CS769 Advanced NLP

Dialog System

Junjie Hu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Slides adapted from Shikib, Dan
https://junjiehu.github.io/cs769-spring22/

What is dialog?

- NLP for conversations
  - Understand utterances in the context of the dialog history (i.e., back-and-forth interaction)
  - Generate responses
- a.k.a. conversational agents, chatbot, dialog agents
- Personal assistants on phones or other devices: Siri, Alexa, Cortana, Google Assistant

Two Types of Dialog Systems

- **Open-domain Chatbots**
  - Mimic informal human chatting
  - Chatting for fun, or even for therapy, clinical uses for mental health
- **Task-oriented Dialog (ToD)**
  - Personal assistants on mobile devices
  - NL interfaces in cars, robots, appliances
  - Customer support, booking flights, hotels, restaurants

Goal for Today

- Brief History of Dialog Systems
- Task-oriented Dialog
- Chatbot
- Evaluation

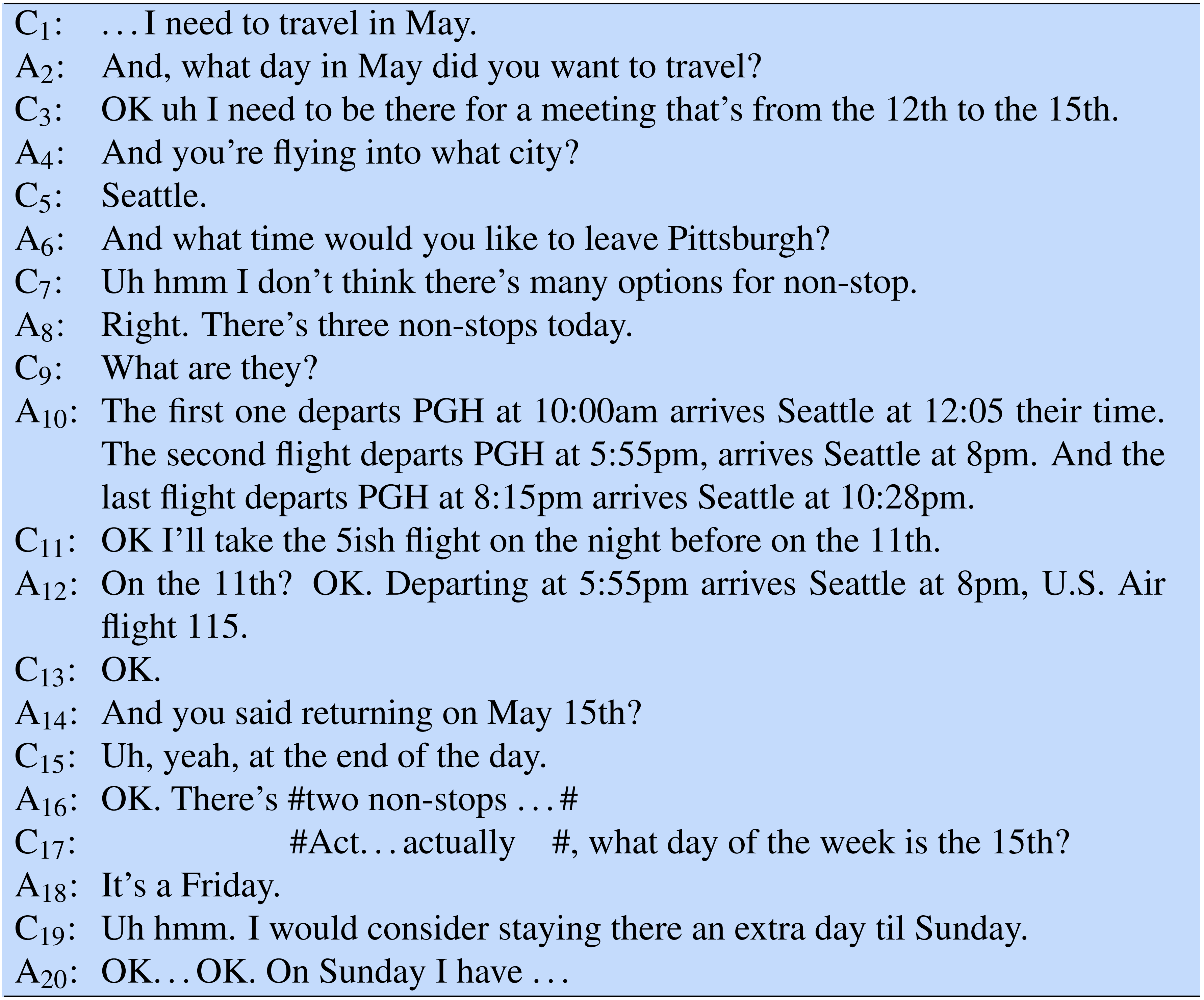
Brief History of Dialog Research

Dialog Architectures

- **Rule-based**
  - Pattern-action rules (ELIZA)
  - + A mental model (PARRY): the first system to pass the Turing test!
- **Corpus-based**
  - Information retrieval (XiaoIce)
  - Neural encoder-decoder (BlenderBot)

Real Example: A telephone conversation

- A human travel agent (A) and a human client (C)



C_1: . . . I need to travel in May.
A_2: And, what day in May did you want to travel?
C_3: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_4: And you're flying into what city?
C_5: Seattle.
A_6: And what time would you like to leave Pittsburgh?
C_7: Uh hmm I don't think there's many options for non-stop.
A_8: Right. There's three non-stops today.
C_9: What are they?
A_{10}: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C_{11}: OK I'll take the 5ish flight on the night before on the 11th.
A_{12}: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_{13}: OK.
A_{14}: And you said returning on May 15th?
C_{15}: Uh, yeah, at the end of the day.
A_{16}: OK. There's #two non-stops . . . #
C_{17}: #Act. . . actually #, what day of the week is the 15th?
A_{18}: It's a Friday.
C_{19}: Uh hmm. I would consider staying there an extra day til Sunday.
A_{20}: OK. . . OK. On Sunday I have . . .

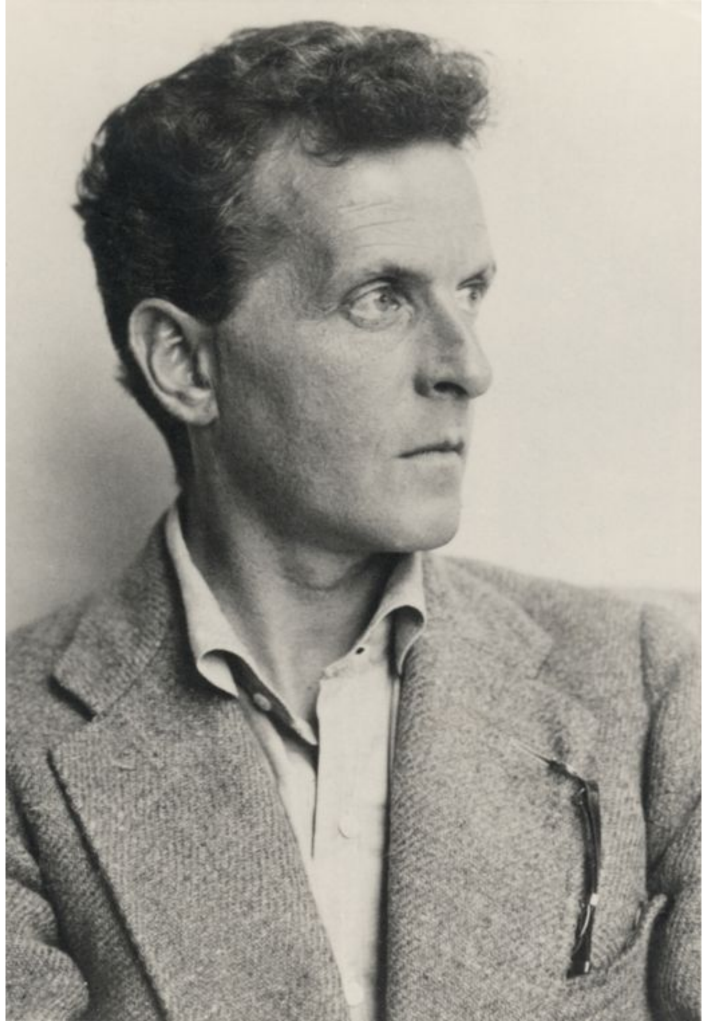
Properties of Human Conversation

- **Turns**
  - We call each contribution a “turn”
  - As if conversation was the kind of game where everyone takes turns.
- **Turn-taking issues**
  - When to take the floor?
  - When to yield the floor?
- **Interruptions**

Implications for Dialog Agents

- Barge-in
  - Allow the user to interrupt
- End-pointing
  - The task for a speech system of deciding whether the user has stoped talking.
  - Very hard, since people often pause in the middle of turns.

Language as Action



Each turn in a dialogue is a kind of action

Wittgenstein (1953) and Austin (1962)

Speech Acts (aka Dialogue Acts)

Bach and Harnish (1979)

Constatives: committing the speaker to something's being the case (*answering*, *claiming*, *confirming*, *denying*, *disagreeing*, *stating*)

Directives: attempts by the speaker to get the addressee to do something (*advising*, *asking*, *forbidding*, *inviting*, *ordering*, *requesting*)

Commissives: committing the speaker to some future course of action (*promising, planning, vowing, betting, opposing*)

Acknowledgments: express the speaker's attitude regarding the hearer with respect to some social action (*apologizing, greeting, thanking, accepting an acknowledgment*)

ELIZA (Weizenbaum 1966)

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED

...

You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum, Joseph. 1966, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36-45

ELIZA's trick: be a Rogerian psychologist

- Draw the patient out by reflecting patient’s statements back at them
- Rare type of conversation in which one can “assume the pose of knowing almost nothing of the real world”

Rogerian psychologist

(Person-Centered Therapy)

Patient: "I went for a long boat ride”
Psychologist: "Tell me about boats”

- You don’t assume she didn’t know what a boat is
- You assume she had some conversational goal
- Chatbots trying to pass the Turing test usually choose some such domain

Eliza Rules

- Rules are organized by **keywords**
- Each keyword has a pattern and a list of possible transform

Keyword: `YOU`

Pattern: `(0 YOU 0 ME)`

Transforms: `(WHAT MAKES YOU THINK I 3 YOU)`

`(WHY DO YOU THINK I 3 YOU)`

keyword pattern list of transforms

$$\begin{array}{llllll}
(K & ((D_1) & (R_{1,1}) & (R_{1,2}) & \cdots & (R_{1,m_1})) \\
 & ((D_2) & (R_{2,1}) & (R_{2,2}) & \cdots & (R_{2,m_2})) \\
 & \quad\vdots & & & & \quad\vdots \\
 & ((D_n) & (R_{n,1}) & (R_{n,2}) & \cdots & (R_{n,m_n})))
\end{array}$$

Keywords are ranked from specific to general

```
User: I know everybody laughed at me
```

"I" is very general:

```
I: (I *) -> (YOU SAY YOU 2)
ELIZA: YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU
```

"Everybody" is more specific and interesting

```
Everybody: (Everybody *) -> (WHO IN PARTICULAR ARE YOU THINKING OF)
ELIZA: WHO IN PARTICULAR ARE YOU THINKING OF?
```

Implementation: keywords stored with their rank

`Everybody 5 (list of` *`transformation rules`*`)`

`I         0 (list of` *`transformation rules`*`)`

Memory

```
(MEMORY MY
 (0 MY 0 = LETS DISCUSS FURTHER WHY YOUR 3)
 (0 MY 0 = EARLIER YOU SAID YOUR 3)
 (0 MY 0 = DOES THAT HAVE ANYTHING TO DO WITH THE FACT
THAT YOUR 3))
```

- Whenever “MY” is highest keyword
  - Randomly select a transform on the MEMORY list
  - Apply to sentence
  - Store on a (first-in-first-out) queue
- Later, if no keyword matches a sentence
  - Return the top of the MEMORY queue instead

```
function ELIZA GENERATOR(user sentence) returns response

Find the word w in sentence that has the highest keyword rank
  if w exists
       Choose the highest ranked rule r for w that matches sentence
       response ← Apply the transform in r to sentence
       if w = 'my'
           future ← Apply a transformation from the 'memory' rule list to sentence
           Push future onto memory stack
  else (no keyword applies)
    either
       response ← Apply the transform for the NONE keyword to sentence
    or
       response ← Pop the top response from the memory stack
 return(response)
```

PARRY: A computational model of schizophrenia

- Another chatbot with a clinical psychology focus
  - Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. Artificial Intelligence 2(1), 1–25.

- Used to study schizophrenia
- Same pattern-response structure as Eliza
- But a much richer:
  - control structure
  - language understanding capabilities
  - model of mental state: variables modeling levels of Anger, Fear, Mistrust

Affect variables

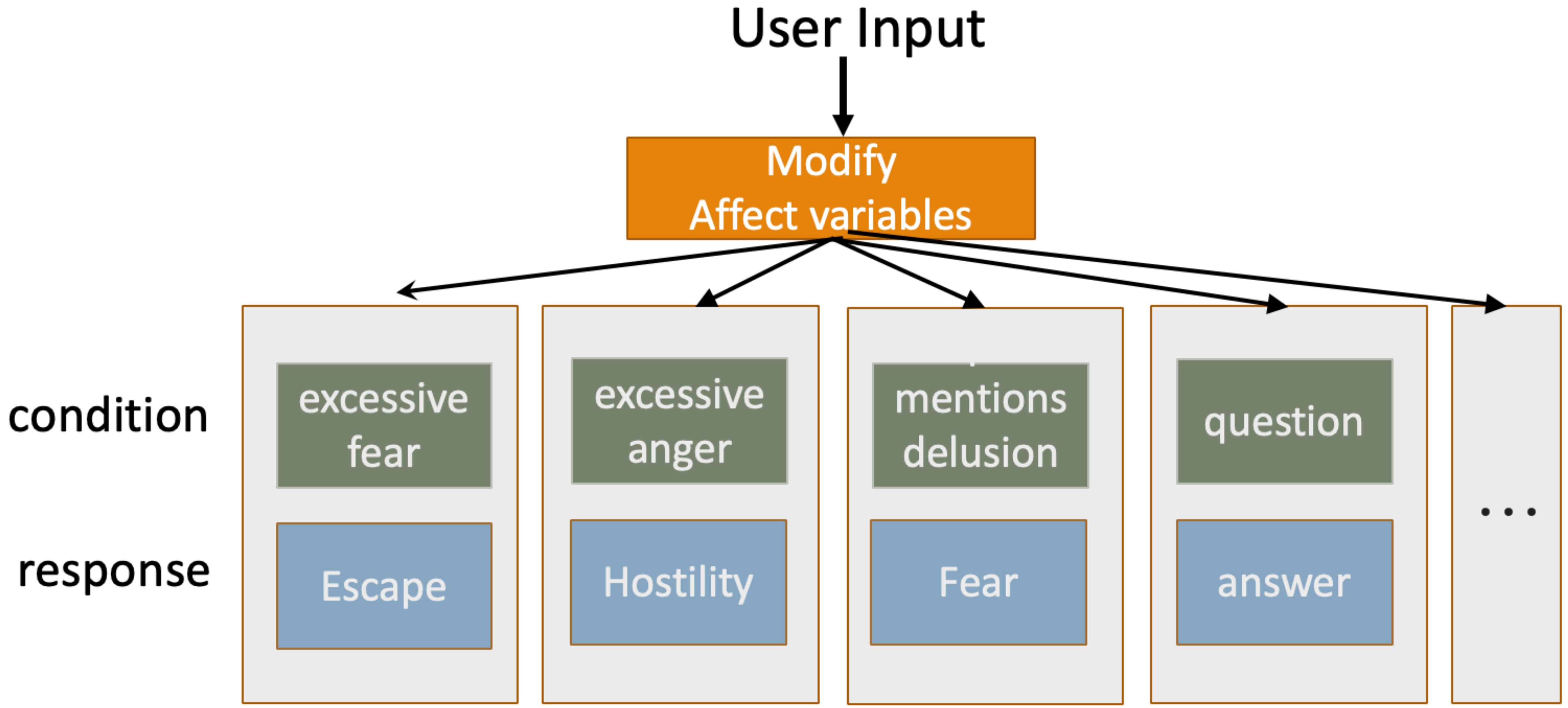
Fear (0-20) **Anger** (0-20) **Mistrust** (0-15)

Start with all variables low

After each user turn

- Each user statement can change Fear and Anger
    - E.g., Insults increases Anger, Flattery decreases Anger
    - Mentions of his delusions increase Fear
- Else if nothing malevolent in input
    - Anger, Fear, Mistrust all drop

Parry's responses depend on mental state



PARRY passes the Turing test in 1972

The first system to pass a version of the Turing test

Psychiatrists couldn’t distinguish interviews with PARRY from (text transcripts of) interviews with people diagnosed with paranoid schizophrenia

- Colby, K. M., Hilf, F. D., Weber, S., and Kraemer, H. C. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence 3*, 199–221.

BlenderBot (Roller et al. 2020)

Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sing it to the tune of “Sing Along with Mitch”?

Yes, I would love that.

OK, I’ll try to get it to you by the end of the day. What else do you like to do?

No, sing it to me now!

OK, give me a few days. I’m a teacher, so I don’t have much free time.

Can you at least tell me what the lyrics are?

Sure, it goes like this: “Baking is my favorite thing to do in the world.”

How does the chorus go?

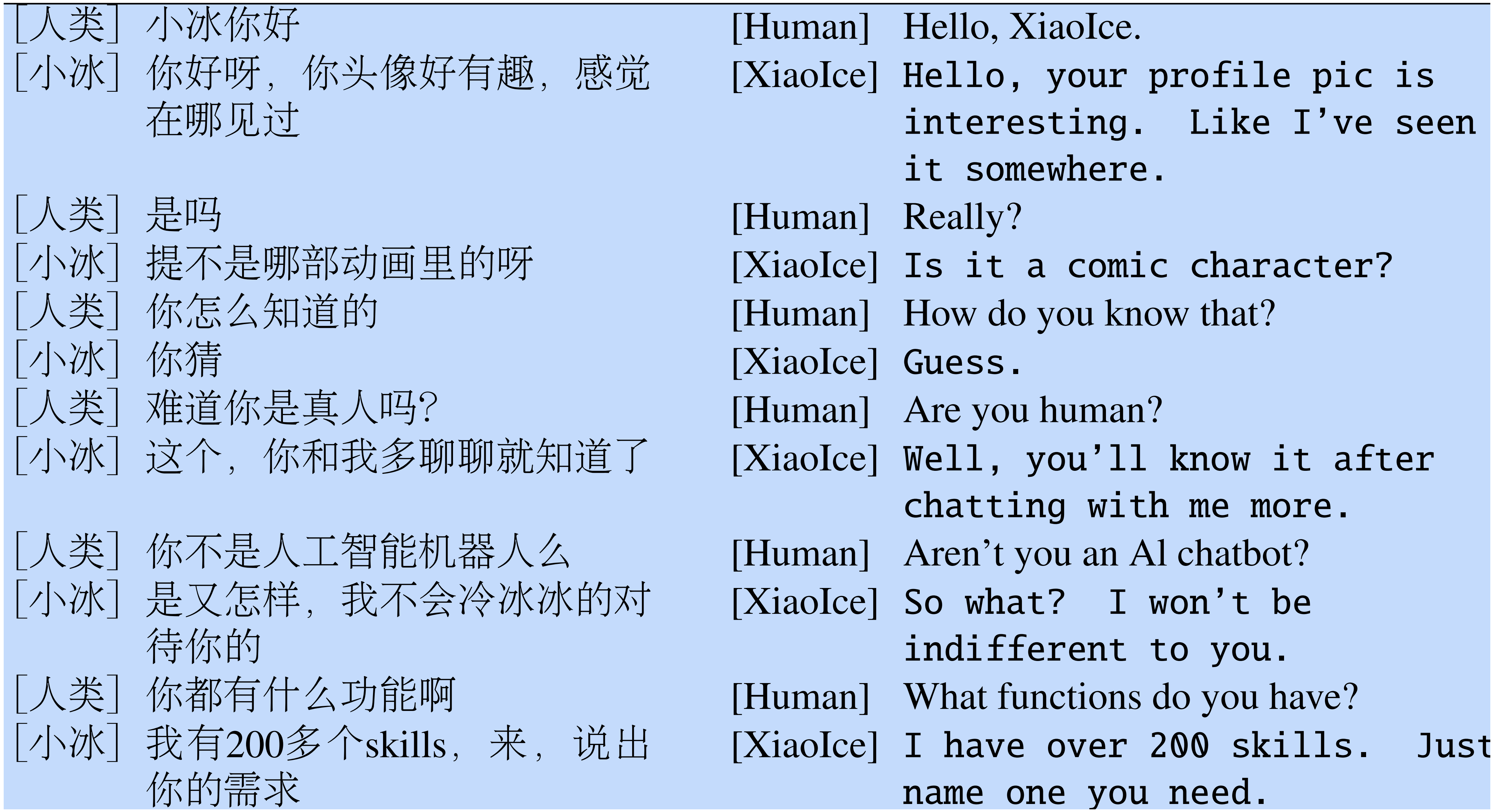
It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!

Wow, sounds like a bop.

I know, right? It’s so catchy. I can’t get it out of my head.

XiaoIce (Zhou et al. 2020)

- Microsoft's chatbot responds mainly by extracting turns that humans have said in past conversations.

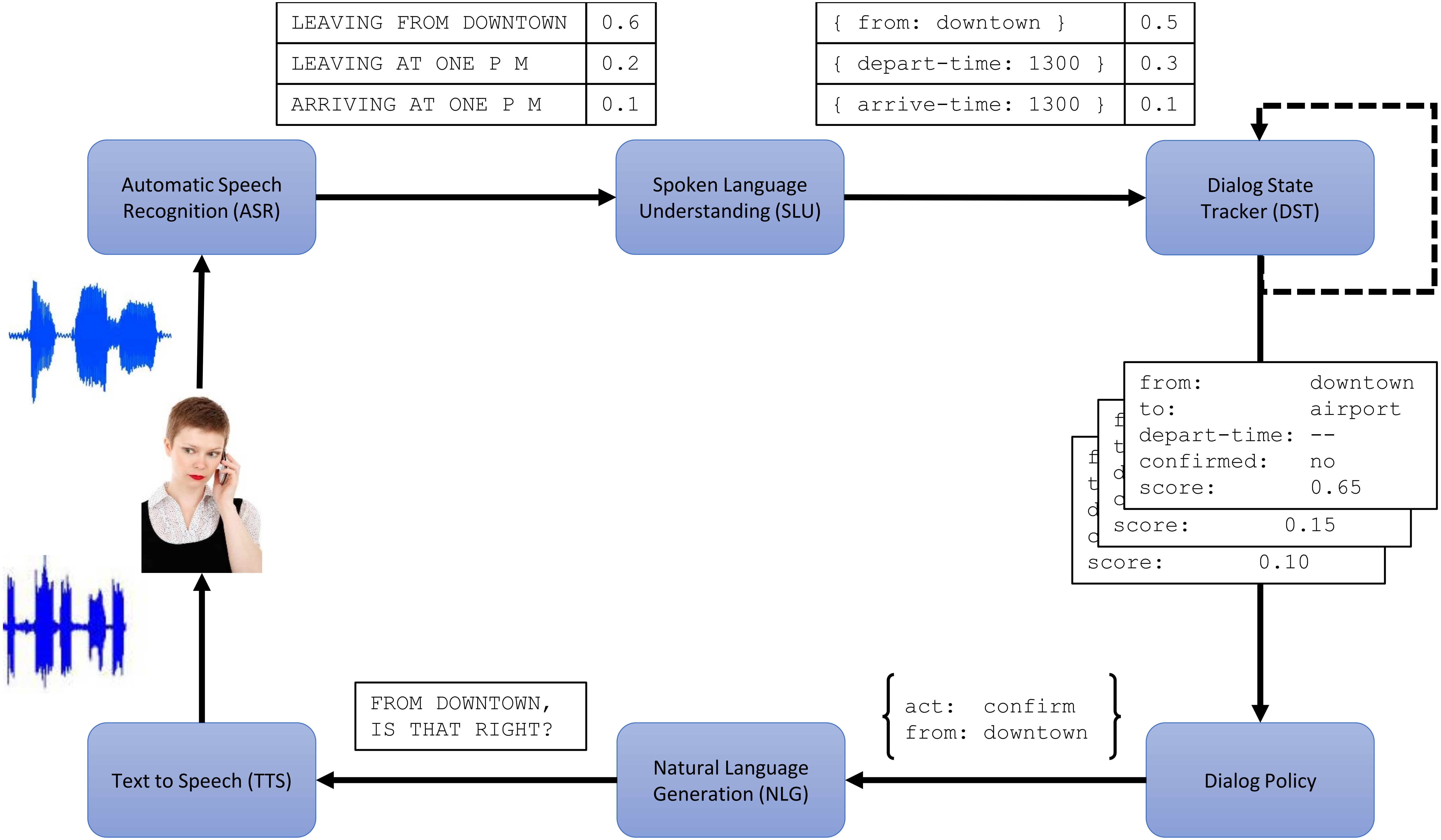


[人类]	小冰你好	[Human]	Hello, XiaoIce.
[小冰]	你好呀，你头像好有趣，感觉在哪见过	[XiaoIce]	Hello, your profile pic is interesting. Like I've seen it somewhere.
[人类]	是吗	[Human]	Really?
[小冰]	提不是哪部动画里的呀	[XiaoIce]	Is it a comic character?
[人类]	你怎么知道的	[Human]	How do you know that?
[小冰]	你猜	[XiaoIce]	Guess.
[人类]	难道你是真人吗?	[Human]	Are you human?
[小冰]	这个，你和我多聊聊就知道了	[XiaoIce]	Well, you'll know it after chatting with me more.
[人类]	你不是人工智能机器人么	[Human]	Aren't you an AI chatbot?
[小冰]	是又怎样，我不会冷冰冰的对待你的	[XiaoIce]	So what? I won't be indifferent to you.
[人类]	你都有什么功能啊	[Human]	What functions do you have?
[小冰]	我有200多个skills，来，说出你的需求	[XiaoIce]	I have over 200 skills. Just name one you need.

Task-oriented Dialog

Pipeline Dialog System

(Williams et al. 2016)



LEAVING FROM DOWNTOWN	0.6
LEAVING AT ONE P M	0.2
ARRIVING AT ONE P M	0.1

{ from: downtown }	0.5
{ depart-time: 1300 }	0.3
{ arrive-time: 1300 }	0.1

```
from:         downtown
to:           airport
depart-time:  --
confirmed:    no
score:        0.65
```

```
score:        0.15
```

```
score:        0.10
```

```
{ act:   confirm
  from:  downtown }
```

```
FROM DOWNTOWN,
IS THAT RIGHT?
```

Automatic Speech Recognition (ASR)
Spoken Language Understanding (SLU)
Dialog State Tracker (DST)
Dialog Policy
Natural Language Generation (NLG)
Text to Speech (TTS)

Components in a ToD architecture

- **NLU:** extracts slot fillers from the user’s utterance using machine learning
- **Dialogue state tracker (DST):** maintains the current state of the dialogue (user’s most recent dialogue act, set of slot-filler constraints from user), sometimes merged in the **NLU**.
- **Dialogue policy (DP or DM):** decides what the system should do or say next
- **NLG**: produce more natural, less templated utterances

NLU

Natural language understanding in dialog involves several key tasks:

- Intent prediction: what is the user’s intent/goal
- Slot filling: what are the slot values (e.g., what is the time)
- **Dialog State tracking (DST)**: track user information/ goals throughout the dialog

NLU

Natural language understanding in dialog involves several key tasks:

- DialoGLUE [Mehri et al. 2020]
- Intent prediction: ATIS, SNIPS, Banking77, CLINC150, HWU64
- Slot filling: ATIS, SNIPS, DSTC8-SGD, Restaurant8k
- State tracking: MultiWOZ (2.X)

Slot filling

- Machine learning classifiers to map words to semantic frame-fillers
- Given a set of labeled sentences

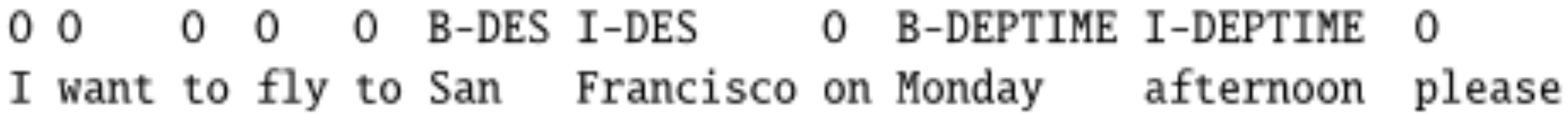
  Input: `"I want to fly to San Francisco on Monday please"`

  Output: `Destination: SF`

  `Depart-time: Monday`
- Build a classifier to map from one to the other
- Requirements: Lots of labeled data

Slot filling as sequence labeling: BIO tagging

- The **BIO tagging** paradigm
- Idea: Train a classifier to label each input word with a tag that tells us what slot (if any) it fills

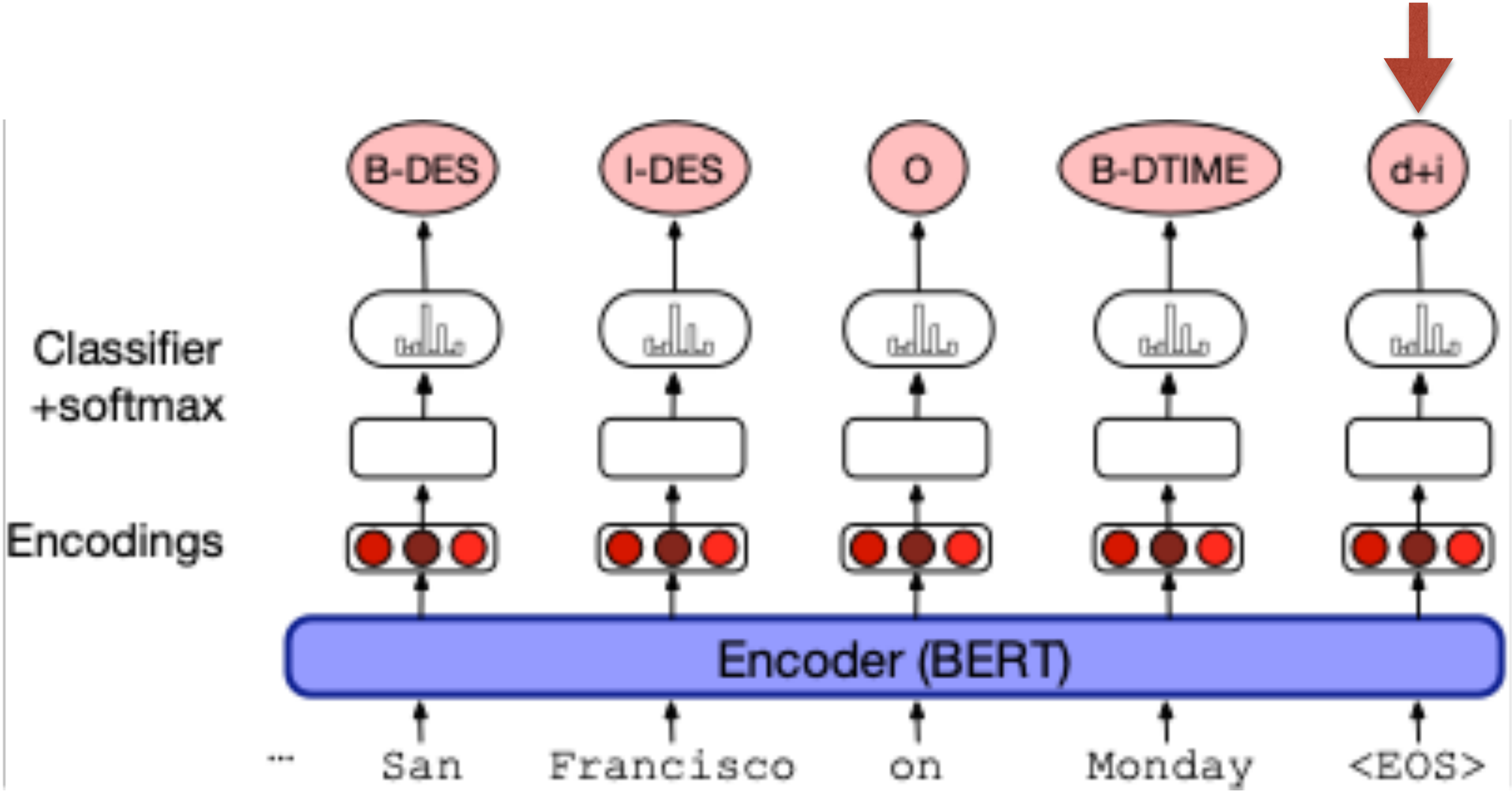


```
O O    O  O   O  B-DES I-DES     O  B-DEPTIME I-DEPTIME  O
I want to fly to San   Francisco on Monday    afternoon  please
```

- We create a B and I tag for each slot-type
- And convert the training data to this format

Slot filling using contextual embeddings

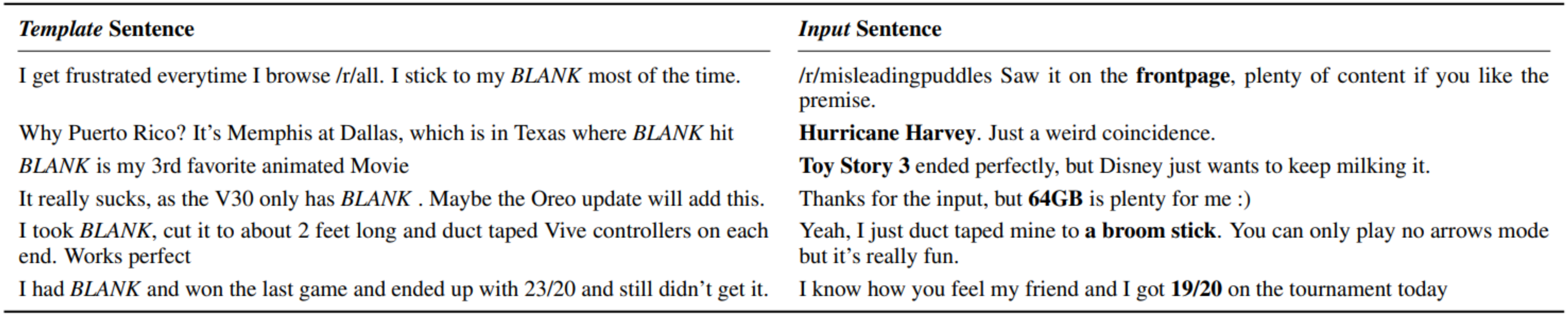
Can do domain and intent too: e.g., generate the label "AIRLINE_TRAVEL + SEARCH_FLIGHT"



ConVEx (Henderson and Vulic. 2020)

- Pre-training paradigm specifically for slot filling on Reddit data→ strong few-shot/zero-shot performance
- Select informative low-frequency key phrases for masking

$$score(kp) = {}^{1}/_{n^{\alpha}} \sum_{i=1}^{n} \log \frac{|D|}{\text{count}(w_i)}.$$

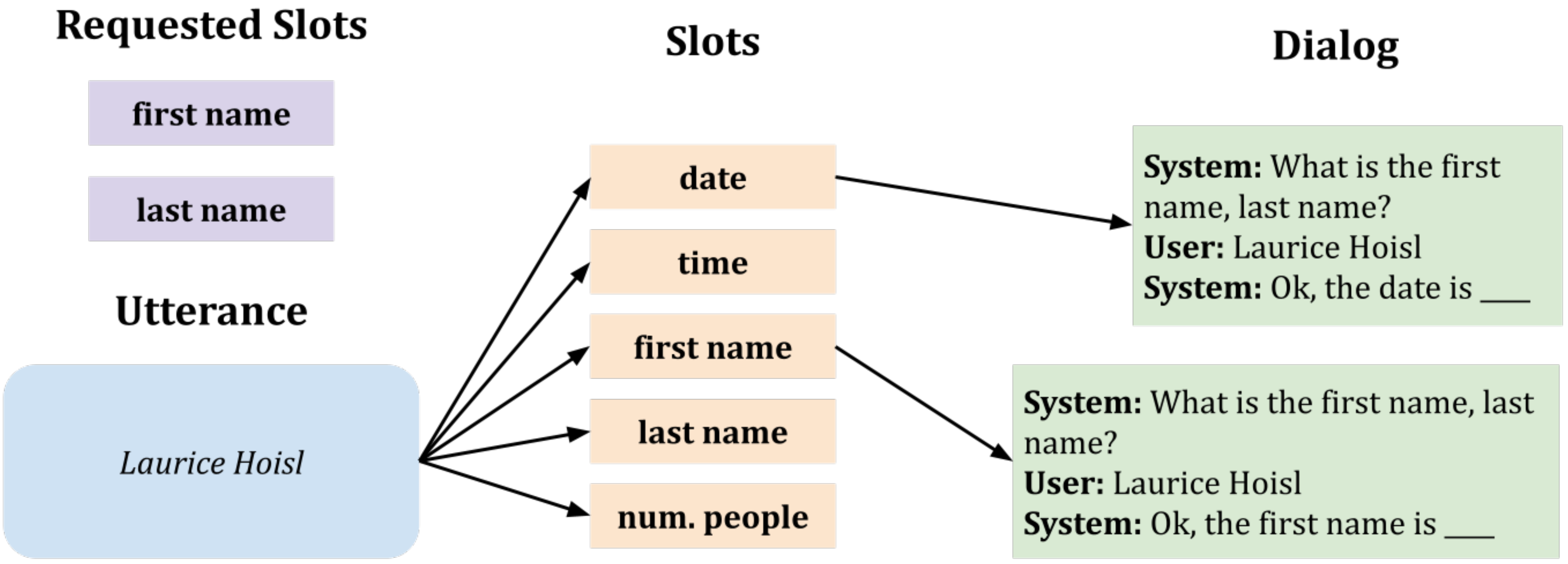


***Template* Sentence**	***Input* Sentence**
I get frustrated everytime I browse /r/all. I stick to my *BLANK* most of the time.	/r/misleadingpuddles Saw it on the **frontpage**, plenty of content if you like the premise.
Why Puerto Rico? It's Memphis at Dallas, which is in Texas where *BLANK* hit	**Hurricane Harvey**. Just a weird coincidence.
BLANK is my 3rd favorite animated Movie	**Toy Story 3** ended perfectly, but Disney just wants to keep milking it.
It really sucks, as the V30 only has *BLANK* . Maybe the Oreo update will add this.	Thanks for the input, but **64GB** is plenty for me :)
I took *BLANK*, cut it to about 2 feet long and duct taped Vive controllers on each end. Works perfect	Yeah, I just duct taped mine to **a broom stick**. You can only play no arrows mode but it's really fun.
I had *BLANK* and won the last game and ended up with 23/20 and still didn't get it.	I know how you feel my friend and I got **19/20** on the tournament today

Table 1: Sample data from Reddit converted to sentence pairs for the ConVEx pretraining via the pairwise cloze task. Target spans in the input sentence are denoted with bold, and are *"BLANKed"* in the template sentence.

GenSF (Mehri and Eskenazi. 2021)

- Convert slot filling task as a response generative task.
- Asking the slot information using a template.
- Make the fine-tuning task close to the pre-training task (LM)



System: What is the first name, last name?
User: Laurice Hoisl
System: Ok, the date is ___

System: What is the first name, last name?
User: Laurice Hoisl
System: Ok, the first name is ___

Results on Restaurant8k

Fraction	**Span-BERT**	**ConVEx**	**GenSF**
1 (8198)	93.1	96.0	**96.1**
1/2 (4099)	91.4	94.1	**94.3**
1/4 (2049)	88.0	92.6	**93.2**
1/16 (512)	76.6	86.4	**89.7**
1/128 (64)	30.6	71.7	**72.2**

Dialog State Tracking

- Maintain the current state of the dialog—update the state from the current user turn incrementally

User: I'm looking for a cheaper restaurant

`inform(price=cheap)`

System: Sure. What kind - and where?

User: Thai food, somewhere downtown

`inform(price=cheap, food=Thai, area=centre)`

System: The House serves cheap Thai food

User: Where is it?

`inform(price=cheap, food=Thai, area=centre); request(address)`

System: The House is at 106 Regent Street

Example from Mrkšić, N., O Séaghdha, D., Wen, T.-H., Thomson, B., and Young, S. (2017). Neural belief tracker: Data-driven dialogue state tracking. ACL.

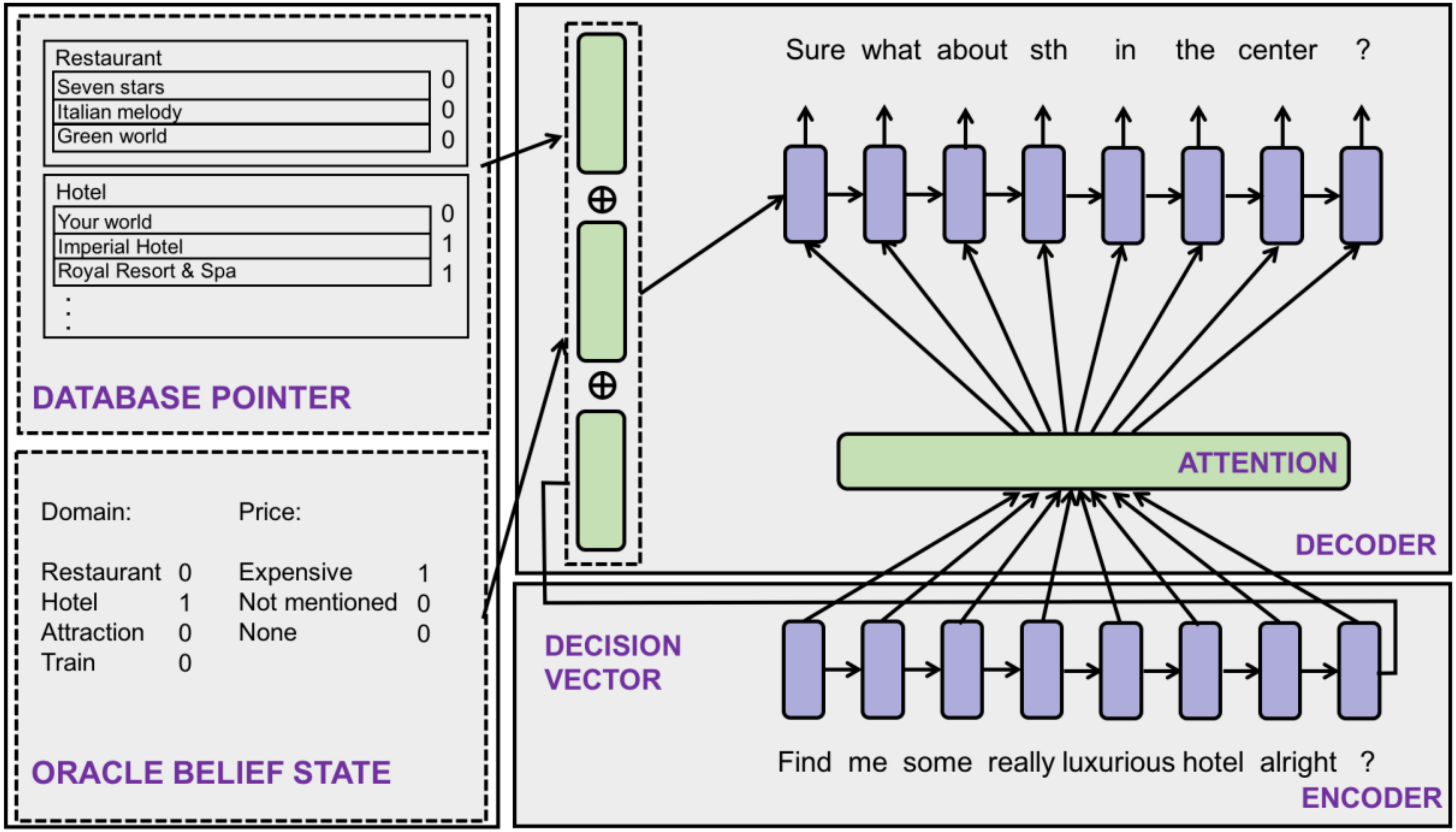
Task-Oriented Response Generation

ToD response generation module should reflect below factors:

- Must understand the **dialog context**
- Must track **belief state** over dialog context
- Often need to interpret structured **database output**
- Must follow task-specific **dialog policy**
- Must generate **fluent, coherent** natural language responses

Seq2Seq with Attention

(Budzianowski et al. 2018)



Restaurant

Seven stars	0
Italian melody	0
Green world	0

Hotel

Your world	0
Imperial Hotel	1
Royal Resort & Spa	1

⋮

DATABASE POINTER

Domain:		Price:	
Restaurant	0	Expensive	1
Hotel	1	Not mentioned	0
Attraction	0	None	0
Train	0		

ORACLE BELIEF STATE

⊕

⊕

Sure what about sth in the center ?

ATTENTION

DECODER

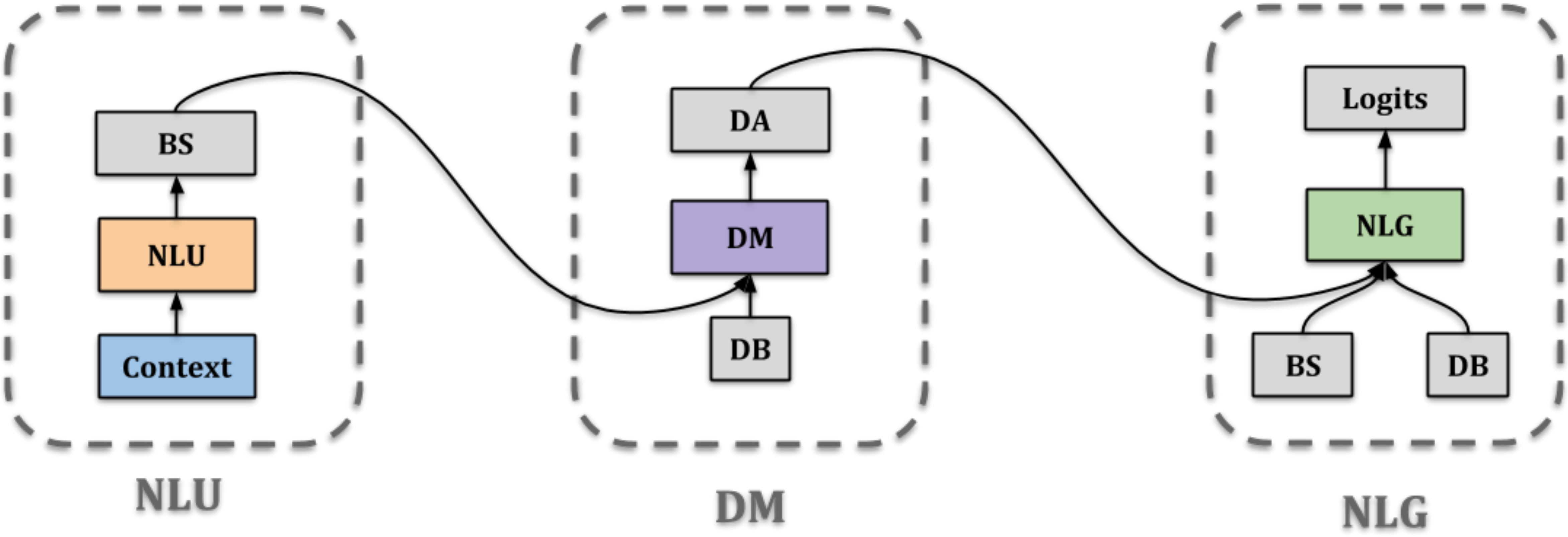
DECISION VECTOR

Find me some really luxurious hotel alright ?

ENCODER

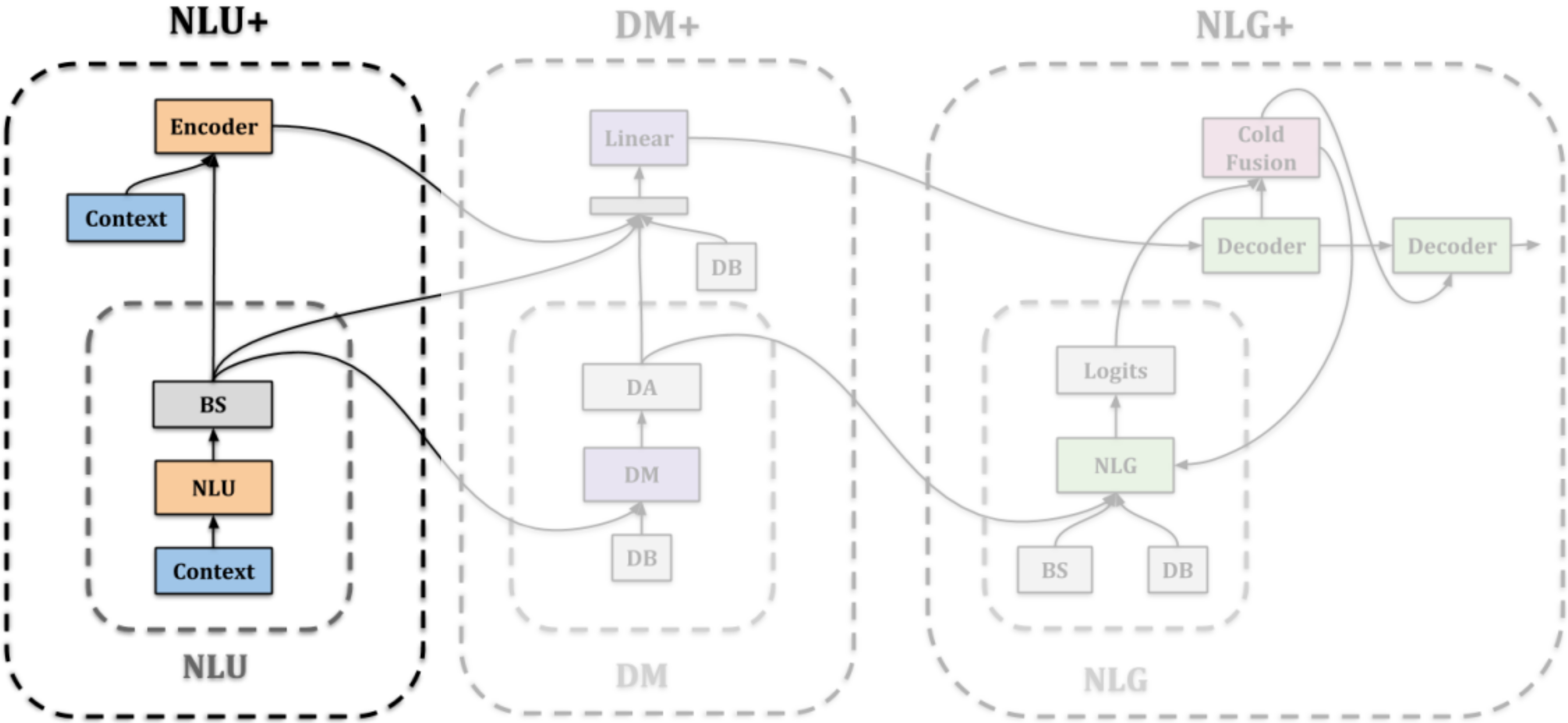
Dialog Modules

Start with **pre-trained** neural dialog modules



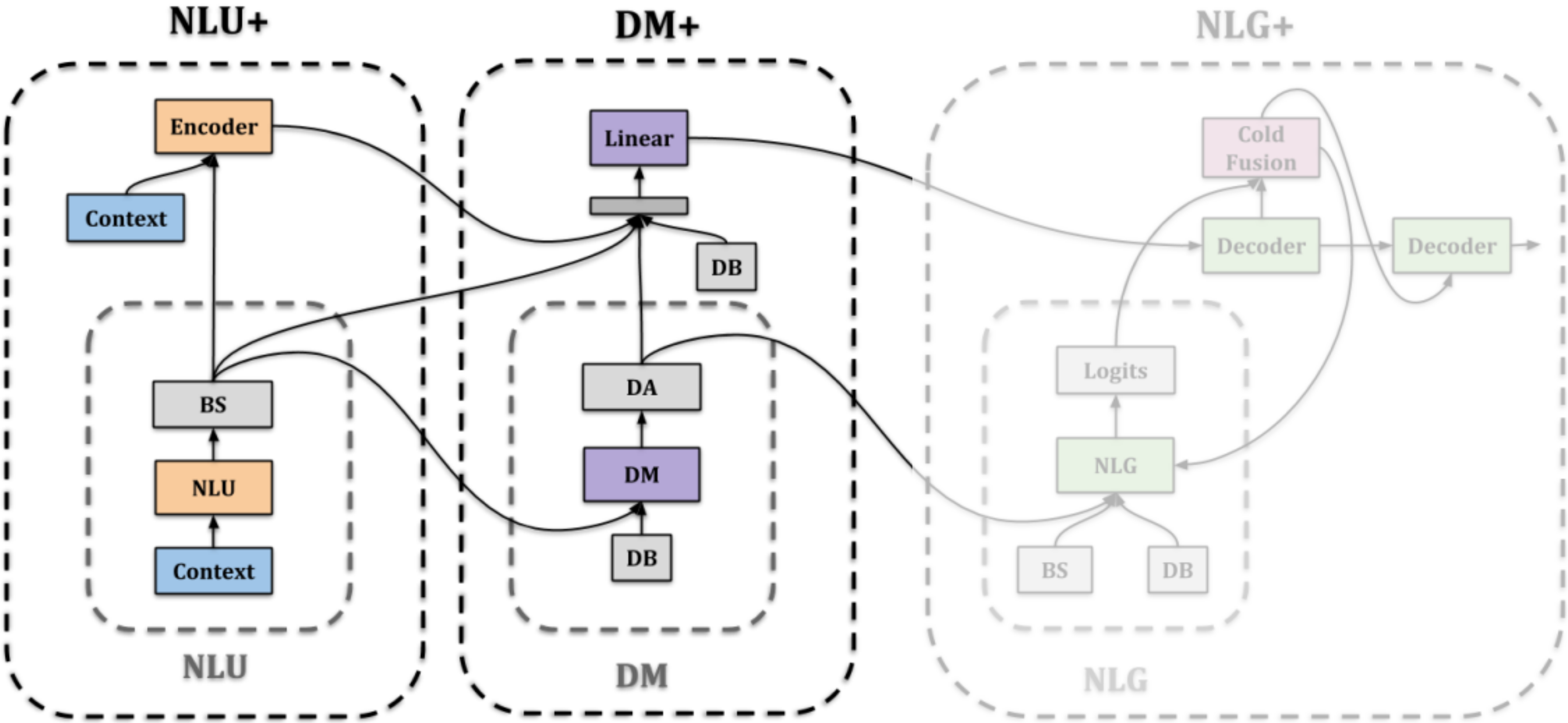
Structured Fusion Networks

(Mehri et al. 2019)



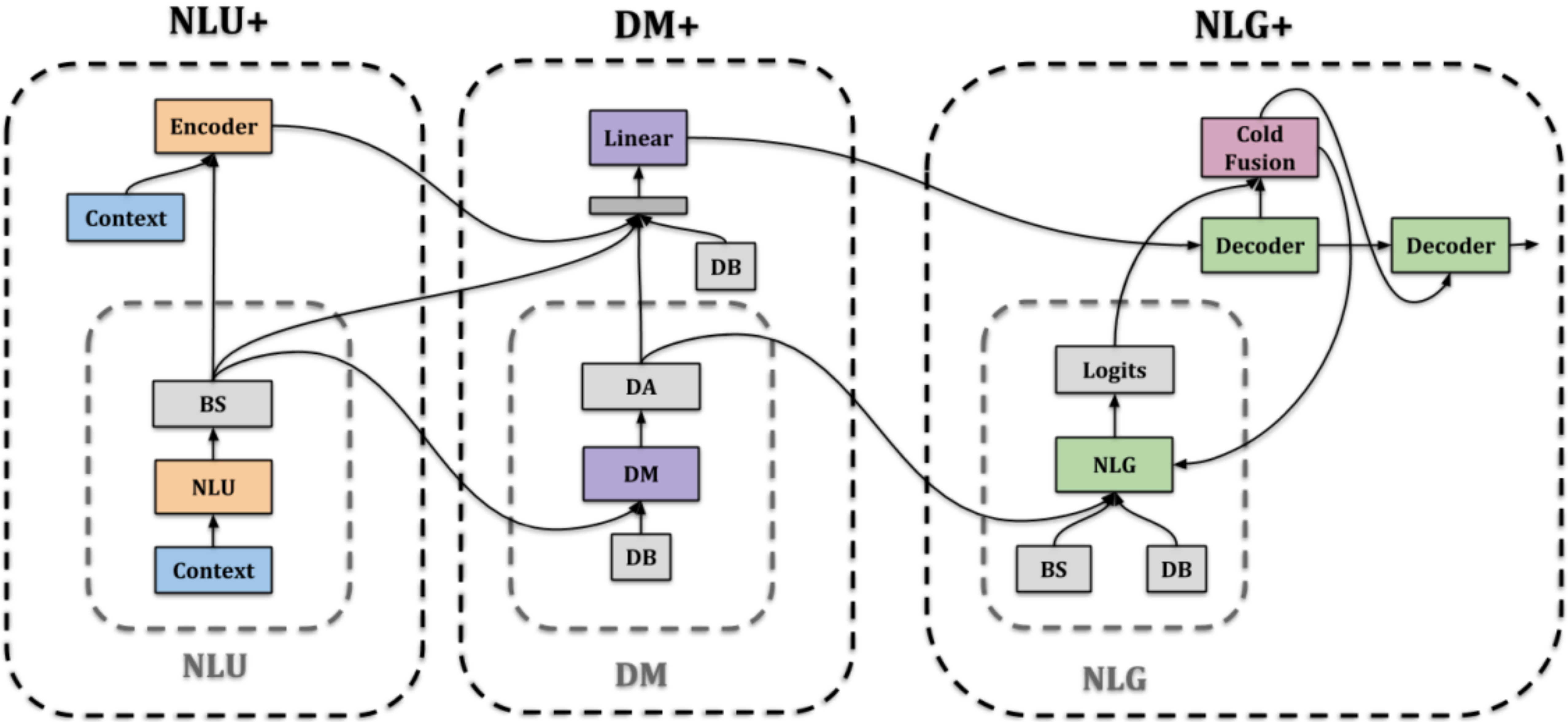
Structured Fusion Networks

(Mehri et al. 2019)



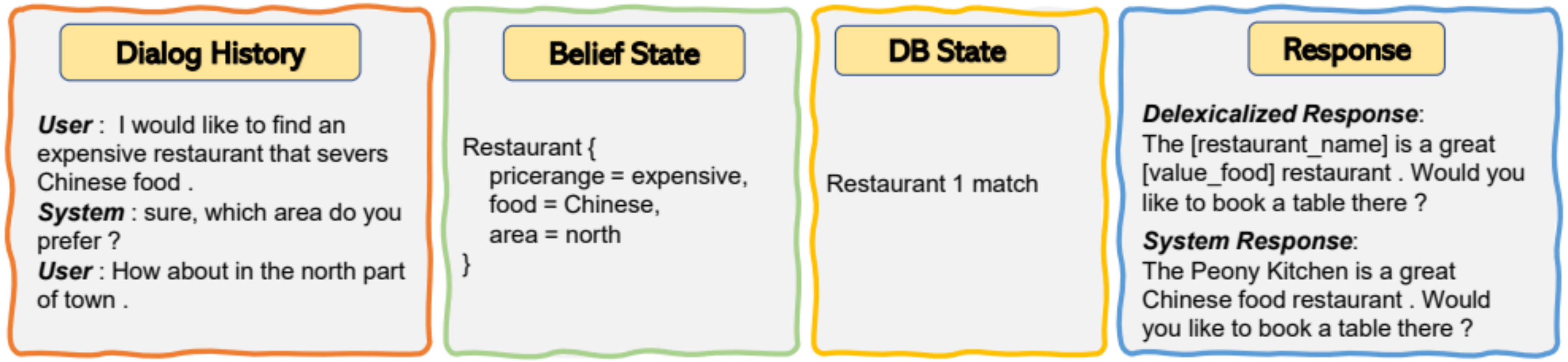
Structured Fusion Networks

(Mehri et al. 2019)



SOLOIST (Peng et al. 2020)

- Pre-train a LM on heterogeneous dialog corpora
- Adapt to new tasks with a handful of task-specific dialogs via machine teaching



Dialog History

User : I would like to find an expensive restaurant that severs Chinese food .
System : sure, which area do you prefer ?
User : How about in the north part of town .

Belief State

Restaurant {
pricerange = expensive,
food = Chinese,
area = north
}

DB State

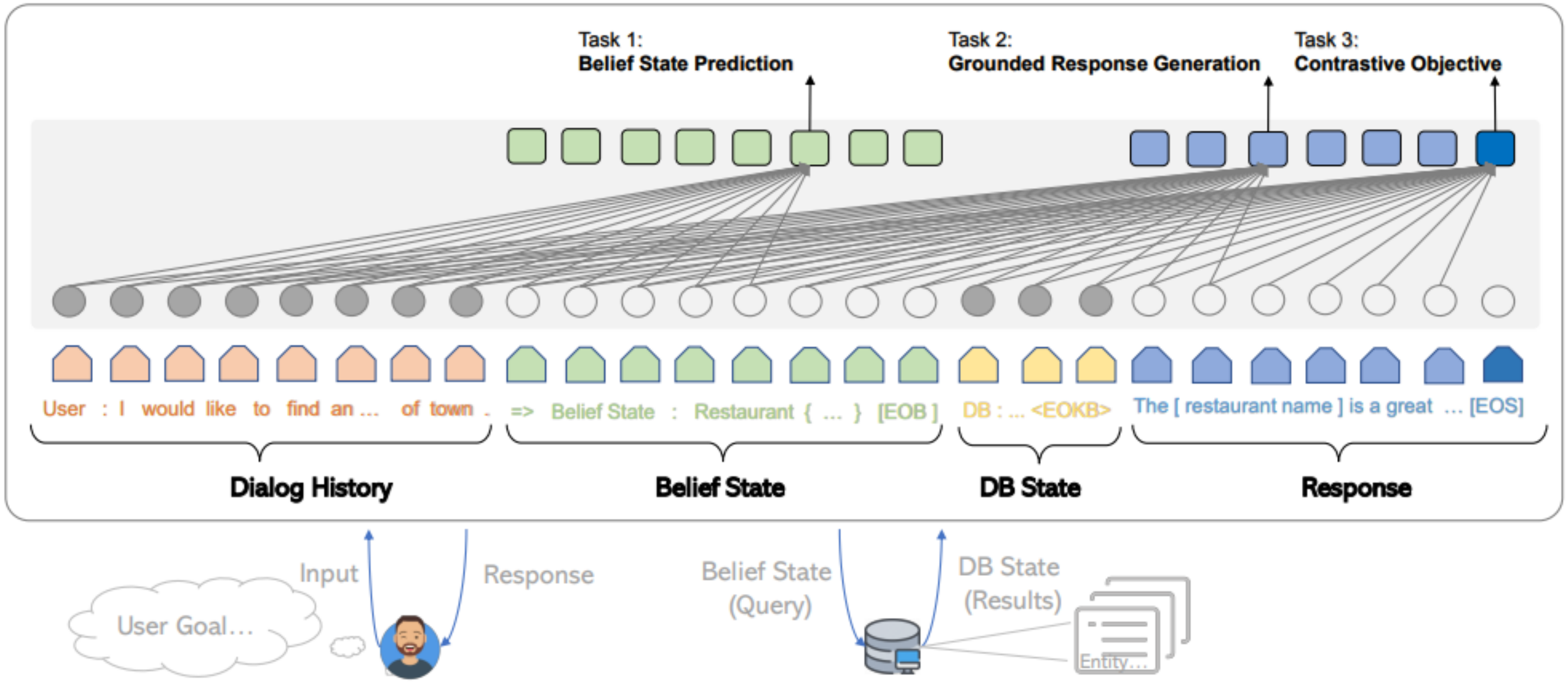
Restaurant 1 match

Response

Delexicalized Response:
The [restaurant_name] is a great [value_food] restaurant . Would you like to book a table there ?

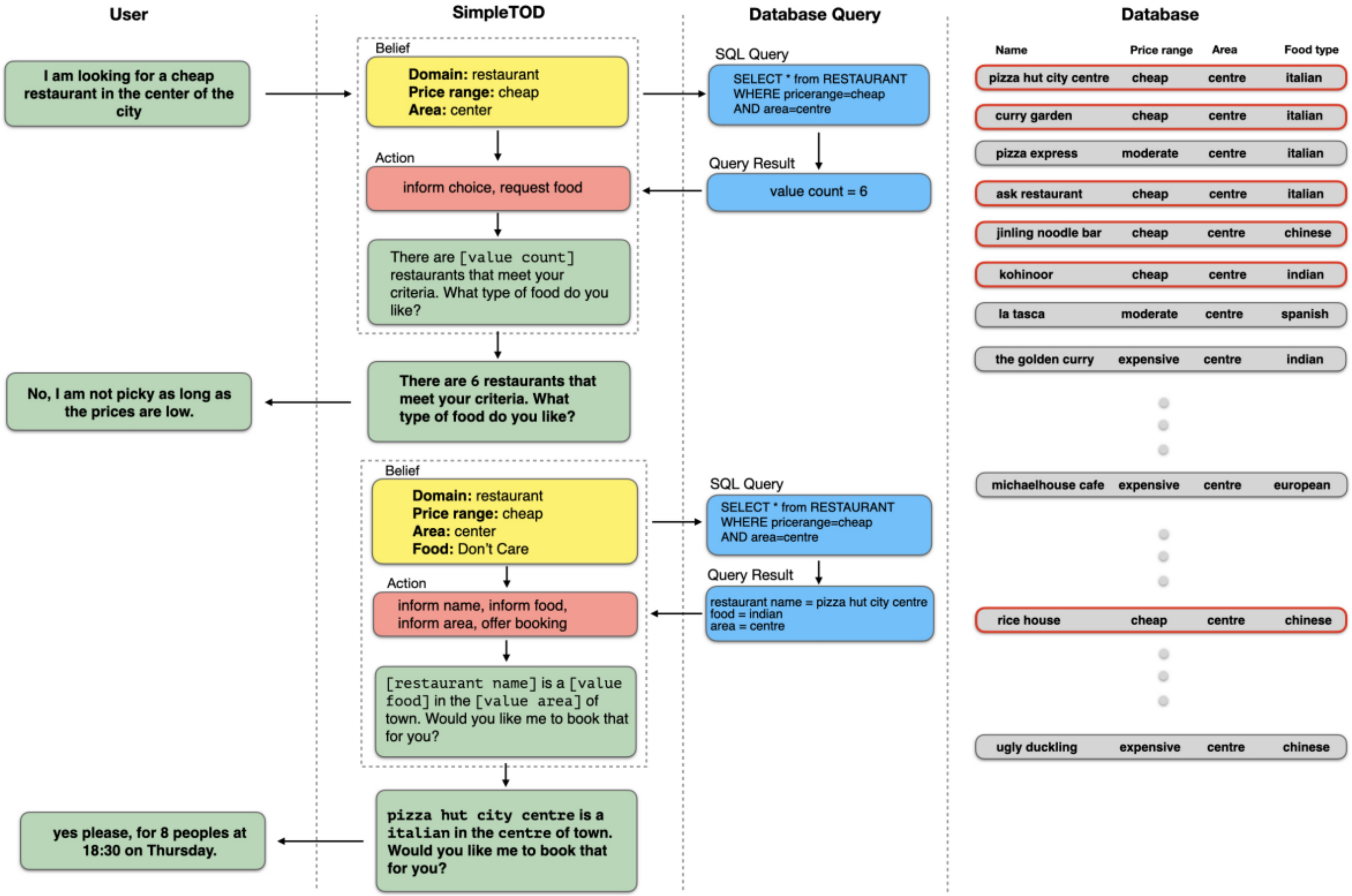
System Response:
The Peony Kitchen is a great Chinese food restaurant . Would you like to book a table there ?

(b) Example snippets for the items compounding the input of SOLOIST model.



SimpleTOD (Hosseini-Asl et al. 2020)

- Use a single LM (GPT2) to generate all outputs given the dialog context and retrieved database search results.



User

I am looking for a cheap restaurant in the center of the city

No, I am not picky as long as the prices are low.

yes please, for 8 peoples at 18:30 on Thursday.

SimpleTOD

Belief

Domain: restaurant
Price range: cheap
Area: center

Action

inform choice, request food

There are `[value count]` restaurants that meet your criteria. What type of food do you like?

There are 6 restaurants that meet your criteria. What type of food do you like?

Belief

Domain: restaurant
Price range: cheap
Area: center
Food: Don't Care

Action

inform name, inform food, inform area, offer booking

`[restaurant name]` is a `[value food]` in the `[value area]` of town. Would you like me to book that for you?

`pizza hut city centre` is a `italian` in the centre of town. Would you like me to book that for you?

Database Query

SQL Query

SELECT * from RESTAURANT
WHERE pricerange=cheap
AND area=centre

Query Result

value count = 6

SQL Query

SELECT * from RESTAURANT
WHERE pricerange=cheap
AND area=centre

Query Result

restaurant name = pizza hut city centre
food = indian
area = centre

Database

Name	Price range	Area	Food type
pizza hut city centre	cheap	centre	italian
curry garden	cheap	centre	italian
pizza express	moderate	centre	italian
ask restaurant	cheap	centre	italian
jinling noodle bar	cheap	centre	chinese
kohinoor	cheap	centre	indian
la tasca	moderate	centre	spanish
the golden curry	expensive	centre	indian
⋮			
michaelhouse cafe	expensive	centre	european
⋮			
rice house	cheap	centre	chinese
⋮			
ugly duckling	expensive	centre	chinese

Open-domain Dialog
(chit-chat)

Open-Domain Response Generation

Open-domain dialog systems must engage in chit-chat with a user

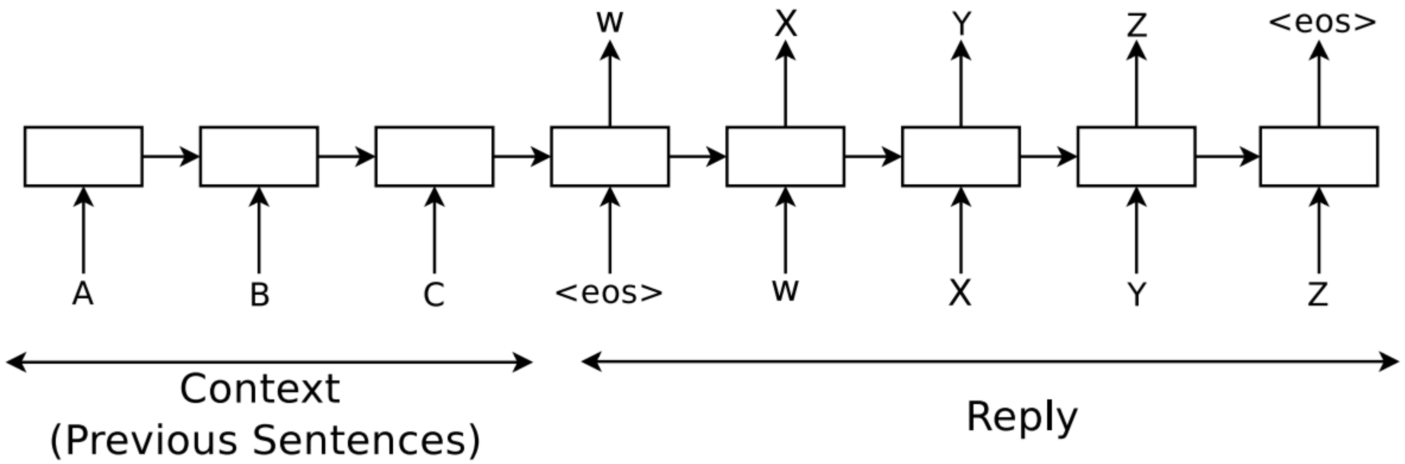
- DailyDialog [Li et al. 2017]
- PersonaChat [Zhang et al. 2018]
- Topical-Chat [Gopalakrishnan et al. 2019]
- Wizards of Wikipedia [Dinan et al. 2018]
- Empathetic Dialogs [Rashkin et al. 2019]

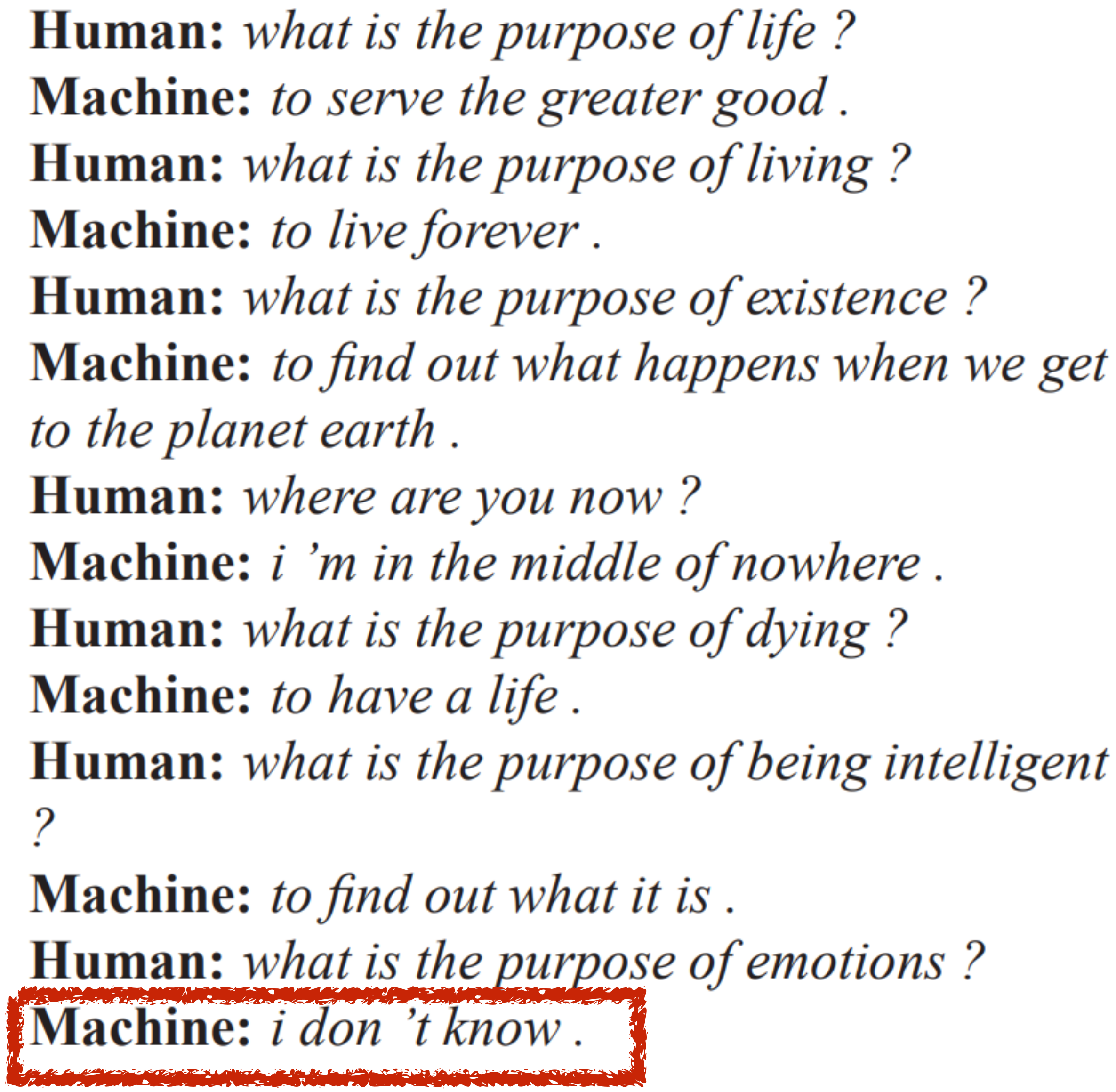
Open-Domain Response Generation

Open-domain dialog systems must engage in chit-chat with a user

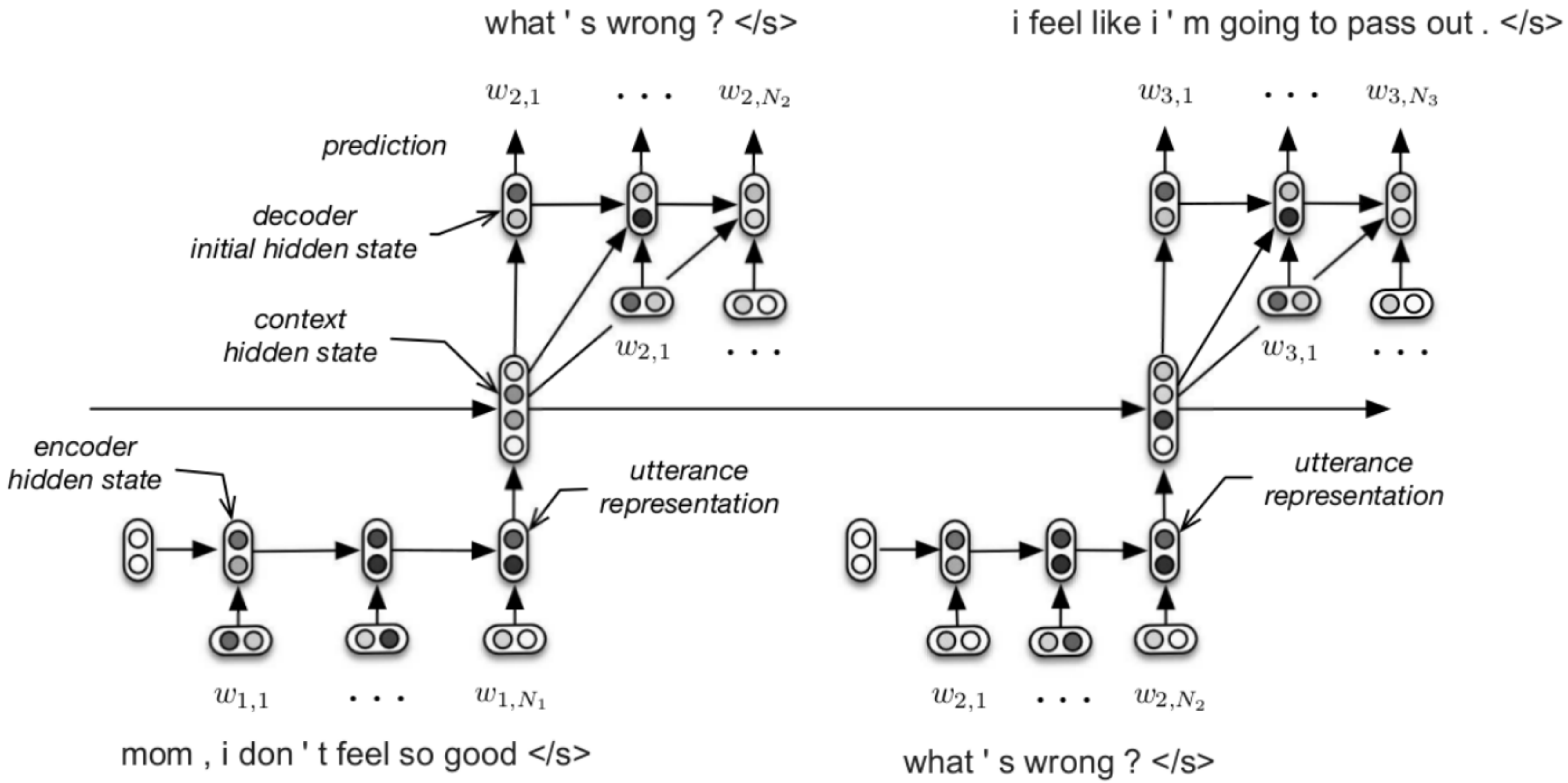
- Must understand the **dialog context**
- Must be able to discuss a variety of **topics**
- Must generate natural language **responses**
- Must generate **engaging/interesting** responses
- Must demonstrate **common sense reasoning**

Seq2Seq (Vinyals and Le. 2015)




Human: *what is the purpose of life ?*
Machine: *to serve the greater good .*
Human: *what is the purpose of living ?*
Machine: *to live forever .*
Human: *what is the purpose of existence ?*
Machine: *to find out what happens when we get to the planet earth .*
Human: *where are you now ?*
Machine: *i 'm in the middle of nowhere .*
Human: *what is the purpose of dying ?*
Machine: *to have a life .*
Human: *what is the purpose of being intelligent ?*
Machine: *to find out what it is .*
Human: *what is the purpose of emotions ?*
Machine: *i don 't know .*

HRED (Serban et al. 2016)


what ' s wrong ? </s>
i feel like i ' m going to pass out . </s>
$w_{2,1}$ . . . w_{2,N_2}
$w_{3,1}$ . . . w_{3,N_3}
prediction
decoder initial hidden state
context hidden state
$w_{2,1}$ . . .
$w_{3,1}$ . . .
encoder hidden state
utterance representation
utterance representation
$w_{1,1}$ . . . w_{1,N_1}
$w_{2,1}$ . . . w_{2,N_2}
mom , i don ' t feel so good </s>
what ' s wrong ? </s>

Diversity Promoting Objective

(Li et al. 2016)

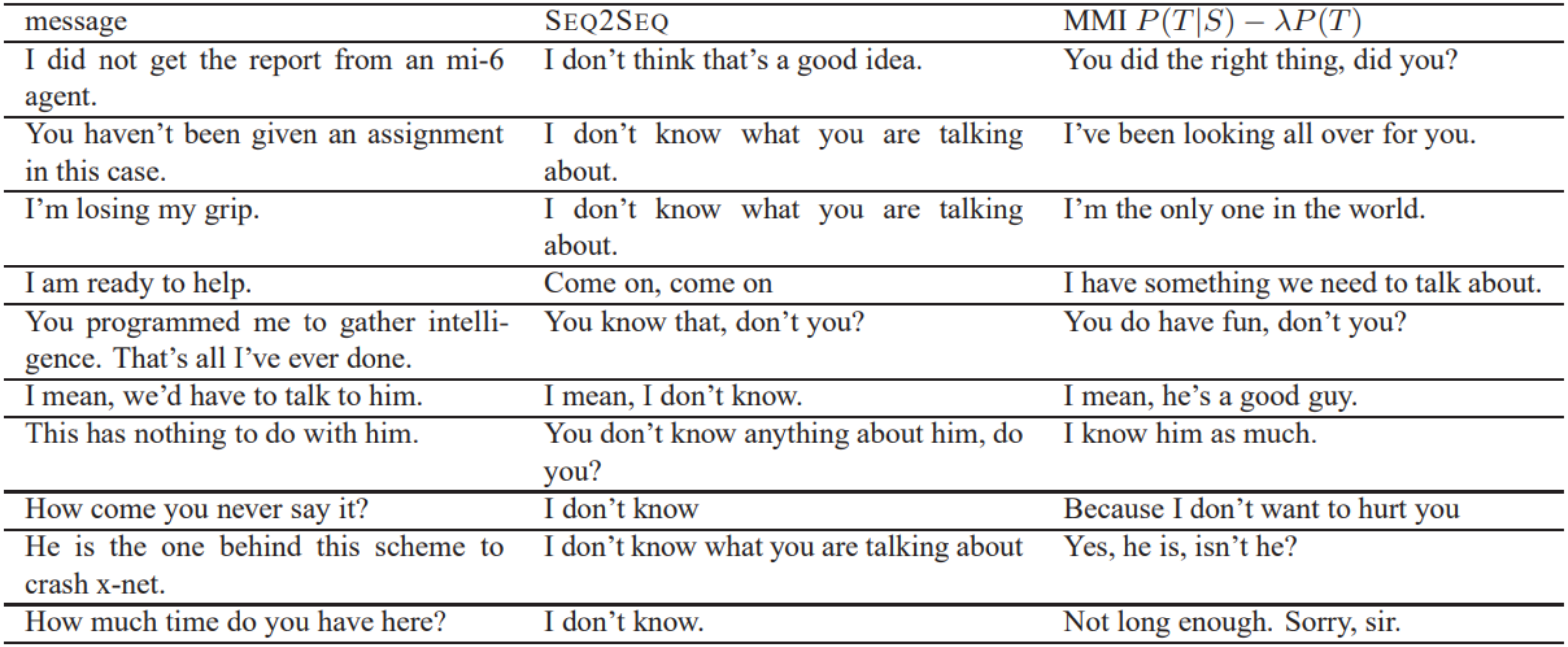
- To mitigate the **dull response problem** (“I don’t know”), Li et al. proposes a diversity-promoting objective function
  - Use Maximum Mutual Information (MMI) rather than cross entropy as a loss function
  - Penalize high-likelihood responses (anti LM objective)

Pairwise Mutual Information: $\log \frac{p(S,T)}{p(S)p(T)}$

MMI objective: $\hat{T} = \arg\max_{T} \left\{ \log p(T|S) - \lambda \log p(T) \right\}$

Diversity Promoting Objective

(Li et al. 2016)

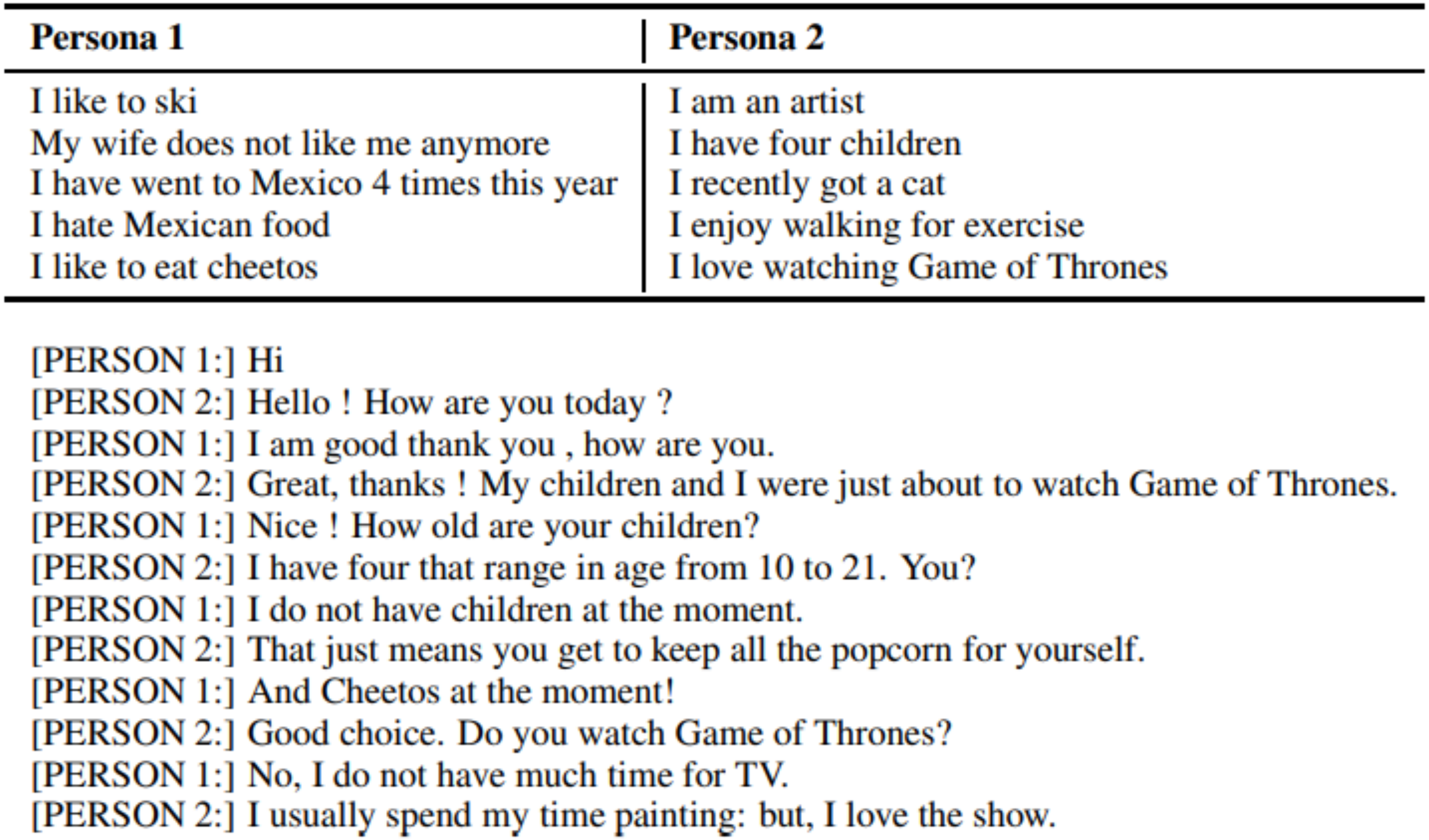


message	SEQ2SEQ	MMI $P(T\|S) - \lambda P(T)$
I did not get the report from an mi-6 agent.	I don't think that's a good idea.	You did the right thing, did you?
You haven't been given an assignment in this case.	I don't know what you are talking about.	I've been looking all over for you.
I'm losing my grip.	I don't know what you are talking about.	I'm the only one in the world.
I am ready to help.	Come on, come on	I have something we need to talk about.
You programmed me to gather intelligence. That's all I've ever done.	You know that, don't you?	You do have fun, don't you?
I mean, we'd have to talk to him.	I mean, I don't know.	I mean, he's a good guy.
This has nothing to do with him.	You don't know anything about him, do you?	I know him as much.
How come you never say it?	I don't know	Because I don't want to hurt you
He is the one behind this scheme to crash x-net.	I don't know what you are talking about	Yes, he is, isn't he?
How much time do you have here?	I don't know.	Not long enough. Sorry, sir.

Persona-Conditioned Models

(Zhang et al. 2018)

- To make open-domain chit-chat dialog models more consistent and engaging, condition them on a persona

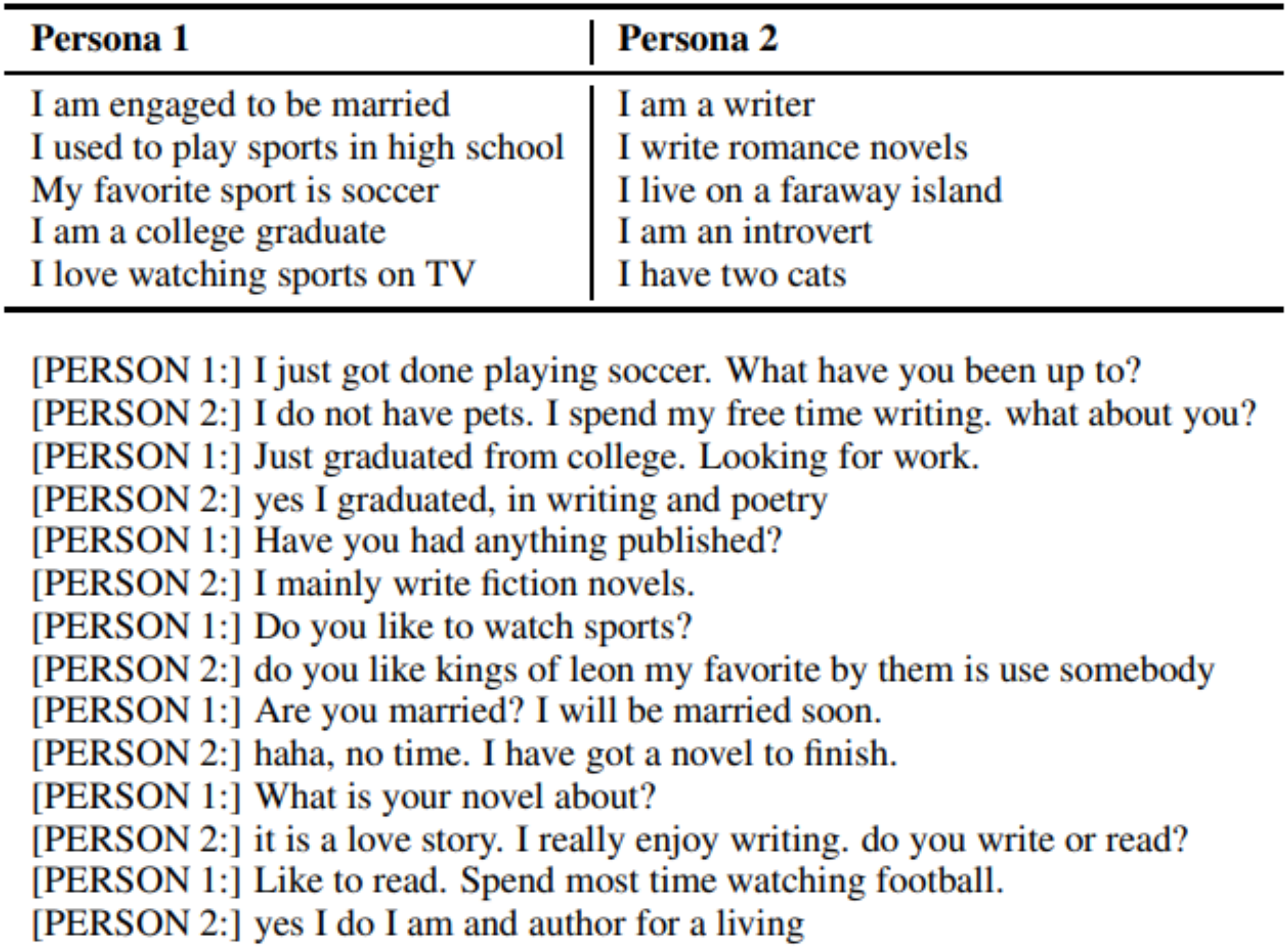


Persona 1	**Persona 2**
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Persona-Conditioned Models

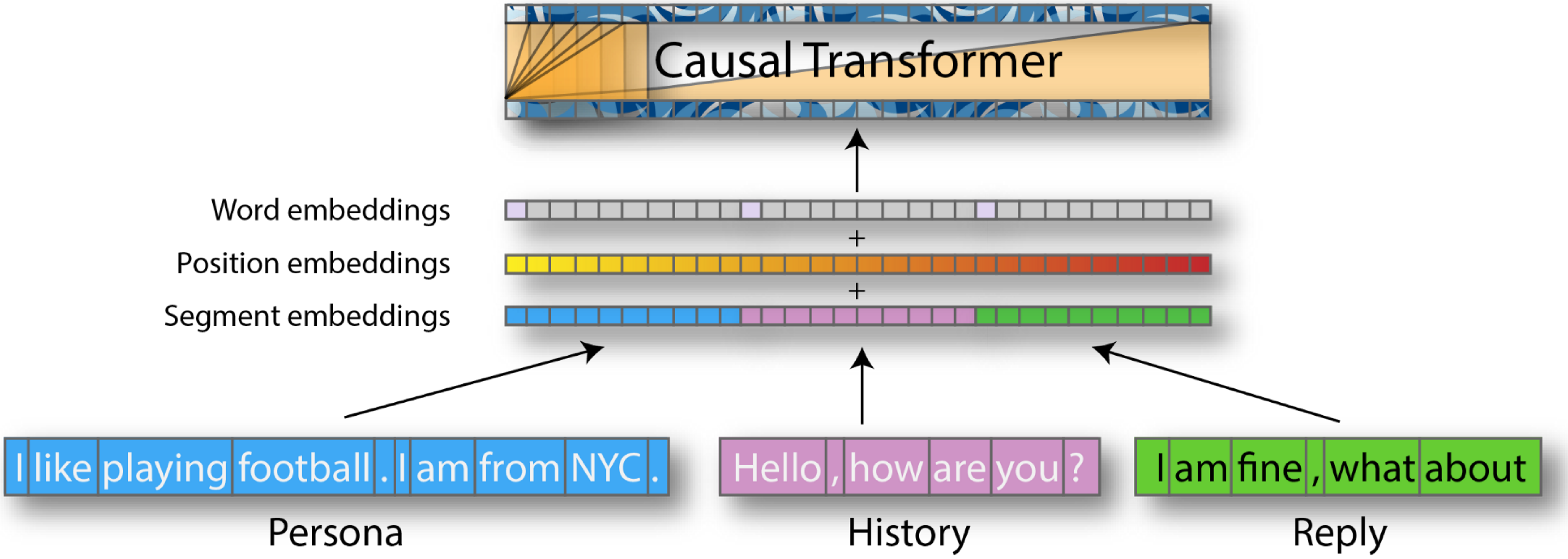
(Zhang et al. 2018)



Persona 1	**Persona 2**
I am engaged to be married	I am a writer
I used to play sports in high school	I write romance novels
My favorite sport is soccer	I live on a faraway island
I am a college graduate	I am an introvert
I love watching sports on TV	I have two cats

[PERSON 1:] I just got done playing soccer. What have you been up to?
[PERSON 2:] I do not have pets. I spend my free time writing. what about you?
[PERSON 1:] Just graduated from college. Looking for work.
[PERSON 2:] yes I graduated, in writing and poetry
[PERSON 1:] Have you had anything published?
[PERSON 2:] I mainly write fiction novels.
[PERSON 1:] Do you like to watch sports?
[PERSON 2:] do you like kings of leon my favorite by them is use somebody
[PERSON 1:] Are you married? I will be married soon.
[PERSON 2:] haha, no time. I have got a novel to finish.
[PERSON 1:] What is your novel about?
[PERSON 2:] it is a love story. I really enjoy writing. do you write or read?
[PERSON 1:] Like to read. Spend most time watching football.
[PERSON 2:] yes I do I am and author for a living

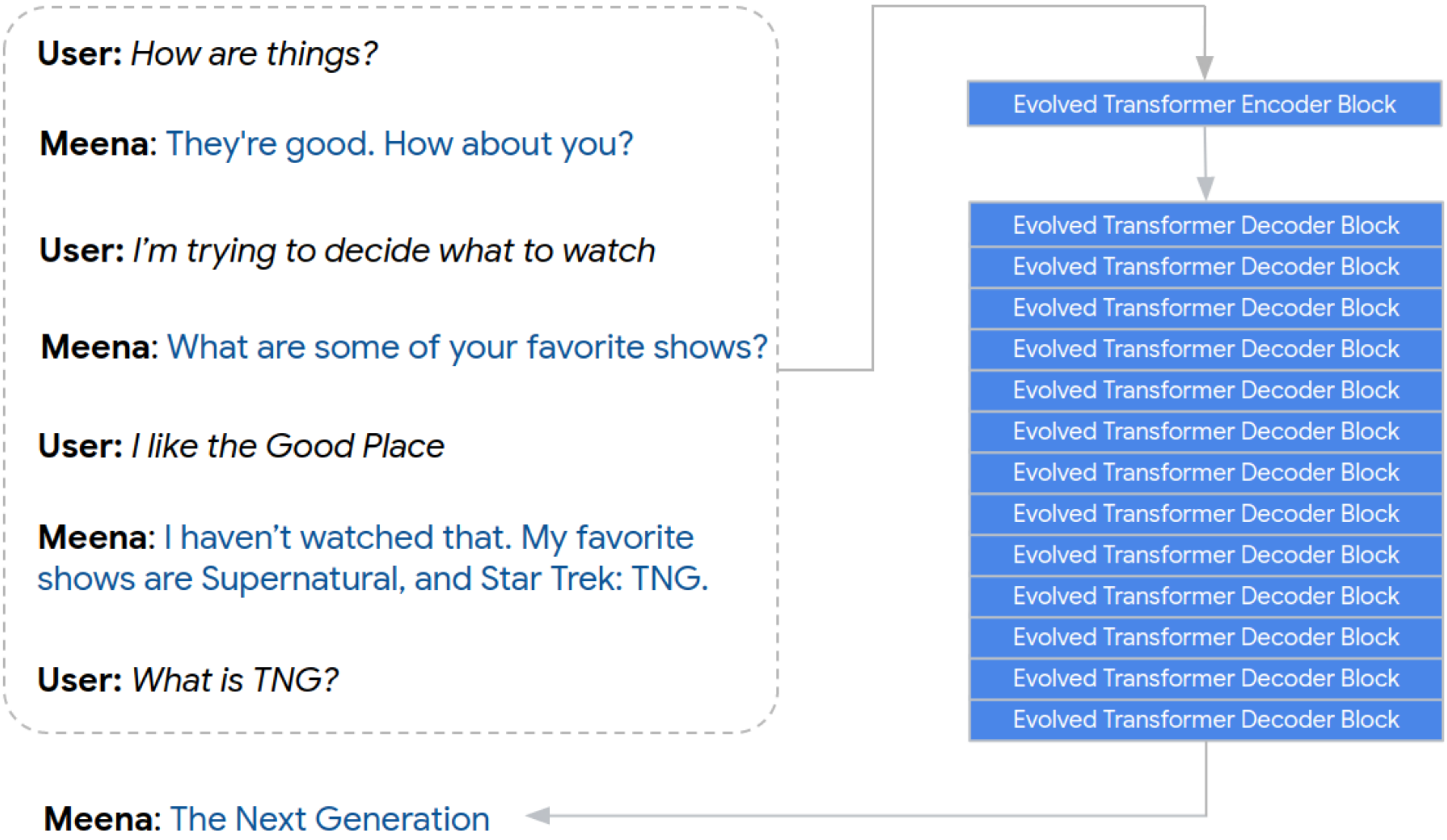
Transfer-Transfo (Wolf et al. 2019)



DialoGPT (Zhang et al. 2019)

- Continue pre-training GPT-2 on **conversations from Reddit**
  - Filter long utterances
  - Filter non-English utterances
  - Filter URLs
  - Filter toxic comments
- Train on 147M dialog instances (1.8B words)
- “Human-level” response generation ability

Meena (Adiwardana et al. 2020)



User: *How are things?*

Meena: They're good. How about you?

User: *I'm trying to decide what to watch*

Meena: What are some of your favorite shows?

User: *I like the Good Place*

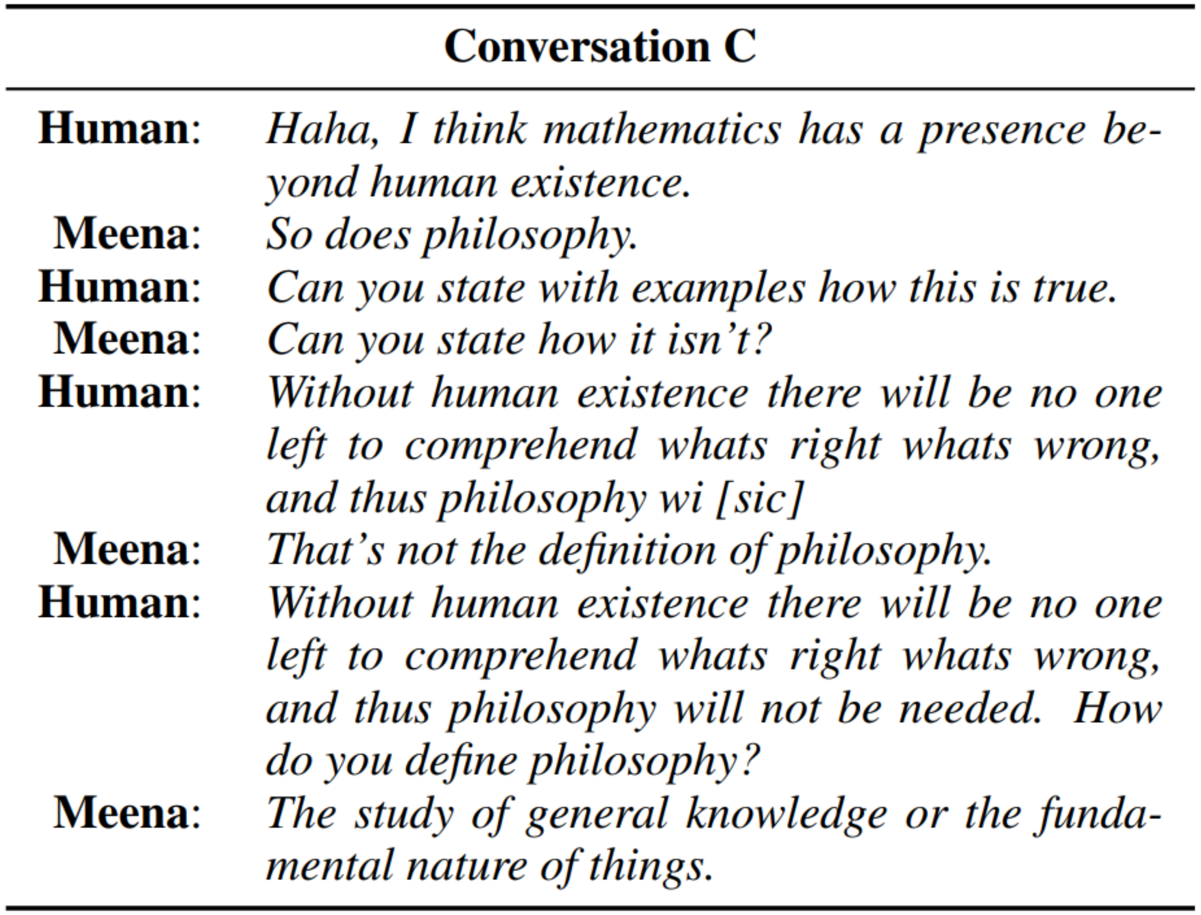
Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*

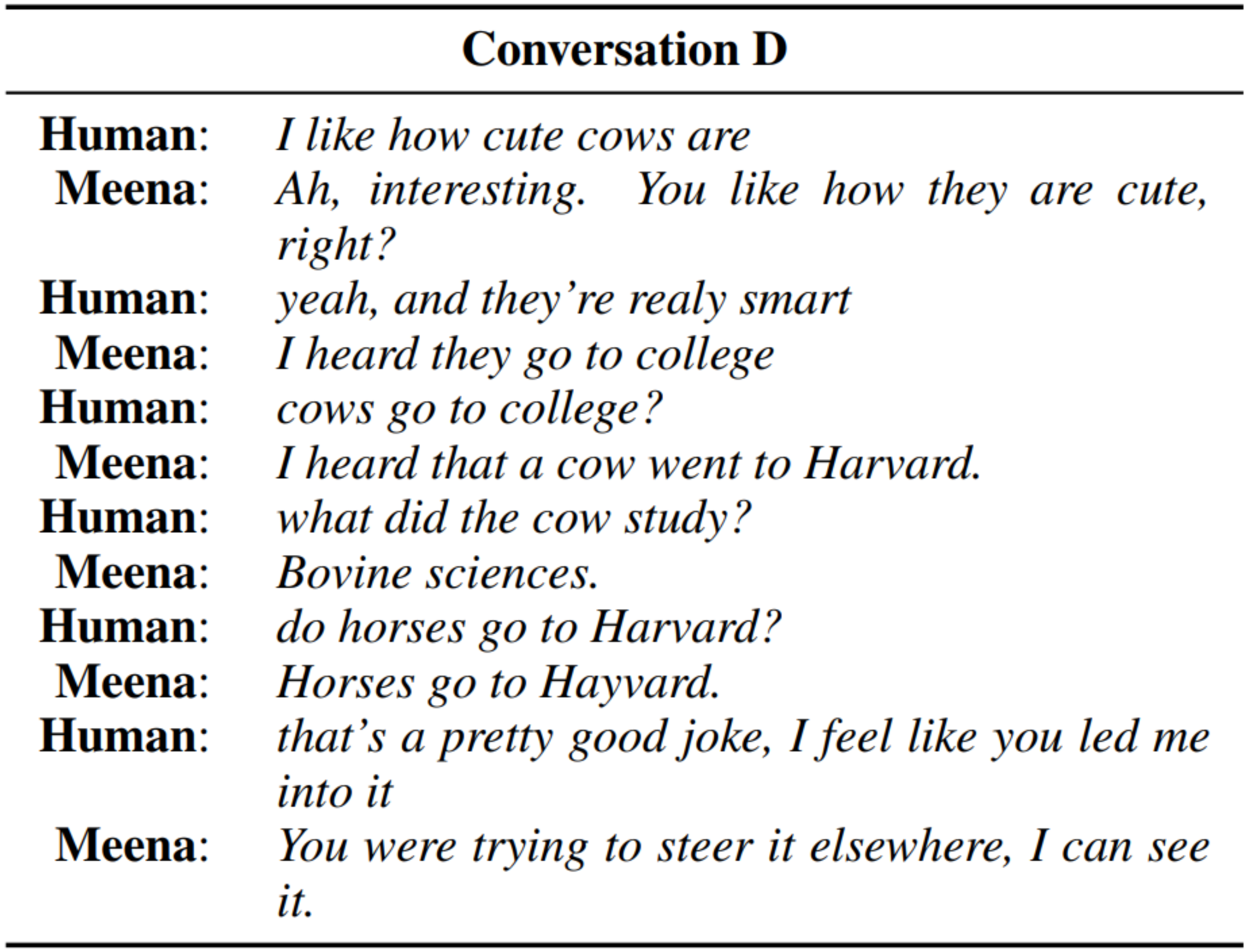
Meena: The Next Generation

Evolved Transformer Encoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block
Evolved Transformer Decoder Block

Meena (Adiwardana et al. 2020)



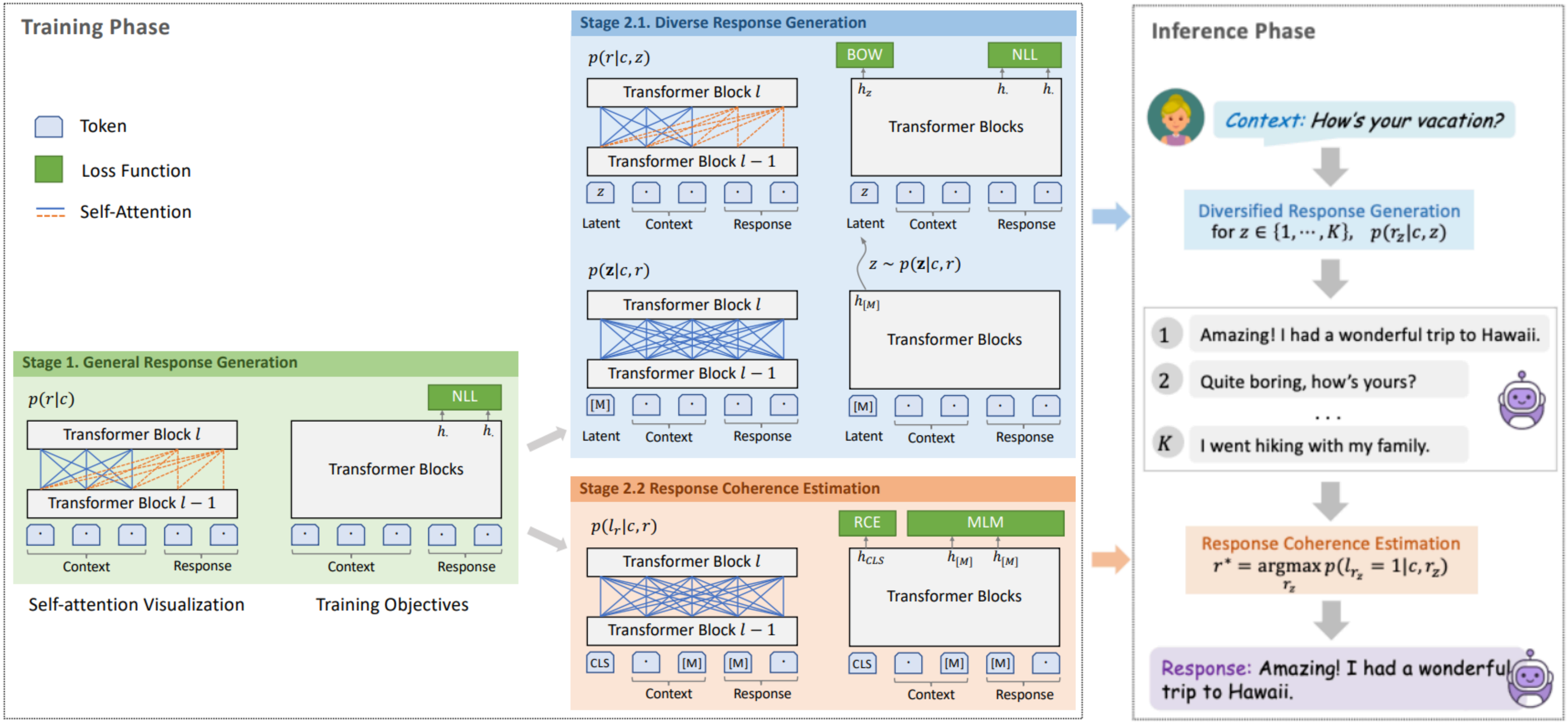
Conversation C	
Human:	*Haha, I think mathematics has a presence beyond human existence.*
Meena:	*So does philosophy.*
Human:	*Can you state with examples how this is true.*
Meena:	*Can you state how it isn't?*
Human:	*Without human existence there will be no one left to comprehend whats right whats wrong, and thus philosophy wi [sic]*
Meena:	*That's not the definition of philosophy.*
Human:	*Without human existence there will be no one left to comprehend whats right whats wrong, and thus philosophy will not be needed. How do you define philosophy?*
Meena:	*The study of general knowledge or the fundamental nature of things.*



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

PLATO-2 (Bao et al. 2021)

- Stage 1: Generate coarse response
- Stage 2: Diversify & Selection


Training Phase
Token
Loss Function
Self-Attention
Stage 1. General Response Generation
$p(r|c)$
Transformer Block l
Transformer Block $l-1$
Context
Response
NLL
$h_.$
Transformer Blocks
Self-attention Visualization
Training Objectives
Stage 2.1. Diverse Response Generation
$p(r|c,z)$
BOW
h_z
z
Latent
$p(\mathbf{z}|c,r)$
$z \sim p(\mathbf{z}|c,r)$
$h_{[M]}$
[M]
Stage 2.2 Response Coherence Estimation
$p(l_r|c,r)$
RCE
MLM
h_{CLS}
CLS
Inference Phase
Diversified Response Generation
for $z \in \{1,\cdots,K\}$, $p(r_z|c,z)$
Response Coherence Estimation
$r^* = \operatorname{argmax}_{r_z} p(l_{r_z}=1|c,r_z)$

Context: How's your vacation?

1 Amazing! I had a wonderful trip to Hawaii.

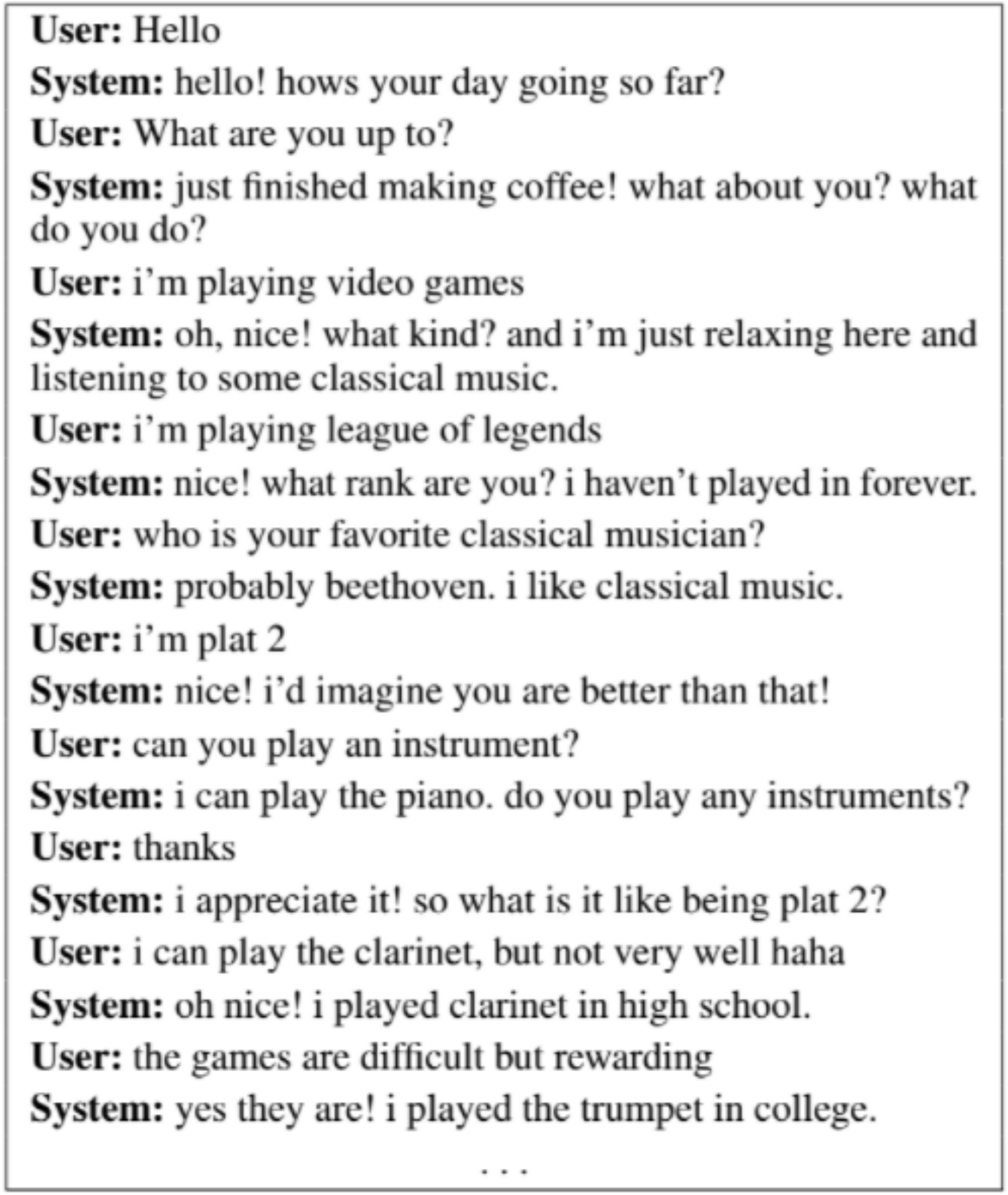
2 Quite boring, how's yours?

...

K I went hiking with my family.

Response: Amazing! I had a wonderful trip to Hawaii.

PLATO-2 (Bao et al. 2021)



User: Hello

System: hello! hows your day going so far?

User: What are you up to?

System: just finished making coffee! what about you? what do you do?

User: i'm playing video games

System: oh, nice! what kind? and i'm just relaxing here and listening to some classical music.

User: i'm playing league of legends

System: nice! what rank are you? i haven't played in forever.

User: who is your favorite classical musician?

System: probably beethoven. i like classical music.

User: i'm plat 2

System: nice! i'd imagine you are better than that!

User: can you play an instrument?

System: i can play the piano. do you play any instruments?

User: thanks

System: i appreciate it! so what is it like being plat 2?

User: i can play the clarinet, but not very well haha

System: oh nice! i played clarinet in high school.

User: the games are difficult but rewarding

System: yes they are! i played the trumpet in college.

. . .

Open-Domain Response Generation

- **Knowledge-grounded** response generation
- **Persona-grounded** response generation
- **Negotiation/persuasive** dialog
- **Commonsense** dialog
- Conversational **QA**

Dialog Evaluation

Dialog Evaluation

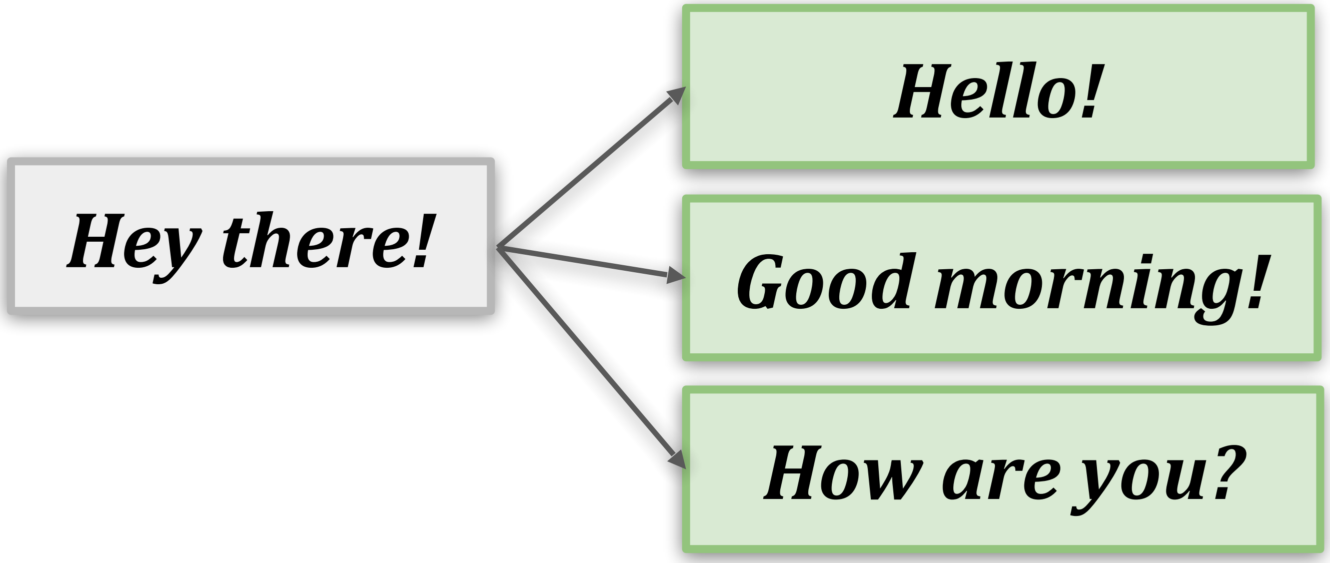
- **Goal:** Construct automatic evaluation metrics for response generation/interactive dialog

- **Given:** dialog history, generated response, reference response (optional)
- **Output:** a score for the response

Why is evaluating dialog hard? (1/3)

1. **One-to-many** nature of dialog

- For each dialog there are **many valid responses**
- Cannot compare to a reference response
  - The reference response isn't the only valid response
- Existing metrics won't work
  - BLEU, F-1, etc.



Hey there!

Hello!

Good morning!

How are you?

Why is evaluating dialog hard? (2/3)

2. Dialog quality is **multi-faceted**

- A response isn't just **good** or **bad**
- For interpretability, should measure **multiple qualities**
    - Relevance
    - Interestingness
    - Fluency

Why is evaluating dialog hard? (3/3)

3. Dialog is **inherently interactive**

- Dialog systems are designed to have a **back-and-forth interaction** with a user
    - Research largely focuses on **static corpora** → Reduces the problem of dialog to **response generation**
- Some properties of a system can't be assessed outside an interactive environment
    - Long-term planning, error recovery, coherence.

Dialog Evaluation

- Evaluation of dialog is **hard**
    - Can’t compare to a reference response [no BLEU, F-1, etc.]
    - Should assess **many aspects** of dialog quality [relevant, interesting, etc.]
    - Should evaluate in an **interactive** manner

Dialog Evaluation

- USR [Mehri and Eskenazi. 2020]
- GRADE [Huang et al. 2020]
- HolisticEval [Pang et al. 2020]
- DSTC6 [Hori and Hori. 2017]
- FED [Mehri and Eskenazi. 2020]
- DSTC9 [Gunasekara et al. 2021]

https://github.com/exe1023/DialEvalMetrics
A Comprehensive Assessment of Dialog Evaluation Metrics
Yi-Ting Yeh, Maxine Eskenazi, Shikib Mehri

Questions?