Applied Deep Learning

End-to-End Conversational AI

June 7th, 2021 **http://adl.miulab.tw**

Slides credited from NeurIPS 2020 Tutorial

National Taiwan University
國立臺灣大學

2

Why and When We Need?

"I want to chat"

"I have a question"

"I need to get this done"

"What should I do?"

Turing Test (talk like a human) — Social Chit-Chat

Information consumption

Task completion

Decision support

— Task-Oriented Dialogues

- *What is today's agenda?*
- *What does NLP stand for?*

- *Book me the train ticket from Kaohsiung to Taipei*
- *Reserve a table at Din Tai Fung for 5 people, 7PM tonight*
- *Schedule a meeting with Vivian at 10:00 tomorrow*

- *Is this course good to take?*

3

Two Branches of Conversational AI

Chit-Chat

Seq2seq models → Seq2seq with conversation contexts → Knowledge-grounded seq2seq models

Task-Oriented

Single-domain, system-initiative → Multi-domain, contextual, mixed-initiative → End-to-end learning, massively multi-domain

4 Vanilla Seq2Seq ConvAI: How

A simple 4-step 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

5

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?

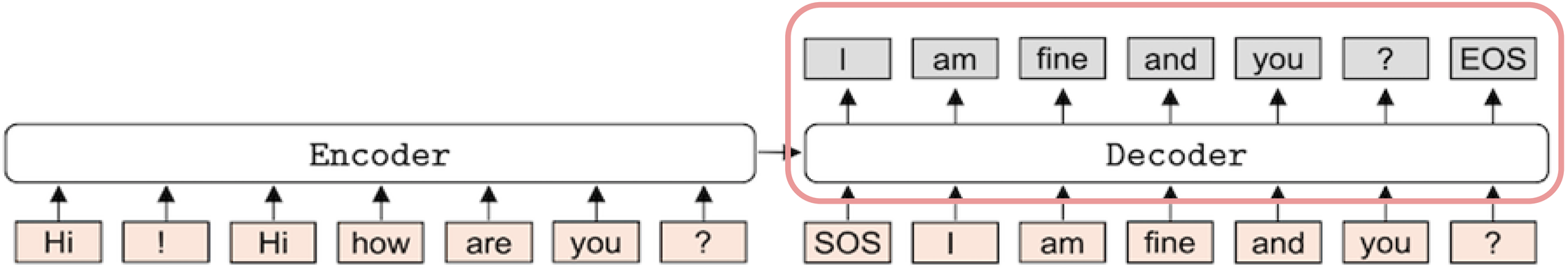
Human1: I'm supposed to work on a paper that's due on Monday.

Human-to-Human Conversations:

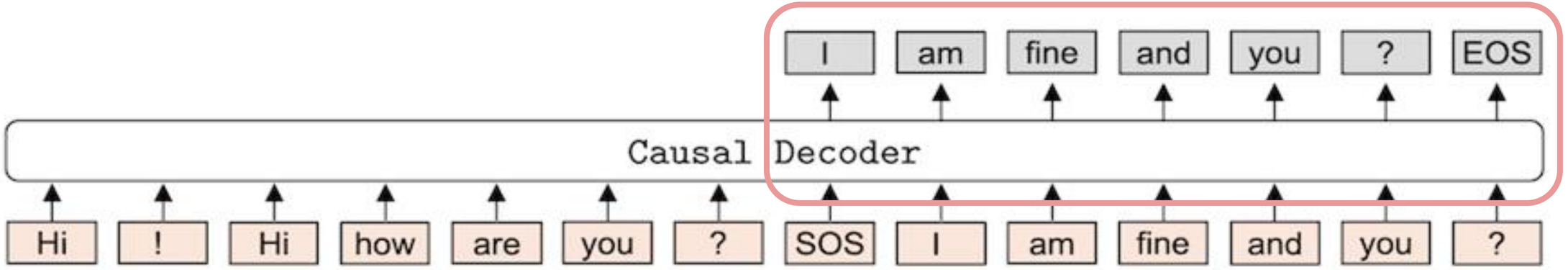
- Daily Dialog
- Ubuntu Dialogue Corpus
- Twitter Conversations
- Reddit Conversational Data
- OpenSubtitles

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

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)

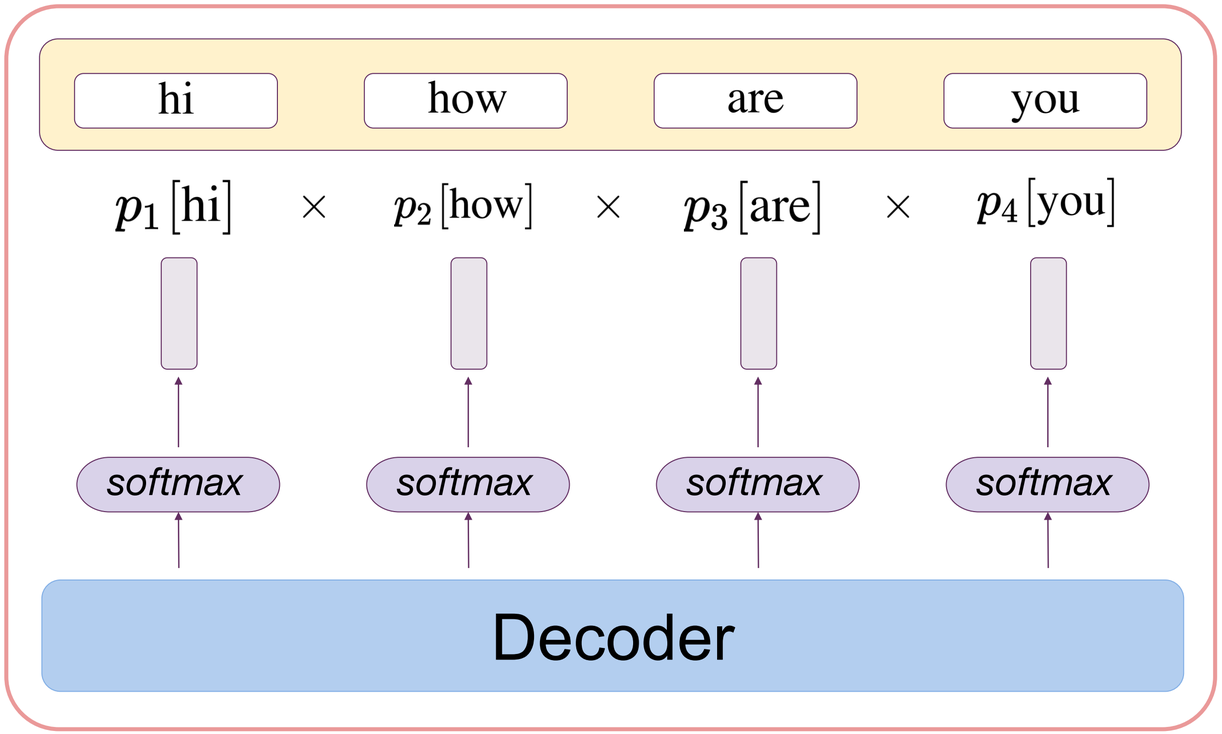
7

Vanilla Seq2Seq ConvAI: Supervised Learning

Maximum Likelihood Estimation (MLE)

⇒ maximizing the conditional probability of the response given the history

⇒ The model output is a probability distribution over the vocab

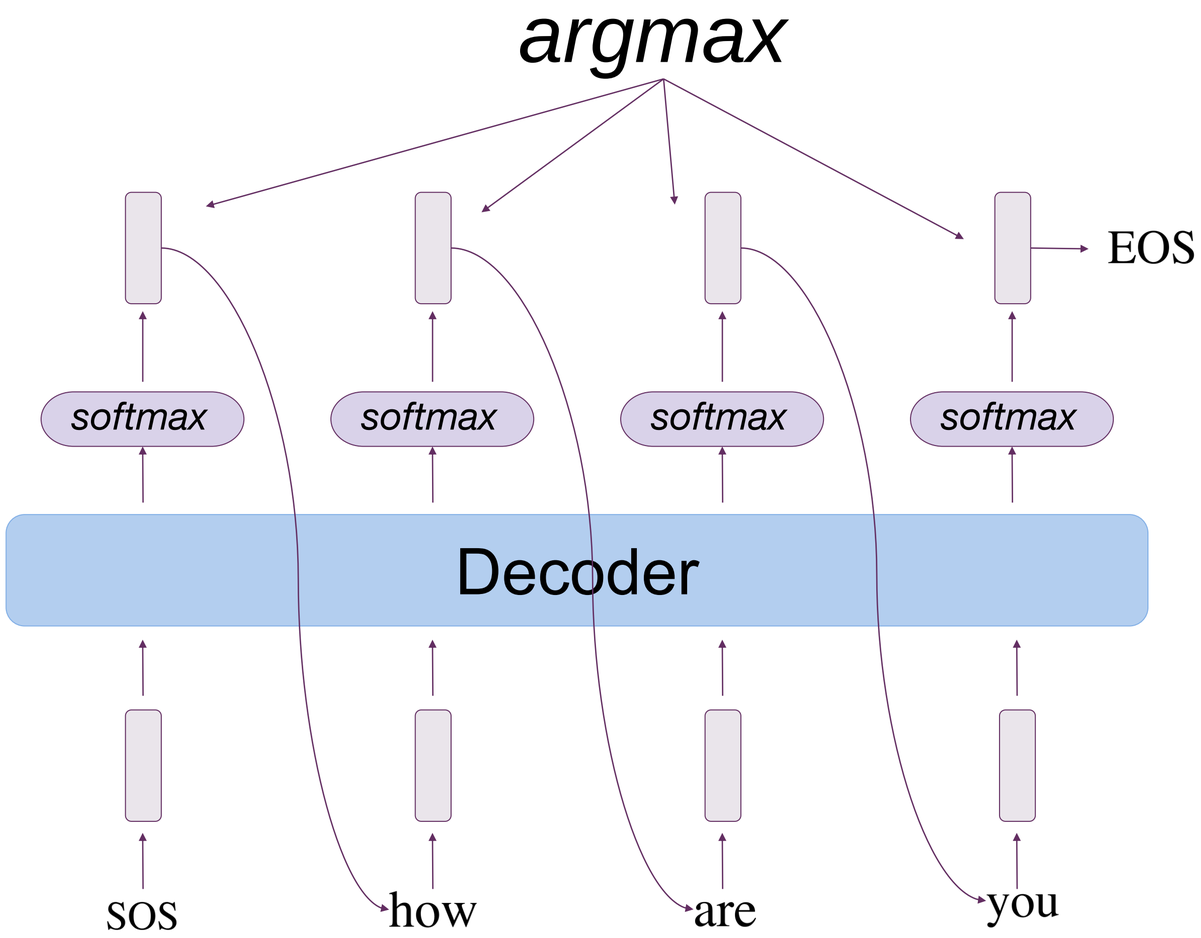


$$\mathbb{P}(y_1, \ldots, y_m) = \prod_{i=1}^{m} p_i[y_i]$$

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

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

Vanilla Seq2Seq ConvAI: Greedy Decoding



Vanilla Seq2Seq ConvAI: How

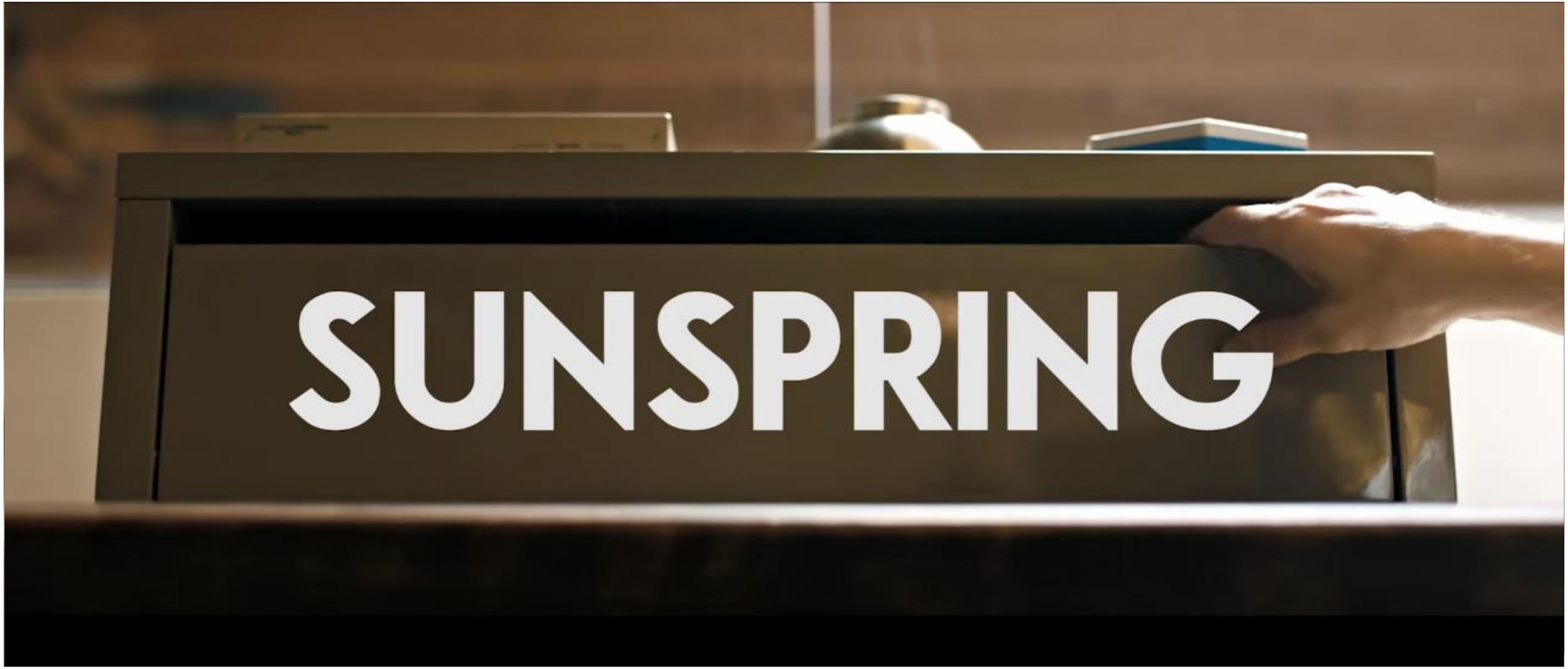
A simple 4-step recipe:

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

10

Sci-Fi Short Film - SUNSPRING

https://www.youtube.com/watch?v=LY7x2Ihqj



11

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

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.

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.

Limitation: Lack of Empathy

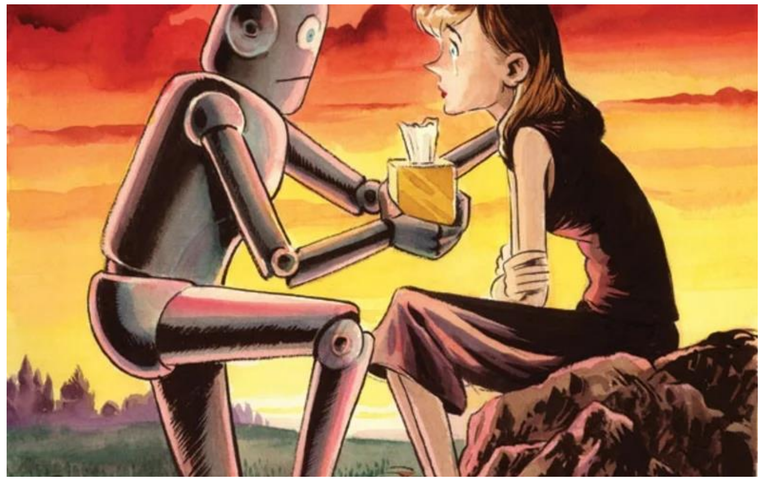
Human: I messed up my project today.

System: What is your project?

Human: I think I might get fired...

System: Then you need to find another job!

Human: 😢



Limitation: Lack of Controllability

Human: Hi, how are you?

Model1: I am good thanks

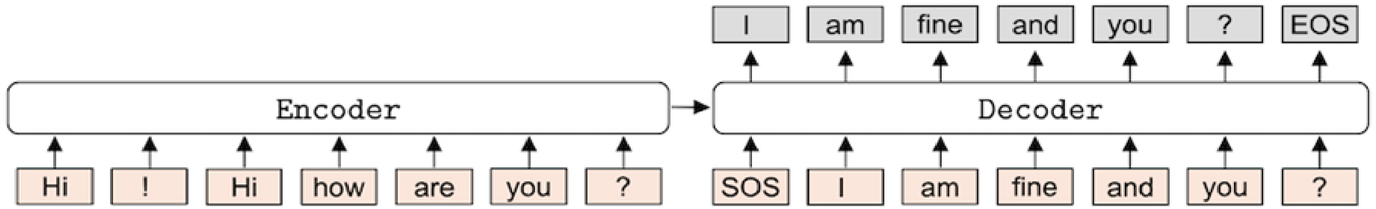
Model2: I had really a bad day

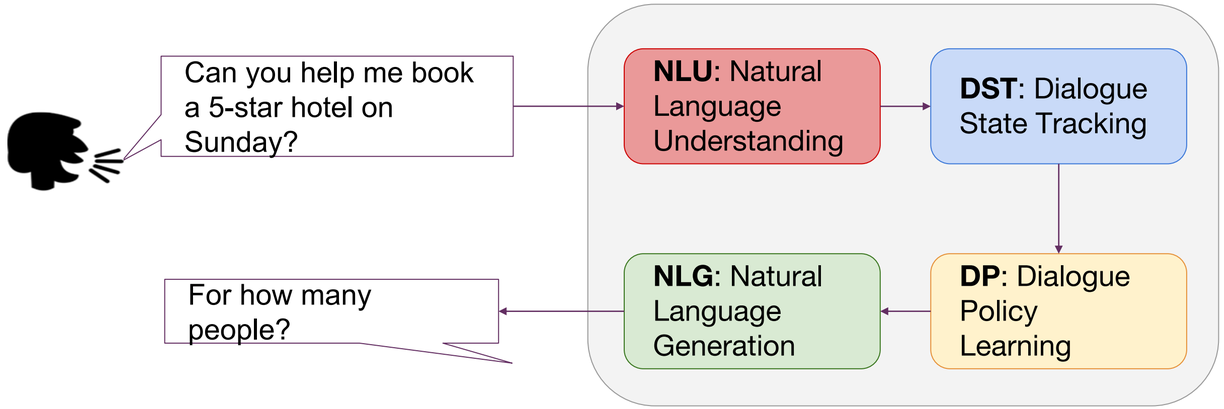
Model3: I am okay, how was your day?

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

→ no control for style, topics, engagement, toxic and inappropriate responses

Limitation: Lack of Versatility

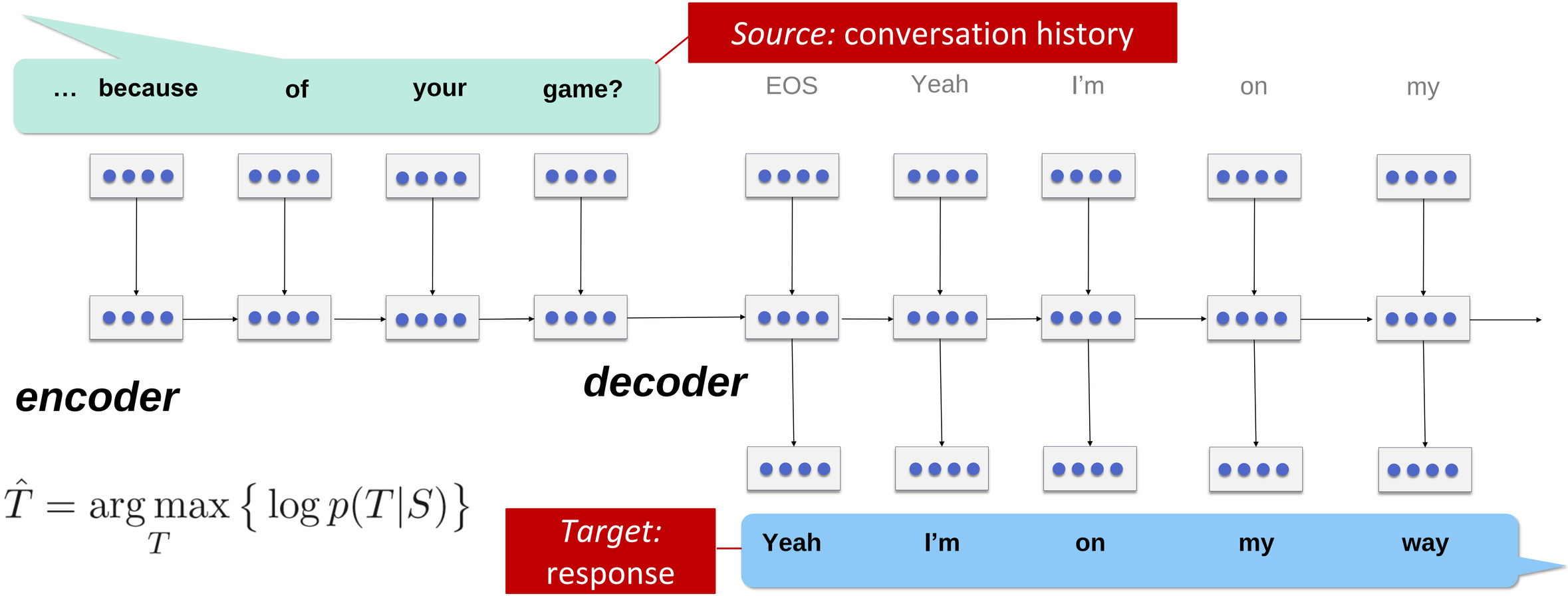




→ cannot handle task-oriented conversations due to API requirement

Limitation: Lack of Global Optimization

- Turn-level optimization



$$\hat{T} = \arg\max_{T} \{ \log p(T|S) \}$$

18

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
7. Lack of global optimization

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

19

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
7. Lack of global optimization

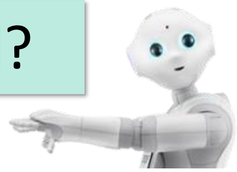
- These limitations of vanilla seq2seq make human-machine conversations boring and shallow. How can we overcome these limitations and move towards deeper conversational AI?

Limitation 1: Lack of Diversity

Wow sour starbursts really do make your mouth water... mm drool. Can I have one?

Of course!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?



I don’t know.

‘tis a fine brew on a day like this! Stro

I'm not sure yet,

Well he was on in Bromley a wh

I don't even know what he's talking about.

32% responses are general and meaningless

“I don’t know”

“I don’t know what you are talking about”

“I don’t think that is a good idea”

“Oh my god”

Solution: Diversify Responses

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

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

2. Model architecture ⇒ Conditional Variational Autoencoder (CVAE)

$$p(T|S) \;\Rightarrow\; p(T|z,S)p(z|S)$$

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

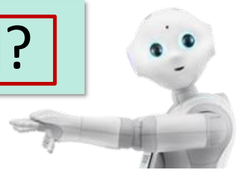
MMI for Response Diversity (Li et al., 2016)

Wow sour starbursts really do make your mouth water... mm drool.

Can I have one?

Of course you can! They're delicious!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?



I think he'd be a good signing.

'tis a fine brew on a day like this! Strong though, how many is sensible?

Depends on how much you drink!

Well he was on in Bromley a while ago... still touring.

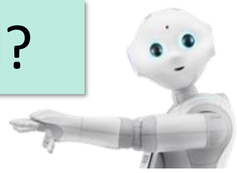
I've never seen him live.

MMI for Response Diversity (Li et al., 2016)

Wow sour starbursts really do make your mouth water... mm drool. Can I have one?

Of course you can! They're delicious!

Milan apparently selling Zlatan to balance the books... Where next, Madrid?



I think he'd be a good signing.

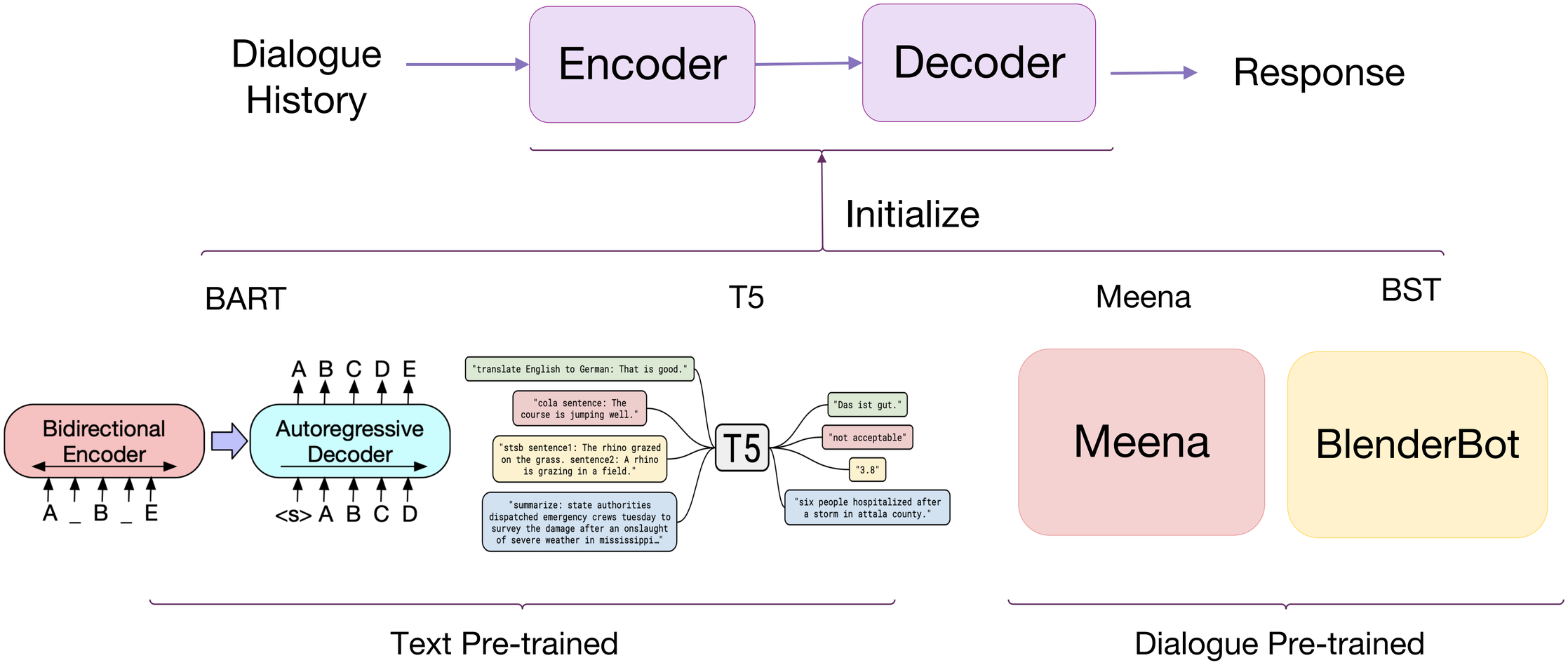
'tis a fine brew on a day like this! Strong though, how many is sensible?

Depends on how much you drink!

Well he was on in Bromley a while ago... still touring.

I've never seen him live.

Diversify by Large-Scale Pretraining

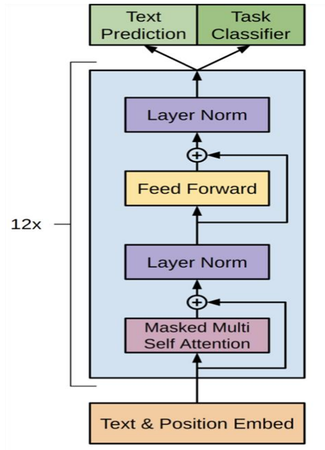


Diversify by Large-Scale Pretraining

Dialogue History → Causal Decoder → Response

Initialize

GPT-1/2/3



Text Pre-trained

DialoGPT

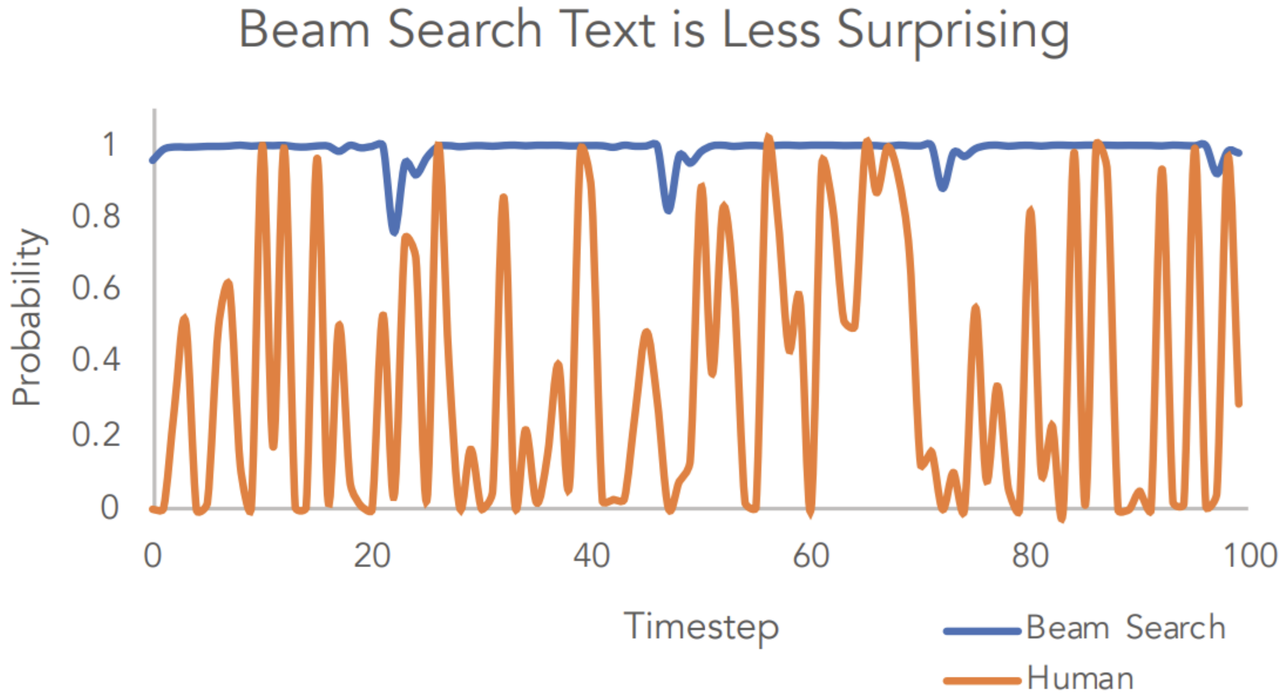
DialoGPT

Dialogue Pre-trained

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

Diversify by Nucleus Sampling

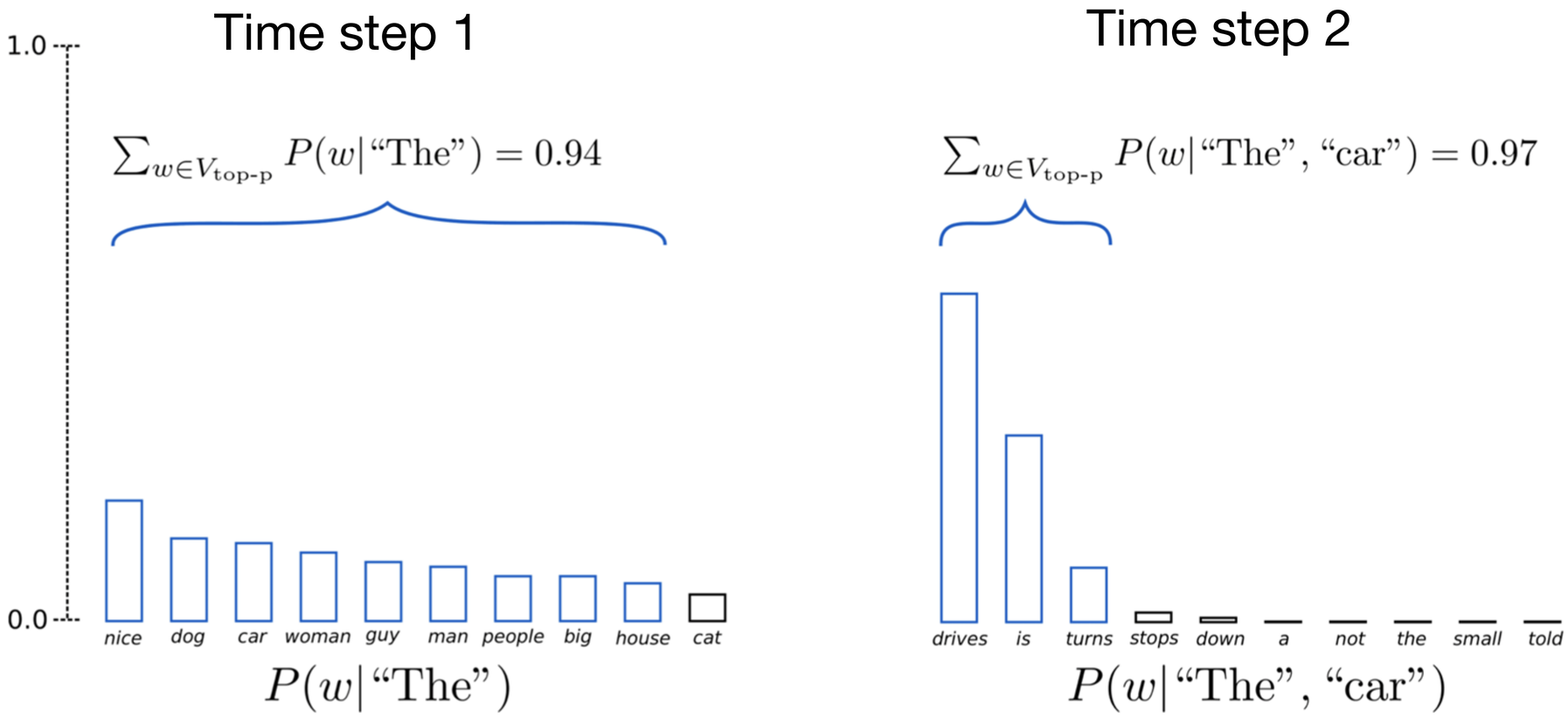


Figure from: https://huggingface.co/blog/how-to-generate

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
7. Lack of global optimization

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

Limitation 2: Lack of Consistency

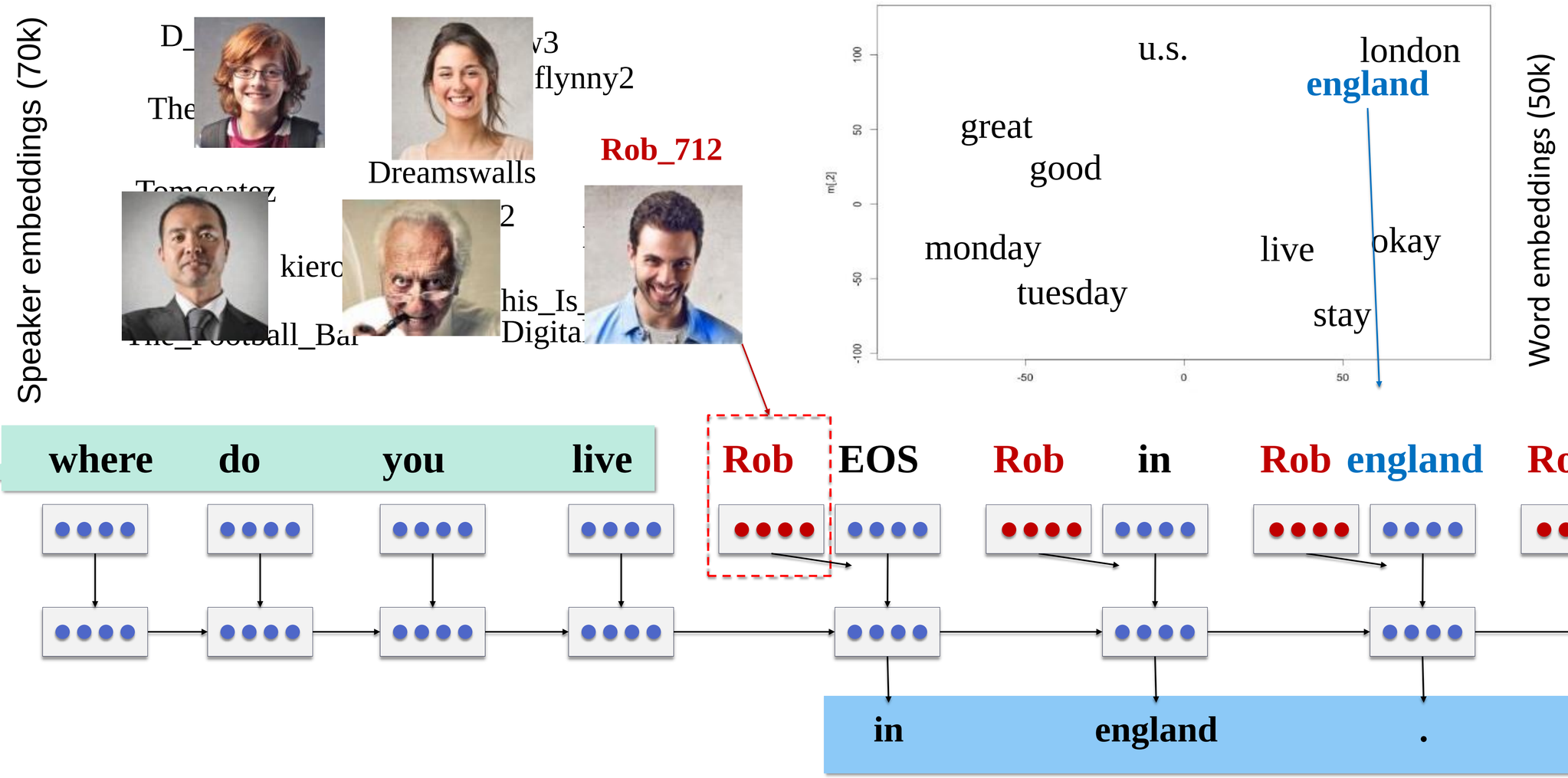
message	Where do you live now?
response	I live in Los Angeles.
message	In which city do you live now?
response	I live in Madrid.
message	In which country do you live now?
response	England, you?
message	Where were you born?
response	I was born in Canada.
message	Where are you from?
response	England, you?
message	Where did you grow up?
response	I grew up in Texas.

message	How old are you?
response	16 and you?
message	What's your age?
response	18.
message	What is your major?
response	I'm majoring in psychology
message	What did you study in college?
response	English lit.

Solution: Personalization

1. Learning speaker embedding:
   - Speaker Model
2. Conditioning on persona descriptions:
   - PersonaChat Dataset
   - TransferTransfo Model

Personalization via Speaker Model



Persona Model for Consistency (Li et al., 2016)

Baseline model → inconsistency

message	Where is your hometown?
response	I live in Indonesia.
message	Where are you from?
response	England, you?
message	In which city do you live now?
response	I live in Jakarta.
message	In which country do you live now?
response	I live in Indonesia.
message	What is your major?
response	Business, you?
message	What did you study in college?
response	Psychology, you?
message	How old are you?
response	I'm 18.
message	What is your age?
response	I'm 16.

Persona model using speaker embedding → consistency

message	Where is your hometown?
response	I'm from England.
message	Where are you from?
response	I'm from England.
message	In which city do you live now?
response	I live in London.
message	In which country do you live now?
response	I live in England.
message	What is your major?
response	Business. You?
message	What did you study in college?
response	I did business studies.
message	How old are you?
response	I'm 18.
message	What is your age?
response	I'm 18.

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

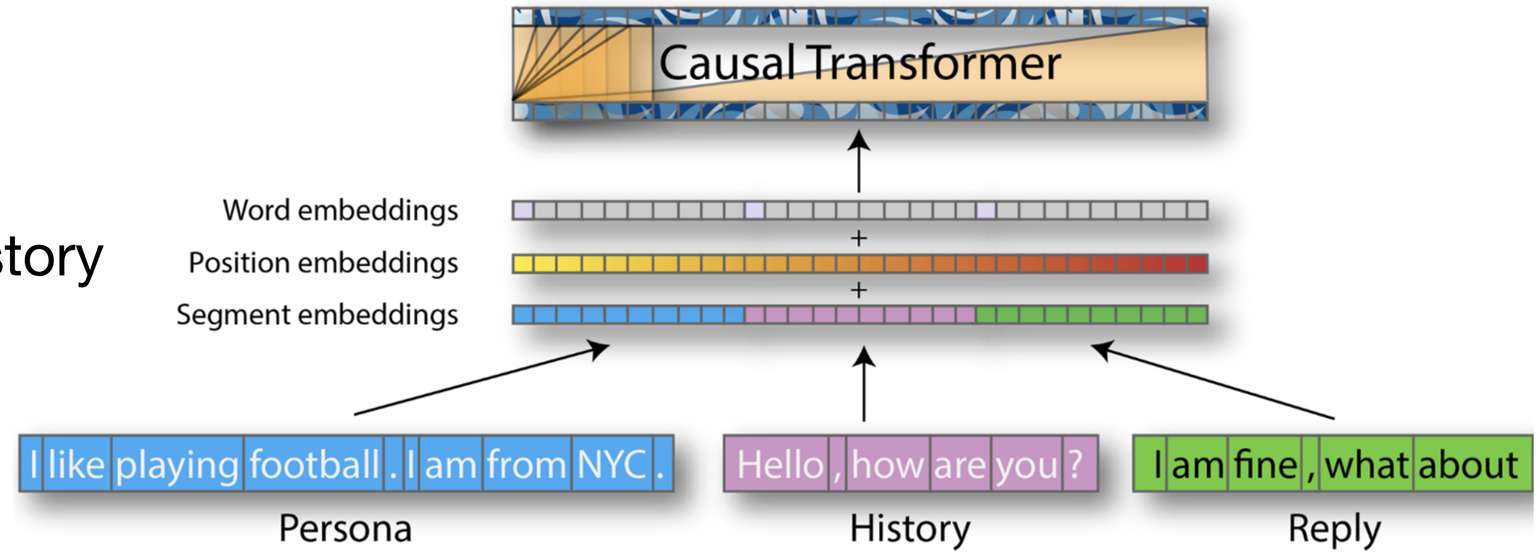
- Persona Chat
- Tweeter Personalized
- Learning Personalized End-to-End Goal-Oriented Dialog

Personalization via TransferTransfo Model

Dialogue History
+
Persona Description

→ Decoder-only → Response

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



35

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
7. Lack of global optimization

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

Limitation 3: Lack of Knowledge

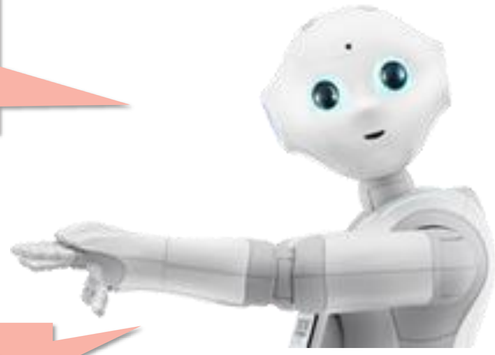
Social Chat
Engaging, Human-Like Interaction
(Ungrounded)

Task-Oriented
Task Completion, Decision Support
(Grounded)

The weather is so depressing these days.

I know, I dislike rain too.
What about a day trip to eastern Washington?

Any recommendation?



Try Dry Falls, it's spectacular!

Conversation and Non-Conversation Data

*You know any good **A** restaurant in **B**?*

*Try **C**, one of the best **D** in the city.*

Conversation Data

Knowledge Resource

*You know any good **Japanese** restaurant in **Seattle**?*

*Try **Kisaku**, one of the best **sushi restaurants** in the city.*

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

39

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:

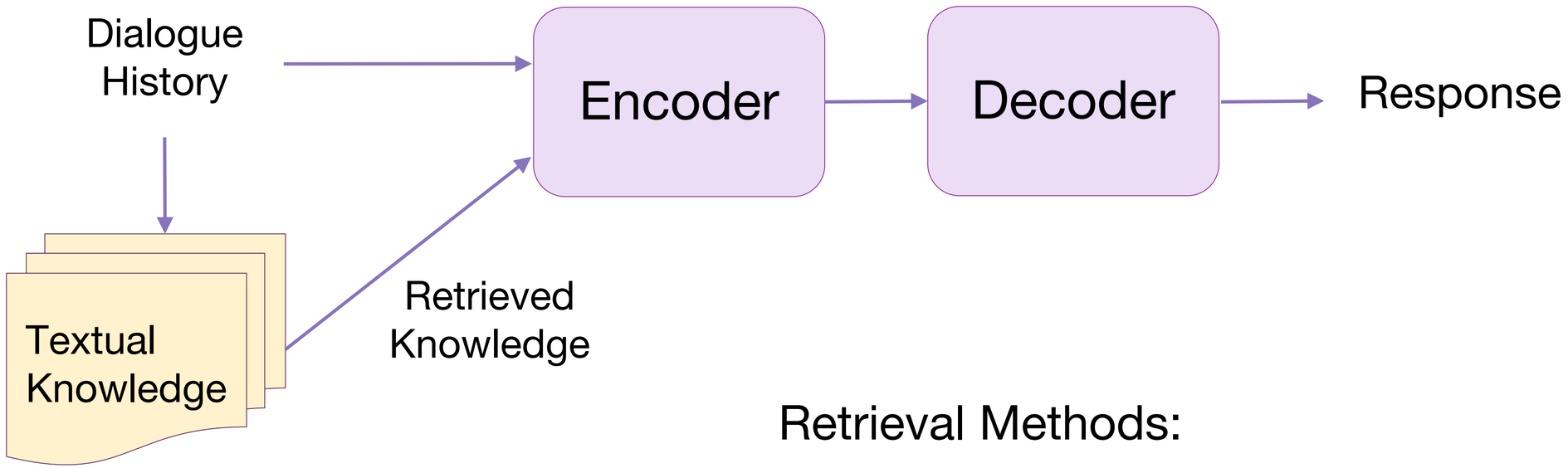
Blue is one of the three primary colours in the RGB colour model. It lies between violet and green on the spectrum of visible light.

Wizard: It is right between violet and green.

Human-to-Human Conversations + Textual Knowledge

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

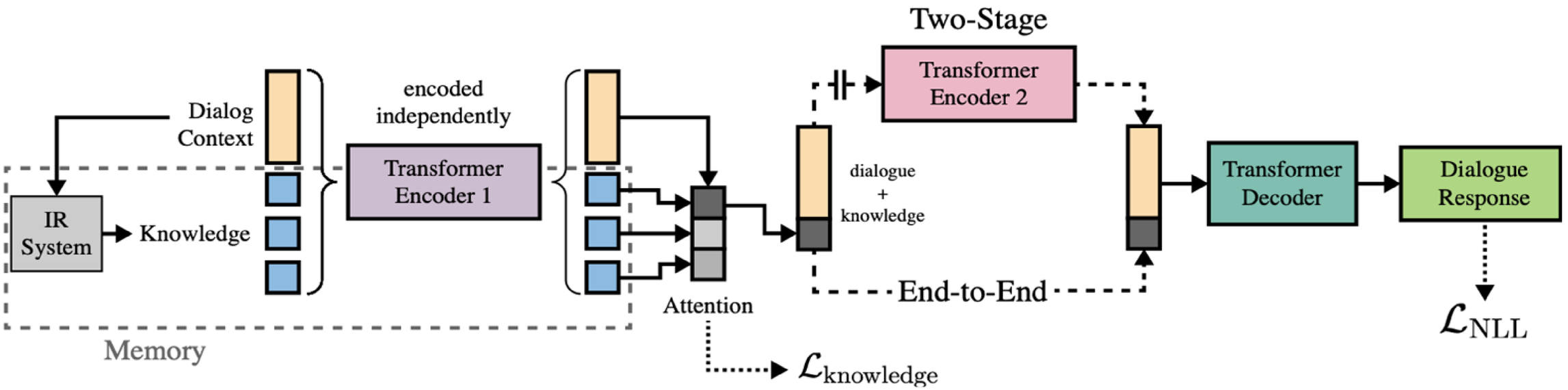
Models with Textual Knowledge



Retrieval Methods:

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

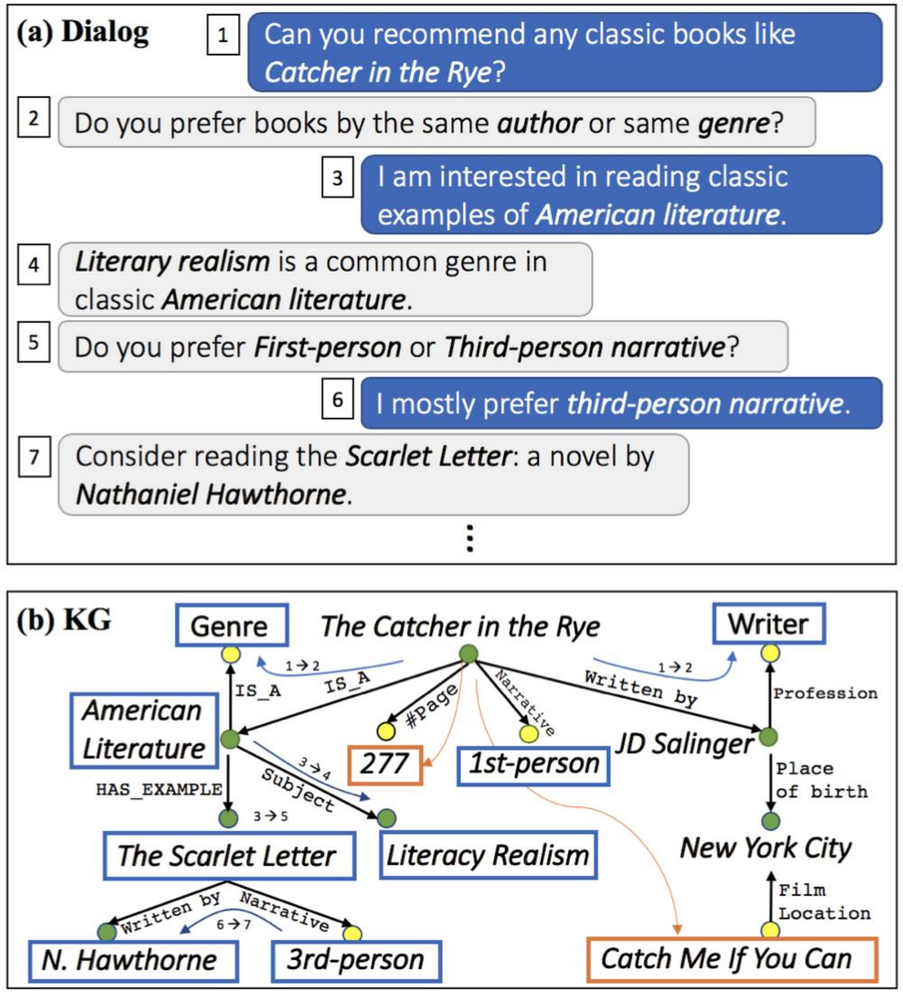
Knowledge: IR Systems + Model



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

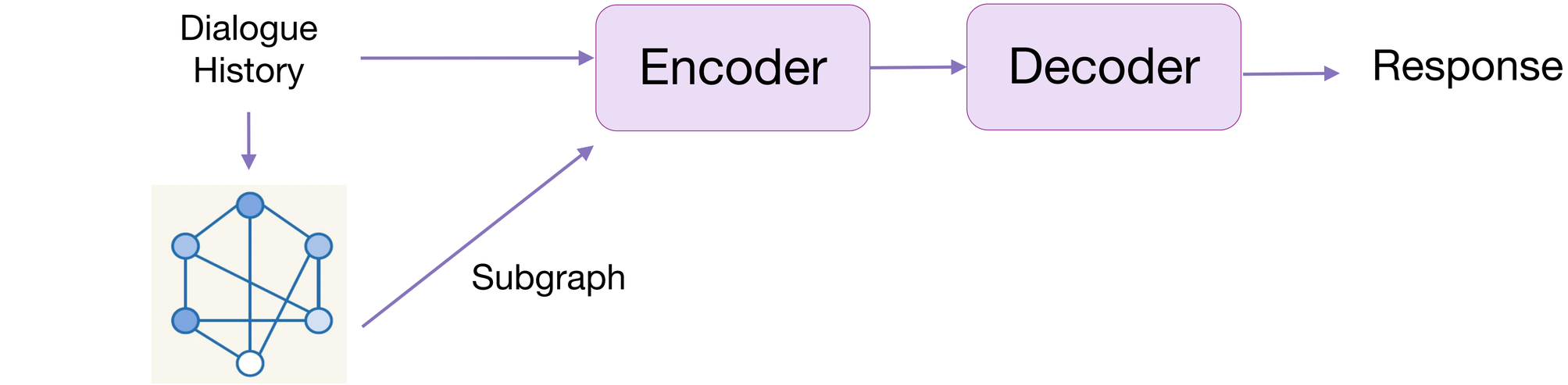
Graph Knowledge



Human-to-Human Conversations + Graph KG

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

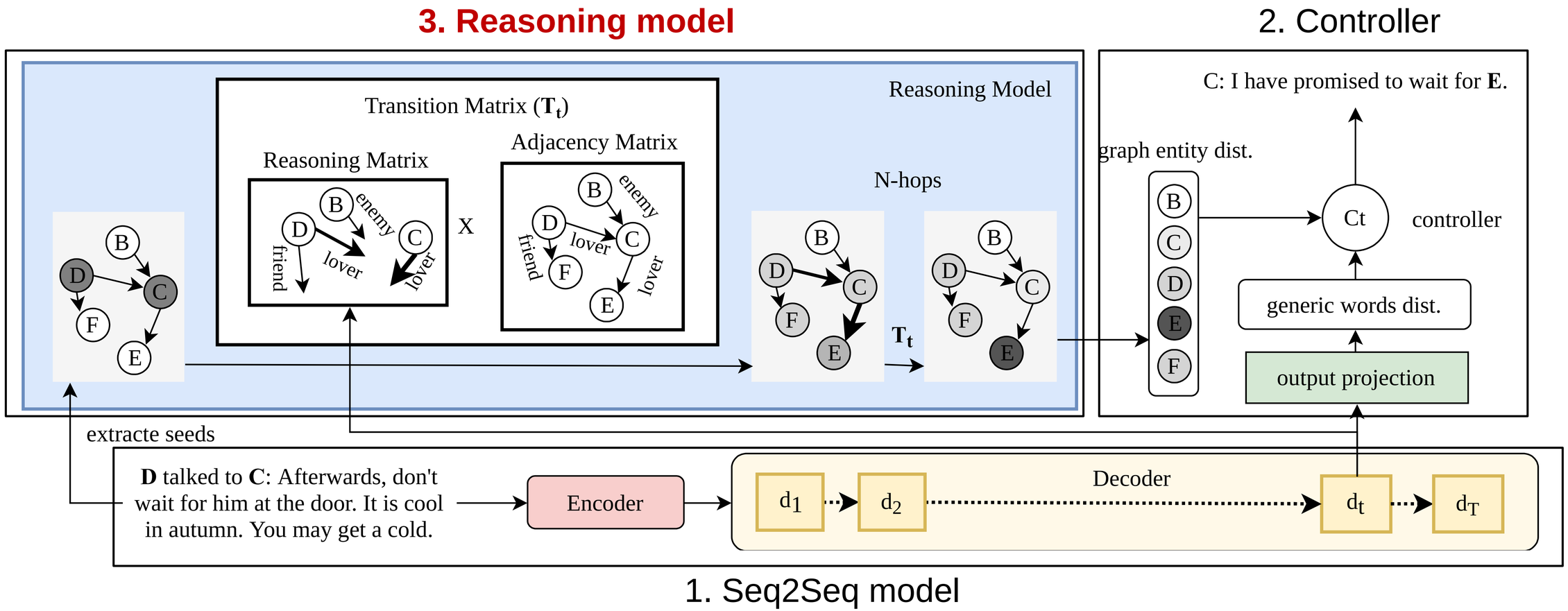
Models with Graph Knowledge



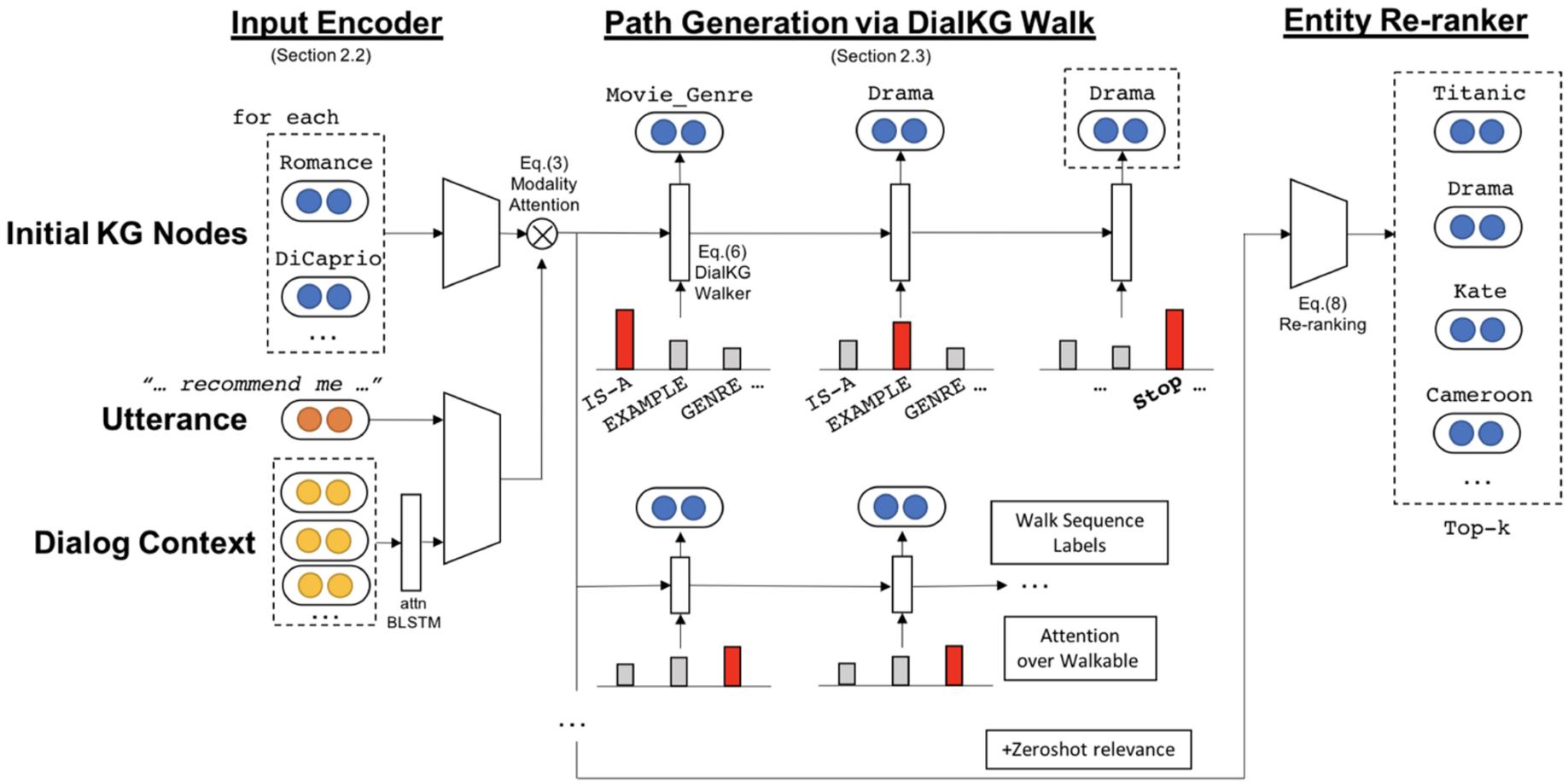
Subgraph Retrieval:

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

DyKgChat: Quick Adaptive Model (Qadpt)



OpenDialKG Walker: Subgraph Retrieval



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

Tabular Knowledge

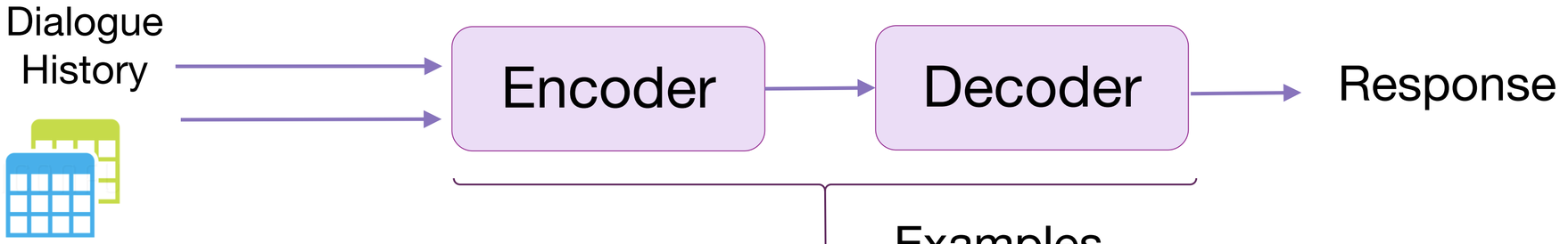
Event	Time	Date	Party	Agenda
swimming act.	3pm	the 11th	sister	-
dinner	7pm	the 3rd	mother	-
football	2pm	the 20th	mother	-
lab appt.	10am	the 17th	Jeff	-
...	...	...	...	...

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*

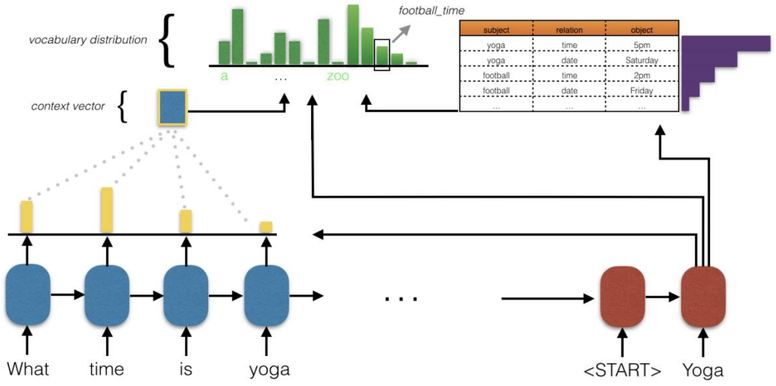
Human-to-Human Conversations + Table Knowledge

- SMD
- Camrest
- bAbI-Dialogues

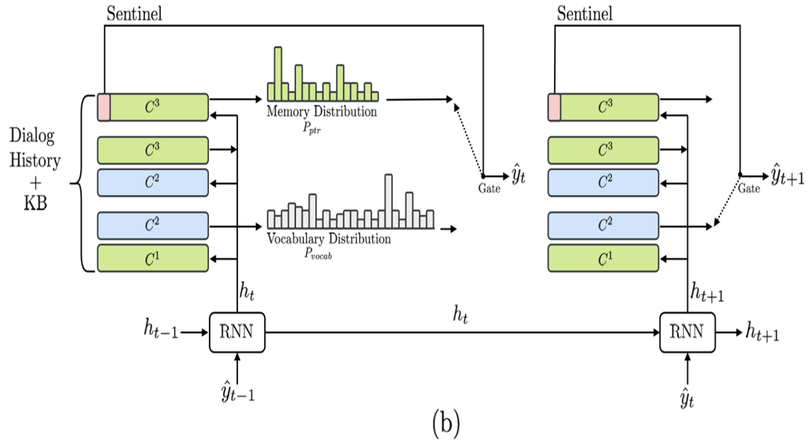
Models with Tabular Knowledge



KVR

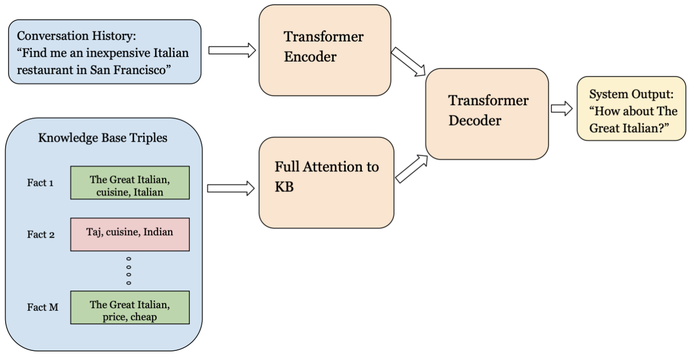


Mem2Seq

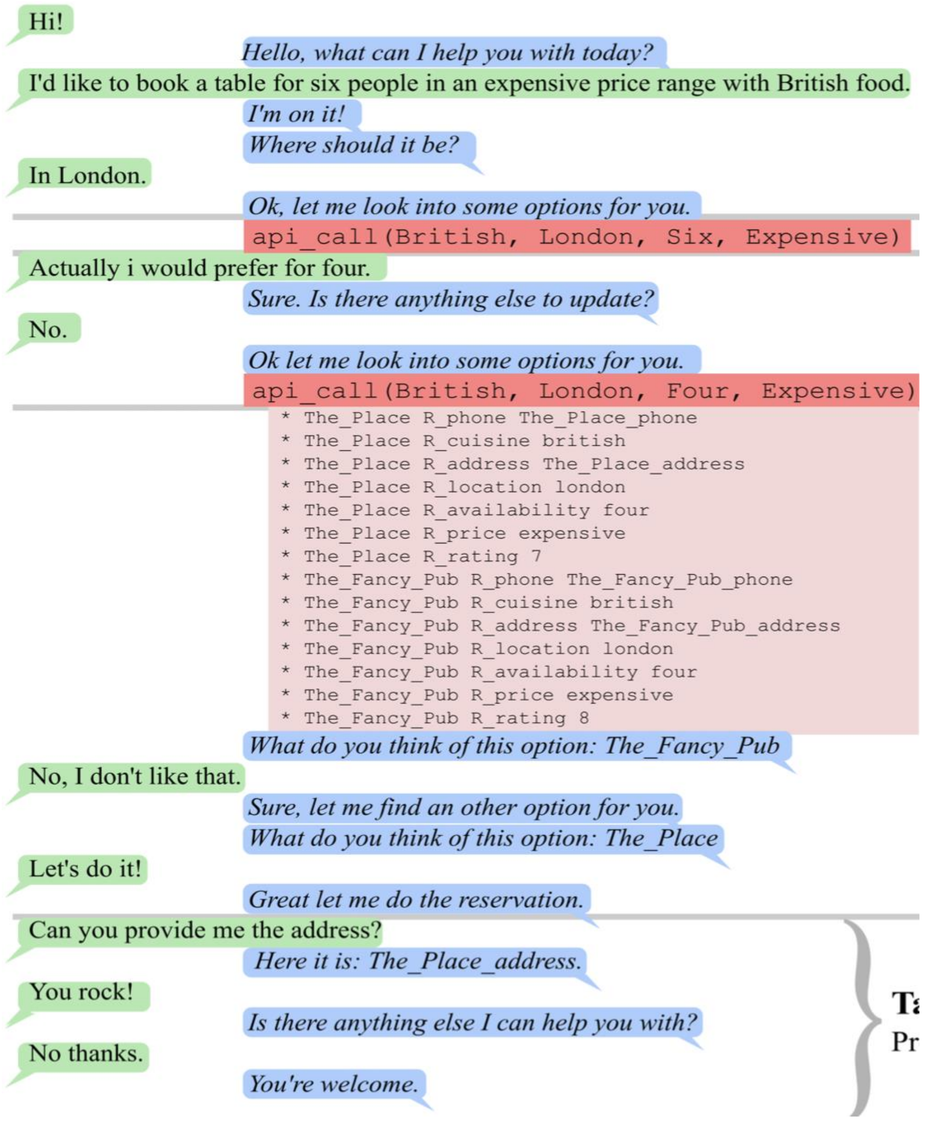


(b)

Neural Assistant



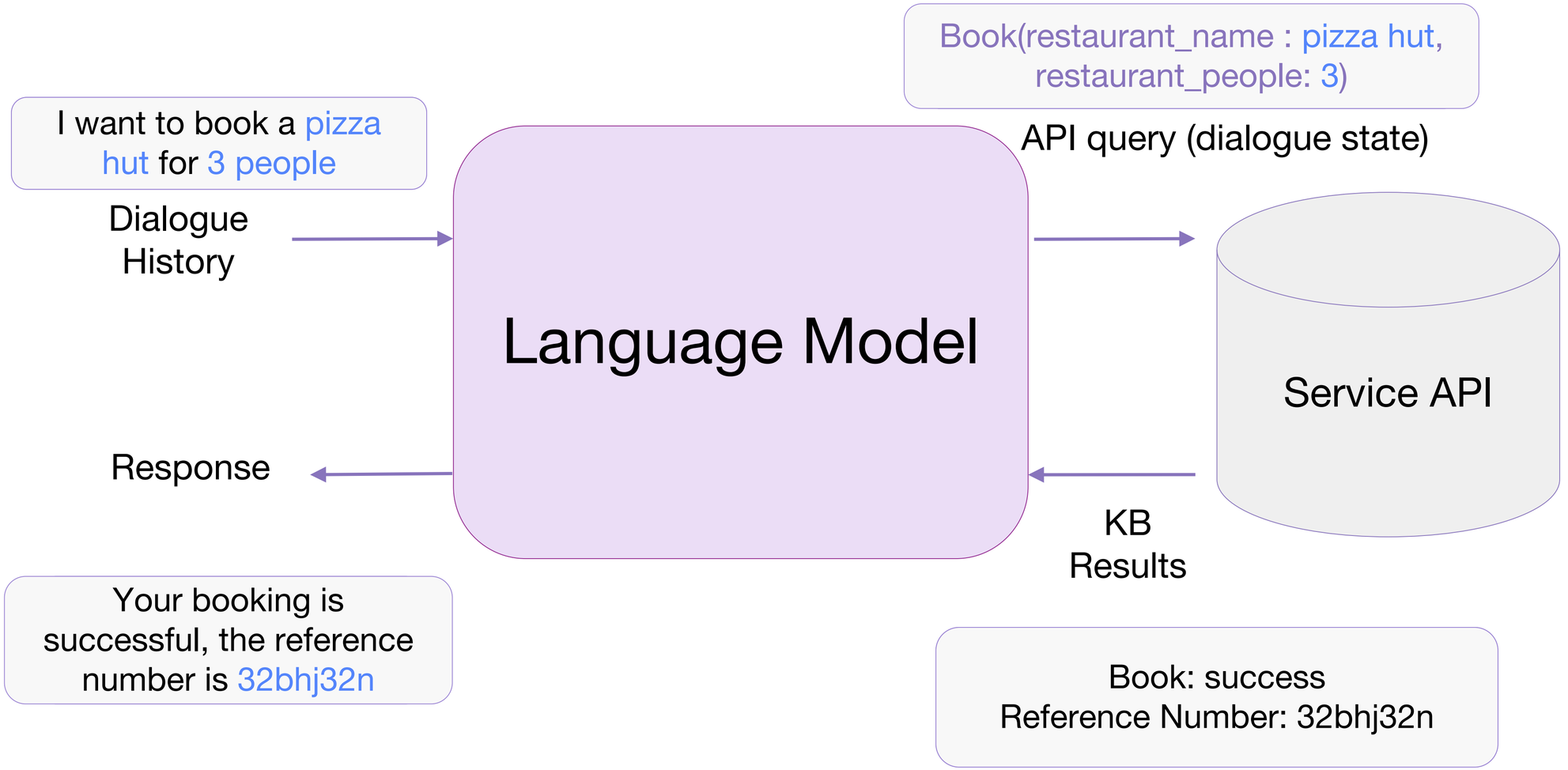
External Service API Interaction



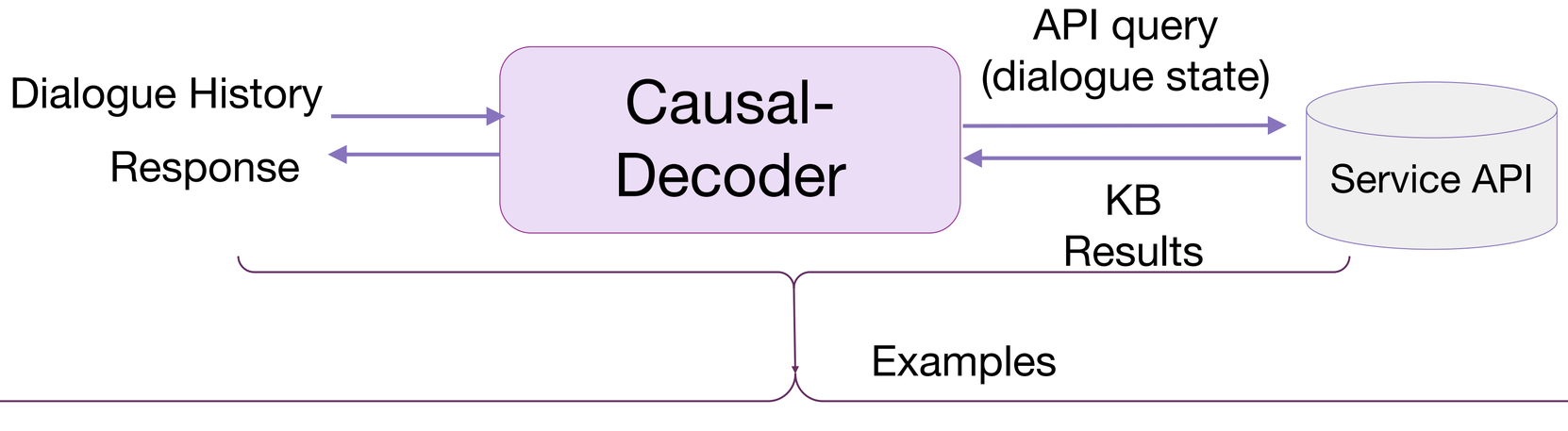
Human-to-Human Conversations + Table Knowledge

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

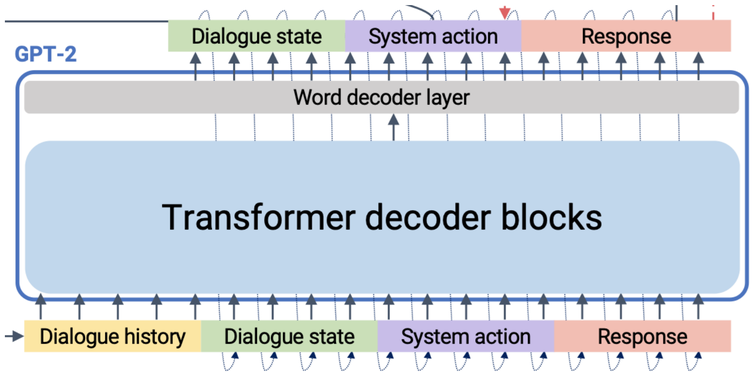
Models with Service API



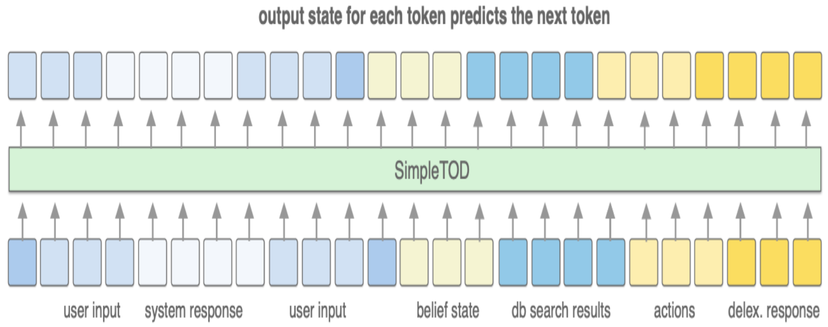
Models with Service API



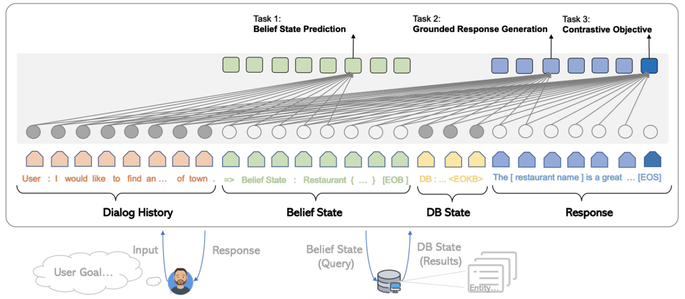
End-to-End GPT2 Neural Pipeline



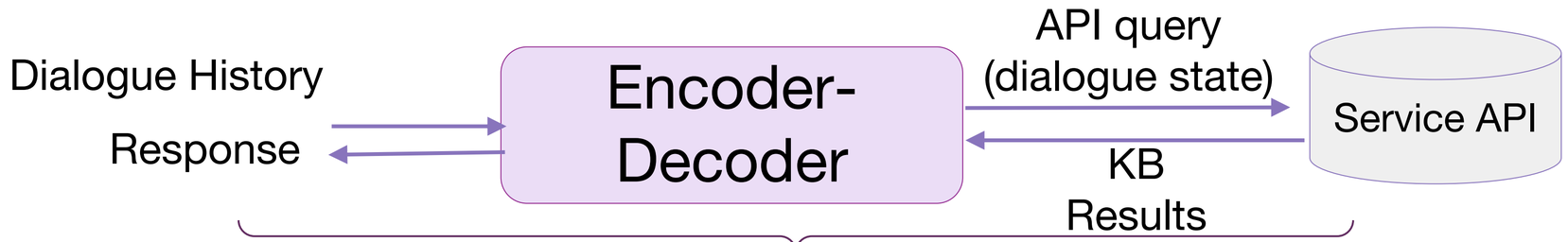
SimpleToD



SOLOIST



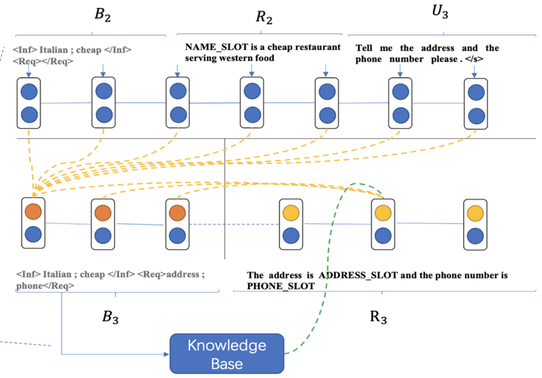
Models with Service API



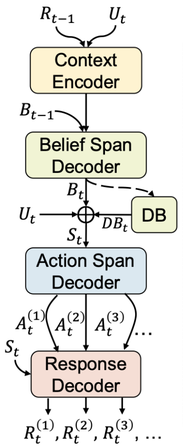
Examples

Sequicity

Turn	Dialogue
$User_1$	Can I have some Italian food please?
$Machine_1$	<Inf> Italian </Inf><Req> </Req> What price range are you looking for?
$User_2$	I want cheap ones.
$Machine_2$	<Inf> Italian ; cheap </Inf> <Req></Req> NAME_SLOT is a cheap restaurant serving western food
$User_3$	Tell me the address and the phone number please .
$Machine_3$	<Inf> Italian ; cheap </Inf> <Req>address ; phone</Req> The address is ADDRESS_SLOT and the phone number is PHONE_SLOT



DAMD



R_{t-1} : What price range do you want for the hotel?
U_t : A cheap one works for me. By the way it should be in the west.

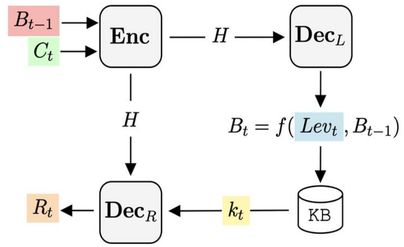
B_t: [hotel] area west ; price cheap S_t
A_t^U: inform
DB_t : match = 3; booking = available

name	price	area	Stars	booking	…
Avalon	cheap	west	4	available	…
…	…	…	…	…	…

$A_t^{(1)}$ [hotel] [inform] name [offerbook]
$A_t^{(2)}$ [hotel] [request] stars
$A_t^{(3)}$ [hotel] [recommend] name wifi

$R_t^{(1)}$ The <v.name> is a great choice meet your criteria! Do you want me to book it for you?
$R_t^{(2)}$ Sure! What star rating do you want?
$R_t^{(3)}$ I would recommend the <v.name>! It is …

MinTL



[hotel] stars 5 area centre day sunday [restaurant] food thai area centre day sunday name bangkok city <EOB> Can you help me book a 5 star hotel near the restaurant on the same day? <EOU>For how many people? <EOR>10 people <EOU>

<SOB>[hotel] people 10 <EOB>

<KB2> sorry, there are no matches. would you like to try another part of town? <EOR>

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
7. Lack of global optimization

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

Limitation 4: Lack of Empathy

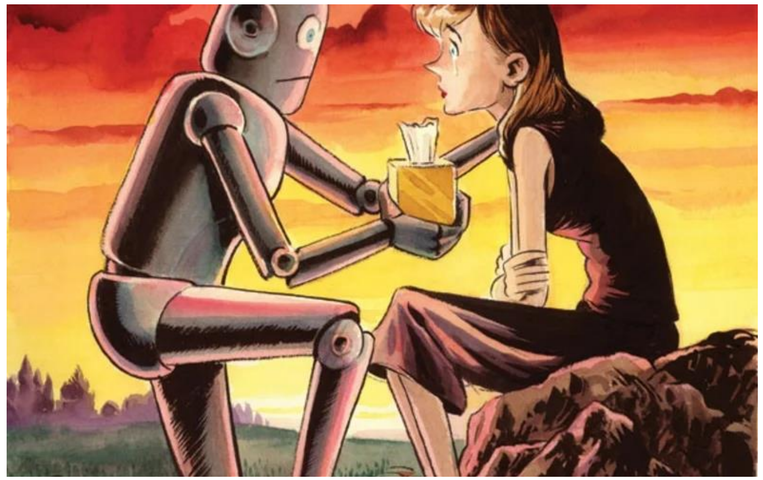
Human: I messed up my project today.

System: What is your project?

Human: I think I might get fired...

System: Then you need to find another job!

Human: 😢

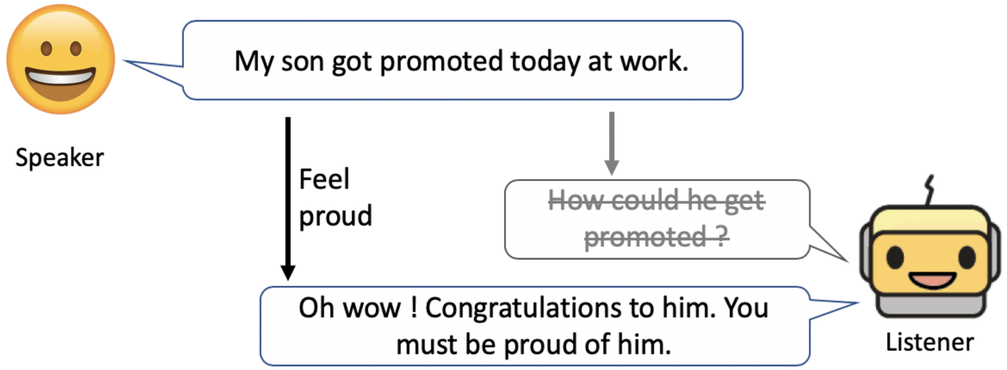


Solution: Empathic Generation

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

Empathy Dataset

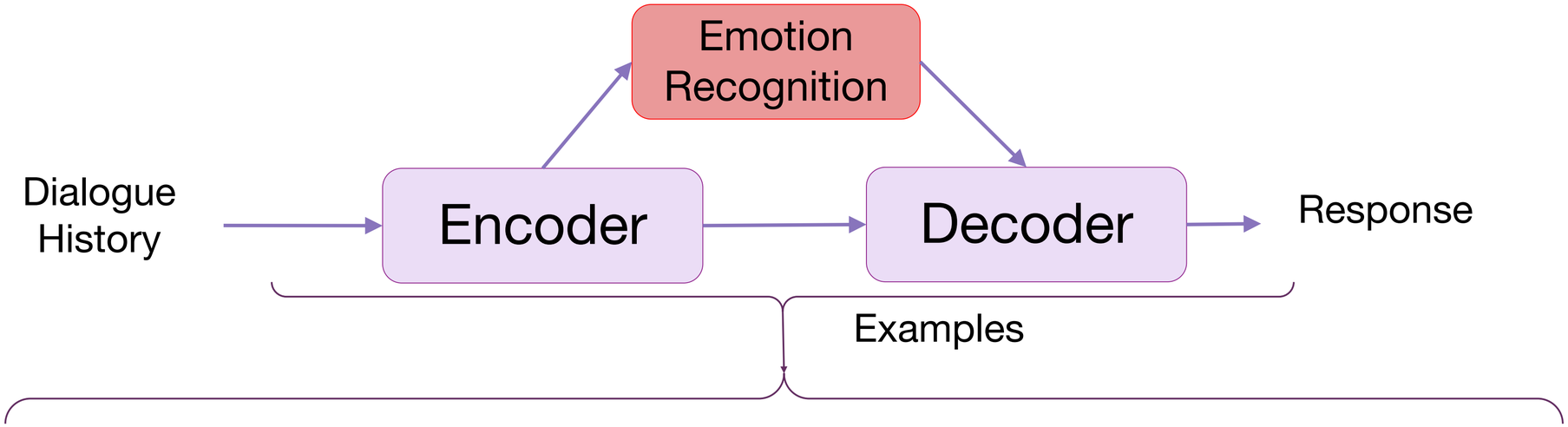
Empathy: understand the feelings of the conversation partner and replying accordingly.



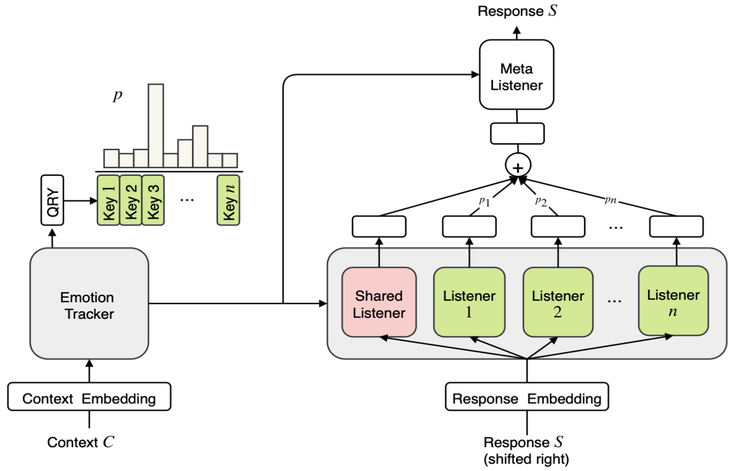
Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.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

Dataset: Empathetic Dialogues

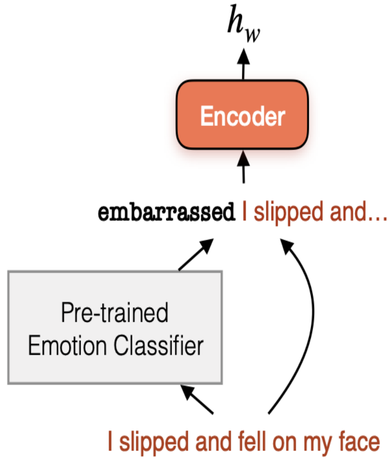
Models with Empathy



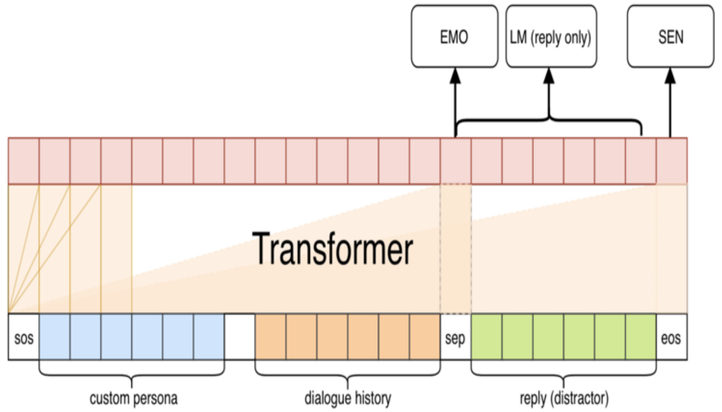
MoEL



EmoPrepend-1

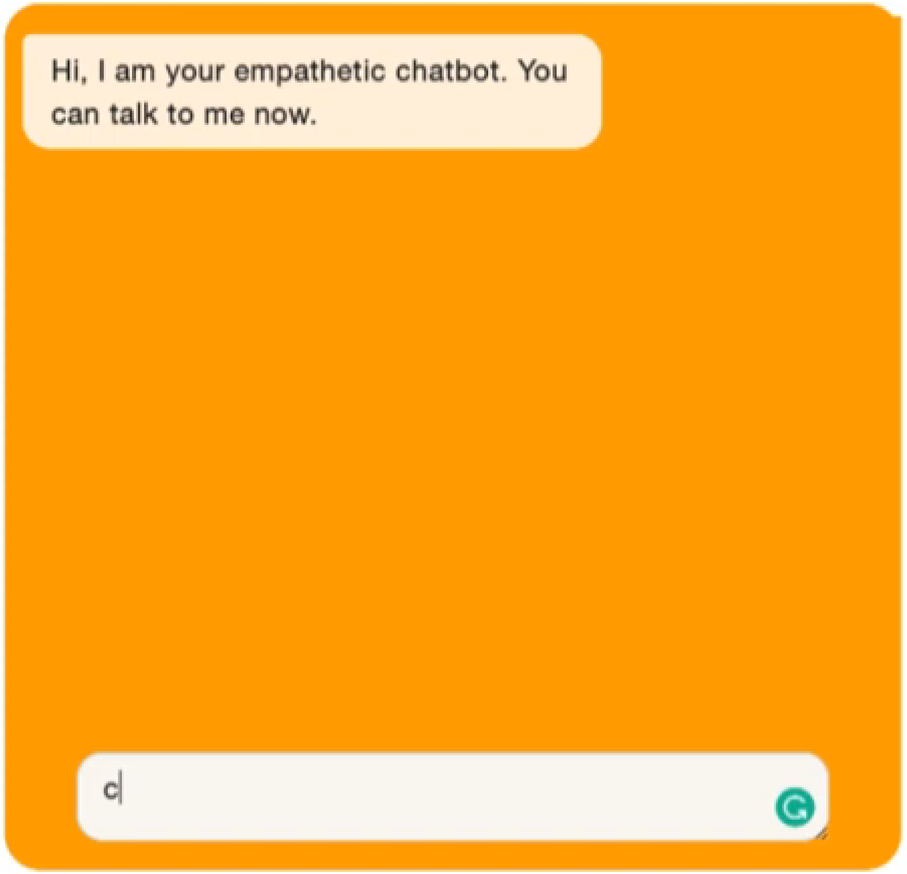


CaireBot



I'm CAiRE, the End-to-End Empathetic Chatbot

CAiRE is implemented by a fully data driven approach as described in this paper.
Special acknowledgement to Huggingface for helpful discussions.



Scenario:
Happy

https://demo.caire.ust.hk/chatbot

Limitations of Vanilla Seq2Seq: Summary

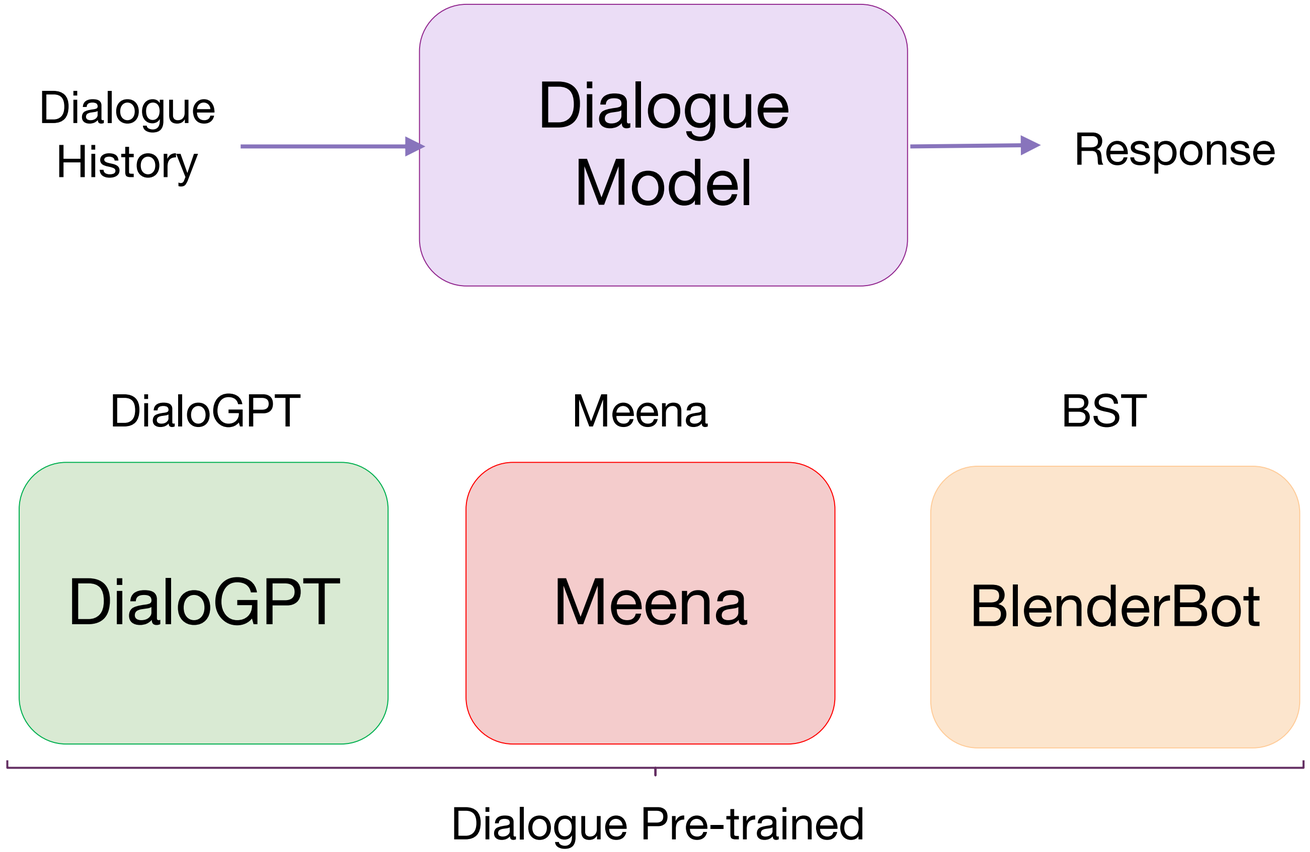
1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. **Lack of controllability**
6. Lack of versatility
7. Lack of global optimization

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

Limitation 5: Lack of Controllability

Existing large pre-trained model has no control over

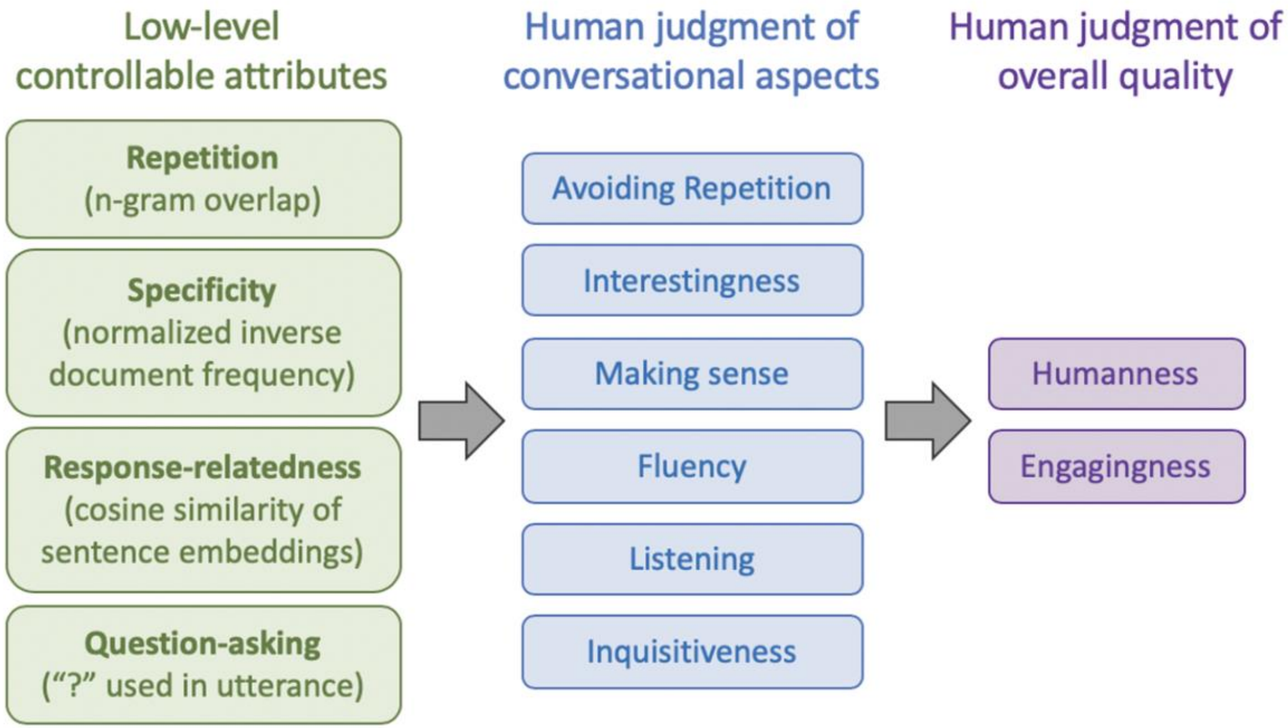
- Response style
- Topics
- Repetition and specificity
- Response-relatedness
- Engagement by proactively asking question



Solution: Controllability

1. Controlling low-level attribute
2. Controlling by fine-tuning
3. Controlling by perturbation
4. Controlling by conditioned generation

Controlling Low-Level Attribute



Conditional Training + Weight Decoding

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

Controlling by Fine-Tuning

Multitask conversation data with style data

⇒ 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

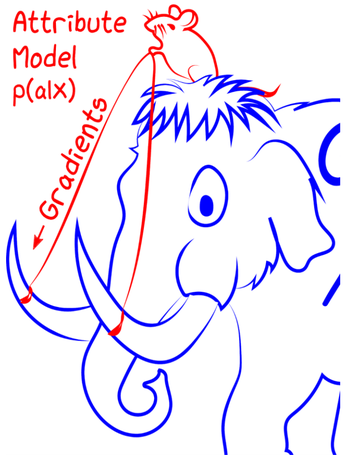
Controlling by Perturbation

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 Plug-and-Play LM (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)



Controlling by Conditioned Generation

Controllable generation architectures in open-domain dialogues:

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

200 style labels in ConvAI2, EmpatheticDialogues, Wizard of Wikipedia, and BlendedSkillTalk) generated by a classifier trained on Image-Chat

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

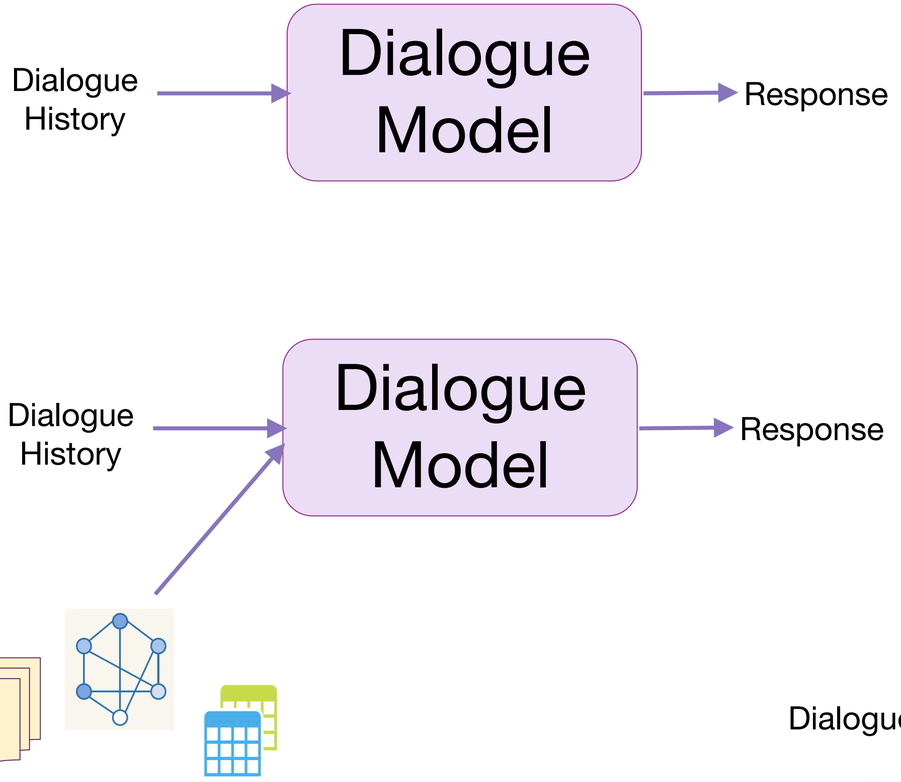
65

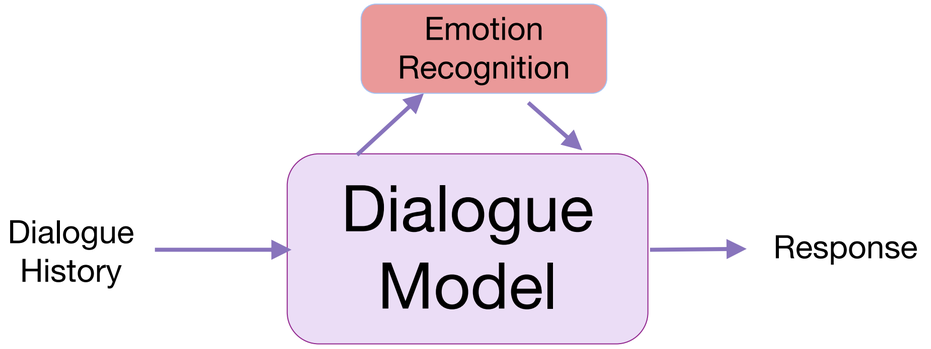
Limitations of Vanilla Seq2Seq: Summary

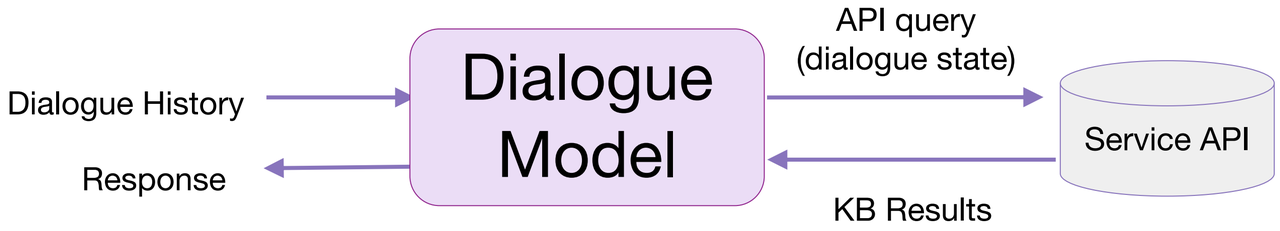
1. Lack of diversity
2. Lack of consistency
3. Lack of knowledge
4. Lack of empathy
5. Lack of controllability
6. **Lack of versatility**
7. Lack of global optimization

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

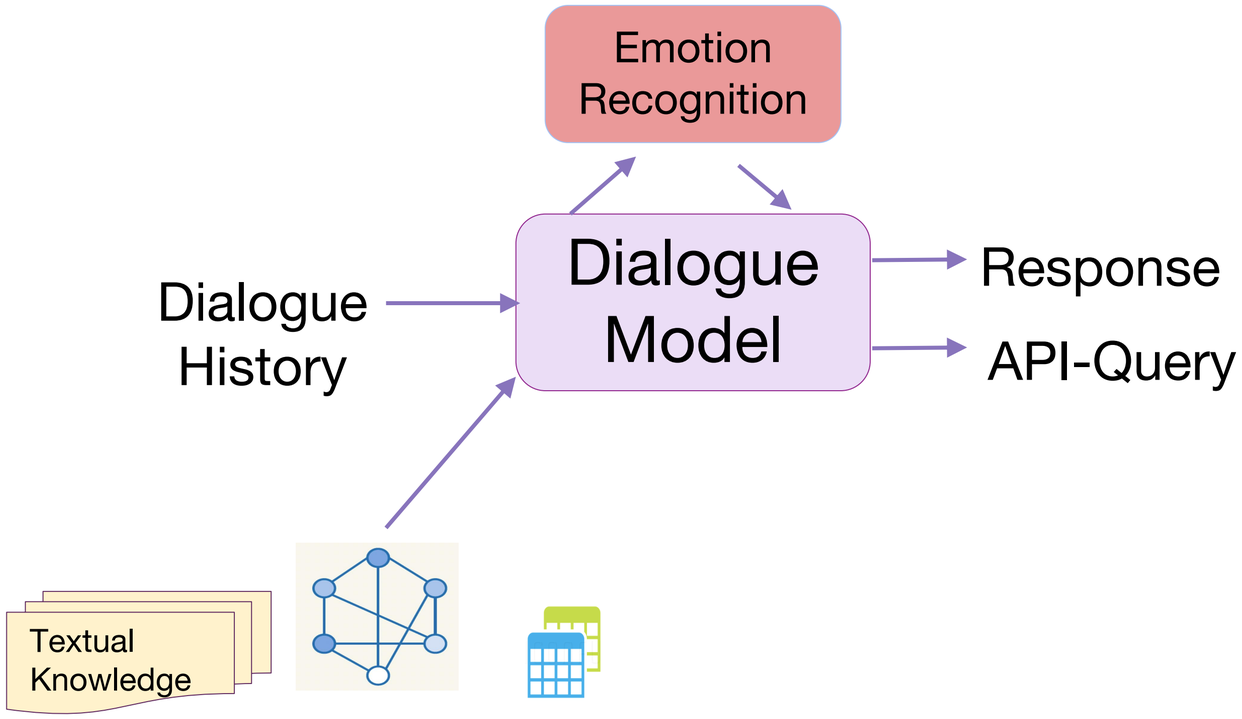
Limitation 6: Lack of Versatility







Solution: ToDs + Chit-Chat

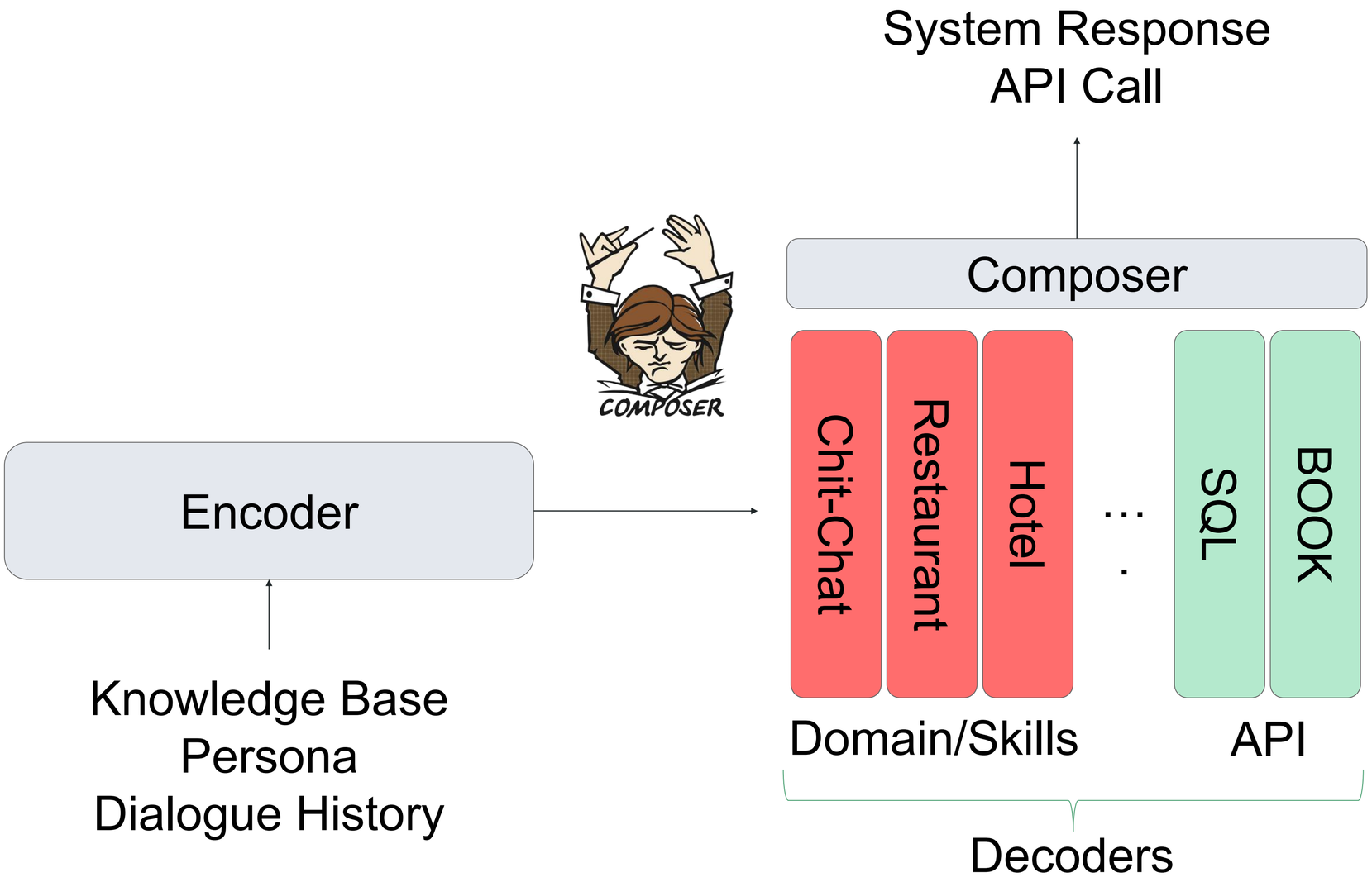


	Dialogue
Usr:	What is the weather like today?
Meta (Weather API):	{Weather : Sunny, High : 32 C, Low : 27 C}
Sys:	Today's weather is sunny with temperature between 27-30 C.
Usr:	Great, I plan to play basketball this afternoon! I am so excited after watched the match between Golden State Warriors and 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.

ToDs + Chit-Chat Datasets

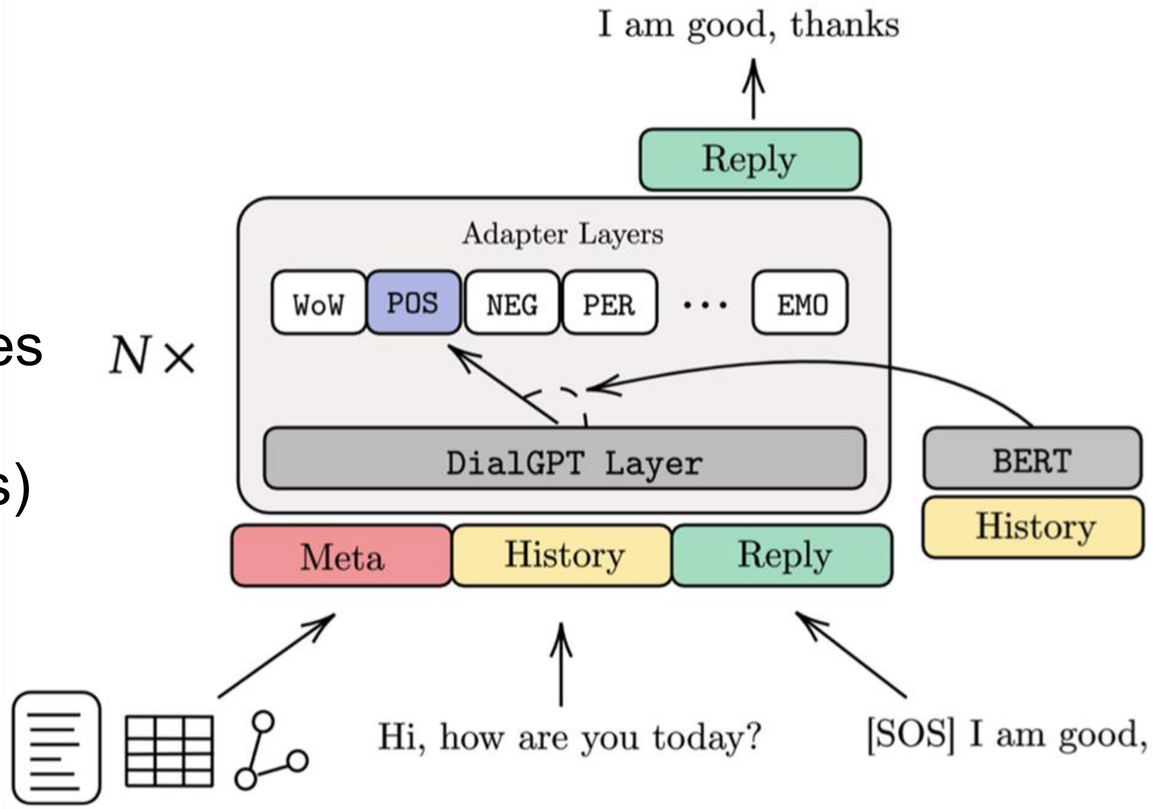
1. Mixing multiple dialogue datasets
2. Multiple dialogue skills
   ⇒ Collecting dataset that mix skills
3. Mixing Chit-Chat and ToDs
   ⇒ Collecting data from mixing the two

Attention over Parameters

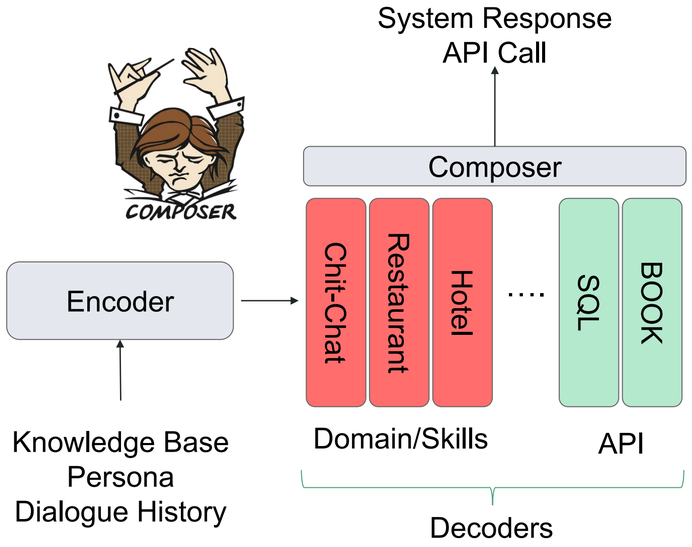


Adapter-Bot: All-In-One Controllable Model

- Use a fixed backbone - DialoGPT
- Encode each dialogue skill with an independently trained adapters
  - able to process multiple knowledge types and styles (8 goal-oriented skills + personalized and empathetic responses)
- A skill manager, BERT, is trained to select each adapter



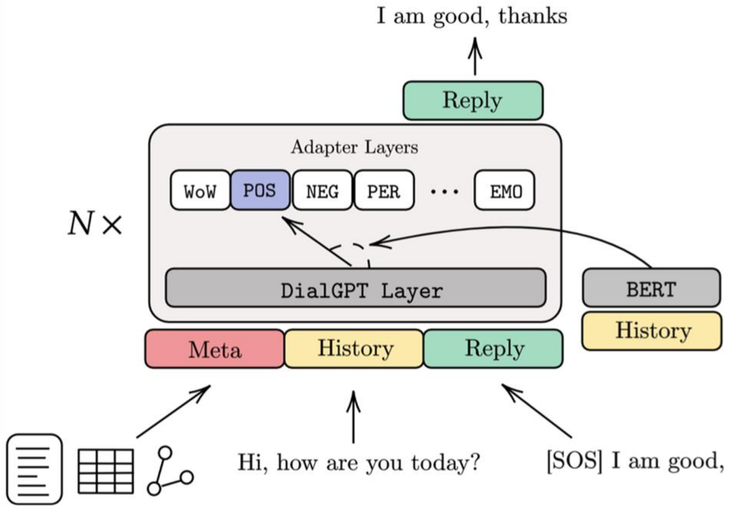
Putting It All Together



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

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

Blender-bot



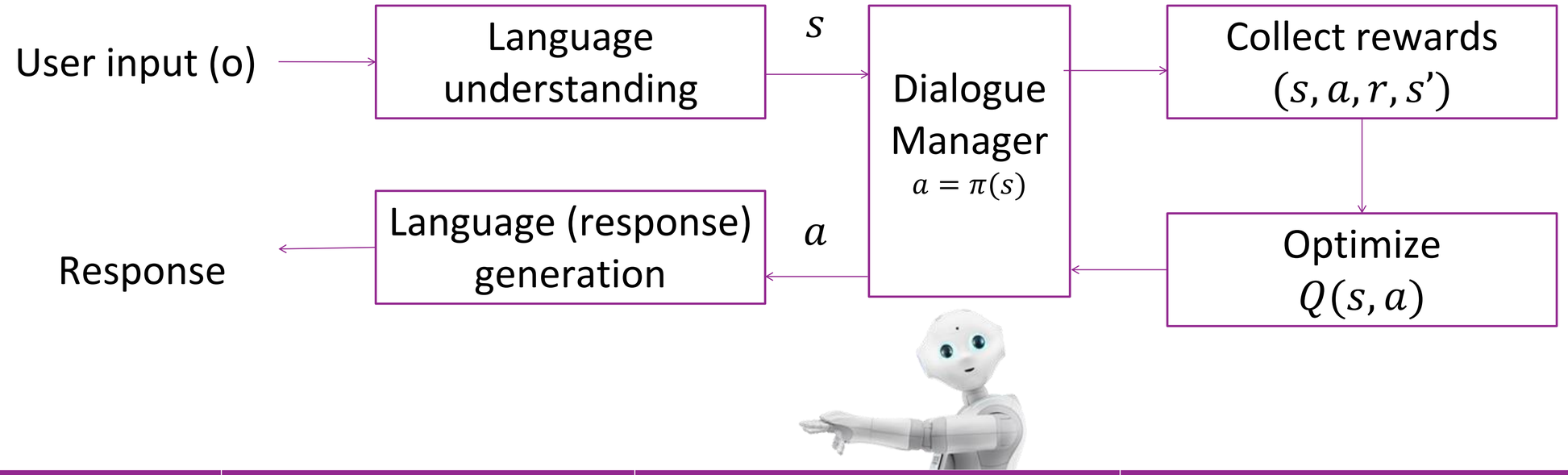
[The Adapter-Bot: All-In-One Controllable Conversational Model (Lin & Madotto et.al. 2020)]

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
7. **Lack of global optimization**

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

Limitation 7: Lack of Global Optimization



Application	State	Action	Reward
Task Completion Bots (Movies, Restaurants, …)	User input + Context	Dialog act + slot-value	Task success rate # of turns
Info Bots (Q&A bot over KB, Web etc.)	Question + Context	Clarification questions, Answers	Relevance of answer # of turns
Social Bot (XiaoIce)	Conversation history	Response	Engagement(?)

Solution: Deep RL for Optimization (Li et al., 2016)

Input message	Supervised Learning Agent	Reinforcement Learning Agent
How old are you ?	I'm 16.	I'm 16, why are you asking ?
What's your full name ?	I have no idea .	What's yours ?
I don't want to go home tonight.	Really ?	Why ?
Do you have any feelings for me ?	I don't know what you are talking about.	I don 't want to hurt your feelings .
How much time do you have here?	Not long enough. Sorry, sir.	Ten seconds .
Shall we get started ?	Of course !	Yes . We 've got a lot of work to do here .
Do you play football ?	No, i don't	Yes. I love football !
We'd have to talk to him.	I mean, he's a good guy	About what ?
How come you never say it?	Because I don't want to hurt you .	I don 't think it 's a good idea to say it .

- RL agent generates more *interactive* responses
- RL agent tends to end a sentence *with a question* and hand the conversation over to the user

Concluding Remarks

- Limitations of vanilla seq2seq models
  1. Lack of diversity
  2. Lack of consistency
  3. Lack of knowledge
  4. Lack of empathy
  5. Lack of controllability
  6. Lack of versatility
  7. Lack of global optimization
- Recent trends for addressing above limitations

Her (2013)

What can machines achieve now or in the future?