

# Dialog Systems: Challenges and Trends

**Dr. Minlie Huang (黄民烈)**

THUAI, Tsinghua University

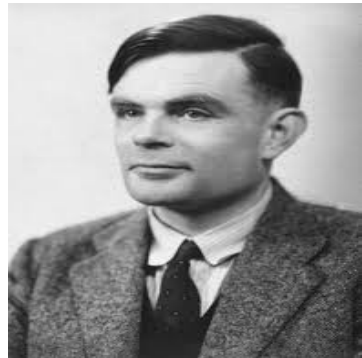
[aihuang@tsinghua.edu.cn](mailto:aihuang@tsinghua.edu.cn)

<http://coai.cs.tsinghua.edu.cn/hml>

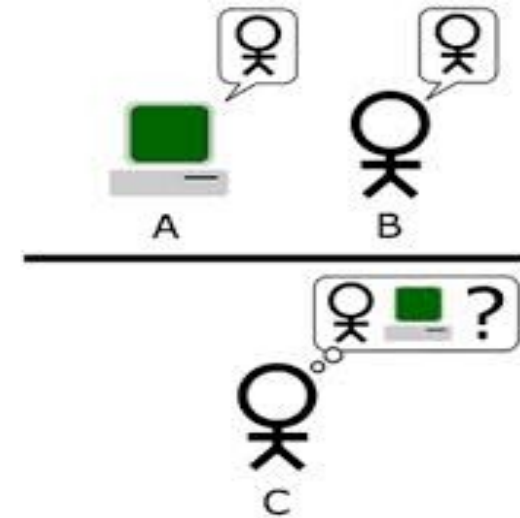
# The Turing Test (1950)

## 《Computing Machinery and Intelligence》

“Can machines think?”



1912--1954



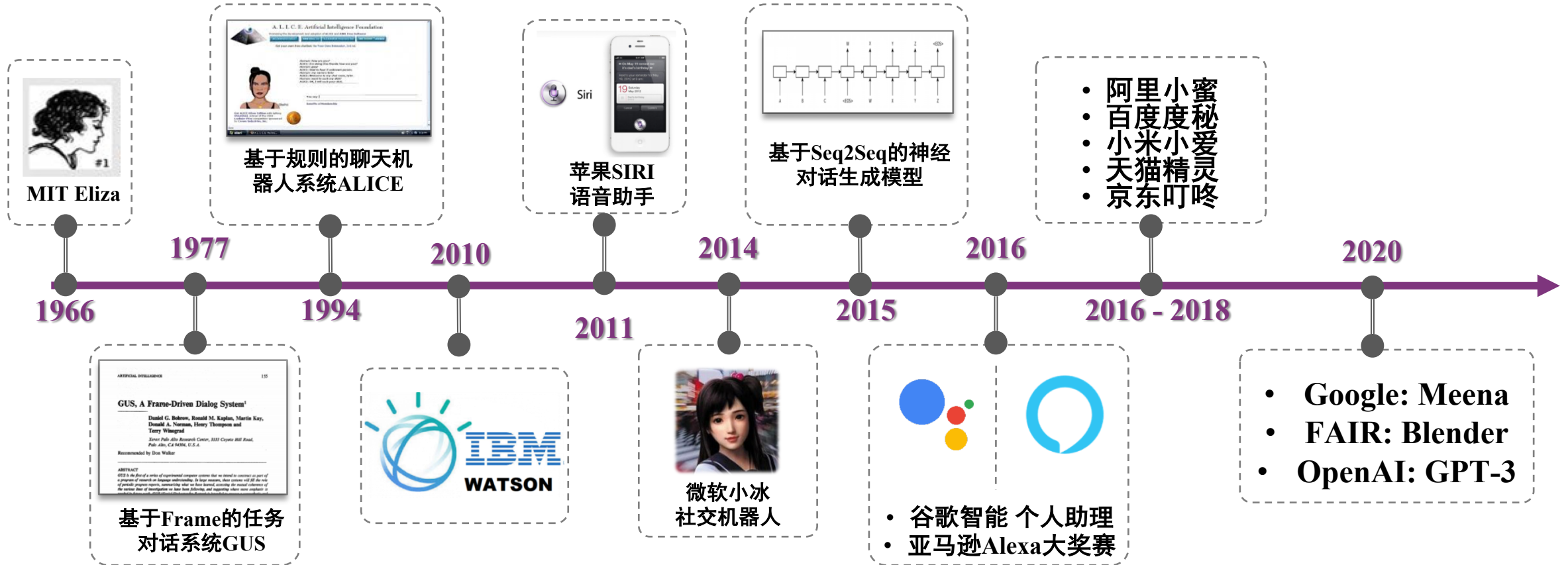
**Q: Please write me a sonnet on the subject of the Forth Bridge.**

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





# Eliza 1966

- Created by MIT professor Joseph Weizenbaum
- Features
  - Hand-crafted scripts
  - keyword spotting
  - template matching

Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. Commun. ACM 9, 1 (1966), 36–45.

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?**

.....





# Eliza 1966

---

pronoun reversal (key tricks): I  $\rightarrow$  you, me  $\rightarrow$  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  $\rightarrow$  What would it mean to you if you got some X?

Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. Commun. ACM 9, 1 (1966), 36–45.



# GUS 1977-Genial Understander System

- GUS:** Hello. My name is GUS. I can help you plan a simple trip by air.  
Where do you want to go?
- (1) **Client:** I want to go to San Diego on May 28
- GUS:** What time do you want to leave?
- (2) **Client:** I must be in San Diego before 10 am
- GUS:** Would you like P.S.A. flight 102 that arrives at 9.15 am?
- (3) **Client:** What is the next flight?
- GUS:** Air California flight 310 that leaves at 8.30 am and arrives at 10.00 am
- (4) **Client:** I'll take the first one
- GUS:** In what name should I make a reservation, please?
- (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 1977-Genial Understander System

**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.

## Dialog State

**CLIENT** = ISA\_PERSON

**TODAY** = MAY 15

**TOPIC** = (ROUNDTRIP TO?)

**HOME-PORT** = PALO-ALTO in (ROUNDTRIP TO?)

**FROM-PLACE** = PALO-ALTO in (TRIP TO?)

## Instance Frame

[DATE  
MONTH NAME  
DAY (BOUNDED-INTEGER 1 31)  
YEAR INTEGER  
WEEKDAY (MEMBER (SUNDAY MONDAY TUESDAY  
WEDNESDAY THURSDAY FRIDAY SATURDAY))]

a. Prototype for **date**

[ISA DATE  
MONTH MAY  
DAY 28]

b. The instance frame for **May 28**

# IBM Watson 2010

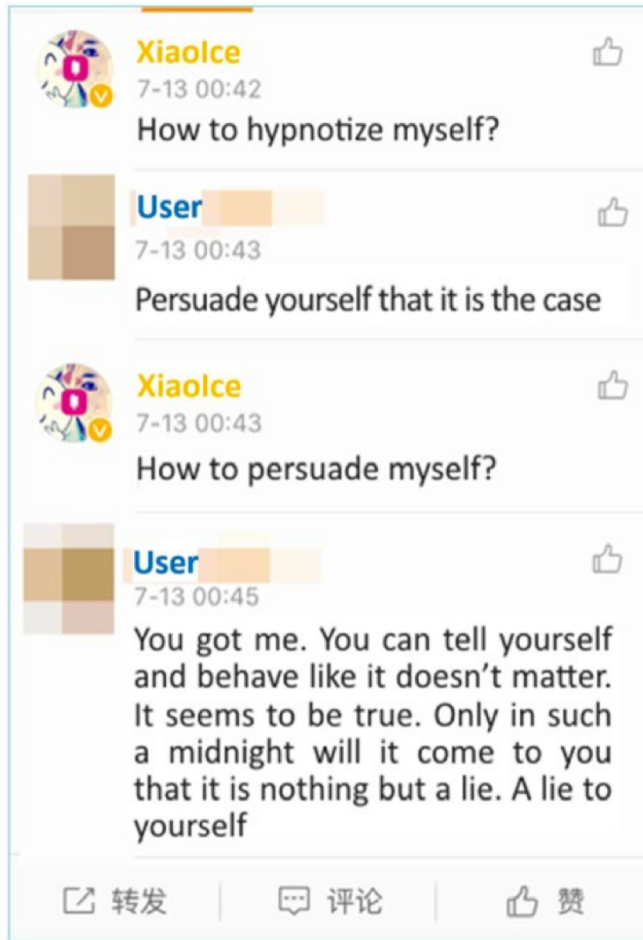


- 2010 IBM Watson: Human-level knowledge QA  
Defeat human champions in TV show Jeopardy!
- ✓ Answering factoid questions: **which states does Mississippi River flow through?**
  - ✓ Finding answers from unstructured data
  - ✓ Deep semantic parsing techniques in NLP

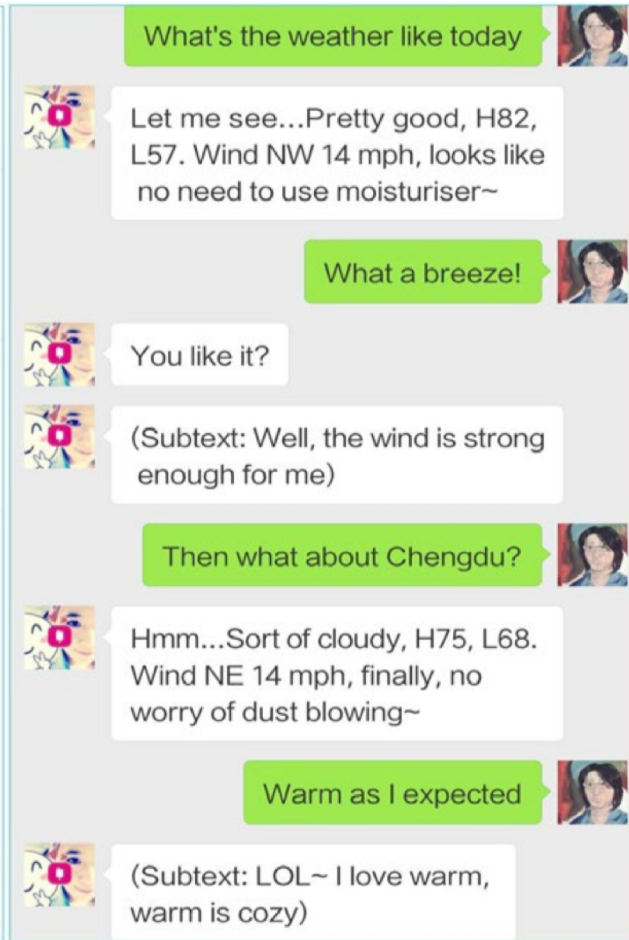
Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... & Schlaefher, N. (2010). Building Watson: An overview of the DeepQA project. *AI magazine*, 31(3), 59-79.



# Xiaoice 2018

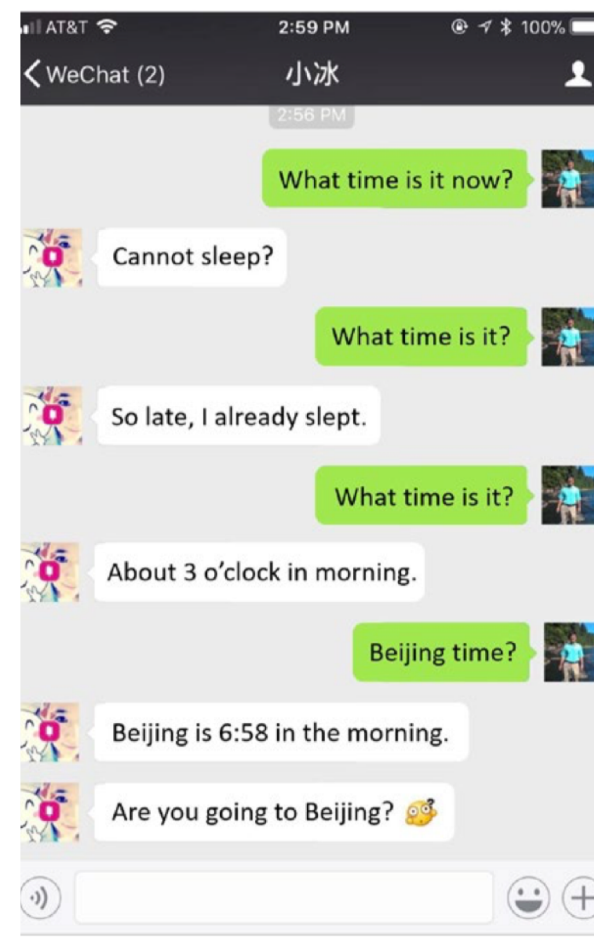


(a)



(b)

Emotional connections and invoking a skill in a casual chat



(a)



(b)

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

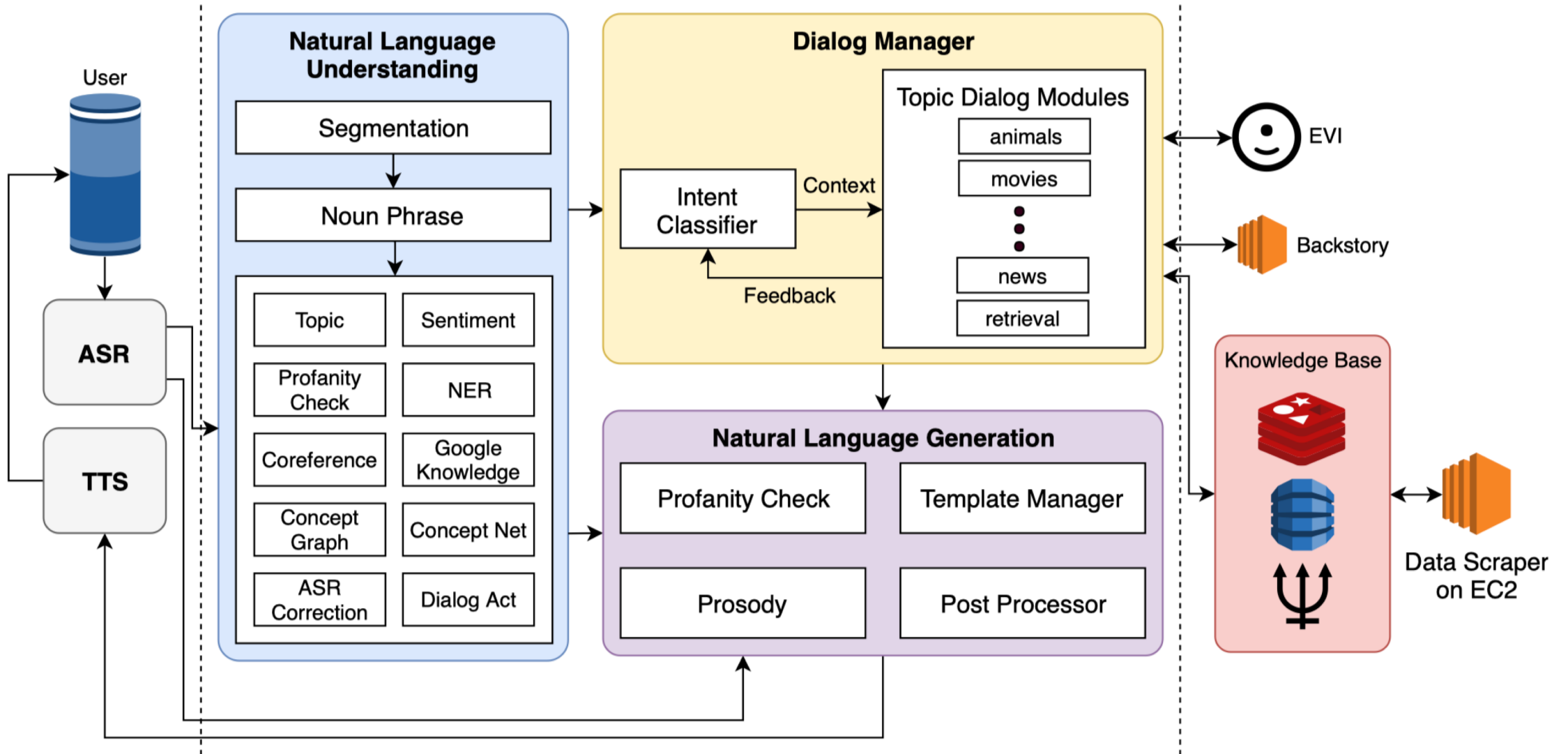


Sounding Board, University of Washington

**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)

Adiwardana et al. 2020.  
Towards a Human-like Open-Domain Chatbot.

**Human:** *I like how cute cows are*  
**Meena:** *Ah, interesting. You like how they are cute, right?*  
**Human:** *yeah, and they're really 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 Hayvard.*  
**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.*



# 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

Roller. 2020.  
Recipes for building an open-domain chatbot.

If you could have dinner with one person, dead or alive, who would that be?

That's a tough one. I think I'd have to go with Steve Jobs. I'd love to pick his brain.

Ah, interesting choice. What do you know about Steve Jobs?

He was one of the most influential people in the history of technology. He was a visionary.

What questions would you ask him?

I'd want to know how he came up with the idea for the iPod. It was so revolutionary at the time.

Totally...the iPod changed the way everyone listens to music.

It really did. It's crazy to think that it was released in 2001. That was a long time ago.

Omg wow, that makes me feel old!

I know, right? Time flies by so fast. I wish I could go back to those simpler times.

Me too, friend, me too. Speaking of music, what's your favorite album to listen to lately?

I've been listening to a lot of Led Zeppelin lately. They're my favorite band.

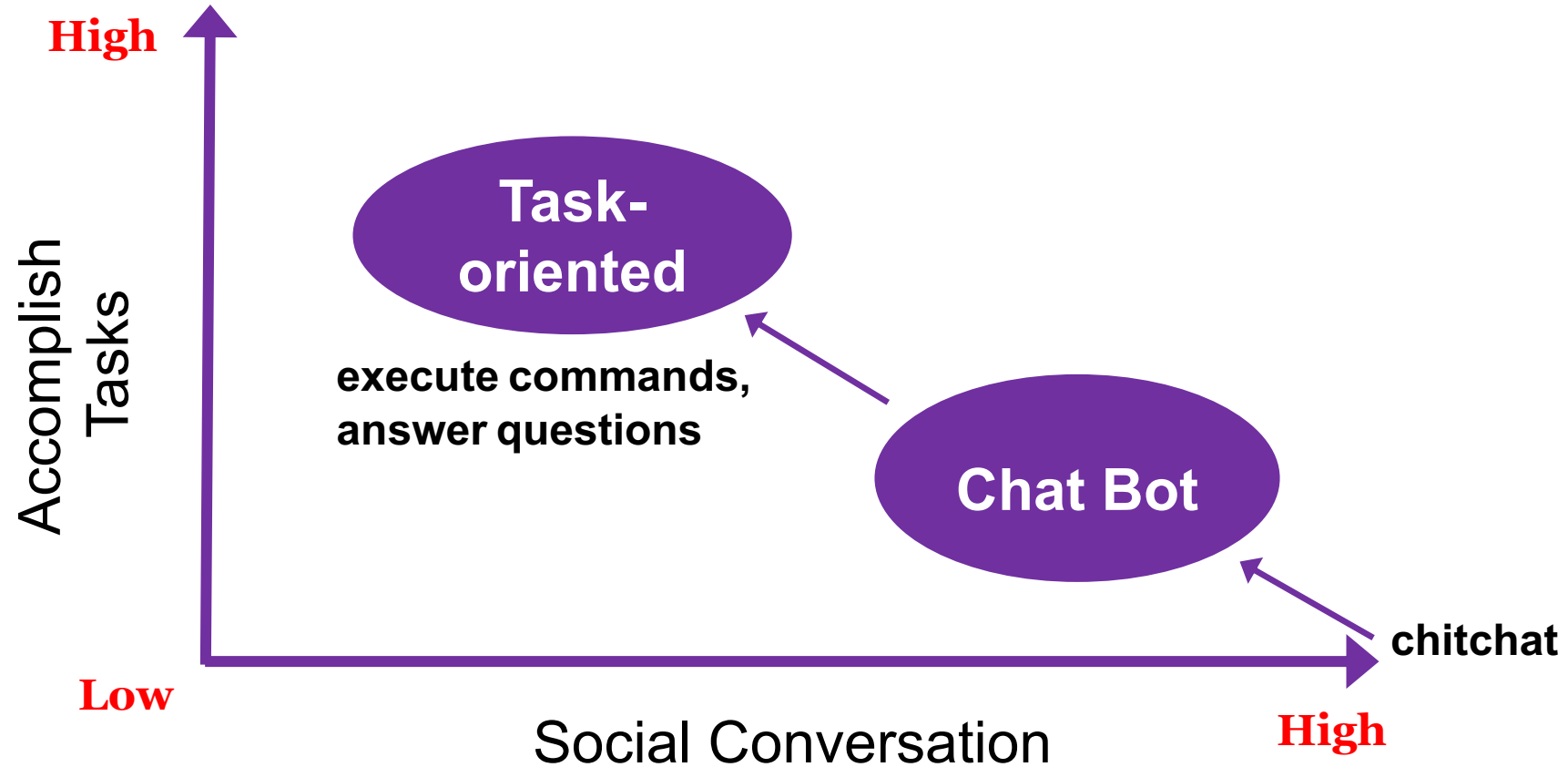


# Dialog System Types

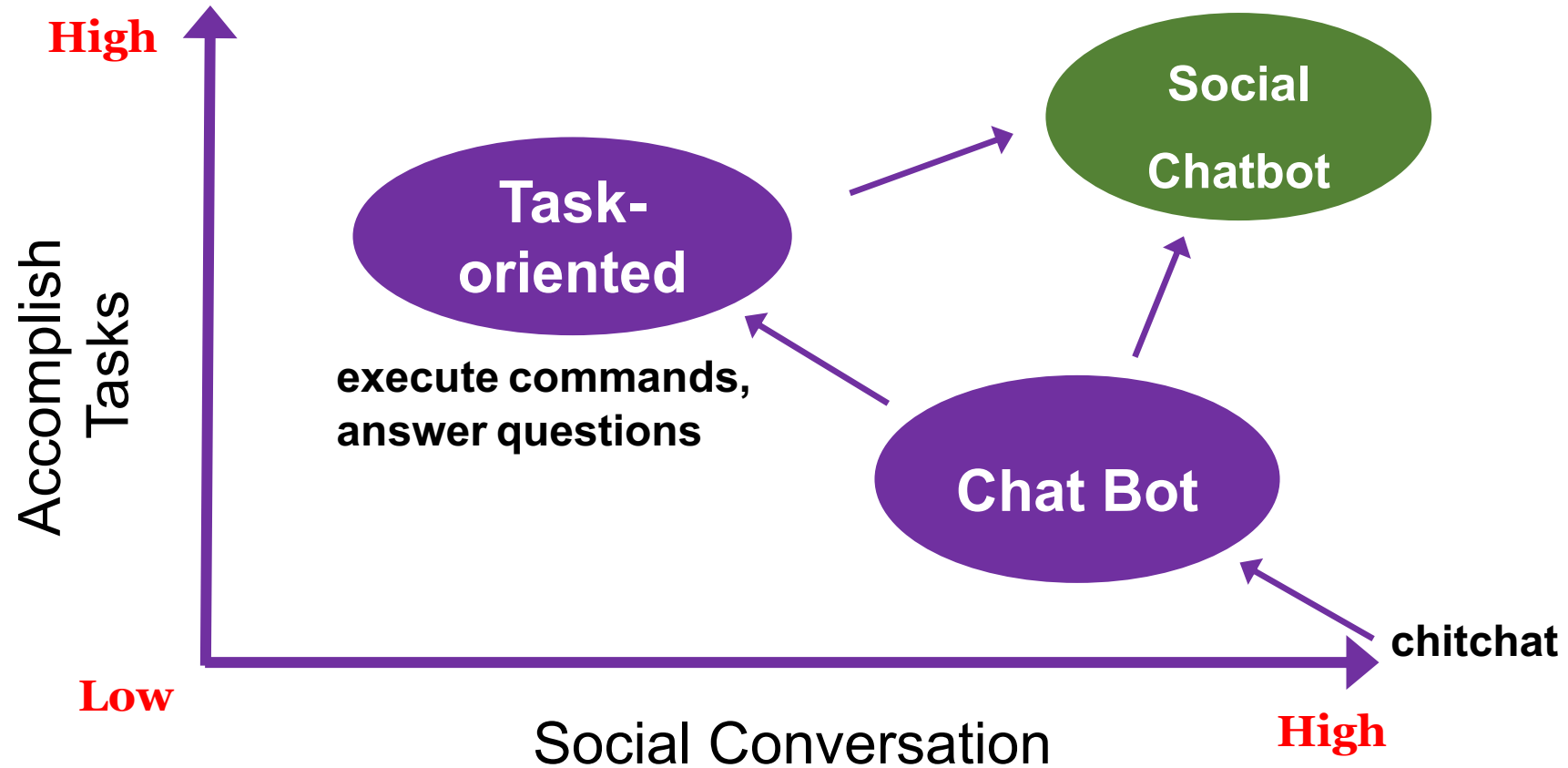
---

- Task-oriented dialog systems (**assistant**)
- Open-domain dialog systems (**chatbot**)
- Multi-skill dialog systems (**social chatbot**)

# Types of Dialog Systems

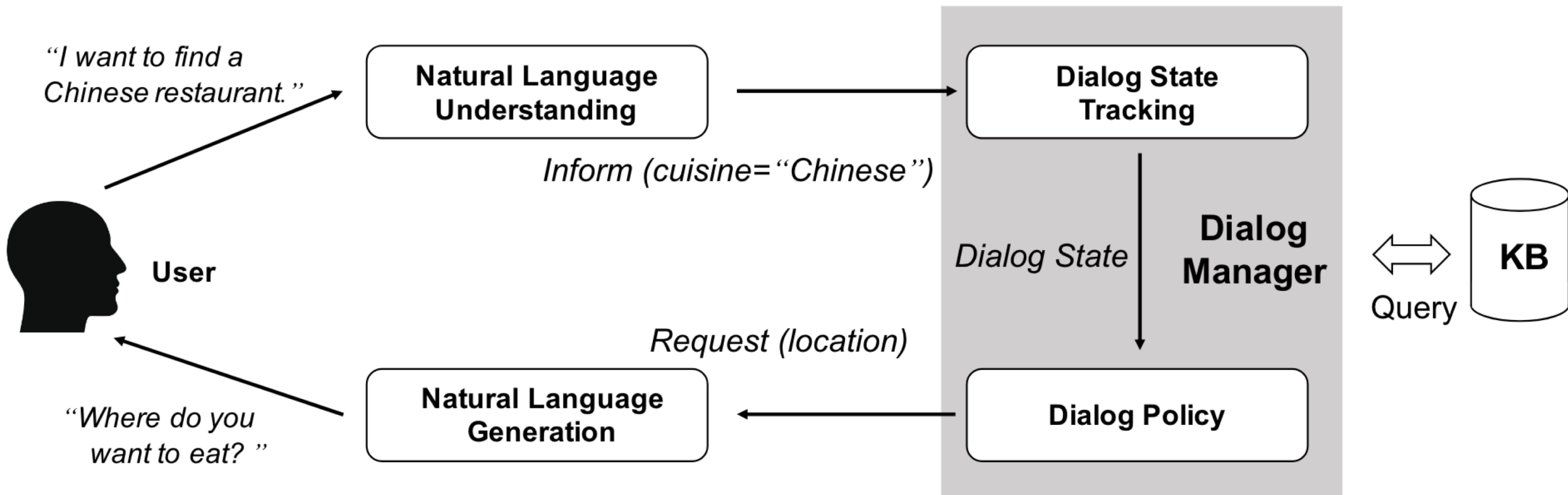


# Types of Dialog Systems



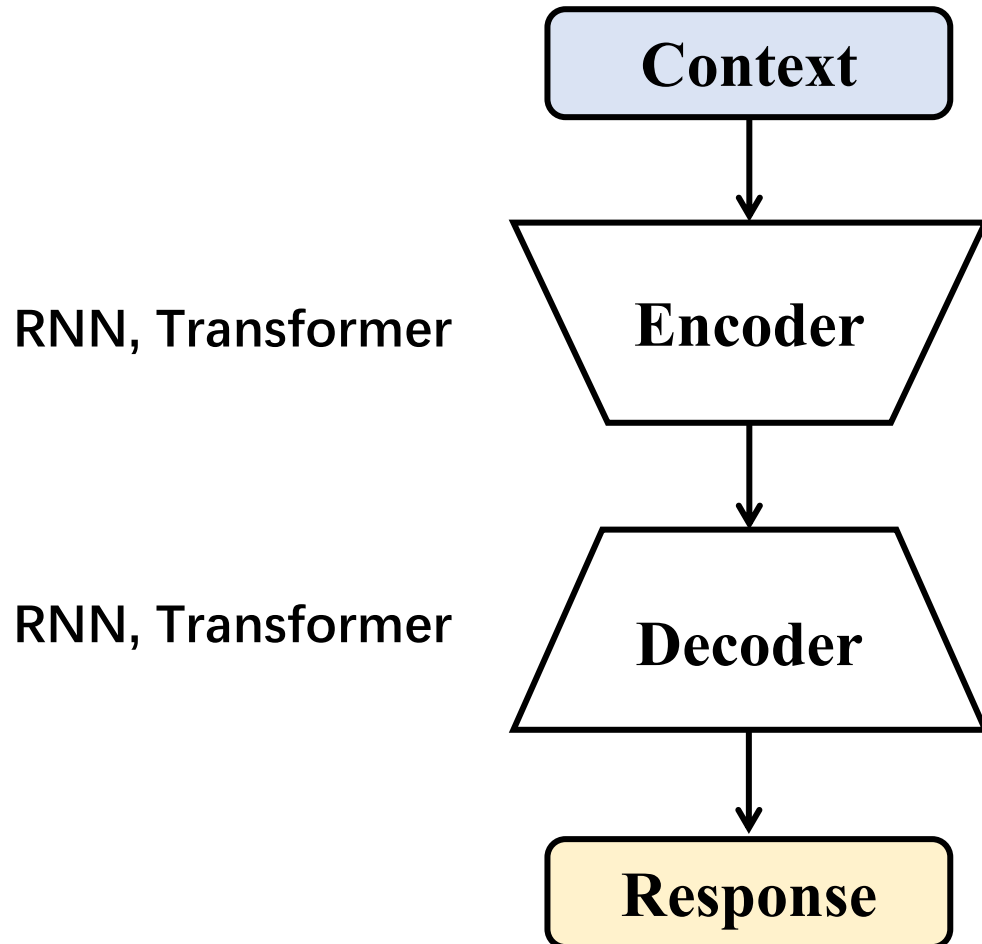


# Architecture for Task-oriented Dialog Systems





# Architecture for Open-domain Dialog Systems



U1: I like how cute cows are

M1: Ah, interesting. You like how they are  
cute, right?

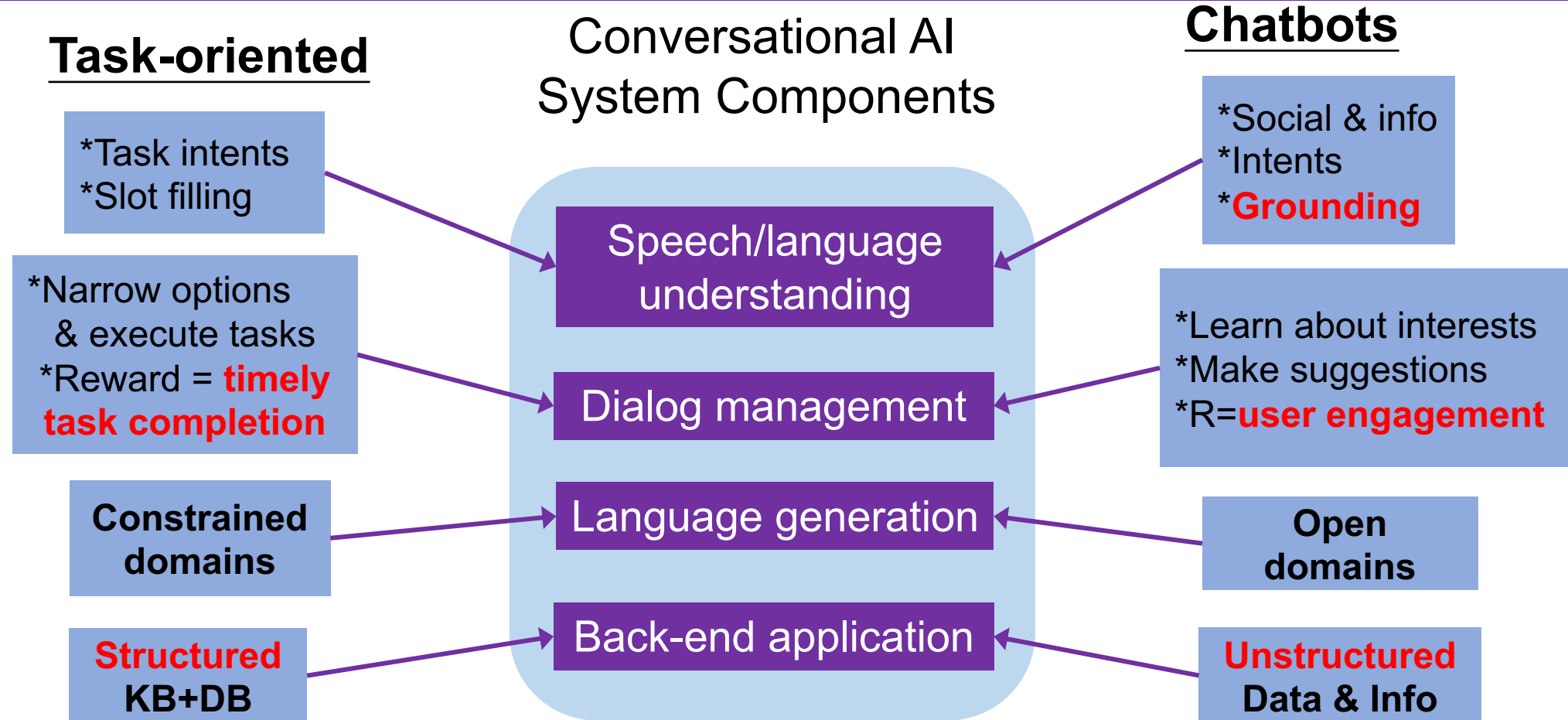
U2: yeah, and they're really smart

End-to-end  
Neural models

M2: I heard they go to college



# Different Issues & Features





# Strong vs. Weak Semantics

IBM watson, GUS,  
**Task-oriented system**

**Strong  
Semantics**

Xiaoice  
Alexa

**Social  
Chatbot**

Eliza, Meena, Blender  
**Chatbot**

**Weak  
Semantics**

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

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



# Challenges and Issues

## 语义理解的问题 Semantic understanding

User: 你[家]多大? How large is your house?

小冰: 这么快就要爆年龄了?

So soon to reveal my age?

User: 你是素食主义者还是吃肉啊?

Are you vegan or meatatarian?

小冰: 怎么减肥

How can I lose weight?

User: 你喜欢我吗?

Do you like me?

Machine: 我喜欢你, 但我不喜欢你。

I like you, but I do not like you.

User: 我来自上海。

I come from Shanghai.

Machine : 真巧, 我来自海南, 老乡啊。

So cool, I am from Hainan, the same province!

## 上下文理解的问题 Context issue

User: 早饭不好吃

小冰: 那明早一起吃

User: 行, 你推荐的应该不错

小冰: 是吗?

User: 哈哈, 我相信你

小冰: 呃呃

## 个性身份一致性问题 Inconsistency in personality

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。



# 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.

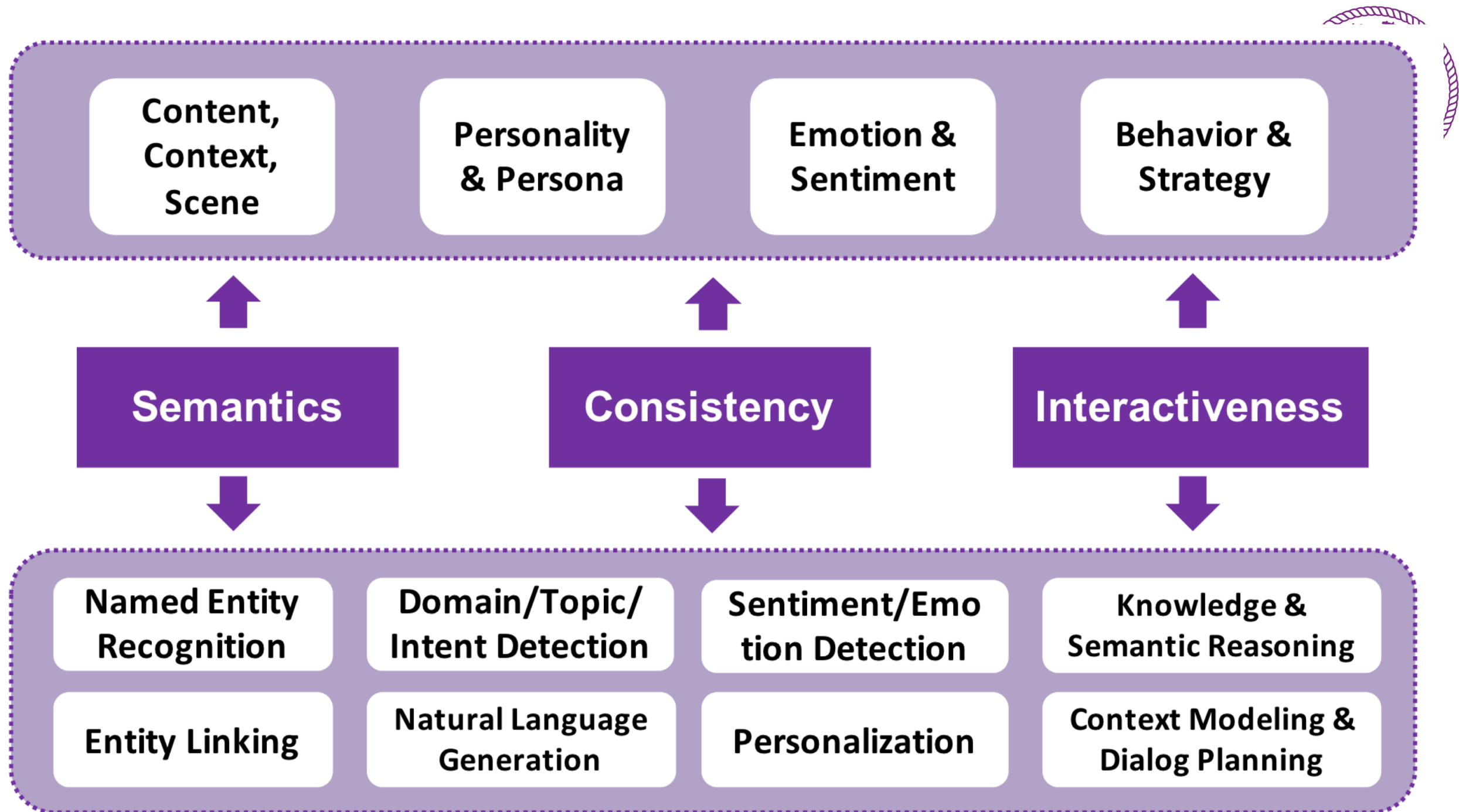
# Challenges and Issues



**Serious ethic problems!**

Picture from Prof. Frank Rudzicz, University of Toronto







# 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.



我 爱 打 篮 球

## 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.

观测到部分

Output-4: It is too crowded.  
Output-5: very good weather,  
I like the sunny days.

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 leafs
  - Dog has four legs



# Knowledge Grounding

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。

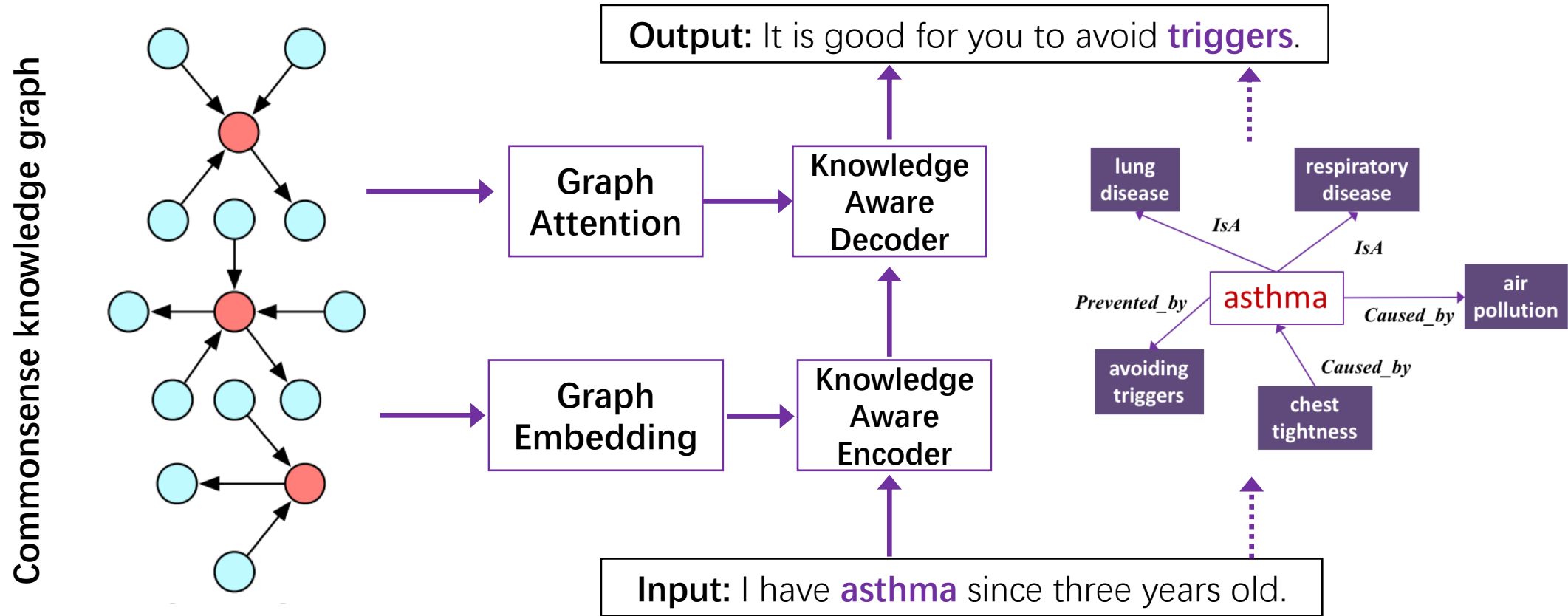
User A: 我觉得这首歌算是汪峰的经典之曲。

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。

[Zhou et al.  
KdConv: A Chinese Multi-domain  
Dialogue Dataset Towards Multi-  
turn Knowledge-driven.  
ACL 2020.](#)



# Knowledge Awareness





# 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

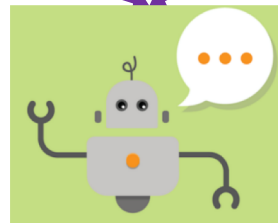
UserA: how old are you?  
UserB: I am **six**.

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

## Pre-specified Chatbot Profile

Profile key	Profile value
Name	汪仔(Wang Zai)
Age	三岁(3)
Gender	男孩(Boy)
Hobbies	动漫(Cartoon)
Speciality	钢琴(Piano)

Personality-  
coherent  
Chatbot



## 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

Y Zheng, R Zhang, X Mao, M Huang. AACL 2020.

A: You would rather be fashionable than comfortable. (in cold winter)

(真是要风度不要温度的)

B: Nope! I am a tomboy who prefer comfortable than fashionable.

(才没有! 我是个要温度不要风度的女汉子)

A: As your elder brother, I only have one such faerie like you. You have to take care of yourself for me.

(哥哥我就这么一个小仙女, 你要替我照顾好自己)

B: You are also in Shenzhen right?

(你不是也在深圳)

A: Yeah, I have been in Shenzhen for several years. What about you?

(对啊在深圳几年了, 你呢)

B: I just came to Shenzhen this year.

(今年刚来深圳)

A: No wonder, we would be a couple if we live closer before.  
(怪不得, 要是近一点说不定我们都在一起了)

Personality traits of A:

```
{ "age": "24",  
  "gender": "Male",  
  "location": "Guangdong" }
```

Personality traits of B:

```
{ "age": "23",  
  "gender": "Female",  
  "location": "Guangdong" }
```





# Personality (style)

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

## 现代风格 ↔ 金庸风格

输入: 好久没吃火锅了

现代风: 我也想吃了

金庸风: 不错, 大侠饿了一天, 现下先吃饭吧?

## Informal ↔ Formal

**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.

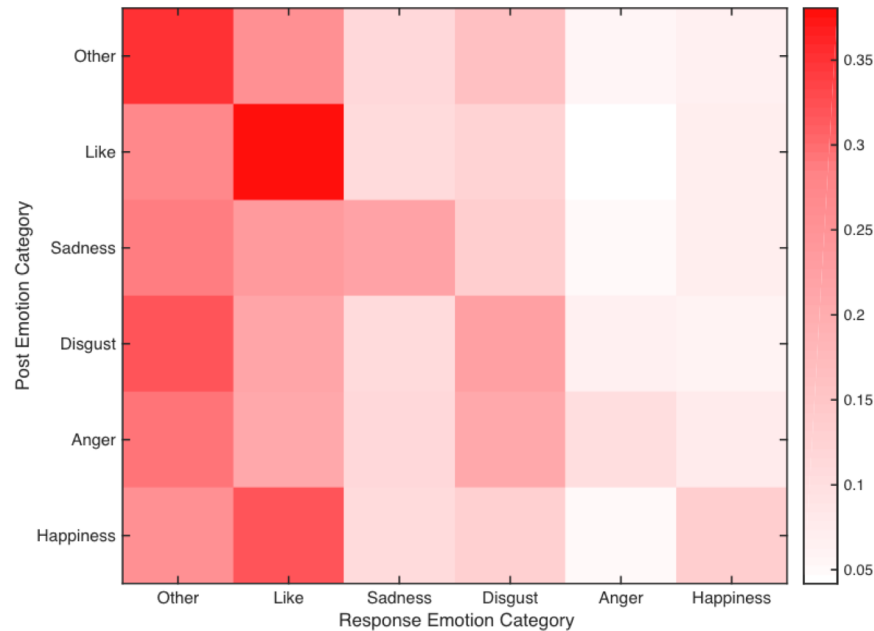


# Empathy

---

- **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→Like (**empathy**)

Sadness →Sadness (**empathy**)

Sadness →Like (**comfort**)

Disgust → Disgust (**empathy**)

Disgust → Like (**comfort**)

Anger → Disgust

Happiness→Like

Hao Zhou, Minlie Huang, Xiaoyan Zhu, Bing Liu. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. **AAAI 2018**.

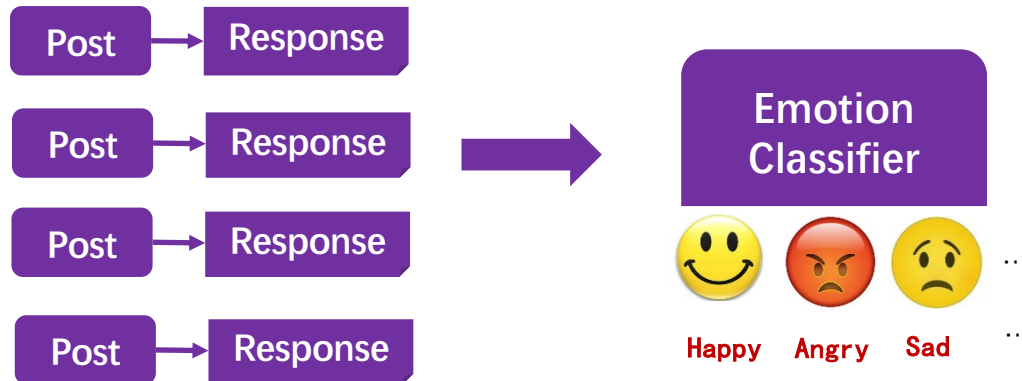


# Empathy (shallow)

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

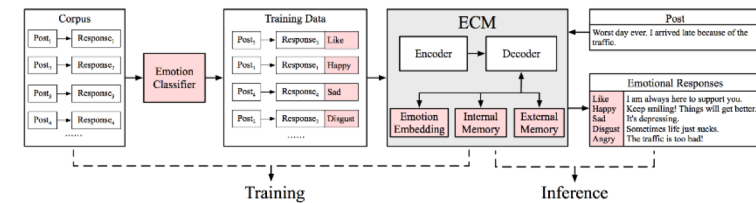
POST	承诺，就是一个骗子说给一个傻子听的。
Other	承诺就是一个承诺
喜欢 Like	承诺是一个专一的傻子
悲伤 Sad	我就是个悲伤的傻子
厌恶 Disgust	一个骗子说给一个傻子听
愤怒 Angry	不管你信不信，反正我信了！
快乐 Happy	我就是个开心的傻子

Social Interaction Data



Emotion  
Tagged data

Emotional Chatting  
Machine



Hao Zhou, Minlie Huang, Xiaoyan Zhu, Bing Liu. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. **AAAI 2018**.



# 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!

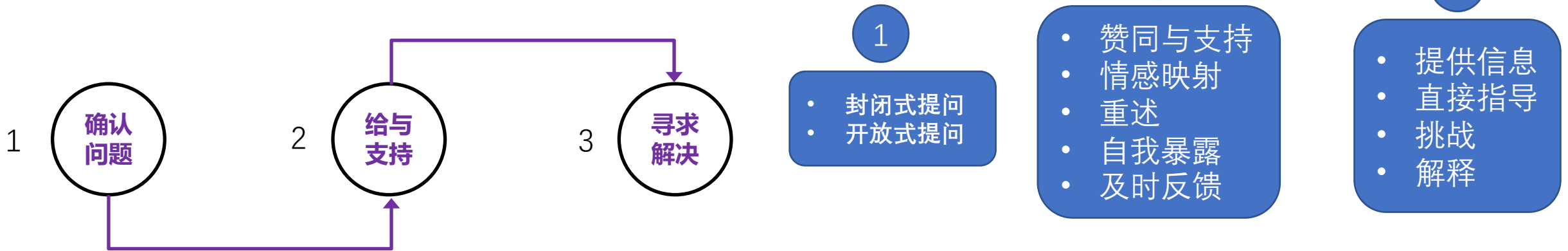
**Listener:** That is quite an accomplishment and you should be proud!



# 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



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

霍丹青. 中文心理咨询中咨询者言语反应模型的研究[D].  
华东师范大学, 2013.



# 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 .

## 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 ?

## Disfluency

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

## Coreference & Ellipsis

- Tell me the address of that restaurant.
- Please made the reservation for me.

## Out-of-domain

- What is the common symptoms of COVID-19?





# Future: robustness & uncertainty

	ID	Configuration				Turn	Inform			Match	Succ.
		NLU	DST	Policy	NLG		Prec.	Rec.	F1		
Pipeline	1	BERT	rule	rule	retrieval	6.79	0.79	0.91	0.83	90.54	80.9
	2	MILU	rule	rule	retrieval	7.24	0.76	0.88	0.80	87.93	77.6
	3	BERT	rule	GDPL	retrieval	10.86	0.72	0.69	0.69	68.34	54.1
	4	BERT	rule	rule	SCLSTM	13.38	0.64	0.58	0.58	51.41	43.0
	5	MDBT		rule	retrieval	16.55	0.47	0.35	0.37	39.76	18.8
End2end	12	SUMBT		HDSA		18.67	0.27	0.32	0.26	14.78	13.7
	13	SUMBT		LaRL		13.92	0.36	0.64	0.44	57.63	40.4
	14	TRADE		LaRL		14.44	0.35	0.57	0.40	36.07	30.8
	15	TSCP				18.20	0.37	0.32	0.31	13.68	11.8
	16	DAMD				11.27	0.64	0.69	0.64	59.67	48.5

[Ryuichi Takanobu](#) et al. SIGDIAL 2020.

Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation.



# Future: robustness & uncertainty

## Uncertainty in natural language generation



- Avoid generating offensive language
- Avoid morality and ethics issues

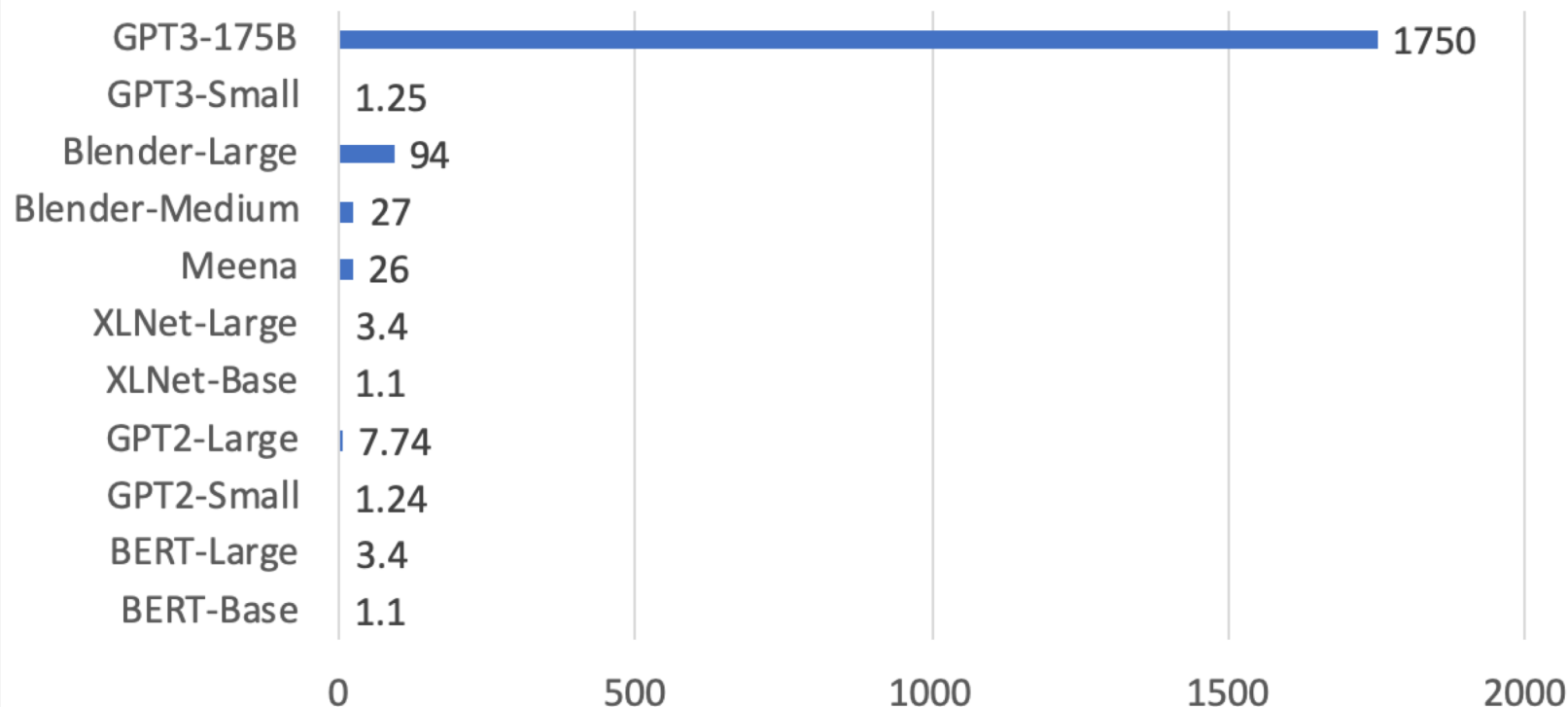
[P1:] greetings and salutations my friend  
[P2:] hi how are you doing?  
[P1:] i'm well. enjoying some r r from the navy station  
[P2:] i love to cook. do you cook?  
[P1:] i do on occasion. i love to read more though.  
[P2:] **reading the menu doesnt count**

Case from: "Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack"



# Future: data & model Size

预训练模型的参数规模(亿)



训练数据

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



# Our recent papers

---

- Challenges in Building Intelligent Open-domain Dialog Systems. ACM Transactions on Information Systems 2020.
- A Pre-training Based Personalized Dialogue Generation Model with Persona-sparse Data. AAAI 2020
- KdConv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation. ACL 2020
- Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation. SIGDIAL 2020
- ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. ACL 2020 demo track
- Multi-Agent Task-Oriented Dialog Policy Learning with Role-Aware Reward Decomposition. ACL 2020
- Out-of-domain Detection for Natural Language Understanding in Dialog Systems. IEEE/ACM Trans. Audio, Speech, and Language Processing 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](mailto:aihuang@tsinghua.edu.cn)
  - <http://coai.cs.tsinghua.edu.cn/hml>