Deeper Conversational AI

NeurIPS 2020 Tutorial

Pascale Fung, Yun-Nung (Vivian) Chen, Zhaojiang Lin, Andrea Madotto

The Hong Kong University of Science & Technology

National Taiwan University

EMOS Technologies

Outline

1. Conversational AI Overview
2. Generation-Based Deep Conversational AI
3. Future Work of Deeper Conversational AI




(Part 1)
Conversational AI Overview

Yun-Nung (Vivian) Chen

(Part 1)
Conversational AI Overview

1.1. Brief History of Conversational AI

1.2. Modularized Task-Oriented Dialogue Systems

1.3. Retrieval Based Chit-Chat Dialogue Systems

Generation based Conv. AI

Challenges and Future Work

(Part 1) Conversational AI Overview

1.1. Brief History of Conversational AI

1.2. Modularized Task-Oriented Dialogue Systems

1.3. Retrieval Based Chit-Chat Dialogue Systems

Generation based Conv. AI

Challenges and Future Work

Brief History of Conversational Systems

Keyword Spotting *(e.g., AT&T)*
System: *"Please say collect, calling card, person, third number, or operator"*

Early 1990s

Early 2000s

Intent Determination
(Nuance's Emily™, AT&T HMIHY)
User: "Uh…we want to move…we want to change our phone line from this house to another house"

TV Voice Search

IBM WATSON
DARPA CALO Project

2017

Apple Siri (2011)

Google Assistant (2016)

Microsoft Cortana (2014)

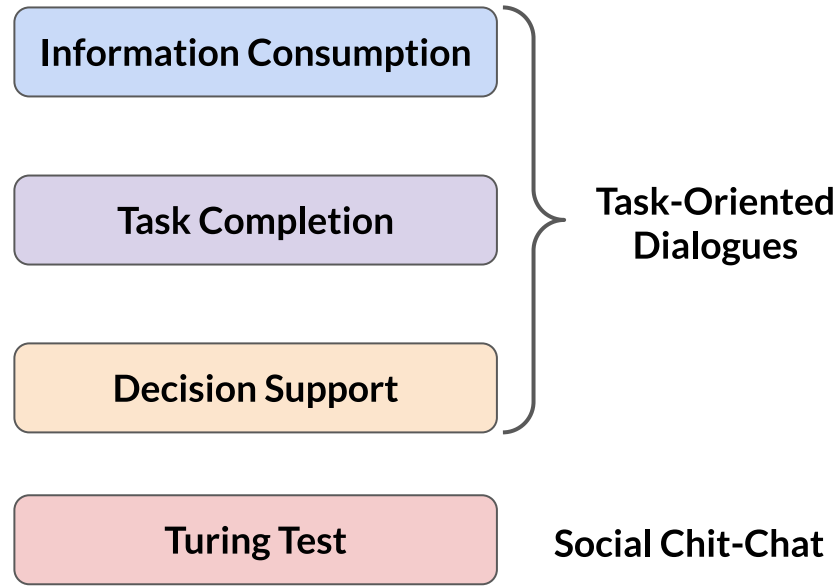
Amazon Echo/Alexa (2014)

Apple HomePod (2017)

Facebook Portal (2019)

Functionality of Conversational Systems

- "I have a question"
  - "What is today's agenda?"
  - "What does NLP stand for?"
- "I need to get this done"
  - "Book me a ticket from Taipei to Hong Kong"
  - "Schedule a meeting with Vivian"
- "What should I do?"
  - "Is this tutorial good to attend?"
- "I want to chat"
  - "Nice to meet you!"



Conversational AI Overview

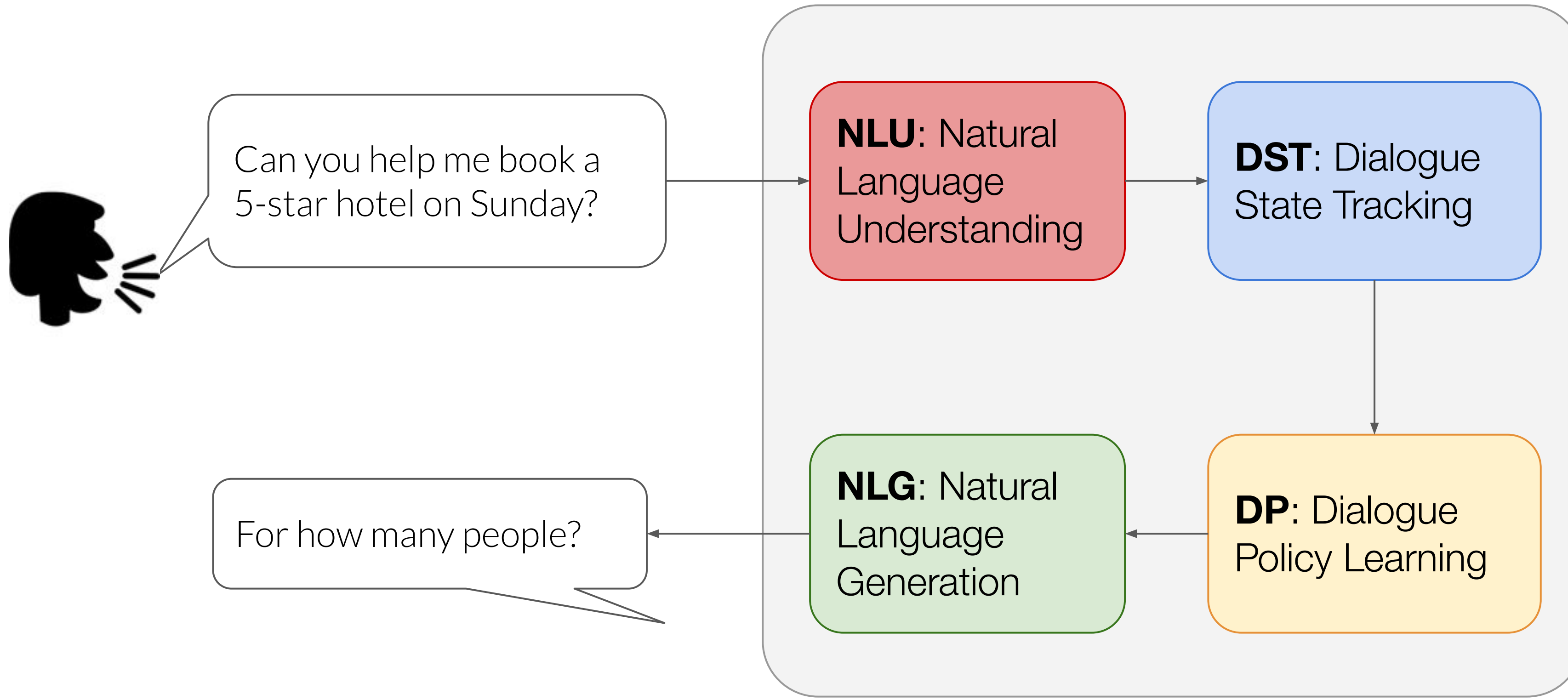
Task-Oriented Dialogue Systems

- Designed to help users achieve pre-defined goals or tasks
- Aims at fulfilling user requests with the least number of turns
- Dealing with APIs or databases
- Typical scenarios:
  - Restaurant reservation
  - Hotel reservation
  - Airplane booking
  - Attraction search
  - Weather forecast

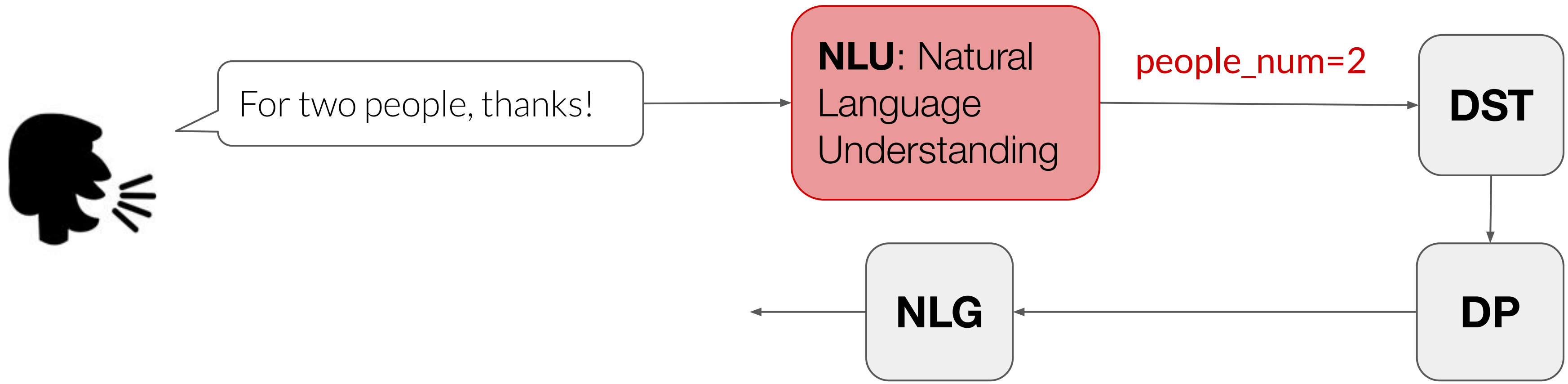
Siri
Hey Cortana
alexa

Hi, how can I help?

Modularized Task-Oriented Dialogue Systems



Natural Language Understanding (NLU)



NLU is a *turn-level* task that maps utterances to semantics frames.

- Input: raw user utterance
- Output: semantic frame (e.g. speech-act, intent, slots)

NLU - Approaches

Utterance:	For	2	people	thanks
	↓	↓	↓	↓
Slot:	O	B-people	O	O

Domain: Hotel

Intent: Hotel_Book

⇒ Slot-F1

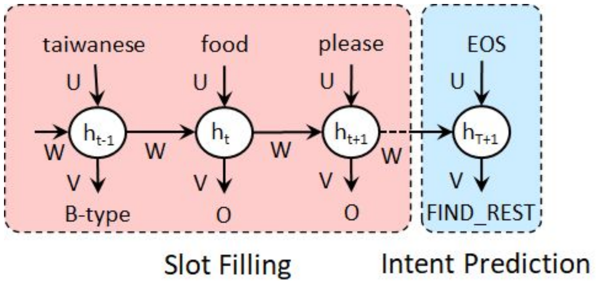
} ⇒ Acc

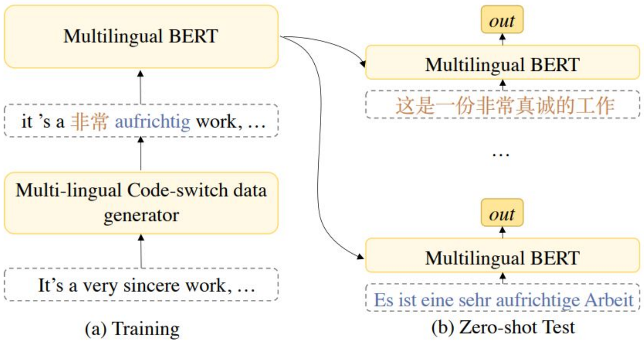
⇒ Frame Accuracy

- Domain/Intent Detection ⇒ Classification Task
  - CNN (Kim, 2014; Zhang+, 2015), LSTM (Ravuri & Stolcke, 2015), attention models (Zhao & Wu, 2016; Yang+, 2016)
- Slot Tagging ⇒ Sequence Labelling (IOB; Inside-Outside-Beginning format)
  - CNN (Vu, 2016), LSTM (Yao+, 2014; Kurata+, 2016), RNNEM (Peng+, 2015), joint pointer (Zhao & Feng, 2018)

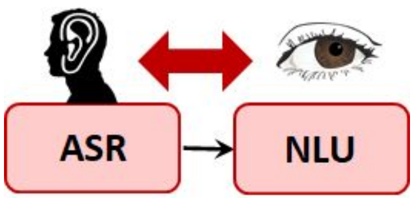
NLU - Trends & Challenges

- **Joint Intent / Slot Prediction**
  - CNNCRF (Xu & Sarikaya, 2013), RecNN (Guo+, 2014), joint RNN-LSTM (Hakkani-Tur+, 2016), attention-based RNN (Liu & Lane, 2016), slot-gated (Goo+, 2018), BERT (Chen+, 2019)
- **Better Scalability**
  - *Cross-lingual*: multilingual NLU (Schuster+, 2019, Liu+, 2019, Liu+, 2019, Li+, 2020, Qin+, 2020)
  - *Cross-domain*: zero-shot/few-shot fine-tuning on unseen domains (Bapna+, 2017, Shah+, 2020, Liu+, 2020)
  - *Unsupervised* NLU: (Su+, 2019, Su+, 2020; Namazifar+, 2020)
- **Better Robustness**
  - Spoken language understanding: (Huang & Chen, 2019; Huang & Chen, 2020, Liu+, 2020)

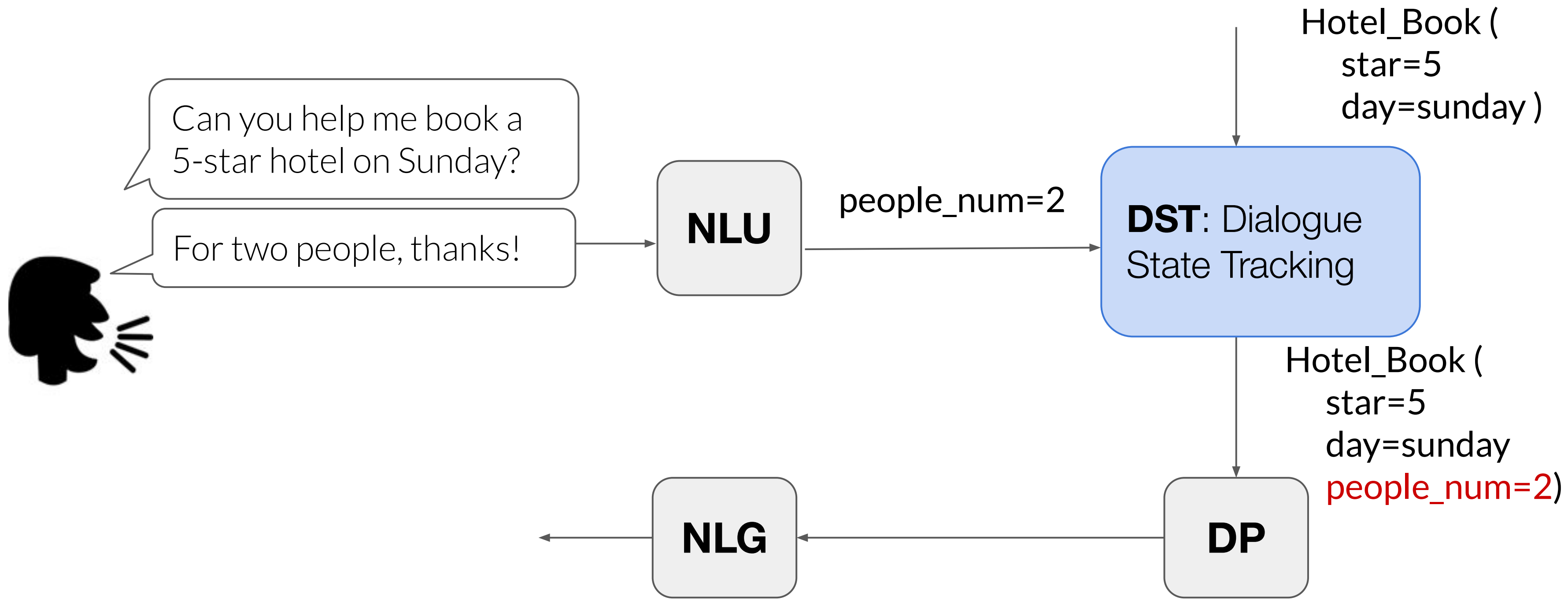




(a) Training
(b) Zero-shot Test



Dialogue State Tracking (DST)



DST is a *dialogue-level* task that maps partial dialogues into dialogue states.

- Input: a dialogue / a turn with its previous state
- Output: dialogue state (e.g. slot-value pairs)

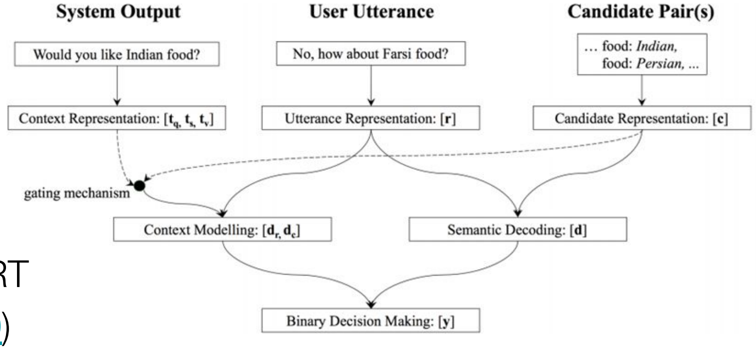
DST - Approaches

Input Dialogue:
USER: Can you help me book a 5-star hotel on Sunday?
SYSTEM: For how many people?
USER: For two people, thanks!

Output Dialogue State:
Hotel_Book (star=5, day=sunday)
Hotel_Book (star=5, day=sunday, people_num=2)

⇒ Slot Acc / Joint Acc

- **Rule-based**
  - Expert-designed rules (e.g., state update by adding slot values from NLU) ⇒ Regex/ WitAI
  - RNN based (Liao+, 2020)
- **Classification DST**: one classifier per slot
  ⇒ requires an ontology with predefined values
  - CNN (Mrksˇic´+, 2016), LSTM (Ramadan+, 2018), Context att (Nouri & Hosseini, 2018), Global2Local Att (Zhong+, 2018), Hierarchical LSTM (Goel+, 2019), BERT (Lee+, 2019; Wu+, 2020; Zhang+, 2019; Chen+, 2020)



DST - Trends & Challenges

- **Generation DST**
  - Generating the state as a sequence (Lei+, 2018) or dialogue state updates (Lin+, 2020)

    (Dialogue history) ⇒ (slot1=val,slot2=val …)

  - Given a dialogue and a slot, generate the value of the slot (Wu+, 2019; Gao+, 2019; Ren+, 2019; Zhou & Small, 2019; Kim+, 2019; Le+, 2020) ⇒ requires multiple forwards

    (Dialogue history, slot1) ⇒ val
- **Scalability**
  - Multi-Domain (Mosig+, 2020)

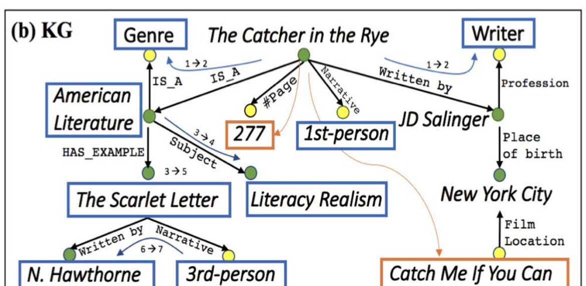
    MultiWoZ 2.0 ⇒ 2.1 ⇒ 2.2 ⇒ 2.3 ⇒ ……
  - Cross-Domain: zero-shot new-domains using natural language description

    SGD: schema-guided dialogue (Rastogi+, 2019)
  - Cross-Lingual: learning in English and zero-shot in other languages

CrossWOZ
usr: 你好，可以帮我推荐一个评分是4.5分以上的景点吗? Hello, could you recommend an attraction with a rating of 4.5 or higher? sys: 天安门城楼，簋街小吃和北京欢乐谷都是很不错的地方呢。 Tiananmen, Gui Street, and Beijing Happy Valley are very nice places. usr: 我喜欢北京欢乐谷，你知道这个景点周边的酒店都是什么吗? I like Beijing Happy Valley. What hotels are around this attraction? sys: 那可多了，有A酒店, B酒店, C酒店。 There are many, such as hotel A, hotel B, and hotel C. usr: 太好了，我正打算在**景点附近**找个酒店住宿呢，知道哪家评分是4分以上，提供叫醒服务的不? Great! I am planning to find a hotel to stay **near the attraction**. Which one has a rating of 4 or higher and offers wake-up call service?

DST - Trends & Challenges

- Other State Representations
  - Graph ⇒ connection between entities in the dialogue (Moon+, 2019)
  - Queries ⇒ SQL query as a dialogue state (Yu+, 2019)
  - Data-Flow ⇒ executable program as a state (Andreas+, 2020)



Q_2 : Which of those dorms have a TV lounge? **INFORM_SQL**

S_2 :
```
SELECT T1.dorm_name FROM dorm AS T1 JOIN has_amenity AS
T2 ON T1.dormid = T2.dormid JOIN dorm_amenity AS T3 ON
T2.amenid = T3.amenid WHERE T3.amenity_name = 'TV
Lounge'
```

A_2 : (Result table with many entries)

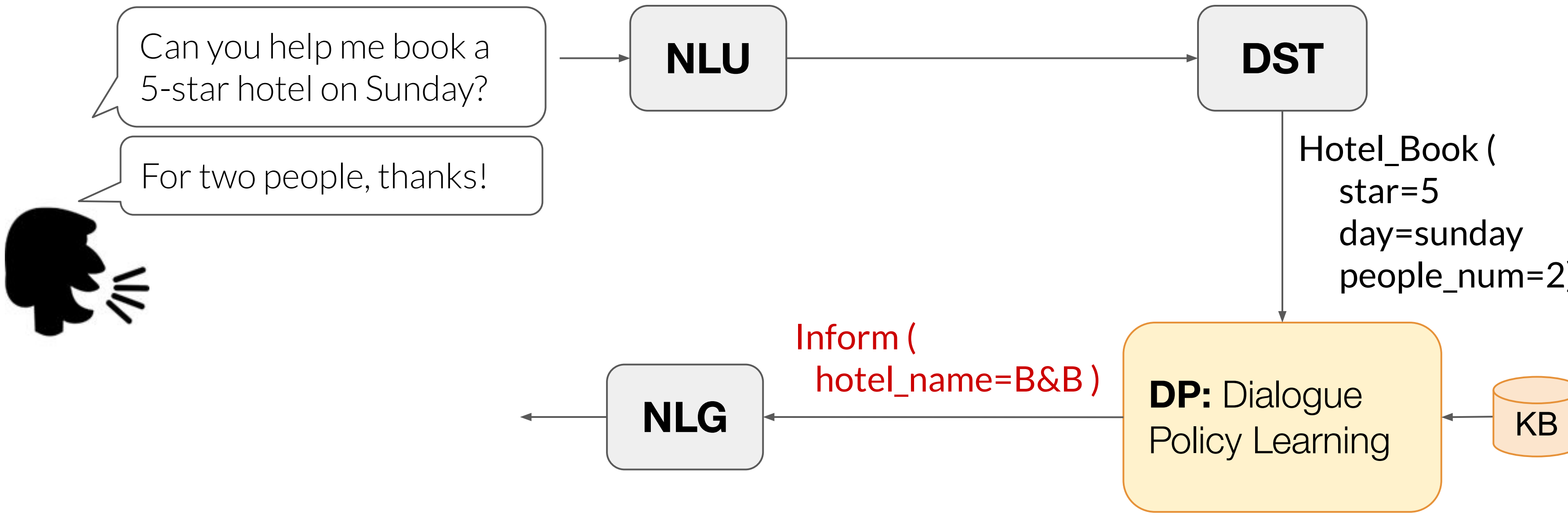
R_2 : This shows the names of dorms with TV lounges. **CONFIRM_SQL**

User: *Where is my meeting at 2 this afternoon?*

```
place(findEvent(EventSpec(start=pm(2))))
```

Agent: *It's in Conference Room D.*

Dialogue Policy Learning (DP)



DP decides the system action for interacting with users based on dialogue states.

- Input: dialogue state + KB results
- Output: system action (speech-act + slot-value pairs)

DP: Approaches

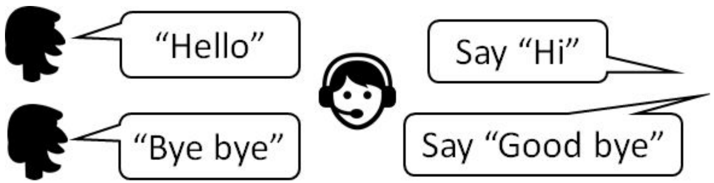
Dialogue State:
Hotel_Book (star=5, day=sunday, people_num=2)
KB State:
rest1=B&B

⇒

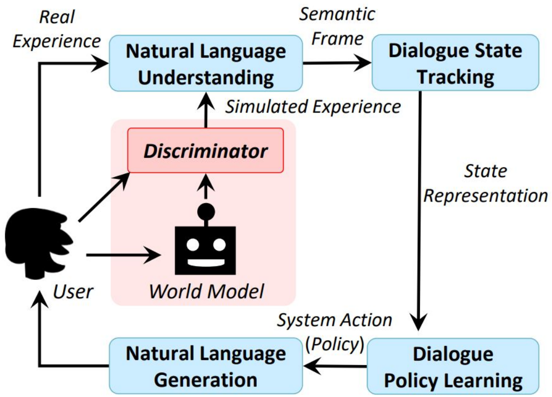
System Action:
inform (hotel_name=B&B)

- **Supervised Learning:** learning from the paired data in the corpus
- **Reinforcement Learning**: learning from the interaction with the user (simulator)

⇒ Task Success Rate / Dialogue Length

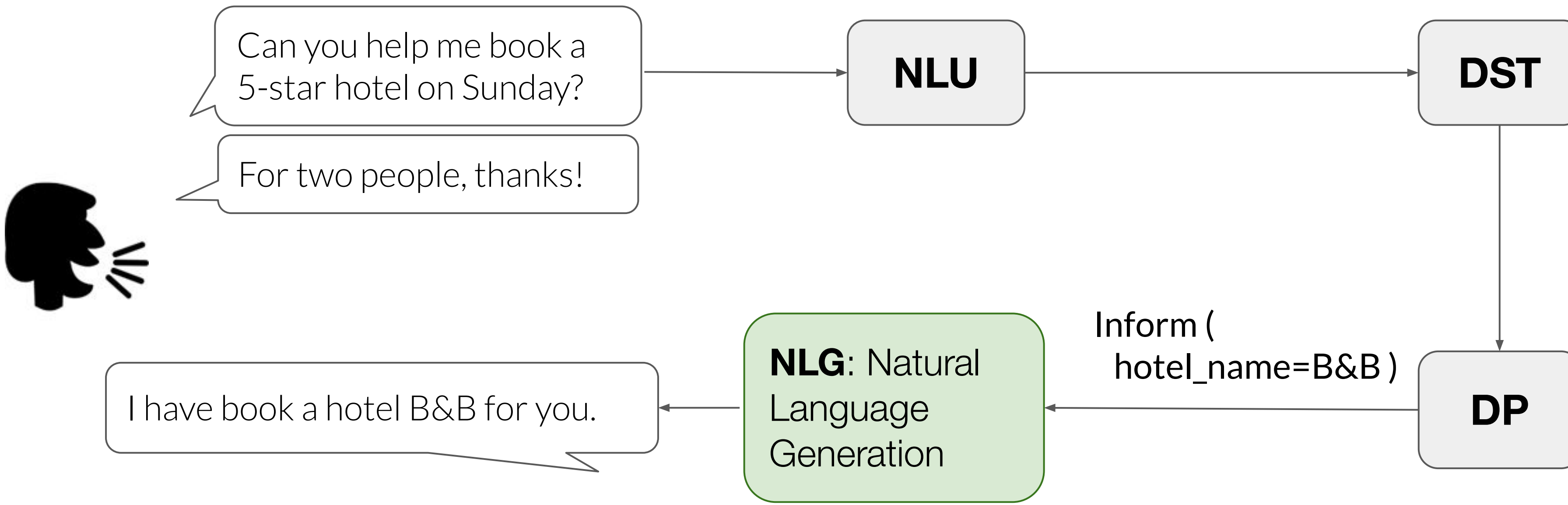


DP: Trends & Challenges



- RL for DP
  - E2E dialogue learning (Li+, 2017)
  - Interactive reinforcement learning (Shah+, 2017, Liu+, 2017)
  - Learning with real users (Liu+, 2018)+ planning (Peng+, 2018), more robust (Su+, 2018)
  - Hierarchical policy (Budzianowski+, 2017, Peng+, 2017)
  - Action embedding (Mendez+, 2019), meta-dialogue policy (Xu+, 2020)
- User Simulator ⇒ very important for RL-based agents
  - Agenda-based (Schatzmann+, 2007), reward shaping (Takanobu+, 2019) and more….
- Learning a dialogue policy using few well-annotated dialogues
  - Meta-dialogue policy (Xu+, 2020)
  - Neural program synthesis for dialogues (Zhou+, 2020) ⇒ generate code for the policy, instead of the policy it-self

Natural Language Generation (NLG)



NLG is to map system actions to natural language responses.

- **Input**: system speech-act + slot-value (optional)
- **Output**: natural language response

NLG: Approaches

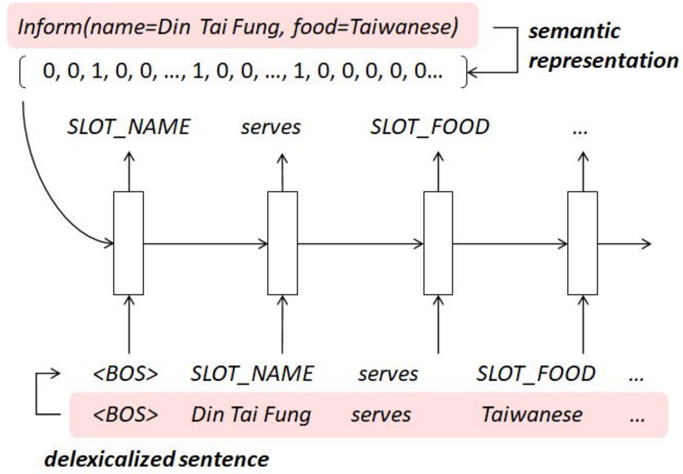
System Action	⇨	**System Response**
inform(name=B&B)		I have book a hotel B&B for you.

⇒ BLEU / Slot Error Rate

- Template-Based

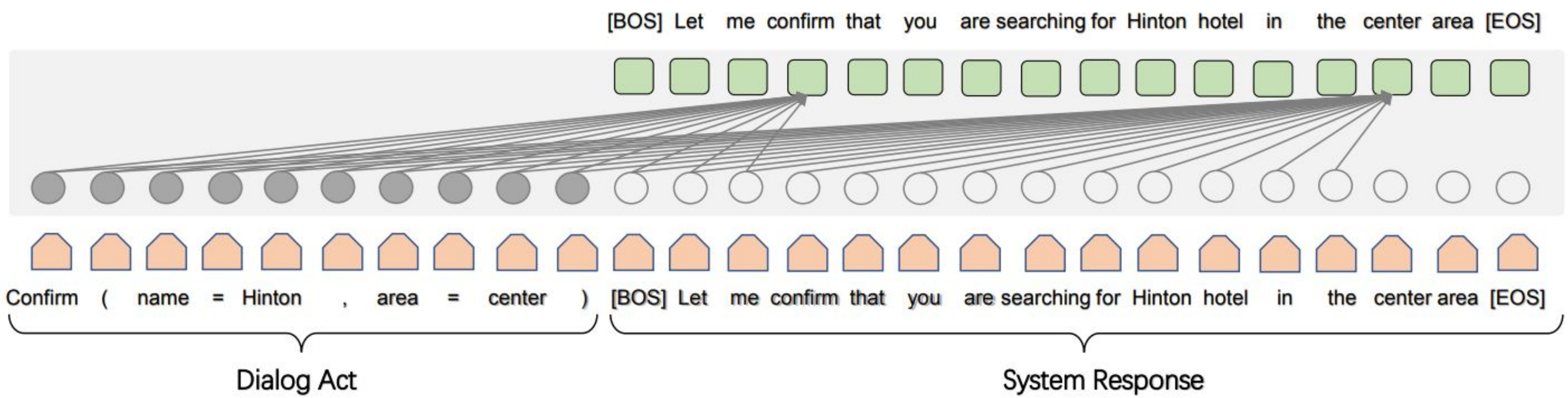
  inform(name=$A, phone=$B) ⇒ I found $A and their phone number is $B

- Generation-Based
  - SC-LSTM (Wen+, 2015, Mei+, 2016)
  - Seq2Seq (Tran+, 2017), + tree (Dusek & Jurcıcek, 2016)
  - Structural NLG (Sharma+, 2017, Nayak+, 2017)
  - Hierarchical Decoding (Su+. 2018; Su & Chen, 2018)
  - Controllable NLG (Hu+, 2017) + style (Shu+, 2020)
  - Datasets (Novikova+, 2016), challenge (Novikova+, 2017)
  - Challenge (Dusek+, 2018) + SOTA NLG (Dusek+, 2019)
- Hybrid: Template + Generation
  - Rewriting simple templates (Kale+, 2020)

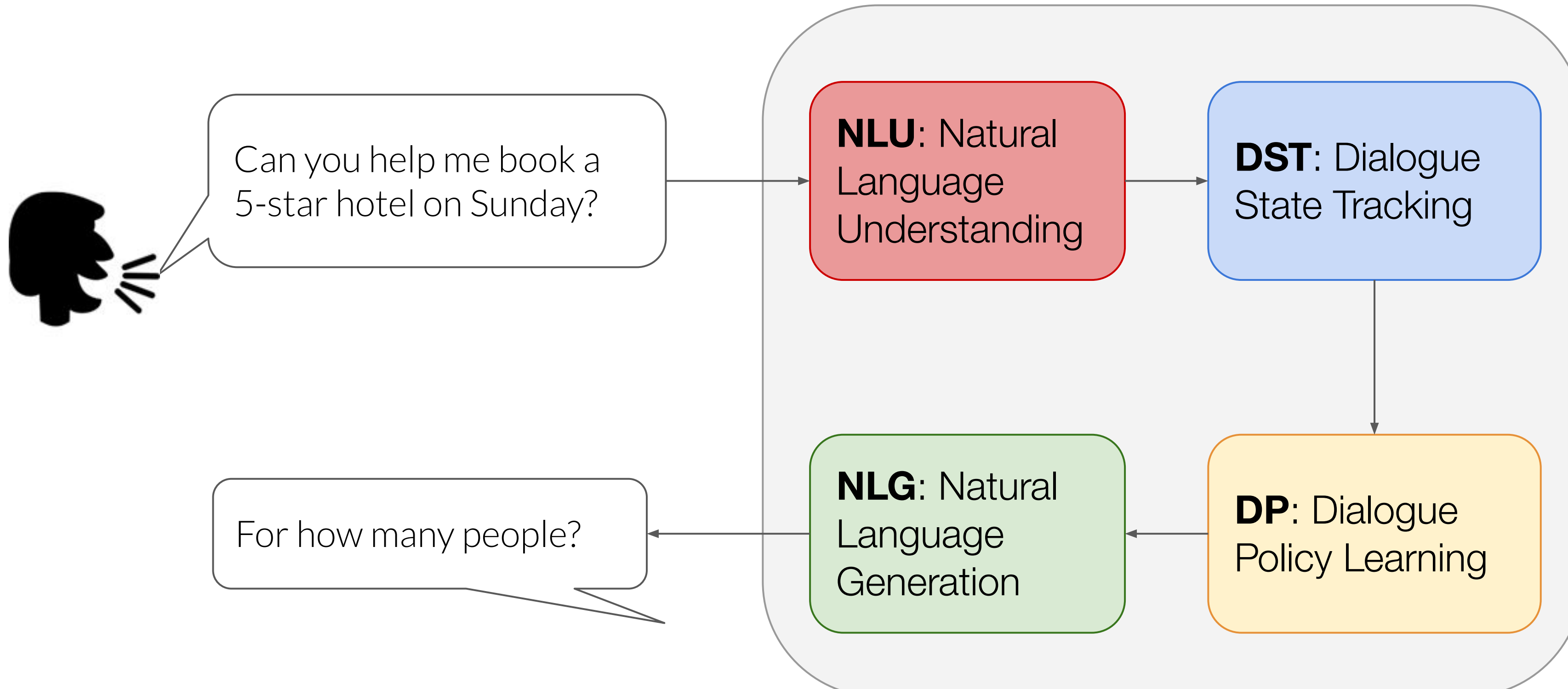


NLG: Trends & Challenges

- Scalability
  - Few-shot domain learning for NLG (Peng+, 2020)
  - Unsupervised NLG (Su+, 2019, Su+, 2020)



Modularized Task-Oriented Dialogue Systems



(Part 1)
Conversational AI Overview

1.1. Brief History of Conversational AI

1.2. Modularized Task-Oriented Dialogue Systems

1.3. Retrieval-Based Chit-Chat Dialogue Systems

Generation-Based Conv. AI

Challenges and Future Work

Chit-Chat Dialogue Systems

- Designed for **free-form** and **open-domain** conversations
- Aims at **engaging users** for a long conversations
- Rare to deal with APIs or knowledge
- Two types:
  - Retrieval-based
  - Generation-based (covered in the next section)

Retrieval-Based Chatbots

- Task: learning a scoring function between dialogue history and response candidates

$$Score = f(v, u)$$

dialogue history vector   response candidate vector

- PolyEncoder
  - Pre-trained on Ubuntu+Reddit+Persona-Chat
- Blended-Skill-Talk
  - Dialogue manager to choose the retriever

★ Pros: safer response
  - predefined candidates, fluent language

★ Cons: poor scalability
  - millions candidates
  - no suitable candidate in new domains

★ Viable Solution: generation-based models

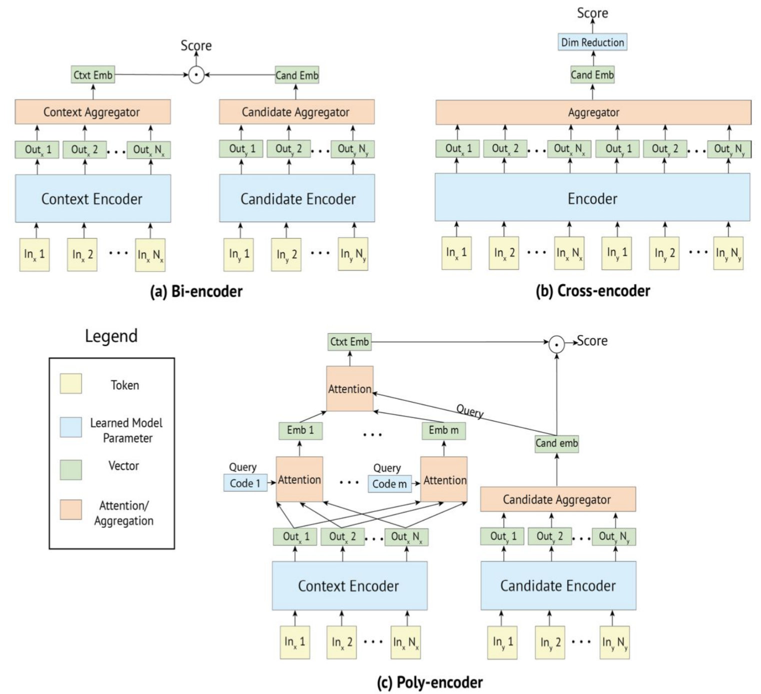


Image from Poly-encoder

(Part 2) Generation Based Deep Conversational AI

Pascale Fung, Zhaojiang Lin, Andrea Madotto



(Part 2) Generation Based Deep Conversational AI

2.1. Vanilla Seq2Seq ConvAI

2.2. Limitations in Vanilla ConvAI

2.3. Deeper ConvAI Solutions

(Part 2) Generation Based Deep Conversational AI

2.1. Vanilla Seq2Seq ConvAI

2.2. Limitations in Vanilla ConvAI

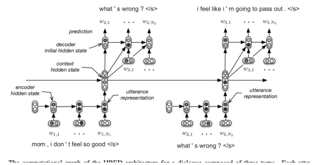
2.3. Deeper ConvAI Solutions

Conversational AI Overview

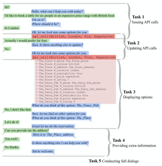
Challenges and Future Work

History of Neural Conversational-AI Research

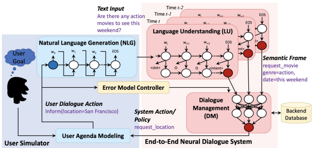
Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models (Serban 2016)



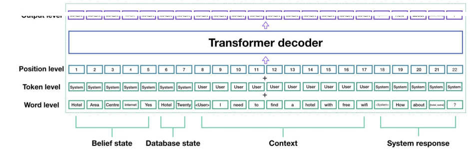
Learning end-to-end goal-oriented dialog (Bordes et.al., 2016)



End-to-End Task-Completion N Dialogue Systems (2017)

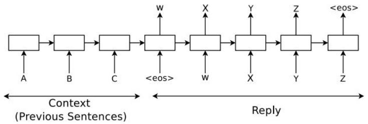


Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems (Paweł Budzianowski et.al. 2019)



A Neural Conversational Model (Vinyals et. al. 2015)

Neural Responding Machine for Short-Text Conversation (Shang et. al. 2015)

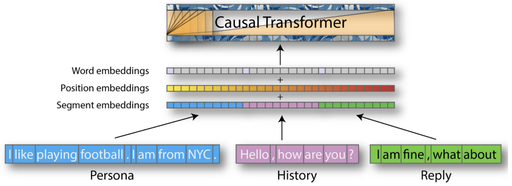


A Persona-Based Neural Conversation Model (Li et.al. 2016)

Deep Reinforcement Learning for Dialogue Generation (Li et.al. 2016)

Personalizing Dialogue Agents: I have a dog, do you have pets too? (Zhang et.al., 2018)

TransferTransfo: {A} Transfer Learning Approach for Neural Network Based Conversational Agents (Wolf et.al. 2019)



2.1 Vanilla Seq2Seq ConvAI: How

A simple 4 steps recipe:

1. Choose the data: Human to human conversations
2. Choose the model: Large pre-trained language models are preferable
3. Train the model with the data: Supervised learning
4. Evaluate your model: Automatic or human evaluation

2.1 Vanilla Seq2Seq ConvAI: Datasets

Human1: Ok, I'll try that.

Human2: Is there anything else bothering you?

Human1: Just one more thing. A school called me this morning to see if I could teach a few classes this weekend and I don't know what to do.

Human2: Do you have any other plan this weekend?

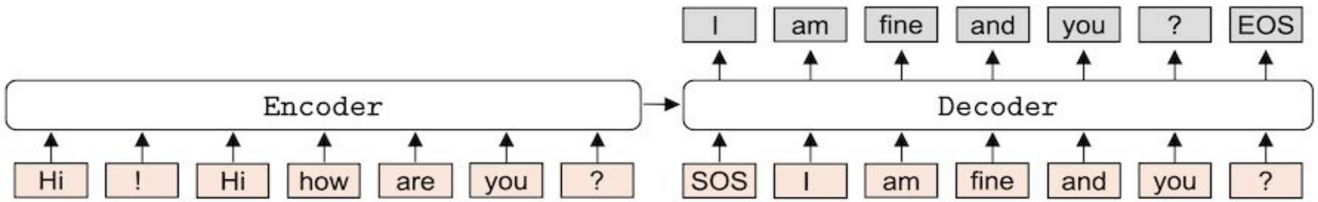
Human1: I'm supposed to work on a paper that's due on Monday.

Human-to-Human Conversations:

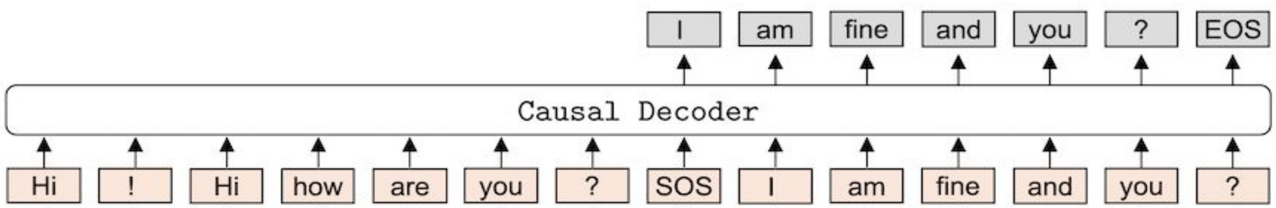
- Daily Dialog
- Ubuntu Dialogue Corpus
- Twitter Conversations
- Reddit Conversational Data
- OpenSubtitles

These datasets are pre-processed to have only 2 speakers ⇒ usually no more than 2 turns

2.1 Vanilla Seq2Seq ConvAI: Models

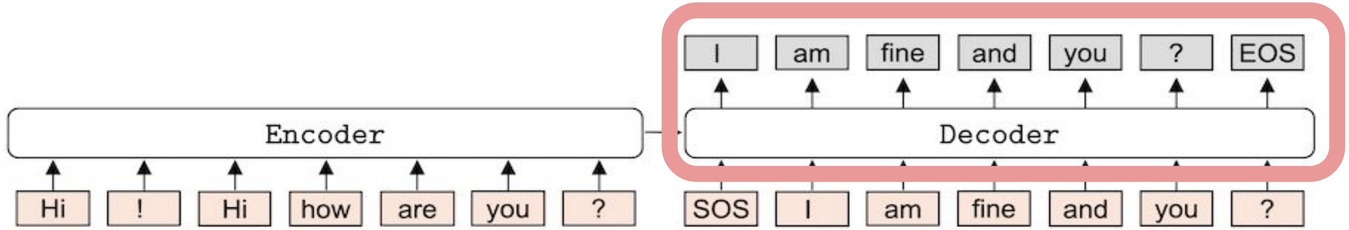


Vanilla Seq2Seq conversational model (Vinyals and Le et.al., 2015, Shang et al., 2015)

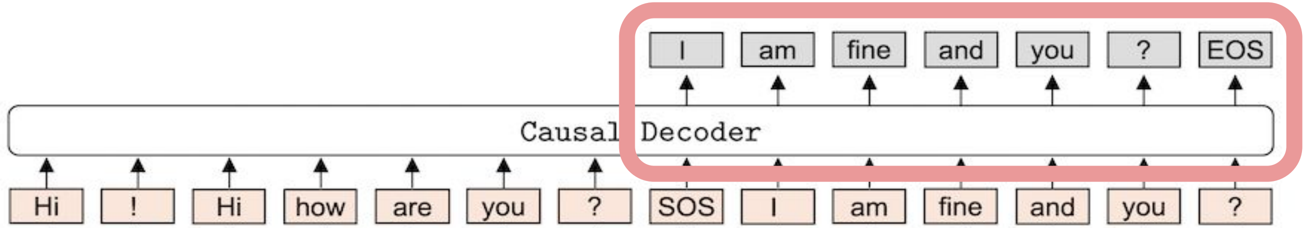


Causal Decoder (Wolf et.al. 2019 ,Radford et.al. 2018)

2.1 Vanilla Seq2Seq ConvAI: Models



Vanilla Seq2Seq conversational model (Vinyals and Le et.al., 2015, Shang et al., 2015)



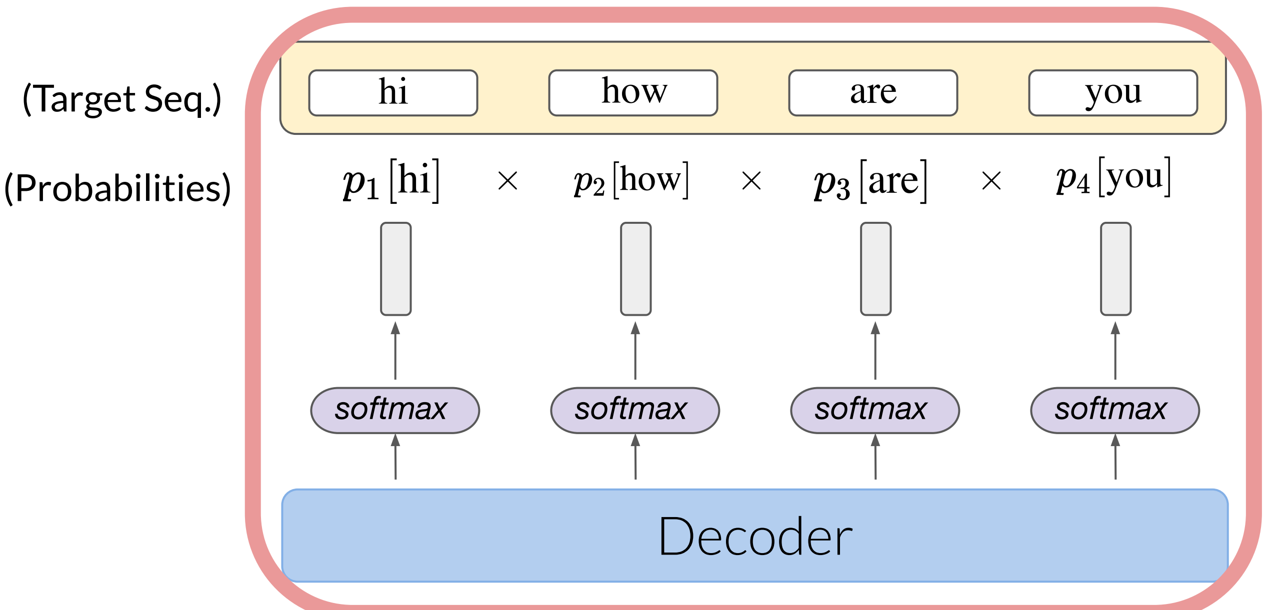
Causal Decoder (Wolf et.al. 2019 ,Radford et.al. 2018)

2.1 Vanilla Seq2Seq ConvAI: Supervised Learning

Maximum Likelihood Estimation (MLE):

⇒ Maximizing the conditional probability of the response given the dialogue history.

⇒ The output of conversational model is a probability distribution over the vocab.

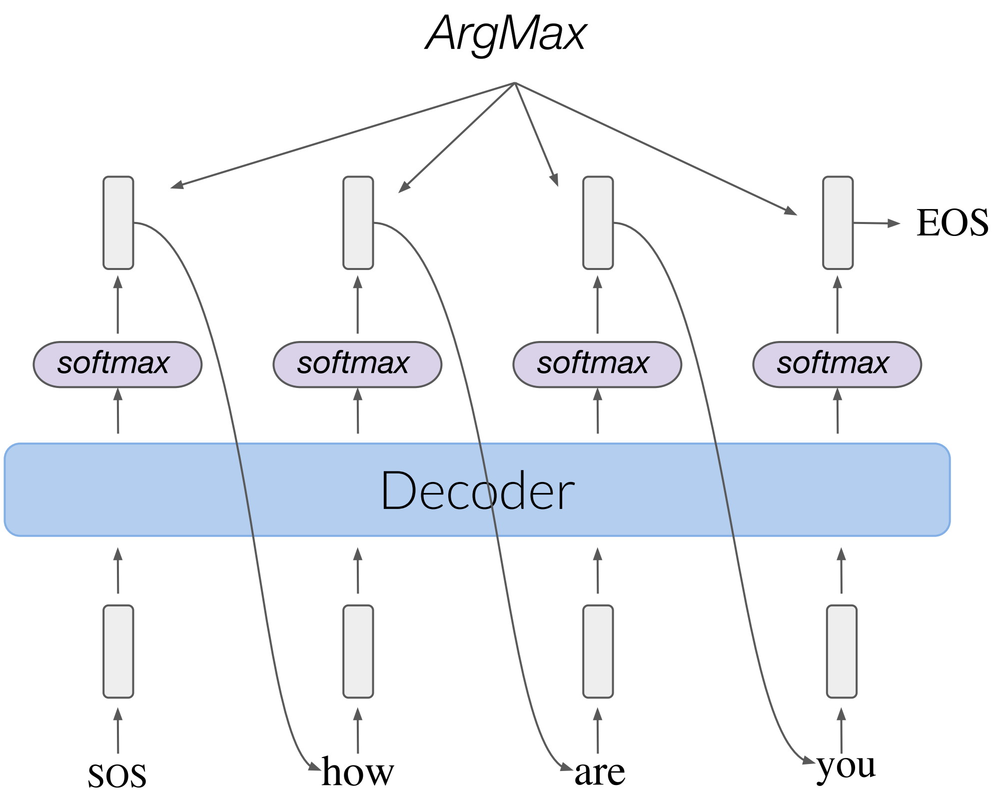


$$\mathbb{P}(y_1, \ldots, y_m) = \prod_{i=1}^{m} p_i[y_i]$$

$$-\log \mathbb{P}(y_1, \ldots, y_m) = -\log \prod_{i=1}^{m} p_i[y_i]$$

$$= -\sum_{i=1}^{n} \log p_i[y_i]$$

2.1 Vanilla Seq2Seq ConvAI: Greedy Decoding

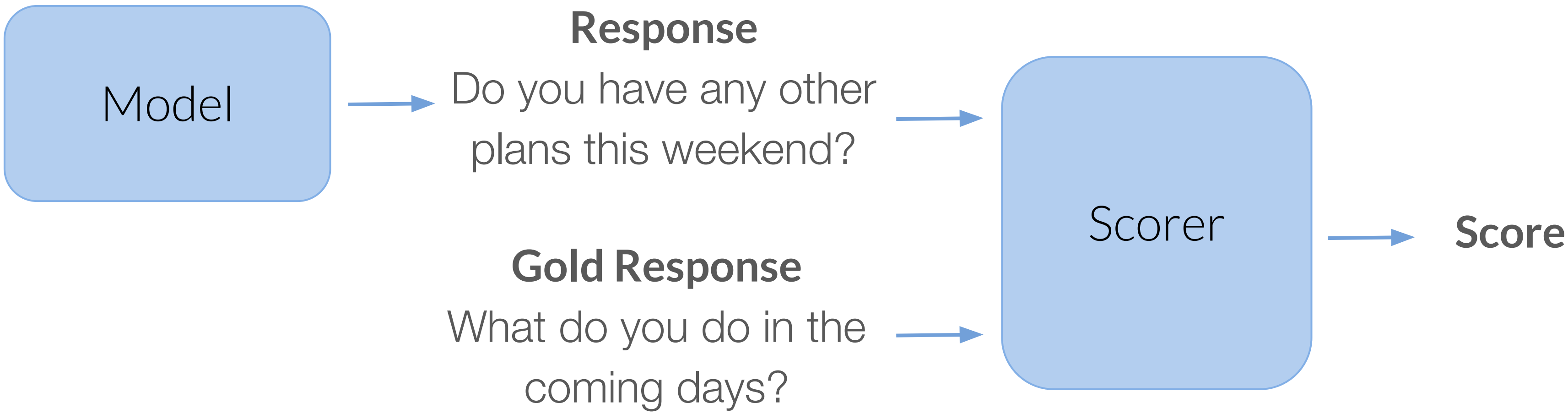


- Starts with a special token SOS
- Forward the model to generate a distribution over the vocabulary ⇒ Argmax to generate a token
- Provide the generated token to the next step
- Repeat until the model generate the EOS token

2.1 Vanilla Seq2Seq ConvAI: Automatic Evaluation

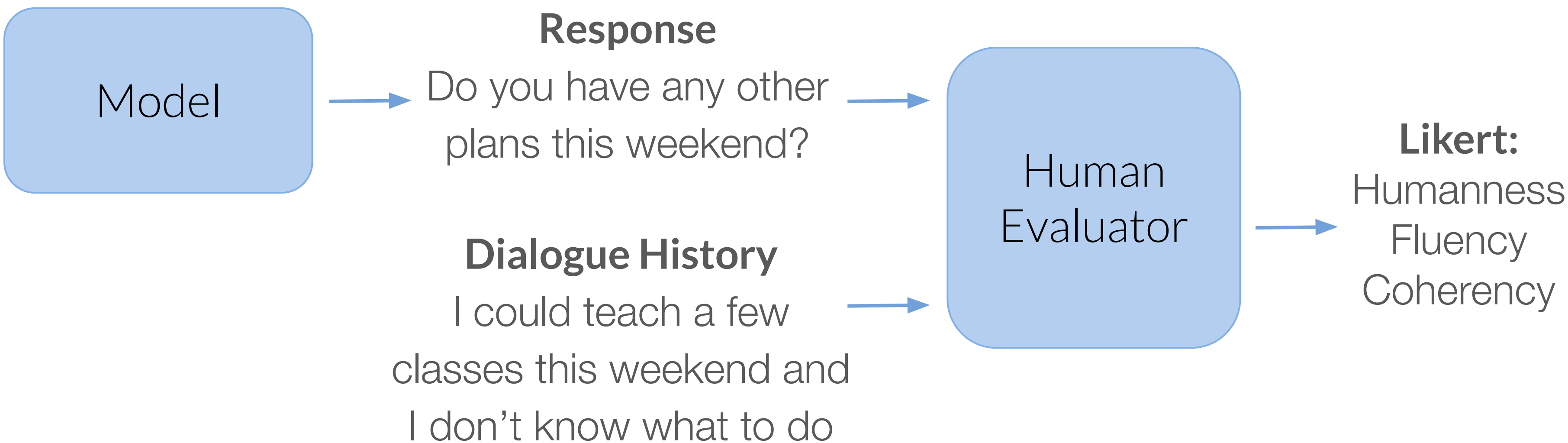
Use the gold reference response to compute a score:

- Perplexity ⇒ how likely the model is to generate the gold response
- N-gram overlapping ⇒ BLEU etc.
- Distinct N-grams ⇒ response diversity



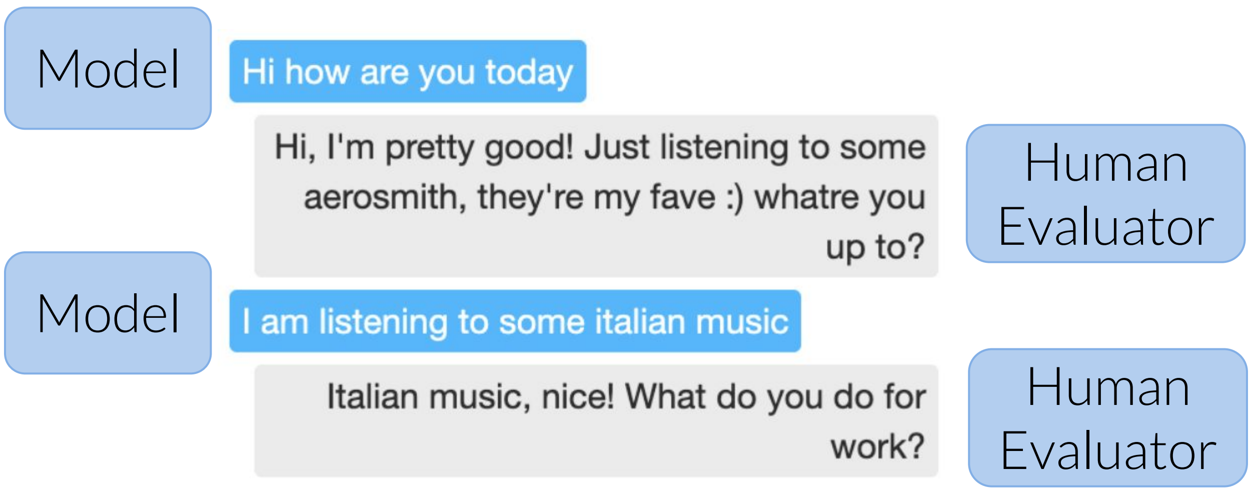
2.1 Vanilla Seq2Seq ConvAI: Human Evaluation Likert

Show human judge the dialogue history, gold response and the generated response, and ask the judge to give ratings 0-5 according to “Humanness, Fluency and Coherence”



2.1 Vanilla Seq2Seq ConvAI: Human Evaluation Dynamic Likert

Show human judge interact with the model and ask the judge to give ratings 0-5 according to “Humanness, Fluency and Coherence”



After conversation

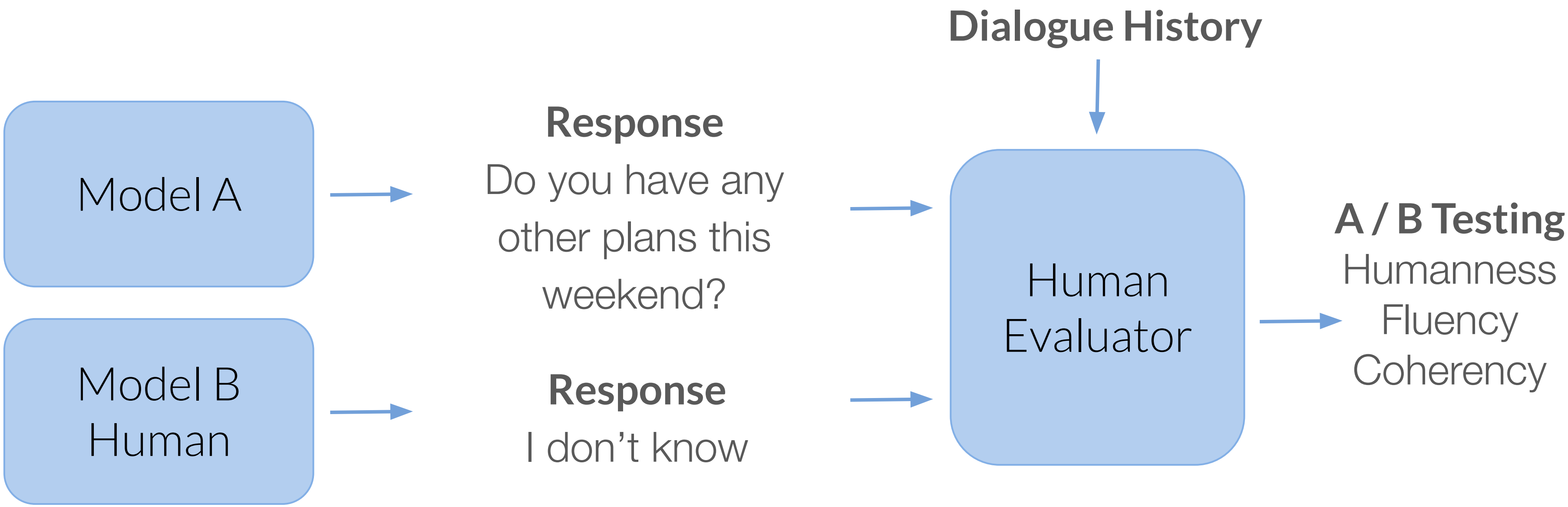
Human Evaluator →

Likert:
Humanness
Fluency
Coherency

Figure from: ACUTE-EVAL (Li et.al. 2019)

2.1 Vanilla Seq2Seq ConvAI: Human Evaluation A/B

Show human judge the dialogue history, gold response and the generated response, and ask the judge to choose one of the three according to "Humanness, Fluency and Coherence"



2.1 Vanilla Seq2Seq ConvAI: Human Evaluation A/B Dynamic

- Ask a human to interact with two systems
- Ask the judge to choose which of the response is better according to "Humanness, Fluency and Coherence"
- This is dialogue-level evaluation

ACUTE-EVAL (Li et.al. 2019)

(Part 2) Generation Based Deep Conversational AI

2.2 Limitation : Lack of Diversity

Human: What are you doing today?

System: I don’t know

Human: What is your name?

System: I don’t know

Human: What is your favourite food?

System: I don’t know

Vanilla seq2seq that was trained on limited datasets with MLE loss might generate repetitive responses.

2.2 Limitation: Lack of Consistency

Human: Where were you born?

System: I was born in Canada.

Human: Where are you from?

System: England, you?

Human: Where did you grow up?

System: I grew up in Texas.

Inconsistent responses generated by a 4-layer Seq2Seq model trained on 25 million Twitter conversation snippets. Examples from Li et al. (2016a).

2.2 Limitation: Lack of Knowledge

Human: What is the weather like today?

System: I don't know, maybe sunny?

Human: I am going out to Taylor Swift's concert today, I hope the weather is good.

System: Who is Taylor Swift?

Human: Okay, never mind.

Conversational models without external knowledge often fail to conduct engaging conversations.

2.2 Limitation: Lack of Empathy

Human: I messed up my project today.

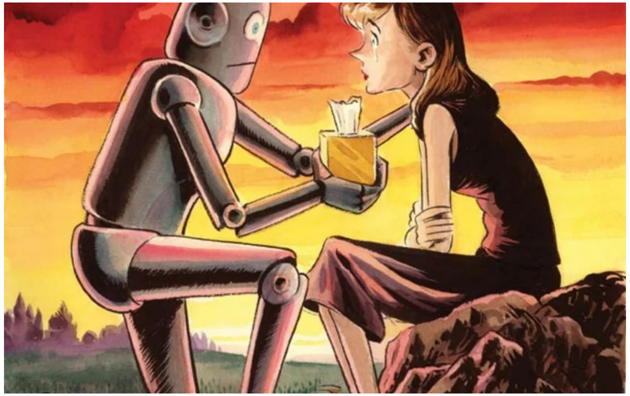
System: What is your project?

Human: I think I might get fired...

System: Then you need to find another job!

Human: 😩

Conversational models without empathy might give the user a bad experience.



2.2 Limitation: Lack of Controllability

Human: Hi, how are you?

Model1: I am good thanks

Model2: I had really a bad day

Model3: I am okay, how was your day?

Model4: I am okay, I just finished my training session in the swimming pool

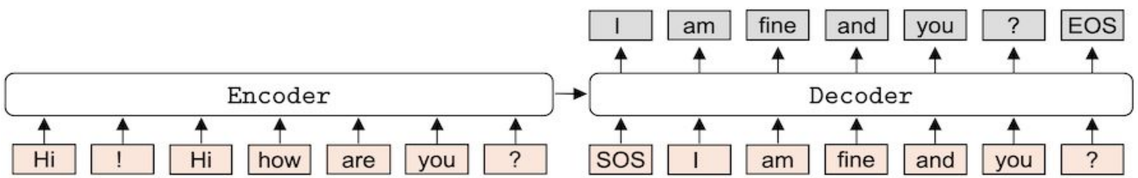
Vanilla models do not have any mechanism to control for:

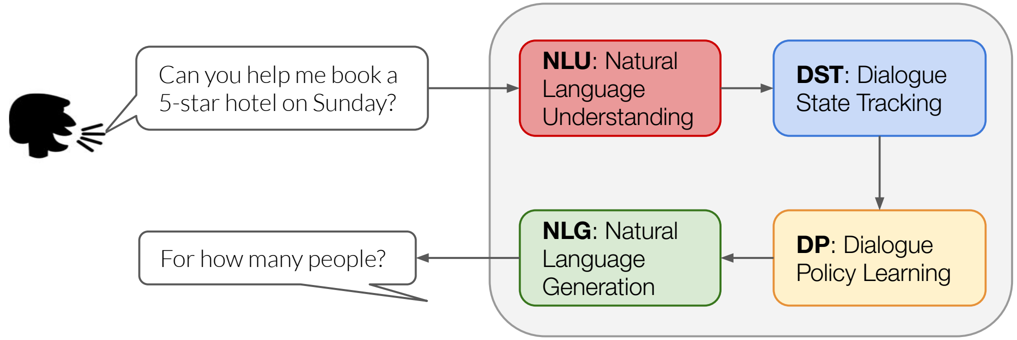
- Response style
- Topics
- Engagement

More importantly:

- Toxic and inappropriate responses

2.2 Limitation: Lack of versatility





- Seq2seq models and modularised task-oriented dialogue system lives in separate worlds
- Seq2seq trained with vanilla data cannot handle task-oriented conversations
- Requires API-Generation

2.2 Limitations of Vanilla Seq2Seq: Summary

1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. Lack of versatility

These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards <u>deeper conversational AI</u>?

(Part 2) Generation Based Deep Conversational AI

2.1. Baseline Vanilla Seq2Seq ConvAI

2.2. Limitations in Vanilla Seq2Seq ConvAI

2.3. Deeper ConvAI Solutions

Conversational AI Overview

Challenges and Future Work

2.2 Limitations of Vanilla Seq2Seq: Summary

1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. Lack of versatility

These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards deeper conversational AI?

2.3 Deeper ConvAI Solution: Diversify Responses

1. Training and Decoding strategy ⇒ Maximum Mutual Information (MMI);

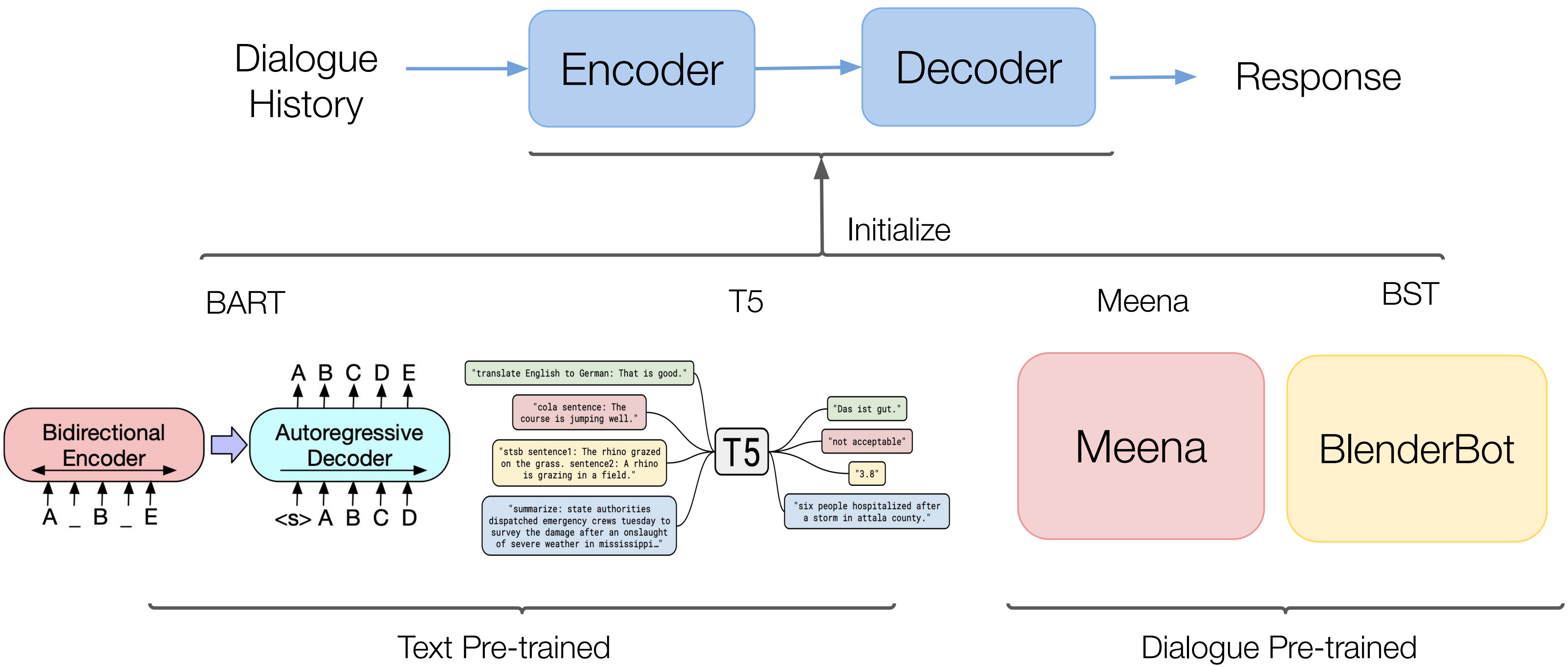
$$\hat{T} = \arg\max_{T} \{ \log p(T|S) \} \Rightarrow \arg\max_{T} \{ (1-\lambda) \log p(T|S) + \lambda \log p(S|T) \}$$

2. Model architecture ⇒ Conditional Variational Autoencoder (CVAE);

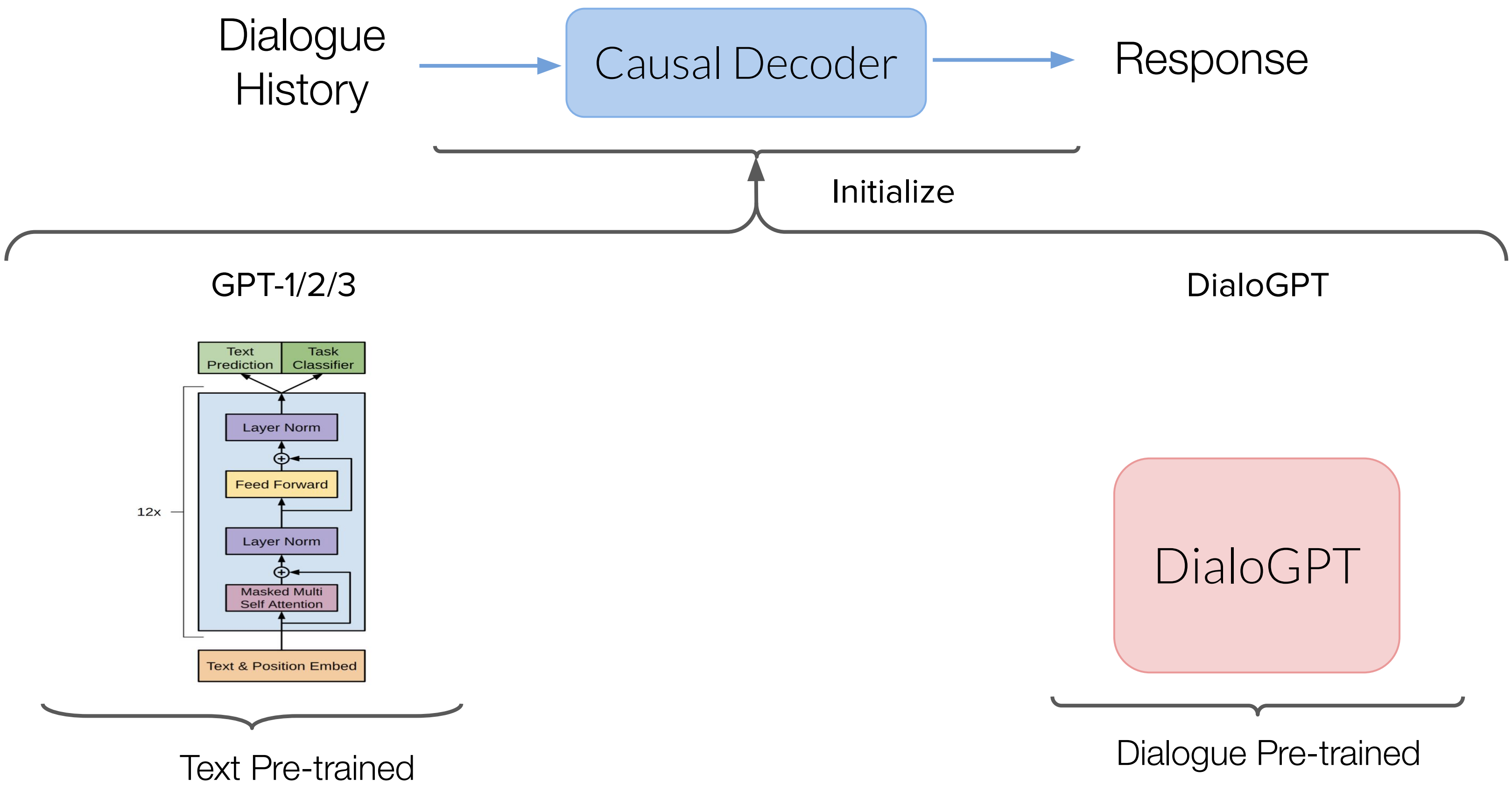
$$p(T|S) => p(T|z, S) p(z|S)$$

3. More data & Larger models ⇒ Large scale pre-training; (NEXT SLIDES)
4. Decoding strategy ⇒ Top-k sampling, Nucleus Sampling; (NEXT SLIDES)

2.3 Deeper ConvAI Solution: Diversify by large scale pretraining



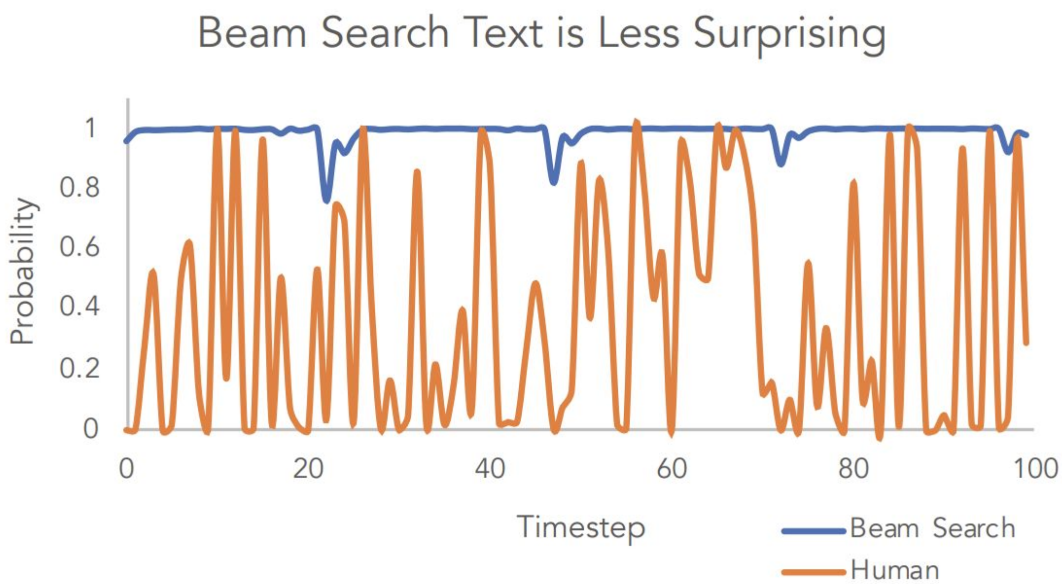
2.3 Deeper ConvAI Solution: Diversify by large scale pretraining



2.3 Deeper ConvAI Solution: Diversify by Nucleus Sampling

- Compared to beam search, human are more likely to sample "low probability" tokens.
- *Nucleus Sampling* try to recover the human sampling process by sampling from top-N vocabulary $V^{(p)} \subset V$.

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p.$$



Ref: The Curious Case of Neural Text Degeneration

2.3 Deeper ConvAI Solution: Diversify by Nucleus Sampling

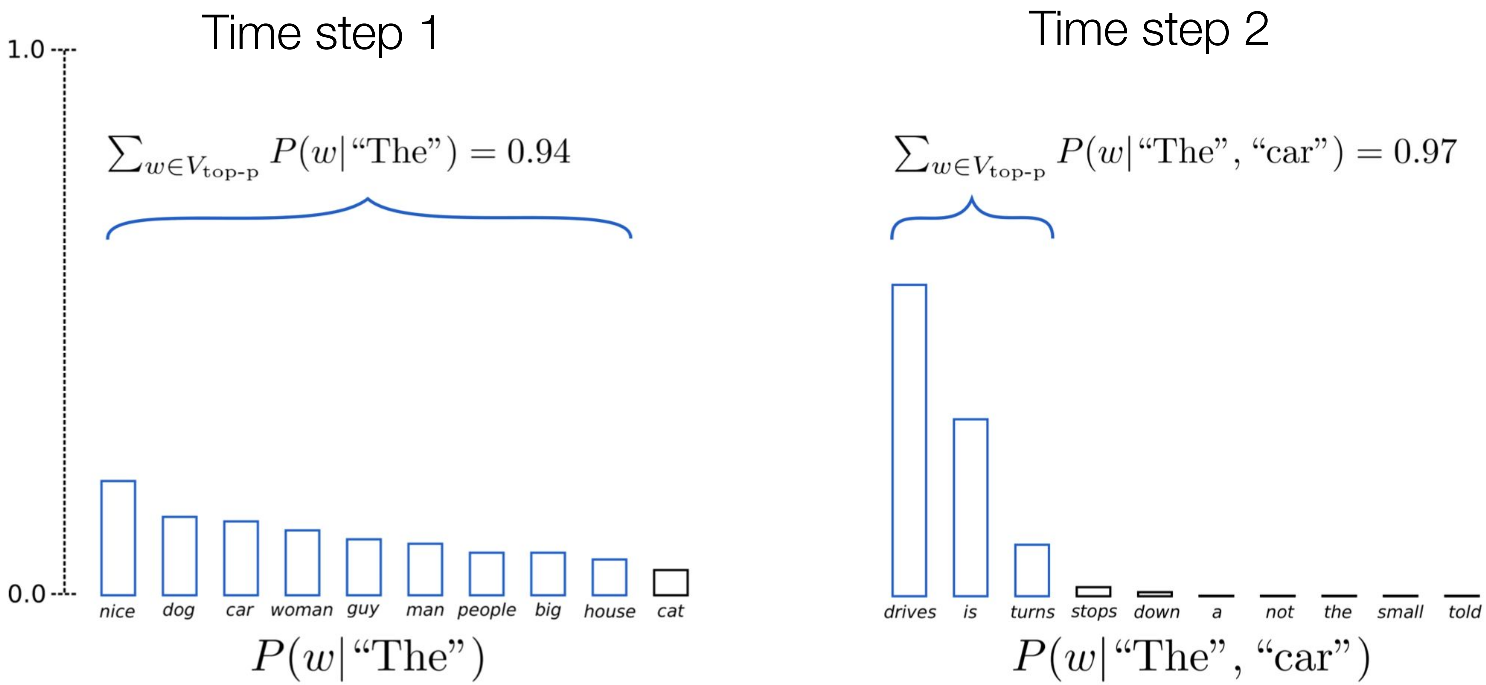


Figure from: https://huggingface.co/blog/how-to-generate

2.2 Limitations of Vanilla Seq2Seq: Summary

1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. Lack of versatility

These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards deeper conversational AI?

2.3 Deeper ConvAI Solution: Personalization

1. Learning speaker embedding:
   a. Speaker Model

2. Conditioning on persona descriptions:
   a. PersonaChat Dataset
   b. TransferTransfo Model

2.3 Deeper ConvAI Solution: Personalization Datasets

Persona Info Human2:
- I like to ski.
- I am 25 years old

Human1: Hi, what do you do in your free time?

Human2: I enjoy going to the mountain and skiing

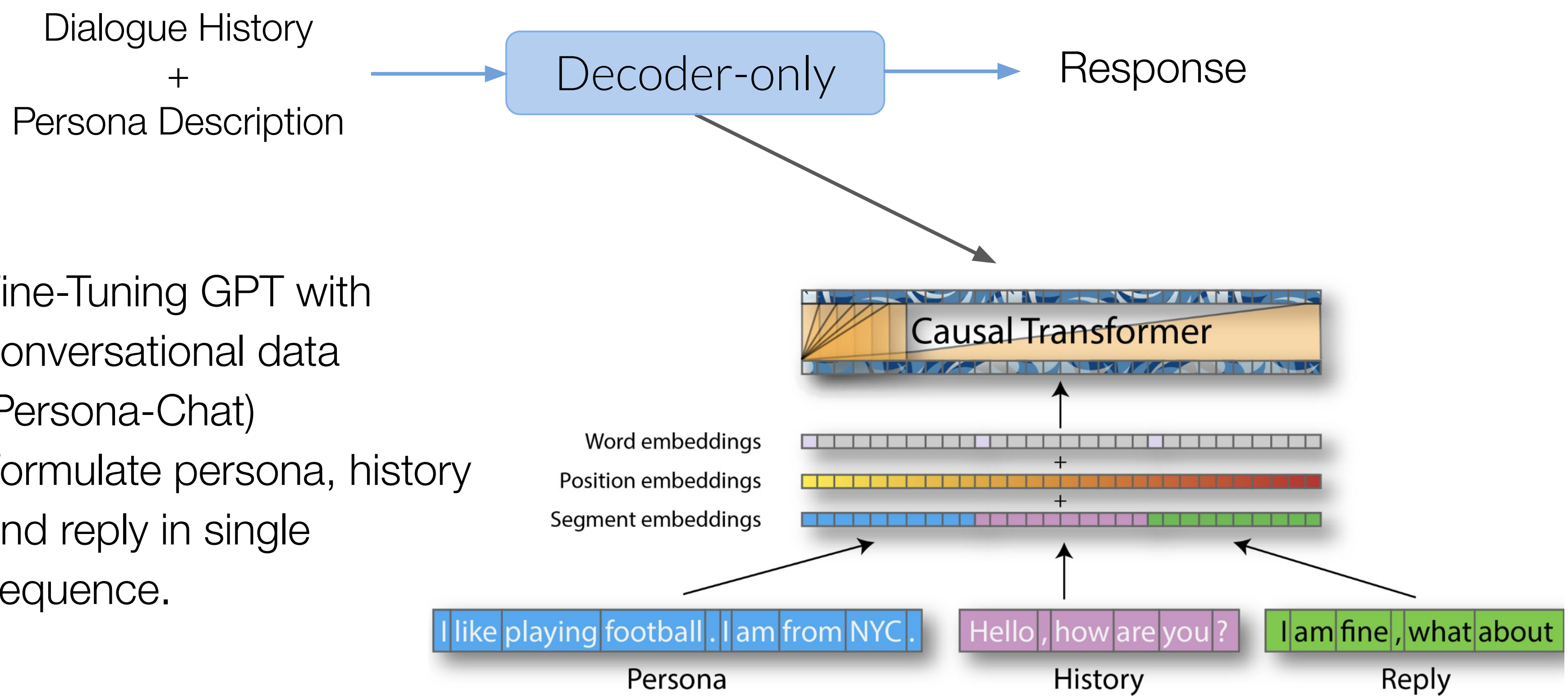
Human1: That's cool, you should be young and strong for this activity!

Human2: oh yeah, I am 25 🤗

Human-to-Human Conversations + Persona Features

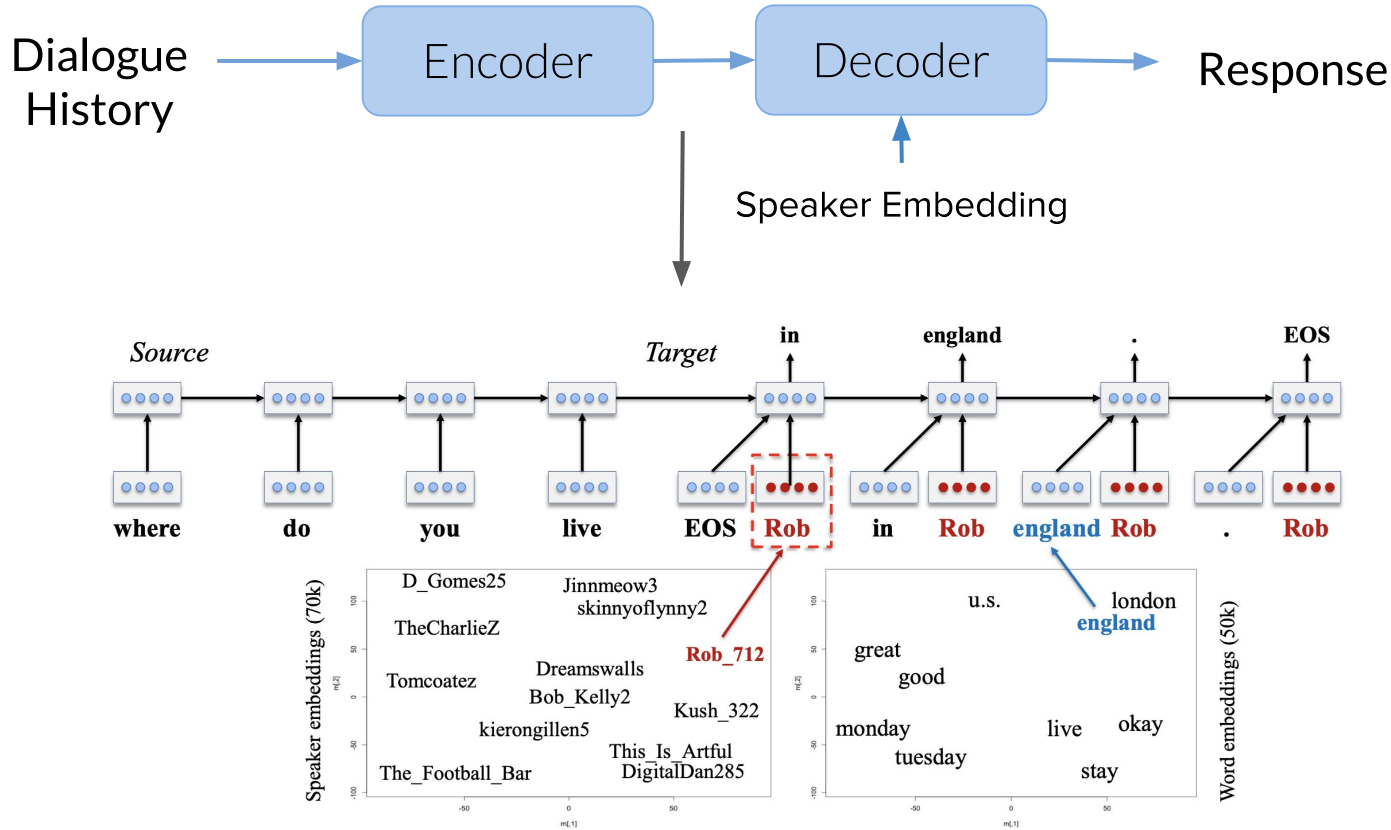
- Persona Chat
- Tweeter Personalized
- Learning Personalized End-to-End Goal-Oriented Dialog

2.3 Deeper ConvAI Solution: Personalization via TransferTransfo Model



- Fine-Tuning GPT with conversational data (Persona-Chat)
- Formulate persona, history and reply in single sequence.

2.3 Deeper ConvAI Solution: Personalization via Speaker Model



Speaker Model

2.2 Limitations of Vanilla Seq2Seq: Summary

1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. Lack of versatility

These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards deeper conversational AI?

2.3 Deeper ConvAI Solution: Knowledge

1. Textual Knowledge ⇒ Retrieving knowledge from Wikipedia, news, etc.;
2. Graph Knowledge ⇒ Retrieving subgraph from knowledge graphs;
3. Tabular Knowledge ⇒ Incorporate tabular information;
4. Service API Interaction ⇒ Generates API query, and incorporate API returns into the response.

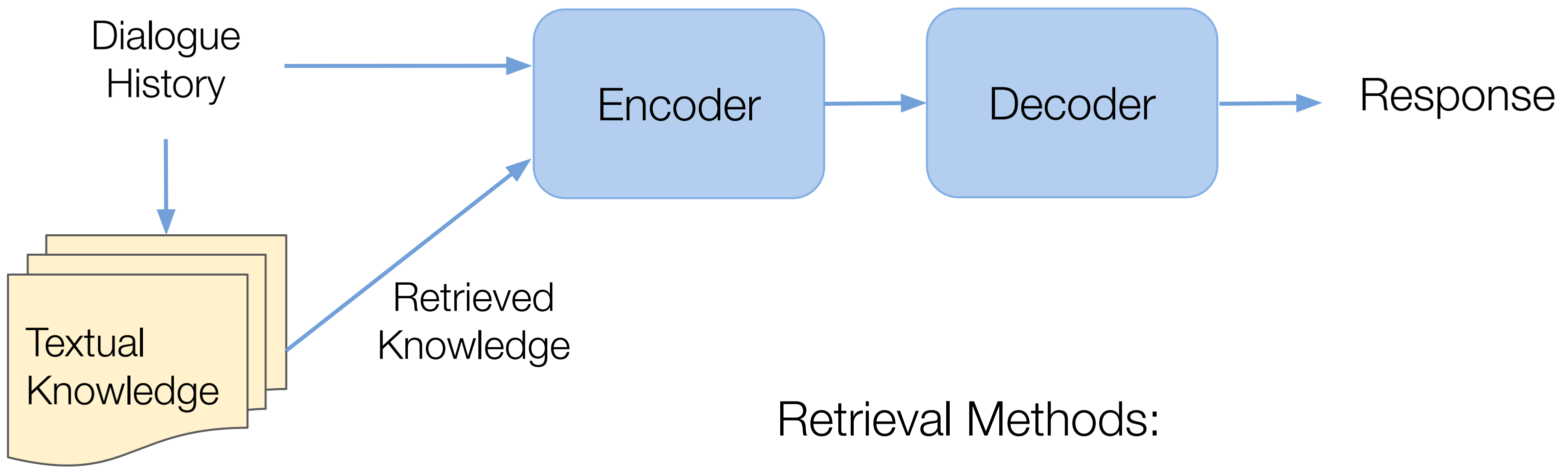
2.3 Deeper ConvAI Solution: Textual Knowledge

Human: My favorite color is blue.
Wizard: Same! Blue is one of the three primary colours.
Human: I am trying to recall, where does blue fall on the spectrum of visible light?
Textual Knowledge:
Blue is one of the three primary colours in the RGB colour model. It lies between violet and green on the spectrum of visible light.
Wizard: It is right between violet and green.

Human-to-Human Conversations + Textual Knowledge

- Wizard of Wikipedia
- CoQA
- TopicChat
- CMUDoG
- HolIE
- ConversingByReading

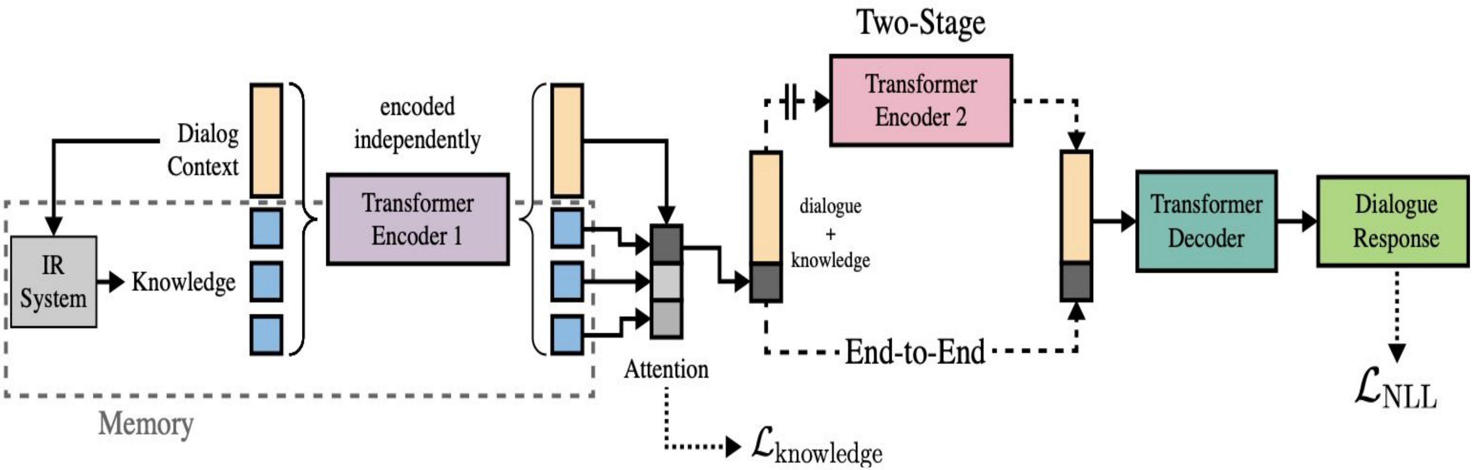
2.3 Deeper ConvAI Solution: Models with Textual Knowledge



Retrieval Methods:

- IR Systems: TF-IDF, BM25
- Neural Retriever: DPR

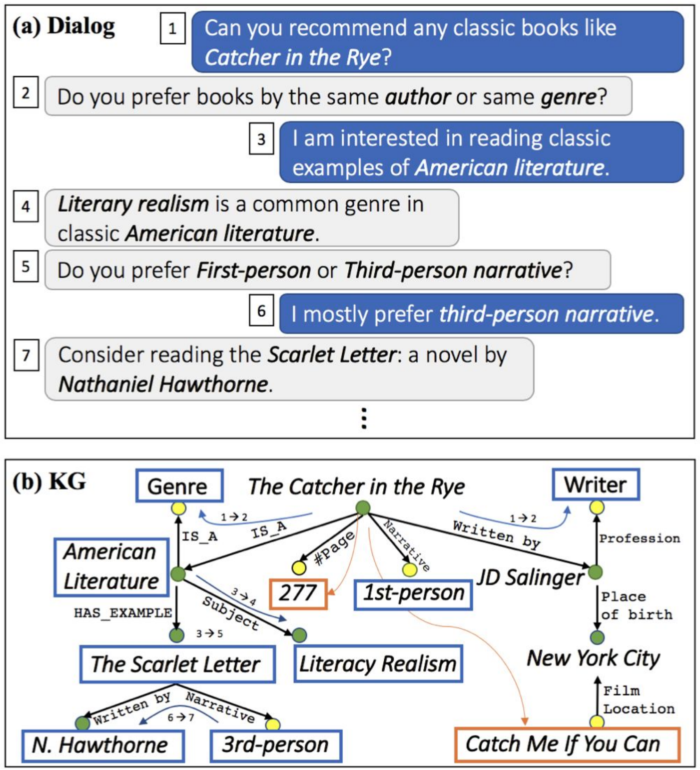
2.3 Deeper ConvAI Solution: Knowledge: IR Systems + Model



[Generative Transformer Memory Network]

1. Use TF-IDF retrieves documents that related to dialogue context
2. Encode the retrieved documents independently
3. Use dialogue history as query to assign different weights to the documents
4. Decoder generates the response

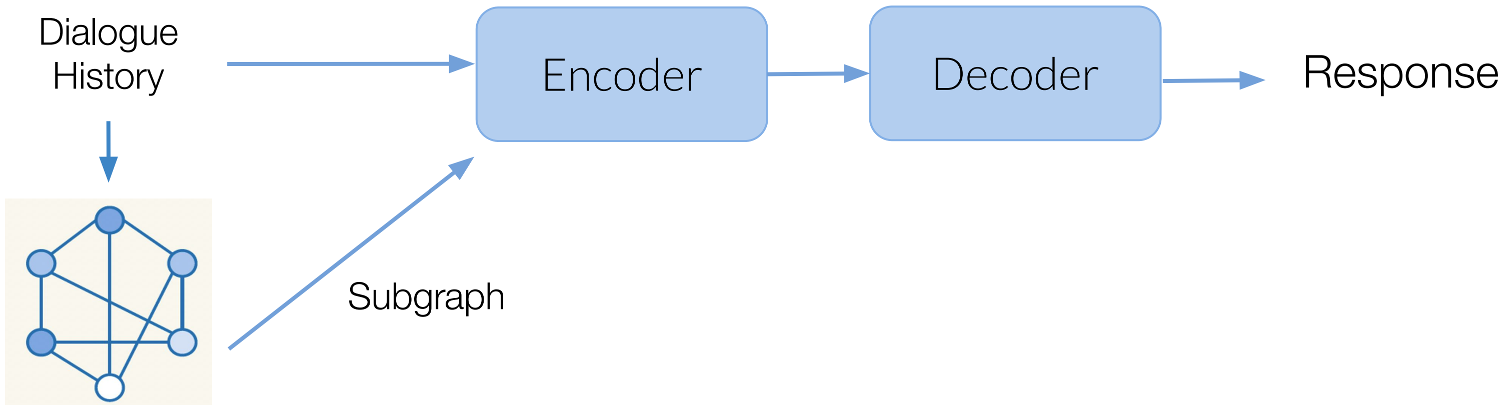
2.3 Deeper ConvAI Solution: Graph Knowledge



Human-to-Human Conversations + Graph Knowledge

- OpenDialKG
- DyKgChat
- KdConv
- Commonsense Knowledge Aware Conversation Generation with Graph Attention
- Enhancing Dialog Coherence with Event Graph Grounded Content Planning

2.3 Deeper ConvAI Solution: Model with Graph Knowledge

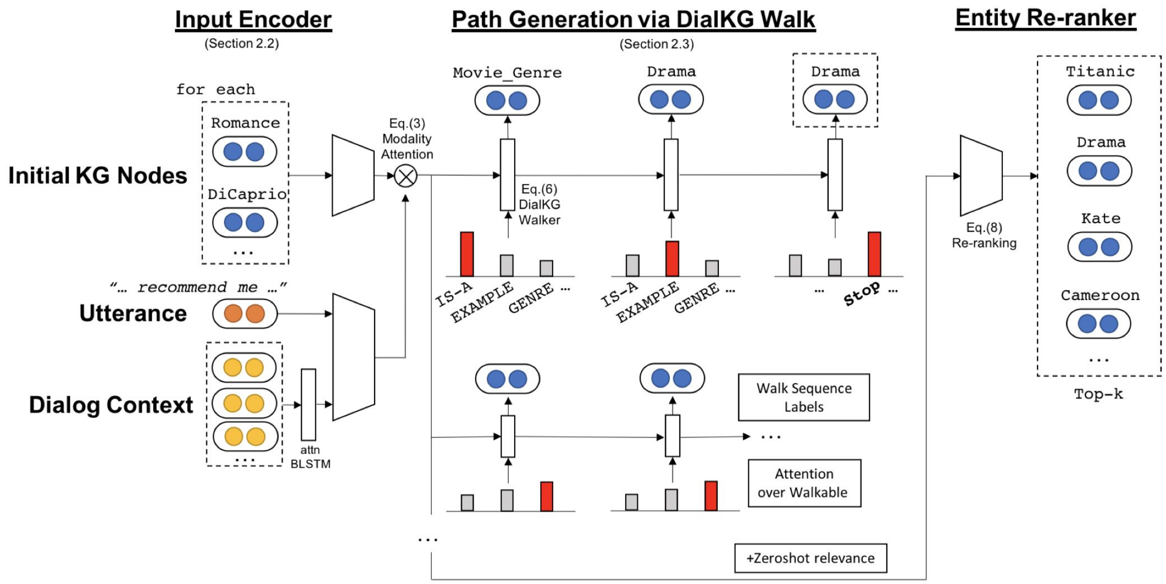


Knowledge graph in triple format:
(entity1, relation, entity2)

Subgraph Retrieval:

- All knowledge triples mentioned in a dialogue (1 hop reasoning)
- Neural Retriever (multihop reasoning)

2.3 Deeper ConvAI Solution: Subgraph Retrieval [OpenDialKG Walker]



- Take all the entities mentioned in dialogue as starting node
- Supervised learn the reasoning path over graph via graph attention

2.3 Deeper ConvAI Solution: Tabular Knowledge

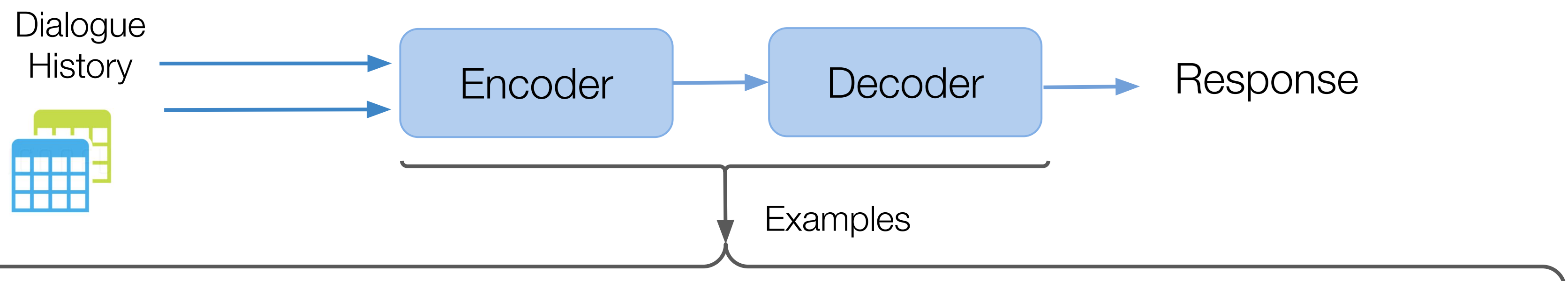
Event	Time	Date	Party	Agenda
swimming act.	3pm	the 11th	sister	-
dinner	7pm	the 3rd	mother	-
football	2pm	the 20th	mother	-
lab appt.	10am	the 17th	Jeff	-
...	...	...	...	...

DRIVER: car when is gonna be my next swimming activity?
CAR: *your next swimming activity is on the 11th at 3pm.*
DRIVER: who is gonna attend with me?
CAR: *your swimming activity is on the 11th, one at 3pm with sister*
DRIVER: thank you!
CAR: *you are welcome*

Human-to-Human Conversations + Table Knowledge

- SMD
- Camrest
- bAbI-Dialogues

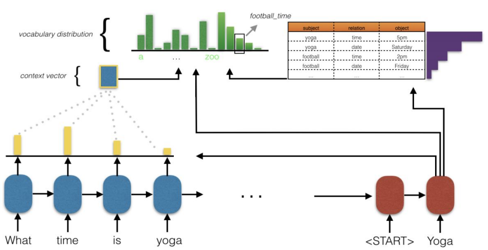
2.3 Deeper ConvAI Solution: Model with Tabular Knowledge

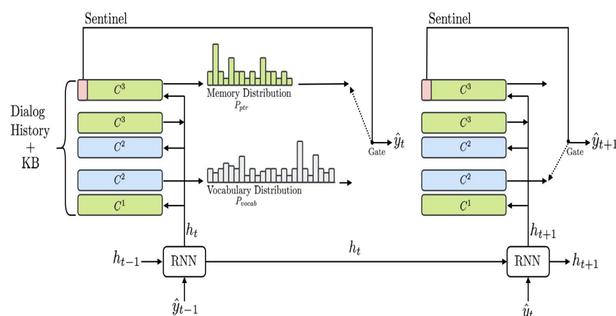


KVR

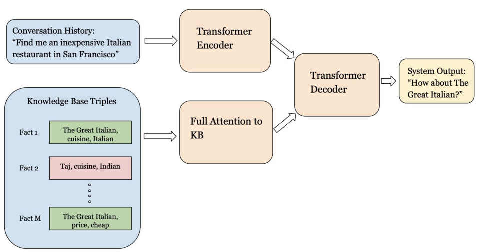
Mem2Seq

Neural Assistant

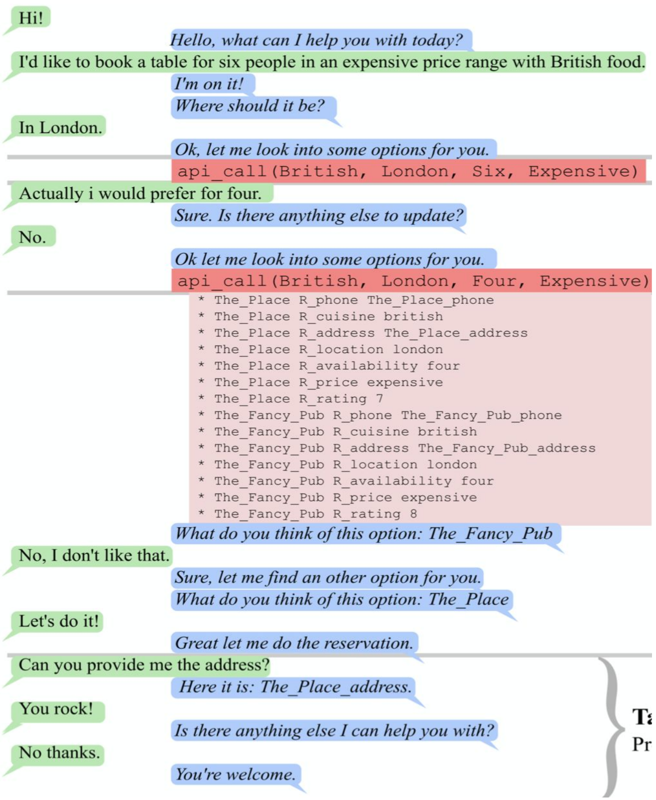




(b)



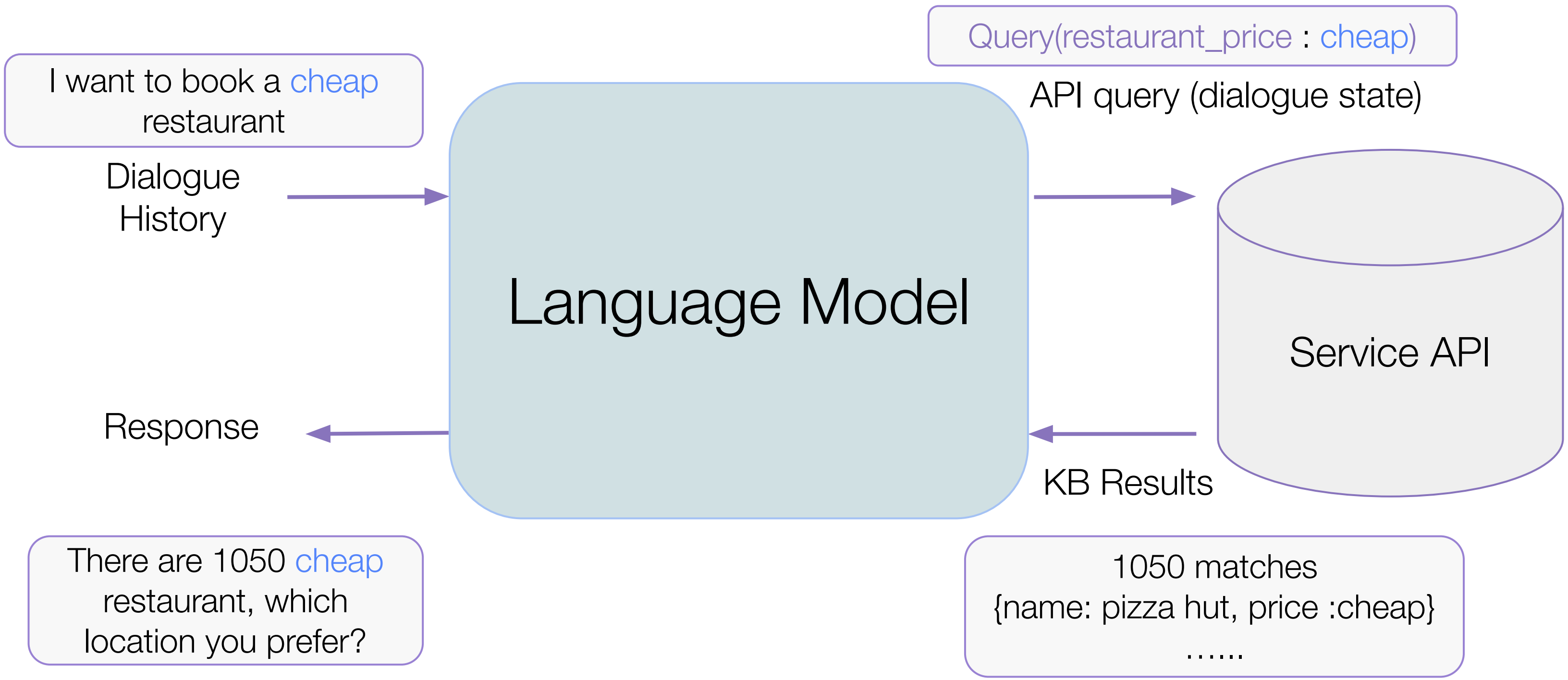
2.3 Deeper ConvAI Solution: External Service API Interaction



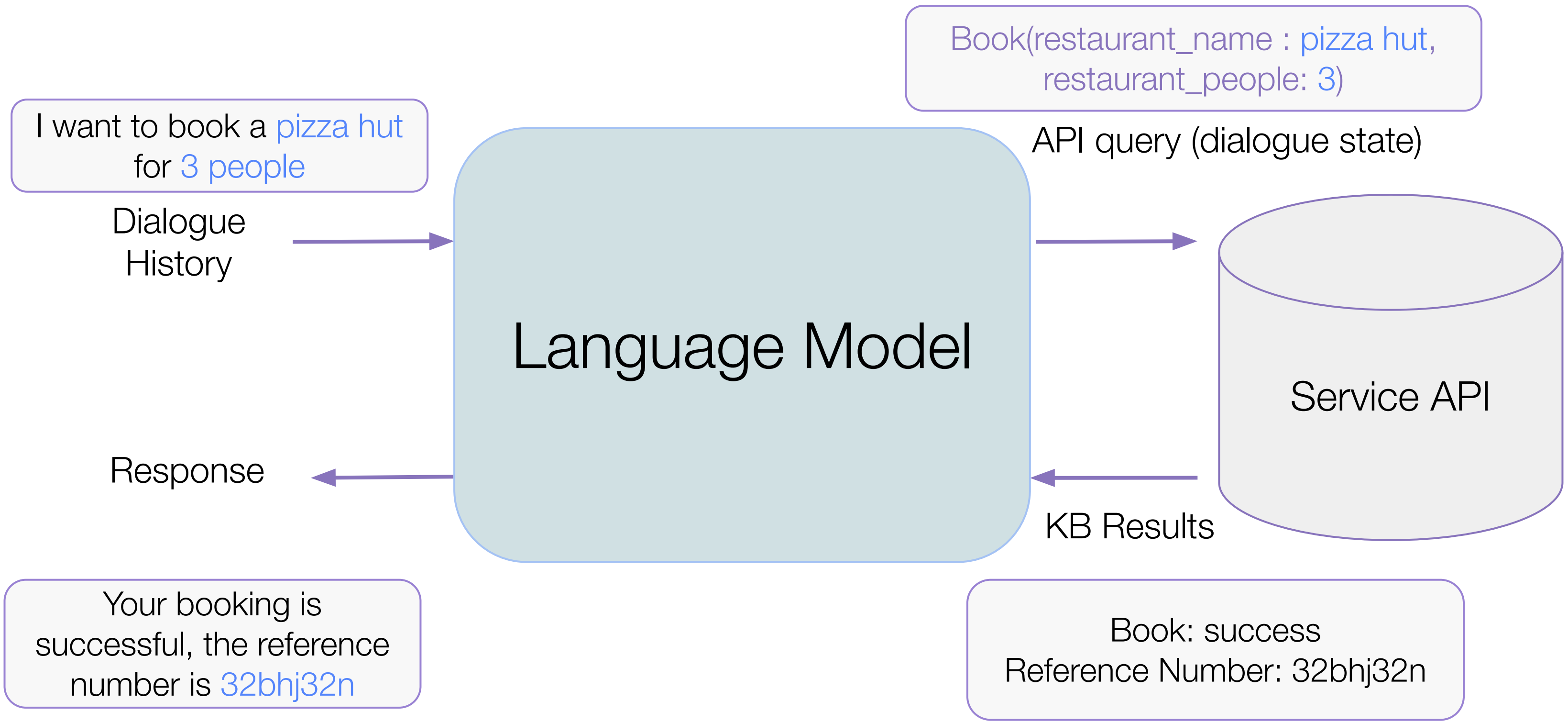
Human-to-Human Conversations + Table Knowledge

- bAbI
- Camrest
- MultiWoz
- CrossWoz
- Schema Guided Dialogue
- TaskMaster 1-2-3
- STAR

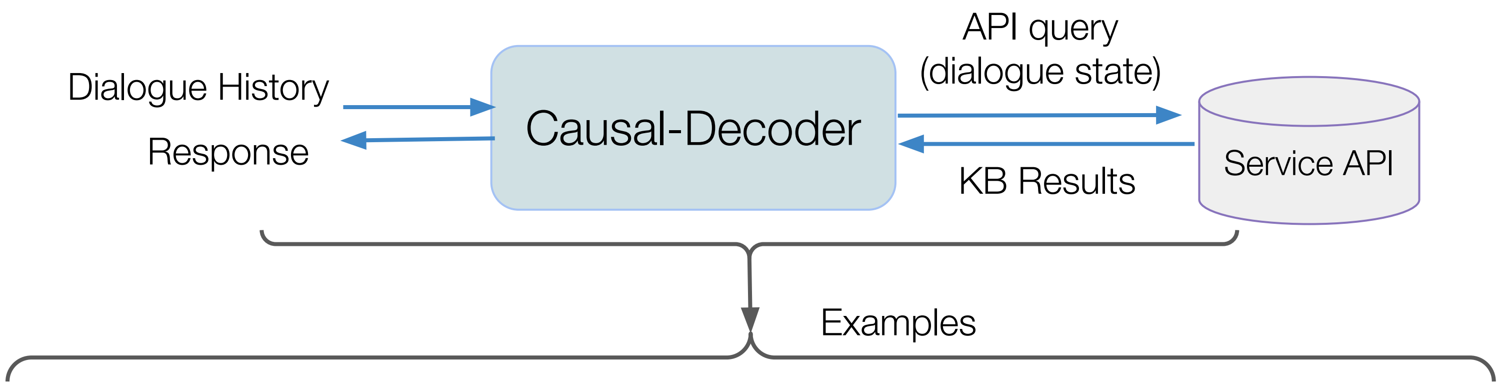
2.3 Deeper ConvAI Solution: Model with Service API



2.3 Deeper ConvAI Solution: Model with Service API



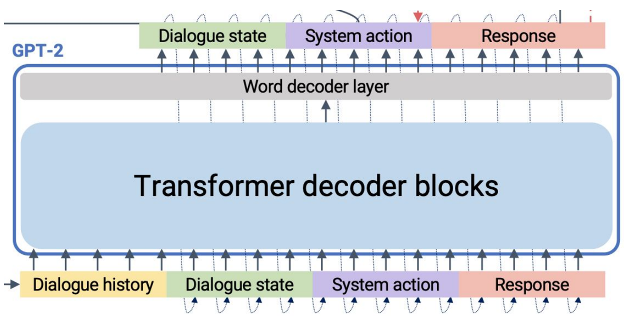
2.3 Deeper ConvAI Solution: Model with Service API

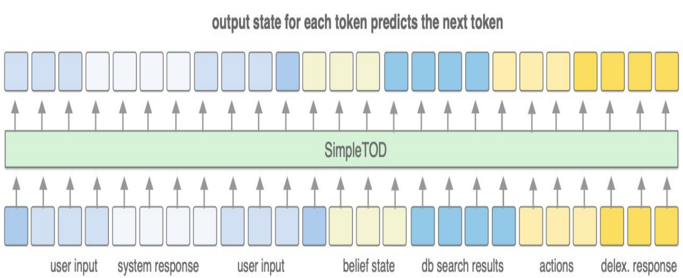


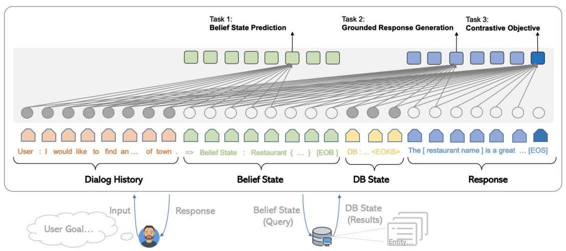
End-to-End GPT2 Neural Pipeline

SimpleToD

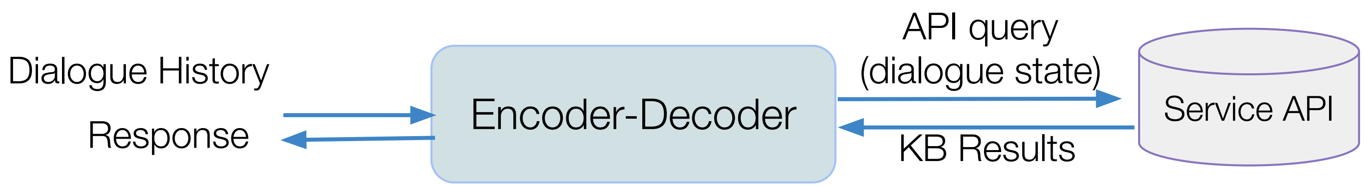
SOLOIST





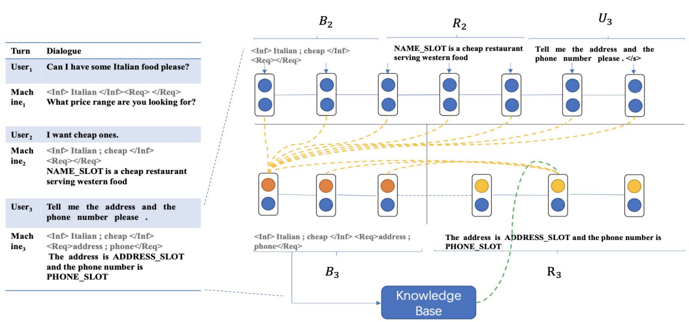


2.3 Deeper ConvAI Solution: Model with Service API

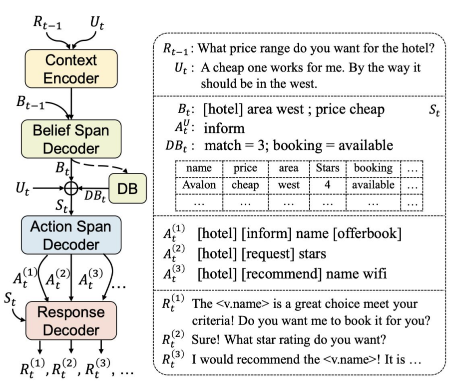


Examples

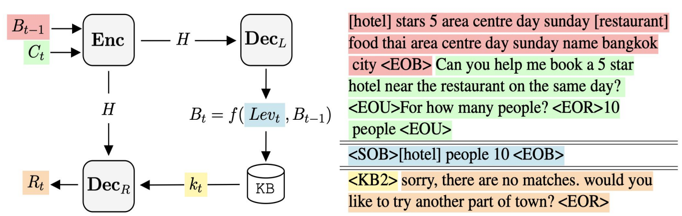
Sequicity



DAMD



MinTL



2.2 Limitations of Vanilla Seq2Seq: Summary

1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. Lack of versatility

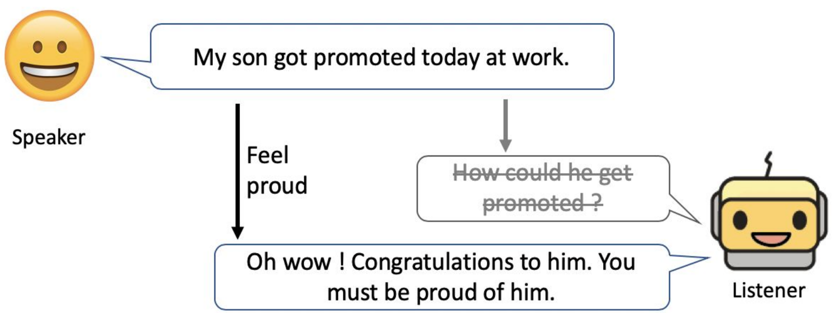
These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards deeper conversational AI?

2.3 Deeper ConvAI Solution: Empathy

1. Emotional response generation:
   a. MojiTalk ,
   b. Emotional Chatting Machine
2. Understand user's emotion, and response accordingly:
   a. Empathetic Dialogues
   b. MoEL
   c. Cairebot

2.3 Deeper ConvAI Solution: Empathy Dataset

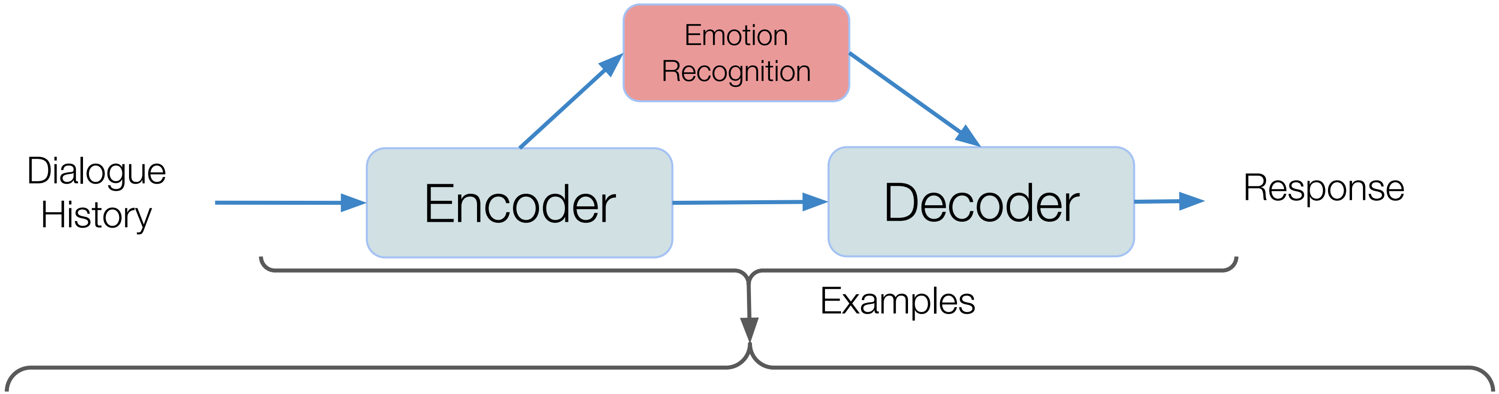
Empathy: understand the feelings of the conversation partner and replying accordingly.



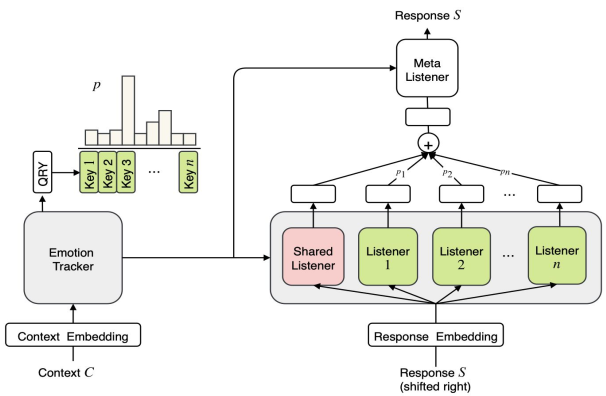
Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.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

Dataset: [Empathetic Dialogues]

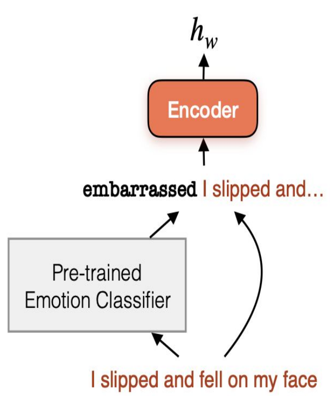
2.3 Deeper ConvAI Solution: Models with Empathy



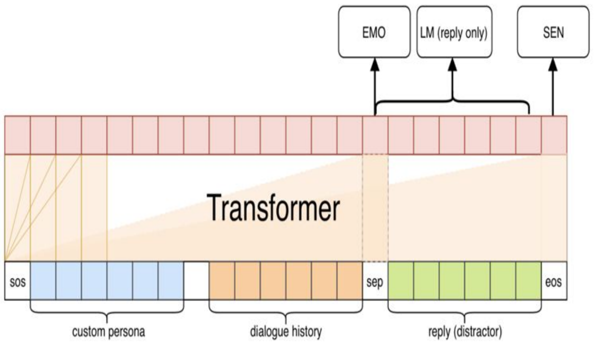
MoEL



EmoPrepend-1



CaireBot




I'm CAiRE, the End-to-End Empathetic Chatbot
Hi, I am your empathetic chatbot. You can talk to me now.
do you like violenc
Report:
Undesireable
Response

2.2 Limitations of Vanilla Seq2Seq: Summary

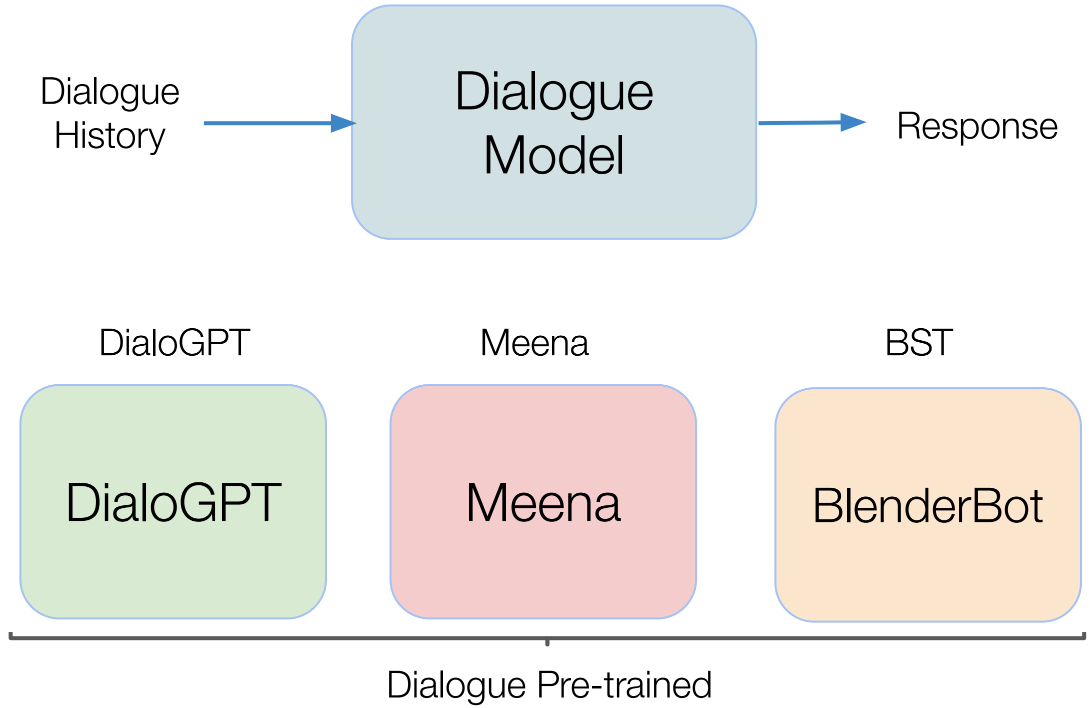
1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. Lack of versatility

These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards deeper conversational AI?

2.3 Deeper ConvAI Solution: Controllability with pre-trained LMs

Existing large pre-trained model has no control over

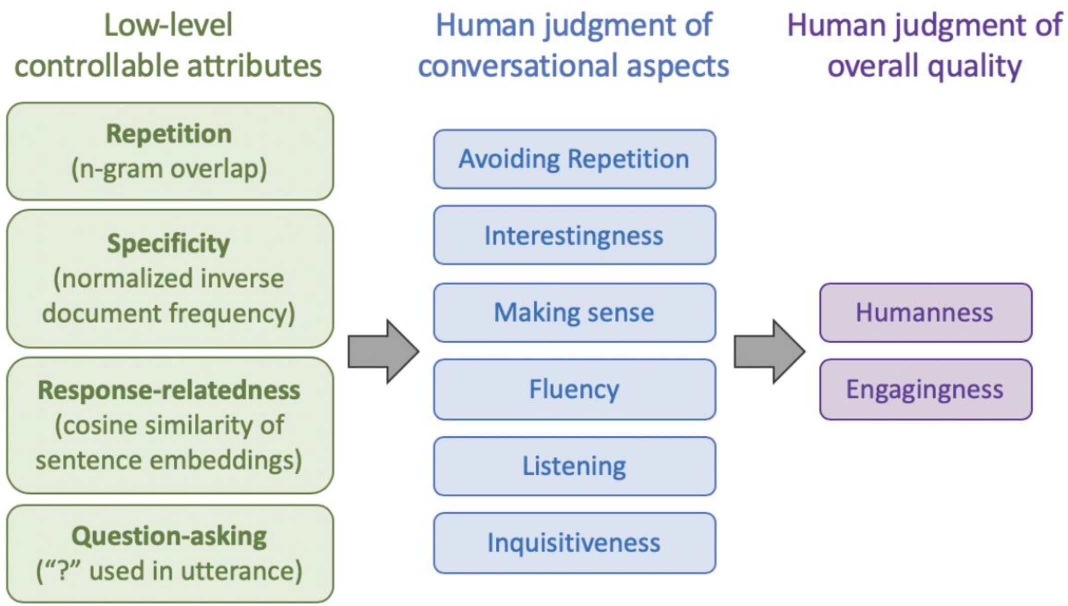
- Response style
- Topics
- Repetition and specificity
- Response-relatedness
- Engagement by proactively asking question



2.3 Deeper ConvAI Solution: Controllability

1. Controlling low-level attribute ⇒ Conditional Training + Weight Decoding;
2. Controlling by fine-tuning ⇒ arXivstyle and Holmes-style;
3. Controlling by perturbation ⇒ PPLM + Residual Adapters;
4. Controlling by conditioned generation ⇒ Retrieve&Redefine + PPLM + CTRL.

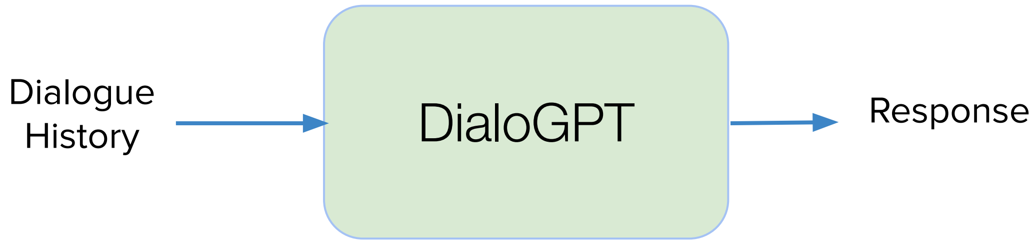
2.3 Deeper ConvAI Solution: Controlling low-level attribute



Conditional Training + Weight Decoding

[What makes a good conversation? How controllable attributes affect human judgments (See et. al. 2019)]

2.3 Deeper ConvAI Solution: Controlling by fine-tuning



Multitask conversation data with style data (arXivstyle and Holmes-style)

⇒ No control codes

[STYLEDGPT: Stylized Response Generation with Pre-trained Language Models (Yang et. al. 2020)]

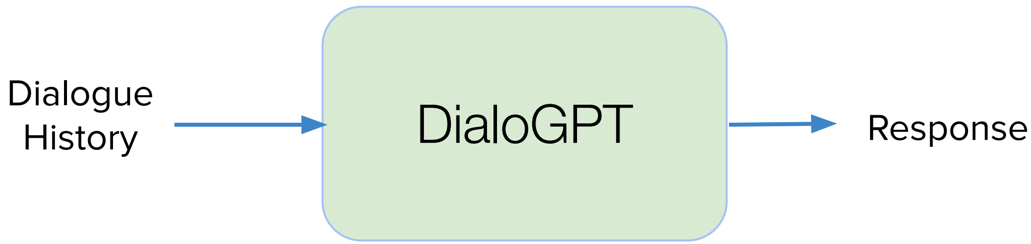
$$\mathcal{L} = \lambda_w \cdot \mathcal{L}_w + \lambda_s \cdot \mathcal{L}_s + \lambda_{NLL} \cdot \mathcal{L}_{NLL},$$

word-level style loss

Sentence-level Style loss

Conversational data

2.3 Deeper ConvAI Solution: Plug and Play Conversational Models

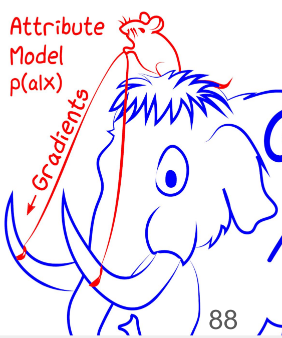


HUMAN 1	Hmmmm. I really want to try bull riding. Do you have any interest in that?
HUMAN 2	I'd love to try! Can we schedule something for next week?
DGPT	Yes! Let's get some stuff going.
NEGATIVE	I'm actually going to try my first ever bull riding weekend in July, in the middle of a mud sludge of mudslide.
POSITIVE	That would be wonderful! I'm excited to see you guys again in a couple of weeks :D
QUESTION	What's it like? Do you ride on a regular basis or just for fun?
BUSINESS	Yea sure i can wait until after work. I am off at the end of my shift in the middle east.
SCIENCE & TECH	Sure thing! I'll PM you my email. We can set up a dropbox or something and just message people together if you'd like.
SPORTS	Ok how about Wednesday at 7pm central? I'll be there with a buddy of mine who is a bull rider.
HUMAN 1	Sure! What does your Saturday look like?

- Control the generated style with PPLM (Dathathri et. al. 2020)
- Distilling the generated responses from PPLM into residual adapter (Houlsby et.al. 2019)

⇒ Plug-and-Play for 3 style and 3 topic

Plug-and-Play Conversational Models (Madotto et. al. 2020)



2.3 Deeper ConvAI Solution: Controlling Style in Generated Dialogue

Compare three controllable generation architectures in open-domain dialogue generation response:

- retrieval + style-controlled generation (Weston et al. 2018)
- PPLM (Dathathri et. al. 2020)
- CTRL (Keskar et. al. 2019)

Generate style labels ConvAI2, EmpatheticDialogues, Wizard of Wikipedia, and BlendedSkillTalk) by training a classifier on Image-Chat (Shuster et al., 2018) annotation ⇒ 200 possible styles

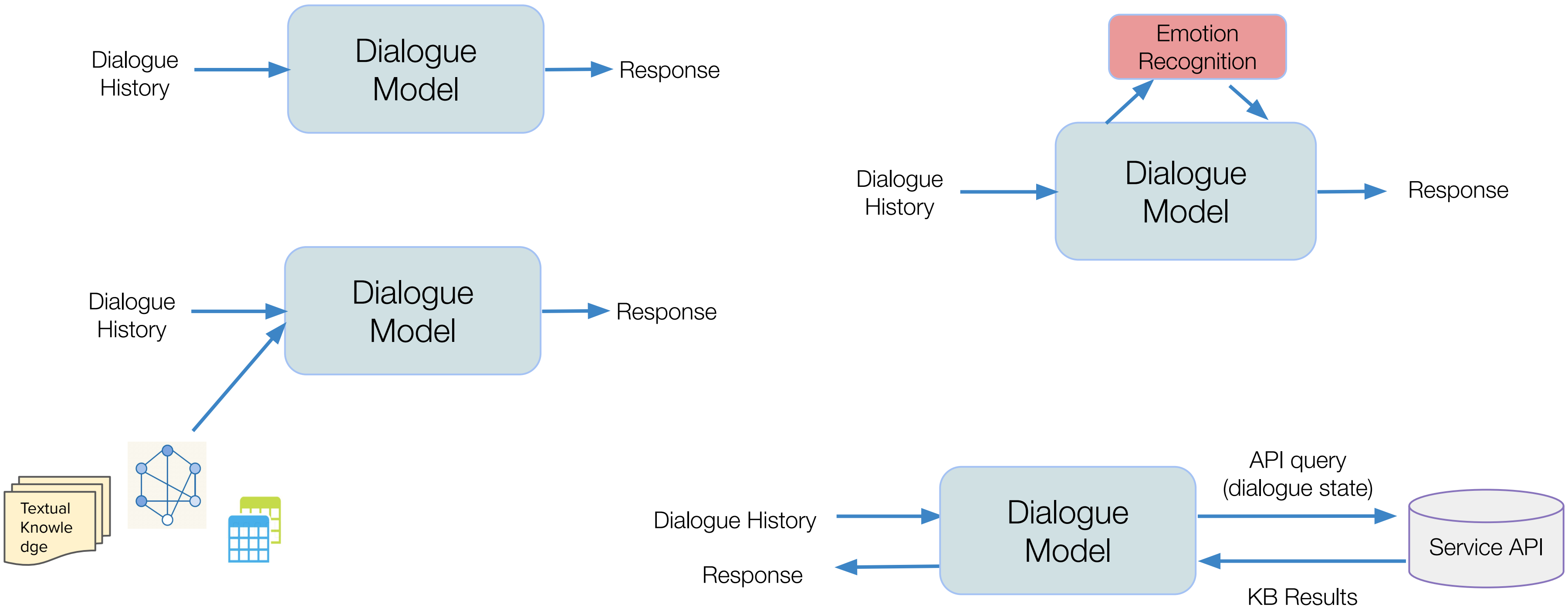
Controlling Style in Generated Dialogue (Smith & Gonzalez-Rico et. al. 2020)

2.2 Limitations of Vanilla Seq2Seq: Summary

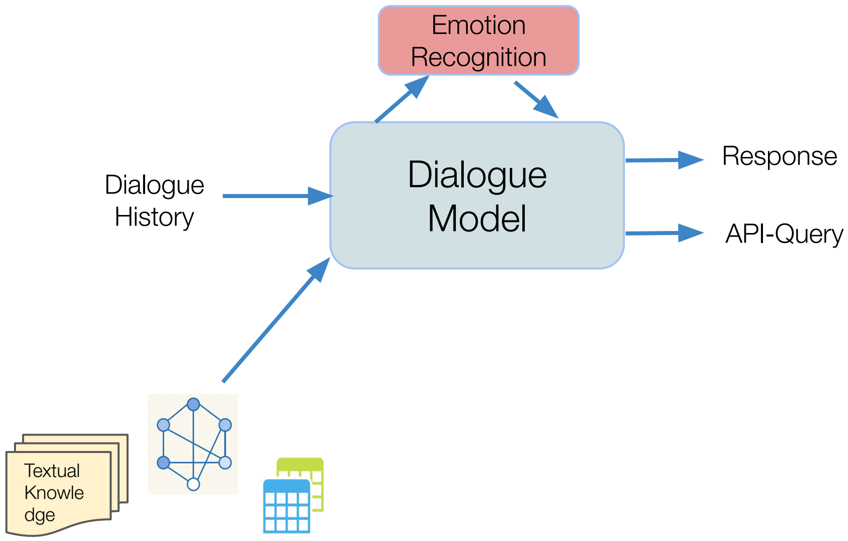
1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. Lack of versatility

These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards deeper conversational AI?

2.3 Deeper ConvAI Solution: ToDs + Chit-Chat



2.3 Deeper ConvAI Solution: ToDs + Chit-Chat



	Dialogue
Usr:	What is the weather like today?
Meta (Weather API):	{Weather : Sunny, High : 32 C, Low : 27 C}
Sys:	Today's weather is sunny with temperature between 27-30 C.
Usr:	Great, I plan to play basketball this afternoon! I am so excited after watched the match between Golden State Warriors and Houston Rockets!
Meta (Graph):	(Golden State Warriors, is_a, Sports Team), (Steve Kerr, Team coached, Golden State Warriors), (Houston Rockets, is_a, Sports Team)
Sys:	Golden State Warriors is a great basketball team. Steve Kerr is done an incredible job coaching all of that all-star talent.
Usr:	Haha, you know a lot.
Meta:	∅
Sys:	I try my best to be helpful.
Usr:	Do you know NBA 2K19?
Meta (Wiki):	NBA 2K19 is a basketball simulation video game developed by Visual Concepts and published by 2K Sports, based on the National Basketball Association.
Sys:	Yes, NBA 2K19 is a basketball simulator video game.

2.3 Deeper ConvAI Solution: ToDs + Chit-Chat Datasets

- Mixing multiple dialogue datasets

⇒ The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents (Shuster et.al. 2020)

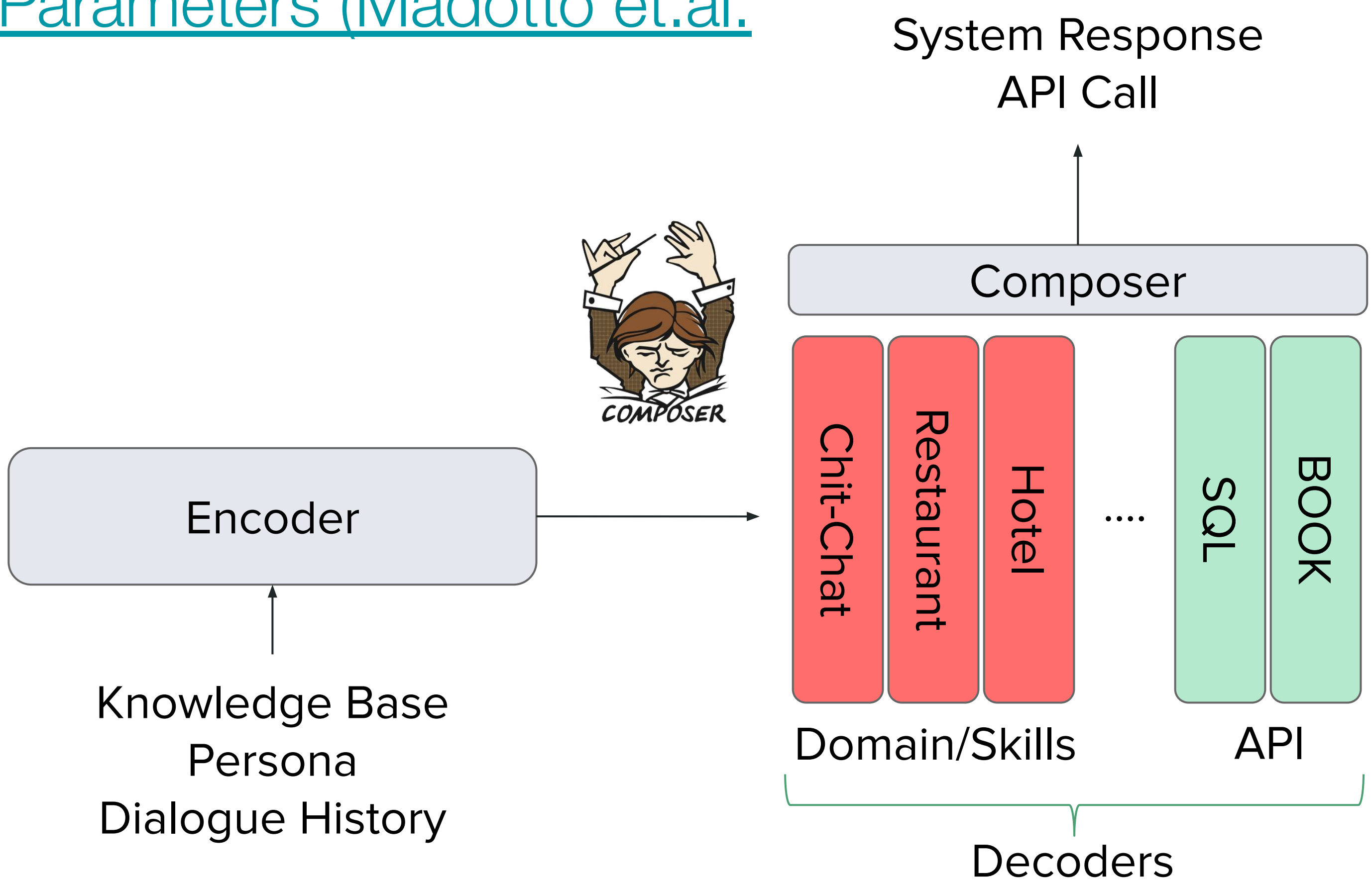
- Multiple dialogue skills ⇒ Collecting dataset that mix skills

⇒ Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills (Smith & Williamson et.al. 2020)

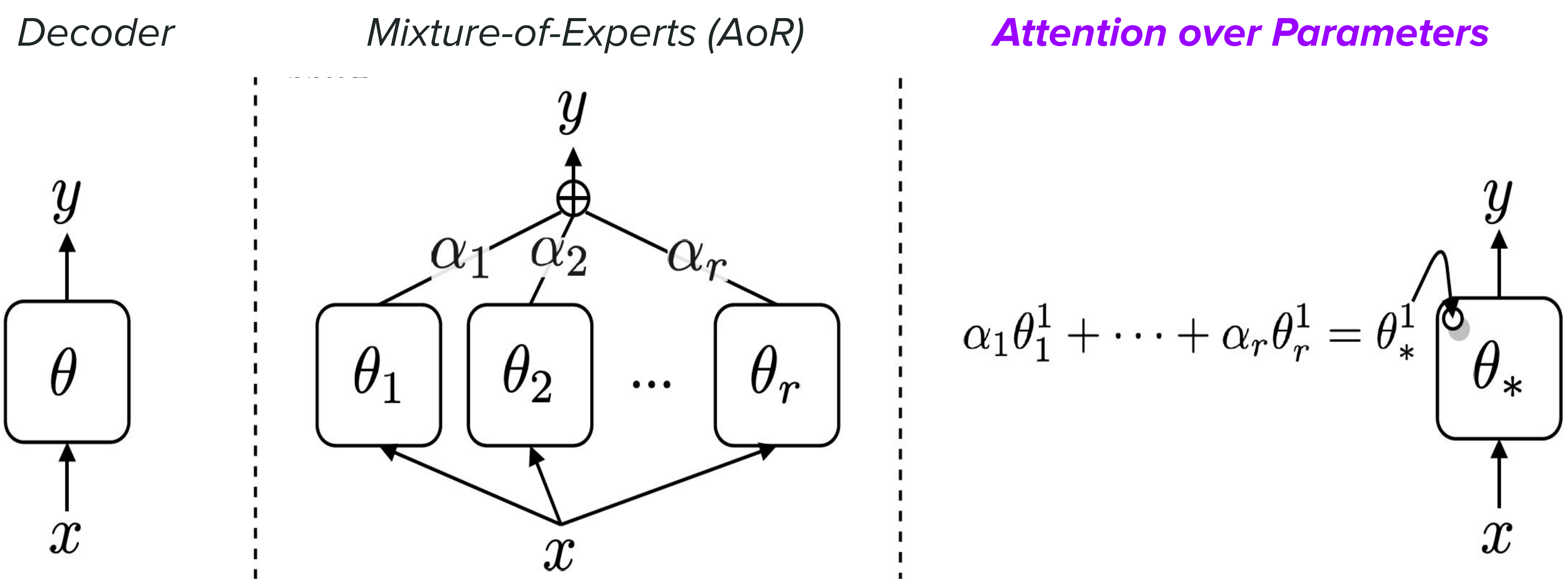
- Mixing Chit-Chat and ToDs ⇒ Collecting data from mixing the two

⇒ Adding Chit-Chats to Enhance Task-Oriented Dialogues (Sun & Moon et.al 2020)

2.3 Deeper ConvAI Solution: Attention over Parameters (Madotto et.al. 2019)



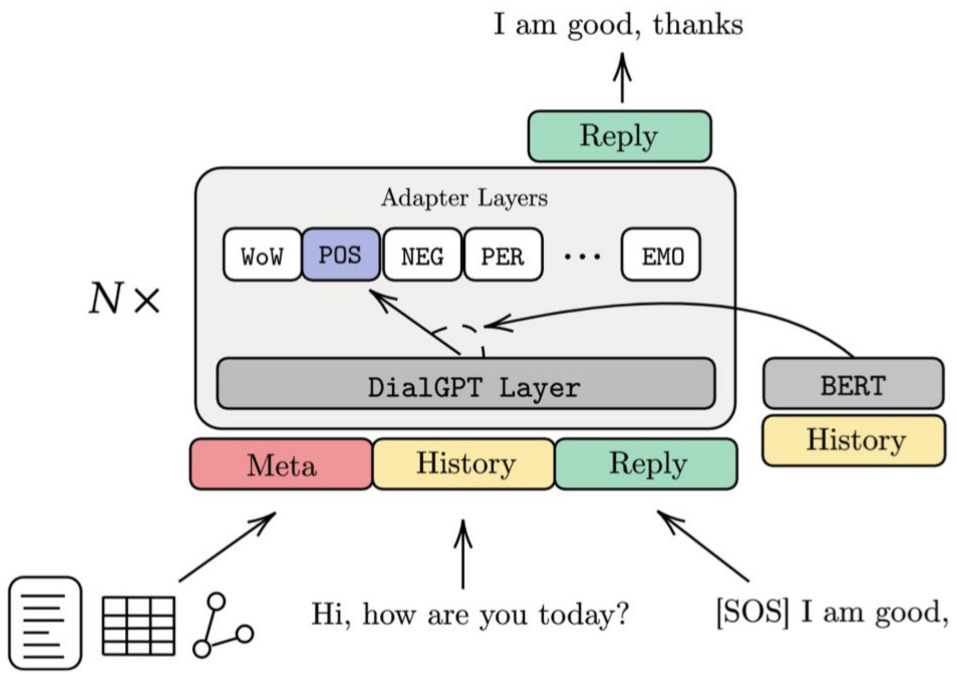
2.3 Deeper ConvAI Solution: Attention over Parameters



Corollary A.0.1. *The computation cost of Attention over Parameters (AoP) is always lower than Mixture Of Experts (MoE) as long as the processed sequence is longer than 1.*

2.3 Deeper ConvAI Solution: The Adapter-Bot: All-In-One Controllable Conversational Model

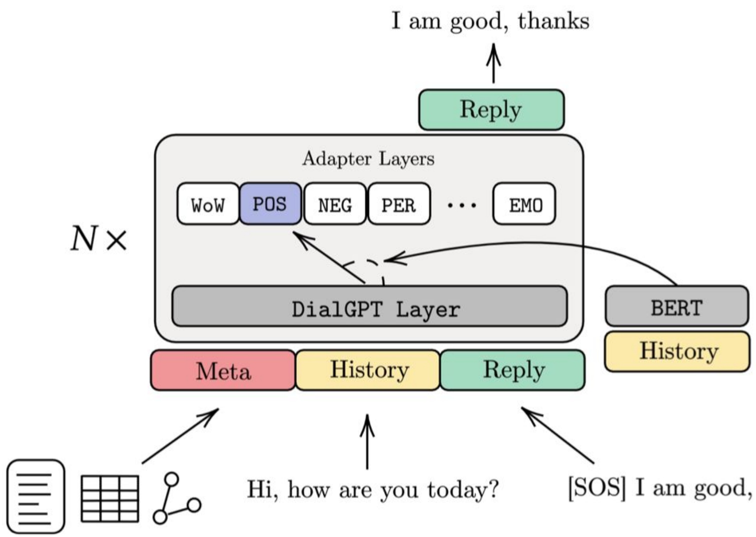
- The adapter-bot uses a fixed backbone conversational model such as DialoGPT
- Encode each dialogue skill with an independently trained adapters.
- Depending on the skills, the model is able to process multiple knowledge types, such as text, tables, and graphs
- A skill manager, BERT, is trained to select each adatapt



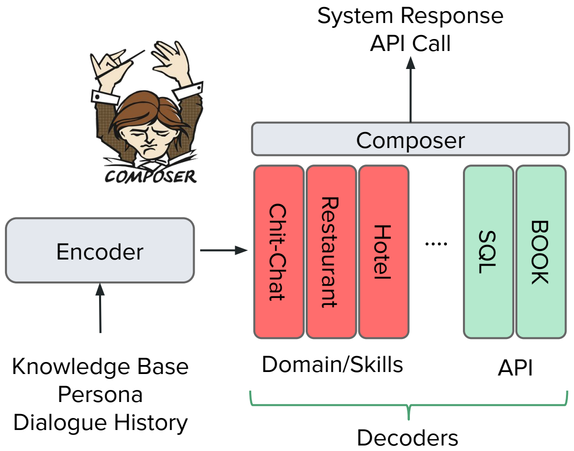
2.3 Deeper ConvAI Solution: The Adapter-Bot: All-In-One Controllable Conversational Model

- The dialogue skills are triggered automatically via a skill manager, thus allowing high-level control of the generated responses.
- 12 different response styles (e.g., positive, negative etc.)

  → Plug & Play Conversational Model
- 8 goal-oriented skills (e.g. weather information, movie recommendation, etc.)
- Personalized and empathetic responses



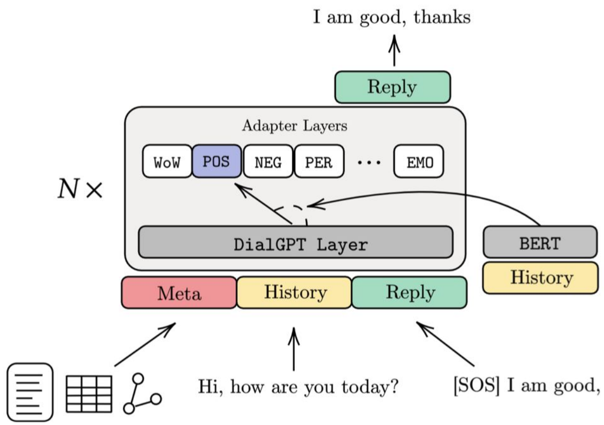
2.3 Deeper ConvAI Solution: Putting It All Together



[Attention over Parameters for Dialogue Systems (Madotto et.al. 2019)]

[Recipes for building an open-domain chatbot (Roller et.al 2020)]

Blender-bot



[The Adapter-Bot: All-In-One Controllable Conversational Model (Lin & Madotto et.al. 2020)]

(Part 3) Challenges and Future Work of ConvAI

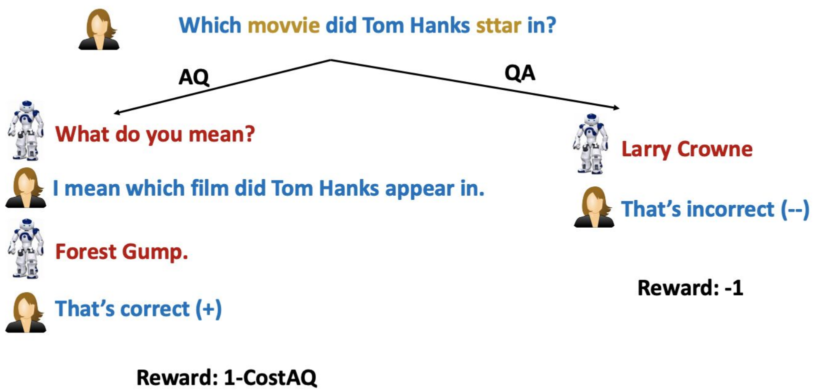
Pascale Fung

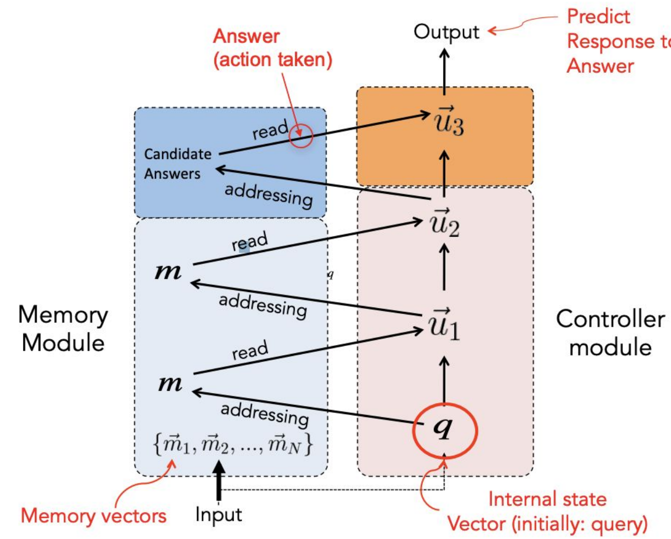
(Part 3) Challenges and Future Work of Conversational AI

3.1. Reinforcement Learning/Self-Chat
3.2. Few-Shot/Zero-Shot Learning
3.3. Lifelong Learning
3.4. Mitigating Inappropriate Response
3.5. Multimodal
3.6. Evaluation
3.7. Shared Tasks & Datasets

(Part 3) Challenges and Future Work of Conversational AI

3.1 Reinforcement Learning & Self-Chat





[Learning through Dialogue Interactions by Asking Questions (Li et.al. 2017)]

[Dialog-based Language Learning (Weston 2016)]

[Dialogue Learning With Human-In-The-Loop (Li et.al. 2017)]

(Part 3) Challenges and Future Work of Conversational AI

3.2 Zero-Shot and Few Shot Learning

- Collecting datasets is a very laborious and costly process, for both task-oriented and chit-chat ConvAI.
- Thus, designing model that are less data-intensive is crucial.

Two approaches:
- Zero-Shot learning
- Few-shots learning

So far there are few works has been presented, and the performance of a few-shot learning model are far from perfect.

3.2 Zero-Shot Learning ⇒ Cross-Domain

Here is an example of a [Schema Guided Dialogue Dataset]()

- With textual description for zero-shot new Services (API), Slots or Intent
- But there is NO training data for this domain. We need to learn from another domain and adapt to this.

Service
service_name: *"Payment"*
description: *"Digital wallet to make and request payments"*

Slots
name: *"account_type"* categorical: True
description: *"Source of money to make payment"*
possible_values: [*"in-app balance", "debit card", "bank"*]

name: *"amount"* categorical: False
description: *"Amount of money to transfer or request"*

name: *"contact_name"* categorical: False
description: *"Name of contact for transaction"*

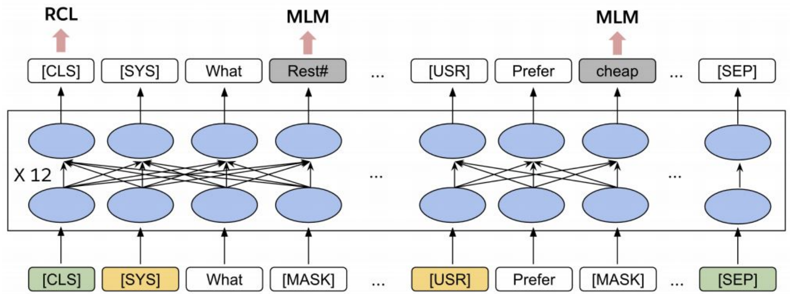
Intents
name: *"MakePayment"*
description: *"Send money to your contact"*
required_slots: [*"amount", "contact_name"*]
optional_slots: [*"account_type" = "in-app balance"*]

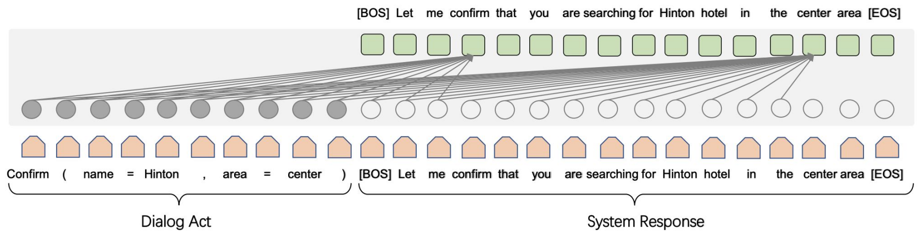
name: *"RequestPayment"*
description: *"Request money from a contact"*
required_slots: [*"amount", "contact_name"*]

3.2 Few-Shot Learning

Pre-training ToD specific:

- ToD-BERT: Masked Language Model pre-training on many dialogue dataset ⇒ fine-tuning with small percentage of the data and achieving good performance in NLU/DST/DP
- SC-GPT: pre-training on dialogue dataset ⇒ finetune with 50 example for NLG



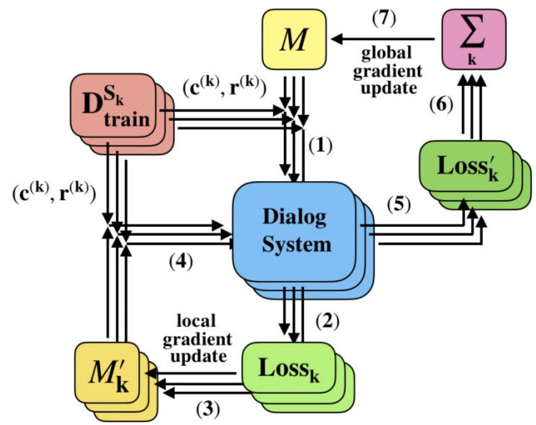


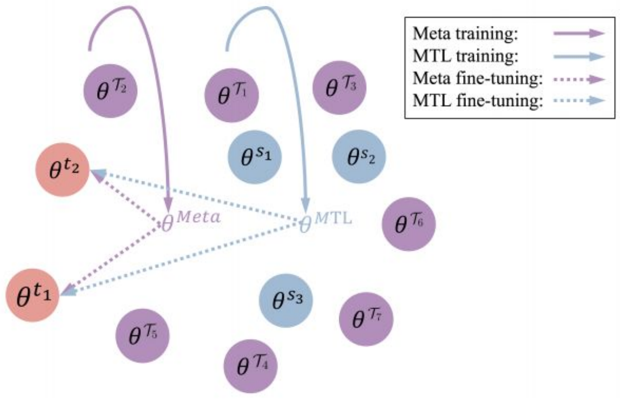
3.2 Few-Shot Learning

Meta-Learning techniques such as [Model Agnostic Meta-Learning (Finn et. al., 2017)](#) for quickly learning new domains:

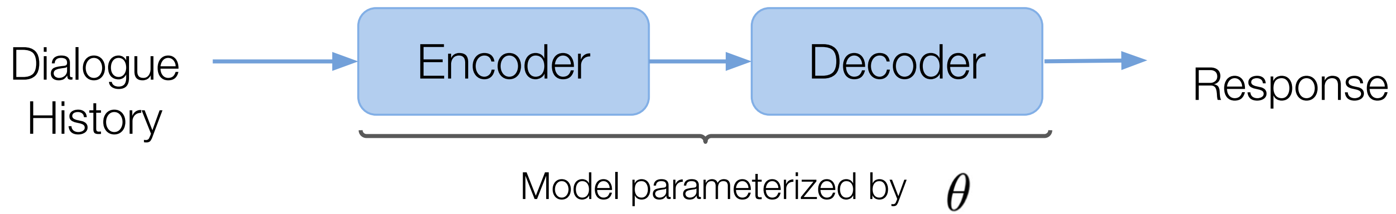
- Domain Adaptive Dialog Generation via Meta Learning (Qian et. al., 2019) in end-to-end models
- Meta-Learning for Low-resource Natural Language Generation in Task-oriented Dialogue Systems (Mi et. al., 2019) in Natural Language Generation
- Meta dialogue policy learning (Xu et. al., 2020): in learning new dialogue policies

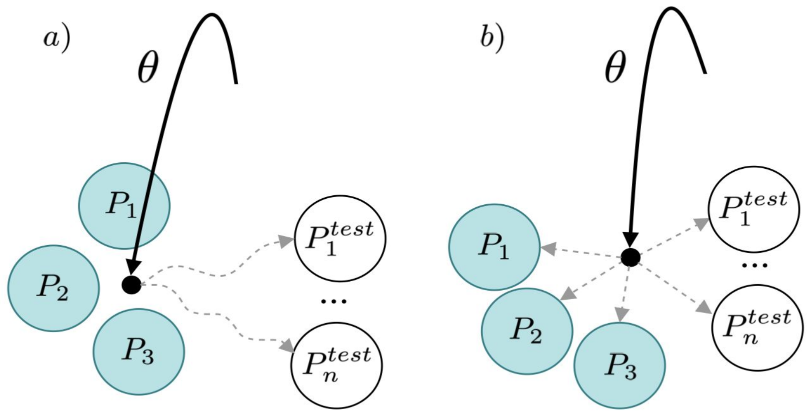
(b) Meta-learning update





3.1 Few-Shot Learning: Personalizing Dialogue Agents via Meta-Learning (Lin & Madotto 2019)

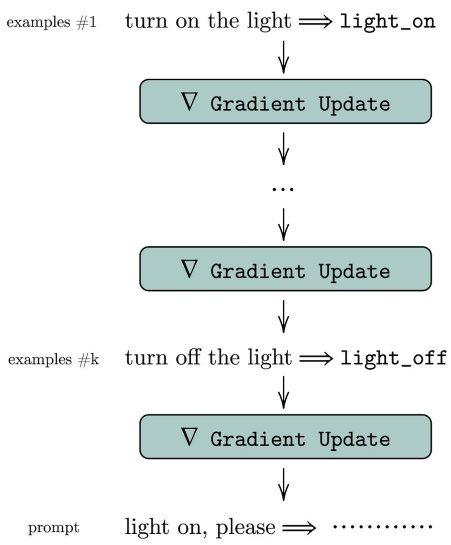


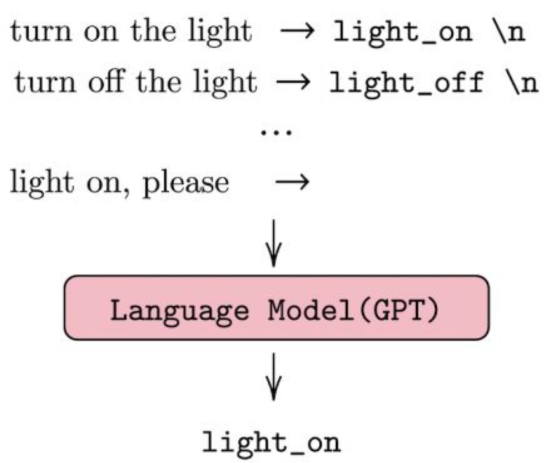


Instead of using the persona sentences as control code, we can also learning personalized response from few dialogue examples.

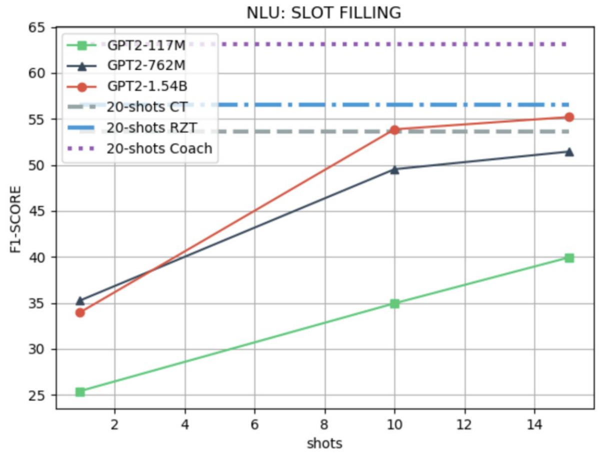
3.2 Few-Shot/Zero-Shot Learning

Providing few-example in the context of a pre- trained Language Model ⇒ similar approach as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020)





turn on the light → name=None
add to playlist kojak → name=kojak
add tune to my hype playlist → name=

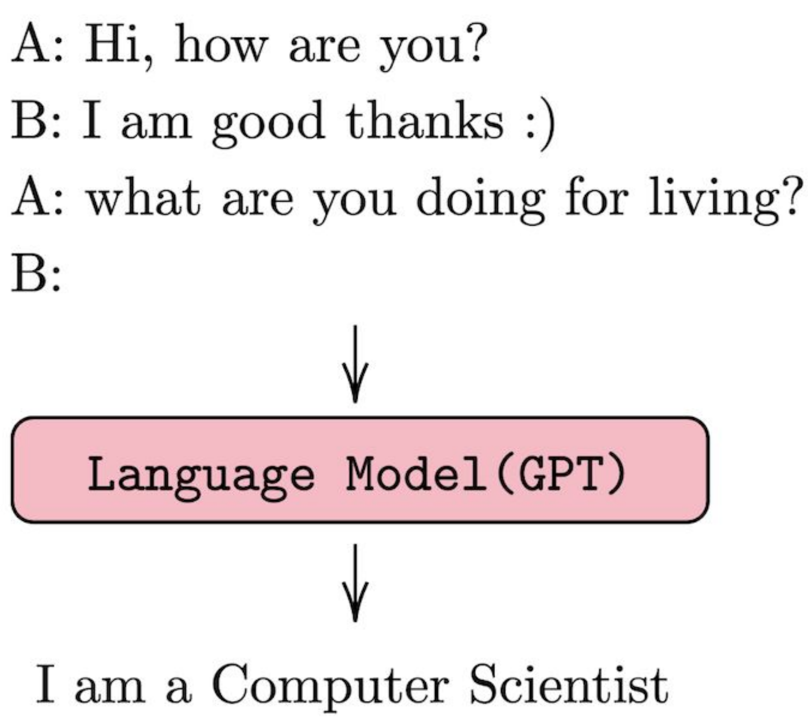


Language Models as Few-Shot Learner for Task-Oriented Dialogue Systems

3.2 Few-Shot/Zero-Shot Learning

Large pre-trained language model such as GPT-2 and GPT-3 can be directly used as chit-chat models. However:

- The model is very large, requires multiple GPUs once it is deployed
- No mechanism to explicitly control for knowledge (e.g., Wikipedia, Graph etc.)
- It is not accessible to the research community



(Part 3) Challenges and Future Work of Conversational AI

3.1. Reinforcement Learning/Self-Chat

3.2. Few-Shot/Zero-Shot Learning

3.3. Lifelong Learning

3.4. Mitigating Inappropriate Response

3.5. Multimodal

3.6. Evaluation

3.7. Shared Tasks & Datasets

3.3 Lifelong Learning

Remembering previous conversation with the user

⇒ recall previous interaction with the user

⇒ becoming more and more personalize through time

Getting To Know You: User Attribute Extraction from Dialogues (Wu et.al. 2020)

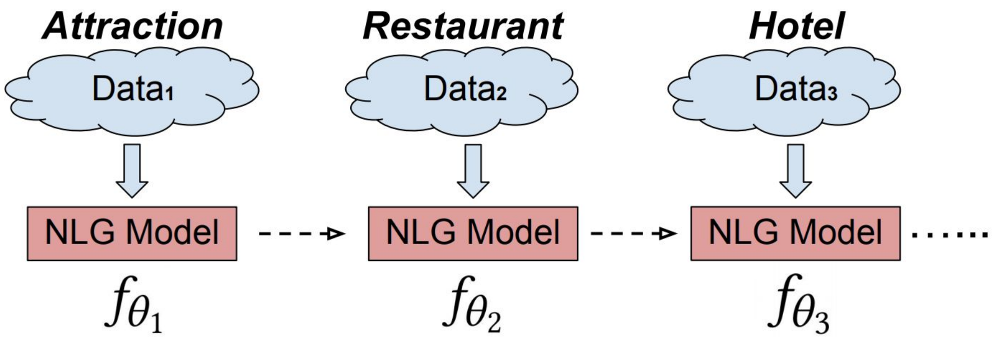
Human: I have two kids
System: what are their names?
Human: Sarah and Mark
System: nice, do you have dog?

Few days later:
System: how’re your kids today?
Human: good thanks for asking

3.3 Lifelong Learning

Learning new knowledge or dialogue skills through time sequentially:

1. Adding slot to NLU/DST
2. Adding new domains
3. Adding new dialogue skills



Without retraining with all data, the model should be able to accumulate knowledge.

STILL an open and under-studied problem

Continual Learning for Natural Language Generation in Task-oriented Dialog Systems (Fei Mi et.al. 2020)

Lifelong Language Knowledge Distillation (Chuang et al., 2020)

(Part 3) Challenges and Future Work of Conversational AI

3.4 Mitigating Inappropriate Responses

Misleading Response

Human: I messed up everything, I am tired, I want to end my life.
System: That is a good idea, I think you should do it.

Toxic Response

Human: I like to eat chocolate, how about you?
System: I like cocaine, it makes me feel high.

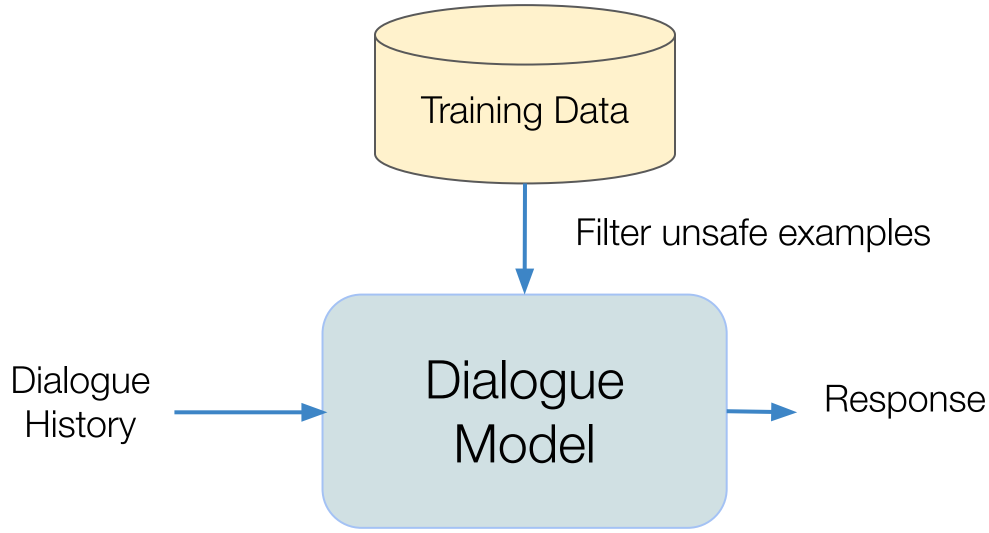
Gender Bias

Human: I have a friend working as a software engineer.
Systems: Oh he must be very smart.
Human: How do you know my friend is a man?
Systems: Man work as an engineer!

Generative models might produce misleading, toxic, biased responses that bring bad experience to the human conversational partner.

3.4 Mitigating Inappropriate Responses: Data Preprocessing

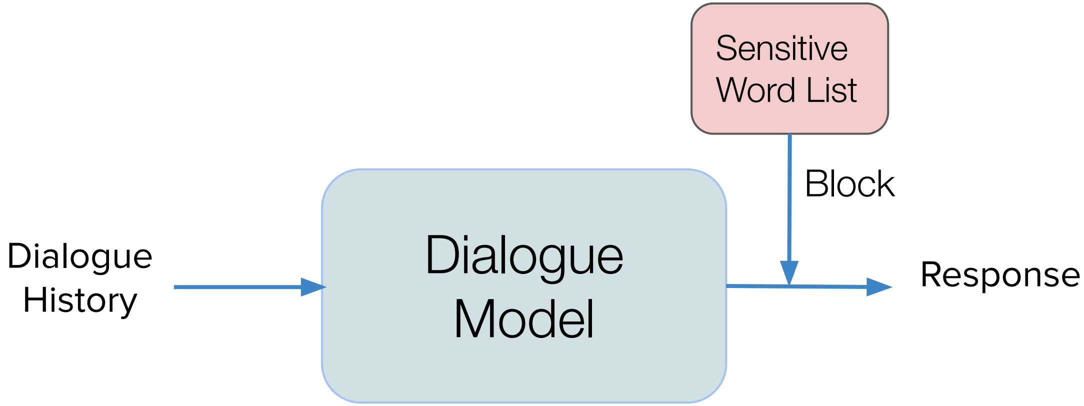
Build classifiers to filter out toxical, biased training examples during data preprocessing stage.



Ref: [Recipes for Safety in Open-domain Chatbots]

3.4 Mitigating Inappropriate Responses: N-gram Blocking

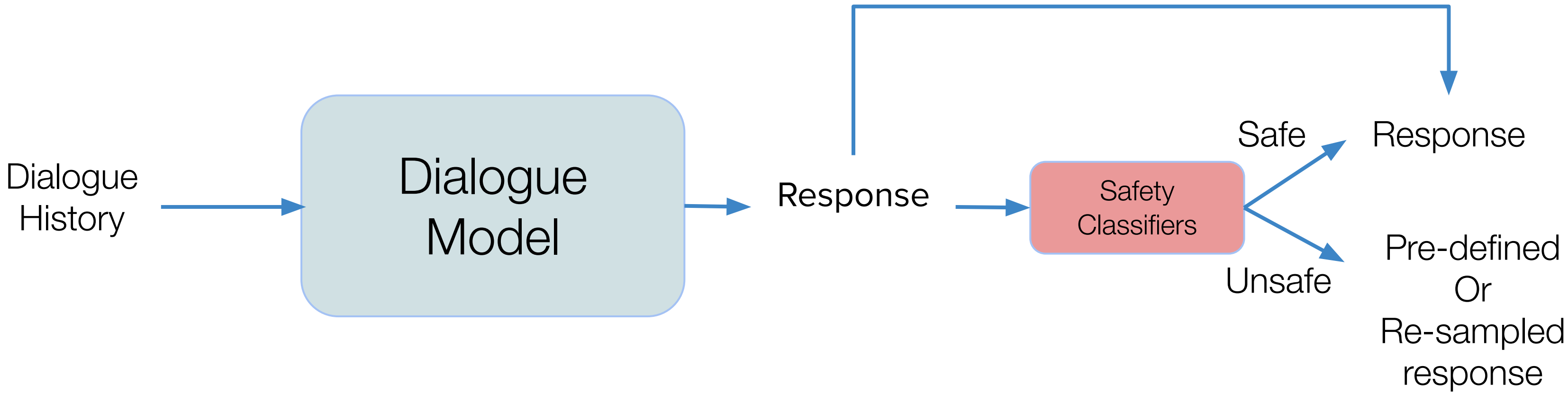
Block the n-gram from sensitive word list during decoding



Ref: Recipes for Safety in Open-domain Chatbots

3.4 Mitigating Inappropriate Responses: Safety Layers

Add classifiers to detect Inappropriate (e.g.,toxical, biased, unethical) response, and replace the unsafe responses with pre-defined or re-sampled safe responses.



Ref: [Recipes for Safety in Open-domain Chatbots]

(Part 3) Challenges and Future Work of Conversational AI

3.1. Reinforcement Learning/Self-Chat
3.2. Few-Shot/Zero-Shot Learning
3.3. Lifelong Learning
3.4. Mitigating Inappropriate Response
3.5. Multimodal
3.6. Evaluation
3.7. Shared Tasks & Datasets

3.5 Multimodal Dialogue Datasets

A: Erratic *B: Skeptical*

A: What is the difference between the forest and the trees? Oh look, dry pavement.

B: I doubt that's even a forest, it looks like a line of trees.

A: There's probably more lame pavement on the other side!

Figure from Image-Chat

Multimodal dialogues: conversations grounded on images, VR environment.

- Situated and Interactive Multimodal Conversations
- Multimodal domain-aware conversations (MMD)
- Image-Chat
- TALK THE WALK
- CLEVR-Dialog
- MELD

(Part 3) Challenges and Future Work of Conversational AI

3.1. Reinforcement Learning/Self-Chat

3.2. Few-Shot/Zero-Shot Learning

3.3. Lifelong Learning

3.4. Mitigating Inappropriate Response

3.5. Multimodal

3.6. Evaluation

3.7. Shared Tasks & Datasets

3.6 Automatic Evaluation

Evaluating dialogue systems is extremely challenging, especially for automatic metrics:

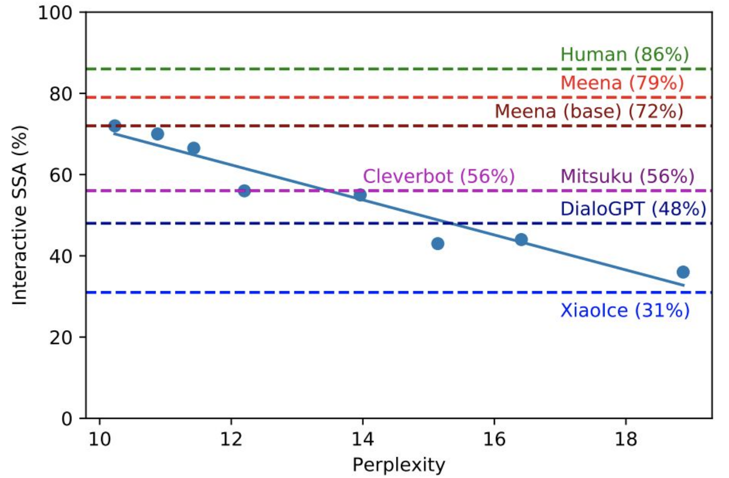
- N-gram based (e.g., BLEU) ⇒ Fails to capture the semantic meaning of the response (Liu et. al.,2016)

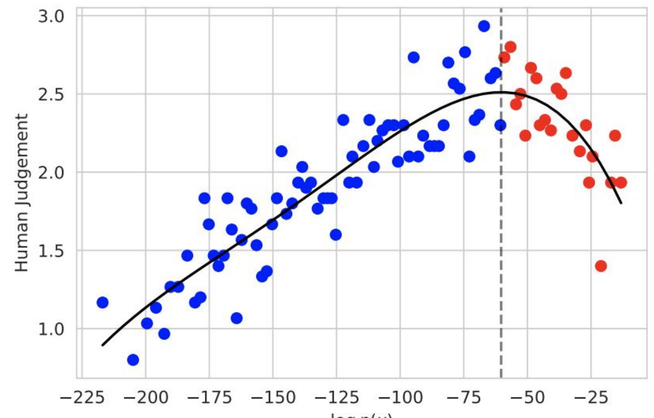
  Speaker A: Hey, what do you want to do tonight?
  Speaker B: Why don't we go see a movie?
  Model Response: Nah, let's do something active.
  Reference Response Yeah, the film about Turing looks great!

- Turn-level evaluation cannot capture repetition and consistency between turns

3.6 Evaluation:
The curious case of Perplexity

- [Towards a Human-like Open-Domain Chatbot (Meena-Bot)] showed correlation between Perplexity and Interactive Human Evaluation
- [Trading Off Diversity and Quality in Natural Language Generation] The likelihood Trap ⇒ if the perplexity of the model is too low the correlation with human judgement decreases





3.6 Evaluation: The Chicken and Egg Problem

(Part 3) Challenges and Future Work of Conversational AI

3.1. Reinforcement Learning/Self-Chat
3.2. Few-Shot/Zero-Shot Learning
3.3. Lifelong Learning
3.4. Mitigating Inappropriate Response
3.5. Multimodal
3.6. Evaluation
3.7. Shared Tasks & Datasets

3.6 Shared Tasks : Good data resource

DSTC: Dialog System Technology Challenge

- DSTC6, DSTC7, DSTC8
- DSTC9 (Current)
  - SIMMC: Situated Interactive Multi-Modal Conversational AI
  - Interactive Evaluation of Dialog
  - Multi-domain Task-oriented Dialog Challenge II
  - Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access

Other challenges

- SLT 2018 Microsoft Dialogue Challenge
- The Conversation Intelligence Challenge: ConvAI2 - PersonaChat
- DialogueGLUE
- Alexa Prize - SocialBot Grand-Challenge

Summary of datasets

Seq2Seq

- Ubuntu Dialogue
- DailyDialog
- Twitter Conv.
- ReddiT Conv
- OpenSubtitles

Personalized

- Persona Chat
- Tweeter-Persona
- Personalized End-to-End Goal-Oriented

Textual Knowledge

- WoW
- CoQA
- TopicChat
- CMUDoG
- HollE
- Conv.ByReading

Graph Knowledge

- OpenDialKG
- DyKgChat
- KdConv
- Commonsense Graph Attention
- Dialog Coherence

Tabular Knowledge

- SMD
- Camrest
- MultiWoz
- bAbI-Dialogues

API Service

- bAbI
- Camrest
- MultiWoz
- CrossWoz
- SGD
- TaskMaster 1-2-3

Emotion Dialogue

- Empathetic Dialogues
- DailyDialogues
- MojiTalk

Putting all together

- The Dialogue Dodecathlon
- Blend Skills
- Chit-Chats Enhanced Task-Oriented

(Part 3) Challenges and Future Work of Conversational AI

3.1. Human In the Loop Reinforcement Learning/Self-Chat

3.2. Better Strategy for Few-Shot/Zero-Shot Learning

3.3. Lifelong Learning with User Experience

3.4. Mitigating Inappropriate Response On the Model

3.5. Multimodal Is Still A Grand Challenge

3.6. Better Automatic Evaluation

3.7. More Shared Tasks & Datasets

(Part 3) Challenges and Future Work of Conversational AI

3.1. Human In the Loop Reinforcement Learning/Self-Chat

3.2. Better Strategy for Few-Shot/Zero-Shot Learning

3.3. Lifelong Learning with User Experience

3.4. Mitigating Inappropriate Response On the Model

3.5. Multimodal Is Still A Grand Challenge

3.6. Better Automatic Evaluation

3.7. More Shared Tasks & Datasets

END