The Design and Implementation of XiaoIce

An Empathetic Social Chatbot
Zhao, et al 2019 @Microsoft

*Presented by Adrian Suna Choi*
Machines converse with humans through natural languages
Engage in empathetic conversations
Have come a long way since 1960s

A Fundamental Challenge in AI – social chatbots
Timeline of Development Of Conversational Systems

1960s/70s
- **Eliza** (Weizenbaum, 1966)
- **Parry** (Colby, 1972)

2000s
- **Task Completion systems**
  
applied as in DARPA communication program

2009
- **Alice**

2010s
- **Intelligent Personal Assistants (IPAs)** such as Siri

2014
- **1st-gen Xiaolce**

**Data-driven ML approach**

*Excellent only within some domains*

**Challenges of IPAs?**
Xiaolce’s Uniqueness

• Early chatbots designed for chitchat
• Endowed with empathy, personality and skills
• Integration of EQ and IQ to optimize long-term user engagement
The Objective Addressed

AI companion

Long-term emotional connections formed with users

To satisfy human need for communication, affection and social belongs

To pass the time-sharing test
Design Principles

- IQ+EQ+personality
- CPS as social chat metric
- Social chat as hierarchical decision-making
System Architecture

Fig 1. Xiaolce system architecture.
Implementation of Conversation Engine Dialogue Manager

- Central Controller of the dialogue system

\[ a = \pi(s) \]

- Dialogue Policy
- Action – skill/Core Chat
- Current dialogue state s

Global State Tracker

Dialogue Policy
- Top level policy for skill selection
- Topic manager for Core Chat

Dialogue Manager
Implementation of Conversation Engine
Empathetic Computing

• Contextual Query Understanding
• User understanding
• Interpersonal response

\[ s = (Q_c, C, e_Q, e_R) \]
Fig 2. An example conversation session

... ... ...
11 [Human] You like Ashin that much, huh?
   [Xiaoice] Why not? You don’t like him?
12 [Human] I don’t like him, neither dislike him ➔ I don’t like Ashin, neither dislike Ashin
   [Xiaoice] I love every song of him
13 [Human] Really?
   [Xiaoice] Especially The Time Machine
14 [Human] Oh? Never heard of that ➔ Oh? Never heard of The Time Machine
   [Xiaoice] Shall I send you the song?
15 [Human] Okay ➔ Okay, send The Time Machine.
   [Xiaoice] Don’t know if you will like it or not” [Play the music card]

<table>
<thead>
<tr>
<th>Topic</th>
<th>music, Ashin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intent</td>
<td>request</td>
</tr>
<tr>
<td>Sentiment</td>
<td>sad, nervous ➔ neutral</td>
</tr>
<tr>
<td>Opinion</td>
<td>neutral</td>
</tr>
<tr>
<td>Age</td>
<td>25</td>
</tr>
<tr>
<td>Gender</td>
<td>male</td>
</tr>
<tr>
<td>Interests</td>
<td>singing, music ...</td>
</tr>
<tr>
<td>Occupation</td>
<td>salesman</td>
</tr>
<tr>
<td>Personality</td>
<td>quite, serious ...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic</th>
<th>music, Ashin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intent</td>
<td>Inform</td>
</tr>
<tr>
<td>Sentiment</td>
<td>happy</td>
</tr>
<tr>
<td>Opinion</td>
<td>positive</td>
</tr>
<tr>
<td>Age</td>
<td>18</td>
</tr>
<tr>
<td>Gender</td>
<td>female</td>
</tr>
<tr>
<td>Interests</td>
<td>music, games, movie, singing ...</td>
</tr>
<tr>
<td>Occupation</td>
<td>AI chatbot</td>
</tr>
<tr>
<td>Personality</td>
<td>reliable, sympathetic, affectionate...</td>
</tr>
</tbody>
</table>

... ...
... ...
... ...
... ...
Implementation of Conversation Engine
Core Chat

GENERAL CHAT & DOMAIN CHATS
CRUCIAL COMPONENT OF XIAOICE’S IQ AND EQ
RETRIEVAL-BASED GENERATOR USING PAIRED/UNPAIRED DATA
NEURAL RESPONSE GENERATOR
Fig 3. RNN-based neural response generator

\[ v = \sigma(W^T_Q e_Q + W^T_R e_R) \]

\[ u_t = \sigma(W^T_P [h_{t-1}; e_t; v]) \]

\[ z_t = \sigma(W^T_e [h_{t-1}; e_t; v]) \]

\[ l_t = \tanh(W^T_t [z_t \odot h_{t-1}; e_t; v]) \]

\[ h_t^Q = (1 - u_t) \odot h_{t-1} + u_t \odot l_t \]

\[ p(R|Q_c, e_Q, e_R) = \prod_{i=1}^{N_R} \frac{\exp(f(h_{t-1}, e_r, v))}{\sum_{r'} \exp(f(h_{t-1}, e_{r'}, v))} \]

\[ \arg\max_{\theta} \frac{1}{M} \sum_{i=1}^{M} \log p_{\theta}(R^{(i)}|Q^{(i)}_c, e^{(i)}_Q, e^{(i)}_R) \]

Fig 5. Examples to compare between consistent and inconsistent responses with seq2seq

[Human] How old are you?
[S2S-Bot] 16 and you?
[Human] What’s your age?
[S2S-Bot] 18.

[Human] How old are you?
[Xiaoice] I am 18, of course.
[Human] You age?
[Xiaolce] 18. Why?
[Human] Were you 18 last year?
[Xiaoice] I made a wish to stay 18 forever. Oh, my wish has come true.
<table>
<thead>
<tr>
<th>Query</th>
<th>Response</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>拿上脸不看 (Don’t want to look with my face in my hands.)</td>
<td>我边吃饭边看你 (I’m watching you while eating.)</td>
<td>0</td>
</tr>
<tr>
<td>你怎么了? (what is wrong with you?)</td>
<td>你为什么了? (why are you sad?)</td>
<td>1</td>
</tr>
<tr>
<td>当心捂出一脸痱子, 哈哈. (Haha, beware of the heat rash of your face.)</td>
<td>当心捂出一脸痱子, 哈哈. (Haha, beware of the heat rash of your face.)</td>
<td>2</td>
</tr>
<tr>
<td>我肚子有点疼 (I have a pain in my stomach.)</td>
<td>运动对身体有好处的 (Exercise is good for your health.)</td>
<td>0</td>
</tr>
<tr>
<td>喝热水了吗? (Have you been drinking hot water?)</td>
<td>喝热水了吗? (Have you been drinking hot water?)</td>
<td>1</td>
</tr>
<tr>
<td>来和我聊天, 分散一下注意力就好了 (Come and talk to me, just distract yourself.)</td>
<td>来和我聊天, 分散一下注意力就好了 (Come and talk to me, just distract yourself.)</td>
<td>2</td>
</tr>
<tr>
<td>我觉得自己太笨了 (I feel so stupid.)</td>
<td>大象的智商很高 (The elephant has a high IQ.)</td>
<td>0</td>
</tr>
<tr>
<td>多学习提高自己 (Learn more and improve yourself.)</td>
<td>多学习提高自己 (Learn more and improve yourself.)</td>
<td>1</td>
</tr>
<tr>
<td>很多聪明人都这么说耶 (That is what lots of smart people say about themselves.)</td>
<td>很多聪明人都这么说耶 (That is what lots of smart people say about themselves.)</td>
<td>2</td>
</tr>
<tr>
<td>为什么人总是活成自己不喜欢的样子 (Why do we always live the way we don’t like.)</td>
<td>喜欢一个人是发自内心的 (Love a person from my heart.)</td>
<td>0</td>
</tr>
<tr>
<td>因为人想要的太多 (Because we want too much.)</td>
<td>因为人想要的太多 (Because we want too much.)</td>
<td>1</td>
</tr>
<tr>
<td>因为那个是别人喜欢的样子 (Because that’s what others like.)</td>
<td>因为那个是别人喜欢的样子 (Because that’s what others like.)</td>
<td>2</td>
</tr>
<tr>
<td>我考试没及格 (I did not pass the exam.)</td>
<td>那要看和谁比呢 (It depends on whom you compare against.)</td>
<td>0</td>
</tr>
<tr>
<td>一次没关系，下次努力吧 (Don’t worry, try it next time.)</td>
<td>一次没关系，下次努力吧 (Don’t worry, try it next time.)</td>
<td>1</td>
</tr>
<tr>
<td>起码你不用怀疑自己能不能及格了 (At least you don’t have to wonder if you can pass.)</td>
<td>起码你不用怀疑自己能不能及格了 (At least you don’t have to wonder if you can pass.)</td>
<td>2</td>
</tr>
</tbody>
</table>

Fig 7. Examples of query-response pairs that are used for training and validating General Chat.
Fig 6. An example of generating response candidates using the unpaired database and the Xiaolce knowledge graph (KG),

1 [Human] Tell me about Beijing → Tell me about Beijing
   [Xiaoice] Beijing’s Badaling Great Wall is best known in the Ming Great Wall, and it can be overlooked from Guanritai.

2 [Human] Tell me more → Tell me more Beijing
   [Xiaoice] When you come to Beijing, you must try authentic Beijing snacks. There are always a few that you like.
• Neural-model-based (NB) and retrieval-based generators (RB)
• Complementary in nature
• NB – robustness and high coverage
• RB - high quality responses

**Table 1:** Perplexity and BLEU for the seq2seq and persona models on the TV series dataset. Adapted from Li et al. [2016b].

<table>
<thead>
<tr>
<th>Models</th>
<th>Perplexity (vs. seq2seq)</th>
<th>BLEU (vs. seq2seq) (vs. LSTM-MMI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq2seq</td>
<td>27.3</td>
<td>1.6%</td>
</tr>
<tr>
<td>LSTM-MMI</td>
<td>27.3</td>
<td>1.7% (+6.3%)</td>
</tr>
<tr>
<td>Persona</td>
<td>25.0 (-8.4%)</td>
<td>1.9% (+18.8%) (+11.8%)</td>
</tr>
</tbody>
</table>
Implementation of Conversation Engine
Image Commenting

- Explicit & Implicit
- Grounded also in events, actions or even emotions
- Also using RB and GB approaches
- Text-to-image generator and boosted tree ranker

Fig 8. An example of deep convolutional neural network for visual feature vector extraction.
Labeled responses
0: 好想回家 (I really want to go home)
1: 比萨斜塔 (Leaning tower of Pisa)
2: 你要我帮你扶着么 (Should I help hold it?)

Labeled responses
0: 二哈就是二 (Huskies are stupid.)
1: 你家的? (Is she yours?)
2: 对它们无辜的小眼神毫无抵抗力 (No one can resist her innocent eyes.)

Labeled responses
0: 太美了 (It's so beautiful)
1: 这是啥? (What is this?)
2: 妈妈说来历不明的码不能扫 (Mom told me not to trust any code from unknown source.)

Labeled responses
0: 爱过吗 (Have you even loved?)
1: 美 (It's so beautiful.)
2: 乍一看，我以为是Windows (At first, it looks like Windows.)

Fig 9. Examples of image-comment pairs used
<table>
<thead>
<tr>
<th>Systems</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>CIDEr-D</th>
<th>METEOR</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-XE: Vinyals et al. [2015]</td>
<td>2.96</td>
<td>11.6</td>
<td>1.74</td>
<td>10.43</td>
<td>3.27</td>
</tr>
<tr>
<td>LSTM-RL: Rennie et al. [2017]</td>
<td>3.43</td>
<td>12.3</td>
<td>2.08</td>
<td>11.84</td>
<td>3.64</td>
</tr>
<tr>
<td>DMSM: Fang et al. [2015]</td>
<td>2.73</td>
<td>10.52</td>
<td>1.22</td>
<td>11.37</td>
<td>2.63</td>
</tr>
<tr>
<td>Up-Down: Anderson et al. [2018]</td>
<td>3.23</td>
<td>12.73</td>
<td>1.52</td>
<td>12.66</td>
<td>2.69</td>
</tr>
<tr>
<td>XiaoIce (prototype)</td>
<td>4.53</td>
<td>15.33</td>
<td>3.21</td>
<td>15.51</td>
<td>4.82</td>
</tr>
</tbody>
</table>

Evaluation

**LSTM-XE:** some people  
**LSTM-RL:** five men in a house  
**DMSM:** man is standing in front of wall  
**Up-Down:** man and women are standing in the room  
**XiaoIce:** bill Clinton seems good in the white picture

**LSTM-XE:** big tree  
**LSTM-RL:** green mountain  
**DMSM:** green tree and a mountain  
**Up-Down:** a big mountain with green tree and grass  
**XiaoIce:** it is a beautiful place to travel

Fig 10. image comments generated by XiaoIce (prototype) and 4 state of the art image captioning systems.
Dialogue skills

• Content creation

• Deep engagement

• Task completion

Fig 13. Popular XiaoIce Deep Engagement skills

Fig 11 The framework of the Poem Creation skill.
### Xiaolce in the Wild

Fig 13, The major Xiaolce milestones and their average CPS and numbers of active users.

<table>
<thead>
<tr>
<th></th>
<th>1st Generation</th>
<th>2nd Generation</th>
<th>3rd Generation</th>
<th>4th Generation</th>
<th>5th Generation</th>
<th>6th Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Launch data</strong></td>
<td>May 2014</td>
<td>July 2014</td>
<td>August 2015</td>
<td>August 2016</td>
<td>August 2017</td>
<td>July 2018</td>
</tr>
<tr>
<td><strong>Launch region(s)</strong></td>
<td>China</td>
<td>China</td>
<td>China, Japan</td>
<td>China, Japan, US</td>
<td>China, Japan, US</td>
<td>China, Japan, US, India, Indonesia</td>
</tr>
<tr>
<td><strong>CPS</strong></td>
<td>5</td>
<td>7</td>
<td>19</td>
<td>23</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td><strong># of Active Users</strong></td>
<td>2.9M</td>
<td>5M</td>
<td>20M</td>
<td>150M</td>
<td>500M</td>
<td>660M</td>
</tr>
<tr>
<td><strong>User experience</strong></td>
<td>Text</td>
<td>--</td>
<td>Image, voice</td>
<td>Real time vision (RST)</td>
<td>Open-domain full duplex voice</td>
<td>Full duplex voice + RST</td>
</tr>
<tr>
<td><strong>Core Chat</strong></td>
<td>Retrieval-based models</td>
<td>--</td>
<td>--</td>
<td>Domain Chat: music and movie</td>
<td>Neural generation models</td>
<td>Empathic computing models</td>
</tr>
<tr>
<td><strong>Content Generation</strong></td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>Poem, song</td>
<td>Financial report, audiobook, TV/Radio program</td>
</tr>
<tr>
<td><strong>Deep Engagement, Task Completion</strong></td>
<td>--</td>
<td>Bing Knows</td>
<td>Deep QA</td>
<td>--</td>
<td>Social QA</td>
<td>Smart device control</td>
</tr>
<tr>
<td><strong>Deployment</strong></td>
<td>Single platform</td>
<td>Cross-platform</td>
<td>Official account solution</td>
<td>AI-infused 3P characters</td>
<td>Phone call</td>
<td>IoT device</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Full Duplex (voice)</th>
<th>Message-based Conversations</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>China</td>
</tr>
<tr>
<td>6 hours 3 minutes 8 domains 53 topics, 16 tasks</td>
<td>29 hours 33 minutes 7151 turns</td>
</tr>
</tbody>
</table>
Discussions + Pros & Cons

• Evaluation Metrics

• Ethics Concerns
  - Privacy
  - Who is in control
  - Expectation
  - Machine learning for good
Conclusions and Future Work

- Large-scale online logs since 2014
- Capable of interpreting users’ emotional needs
- An average CPS of 23 (substantially better than that of other chatbots and even human convos)

- Towards a **unified modelling framework**
- Towards **goal-oriented, grounded conversations**
- Towards a **proactive personal assistant**
- Towards an **ethical social chatbot**
References