Dialog Systems: Challenges and Trends

Dr. Minlie Huang (黄民烈)

THUAI, Tsinghua University
aihuang@tsinghua.edu.cn
http://coai.cs.tsinghua.edu.cn/hml
The Turing Test (1950)

“Can machines think?”

1912--1954

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.
History of Dialog Systems

1966
MIT Eliza

1977
基于Frame的任务对话系统GUS

1994
基于规则的聊天机器人系统ALICE

2010
苹果SIRI语音助手

2014
基于Seq2Seq的神经对话生成模型

2015

2016
• 阿里小蜜
• 百度度秘
• 小米小爱
• 天猫精灵
• 京东叮咚

2016 - 2018
• 谷歌智能个人助理
• 亚马逊Alexa大奖赛

2020
• Google: Meena
• FAIR: Blender
• OpenAI: GPT-3
Eliza 1966

- Created by MIT professor Joseph Weizenbaum
- Features
  - Hand-crafted scripts
  - Keyword spotting
  - Template matching

Human: Well, my boyfriend made me come here.

ELIZA: Your boyfriend made you come here?

Human: He says I'm depressed much of the time.

ELIZA: I am sorry to hear you are depressed.

Human: It's true. I'm unhappy.

ELIZA: Do you think coming here will help you not to be unhappy?

......

pronoun reversal (key tricks): I → you, me → you

• Human: He says I am depressed much of the time.
• Eliza: I am sorry to hear you are depressed.

• Human: I need some help, that much seems certain.
• Eliza: what would it mean to you if you got some help?
  • Pattern: I need some X → What would it mean to you if you got some X?

GUS 1977-Genial Understander System

GUS: Hello. My name is gus. I can help you plan a simple trip by air.

(1) Client: I want to go to San Diego on May 28

(2) Client: I must be in San Diego before 10 am

(3) Client: What is the next flight?

(4) Client: I'll take the first one

(5) Client: Dan Bobrow

GUS: I have confirmed the following flight:
P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.

What date do you want to return on?

GUS: Hello. My name is GUS. I can help you plan a simple trip by air.

GUS: Where do you want to go?

Client: I want to go to San Diego on May 28.

Datatable:

<table>
<thead>
<tr>
<th>DATE</th>
<th>MONTH</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAY</td>
<td>(BOUNDDED-INTEGER 1 31)</td>
<td></td>
</tr>
<tr>
<td>YEAR</td>
<td>INTEGER</td>
<td></td>
</tr>
<tr>
<td>WEEKDAY</td>
<td>(MEMBER (SUNDAY MONDAY TUESDAY WEDNESDAY THURSDAY FRIDAY SATURDAY))</td>
<td></td>
</tr>
</tbody>
</table>

a. Prototype for date

[ISA DATE
MONTH MAY
DAY 28]

b. The instance frame for May 28
IBM Watson 2010

2010 IBM Watson: Human–level knowledge QA
Defeat human champions in TV show Jeopardy!

✓ Answering factoid questions: which states does Mississippi River flow through?
✓ Finding answers from unstructured data
✓ Deep semantic parsing techniques in NLP

Emotional connections and invoking a skill in a casual chat

Showing IQ+EQ in a social chatbot
Alexa Prize Challenges

Challenge 2018

Gunrock from the University of California, Davis
Average score of 3.1 and average duration of almost 10 minutes

Challenge 2017

Sounding Board, University of Washington
Average score of 3.17 and average duration of 10 minutes and 22 seconds
Gunrock (Alexa Prize 2018)
Meena 2020 (Google)

• Model: 2.6B parameters
• Data: 40B words
• 30 days on a TPUv3 Pod (2,048 TPU cores)

---

Human: *I like how cute cows are*
Meena: *Ah, interesting. You like how they are cute, right?*
Human: *yeah, and they’re realy smart*
Meena: *I heard they go to college*
Human: *cows go to college?*
Meena: *I heard that a cow went to Harvard.*
Human: *what did the cow study?*
Meena: *Bovine sciences.*
Human: *do horses go to Harvard?*
Meena: *Horses go to Hayward.*
Human: *that’s a pretty good joke, I feel like you led me into it*
Meena: *You were trying to steer it elsewhere, I can see it.*

Adiwardana et al. 2020.
Towards a Human-like Open-Domain Chatbot.
Blender 2020 (Facebook)

• Model: 9.4B/2.7B/0.9B
• Data: 1.50B comments 56.8B label BPE tokens 88.8B context tokens
• Blending skills
  • Personality
  • Knowledge
  • Empathy

Recipes for building an open-domain chatbot.
Dialog System Types

• Task-oriented dialog systems (assistant)
• Open-domain dialog systems (chatbot)
• Multi-skill dialog systems (social chatbot)
Types of Dialog Systems

- **Chat Bot**
  - Social Conversation: High
  - Accomplish Tasks: Low
  - Purpose: chitchat

- **Task-oriented**
  - Social Conversation: Low
  - Accomplish Tasks: High
  - Purpose: execute commands, answer questions
Types of Dialog Systems

- **Task-oriented**
  - execute commands, answer questions

- **Social Chatbot**

- **Chat Bot**

- **Social Conversation**
  - chitchat

- **Accomplish Tasks**
  - High
  - Low
Architecture for Task-oriented Dialog Systems

"I want to find a Chinese restaurant."

User

Natural Language Understanding

Inform (cuisine="Chinese")

Dialog State Tracking

Dialog State

Dialog Manager

Dialog Policy

"Where do you want to eat?"

Natural Language Generation

Query

KB

Architecture for Open-domain Dialog Systems

U1: I like how cute cows are
M1: Ah, interesting. You like how they are cute, right?
U2: yeah, and they’re realy smart
M2: I heard they go to college
Different Issues & Features

Task-oriented

*Task intents
*Slot filling

*Narrow options & execute tasks
*Reward = timely task completion

Constrained domains

Structured KB+DB

Conversational AI System Components

Speech/language understanding

Dialog management

Language generation

Back-end application

Chatbots

*Social & info
*Intents
*Grounding

*Learn about interests
*Make suggestions
*R=user engagement

Open domains

Unstructured Data & Info
Strong vs. Weak Semantics

IBM Watson, GUS, Task-oriented system

Strong Semantics

Entity Recognition
Intent Classification
Semantic Parsing
Slot Filling for state tracking
Template-based language generation
Symbolic-based methods

Social Chatbot

Xiaoice, Alexa

Weak Semantics

Eliza, Meena, Blender Chatbot

Data-driven
End-to-end
Deep, large neural models
Social Chatbots

• **Social Chatbot = IQ + EQ**
  - IQ: question answering, task completion, recommendation, etc.
  - EQ: emotion, affect, social interactions, etc.

• **Task completion + Socialization**
  - Information need
  - Social need

**Challenges in Building Intelligent Open-domain Dialog Systems**
M Huang, X Zhu, J Gao
ACM Transactions on Information Systems (TOIS) 38 (3), 1-32
用户：你几岁了？
小冰：反正我是九零后!

用户：那你是哪一年出生的？
小冰：75年阳历一月出生。

用户：你是九零后，怎么会75年出生？
小冰：生在九零后，在深圳只能被当做八零后了。

用户：你家多大？
小冰：这么快就要爆年龄了？

用户：你是素食主义者还是吃肉啊？
小冰：怎么减肥

用户：早饭不好吃
小冰：那明早一起吃

用户：行，你推荐的应该不错
小冰：是吗？

用户：哈哈，我相信你
小冰：呃呃

用户：你喜欢我吗？
机器：我喜欢你，但我不喜欢你。

用户：我来自上海。
机器：真巧，我来自海南，老乡啊。
Challenges and Issues

Human: What's your favorite band?
Meena: Avenged sevenfold.

Human: Ooh, that's a good answer! What's your least favorite band?
Meena: Avenged sevenfold.

Human: Oh, I thought you liked Avenged sevenfold?
Meena: Yeah. That's my favorite band.

From: Gary Marcus. Then next decades in AI: the four steps towards robust artificial intelligence. 2020.
Challenges and Issues

Human: What are some of your new year resolutions?
Meena: Lose 60 pounds Learn Cantonese Learn French
Human: Don't you think, that's pretty far fetched? :P
Meena: Not really. I'm currently down 35 pounds, I can already speak Cantonese to a basic level, and I'm already learning French in school.
Human: Wao! That's impressive!
Meena: I'm learning French on the side, it's really easy for me since I'm a native Cantonese speaker.

From: Gary Marcus. Then next decades in AI: the four steps towards robust artificial intelligence. 2020.
Challenges and Issues

Serious ethic problems!

Picture from Prof. Frank Rudzicz, University of Toronto
Challenges in Socialbots

- Content, Context, Scene
- Personality & Persona
- Emotion & Sentiment
- Behavior & Strategy

Semantics
- Named Entity Recognition
- Entity Linking

Consistency
- Domain/Topic/Intent Detection
- Natural Language Generation
- Sentiment/Emotion Detection
- Personalization

Interactiveness
- Knowledge & Semantic Reasoning
- Context Modeling & Dialog Planning
Why is it so hard for Dialog?

• Dialog is not a **stand-alone** task
  • Entity recognition / linking
  • Text classification (intent, domain, sentiment, emotion, etc.)
  • Knowledge representation and reasoning
  • Natural language inference
  • Natural language generation

• Dialog involves **many** applications in one scenario
  • Chitchat, task completion, question answering, recommendation, information seeking, etc.
Why is it so hard for Dialog?

- Input is incomplete (输入信息不完备)
- One-to-many in semantics (一对多映射)

**Machine Translation**
输入包含了输出信息所需的完整语义

I love playing basketball.

**Dialog System**
多种可能，大部分观测不到

Input: Are you happy to be here?

Output-1: yes, very happy.
Output-2: ohh, totally not.
Output-3: I am lost in this place.
Output-4: It is too crowded.
Output-5: very good weather, I like the sunny days.

观察到部分

未观测部分
Present: what are we doing now?

• **Semantics (Knowledge)**
  • 有知识 (IJCAI 2018 distinguished paper)
  • 知识赋能 (Knowledge-grounded) (ACL 2020; to EMNLP 2020)

• **Consistency (Personality)**
  • 有人设 (Personality coherent) (IJCAI 2018)
  • 有个性 (Personal traits) (AAAI 2020)
  • 有风格 (Stylized) (to EMNLP 2020)

• **Interactiveness (Empathy)**
  • 懂情感 (Emotion-aware) (AAAI 2018)
  • 主动性 (Proactive behaviors) (ACL 2018)
  • 有话题 (Topic planning) (SIGIR 2018)
Knowledge Grounding

• Knowledge provides *kind of* understanding
• Contentful generation: aligning generation to knowledge

• World facts: entities, relations, (Jobs – founder – Apple)

• Commonsense knowledge consists of facts about the everyday world, that all humans are expected to know
  • Lemons are sour
  • Tree has leaves
  • Dog has four legs
User A: 知道《飞得更高》这首歌吗？
User B: 知道呀，这首歌入选了中歌榜中国年度最受华人欢迎十大金曲。
User A: 具体的发行时间你记得吗？
User B: 记得，是在 2005 年 3 月 19 日。
User A: 我觉得这首歌算是 汪峰 的经典之曲。
User B: 我也那么认为，编曲填词都由他自己完成，真的算是经典之作了。
Input: I have asthma since three years old.

Output: It is good for you to avoid triggers.

Zhou et al. Commonsense Knowledge Aware Conversation Generation with Graph Attention. *IJCAI-ECAI 2018 distinguished paper*
Personality

• Passing the **Turning Test**?
  • Deep semantic understanding
  • Existing chatting machine lacks *identity* or *personality*

• Personality is a well-defined concept in psychology ([Norman, 1963; Gosling et al., 2003](#))

• Extremely **subtle, implicit** in language expression:
  • Age, gender, language, speaking style, level of knowledge, areas of expertise
Personality

- Personality is important for game, custom service, virtual agent
- A coherent personality makes a system more trustable
- Personality is important for making effective social interactions

- **Level I: persona**
  - Extrinsic
  - Age, gender, social relations

- **Level II: personality**
  - Intrinsic
  - Introvert, extrovert, style

---

**From XiaoIce**

User: 你几岁了？
小冰: 我啊。。。反正是九零后！
User: 那你是几几年出生的？
小冰: 错，75年阳历一月出生。
User: 你是九零后，怎么会75年出生？
小冰: 生在九零后，在深圳只能被当做八零后了。
Personality (profile)

- Deliver coherent conversations w.r.t. identity/personality

Generic Dialogue Data for Training

UserA: how old are you?
UserB: I am six.

UserA: do you like to play piano?
UserB: I play violin.

Pre-specified Chatbot Profile

<table>
<thead>
<tr>
<th>Profile key</th>
<th>Profile value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Wang Zai</td>
</tr>
<tr>
<td>Age</td>
<td>3 years</td>
</tr>
<tr>
<td>Gender</td>
<td>Boy</td>
</tr>
<tr>
<td>Hobbies</td>
<td>Cartoon</td>
</tr>
<tr>
<td>Speciality</td>
<td>Piano</td>
</tr>
</tbody>
</table>

Generated Dialogues

User: how old are you?
Machine: I am three years old.

User: do you like to play piano?
Machine: Yes, I play piano.
Personality (trait)

- Learning dialog models from user’s personal traits
  - Age
  - Gender
  - Location
  - Personal interests

A Pre-training Based Personalized Dialogue Generation Model with Persona-sparse Data
Personality (style)

• From informal to formal?
• From impolite to polite?
• From modern to Jinyong?

现代风格↔金庸风格

Input: It’s only 9:57 pm and I’m already falling asleep.
Informal: it’s almost 9 am here and i just got up..

Formal: Yes, it is 9:06 pm here, and I am still on the couch.
Empathy

• **Emotion intelligence** is a key human behavior for intelligence (Salovey and Mayer, 1990; Picard and Picard, 1997)

• Understanding **emotion and affect** is important for dialogue and conversation
  
  • Enhance user performance
  
  • Improve user satisfaction
  
  • Less breakdowns
Empathy (shallow)

Like $\rightarrow$ Like (empathy)
Sadness $\rightarrow$ Sadness (empathy)
Sadness $\rightarrow$ Like (comfort)
Disgust $\rightarrow$ Disgust (empathy)
Disgust $\rightarrow$ Like (comfort)
Anger $\rightarrow$ Disgust
Happiness $\rightarrow$ Like

Empathy (shallow)

Perceiving and Expressing emotion by machine
Closer to human-level intelligence

Social Interaction Data

<table>
<thead>
<tr>
<th>Post</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Emotion Classifier

Happy  Angry  Sad ...

Emotion Tagged data

Emotional Chatting Machine

POST

承 诺  就 是 一 个  骗 子 说  给 一 个  傻 子  听 的 。

Other

喜  欢  Like

悲  伤  Sad

厌  恶  Disgust

愤怒  Angry

快乐  Happy

我  就  是 一 个  悲  伤  的  傻  子

我  就  是 一 个  专  一  的  傻  子

不管  你  信  不  信  ,  反  正  我  信  了  !

我  就  是  一 个  开  心  的  傻  子

Empathy (shallow)

**Label: Afraid**
**Situation:** Speaker felt this when...
“I’ve been hearing noises around the house at night”
**Conversation:**
**Speaker:** I’ve been hearing some strange noises around the house at night.
**Listener:** Oh no! That’s scary! What do you think it is?
**Speaker:** I don’t know, that’s what’s making me anxious.
**Listener:** I’m sorry to hear that. I wish I could help you figure it out

**Label: Proud**
**Situation:** Speaker felt this when...
“I finally got that promotion at work! I have tried so hard for so long to get it!”
**Conversation:**
**Speaker:** I finally got promoted today at work!
**Listener:** Congrats! That’s great!
**Speaker:** Thank you! I’ve been trying to get it for a while now!
**Listener:** That is quite an accomplishment and you should be proud!

Empathy (deep)

- 借鉴心理咨询的共情系统：情绪、话题、策略

谈话的三个阶段

1. 确认问题
2. 给与支持
3. 寻求解决

言语分类类型引自Hill’s Helping Skill System 中文修订版

Future: where to go?

• Robustness and uncertainty
• Huge data, huge model?
**Future: robustness & uncertainty**

**Paraphrase**
- Hi, I'm looking for a train that is going to Cambridge and arriving there by 20:45, is there anything like that?
- Yes, to Cambridge, and I would like to arrive by 20:45.

**Disfluency**
- I want a flight to Boston, um, to Denver.

**Coreference & Ellipsis**
- Tell me the address of that restaurant.
- Please make the reservation for me.

**ASR error**
- Could you book the went was (Wartworth) for one night one person?
- 能帮我查一下好萌(澳门)的天气吗？

**Out-of-domain**
- What is the common symptoms of COVID-19?
### Future: robustness & uncertainty

<table>
<thead>
<tr>
<th>ID</th>
<th>Configuration</th>
<th>Turn</th>
<th>Inform</th>
<th>Match</th>
<th>Succ.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Prec.</td>
<td>Rec.</td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>BERT</td>
<td>6.79</td>
<td>0.79</td>
<td>0.91</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>MILU</td>
<td>7.24</td>
<td>0.76</td>
<td>0.88</td>
<td>0.80</td>
</tr>
<tr>
<td>3</td>
<td>BERT</td>
<td>10.86</td>
<td>0.72</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>BERT</td>
<td>13.38</td>
<td>0.64</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>5</td>
<td>MDBT</td>
<td>16.55</td>
<td>0.47</td>
<td>0.35</td>
<td>0.37</td>
</tr>
<tr>
<td>12</td>
<td>SUMBT</td>
<td>18.67</td>
<td>0.27</td>
<td>0.32</td>
<td>0.26</td>
</tr>
<tr>
<td>13</td>
<td>SUMBT</td>
<td>13.92</td>
<td>0.36</td>
<td>0.64</td>
<td>0.44</td>
</tr>
<tr>
<td>14</td>
<td>TRADE</td>
<td>14.44</td>
<td>0.35</td>
<td>0.57</td>
<td>0.40</td>
</tr>
<tr>
<td>15</td>
<td>TSCP</td>
<td>18.20</td>
<td>0.37</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>16</td>
<td>DAMD</td>
<td>11.27</td>
<td>0.64</td>
<td>0.69</td>
<td>0.64</td>
</tr>
</tbody>
</table>

---

Ryuichi Takanobu et al. SIGDIAL 2020.  
Future: robustness & uncertainty

Uncertainty in natural language generation

- Avoid generating offensive language
- Avoid morality and ethics issues

Case from: “Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack”
Future: data & model Size

<table>
<thead>
<tr>
<th>Model</th>
<th>parameter size (G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT3-175B</td>
<td>1.25</td>
</tr>
<tr>
<td>GPT3-Small</td>
<td>94</td>
</tr>
<tr>
<td>Blender-Large</td>
<td>27</td>
</tr>
<tr>
<td>Blender-Medium</td>
<td>26</td>
</tr>
<tr>
<td>Meena</td>
<td>3.4</td>
</tr>
<tr>
<td>XLNet-Large</td>
<td>1.1</td>
</tr>
<tr>
<td>XLNet-Base</td>
<td>7.74</td>
</tr>
<tr>
<td>GPT2-Large</td>
<td>1.24</td>
</tr>
<tr>
<td>GPT2-Small</td>
<td>3.4</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>1.1</td>
</tr>
<tr>
<td>BERT-Base</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**Training Data**

- **GPT3**: 500B tokens; Common Crawl (410B), WebText2 (19B), Books1 (12B), Books2 (55B), Wikipedia (3B)
- **Blender**: Reddit (1.5B comments, 56.8B BPE tokens, 88.8B context tokens)
- **Meena**: 40B words, social media
- **XLNet**: 130GB Book, ClueWeb, Common Crawl
- **GPT2**: WebText (8M documents, 40GB text)
- **BERT**: 3.3B words; BookCorpus, English Wikipedia
Our recent papers

- A Pre-training Based Personalized Dialogue Generation Model with Persona-sparse Data. AAAI 2020
- Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition. ACL 2020
- CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset. TACL 2020

More is on http://coai.cs.tsinghua.edu.cn/hml
Thanks for your attention

• Recruiting post-docs, PhDs, & interns

• Minlie Huang, Tsinghua University
  • aihuang@tsinghua.edu.cn
  • http://coai.cs.tsinghua.edu.cn/hml