is conference

  “Which conference?” [Elaborating question]

- Or simply repeat the focus word
  
  “I went to Okinawa.”

  - Which Okinawa
  - Whose Okinawa
  - Okinawa, when?
  - Okinawa where?

  “Okinawa?” [Partial repeat]
Statement Assessment based on Sentiment Analysis

- Sentimental attribute annotated for each word
- Assessment selection based on (summed) attribute values

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective (fact)</strong></td>
<td>That’s nice</td>
<td>That’s bad</td>
</tr>
<tr>
<td></td>
<td>素敵ですね</td>
<td>大変ですね</td>
</tr>
<tr>
<td><strong>Subjective (comment)</strong></td>
<td>Wonderful</td>
<td>That’s a pity</td>
</tr>
<tr>
<td></td>
<td>いいですね</td>
<td>残念ですね</td>
</tr>
</tbody>
</table>

“I went a party.” → “That’s nice”
“But I was tired.” → “That’s a pity”
Formulaic Response

• Used as a back-off
  – “I see.”
  – “Really?”
  – “Isn’t it?”

• Function similar to backchannels
Flow of Attentive Listening System

- Speech recognition
  - Focus detection
  - Sentiment analysis
  - Prosody

- Flow of processes:
  - Elaborating Question
  - Partial Repeat
  - Statement Assessment
  - Formulaic Response
  - Backchannel

- Response Selection
Response Selection among Candidates

- There are many possible responses
- No ground truth (Even the corpus is not ground truth)

“Last Sunday, I went to a high-school reunion.”

<table>
<thead>
<tr>
<th>Formulaic response</th>
<th>“Really?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment</td>
<td>“That’s nice”</td>
</tr>
<tr>
<td>Partial repeat</td>
<td>“High-school reunion?”</td>
</tr>
<tr>
<td>Elaborating question</td>
<td>“Which reunion?”</td>
</tr>
</tbody>
</table>

Not selection problem, but validation problem (acceptable given linguistic & dialogue context?)
Response Selection among Candidates

- Many possible responses other than corpus occurrence
- Annotated acceptable responses

<table>
<thead>
<tr>
<th></th>
<th>Corpus occurrence</th>
<th>Acceptable ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formulaic response</td>
<td>45%</td>
<td>90%</td>
</tr>
<tr>
<td>Assessment</td>
<td>21%</td>
<td>60%</td>
</tr>
<tr>
<td>Partial repeat</td>
<td>22%</td>
<td>64%</td>
</tr>
<tr>
<td>Elaborating question</td>
<td>11%</td>
<td>28%</td>
</tr>
</tbody>
</table>

- Formulaic responses are mostly acceptable.
- Assessments and partial repeats are possible in a majority case.
Evaluation of Generated Responses

<table>
<thead>
<tr>
<th>Type</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formulaic response</td>
<td>99%</td>
<td>91%</td>
<td>0.95</td>
</tr>
<tr>
<td>Assessment</td>
<td>51%</td>
<td>73%</td>
<td>0.60</td>
</tr>
<tr>
<td>Partial repeat</td>
<td>68%</td>
<td>80%</td>
<td>0.74</td>
</tr>
<tr>
<td>Elaborating question</td>
<td>46%</td>
<td>41%</td>
<td>0.43</td>
</tr>
<tr>
<td>Weighted average</td>
<td>70%</td>
<td>73%</td>
<td>0.71</td>
</tr>
</tbody>
</table>

- Significantly better than the chance rate
- Still many irrelevant elaborating questions
Comparison with Standard Corpus-based Training

- Randomly generated according to distribution in the corpus
- Training with the corpus occurrence only
- Training with the enhanced annotation of acceptance
Challenge: **Total Turing Test** of Attentive Listening System

- Almost all senior subjects believed to be talking to ERICA during data collection in WOZ setting.

1. Can we generate same responses for a corpus collected via WOZ? **[objective evaluation]**

   → 70%

1. Can autonomous ERICA satisfy subjects in a same as WOZ? **[subjective evaluation]**

   2.1 Offline video/audio evaluation → ???

   2.2 Online system experience
(Preliminary) Subjective Offline Video/Audio Evaluation

• Video (Audio) prepared by replacing the operator’s voice with the system’s response
• Third-party subjects evaluated several questionnaire items, and compared against the baseline

• Overall evaluation is not good (around 0 [-3～3 scale])
  – Precision of 70% is not sufficient.
    • Irrelevant questions and assessments give bad impression.
  – Response is monotonous.
  – TTS and turn-taking is not natural enough?
  – No backchannels in this experiment!!
Conclusions & Practical Issues

• Considering arbitrary nature is important
• Enhanced annotation requires much effort
• Machine learning gives some improvement
• 70% in recall & precision
• But the system is not yet in a satisfactory level
Generation of Backchannels
Non-lexical utterances
--“Voice” beyond “Speech”--

• Continuer Backchannels: “うん”
  – listening, understanding, agreeing to the speaker
• Assessment Backchannels: “はー”、“ふーん”
  – Surprise, interest and empathy
• Fillers: “あのー”、“えーと”
  – Attention, politeness
• Laughter
  – Funny
  – Socializing
  – Self-pity
Backchannels (BC)

• Feedback for smooth communication
  – Indicate that the listener is listening, understanding, agreeing to the speaker
  – “right”, “はい”, “うん”

• Express listener’s reactions
  – Surprise, interest and empathy
  – “wow”, “あー”, “へー”

• Produce a sense of rhythm and feelings of synchrony, contingency and rapport
Factors in Backchannel Generation

• Timing *(when)*
  – Usually at the end of speaker’s utterances
  – Should predict before end-point detection

• Lexical form *(what)*
  – Machine learning using prosodic and linguistic features [Interspeech16]

• Prosody *(how)*
  – Adjust according to preceding user utterance [IWSDS15]

  – Many systems use same recorded pattern
  – giving monotonous impression to users
Categories and Occurrence Counts of Backchannels

<table>
<thead>
<tr>
<th>category</th>
<th>occurrence at IPU (clause) boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un うん</td>
<td>12% (10%)</td>
</tr>
<tr>
<td>Un x2 うんうん</td>
<td>7% ( 9%)</td>
</tr>
<tr>
<td>Un x3 うんうんうん</td>
<td>13% (19%)</td>
</tr>
<tr>
<td>Assessments</td>
<td>8% (14%)</td>
</tr>
<tr>
<td>None</td>
<td>60% (47%)</td>
</tr>
</tbody>
</table>

Backchannels are observed at 40% of IPUs with different forms in a good balance.
Additional Annotation of Backchannels

• Generation of backchannels and choice of their form are arbitrary
• Evaluation with exactly observed patterns may not be meaningful
  
  • Augment the annotation
    – Three human annotators judge which backchannel forms are acceptable, given dialogue context
    – Accept only when ALL three annotators agree
    – The added forms are regarded as correct in evaluation
Selection Problem ➔ Validation Problem

Not to generate

un

un x2

un x3

assessment

un?

un x2?

un x3?

assessment?

max>θ?

output

output

0.5

0.6

0.2

0.1

Not output
Prediction Performance by using Linguistic and Prosodic Features

<table>
<thead>
<tr>
<th>Category</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>un</td>
<td>0.311</td>
<td>0.657</td>
<td>0.422</td>
</tr>
<tr>
<td>un x2</td>
<td>0.382</td>
<td>0.820</td>
<td>0.521</td>
</tr>
<tr>
<td>un x3</td>
<td>0.672</td>
<td>0.333</td>
<td>0.454</td>
</tr>
<tr>
<td>assessments</td>
<td>0.467</td>
<td>0.342</td>
<td>0.405</td>
</tr>
<tr>
<td>not-to-generate</td>
<td>0.775</td>
<td>0.769</td>
<td>0.772</td>
</tr>
<tr>
<td>average</td>
<td>0.643</td>
<td>0.643</td>
<td>0.643</td>
</tr>
</tbody>
</table>

- Precision of simple continuers (un, un x2) is very high because they are acceptable in many cases.
- Reasonable performance for “not-to-generate” decision.
Subjective Offline Evaluation of Generated Backchannels

- Voice files of backchannels (one for each category) recorded by a voice actress (for TTS)
- Audio channel of counselors replaced by the generated backchannels

- 9 subjects listened 8 segments of dialogue, and evaluated on 7 items with 7-point scales
- Compared with
  - Weighted random generation
  - Counselors’ choice (voice replaced)
Subjective Evaluation of Backchannels

<table>
<thead>
<tr>
<th>Question</th>
<th>random</th>
<th>proposed</th>
<th>counselor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are backchannels natural?</td>
<td>-0.42</td>
<td>1.04</td>
<td>0.79</td>
</tr>
<tr>
<td>Are backchannels in good tempo?</td>
<td>0.25</td>
<td>1.29</td>
<td>1.00</td>
</tr>
<tr>
<td>Did the system understand well?</td>
<td>-0.13</td>
<td>1.17</td>
<td>0.79</td>
</tr>
<tr>
<td>Did the system show empathy?</td>
<td>0.13</td>
<td>1.04</td>
<td>0.46</td>
</tr>
<tr>
<td>Would like to talk to this system?</td>
<td>-0.33</td>
<td>0.96</td>
<td>0.29</td>
</tr>
</tbody>
</table>

- obtained higher rating than random generation
- even comparable to the counselor’s choice, though the scores are not sufficiently high
  - Same voice files are used for each backchannel form
  - Need to change the prosody as well
  - Tuning of precise timing is also needed
Challenge: **Total Turing Test of Backchanneling System**

1. Can we generate same responses for a corpus collected via WOZ? *[objective evaluation]*
   \[ \rightarrow 64\% \]

1. Can autonomous ERICA satisfy subjects in a same level as WOZ? *[subjective evaluation]*
   2.1 Offline video/audio evaluation
   \[ \rightarrow \bigcirc \text{ backchannel forms} \]
   \[ \rightarrow \times \text{ prosody & precise timing} \]

2.2 Online system experience
   \[ \rightarrow ??? \text{ (demo)} \]
Generating Fillers

• No filler

• Filler before moving to next question
Demonstration of Attentive Listening System
Current Lessons Learned

• Backchannels are effective, but proper precise timing is critical. (<200ms)

• Repeat of named entities is effective for showing understanding, but vulnerable to ASR errors.

• Proper assessment is expected at the end of talk, but often difficult.
  – People want to share their joy/sadness

• When above two works, dialogue is engaging.
Job Interview System
Job Interview

• Interview is an essential process in hiring persons and accepting (graduate) students
• Purpose
  – Check communication skill (when inclined to hire)
  – Find something special (when uncertain to hire)
• Face-to-face is norm
• Currently,
  – Students (and Companies) spend a lot in rehearsal and preparation
Challenge: Total Turing Test of Job Interview System

• Can robot be an interviewer?

• Some Japanese companies are introducing robots for interview in the initial stage
  – But mostly based on prepared question scenario
  – Interviewee can easily prepare (rehearse) well

1. Can we generate adaptive (non-scenario-based) questions? [corpus-based evaluation]
2. Can autonomous ERICA make subjects feel like real interview? [subjective evaluation]
Flow of Job Interview System

- Hand-crafted flow
- Optional questions
- Backchannel

Input processes:
- Speech recognition
- Focus detection
- Prosody
Current Implementation

• Flow of basic questions
  – Motivation for application
  – strong/weak points of the interviewee...

• Optional additional questions
  – “Why our company instead of other companies?”
  – “Can you tell me a specific example?”

• Selection of optional questions
  – Machine learning is difficult
  – Heuristics based on duration of turns
Demonstration of Job Interview System
Other Topics
Flexible Turn-taking

• Natural ⇐ push-to-talk, magic words
  – TRP predictor (pause / prosody)
• Fuzzy decision ⇐ Binary decision
  – Use fillers and backchannels when ambiguous
  – TTS output cannot be stopped

<table>
<thead>
<tr>
<th>User status</th>
<th>System action</th>
</tr>
</thead>
<tbody>
<tr>
<td>User definitely holds a turn</td>
<td>nothing</td>
</tr>
<tr>
<td>User maybe holds a turn</td>
<td>continuer backchannel</td>
</tr>
<tr>
<td>User maybe yields a turn</td>
<td>filler to take a turn</td>
</tr>
<tr>
<td>User definitely yields a turn</td>
<td>response</td>
</tr>
</tbody>
</table>
Non-verbal information

• Valence Recognition
  – Positive/negative feeling on what is talked about
  → proper assessment (including prosody) in attentive mode

• Engagement Recognition
  – Positive/negative attitude to keep the current dialogue
  → change topics, turn-taking behaviors,
    manner of system reply (including prosody)

• Ice-breaking
  – Rapport with the first-comer
  → Switch dialogue to main topic
Character Modeling ➔ Desire

- Attentive / Inattentive
- Extrovert / Introvert
- Polite / Casual

(cf.) Big Five

Myers-Briggs Type Indicator (MBTI)
Evaluation Criteria

• Total Turing Test (Level 1)
  – Comparable to WOZ setting

• Total Turing Test (Level 2)
  – Comparable to “human-like interaction experience”
  – measured by Engagement level
References


